Challenges in Building Large-Scale Information Retrieval Systems

Jeff Dean
Google Senior Fellow
jeff@google.com

Why Work on Retrieval Systems?

- Challenging blend of science and engineering
  - Many interesting, unsolved problems
  - Spans many areas of CS:
    - architecture, distributed systems, algorithms, compression, information retrieval, machine learning, UI, etc.
  - Scale far larger than most other systems

- Small teams can create systems used by hundreds of millions

Retrieval System Dimensions

- Must balance engineering tradeoffs between:
  - number of documents indexed
  - queries / sec
  - index freshness/update rate
  - query latency
  - information kept about each document
  - complexity/cost of scoring/retrieval algorithms
- Engineering difficulty roughly equal to the product of these parameters
- All of these affect overall performance, and performance per $

1999 vs. 2014

• # docs: ~70M to many billion	~100X
• queries processed/day:	~1000X
• per doc info in index:	~3X
• update latency: months to minutes	~10000X
• avg. query latency: <1s to <0.15s	~6X
• More machines * faster machines:	~1000X

Constant Change

- Parameters change over time
  - often by many orders of magnitude
- Right design at X may be very wrong at 10X or 100X
  - ... design for ~10X growth, but plan to rewrite before ~100X
- Continuous evolution:
  - 7 significant revisions in last 10 years
  - often rolled out without users realizing we've made major changes

Rest of Talk

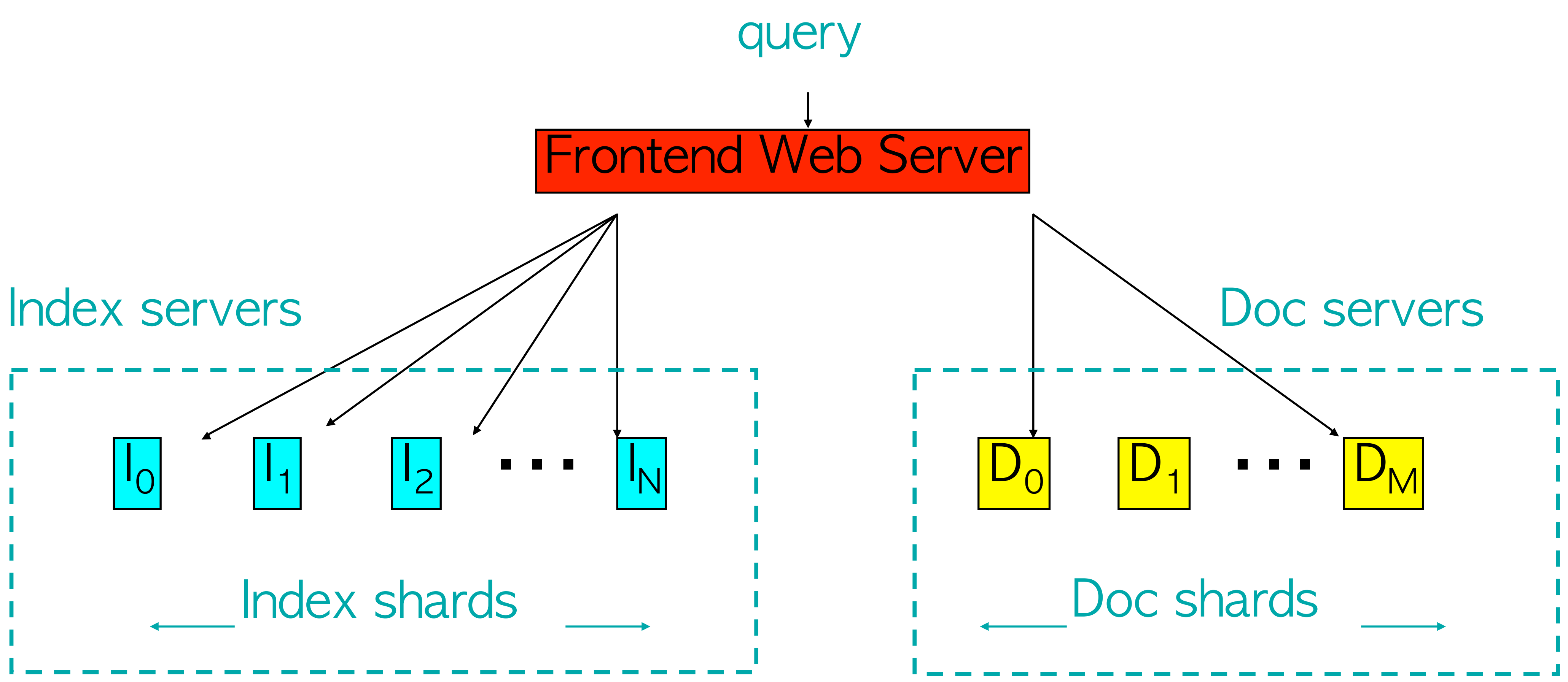
- Evolution of Google's search systems
  - several gens of crawling/indexing/serving systems
  - brief description of supporting infrastructure

  - Joint work with many, many people

- Interesting directions and challenges

“Google” Circa 1997 (google.stanford.edu)

Research Project, circa 1997



Ways of Index Partitioning

- **By doc**: each shard has index for subset of docs
  - pro: each shard can process queries independently
  - pro: easy to keep additional per-doc information
  - pro: network traffic (requests/responses) small
  - con: query has to be processed by each shard
  - con: O(K*N) disk seeks for K word query on N shards

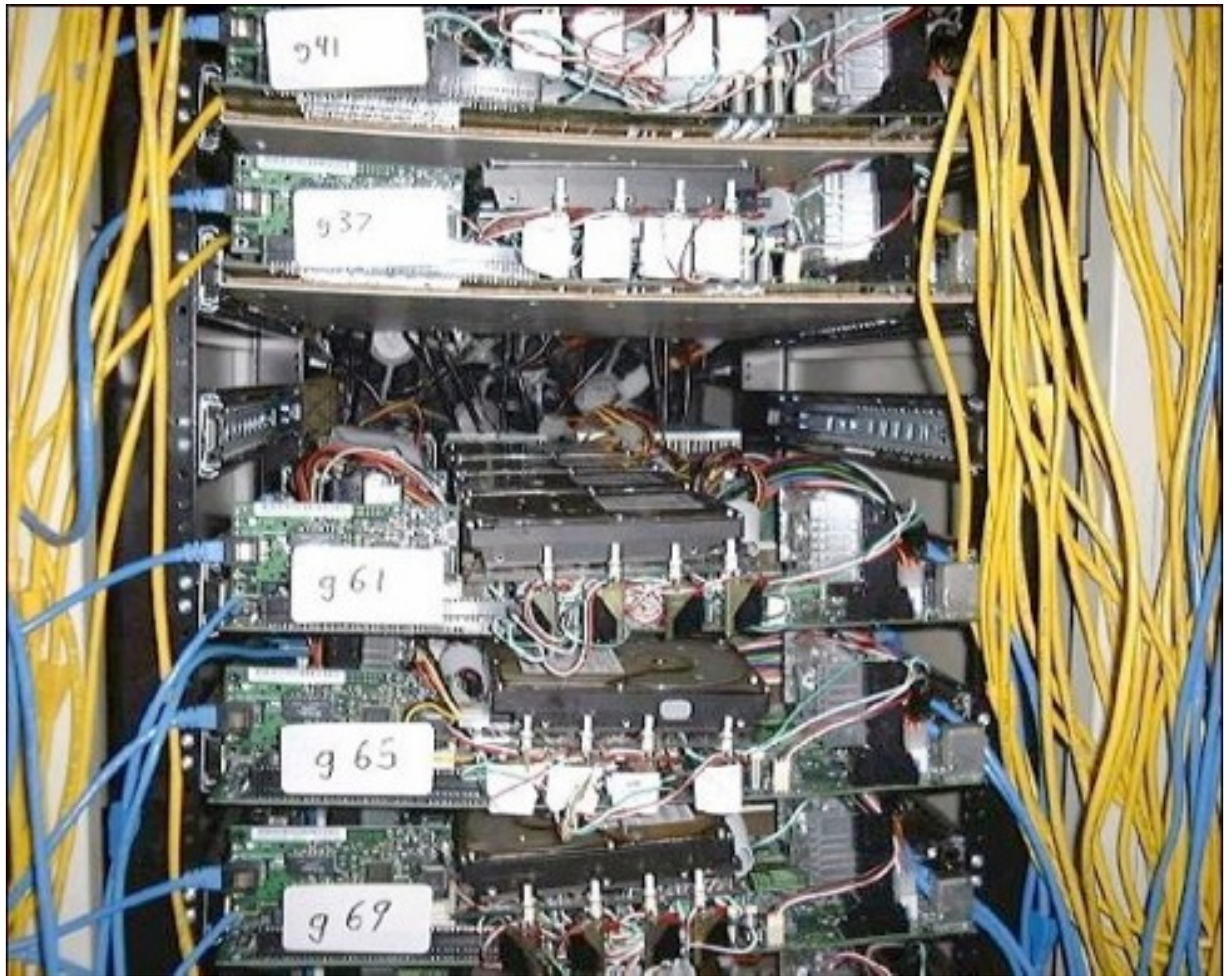
  In our computing environment, by doc makes more sense

- **By word**: shard has subset of words for all docs
  - pro: K word query => handled by at most K shards
  - pro: O(K) disk seeks for K word query
  - con: much higher network bandwidth needed
    - data about each word for each matching doc must be collected in one place
  - con: harder to have per-doc information

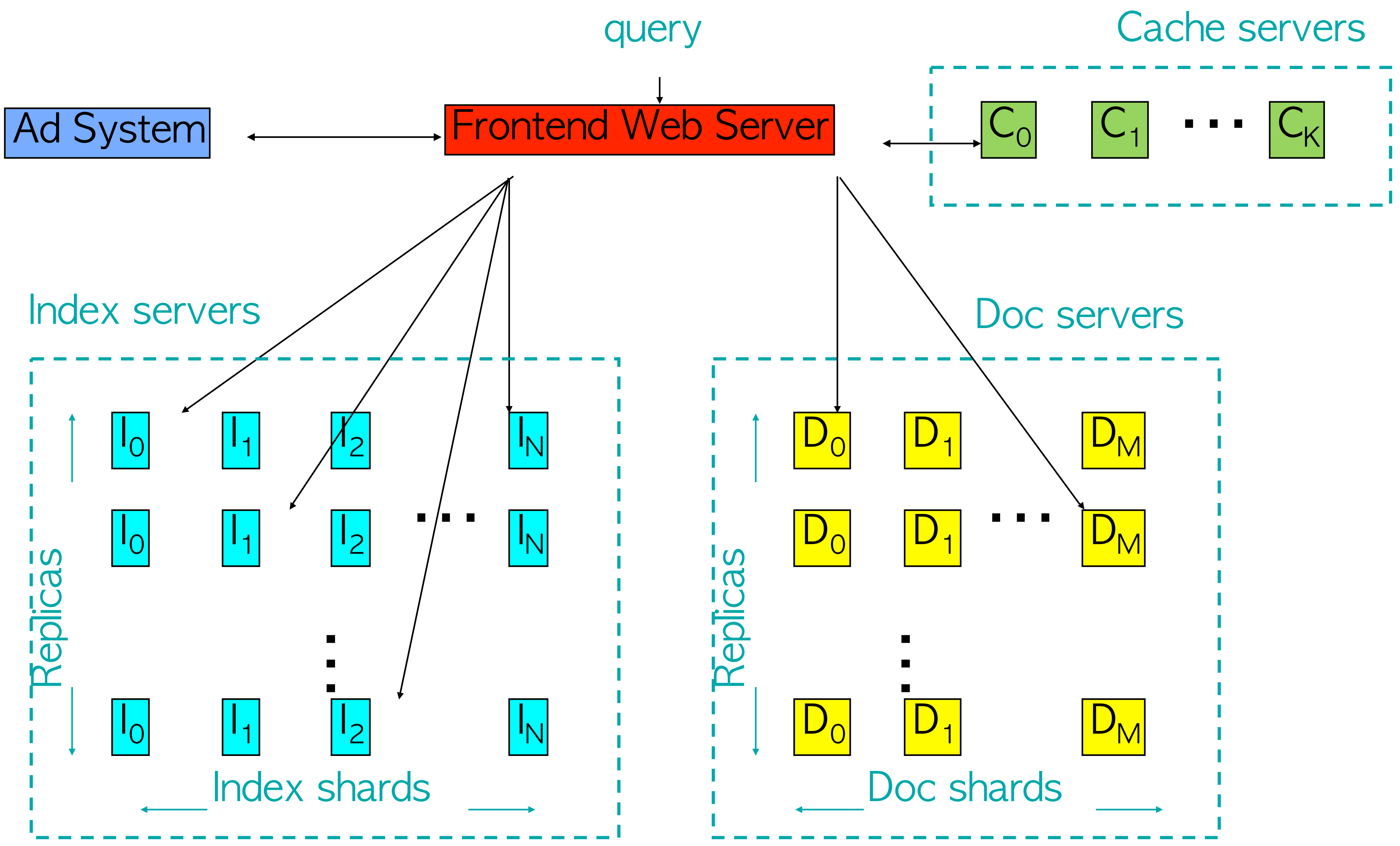
Basic Principles

- Documents assigned small integer ids (docids)
  - good if smaller for higher quality/more important docs
- Index Servers:
  - given (query) return sorted list of (score, docid, ...)
  - partitioned (“sharded”) by docid
  - index shards are replicated for capacity
  - cost is O(# queries * # docs in index)
- Doc Servers
  - given (docid, query) generate (title, snippet)
  - map from docid to full text of docs on disk
  - also partitioned by docid
  - cost is O(# queries)

“Corkboards” (1999)



Serving System, circa 1999



Caching

- Cache servers:
  - cache both index results and doc snippets
  - hit rates typically 30-60%
    - depends on frequency of index updates, mix of query traffic, level of personalization, etc
- Main benefits:
  - performance! 10s of machines do work of 100s or 1000s
  - reduce query latency on hits
    - queries that hit in cache tend to be both popular and expensive (common words, lots of documents to score, etc.)
- Beware: big latency spike/capacity drop when index updated or cache flushed

Crawling (circa 1998-1999)

- Simple batch crawling system
  - start with a few URLs
  - crawl pages
  - extract links, add to queue
  - stop when you have enough pages

- Concerns:
  - don’t hit any site too hard
  - prioritizing among uncrawled pages
    - one way: continuously compute PageRank on changing graph
  - maintaining uncrawled URL queue efficiently
    - one way: keep in a partitioned set of servers
  - dealing with machine failures

Indexing (circa 1998-1999)

- Simple batch indexing system
  - Based on simple unix tools
  - No real checkpointing, so machine failures painful
  - No checksumming of raw data, so hardware bit errors caused problems
    - Exacerbated by early machines having no ECC, no parity
    - Sort 1 TB of data without parity: ends up "mostly sorted"
    - Sort it again: "mostly sorted" another way

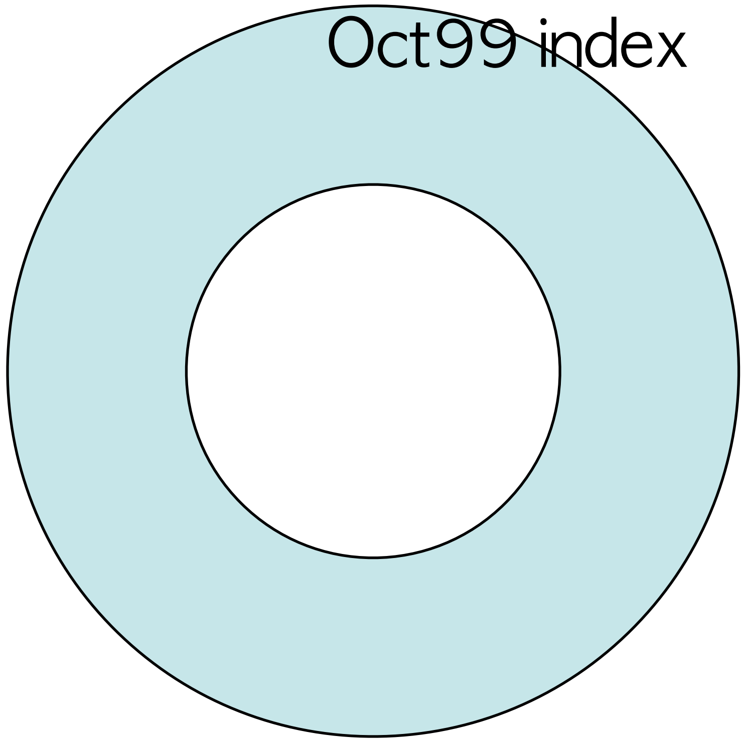
- “Programming with adversarial memory”
  - Led us to develop a file abstraction that stored checksums of small records and could skip and resynchronize after corrupted records

Index Updates (circa 1998-1999)

- 1998-1999: Index updates (~once per month):
  - Wait until traffic is low
  - Take some replicas offline
  - Copy new index to these replicas
  - Start new frontends pointing at updated index and serve some traffic from there

Index Updates (circa 1998-1999)

- Index server disk:
  - outer part of disk gives higher disk bandwidth

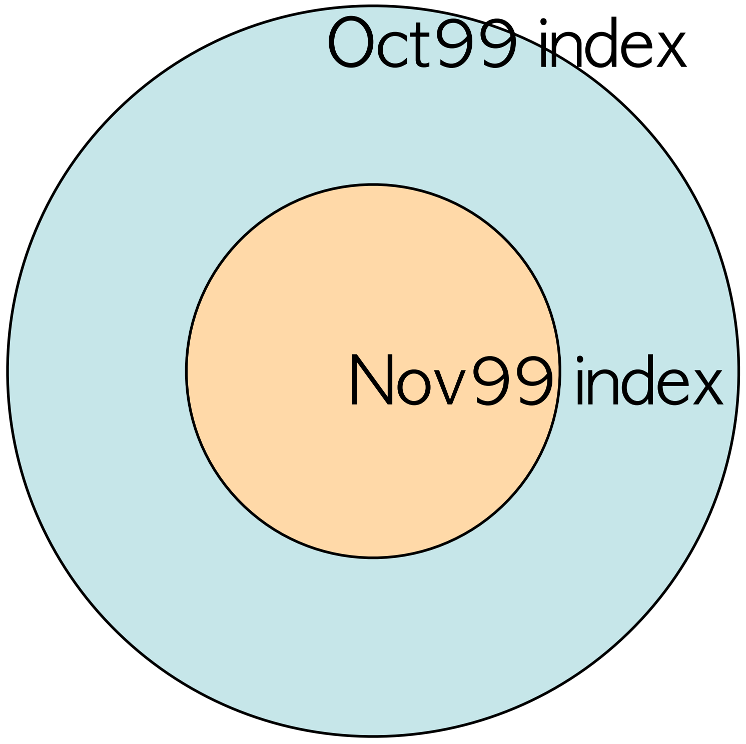


Index Updates (circa 1998-1999)

- Index server disk:
  - outer part of disk gives higher disk bandwidth

1. Copy new index to inner half of disk (while still serving old index)

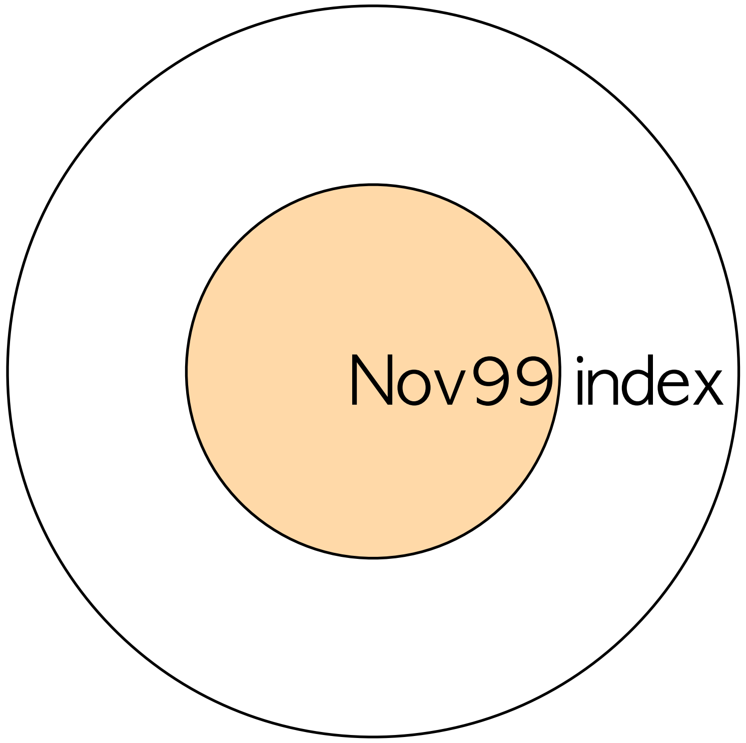
2. Restart to use new index



Index Updates (circa 1998-1999)

- Index server disk:
  - outer part of disk gives higher disk bandwidth

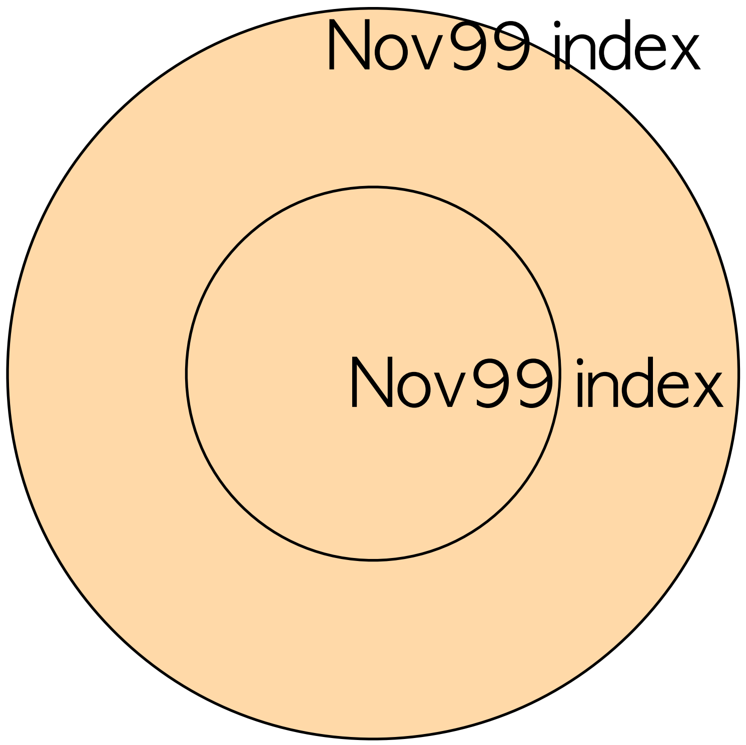
1. Copy new index to inner half of disk (while still serving old index)
2. Restart to use new index
3. Wipe old index



Index Updates (circa 1998-1999)

- Index server disk:
  - outer part of disk gives higher disk bandwidth

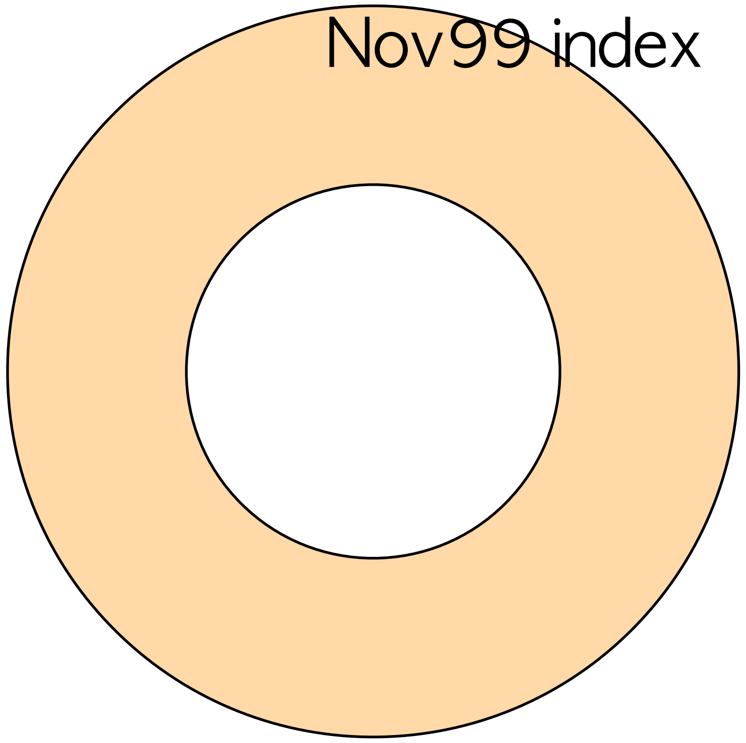
1. Copy new index to inner half of disk (while still serving old index)
2. Restart to use new index
3. Wipe old index
4. Re-copy new index to faster half of disk



Index Updates (circa 1998-1999)

- Index server disk:
  - outer part of disk gives higher disk bandwidth

1. Copy new index to inner half of disk (while still serving old index)
2. Restart to use new index
3. Wipe old index
4. Re-copy new index to faster half of disk
5. Wipe first copy of new index



Index Updates (circa 1998-1999)

- Index server disk:
  - outer part of disk gives higher disk bandwidth

1. Copy new index to inner half of disk (while still serving old index)

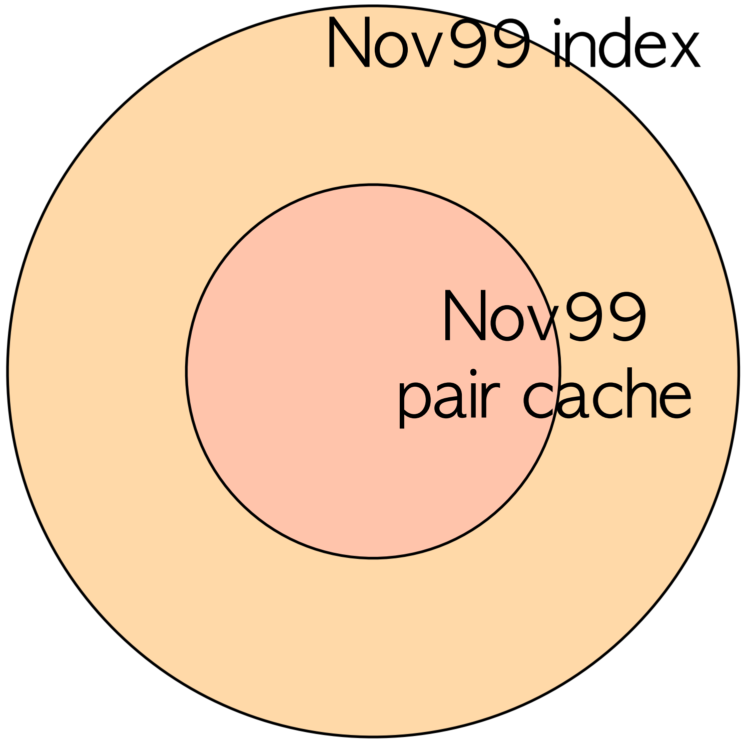
2. Restart to use new index

3. Wipe old index

4. Re-copy new index to faster half of disk

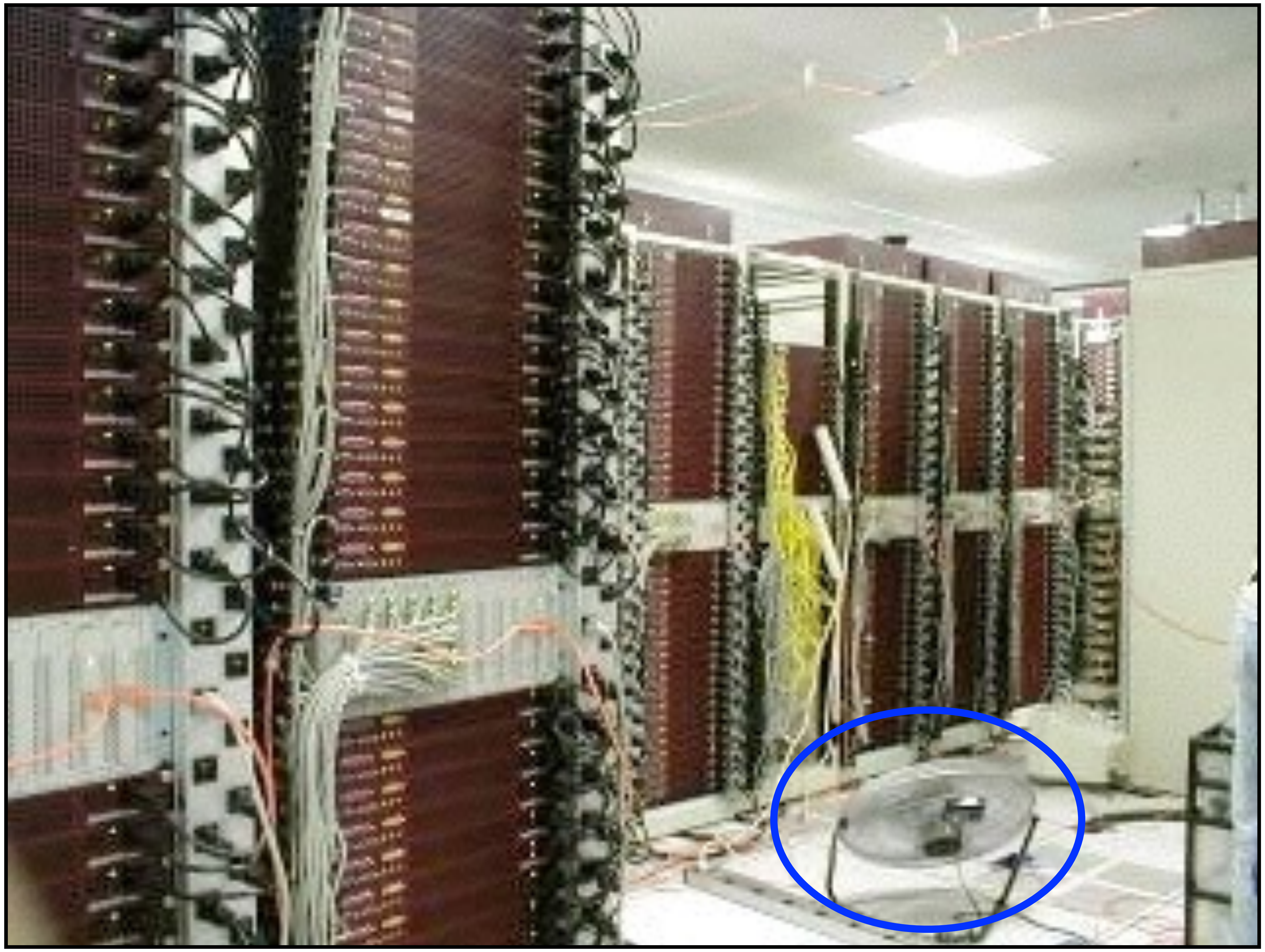
5. Wipe first copy of new index

6. Inner half now free for building various performance improving data structures



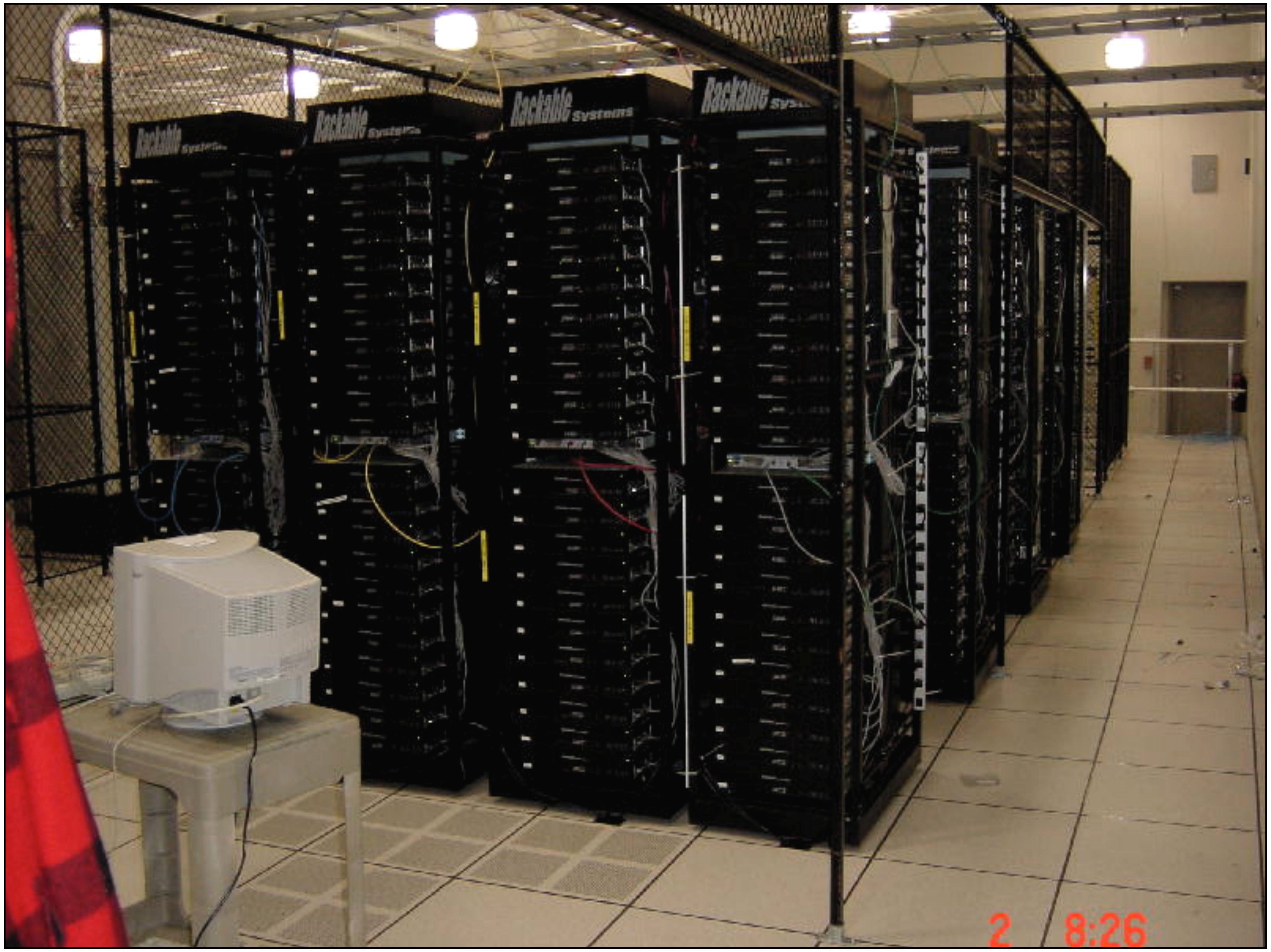
Pair cache: pre-intersected pairs of posting lists for commonly co-occurring query terms (e.g. “new” and “york”, or “barcelona” and “restaurants”)

Google Data Center (2000)



Google (new data center 2001)

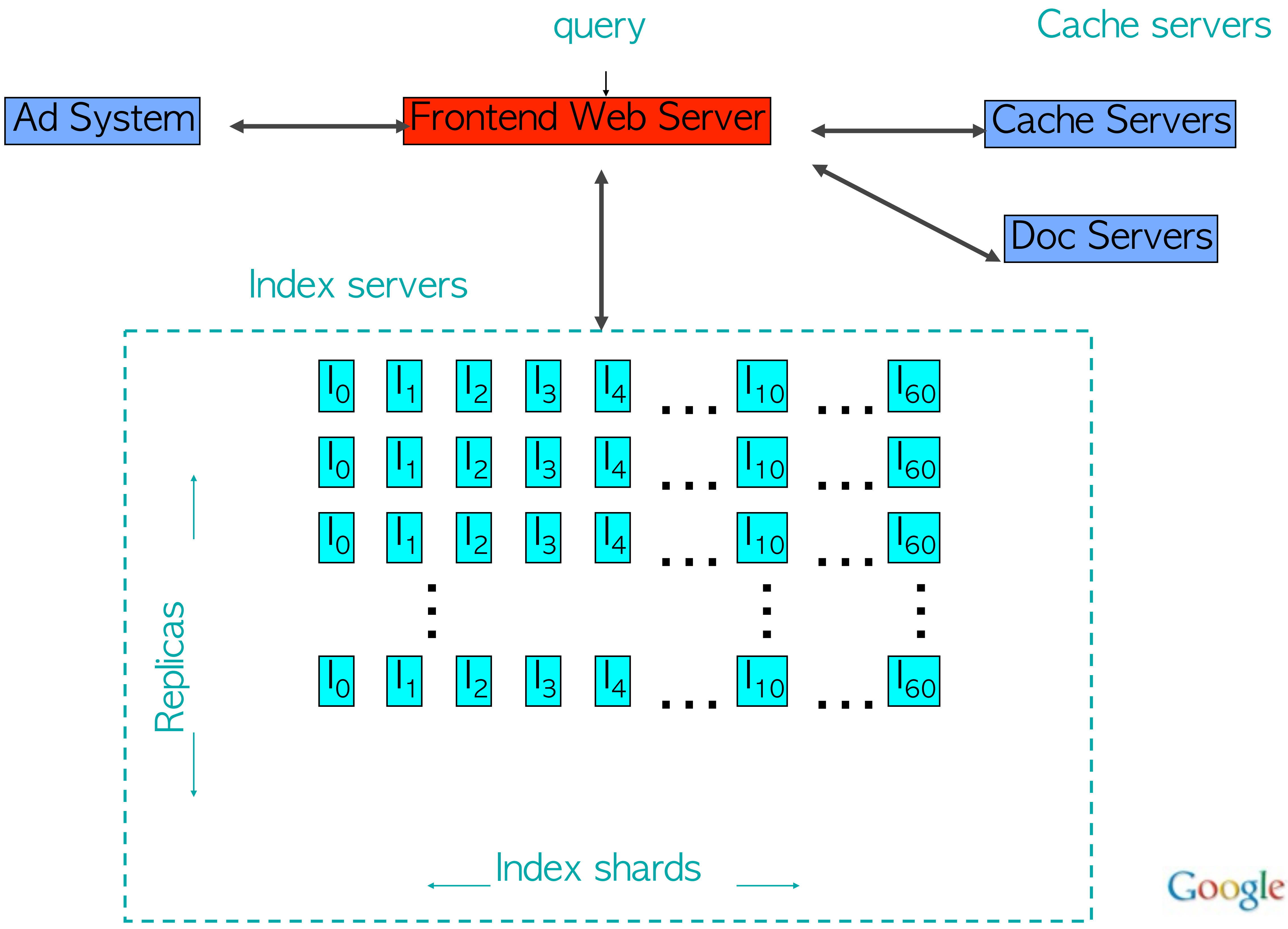
Google Data Center (3 days later)



Increasing Index Size and Query Capacity

- Huge increases in index size in ’99, ’00, ’01, ...
  - From ~50M pages to more than 1000M pages
- At same time as huge traffic increases
  - ~20% growth per month in 1999, 2000, ...
  - ... plus major new partners (e.g. Yahoo in July 2000 doubled traffic overnight)
- Performance of index servers was paramount
  - Deploying more machines continuously, but...
  - Needed ~10-30% software-based improvement every month

Dealing with Growth



Google

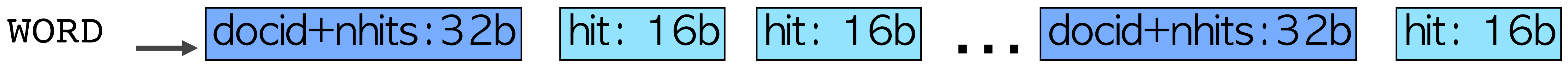
Implications

- Shard response time influenced by:
  - # of disk seeks that must be done
  - amount of data to be read from disk

- Big performance improvements possible with:
  - better disk scheduling
  - improved index encoding

Index Encoding circa 1997-1999

- Original encoding ('97) was very simple:



  - hit: position plus attributes (font size, title, etc.)
  - Eventually added skip tables for large posting lists
- Simple, byte aligned format
  - cheap to decode, but not very compact
  - ... required lots of disk bandwidth

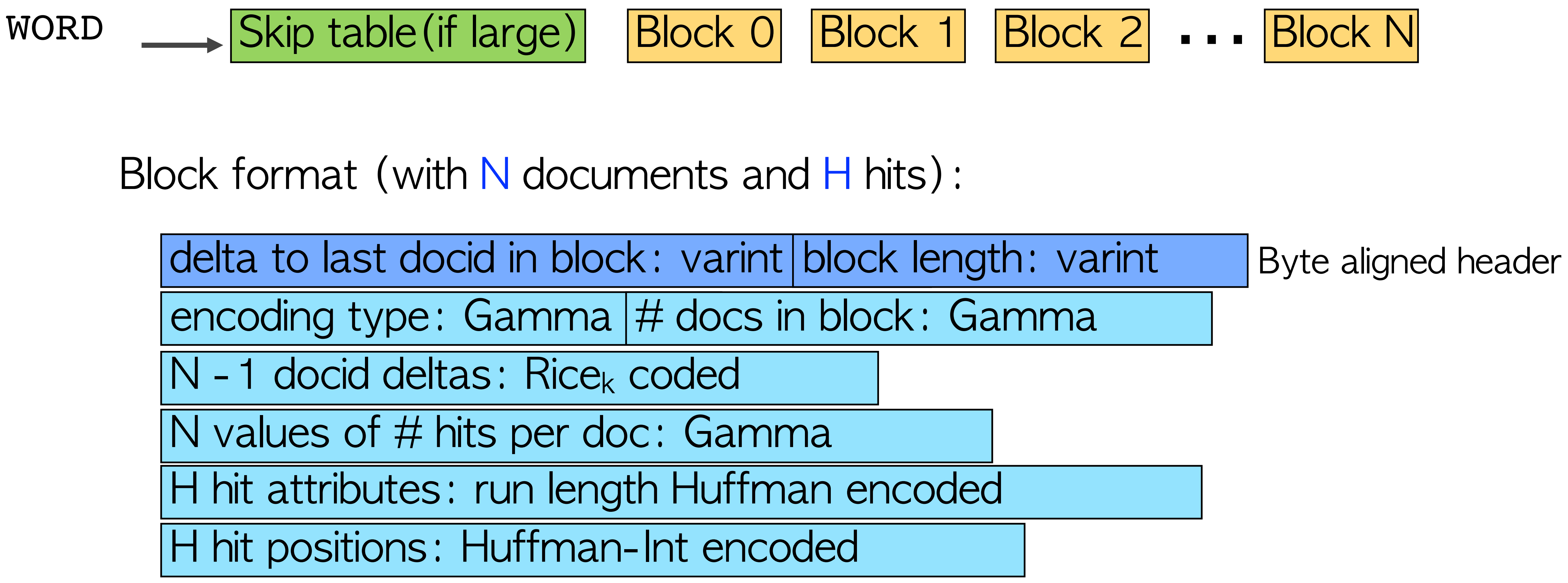
Encoding Techniques

- Bit-level encodings:
  - Unary: N ‘1’s followed by a ‘0’
  - Gamma: $\log_2(N)$ in unary, then floor(log2(N)) bits
  - Rice$_K$: floor(N / 2^K) in unary, then N mod 2^K in K bits
    - special case of Golomb codes where base is power of 2
  - Huffman-Int: like Gamma, except $\log_2(N)$ is Huffman coded instead of encoded w/ Unary

- Byte-aligned encodings:
  - varint: 7 bits per byte with a continuation bit
    - 0-127: 1 byte, 128-4095: 2 bytes, ...
  - ...

Block-Based Index Format

- Block-based, variable-len format reduced both space and CPU

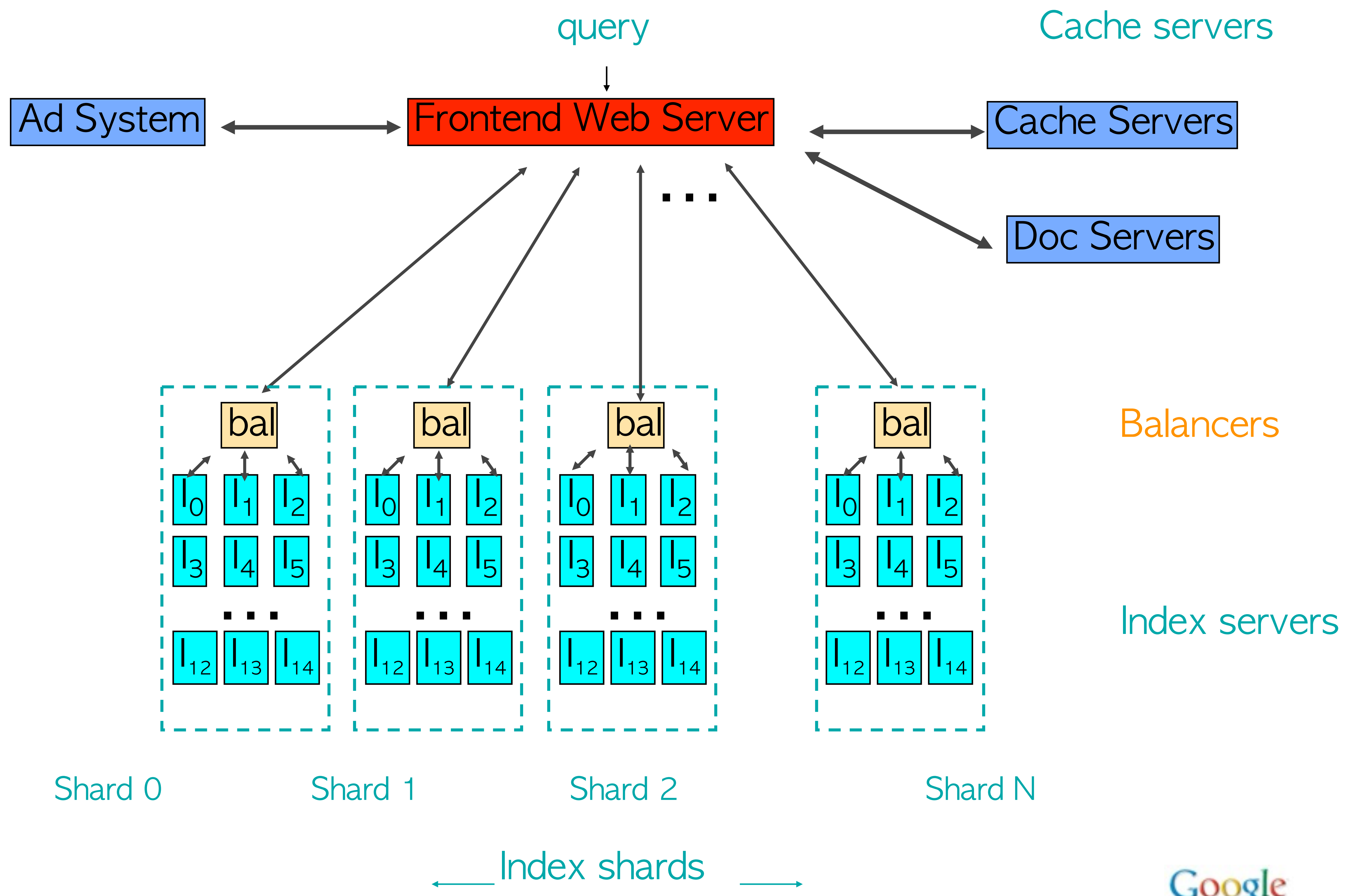

WORD
Skip table (if large)
Block 0
Block 1
Block 2
. . .
Block N
Block format (with N documents and H hits):
delta to last docid in block: varint
block length: varint
Byte aligned header
encoding type: Gamma
# docs in block: Gamma
N - 1 docid deltas: Rice$_k$ coded
N values of # hits per doc: Gamma
H hit attributes: run length Huffman encoded
H hit positions: Huffman-Int encoded

- Reduced index size by ~30%, plus much faster to decode

Implications of Ever-Wider Sharding

- Must add shards to keep response time low as index size increases
- ... but query cost increases with # of shards
  - typically >= 1 disk seek / shard / query term
  - even for very rare terms
- As # of replicas increases, total amount of memory available increases
  - Eventually, have enough memory to hold an **entire copy of the index in memory**
    - radically changes many design parameters

Early 2001: In-Memory Index



Google

In-Memory Indexing Systems

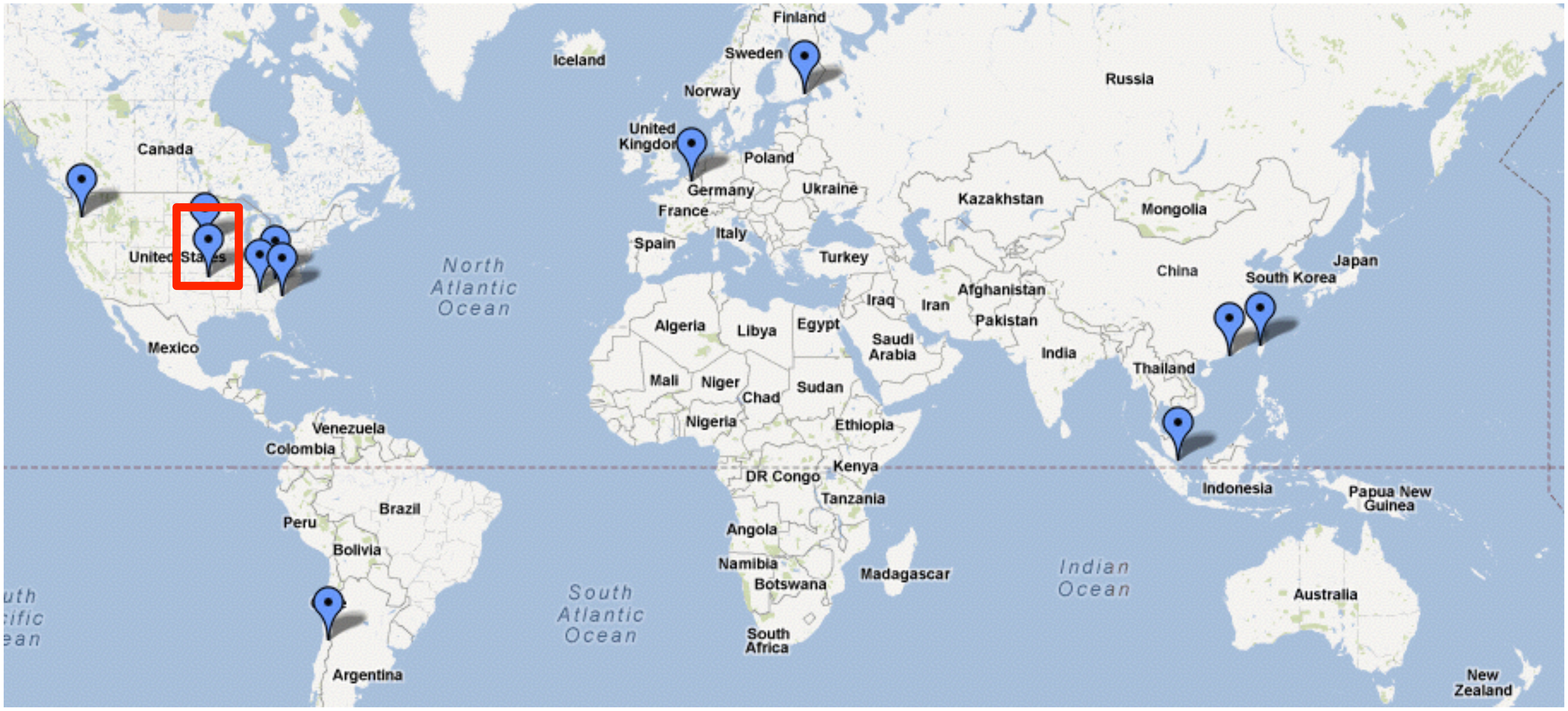
- Many positives:
  - big increase in throughput
  - big decrease in latency
    - especially at the tail: expensive queries that previously needed GBs of disk I/O became much faster

```
        e.g. [ "circle of life" ]
```

- Some issues:
  - Variance: touch 1000s of machines, not dozens
    - e.g. randomized cron jobs caused us trouble for a while
  - Availability: 1 or few replicas of each doc's index data
    - Queries of death that kill all the backends at once: very bad
    - Availability of index data when machine failed (esp for important docs): replicate important docs

Google's Computational Environment Today

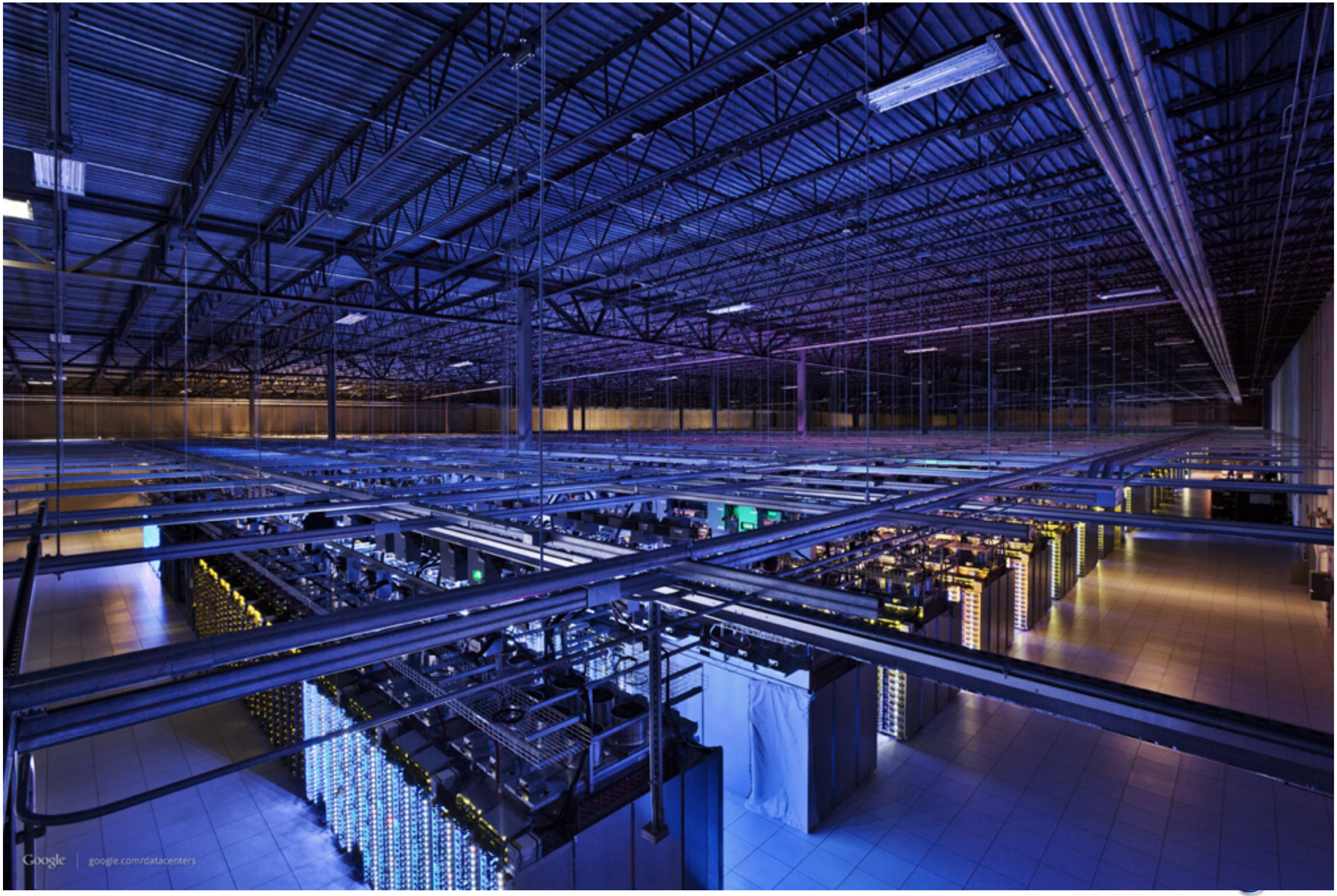
- Many datacenters around the world



Google

Zooming In...

Zooming In...

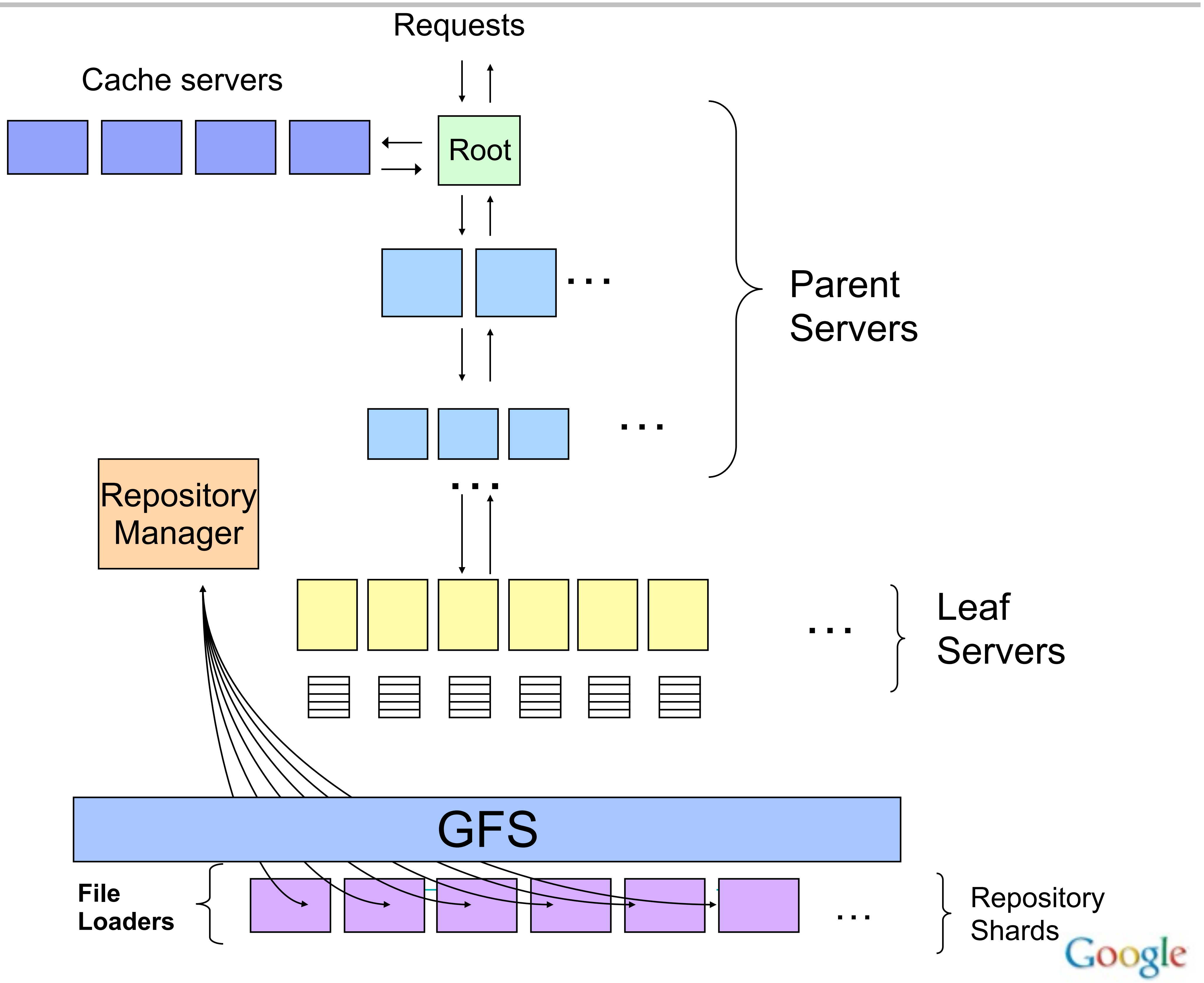


Lots of machines...

Save a bit of power: turn out the lights...

Cool...

Serving Design, 2004 edition

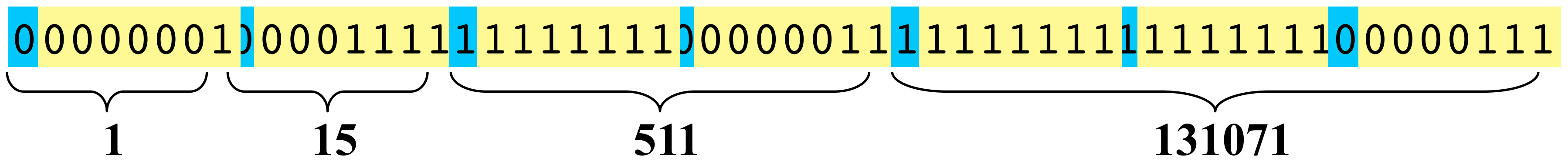


New Index Format

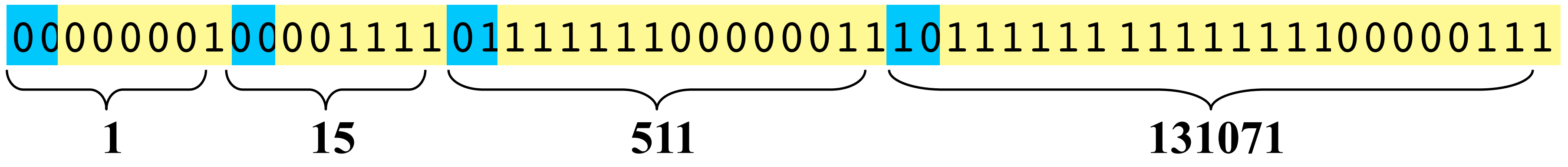
- Block index format used two-level scheme:
  - Each hit was encoded as (docid, word position in doc) pair
  - Docid deltas encoded with Rice encoding
  - Very good compression (originally designed for disk-based indices), but slow/CPU-intensive to decode
- New format: single flat position space
  - Data structures on side keep track of doc boundaries
  - Posting lists are just lists of delta-encoded positions
  - Need to be compact (can’t afford 32 bit value per occurrence)
  - … but need to be very fast to decode

Byte-Aligned Variable-length Encodings

- Varint encoding:
  - 7 bits per byte with continuation bit
  - Con: Decoding requires lots of branches/shifts/masks



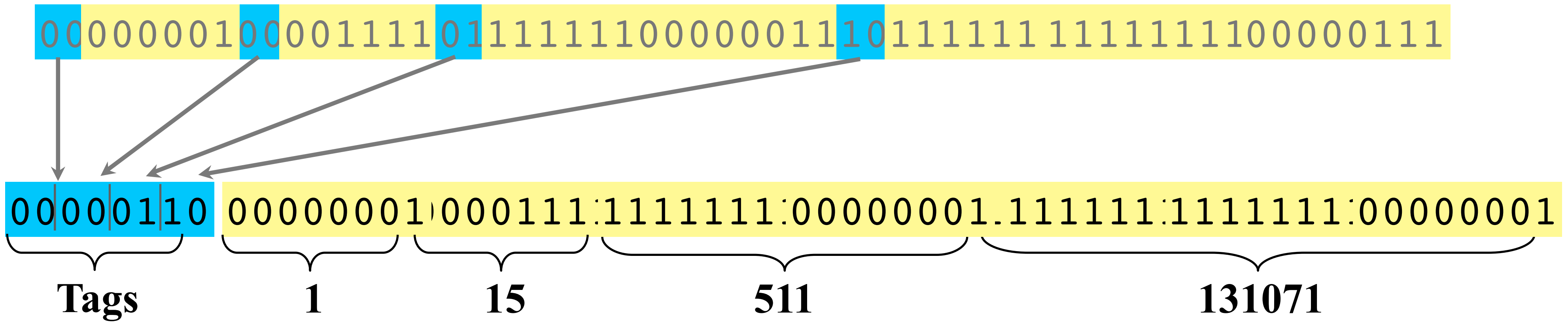
- Idea: Encode byte length as low 2 bits
  - Better: fewer branches, shifts, and masks
  - Con: Limited to 30-bit values, still some shifting to decode



Google

Group Varint Encoding

- Idea: encode groups of 4 values in 5-17 bytes
  - Pull out 4 2-bit binary lengths into single byte prefix

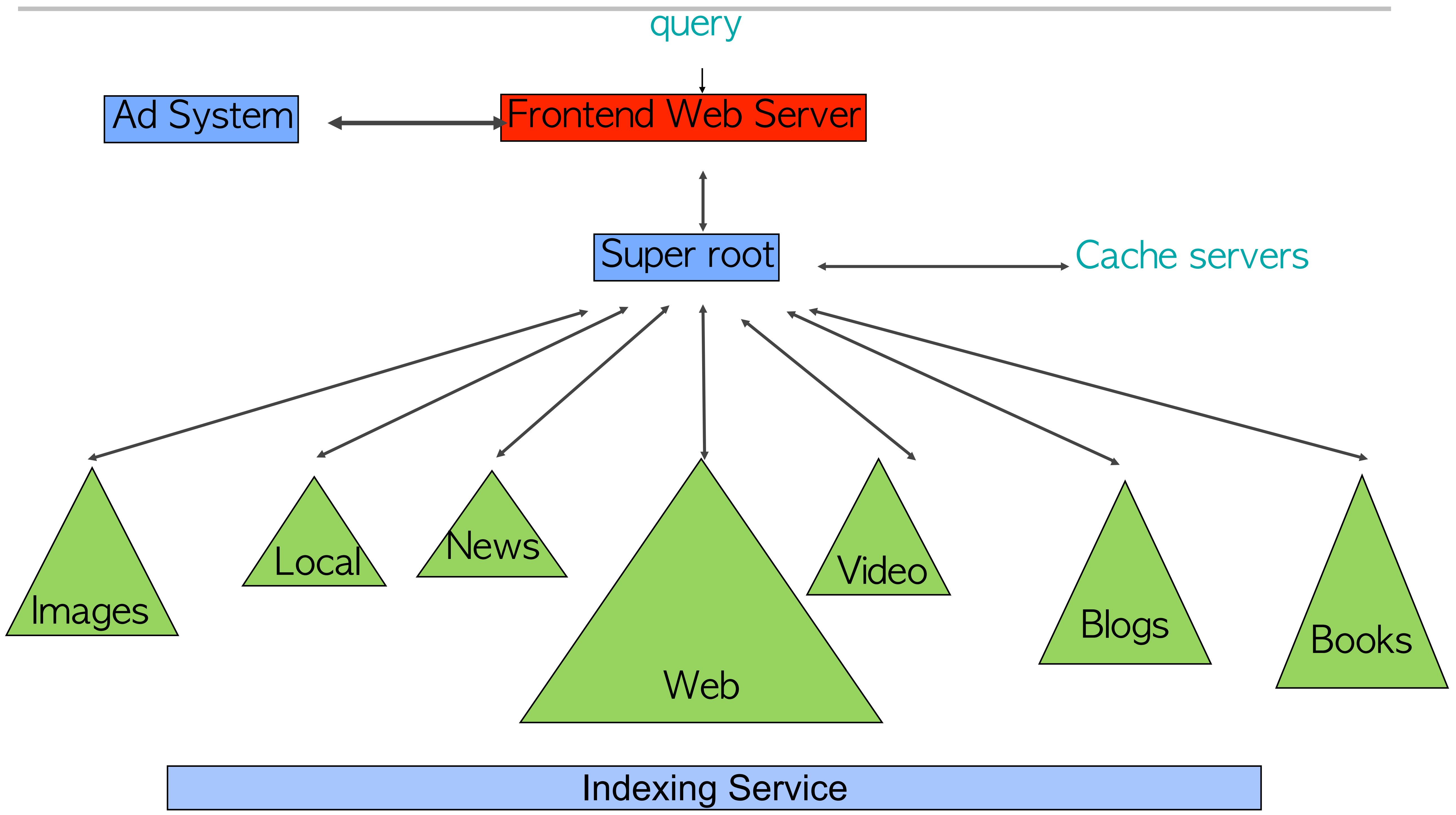


- Decode: Load prefix byte and use value to lookup in 256-entry table:

```
                 ...
00000110 → Offsets: +1,+2,+3,+5; Masks: ff, ff, ffff, ffffff
                 ...
```

- Much faster than alternatives:
  - 7-bit-per-byte varint: decode ~180M numbers/second
  - 30-bit Varint w/ 2-bit length: decode ~240M numbers/second
  - Group varint: decode ~400M numbers/second

2007: Universal Search



Index that? Just a minute!

- Low-latency crawling and indexing is tough
  - crawl heuristics: what pages should be crawled?
  - crawling system: need to crawl pages quickly
  - indexing system: depends on global data
    - PageRank, anchor text of pages that point to the page, etc.
    - must have online approximations for these global properties
  - serving system: must be prepared to accept updates while serving requests
    - very different data structures than batch update serving system

Often use hybrid systems (1+2):

1. Very fast update system, but with modest capacity (high $$$/doc/query)

2. Slower updating system, but with huge capacity and relatively immutable data structures (low $/doc/query)

Understanding Text

Query

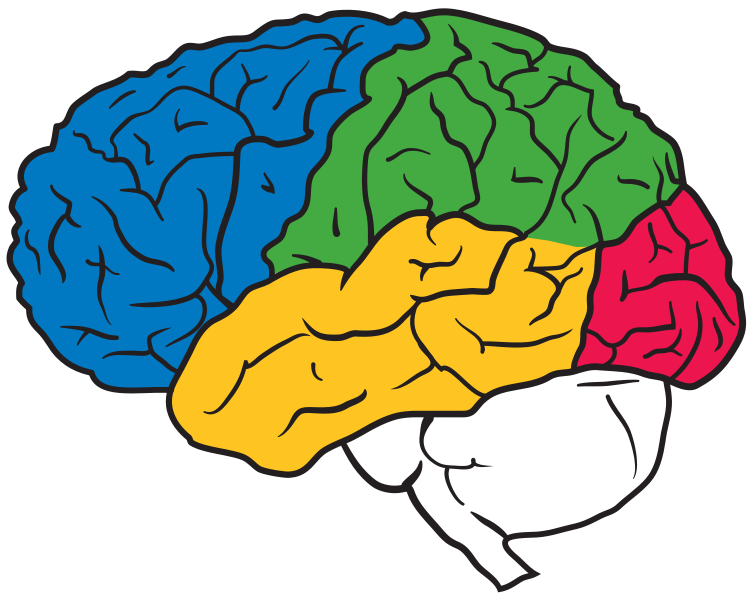
[car parts for sale]

Document 1

… car parking available for a small fee.
… parts of our floor model inventory for sale.

Document 2

Selling all kinds of automobile and pickup truck parts, engines, and transmissions.

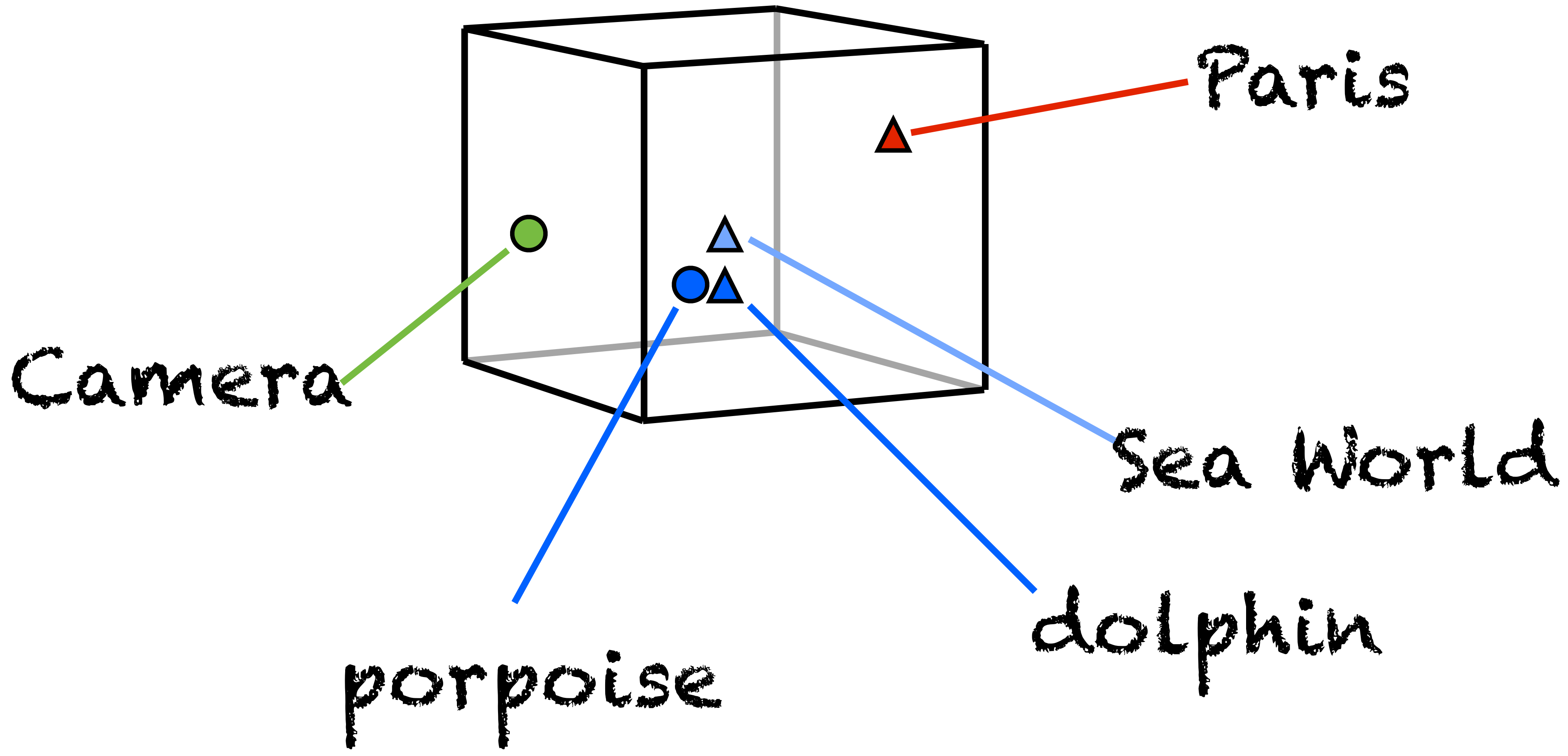


- Embeddings & Neural Nets
- How we use them
- What's next?

go/brain

N-dimensional Embeddings

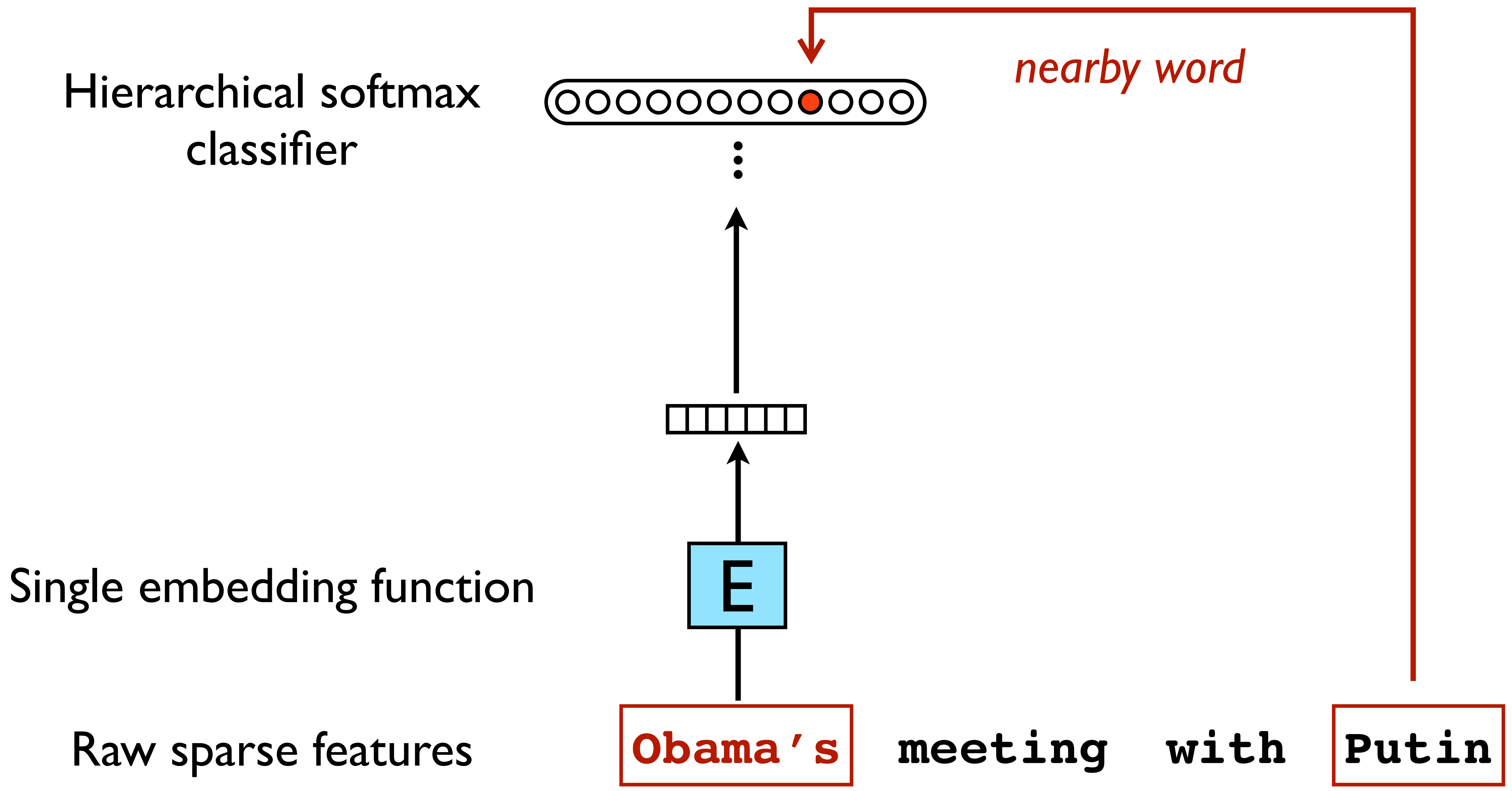
3-D embedding space



Embedding Function: A look-up-table that maps sparse features into dense floating point vectors.

Usually use many more dimensions than 3 (e.g. 1000-D embeddings)

Skipgram Text Model



Mikolov, Chen, Corrado and Dean. *Efficient Estimation of Word Representations in Vector Space*, http://arxiv.org/abs/1301.3781

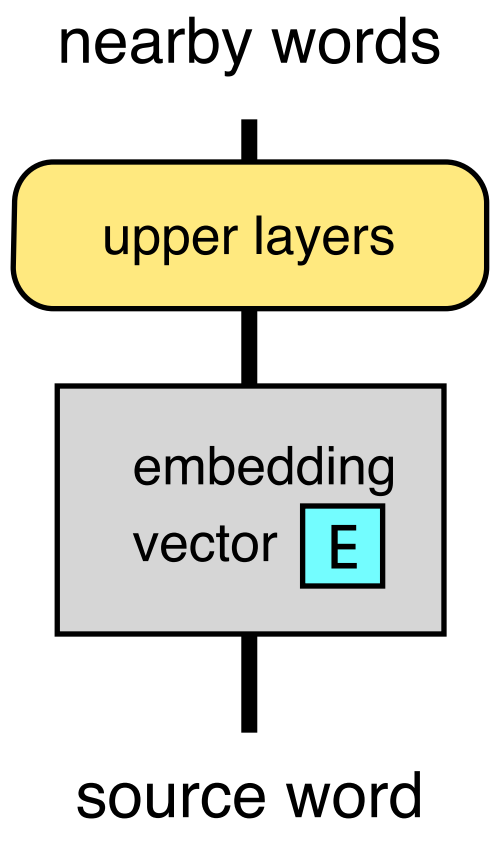
Open source implementation: https://code.google.com/p/word2vec/

Google

Nearest neighbors in language embeddings space are closely related semantically.

- Trained language model on Wikipedia corpus.

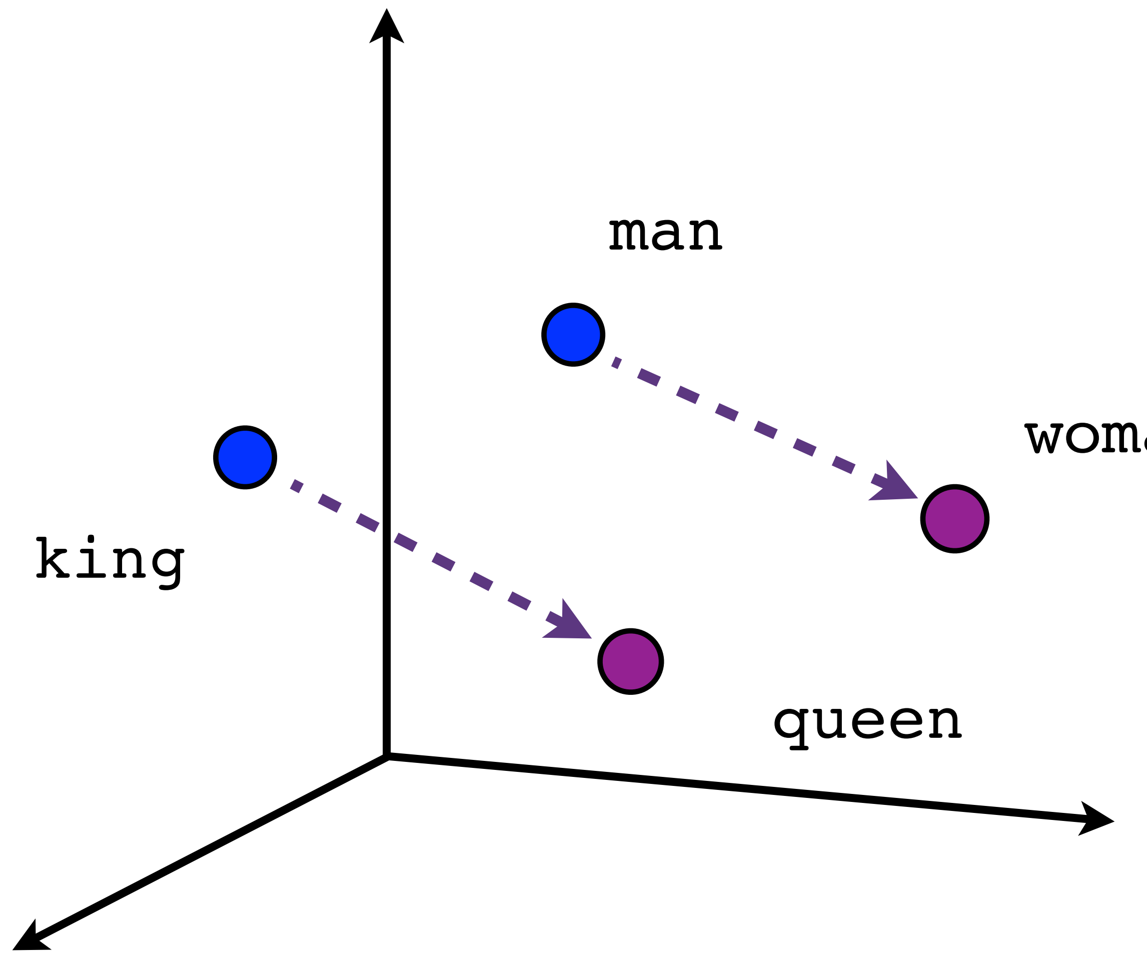
tiger shark	car	new york
bull shark	cars	new york city
blacktip shark	muscle car	brooklyn
shark	sports car	long island
oceanic whitetip shark	compact car	syracuse
sandbar shark	autocar	manhattan
dusky shark	automobile	washington
blue shark	pickup truck	bronx
requiem shark	racing car	yonkers
great white shark	passenger car	poughkeepsie
lemon shark	dealership	new york state

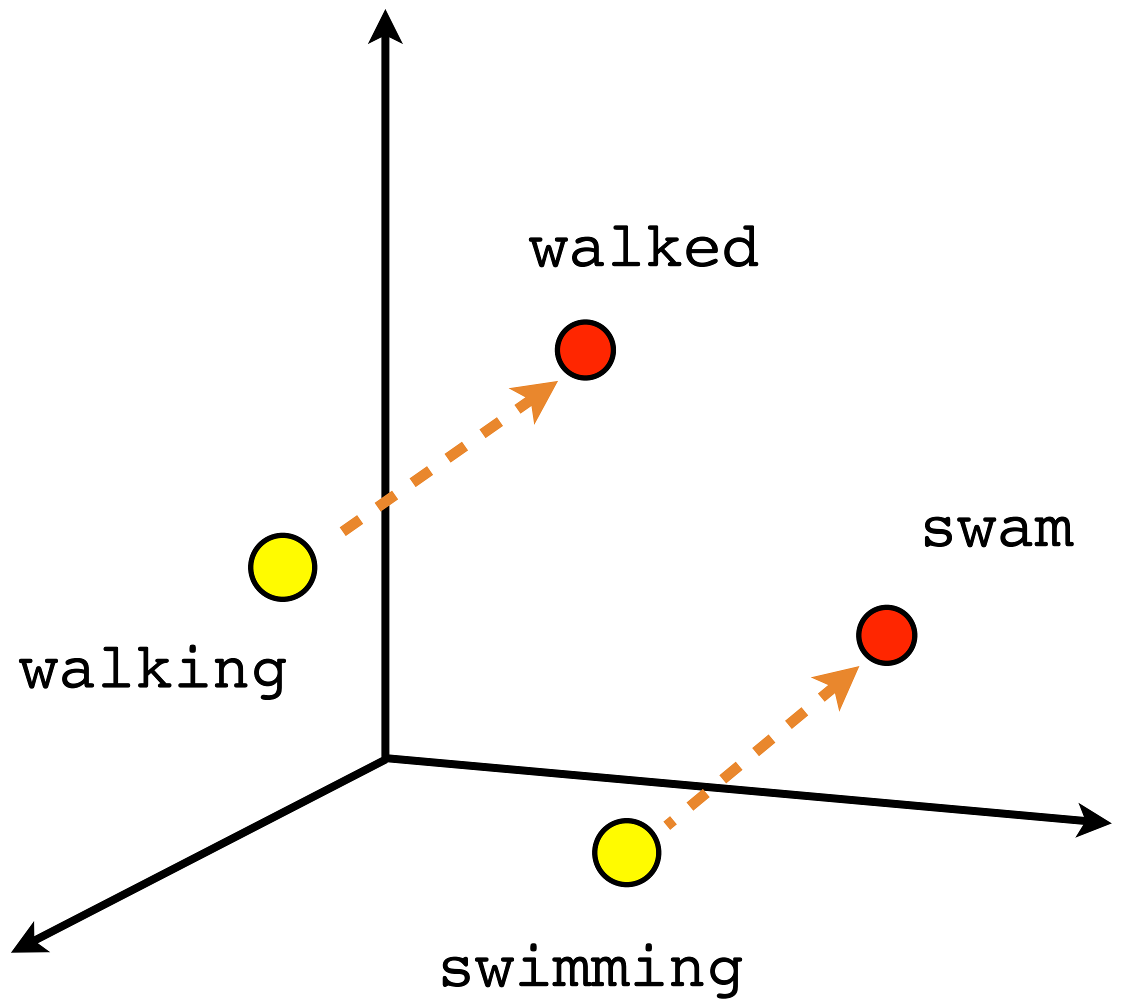


Google

* 5.7M docs, 5.4B terms, 155K unique terms, 500-D embeddings

Embeddings are Powerful





Solving Analogies

- Embedding vectors trained for the language modeling task have very interesting properties (especially the skip-gram model).

$$E(\textit{hotter}) - E(\textit{hot}) \approx E(\textit{bigger}) - E(\textit{big})$$

$$E(\textit{Rome}) - E(\textit{Italy}) \approx E(\textit{Berlin}) - E(\textit{Germany})$$

Solving Analogies

- Embedding vectors trained for the language modeling task have very interesting properties (especially the skip-gram model).

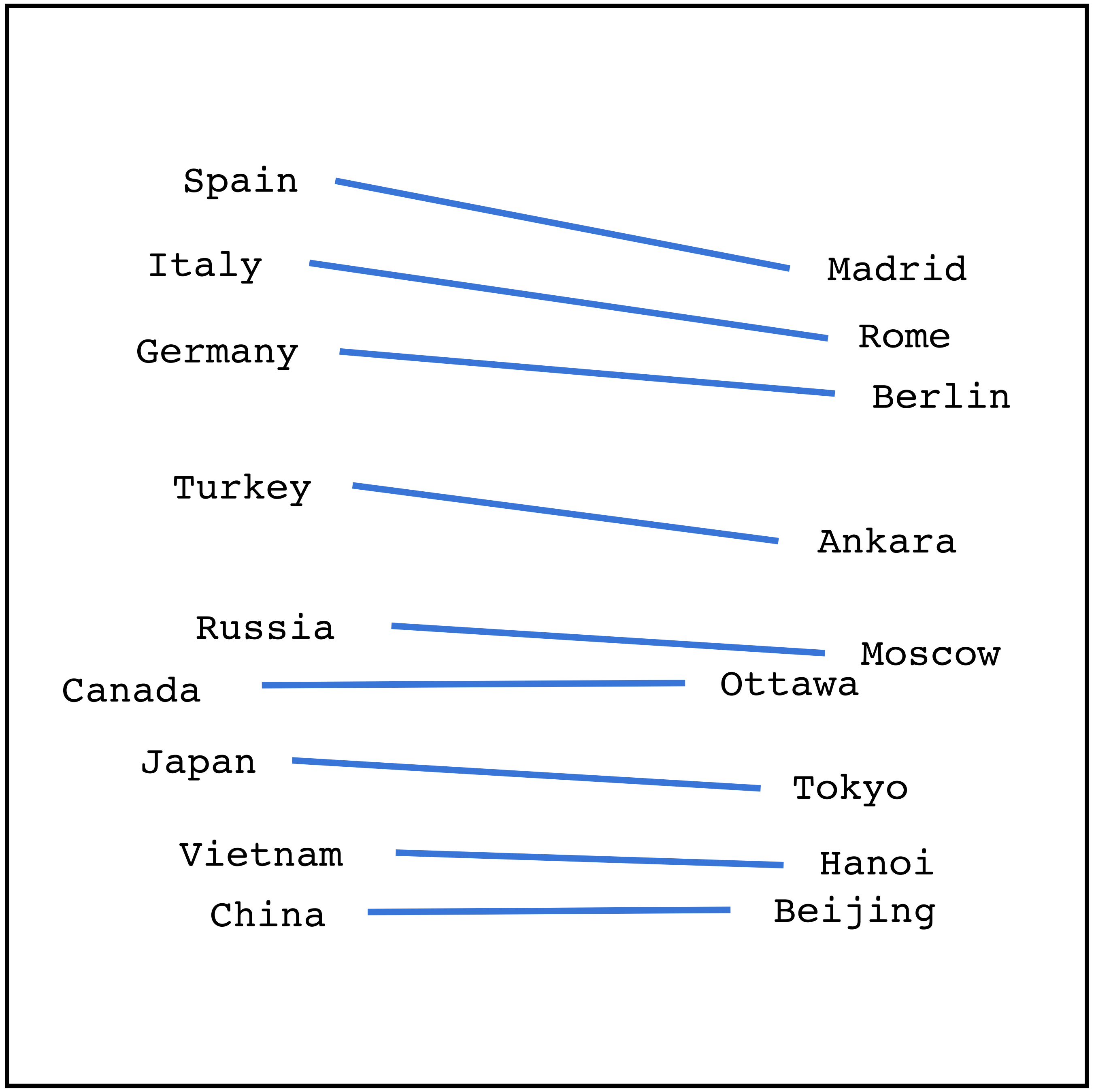
$$E(\textit{hotter}) - E(\textit{hot}) + E(\textit{big}) \approx E(\textit{bigger})$$

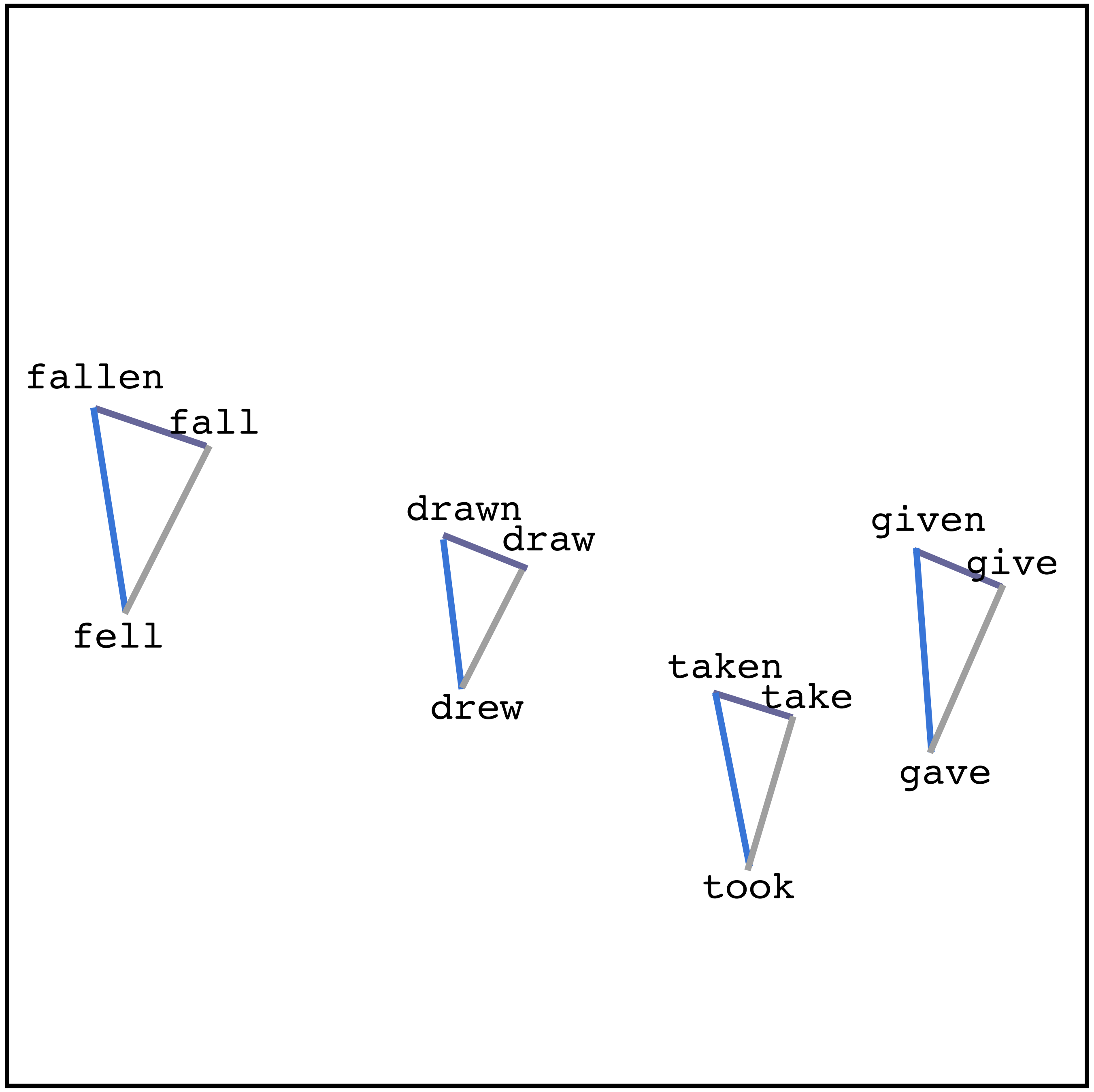
$$E(\textit{Rome}) - E(\textit{Italy}) + E(\textit{Germany}) \approx E(\textit{Berlin})$$

Skip-gram model w/ 640 dimensions trained on 6B words of news text achieves 57% accuracy for analogy-solving test set.

Details in: *Efficient Estimation of Word Representations in Vector Space*. Mikolov, Chen, Corrado and Dean. NIPS 2013.

Embeddings are Powerful





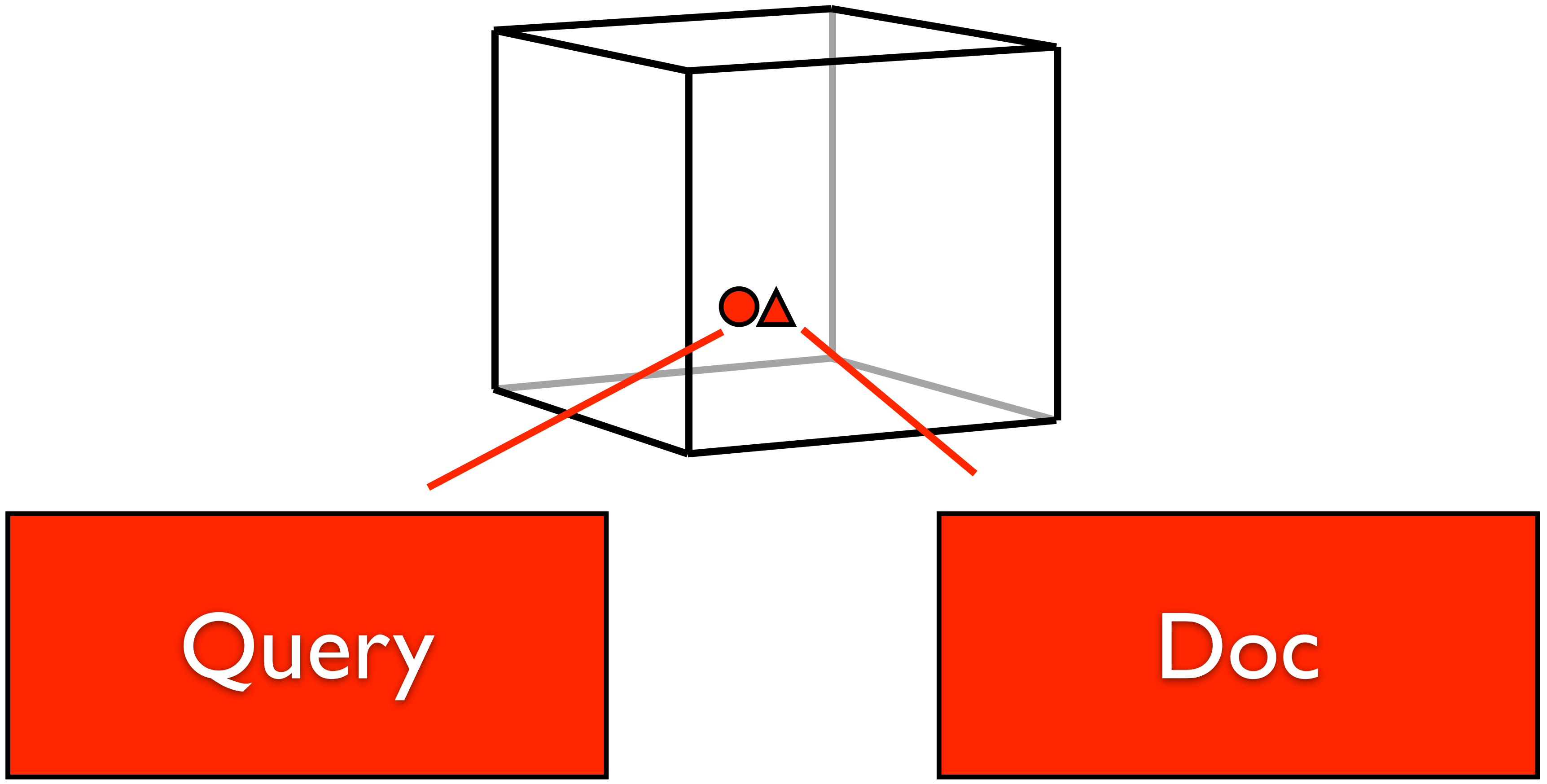
Google

N-dimensional Embeddings

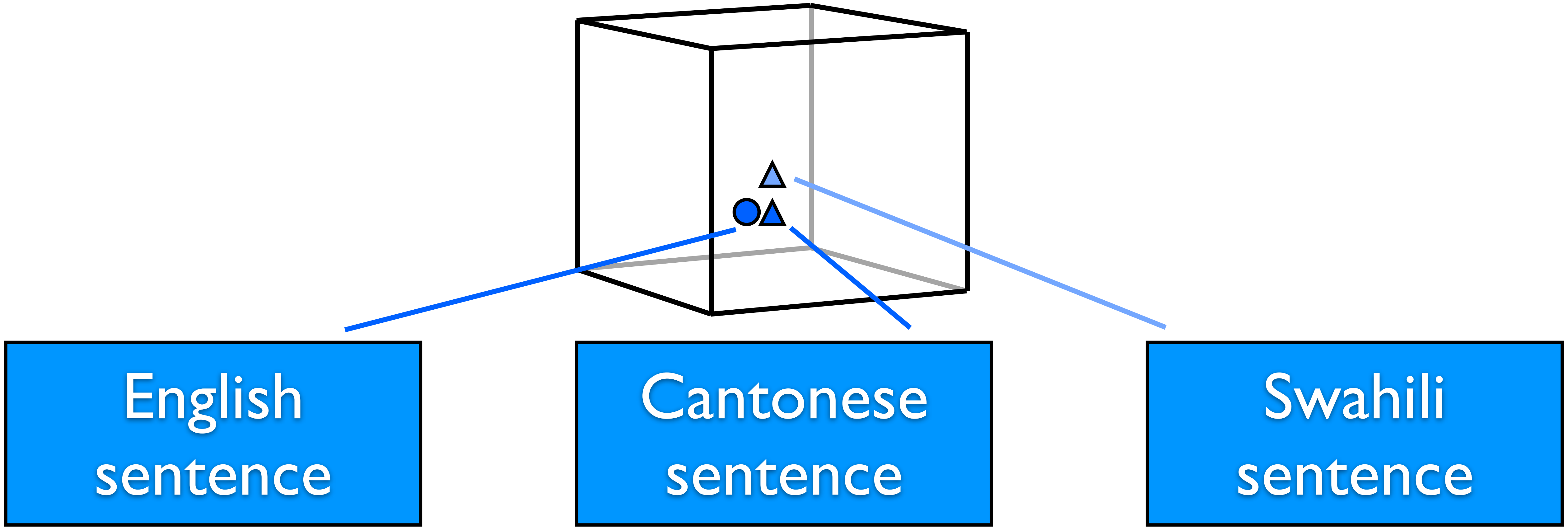
Plenty of applications:

- Machine Translation
- Synonym handling
- Sentiment analysis
- Ads
- Semantic visual models
- Entity disambiguation

Embedding Longer Pieces of Text

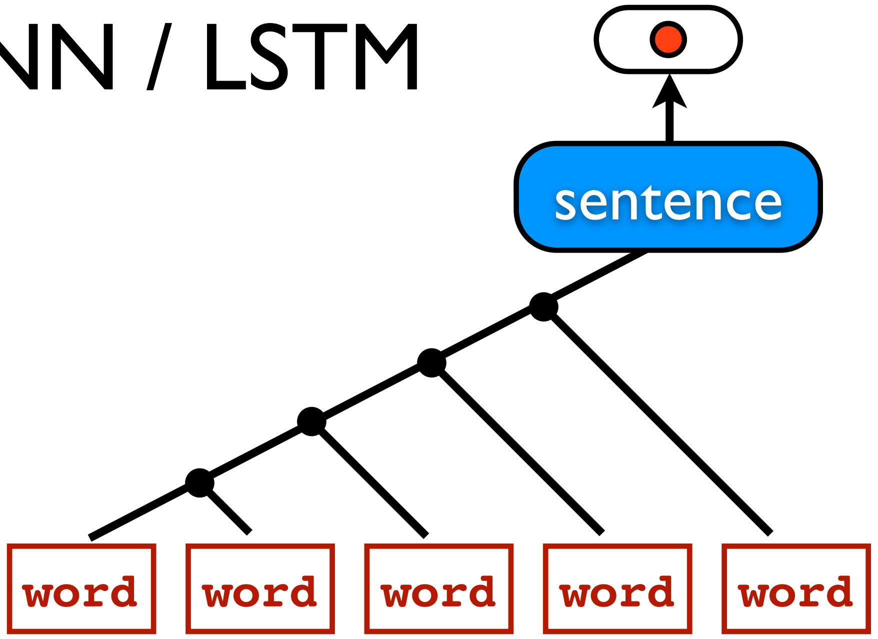


Embedding Longer Pieces of Text

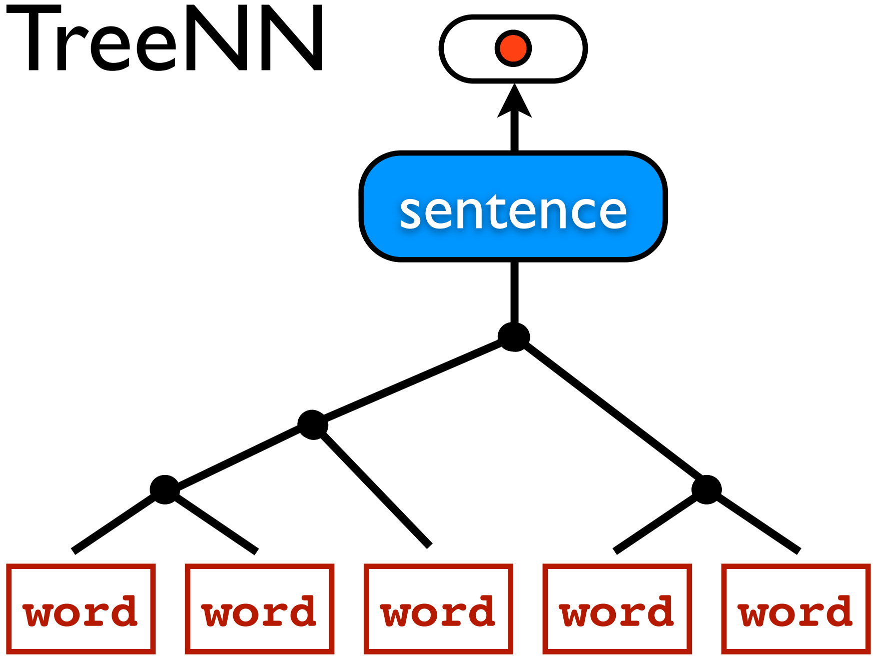


Embedding Longer Pieces of Text

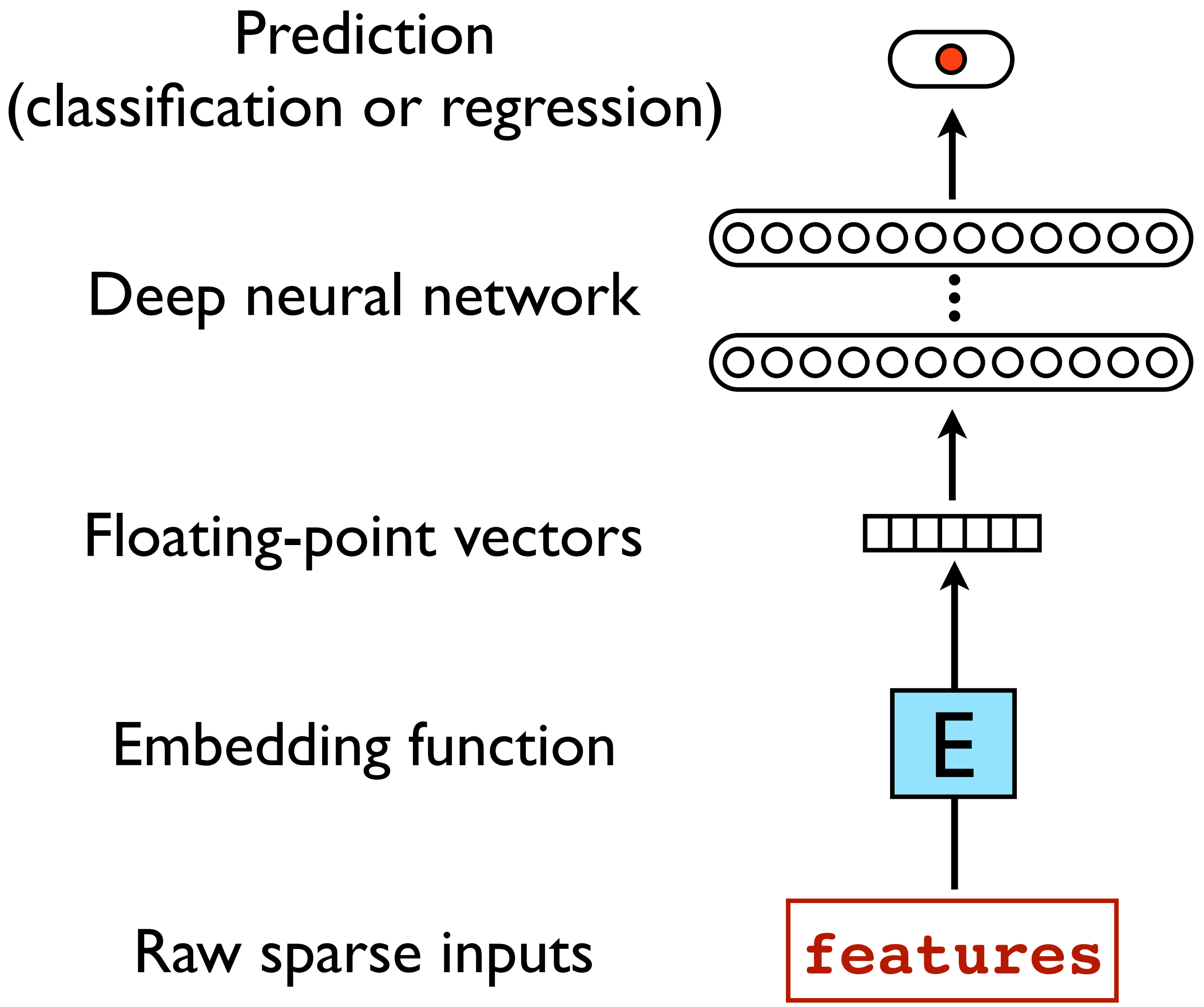
RNN / LSTM



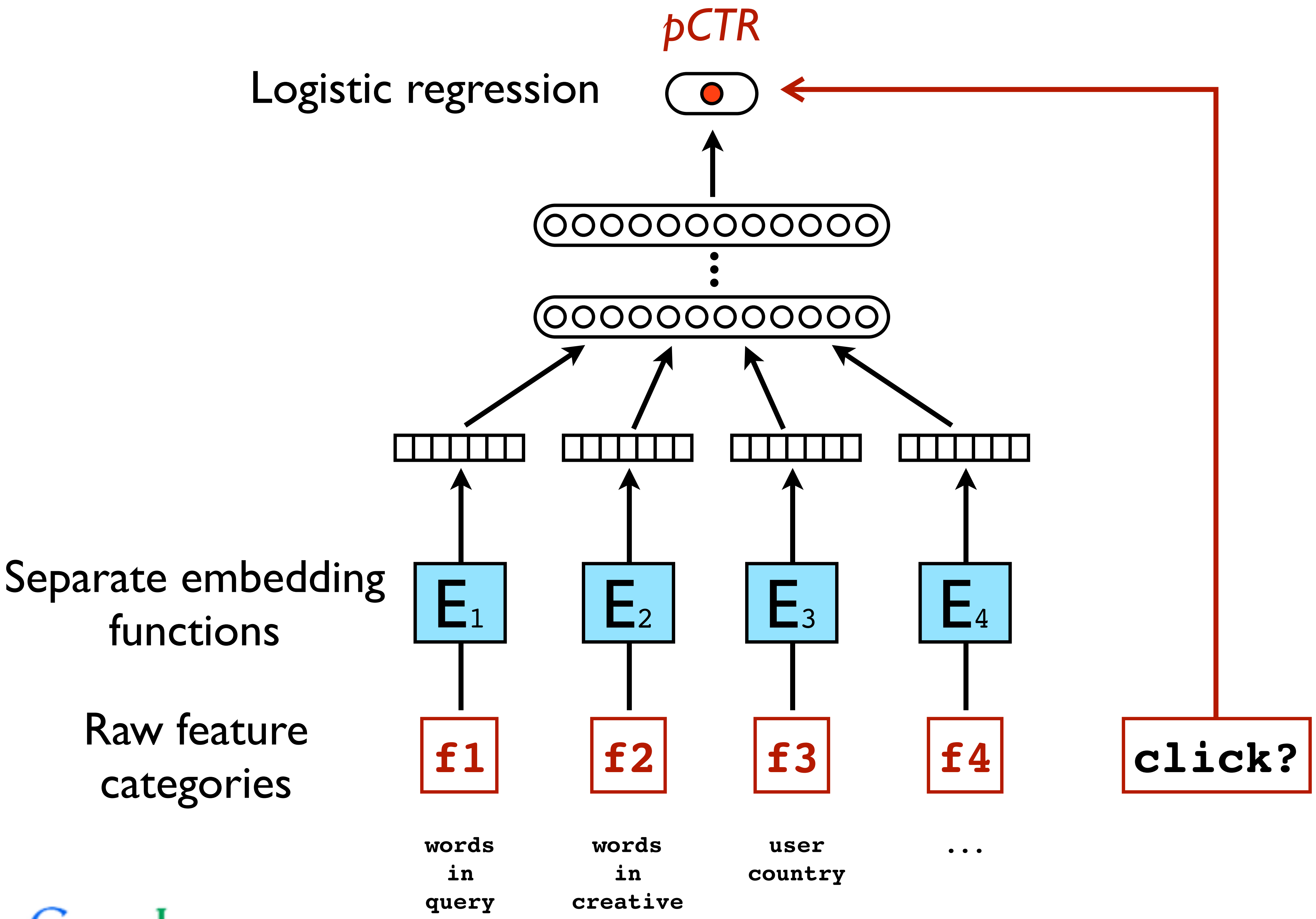
TreeNN



Using Embeddings in Larger Neural Models



Predicting clicks



Google

Embeddings and Neural Nets Show Considerable Promise

In Conclusion...

- Designing and building large-scale retrieval systems is a challenging, fun endeavor
  - new problems require continuous evolution
  - work benefits many users
  - new retrieval techniques often require new systems

- Thanks for your attention!

Thanks! Questions...?

- Further reading:

Ghemawat, Gobioff, & Leung. *Google File System*, SOSP 2003.

Barroso, Dean, & Hölzle. *Web Search for a Planet: The Google Cluster Architecture*, IEEE Micro, 2003.

Dean & Ghemawat. *MapReduce: Simplified Data Processing on Large Clusters*, OSDI 2004.

Chang, Dean, Ghemawat, Hsieh, Wallach, Burrows, Chandra, Fikes, & Gruber. *Bigtable: A Distributed Storage System for Structured Data*, OSDI 2006.

Brants, Popat, Xu, Och, & Dean. *Large Language Models in Machine Translation*, EMNLP 2007.

Mikolov, Chen, Corrado and Dean. *Efficient Estimation of Word Representations in Vector Space*, http://arxiv.org/abs/1301.3781

Dean, Corrado, et al. , *Large Scale Distributed Deep Networks,* NIPS 2012.

- These and many more available at:

  http://labs.google.com/papers.html

  http://research.google.com/people/jeff