

Full Text Indexing

versus

Tagging

Or how to tag a full text index?

Let's start with some

Kitchen Psychology

Let's start with some Kitchen Psychology

„When did you
take that picture
of my desk?“
Remember 'A'



„Let's login and
start working.“
Remember 'B'



Let's start with some
Kitchen Psychology



„There you are
my beloved pair
of red socks.“
Remember 'A'

„Who switched
the blue and the
yellow pair?“
Remember 'B'



Here's
the result

A+A:

Keep piling. Google loves you
and you love google.

B+B:

You are definitely a filer,
del.icio.us is so successful thanks
to you.

A+B or B+A:

Well, I would be worried...

Filer & Piler

People tend to be either filer or piler.
Classic: look at your inbox!

Filer & Piler

People tend to be either filer or piler.
Classic: look at your inbox!

Let's start with the pilers...

Full Text Indexing

The goal of creating a full text index is to minimize the time to find a (relevant) document to a specific query.

Commonly used data structure:
inverted index

Inverted Index

D1: „**You can make money without doing evil**“

D2: „**You can be serious without a suit**“

Would look in a simplified
inverted index:

you	(1,1)	(2,1)
can	(1,2)	(2,2)
make	(1,3)	
money	(1,4)	
without	(1,5)	(2,5)
do	(1,6)	
evil	(1,7)	
serious	(2,4)	
suit	(2,7)	

Inverted Index

D1: „**You can make money without doing evil**“

D2: „**You can be serious without a suit**“

Would look in a simplified
inverted index

e.g. Stopword

e.g. Stemming

you	(1,1) (2,1)
can	(1,2) (2,2)
make	(1,3)
money	(1,4)
without	(1,5) (2,5)
do	(1,6)
evil	(1,7)
serious	(2,4)
suit	(2,7)

An Index may contains



An Index may contains



An Index may contains



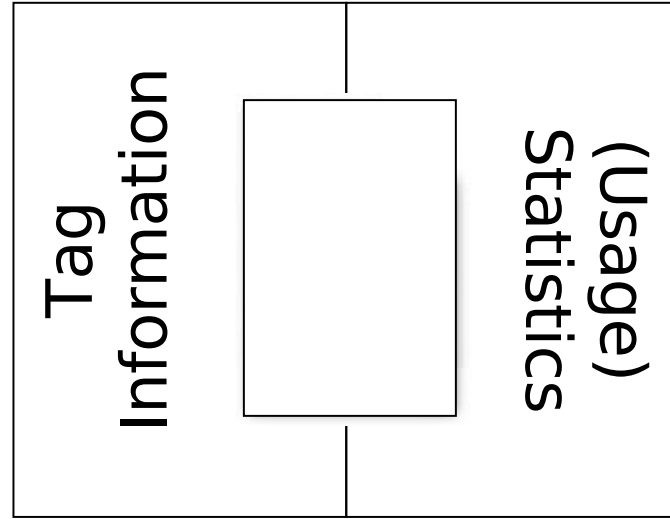
Tagging

“A tag is simply **a word** you use to describe a bookmark. Unlike folders, you make up tags **when** you need them and you can use as **many** as you like. The result is a better way to organize your bookmarks and a great way to **discover** interesting things on the Web.” *del.icio.us*

Tagging



Tagging



Tagging Data structure

This page is intentionally left blank

Tagging

other aspects

Tags may express subjective aspects like

- “mustRead“, “blogThis“ (actions)
- “WebtuesdayFeb07“ (temporal, share, separate)
- “*****“ (rating)

Multimedia Tagging

- Peekaboom, google image labler

Deep Tagging

(Tag Spam? Autotagging? A unique pers. Tag?)

Now

that we know all this

let's do a theoretical experiment:

- What happens, if – in a combined effort – all producers add all their pages with all tokens as tags to del.icio.us?

Now

that we know all this

let's do a theoretical experiment:

- What happens, if – in a combined effort – all producers add all their pages with all tokens as tags to del.icio.us?
- What if google would allow everyone to freely tag every document?

Key Differences

- i. Indexing/Matching
- ii. Ranking
- iii. Timing

Key

Differences - Matching

Segmentation

- How to Tag *San Francisco*?
- Or *La Punt-Chamues-ch*?

Normalisation

- <http://del.icio.us/tag/oel> vs. <http://del.icio.us/tag/öI> vs. <http://del.icio.us/tag/ÖI>
- Language detection
- Stemming/Decompounding

Extension/Reduction

- Stopwords
- Synonyms

Key

Differences - Ranking

Search engine

Lot's of different algorithms, taking all sorts of information into account, e.g.

- vector space model
- page rank

Tagging

the matching is exact and all information that can be used for ranking lies outside the documents, e.g.

- Nof times tagged, nof tags (Popularity)
- Recency
- Co-occurrence

Key Differences - Timing

Full Text Index

Producer (offline) – Consumer (online) Model.
Timing is unpredictable. Change the content of your page and pray.

Tagging

Almost synchronous?
The producer is also the consumer

A Tagging Time-Lag Exp

Site:

flickr.com

Document:

<http://www.flickr.com/photos/fxh/369274632/>

Tag:

averyunlikelywebtuesdaytag

Users:

owner, friend, the public

Can we combine the 2 approaches?

Why would we want this in 1st place?

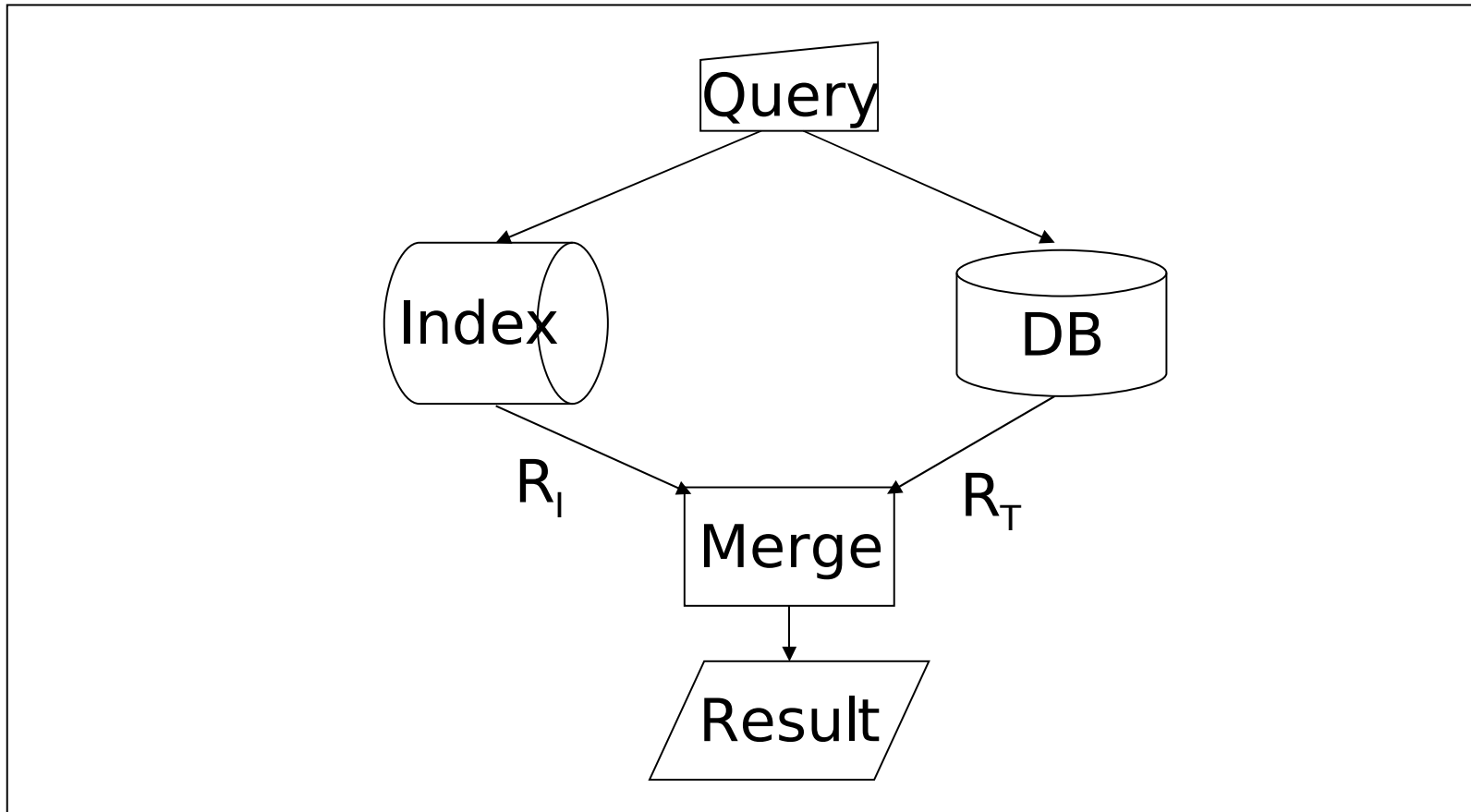
- Improve the search experience
- Reduce the number of places we need to consult
- Control the personalisation of your search result

Possible Strategies

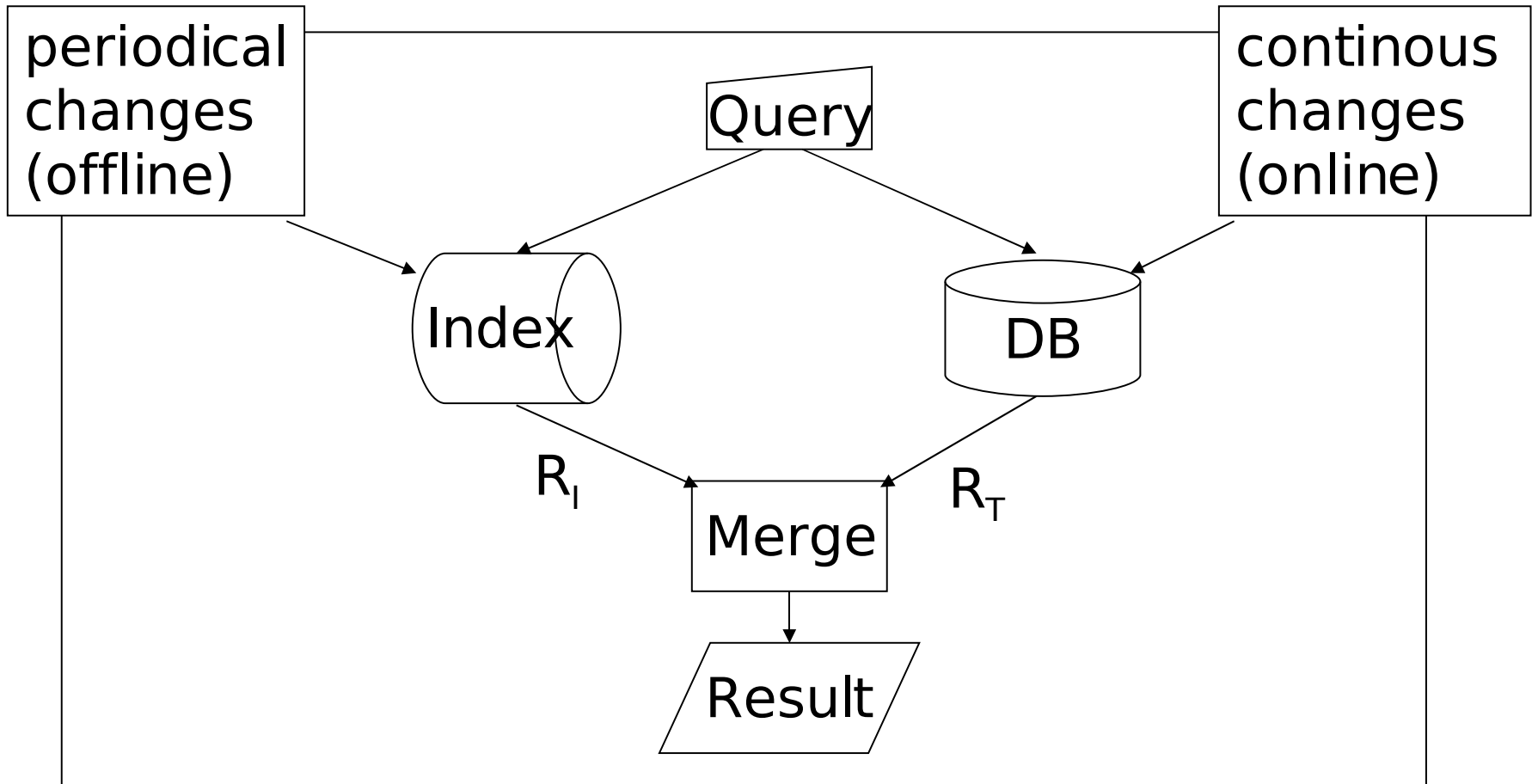
- i. Tags and Fulltext into DB
- ii. Tags into DB, Fulltext in inv.
Index
- iii. Tags and Fulltext into inv. Index

Strategy

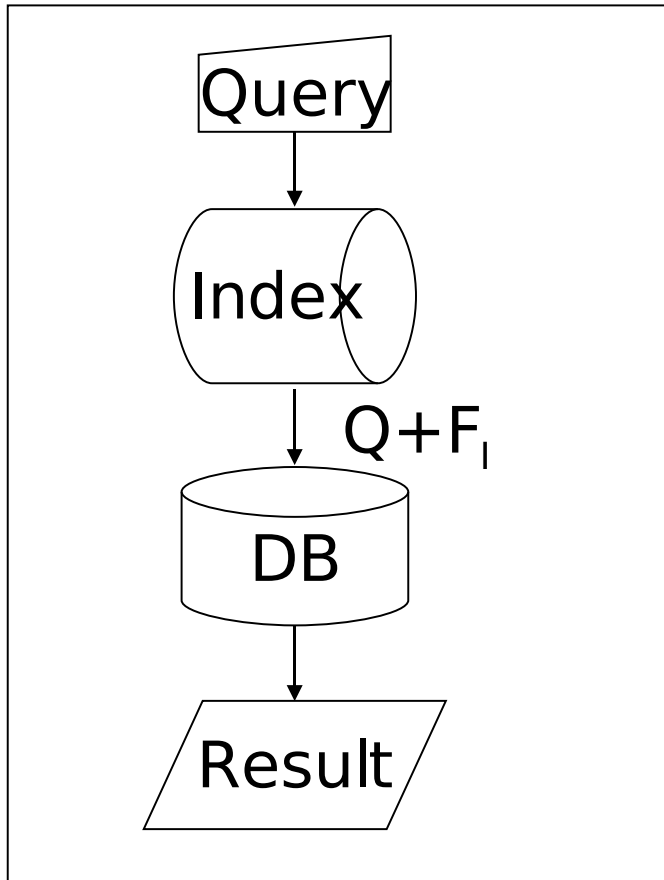
DB & Index (1)



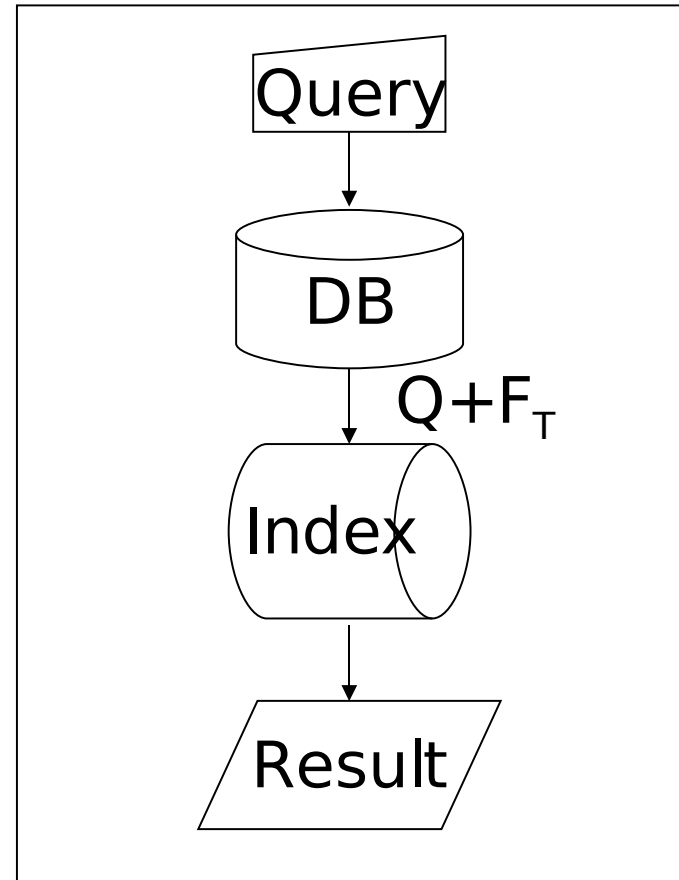
Strategy DB & Index (1)



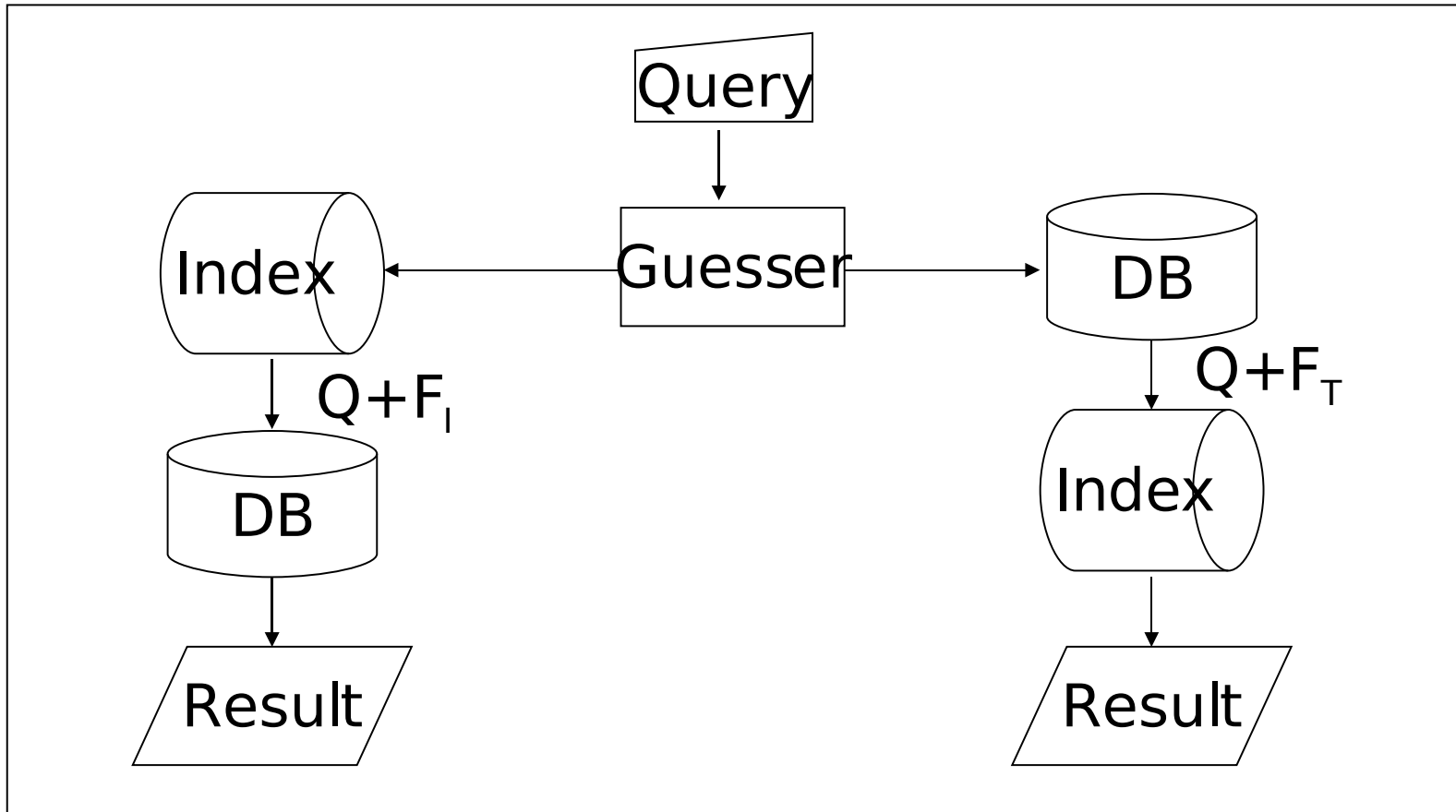
Strategy DB & Index (2)



or



Strategy DB & Index (2)



Strategy

Index only

Update the full text index by adding the tags as new tokens.

Advantages: full usage of ranking, matching and normalisation possible, maintain one world, no code/rule duplication

Disadvantages: Incremental updates of inverted lists are expensive in current inverted index structures and block queries for a short time

Nevertheless

an approach with lucene

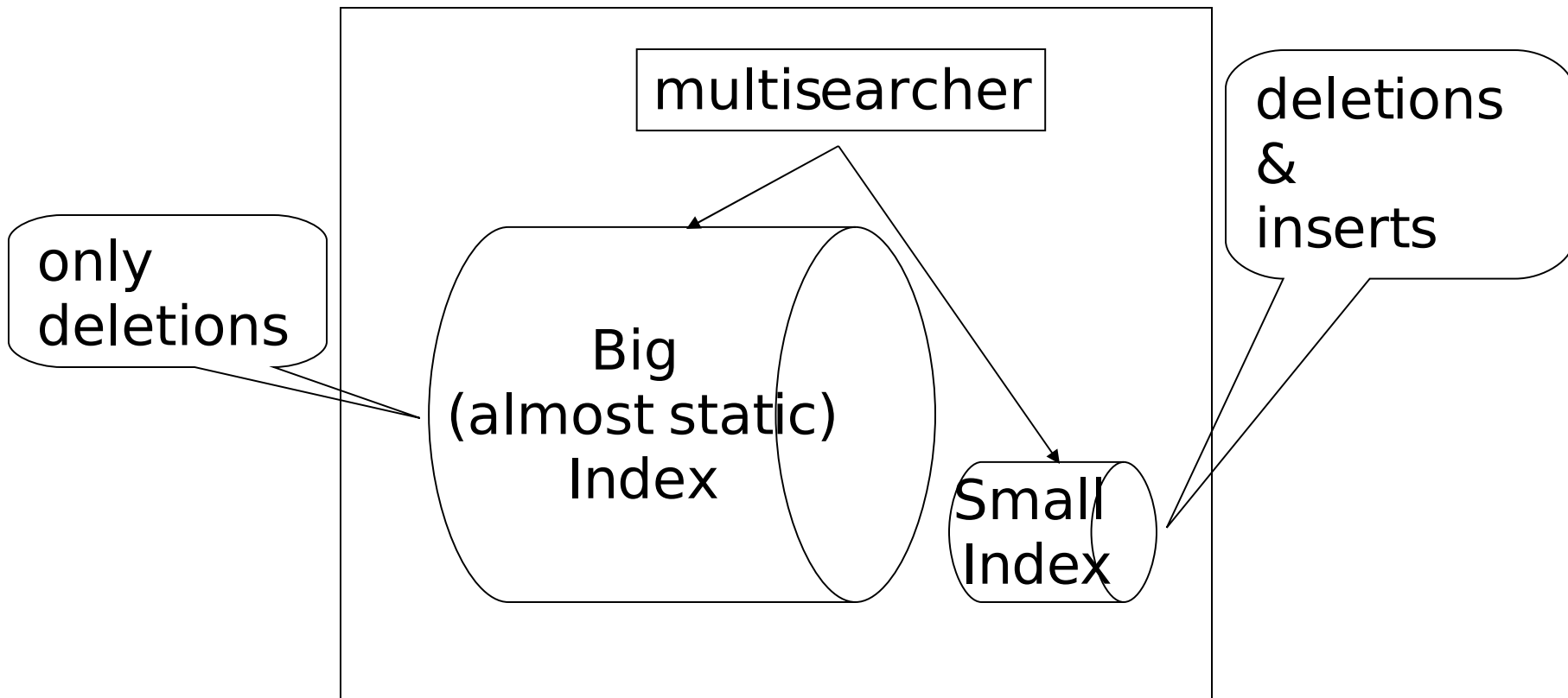
Preconditions

- Take an existing full text search engine
- Do not change its core or data structure
- Allow all sorts of changes (not just adding tags)

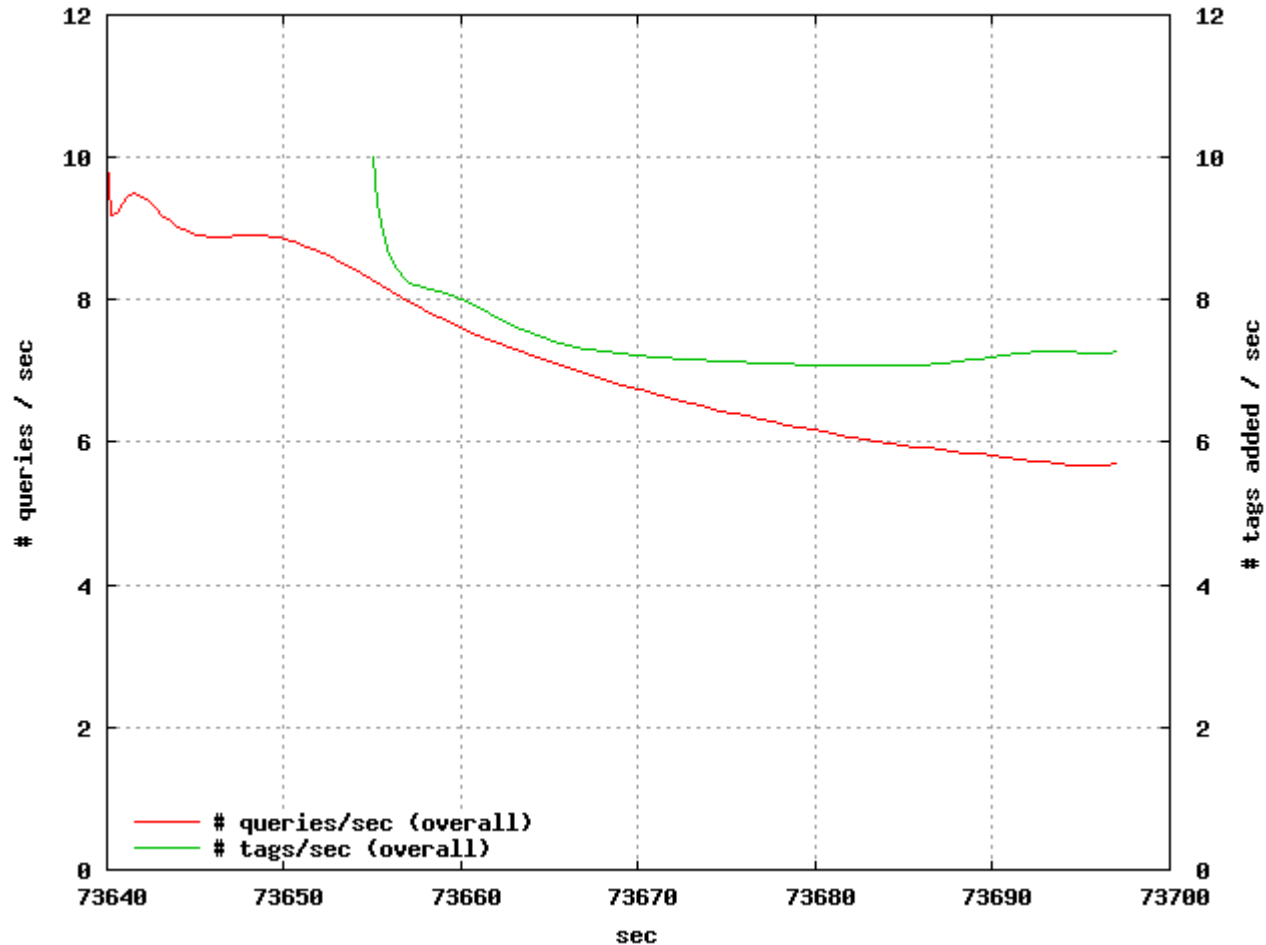
Some Assumptions

- The number of changes \ll the number of retrievable items
- The number of changes \ll the number of queries
- The index is not your main data repository

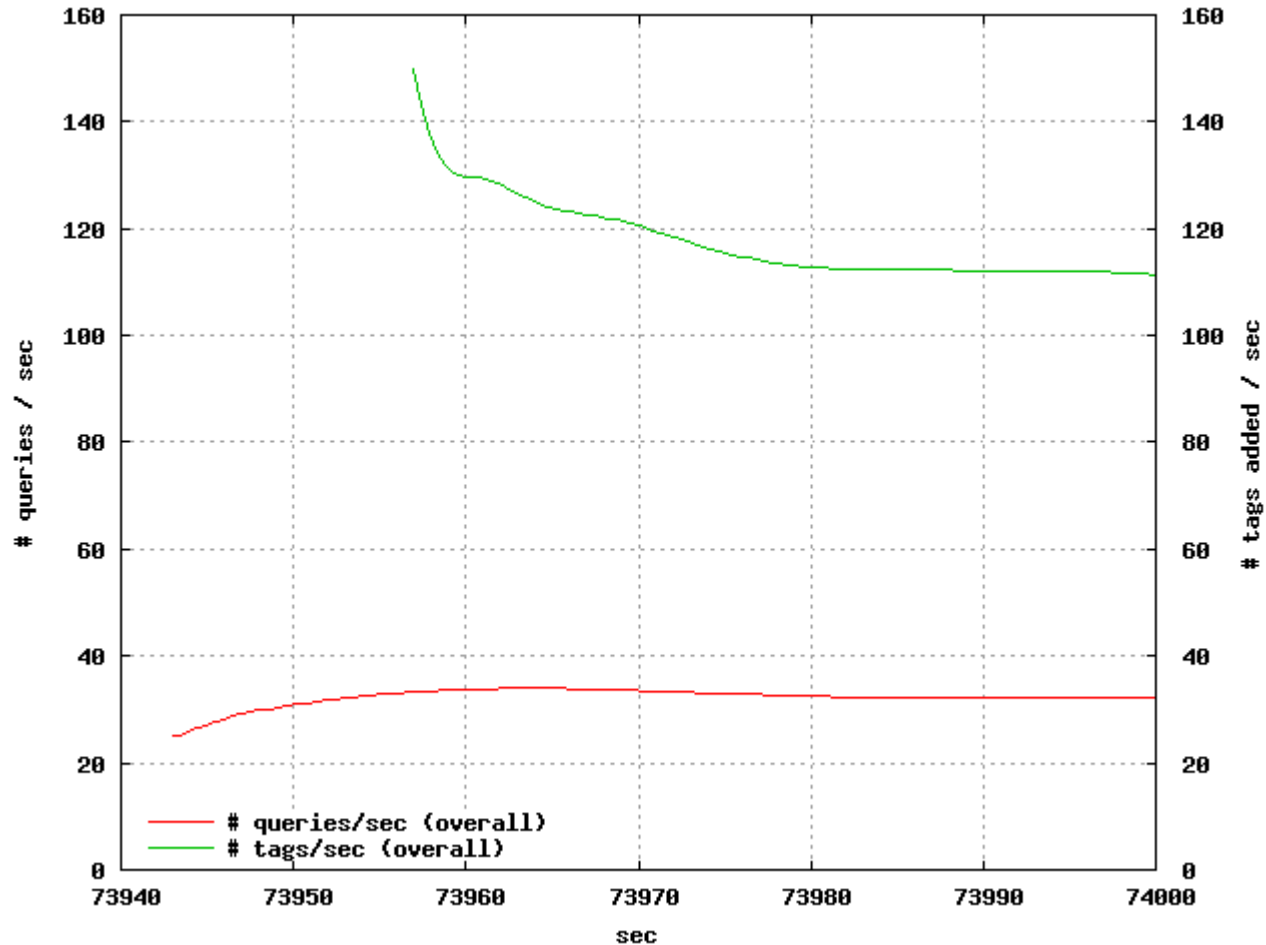
The Basic Layout



Let's be brave/stupid



Let's be brave/stupid



Let's
be brave/stupid

The Way we look at things

Moses: ***The Law*** is everything (BC)

The Way we look at things

Moses: ***The Law*** is everything (BC)

Jesus: ***Love*** is everything (@C)

The Way we look at things

Moses: **The Law** is everything (BC)

Jesus: **Love** is everything (@C)

Marx: **Money** is everything (2 Centuries ago)

The Way we look at things

Moses: **The Law** is everything (BC)

Jesus: **Love** is everything (@C)

Marx: **Money** is everything (2 Centuries ago)

Freud: **Sex** is everything (1 Century ago)

The Way we look at things

Moses: **The Law** is everything (BC)

Jesus: **Love** is everything (@C)

Marx: **Money** is everything (2 Centuries ago)

Freud: **Sex** is everything (1 Century ago)

Google: **Search** is everything (10 years ago)

The Way we look at things

Moses: **The Law** is everything (BC)

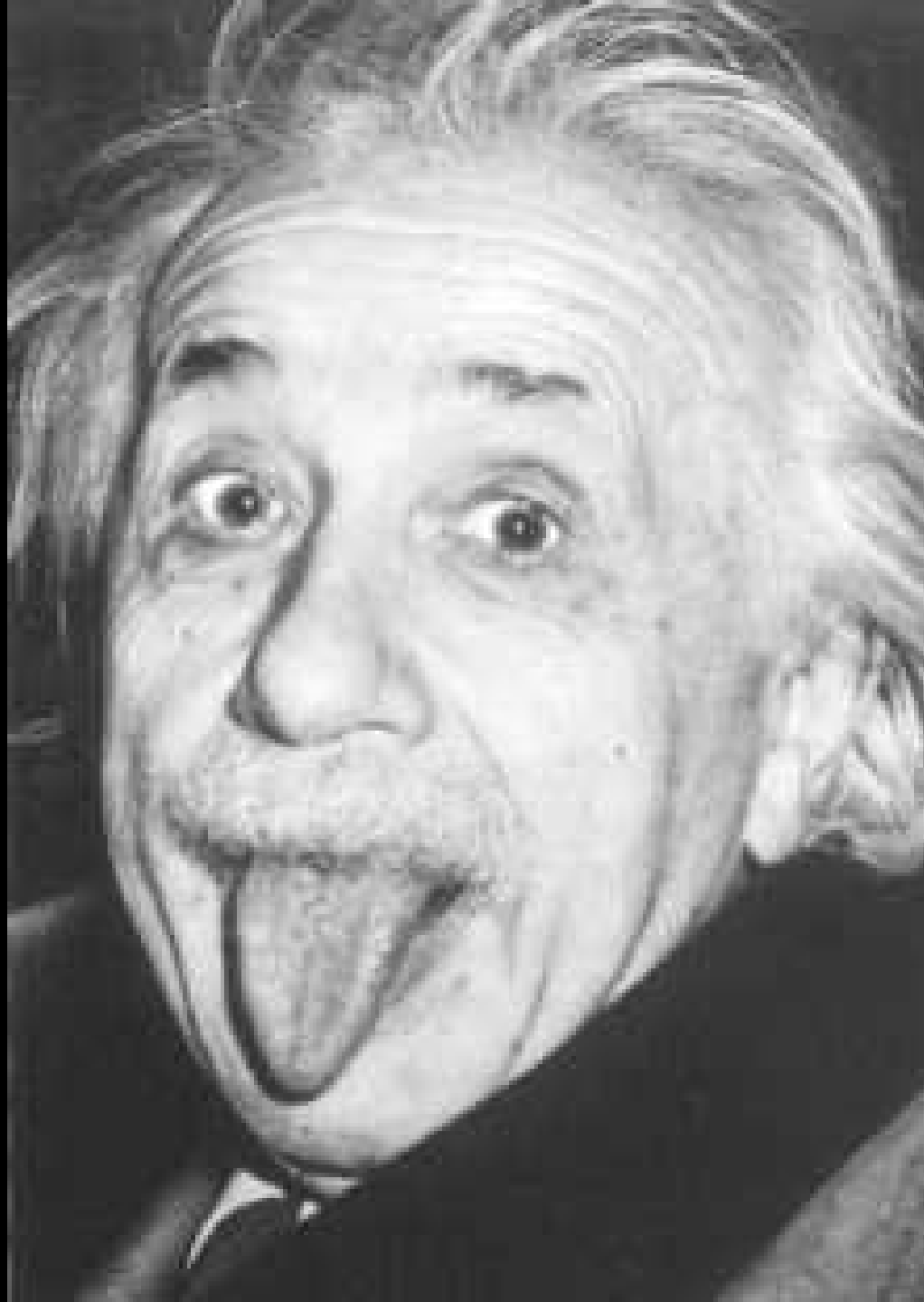
Jesus: **Love** is everything (@C)

Marx: **Money** is everything (2 Centuries ago)

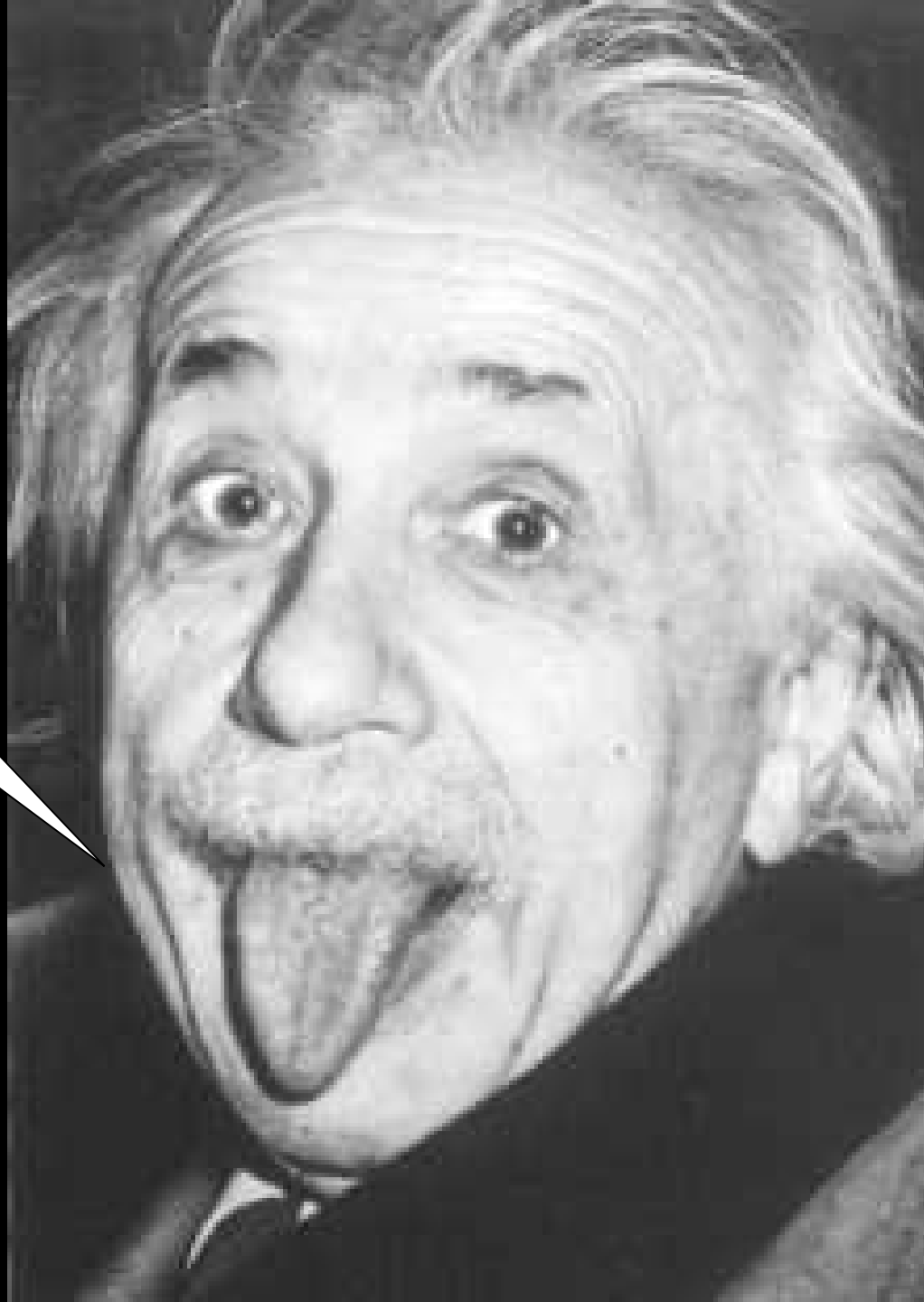
Freud: **Sex** is everything (1 Century ago)

Google: **Search** is everything (10 years ago)

Del.icio.us: **Tagging** is everything (4years ago)



“Everything
is *relative*”



Thank you

