

APACHE
LUCENE
EUROCON



Natural language search in Solr

Tommaso Teofili, Sourcesense
t.teofili@sourcesense.com, October 19th 2011

Presented by

lucid
IMAGINATION



Agenda

- An approach to natural language search in Solr
- Main points
 - Solr-UIMA integration module
 - Custom Lucene analyzers for UIMA
 - OSS NLP algorithms in Lucene/Solr
 - Orchestrating blocks to build a sample system able to understand natural language queries
- Results

My Background



- Software engineer at Sourcesense
 - Enterprise search consultant
- Member of the Apache Software Foundation
 - UIMA
 - Clerezza
 - Stanbol
 - DirectMemory
 - ...

Google in '99

Google!

Search the web using Google!

Google Search

I'm feeling lucky


[Why use Google!](#)
[Press about Google!](#)
[Help!](#)

[Company Info](#)
[Jobs at Google](#)
[Google! Logos](#)
[Making Google! the Default](#)

Copyright ©1999 Google Inc.

Google today

+Tommaso **Web** Images Videos Maps News Gmail More ▾

Google 

Search

Everything


Showing results for [stock google](#)
Search instead for [stokc googl](#)

GOOG - Google Inc. (NASDAQ)
[Google Finance](#) [Yahoo Finance](#) [MSN Money](#) [DailyFinance](#) [CNN Money](#) [Reuters](#)

546.41 **+9.24** (1.72%) Oct 11 12:00pm ET

Open: 533.46	Volume: 1,381,894
High: 546.80	Avg Vol: 2,947,000
Low: 533.46	Mkt Cap: 176.43B


[Disclaimer](#)



Any time

[Past hour](#)



[Past 24 hours](#)

Google Finance: Stock market quotes, news, currency conversions ...
www.google.com/finance 


Get real-time **stock** quotes & charts, financial news, currency conversions, or track your

Google today

+Tommaso **Web** Images Videos Maps News Gmail More ▾

Search About 1,460,000,000 results (0.39 seconds)

Everything ▶ [25 Greatest Science Books of All Time | DISCOVER Magazine](#)
[discovermagazine.com/2006/dec/25-greatest-science-books](#) 
8 Dec 2006 – DISCOVER presents the essential reading list for anyone interested in **science**. Visit Discover Magazine to read this article and other exclusive ...

Images

Maps

Videos

News


Shopping


More


All results

Timeline

More search tools

[Amazon.com: Books on Science and Religion](#)
[www.amazon.com/Books-on-Science-and.../R30N8XYUJER5FC](#) 
These are a list of **books** based on the synergies between **science** and spirituality.

[NSTA :: Outstanding Science Trade Books for Students K-12](#)
[www.nsta.org › Publications and Products](#) 
Reading **science** trade **books** is the perfect way for students to build literacy skills while learning **science** content. The **books** that appear in these lists were ...

[Reviews of Books about Science](#)
[www.scibooks.org/](#) 
Reviews of **books about science**. Reviews by Phillip Manning published in The Raleigh News and Observer and other publications.

The Challenge

- Improved recall/precision
 - ‘articles about science’ (concepts)
 - ‘movies by K. Spacey’ vs ‘movies with K. Spacey’
- Easier experience for non-expert users
 - ‘people working at Google’ - ‘cities near London’
- Horizontal domains (e.g. Google)
- Vertical domains

Hurdles

- understanding documents' text/user queries
- extract domain-specific/wide entities and concepts
- index/search performance

Use Case

- search engine for an online movies magazine
- Solr based
- non technical users
- time / cost
 - Solr 3.x setup : 2 mins
 - NLS setup / tweak : 5 days
- expecting
 - improved recall / precision
 - more time (clicks) on site (\$)

Online movies magazine

MOVIE INFO

Contemporary high schooler Marty McFly (Michael J. Fox) doesn't have the most pleasant of lives. Browbeaten by his principal at school, Marty must also endure the acrimonious relationship between his nerdy father (Crispin Glover) and his lovely mother (Lea Thompson), who in turn suffer the bullying of middle-aged jerk Biff (Thomas F. Wilson), Marty's dad's supervisor. The one balm in Marty's life is his friendship with eccentric scientist Doc (Christopher Lloyd), who at present is working on a time machine. Accidentally zapped back into the 1950s, Marty inadvertently interferes with the budding romance of his now-teenaged parents. Our hero must now reunite his parents-to-be, lest he cease to exist in the 1980s. It won't be easy, especially with the loutish Biff, now also a teenager, complicating matters. Beyond its dazzling special effects, the best element of Back to the Future is the performance of Michael J. Fox, who finds himself in the quagmire of surviving the white-bread 1950s with a hip 1980s mindset. Back to the Future cemented the box office success of the film's director, Robert Zemeckis, who went on to helm two more films: Forrest Gump and Cast Away. Hal Erickson, Rovi

PG, 1 hr. 56 min.

In Theaters

Action & Adventure, Science Fiction & Fantasy, On DVD: Dec 17, 2002

Back to the Future :: Movie

- [Search in Reuters](#)
- [Search in Google](#)
- [Search in Wikipedia](#)
- [Search in Technorati](#)

General approach

- Natural language processing
- Processing documents at indexing time
 - document text analysis
 - write enriched text in (dedicated) fields
 - add custom types / payloads to terms
- Processing queries at searching time
 - query analysis
 - higher boosts to entities/concepts
 - in-sentence search
 - ...

NLP

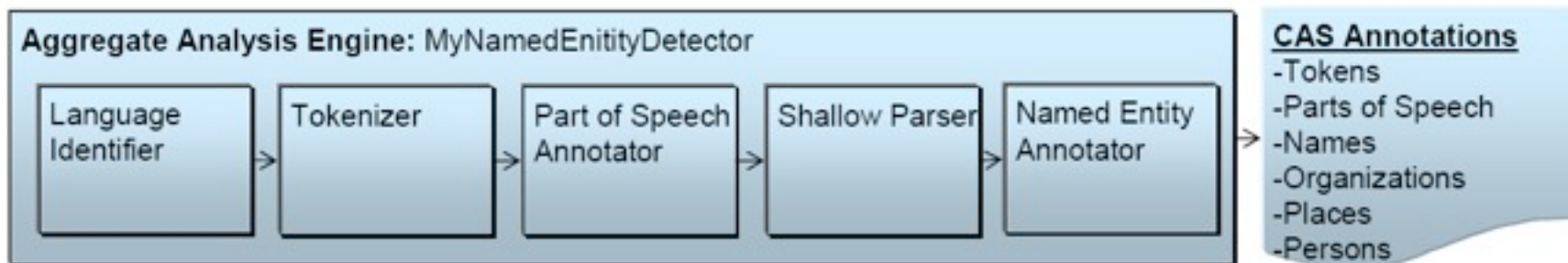
- AI discipline
 - **Computers understanding and managing information written in human language**
- analyze text at various levels
- incrementally enrich / give structure
- extract concepts and named entities

Technical detail

- NLP algorithms plugged via Apache UIMA
- Indexing time
 - UpdateProcessor plugin (solr/contrib/uima)
 - Custom tokenizers/filters
- Search time
 - Custom QParserPlugin

Why Apache UIMA?

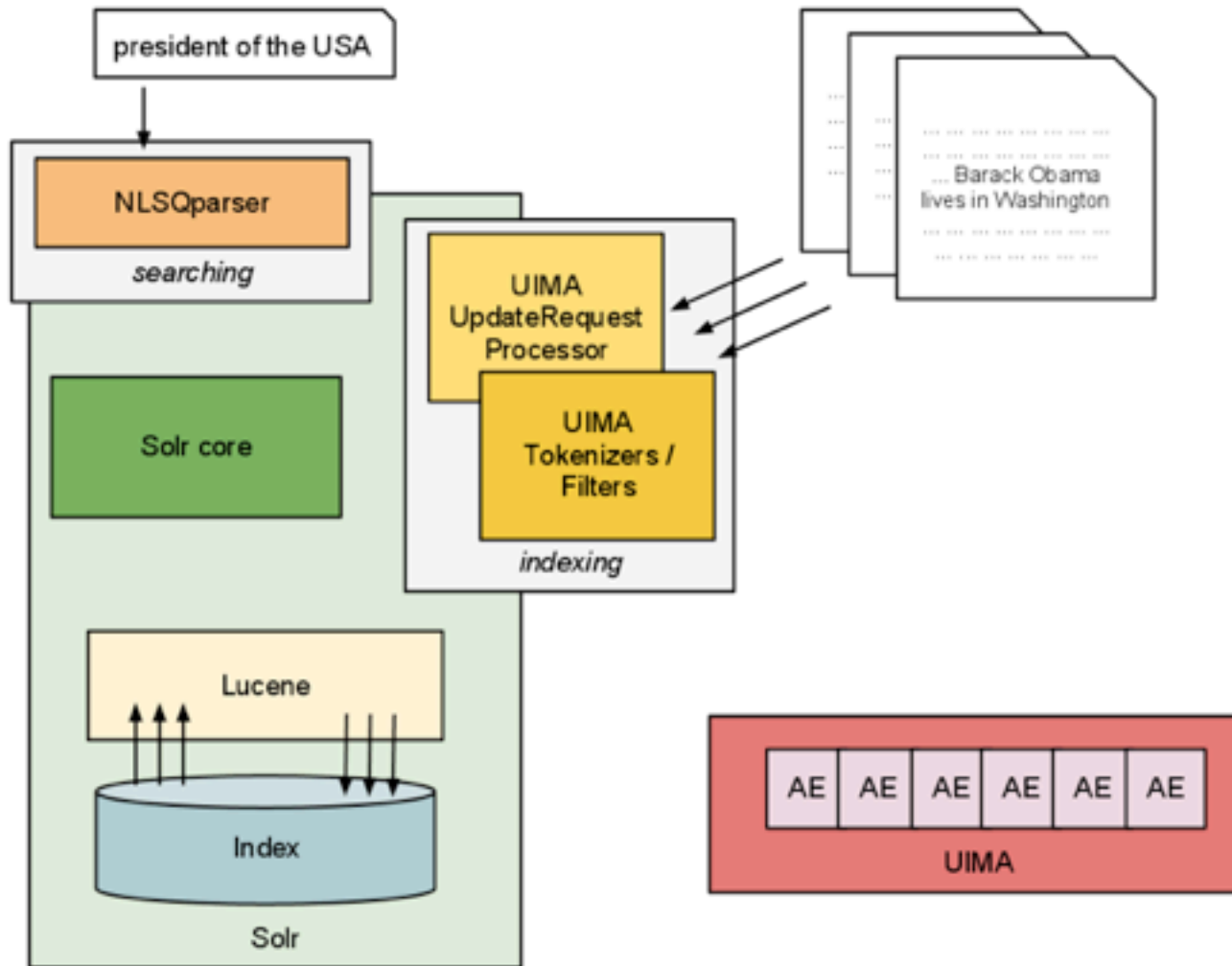
- OASIS standard for UIM
- TLP since March 2010
- Deploy pipelines of Analysis Engines
- AEs wrap NLP algorithms
- Scaling capabilities



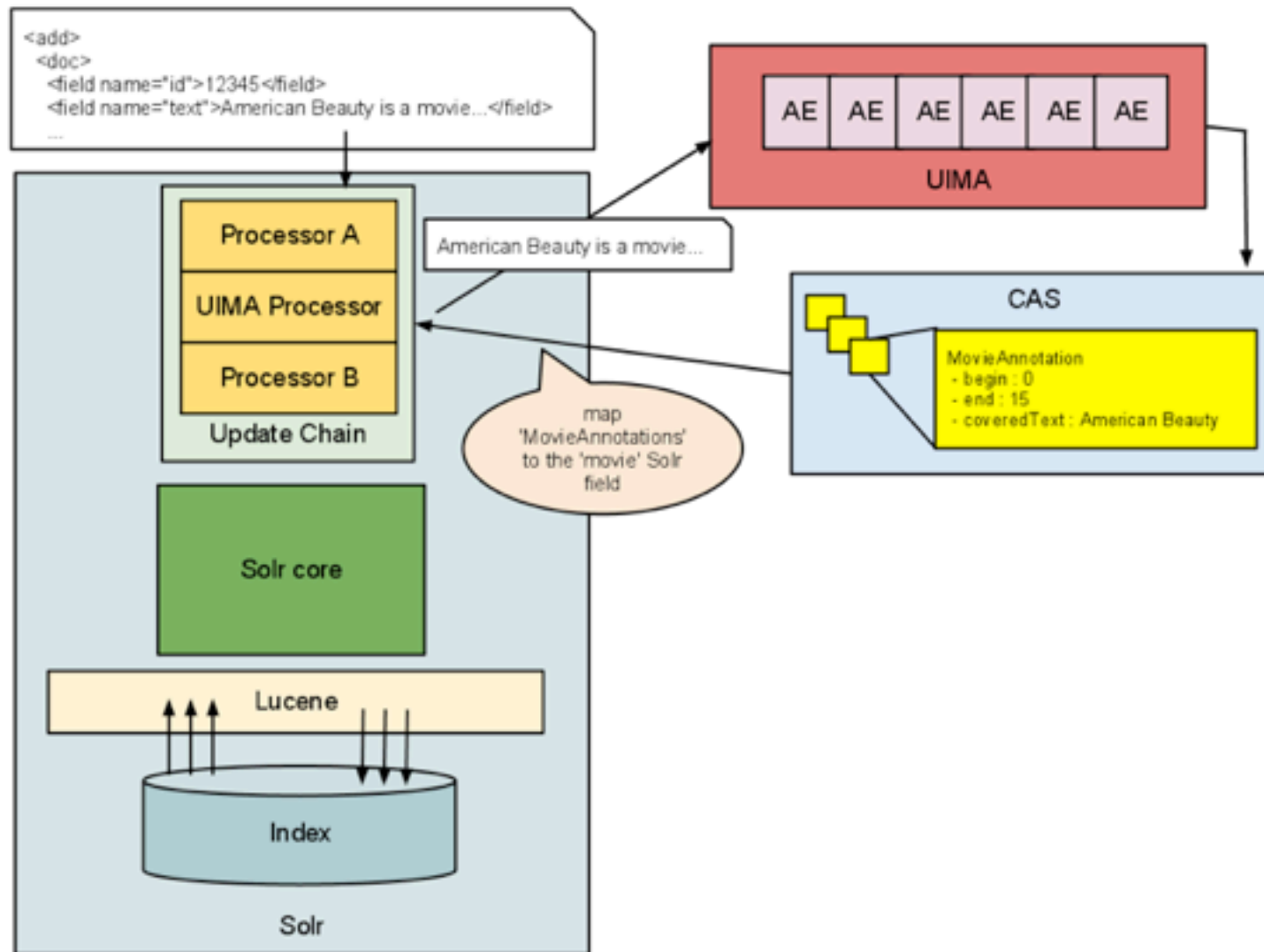
NLP and OSS

- Sentence Split
 - OpenNLP, UIMA Addons, StanfordNLP
- PoS tagging
 - OpenNLP, UIMA Addons, StanfordNLP
- Chunking/Parsing
 - OpenNLP, StanfordNLP
- NER
 - OpenNLP, UIMA Addons, Stanbol, StanfordNLP
- Clustering/Classifying
 - Mahout, OpenNLP, StanfordNLP
- ...

Solr NLS architecture



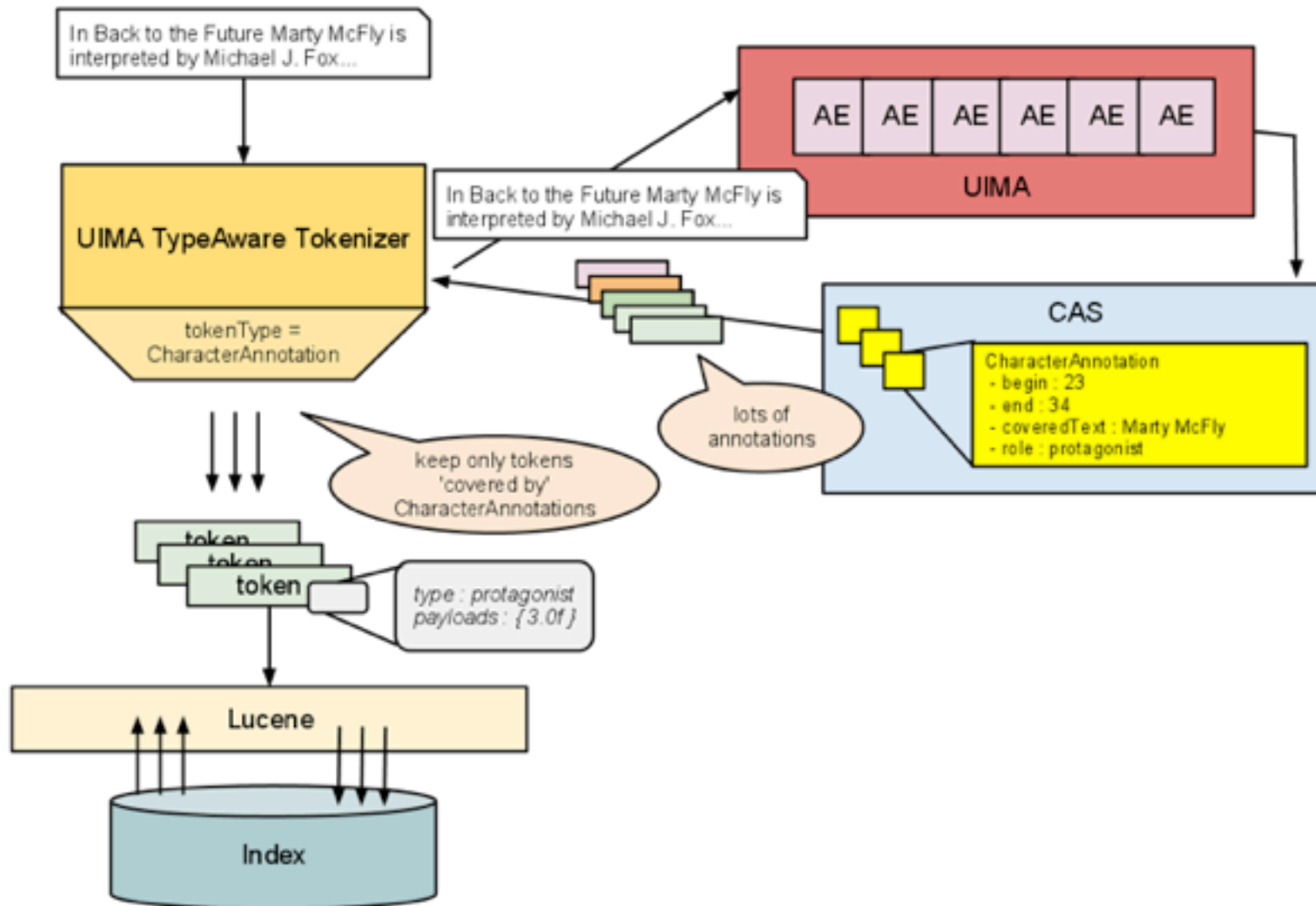
UIMA Update Processor



Lucene analysis & UIMA

- Type : denote lexical types for tokens
- Payload : a byte array stored at each term position
- tokenize / filter tokens covered by a certain annotation type
- store UIMA annotations' features in types / payloads

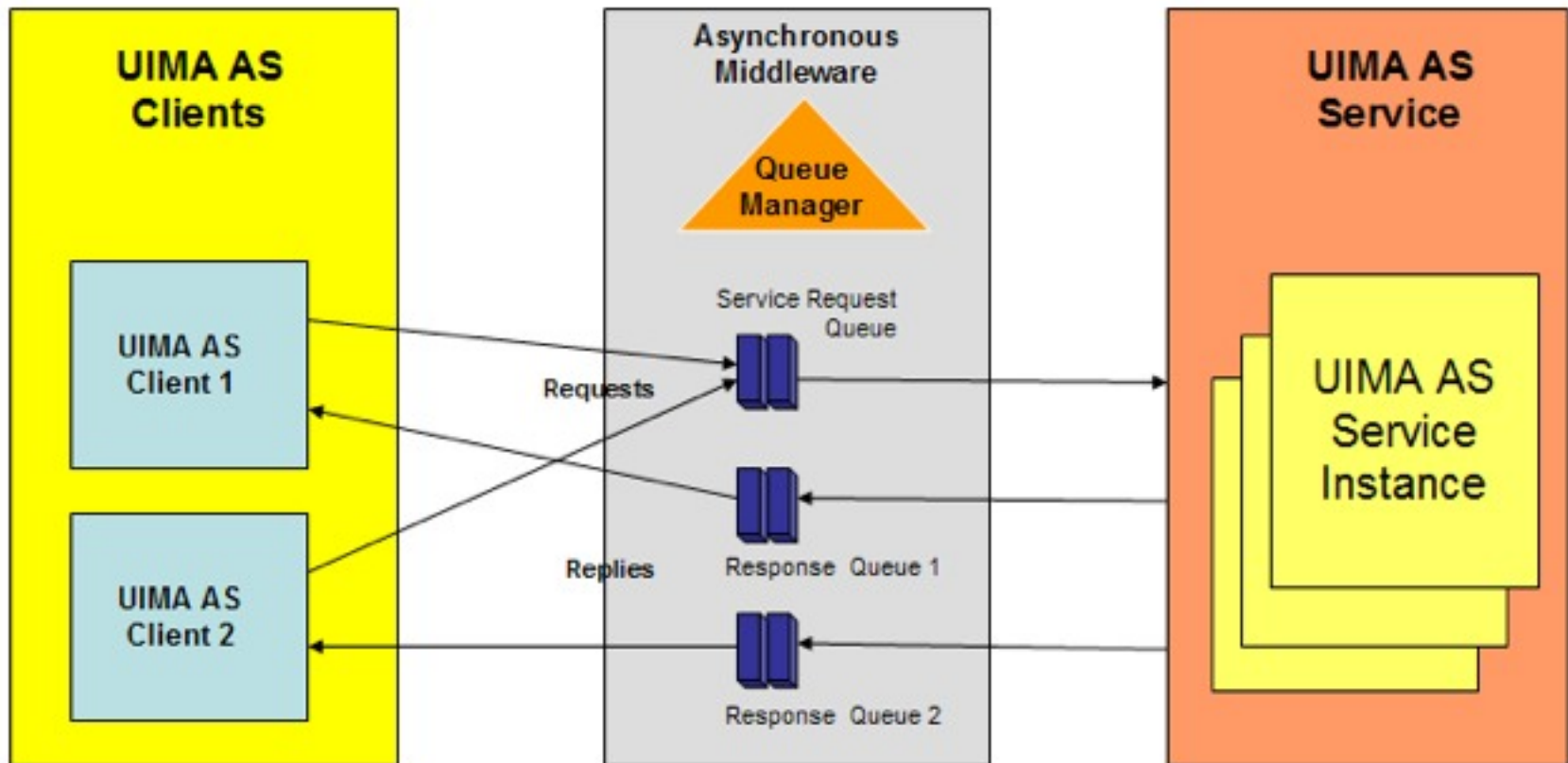
UIMA type-aware tokenizer



Solr NLS QParser

- analyze user query
- extract (and query on) concepts / entities
- use types/PoS in the query for
 - **boosting terms**
 - **synonym expansion**
- search within sentences
- faceting / clustering using entities
- identify 'place queries' and expand Solr spatial queries (for filtering / boosting)

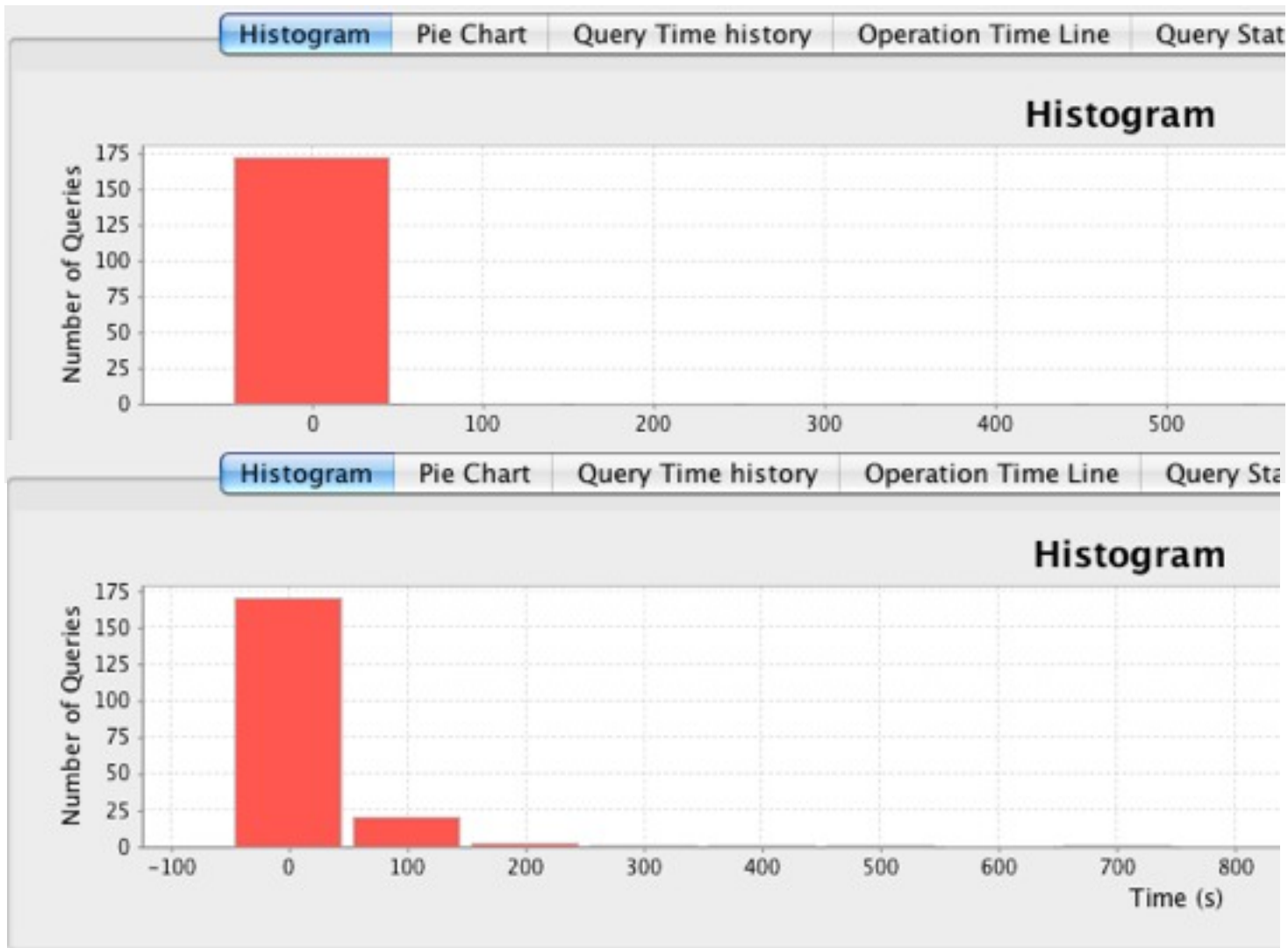
Scaling architecture



Performance

- basic (in memory)
 - slower with NRT indexing
 - search could be significantly impacted
- ReST (SimpleServer)
 - faster
 - need to explicitly digest results
- UIMA-AS
 - fast also with NRT indexing
 - fast search
 - scales nicely with lots of data

DisMax vs NLS



Wrap up

- general purpose architecture
- generally improved recall / precision
- NLP algorithms accuracy make the difference
- lots of OSS alternatives
- performances can be kept good

Sources

■ Resources

- <http://svn.apache.org/repos/asf/lucene/dev/trunk/solr/contrib/uima/>
- <https://github.com/tteofili/le11-nls>

■ Links

- <http://wiki.apache.org/solr/SolrUIMA>
- <http://googleblog.blogspot.com/2010/01/helping-computers-understand-language.html>

Thanks

- <http://www.sourcesense.com>
- t.teofili@sourcesense.com
- [@tteofili](#)