

Apache Lucene in Industry

Charlie Hull - Managing Director
8th September 2016
Lucene4IR Workshop, Glasgow

charlie@flax.co.uk
www.flax.co.uk/blog
+44 (0) 8700 118334
Twitter: @FlaxSearch



- ◆ We build, tune and support fast, accurate and highly scalable search, analytics and Big Data applications
- ◆ We use (and create) **open source** software
- ◆ We're independent, honest and have 15+ years experience
- ◆ We also:
 - Run and attend many events & conferences
 - Write extensively about search & related matters
 - Train and mentor



Ancient history


- ◆ 1999-2001 – Muscat & Brightstation

Ancient history

- ◆ 1999-2001 – Muscat & Brightstation

- ◆ 1999 - Doug Cutting writes 

Ancient history

- ♦ 1999-2001 – Muscat & Brightstation
- ♦ 1999 - Doug Cutting writes 
- ♦ 2001 – Flax founded as Lemur Consulting Ltd.



Xapian

- Explaining open source search
- Projects



More ancient history

◆ 2005-7

– The rise of  &



More ancient history

◆ 2005-7

– The rise of  &



– More  Xapian projects



(Slightly less) ancient history

◆ 2010-11

– OK,  wins


(Slightly less) ancient history

◆ 2010-11

- OK,  wins
- We stop having to explain open source search

(Slightly less) ancient history

♦ 2010-11


- OK,  wins
- We stop having to explain open source search
- Projects



CAMBRIDGE 

(Slightly less) ancient history

♦ 2010-11

- OK,  wins
- We stop having to explain open source search
- Projects



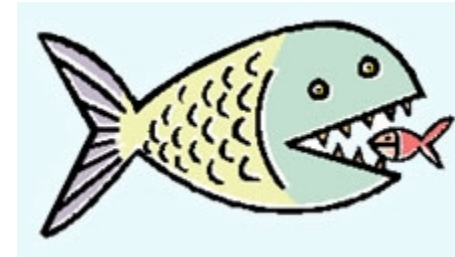
CAMBRIDGE 

- Partnership with Lucid Imagination
 - Founded by Lucene committers
 - VC funded
 - Offers support, training and packaged search



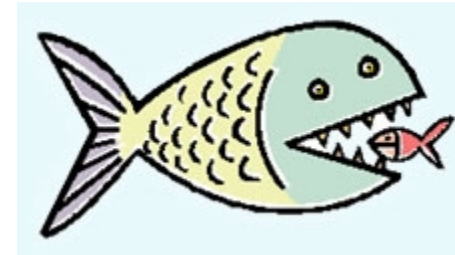
...in the meantime...

- ♦ Massive change in the enterprise search market
 - Microsoft buys FAST
 - Oracle buys Endeca
 - IBM buys Vivisimo
 - Dassault buys Exalead
 - Hewlett Packard buys Autonomy (that went well!)

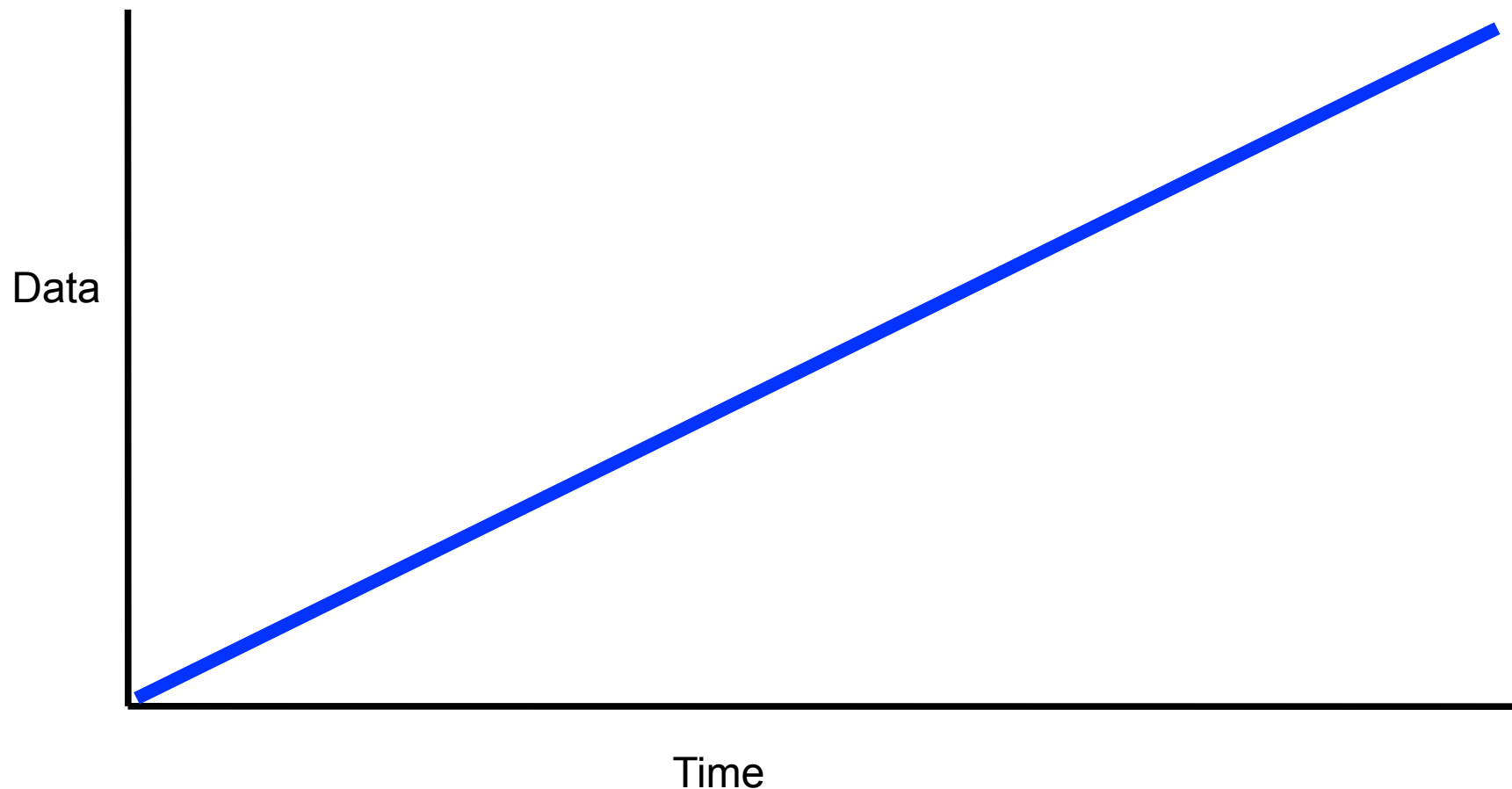


...in the meantime...

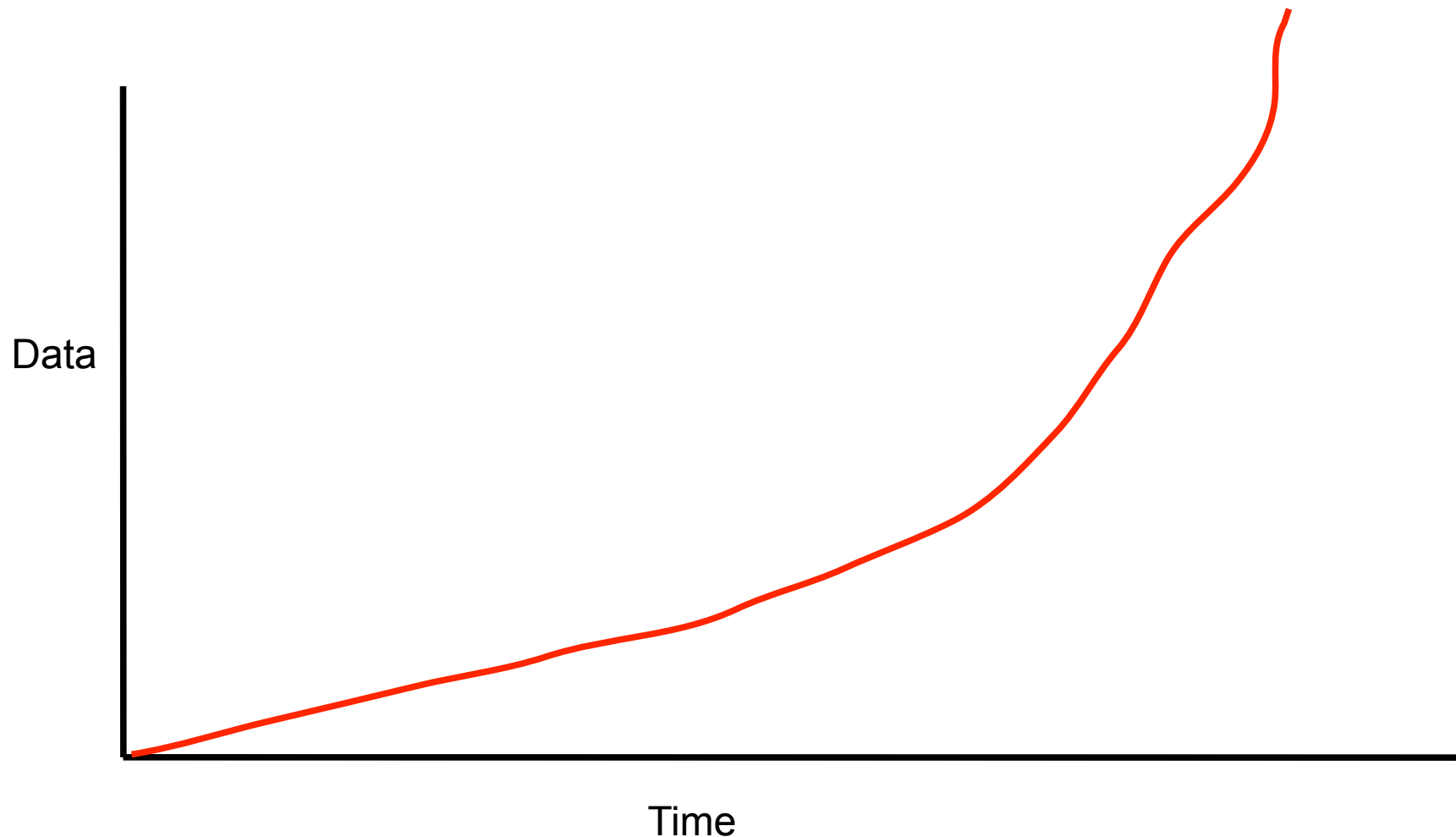
- ♦ Massive change in the enterprise search market
 - Microsoft buys FAST
 - Oracle buys Endeca
 - IBM buys Vivisimo
 - Dassault buys Exalead
 - Hewlett Packard buys Autonomy (that went well!)
- ♦ Why?
 - We suspect the rise of open source search is a factor
 - Powerful search is no longer only available to those with 6 figure budgets
 - Everyone's talking Big Data...



Charlie's All Purpose Big Data Graph



Charlie's All Purpose Big Data Graph



And now

- ◆ 2012-16

- Leading Lucene specialists & committers

And now

◆ 2012-16

- Leading Lucene specialists & committers
- Run Lucene hackdays & meetups

And now

◆ 2012-16

- Leading Lucene specialists & committers
- Run Lucene hackdays & meetups
- Clients



CabinetOffice



GIG
CYMRU
NHS
WALES

Gwasanaeth
Gwybodeg
Informatics
Service



CAMBRIDGE
UNIVERSITY PRESS

And now

◆ 2012-16

- Leading Lucene specialists & committers
- Run Lucene hackdays & meetups
- Clients



CabinetOffice



GIG
CYMRU
NHS
WALES

Gwasanaeth
Gwybodeg
Informatics
Service



CAMBRIDGE
UNIVERSITY PRESS

- Still a  **Lucidworks** partner

Why Apache Lucene?

- ◆ The most widely used open source search engine

Why Apache Lucene?

- ◆ The most widely used open source search engine
- ◆ Hugely flexible, feature-rich, scalable & performant

Why Apache Lucene?

- ◆ The most widely used open source search engine
- ◆ Hugely flexible, feature-rich, scalable & performant
- ◆ Very large and healthy community, many books & examples
 - Supported by the Apache Foundation

Why Apache Lucene?

- ◆ The most widely used open source search engine
- ◆ Hugely flexible, feature-rich, scalable & performant
- ◆ Very large and healthy community, many books & examples
 - Supported by the Apache Foundation
- ◆ Used by some of the world's largest companies

SONY

SIEMENS



We don't use Lucene directly...

- ◆ We use a search server
 - Apache Solr
 - Mature & stable
 - Highly scalable with SolrCloud
 - Many interfaces & plugins



We don't use Lucene directly...

◆ We use a search server

– Apache Solr

- Mature & stable
- Highly scalable with SolrCloud
- Many interfaces & plugins



– Elasticsearch

- Slightly easier to get started with
- Lots of cool analytics & visualisations (ELK Stack)
- Open source – but not open development



But hang on....



Lemur



Terrier

But hang on....

 Lemur  Terrier

- Sorry, no one in industry has even heard of them!

The sort of projects we do



- ◆ Upgrading a media monitoring system
 - Very old (2001) media monitoring system based on Verity
 - Slightly less old installation of Autonomy IDOL used for Infomedia's Media Archive
 - Drawbacks:
 - Verity at almost max capacity needing constant attention
 - Old and complex workflow for receiving and processing articles
 - Different platforms for monitoring and archive searches meant we were 'bi-lingual', using two different query languages in-house.
 - Verity no longer supported by the owning company (HP)
 - Verity not scalable

The sort of projects we do



♦ Autonomy IDOL replaced by Solr

- Parses IQL to Lucene queries
- SolrCloud distributes the index & queries across several servers
- Setup: 75 million documents hosted on 8 servers, 6 cores/24GB memory and 125 GB storage per server. This setup is doubled to have full redundancy
- Features added to standard Solr by Flax:
 - Custom highlighting, Framework to handle multiple languages, Extended error logging, Cluster management, Performance enhancements for complex wildcard queries

The sort of projects we do



♦ Autonomy IDOL replaced by Solr

- Parses IQL to Lucene queries
- SolrCloud distributes the index & queries across several servers
- Setup: 75 million documents hosted on 8 servers, 6 cores/24GB memory and 125 GB storage per server. This setup is doubled to have full redundancy
- Features added to standard Solr by Flax:
 - Custom highlighting, Framework to handle multiple languages, Extended error logging, Cluster management, Performance enhancements for complex wildcard queries

♦ Verity replaced by **Flax Monitor**

- Parses IQL to Lucene queries
- Runs on 2 servers
- Uses Luwak, Flax's 'stored search' library:
 - Built on Apache Lucene (as is Solr)
 - Also used by Bloomberg, Booz Allen Hamilton & others
 - In use for 1m stored searches (some 250k characters), 1m stories/day
 - 40x faster than Elasticsearch Percolator
 - Open source at <https://github.com/flaxsearch/luwak>

The sort of projects we do



- ◆ Improving bulk indexing with Elasticsearch
 - More than 1 billion documents from web crawling
 - Business information for emerging markets
 - Stored in Hadoop and indexed by a 10-node ES cluster
 - Indexing was taking 30 days using `elasticsearch-hadoop`

The sort of projects we do



◆ Improving bulk indexing with Elasticsearch

- More than 1 billion documents from web crawling
- Business information for emerging markets
- Stored in Hadoop and indexed by a 10-node ES cluster
- Indexing was taking 30 days using `elasticsearch-hadoop`

◆ Our solution:

- Reduce the map jobs
- Turn off Elasticsearch merge throttling
- Write our own own Hadoop OutputFormat implementation using an Elasticsearch TransportClient and BulkProcessor.
- Indexing now takes 17 hours

The sort of projects we do



- ◆ A 18 month project to improve search for bioinformatics
 - Funded by BBSRC
 - Two days a week on site at EBI



The sort of projects we do



◆ A 18 month project to improve search for bioinformatics

- Funded by BBSRC
- Two days a week on site at EBI



◆ Developed open source enhancements and plugins for Solr and Elasticsearch

- Better faceting
- Xjoin
- Ontology indexing

The sort of projects we do



- ◆ A 18 month project to improve search for bioinformatics

- Funded by BBSRC
- Two days a week on site at EBI



- ◆ Developed open source enhancements and plugins for Solr and Elasticsearch

- Better faceting
- Xjoin
- Ontology indexing

- ◆ Ran workshop events, papers at conferences, blogs

The sort of projects we do



- ◆ A 18 month project to improve search for bioinformatics

- Funded by BBSRC
- Two days a week on site at EBI



- ◆ Developed open source enhancements and plugins for Solr and Elasticsearch

- Better faceting
- Xjoin
- Ontology indexing

- ◆ Ran workshop events, papers at conferences, blogs

- ◆ Created links with other institutions e.g. NCBI

Home truths

Home truths

- ◆ Open source does not mean cheap

Home truths

- ◆ Open source does not mean cheap
- ◆ Most search engines are the same

Home truths

- ◆ Open source does not mean cheap
- ◆ Most search engines are the same
- ◆ Complex features are seldom used & often confusing

Home truths

- ◆ Open source does not mean cheap
- ◆ Most search engines are the same
- ◆ Complex features are seldom used & often confusing
- ◆ Search testing is rarely comprehensive

Home truths

- ◆ Open source does not mean cheap
- ◆ Most search engines are the same
- ◆ Complex features are seldom used & often confusing
- ◆ Search testing is rarely comprehensive
- ◆ Good search developers are hard to find

So what can we do better?

So what can we do better?

- ◆ Learn what works in industry

So what can we do better?

- ◆ Learn what works in industry
- ◆ Improve Lucene with ideas from academia – faster!
 - Learning to Rank, Terrier & Lucene

So what can we do better?

- ◆ Learn what works in industry
- ◆ Improve Lucene with ideas from academia – faster!
 - Learning to Rank, Terrier & Lucene
- ◆ Improve testing of Lucene-based search engines
 - TREC
 - Real-world data
 - Test-based relevance tuning

So what can we do better?

- ◆ Learn what works in industry
- ◆ Improve Lucene with ideas from academia – faster!
 - Learning to Rank, Terrier & Lucene
- ◆ Improve testing of Lucene-based search engines
 - TREC
 - Real-world data
 - Test-based relevance tuning
- ◆ Develop more practitioners & committers
 - Lucene, Solr and Elasticsearch are highly marketable skills

Hackdays

- ◆ This one
- ◆ London Lucene Hackday, October 7th 2016
 - <http://www.meetup.com/Apache-Lucene-Solr-London-User-Group/>
- ◆ Boston Lucene Hackday, October 10th 2016
 - <http://www.meetup.com/New-England-Search-Technologies-NEST-Group>
 - Just before Lucene Revolution <http://lucenerevolution.org/>
- ◆ Run your own!

Thankyou!

Any questions?

charlie@flax.co.uk
www.flax.co.uk/blog
+44 (0) 8700 118334
Twitter: @FlaxSearch