

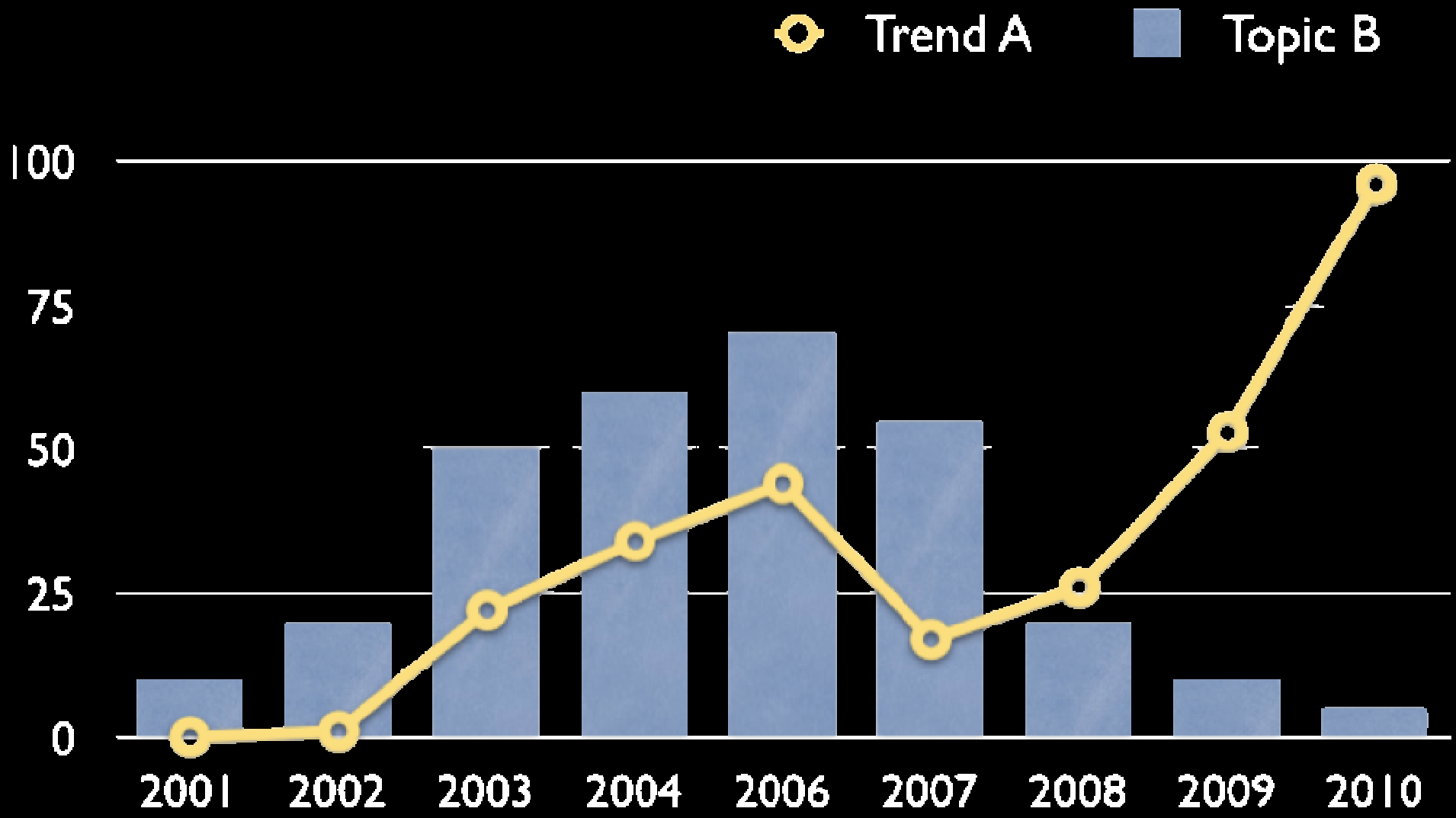
Search and Mine with *hindsight*

Julien Masanès
europarchive.org

Insight

M. Dontcheva, J. Fogarty, et D.S. Weld, "Zoetrope: interacting with the ephemeral"
Proceedings of the 21st annual ACM symposium on User interface software and
technology ACM, 2008.

Zoetrope

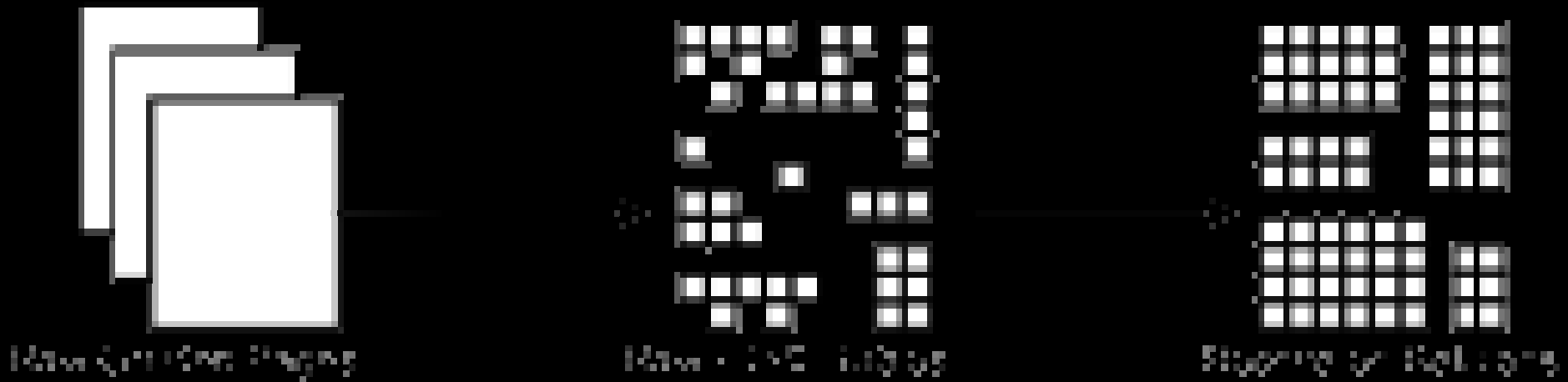


Webscale data

The web represents a unique source of accessible media content of all sorts, that a growing number of scientific communities, agencies and industries are starting to need to mine at large scale.

The ability to acquire, process and mine large scale data from the web is becoming a strategic advantage in many domains from business intelligence to epidemiological tracking and monitoring

'Webtables'



Using Over 14 billions pages from the web to extract tables.

M. Cafarella, A. Halevy, D. Wang, E. Wu, et Y. Zhang, "WebTables: exploring the power of tables on the web," Proc. VLDB Endow., vol. 1, 2008, pp. 549, 538.

Research engine

Key infrastructure to monitor and analyze the evolution of
networked media

More broadly, will become a key tool for research in more and
more domains:

Low noise signal of ecological evolution,

Economical trends,

Emergence of new research of new research fields

Tracking of reputation on the web

Who can do research on Webscale data?

Webscale is already proving to be a challenge for many search group as the infrastructure, the cost and the skills required represent a significant barrier to entry.

But when it comes to doing this through time, all but a few (mainly large search engines) can do it at all.

In other words, only large search engines (none being European) are able to do research at this scale, hence comforting their advance by developing and testing new algorithms for search, ranking, mining etc.

Research challenges (1)

Building in Europe of an open, neutral and sustainable **virtual observatory of the web** for research requires:

Large scale crawling, storage and indexing of web data (10+ Petabytes) not limited to text.

We know the TB, not the PB yet.

Research challenges (2)

Create a baseline **distributed analytics services** (large scale IE, NL, distributed and efficient processing and storage).

We need to standardize and define baseline in this domain to create platform for MMSE, social media research etc.

- Hadoop-style abstractions over internet-wide repository/processing clouds
- Optimized data placement (partitioning and replication) for analytics
- Distributed indices

Research challenges (3

Temporal indexing of significant characteristics of networked content
(from distribution to semantic)

Large spectrum of research in IE/IR, network topology etc.

Fast but not least: make this infrastructure **acceptable by society**
(respect privacy, transparency, IP rights)

Thanks

- Julien Masanès
- European Archive Foundation
- julien@europarchive.org

FIA Valencia, April 2010.