exalead 3s

Search Based Applications at the Confluence of Database and Search Engine Technology

Gregory Grefenstette

Structured Data

- Databases
- Transaction Logs
- Spreadsheets
- LDAP Directories
- ERP Exports
- Application Data

Unstructured Data

- Web Pages
- Email
- Office Documents
- RSS Feeds
- PDFs
- Blogs & Forums

exalead 3S

DATABASES

**Structured Data
Transaction
All tuples**

**Safe,
Precise, SQL
Slow**

**SEARCH
ENGINES**

**Text
Similarity
Ranking**

**Intuitive
Fast
Partial**

exalead

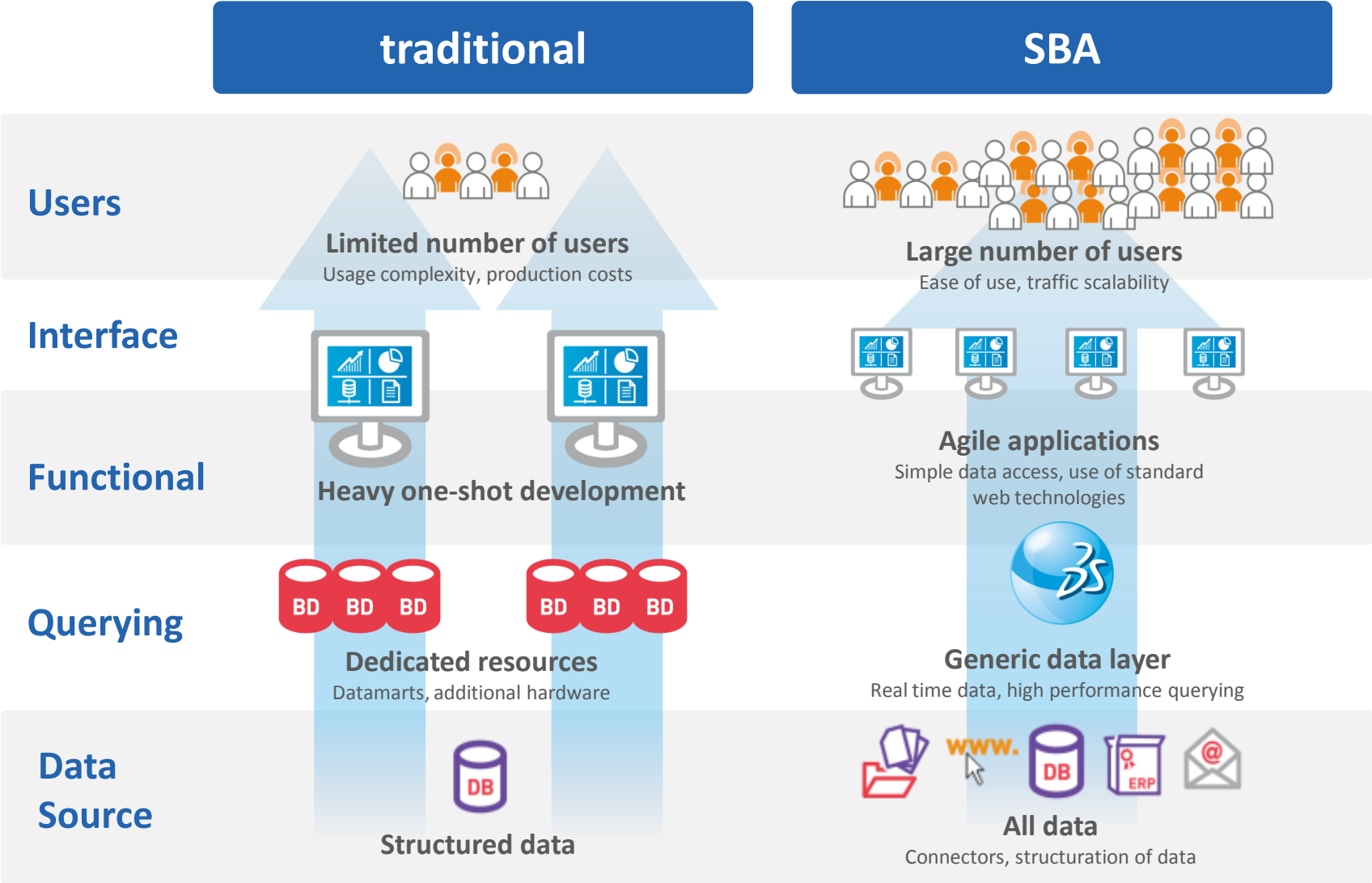# Organisational Data *all over the place*
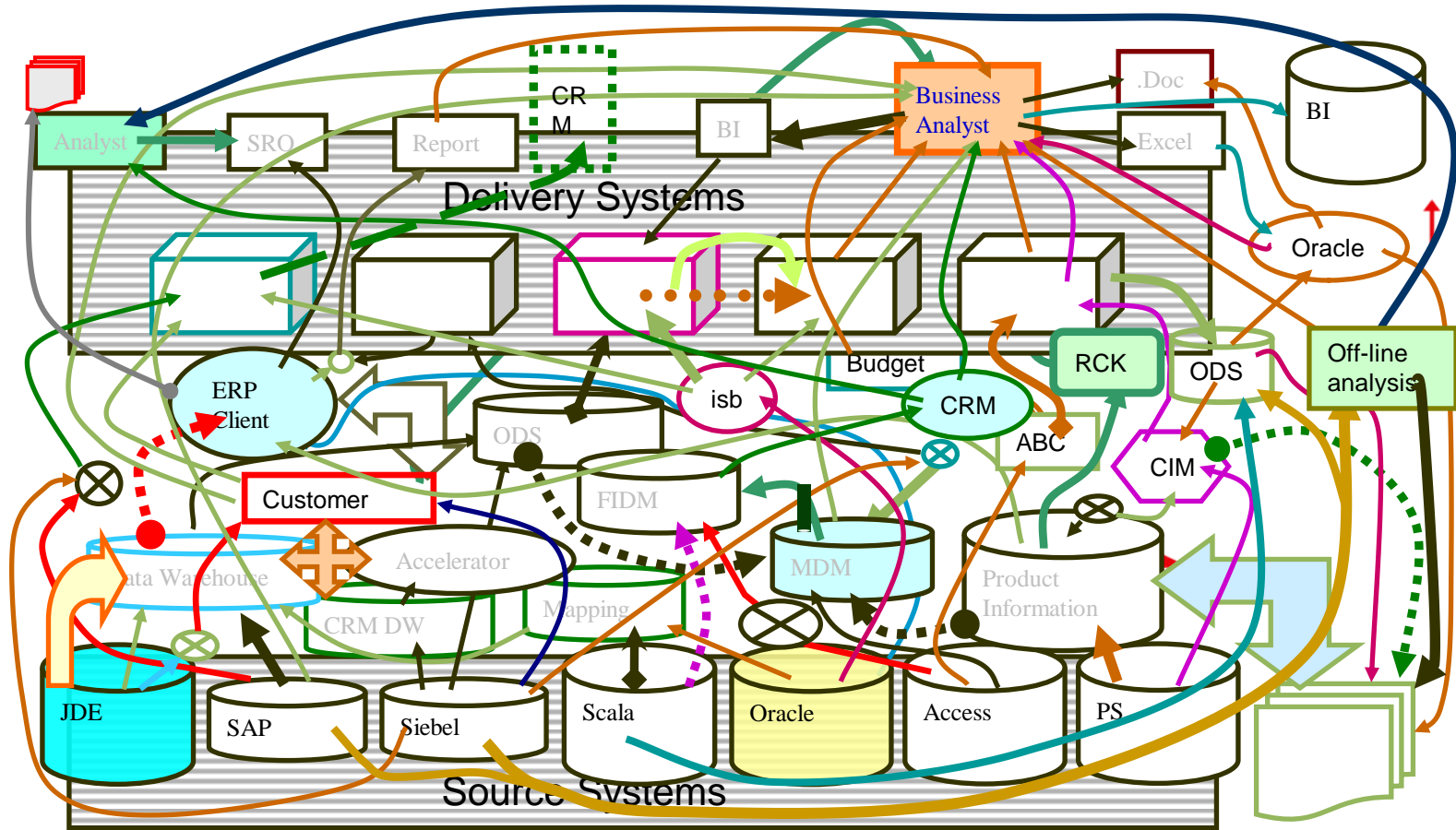
20+ types of systems
6% with 50+ ERP systems alone

Search Based Applications Goals

	traditional	**SBA**
Users	Limited number of users Usage complexity, production costs	Large number of users Ease of use, traffic scalability
Interface		
Functional	Heavy one-shot development	Agile applications Simple data access, use of standard web technologies
Querying	Dedicated resources Datamarts, additional hardware	Generic data layer Real time data, high performance querying
Data Source	Structured data	All data Connectors, structuration of data

exalead

Search Engine now handle rich semantics

› Text fields
  › « … the certification test is
› Numerical fields
  › 3.14159265
› Date
  › 16/11/1957
› GPS coordinates, real w
  › 48.451065619, 1.4392089
› Categories
  › Top/Animals/Pets/Dog

Databases

Search Based Applications

Search engines
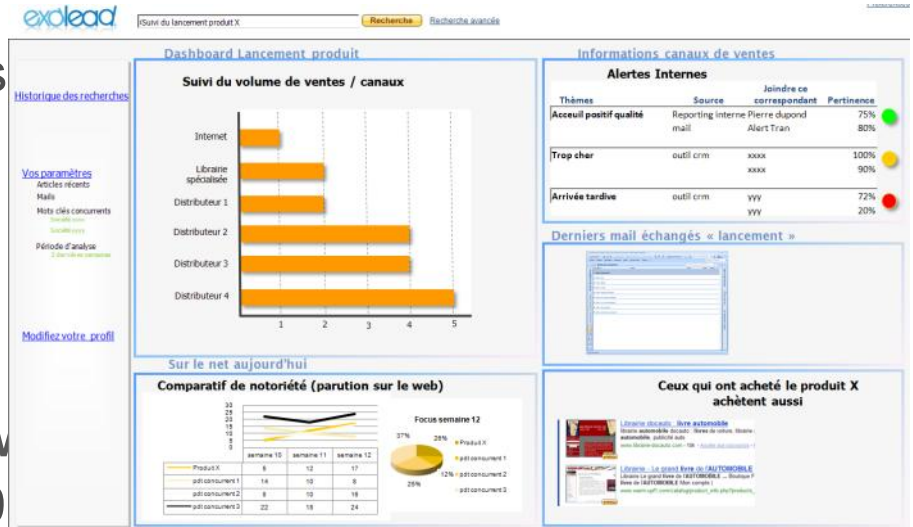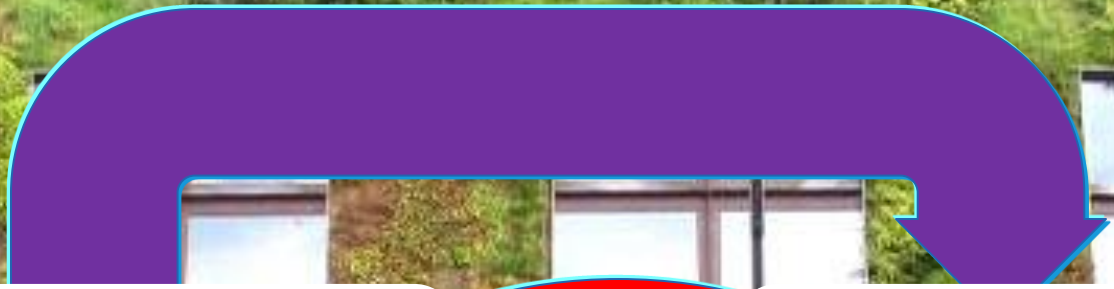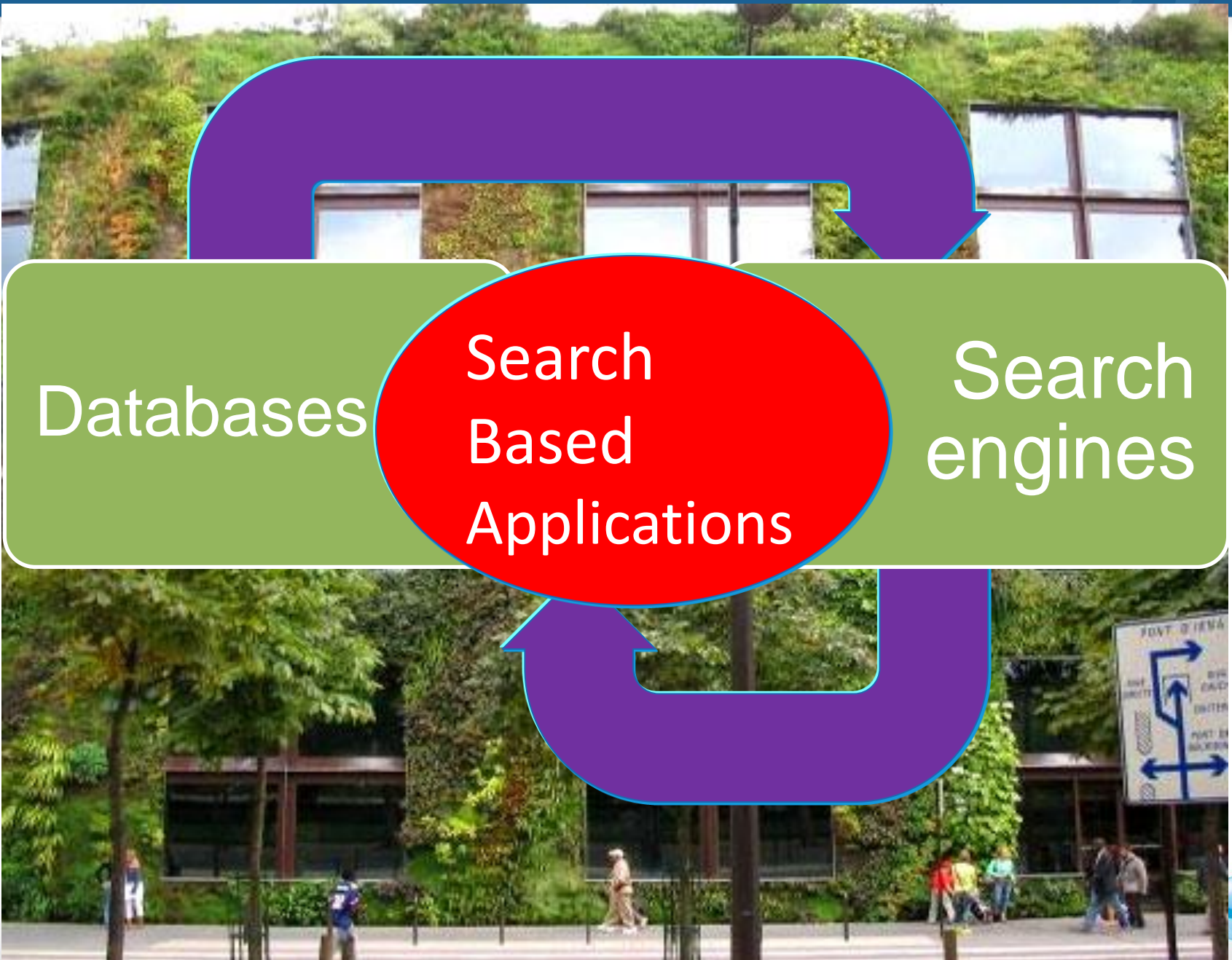
Usability
360°, Google-like, interactif, conversational
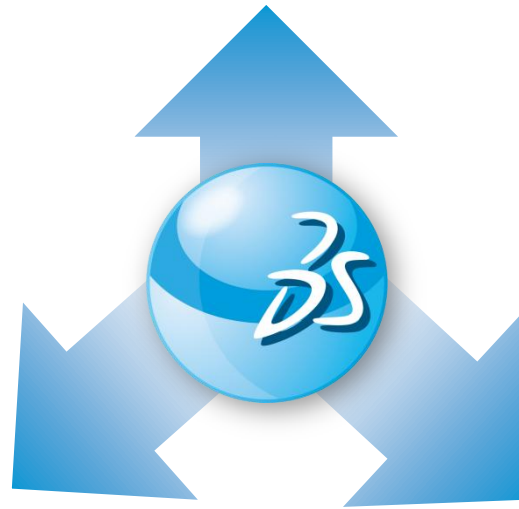
Agility
Flexible, Agile, Days vs Months

Performance
Real time, millions of end-users, Terabytes of information

exalead

How Search Based Applications Work

Connect	Process	Access	Interact
• **Get real-time flows of data** • Get and maintain security information	• **Manage file formats (PDF, office, drawings, XML…)** • Ability to understand free text to relate text to business objects	• **A highly scalable data repository** • **full text search, navigation and reporting capabilities (the Index)**	• **A Framework to create search oriented web applications at the speed of light** • Create virtual feeds of information (RSS, etc…) and associate widgets

Packaged as one, easy to deploy software

Built-in cluster architecture, for high-availability and scalability

Complete DIY

Perfect SBA

exalead
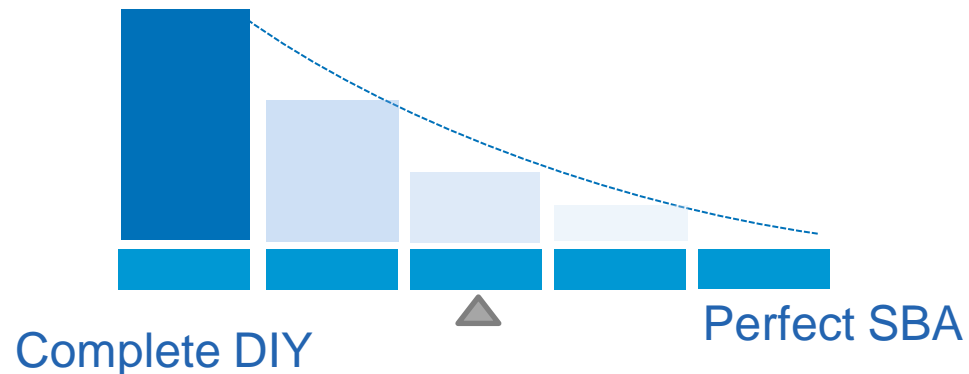
› Beware that a SBA is not

    › A replacement for transactional applications

        • A SBA won't manage your workflows, lifecycles, and won't modify the existing systems

    › A good excuse to drop your business intelligence software

        • A SBA goal is not produce the pixel perfect highly complex final report you have to submit to the SEC

    › Replace all your complex, historical business systems

        • A SBA goal is not to reproduce all the business logic of existing applications. It's to simplify it for information access

› A SBA addresses critical business issues by enabling easy Search & Discovery into key data by key users

exalead 3S

Four Types of Search

exalead

Form Based Search (overlay database)

Traditional search on database is complicated:

- Several fields to fill in
- Need to know field values

exalead

Unique Search Box

Faceted Search

Map Search (GPS, mobile)

› In the past ten years, people have learned to use the unique search box

	Google	Bing	Ask	Yahoo
1	26.79%	46.76%	49.90%	54.15%
2	23.39%	18.81%	13.03%	18.11%
3	18.72%	15.92%	16.09%	12.31%
4	12.78%	8.40%	6.72%	7.08%
5	8.23%	5.23%	6.42%	3.73%
6	4.55%	1.94%	3.77%	2.47%
7	2.76%	1.40%	0.71%	0.97%
8	1.36%	0.71%	2.24%	0.68%
9	1.02%	0.77%	0.81%	0.33%
10	0.41%	0.06%	0.31%	0.18%
avg. length	2.93	2.27	2.39	2.06

http://blog.alessiosignorini.com/2010/02/average-query-length-february-2010/
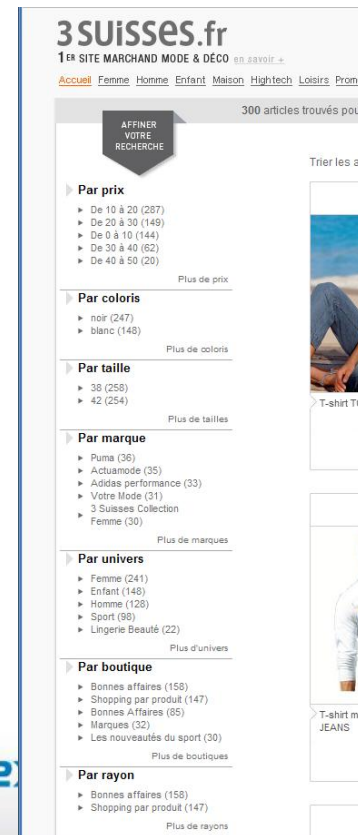
› People have learned to use facets on shopping sites

› People are *learning* to do map search

› Forms are still boring

exalead

Faceted Search and Semantics

› Facets are semantic dimensions

› They are visualisations of semantics that users can understand

› Semantics can come from databases

› Semantics can come from text

› Common semantic dimensions link together structured and unstructured data

› Search Based Applications are possible because of semantics

Type

Anğğe, Flickr

Equality

Relation

exalead
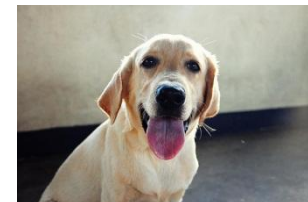
› ## Databases are "structured"

› ## Database semantics comes from row:column

› ### Column defines *Type* of entity
- Examples: client, supplier

› ### Row defines *Relation* between entities
- *Primary key -* main entity

› ### As for *Equality*
- This is the whole problem of Master Data Management

Name	User	Phone	College
Graff	rgraff	392-3900	Pharmacy
Harris	bharris	392-5555	Medicine
Ipswich	zipswich	846-5656	PHHP

Immediate Federated View		
Erik McClain	Address 1	Billing
McClain Erik	Email	CRM
Erik	McClain	Birth Date	Support
Erik	McClain	Phone	ERP

18

› Text is "unstructured"

  › Semantics not explicit

    • Resources and processing needed

› Natural Language Processing

  › *Typing* entities/things

    • Rules, lists, ontologies

  › *Equality*

    • Linguistic variants, morphology, stemming, synonyms

  › *Relations*

    • Parsing, co-occurrence (related terms), Linked Open Data

Google has acquired social search **Aardvark** service Aardvark, says a source that has been briefed on the deal, for around $50 million. We first reported on the discussions between the two companies ...

exalead

Semantics in databases and text

› Database

   › *Type*==column

   › *Relation*==row

   › *Equality*==???

› Text

   › *Type*==ontologies

   › *Relation*==parsing

   › *Equality*==morphology, synonyms

- Search Based Applications
  - Structure from databases
  - Linguistic variation from text
    - *Types*==facets
    - *Relation*==fields
    - *Equality*==text processing

exalead

Semantics in Search Engine

exalead

Database semantics is imported into search engine facets

Semantics is extracted from unstructured text by NLP

autosuggestion

language identification

related terms
related queries

spell checker

faceted search
local search
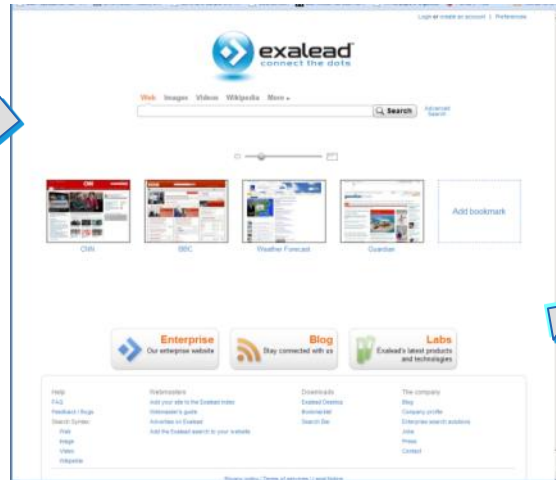
lemmatisation

synonyms
phonetic

cross language

Ontology Matcher

phonetic
lemmatisation
stopwords

sentiment analysis

named entity

Data Sources

INDEX
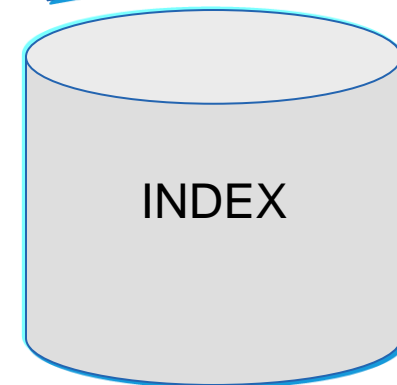
QueryMatcher

FastRules

HTMLRelevantContextExtractor

Categorizer
Clusterer

exalead 3S

Search Pipelines

Understand the user query & enhance results

54 languages supported (used to choose the right tokenizer)

Split words, detect end of phrase, …

Expanded query with user-defined synonym

Determine lemma and stem of the word . Apply Phonetization algorithms

Rewrite special expressions such as *word*

Query is rewritten to be comprehensible by the index

Language Detection → **Parsing (Tokenizer)** → **Synonym Expansion** → **Lemmatization / Phonetic** → **Query Rewriting (Regexp, …)** → **Transformation into index query**

Index

Return results → **Ranking** → **Related Terms** → **Summary** → **Highlighting**

Index return results set matching user query and security rights

Rank results using density, text scoring and ranking formula

Extract Related Terms from the result set

Determine Summary to be displayed for each hit

Highlight the words that are matching the user query

Search Side

exalead 3S

Cross Language Information Retrieval

Ontology Matcher : add an alternative indexing form to a string span

Given an existing taxonomy or ontology, normalized index terms can be produced.

Multimedia Semantics

Multimedia Semantics

› Suggestions

› Spell correction

› Variant forms

- Regular expressions
- Natural language processing

› Synonym expansion, cross lingual search

› Related terms, related queries

› Entities, Classification, facets

› Sentiment Analysis

exalead

Search Based Applications

exalead

Provide interactive reporting

Build decisional tools on unstructured information
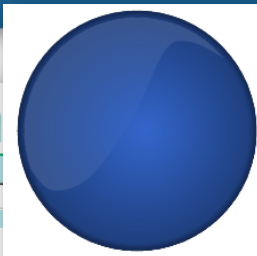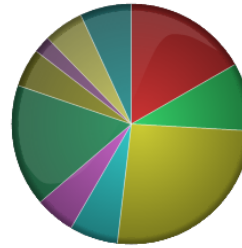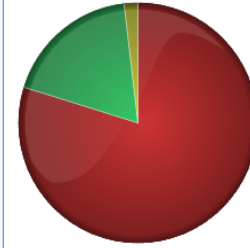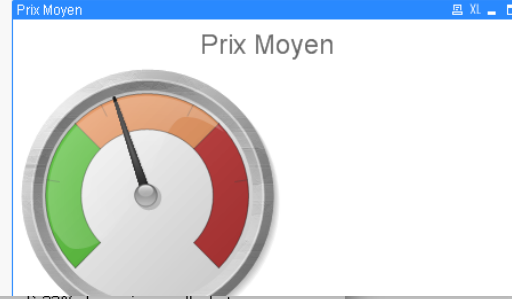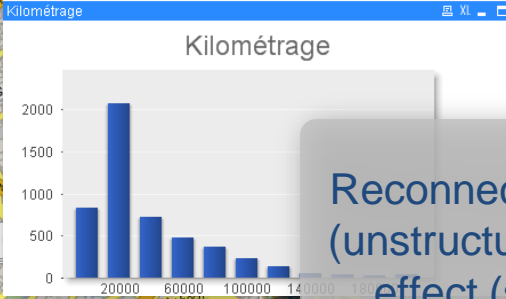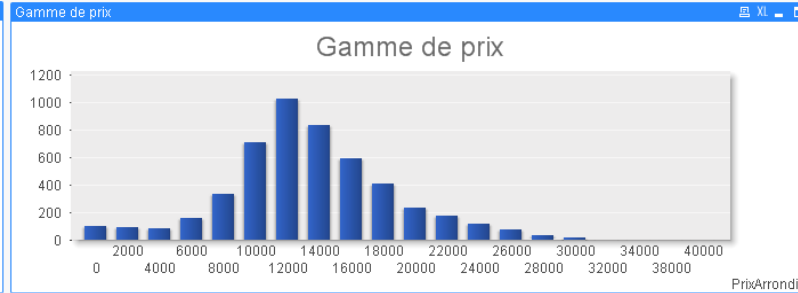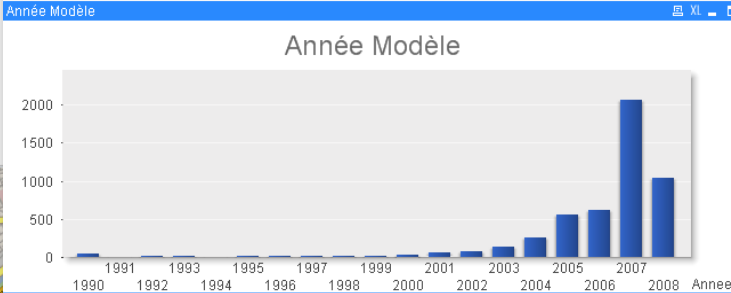
Reconnect the causes (unstructured) with the effect (structured)
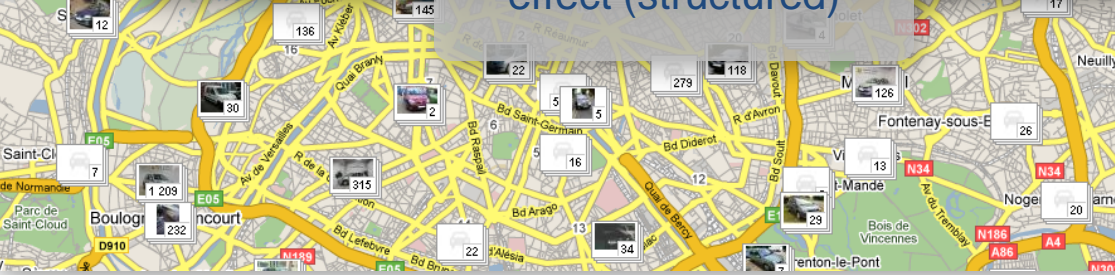
Facets can make search results into Business intelligence

aggregating numerous unstructured data sources

Navigation is refined dynamically

Examples of
Search Based Applications

exalead

Unify structured and unstructured data sources

A unique search bar

Structured results from a Database

Unstructured data from the Web

Merged Facets

Unify structured and unstructured data sources

Structured results from a Database

Unstructured data from the Web

Semantics from Text

Semantic Analytics for Mood-Based Local Search

YPG Database + Web & UGC content + Social Networking

exalead

Map-Based Access and Reporting for Water Quality Data

Description

- Agile MDM project
- Part catalogue normalization & enrichment
- Aggregation of content from ERP & Directories
- ERP (SAP) & Directories (online directories)
- Aggregation of :
  - Part number
  - Technical description (size, length)
  - Description
- Semantic analysis :
  - Merging information from SAP and vertical directories on parts
  - Similar product feature bases on technical description

Current and Future Needs:
Domain Specific Semantics

exalead

Intelligent analysis of query to extract **relevant** terms:

- 2 stars: mandatory field
- Swimming pool: Optional with a boost on matching results

Travel

Description

- B2C Touristic Portal
- Automatic aggregated content
- Exalead 360 + Business Console modules
- Aggregation of :
    - Online news content
    - Hotel directory
    - Entertainment directory
    - Railway station location
- Semantic analysis :
    - Type "Corrida" or "Beaujolais" and the system auto-discover entertainment or railways stations co-cited in the news.

- Search engines can handle the semantics of databases (but not the transactions)

- Facets are semantic dimensions

- Semantics allows for « business intelligence » type reporting

- Search Based Applications use the power of search engines (intuitive, scaling, agility) to extract and merge information from databases and text

- More semantic types are needed for text

MORGAN & CLAYPOOL PUBLISHERS

Search-Based Applications

At the Confluence of Search and Database Technologies

Gregory Grefenstette
Laura Wilber

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES

Gary Marchionini, *Series Editor*

Publisher: Morgan & Claypool

Copyright: 2011

ISBN: Paperback - 9781608455072
Ebook - 9781608455089
Pages: 141
Authors: Gregory Grefenstette &
Laura Wilber, Exalead, S.A.

exalead