Web, Entreprise, Infrastructure, Recherche Convergences et Divergences

Florian Douetteau
Exalead S.A.

Florian.Douetteau@exalead.com

Exalead

- 7 years old
- More than 100 employees (x2)
- Hundreds servers datacenter
- Paris (place de la madeleine, come and see us)
- We do / sell
  - www.exalead.com (web, image, wikipedia, video..)
  - Corporate Search Solutions
  - Web Search Solutions
  - Baagz.com (social bookmarking)

Topics of Discussion

- What Exalead does ?
- How Exalead does ?
- Research Interests

Main point

We might work better together

Subliminal topic

(Web) Search companies are a great place to be !

- Interesting People
- Innovation driven
- Great challenges ahead.

Exalead: What we do, day by day

- Mining the Web
- Build user interface
  - Navigation and search integration
- Selling high-volume indexing clusters and solutions
- Customized corporate search solution.
- OEM Search Software
- Enriching content
  - named entities, statistical analysis, preview…
- Socia! Bookmarking (go to baagz.com !)

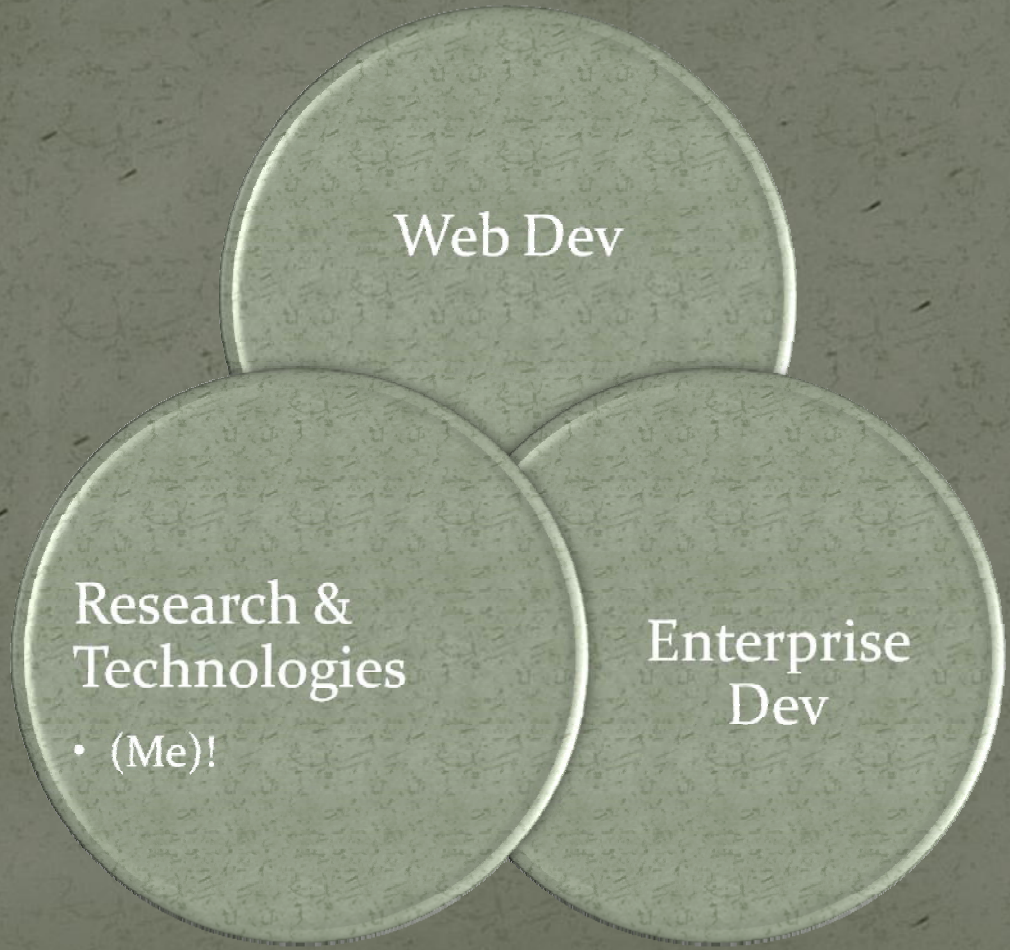
Exalead: what we do (corporate)

- Maintain stable API
  - Document, Search, Administration
- Unified access to unstructured / structure sources
- Integrated Corporate Thesaurus / Directory
- Document, make people understand what « indexing » and « relevance » mean

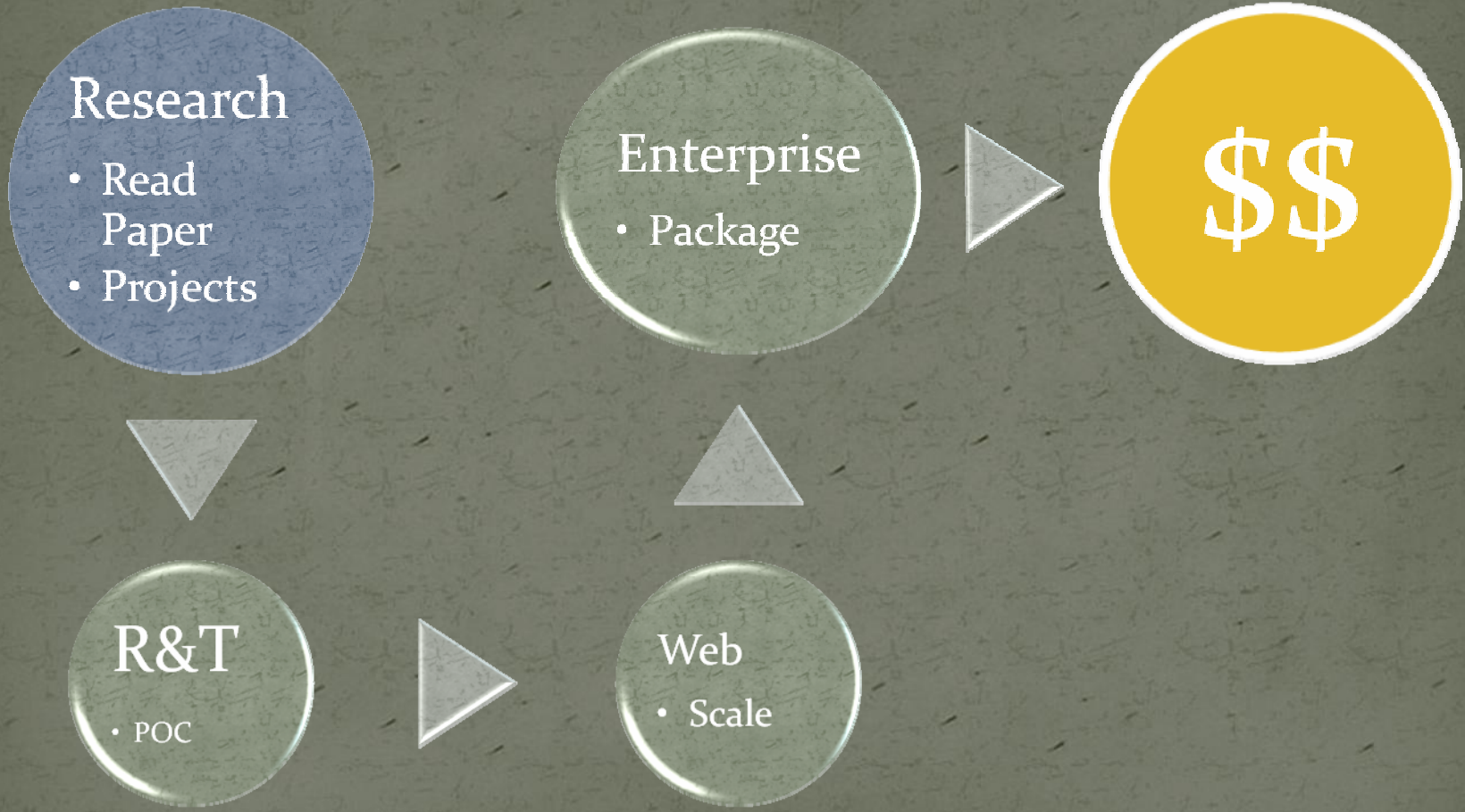
Exalead : what to do (technical)

- Discuss various index layouts
- Analyzing the various computer cache behaviour (L1, L2, Memory, Disk ...)
- Code Profiling
- Data-structures
- Distributed calculus

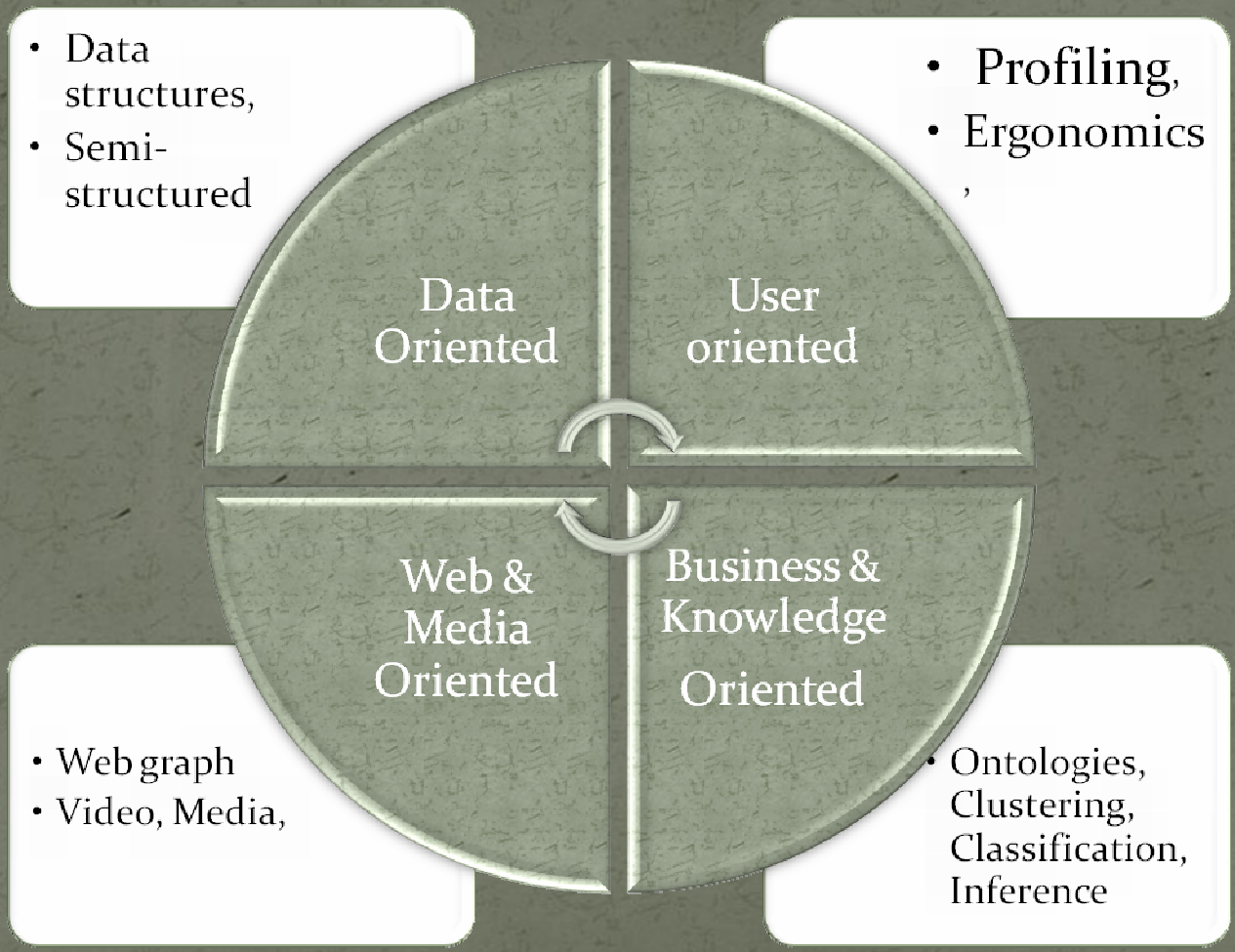
How Exalead Works ?



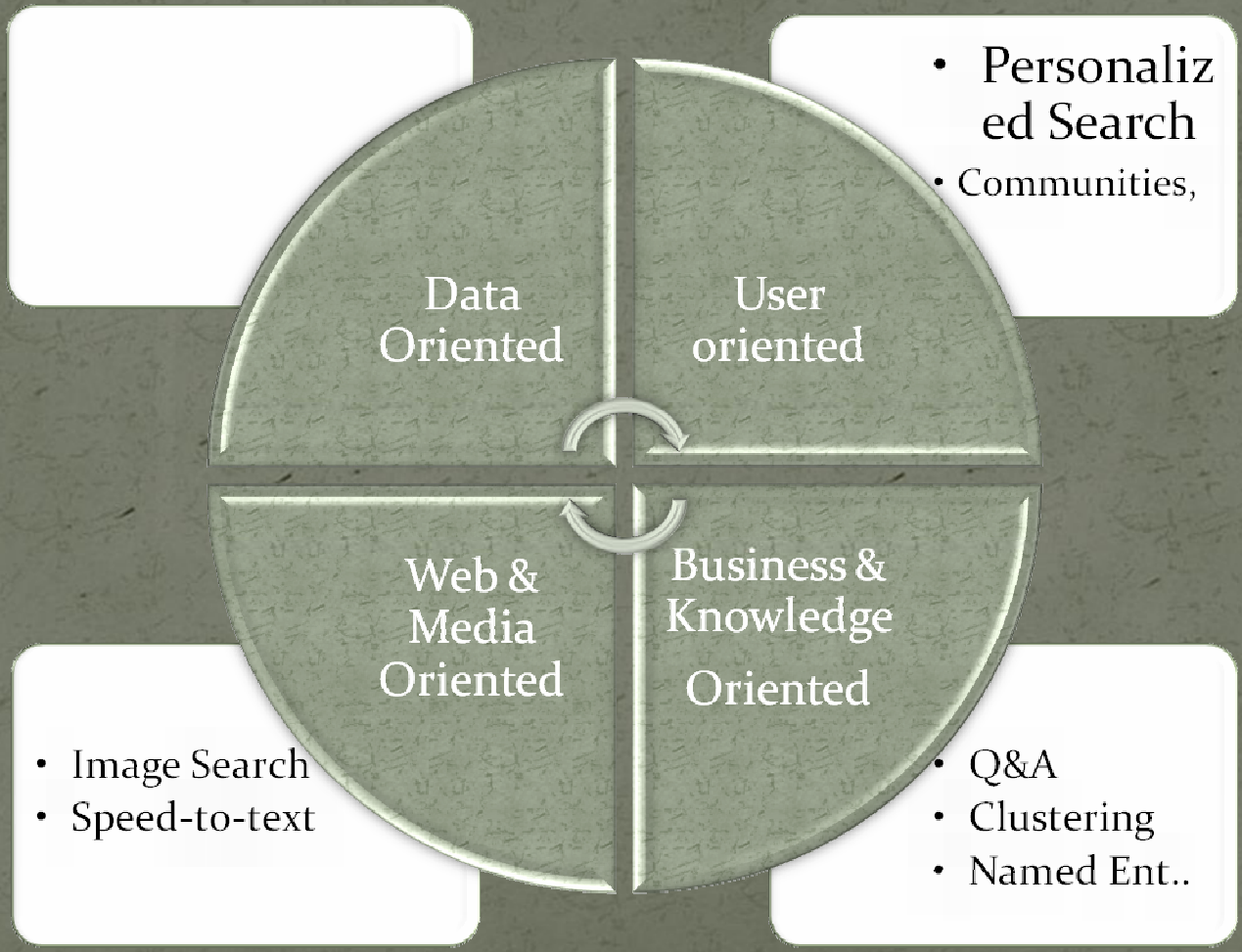
Exalead Incubation



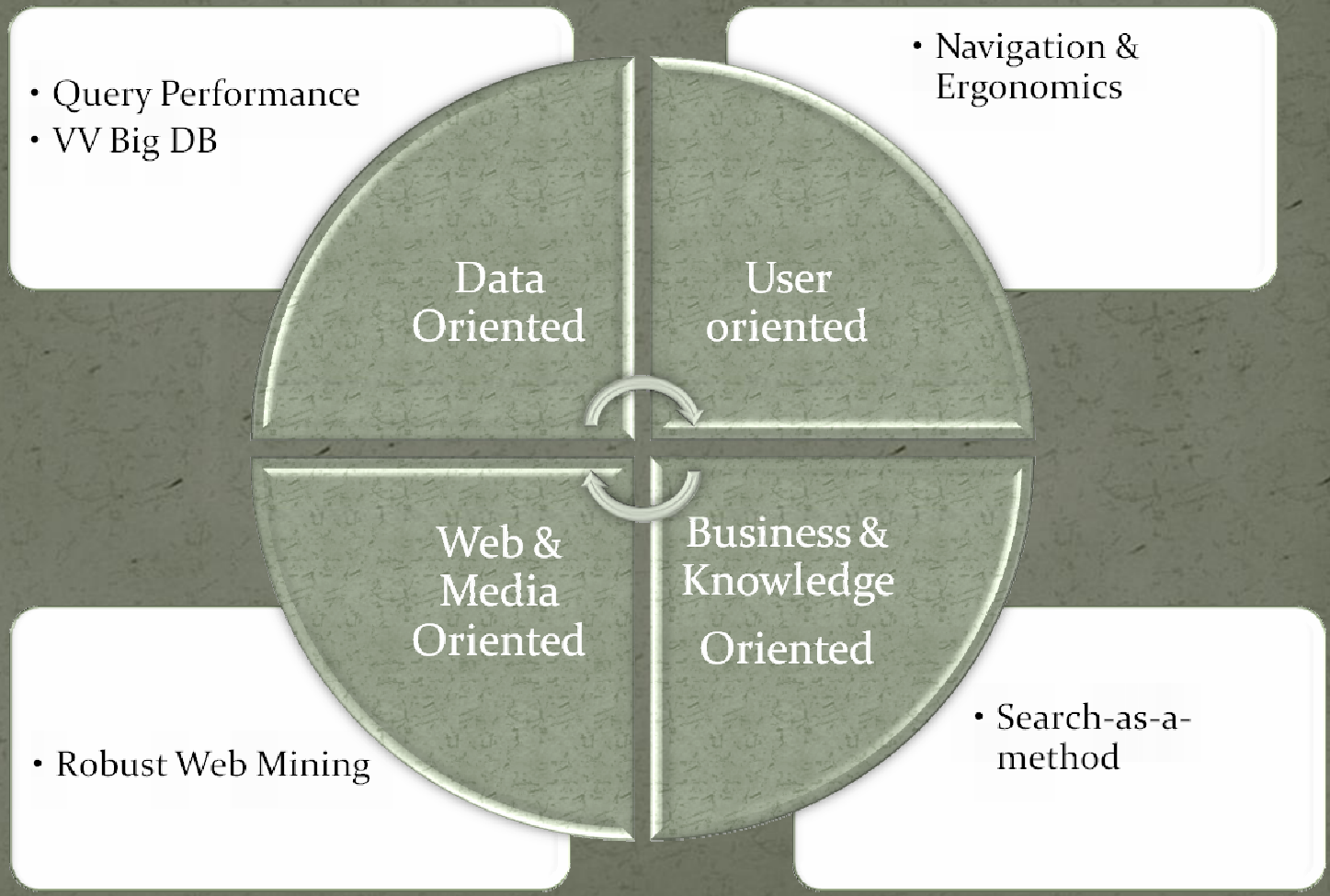
Search Camembert



Some Exalead Classical Interests



Some Exalead Differential Research Interests.



Query Performance Models

- IR is about:
  - Minize the number of information displayed
  - But also minize the amount ressources (memory, disk) required to perform a search.
- Optimize index using static information
- Caching
- Optimize using known information (log, language)
- Query distribution and index slicing strategies (+1000)

Very Very Big databases (+100To)

- When history matters
  - Data increase faster than computing powwer
- Telecom companies, web companies, have increasing amount of logs
- User generated content and storing digital life
- Relational database not always cost effective
- Meet the traditional IR paradigms:
  - Partial result set (relevance)
  - Store and index abstractly mined query
  - Write once need many

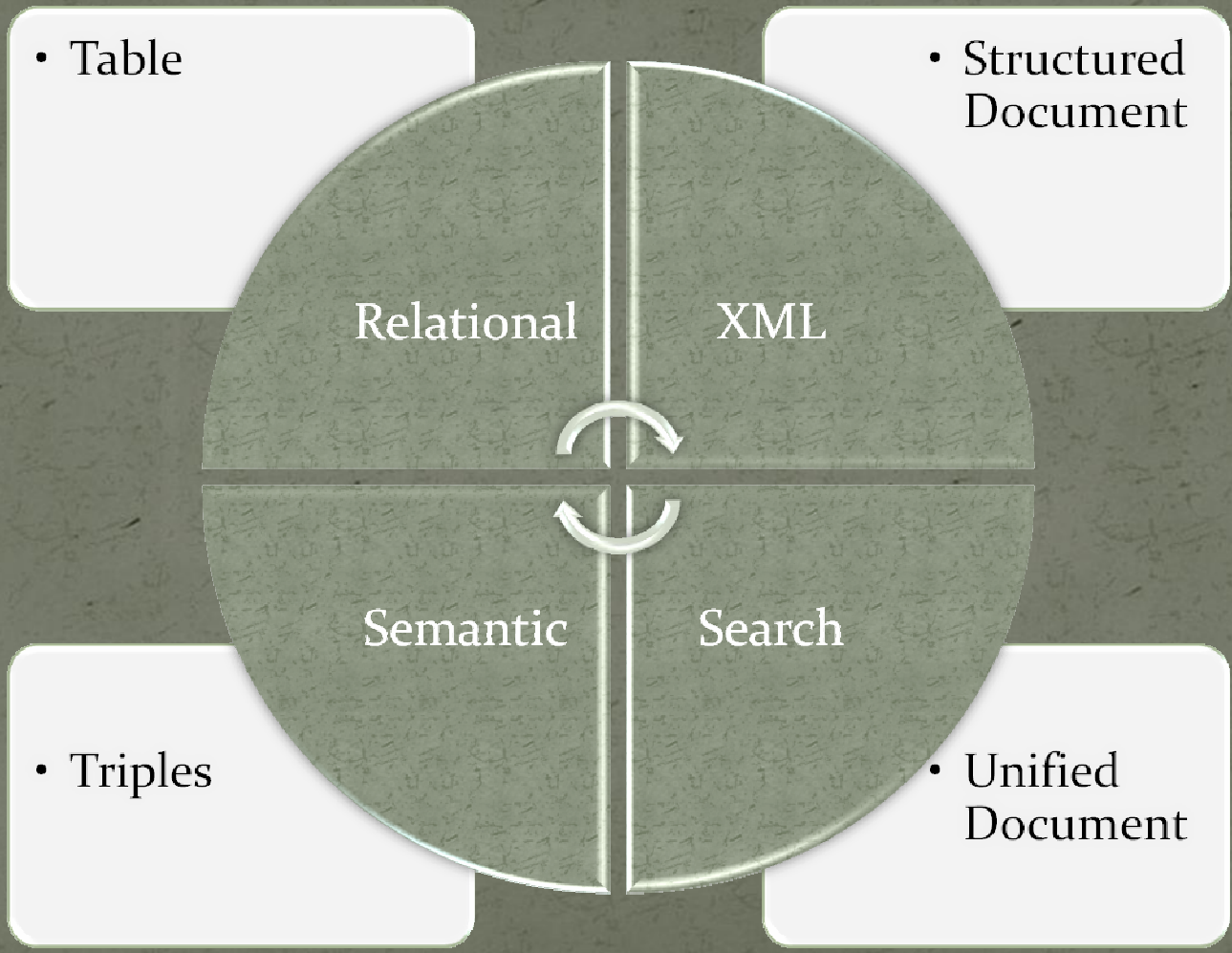
Robust Web mining

- Fetching the web
  - Robot repartition, fairness, etc…
- Real Data (SPAM)
  - Generated content
  - Infinite site
- What is a page, a site ?
  - Semi-Duplicate Content
  - Forums, huge platforms (MySpace, SkyBlog)
- Real Time:
  - Differentiel, Incremental algorithms
  - Time-aware algorithms (buzz)
- Analyzing the web graph
  - Algorithms that must scale:
    - e..g. « MapReduce » paradigm

Topics Maps, Navigation and Ergonomics

- Tested various layouts / interfaces for navigation
- User Lab mandatory for designing efficient interface.
- Too much information
  - user is lost
- Too many ‘dimensions’ to your interface
  - user is lost
- Too many steps
  - user is lost
- **How to break the list of 10 results paradigm ?**

Corporate Information Management Paradigms



Search paradigm:

- Build an Unified document model
- Handling of text (proximity, tidf)
- Enrich document from their textual meaning
- Few writes, many reads
- Classification, clustering, graph analysis

- Search as a way of building applications:
  - Survey
  - Pushed documents, etc…
- Format comparisons of benefits of various paradigms
- Bridge between paradigms

Questions ?