Vivísimo

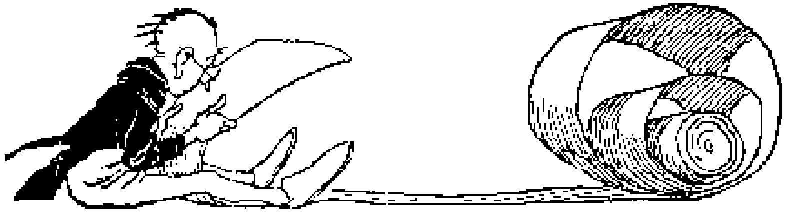
About

- **Raul Valdes-Perez, co-founder & President**
  - 1991 Ph.D. in Computer Science (Carnegie Mellon)
  - CMU CS Faculty member since then, currently on leave
  - Research expertise in AI and Knowledge Discovery
- **Vivisimo, Inc.**
  - Enterprise software company founded June 2000 in Pittsburgh
  - Award-winning web-search site Vivisimo.com
    - Best meta-search site last 2 years in row (Search Engine Watch)
    - Site is just a demonstration of our clustering & EAI products
  - Main Funding: National Science Foundation

Problem: Information Overload & Overlook

Vivísimo

- **Information Overload and Information Overlook**
  - Look for information → get too much back
  - Most people handle overload by overlooking most information



- **Overlooking information has a business/user cost**
  - Corporate employees fail to find needed information
  - Customers don’t solve their problems by themselves, online
  - Publishers lose potential readership
  - Web search providers lose potential click-throughs
  - Users miss out on unexpected discoveries or opportunities

Looking at One Result is Not the Solution

~~I’m Feeling Lucky~~

How to Alleviate Information Overlook?

- **Decrease the available information**
  - Purge the obsolete
  - Censor the worthless
- **Make people …**
  - Smarter
  - Work harder
  - More efficient! (but how?)
- **Provide organized information!**

How to Provide Organized Information?

- **Manually tag (classify, index) all content?**
  - “We have no process for consistently tagging our content. We have 50 different business units. People in one unit do a great job, but others do not use tags at all.” Forrester Report
  - Forrester says $4 per page to make a controlled vocabulary
  - $50 per document to manually tag (large pharma)
  - Tags tend to be broad and bland (one size fits all)
  - → Tagging is costly and leads to mediocre results

Technologies for Automatic Organization

Vivísimo

- **1. Spatial approaches (also temporal)**
  - Assign documents to spatial or temporal coordinates
  - Intuition: people deal well with space/time so embed documents in space-time
  - Simple space: 2D maps (search lists are like 1D space)
  - Prominent example: Pacific Northwest Labs
  - Many commercial variants

- **2. Clustering & categorization**
  - Assign documents to discrete, named groups
  - More than one group is OK & can be hierarchical
  - Intuition: people use groups/hierarchies all the time
  - Prominent examples: NorthernLight, Vivisimo

1. Spatial Approaches

- **Use vector-space representation of documents**
  - The vector dimensions are the document words & phrases
  - Values are #times that a word or phrase appears
  - Very high-dimensional space
  - Each document is a point in this space
- **Then, select 2-3 axes that best scatter the points**
  - E.g., principal components analysis
  - Embed the documents in this 2-3D space
  - Show the map
- **Pros**
  - Can handle a very large number of documents
- **Cons**
  - Textually characterizing the spatial areas is difficult
  - Drill-down is clumsy

2. Categorization and Clustering

Vivísimo

- **Categorization**
  - Places documents in pre-defined categories
  - NorthernLight
    1. Librarians build/maintain controlled vocabulary of categories
    2. Software is designed/trained to assign documents to categories
    3. When querying database, resolve returned document sets into the most frequent categories
  - Yahoo directories & Librarian's Index to the Internet
    - Like above, but step (2) is done manually

- **Clustering**
  - Groups documents and makes spontaneous descriptions for them
  - Many years of work in academia & industry
  - Example: IBM's Intelligent Miner
  - Vivisimo Clustering Engine

Pros/Cons of Using Pre-Defined Categories

Vivísimo

- **Pros**
  - Guarantees human-like categories
  - Algorithm is simple, *once the categories have been created and assigned*

- **Cons**
  - Expensive to develop/maintain (staff time, etc.)
  - Scales poorly across human languages (German, Chinese)
  - Typically must pre-process all the records
    - Some resources are accessed only by searching them
    - Can be logistically complicated
  - Automatic classification into categories is error-prone
  - Reduced chance of surprise/discovery
  - Fixed vocabularies tend to be broad and bland
    - Queries on “Three Rivers Stadium” or “Kingdome”
    - Fixed categories might place in “Construction Industry”
    - Vivisimo says “Implosion” or “Memories of Three Rivers Stadium”

Document Clustering is Hard to Do Well

- **Cluster analysis in general is not so hard**
  - e.g., Warehouse location problem
- **Document clustering is not so hard when …**
  - People don't need to understand the clusters
- **Clustering for post-retrieval browsing is hard**
  - People need to understand the clusters at a glance
- **Most doc-clustering algorithms optimize the wrong things**
  - Mathematical abstractions are distant from the main quality factors
- **Quality of description is paramount**
  - Concise, accurate, natural, hierarchical, and others
  - Hard to formulate mathematically, so develop heuristic algorithm!
  - Needs to use statistics, linguistics, and subject matter knowledge

Modeling the Value of Topic Clusters

- **Intuition**
  - Lots of wasted effort if information is disorganized
  - View few results before exhausting your time allotment

- **Modeling Assumptions**
  - User spends 12 min before giving up or moving on
  - Eye skips over search results or folders sequentially
  - Equal, independent probability that any given result solves the problem
  - Employee costs $60 per hour
  - 1,000 employees
  - 2 searches per employee/day
  - 10 minutes to solve problem elsewhere when search fails

- **Folders let you see 11 docs in detail vs. 6 for ranked lists**

- **Conclusion: savings of $1M+ per year (white paper)**

Empirical Data on Clustering of Search Results

Vivísimo

- **Users click on 2.8 folders, on average (publisher customer)**

- **Increased click-throughs of 30% to 200%**
  - 30% increase at vivisimo.com by adding up to 28 sponsored results at the bottom of 200+ total search results
  - Essentially all of these are accessed via the folders
  - Larger increases when the BEFORE is unclustered ranked lists

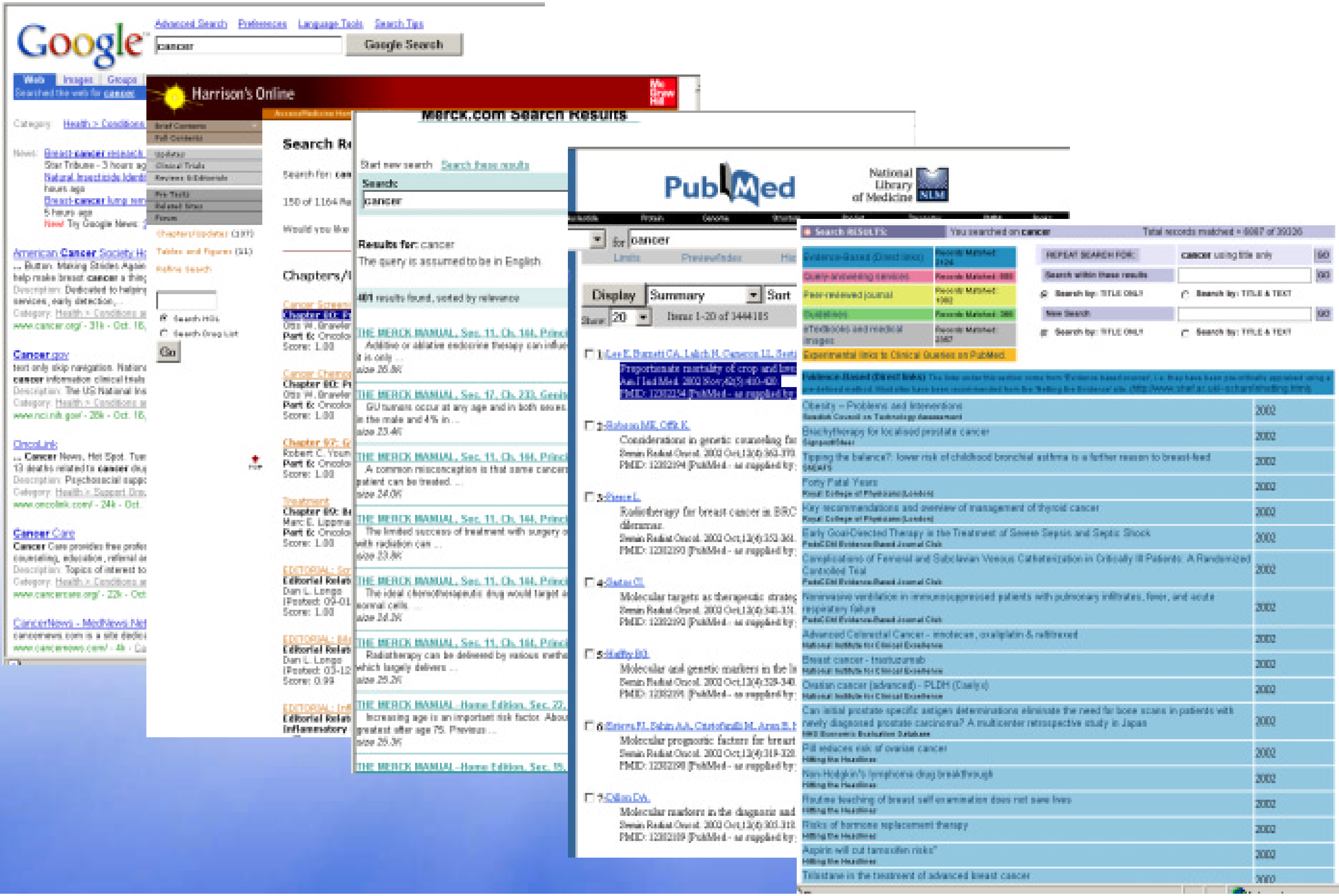
- **Fortune 1000 customer experimental tests**
  - Combination of meta-search and clustering
  - Minutes-to-solve: 1.9 versus 3.7
  - #Searches-to-solve: 1.4 versus 2.3
  - Rating: 1.6 versus 2.4 (1=very easy, 5=impossible)

Hybrid Approach May Offer Best of Both Worlds

- **Automatic clustering on several fields**
  - Document title
  - Summary (aka abstract, snippet, excerpt)
  - Humanly assigned, pre-defined categories
- **Doesn’t matter if some fields (e.g. categories) are missing**
  - Just cluster on what’s available
- **Dynamic (not fixed) categories will prevail …**
  - When they offer better folder descriptions than pre-defined categories
  - At deeper points in the folders hierarchy, where pre-defined categories are not detailed enough
- **Example: clustering PubMed search results**
  - Clusters on titles, abstracts, and Mesh Headings

How the World is Now

Vivísimo



Advanced Search Preferences Language Tools Search Tips

cancer

Google Search

Web Images Groups

Searched the web for cancer

Harrison's Online

Merck.com Search Results

Search Results

Start new search Search these results

Search:

cancer

Results for: cancer

The query is assumed to be in English

401 results found, sorted by relevance

THE MERCK MANUAL, Sec. 11, Ch. 144, Princi

Additive or ablative endocrine therapy can influe

THE MERCK MANUAL, Sec. 17, Ch. 233, Genit

GU tumors occur at any age and in both sexes

THE MERCK MANUAL, Home Edition, Sec. 22,

Increasing age is an important risk factor.

PubMed

National Library of Medicine NLM

for cancer

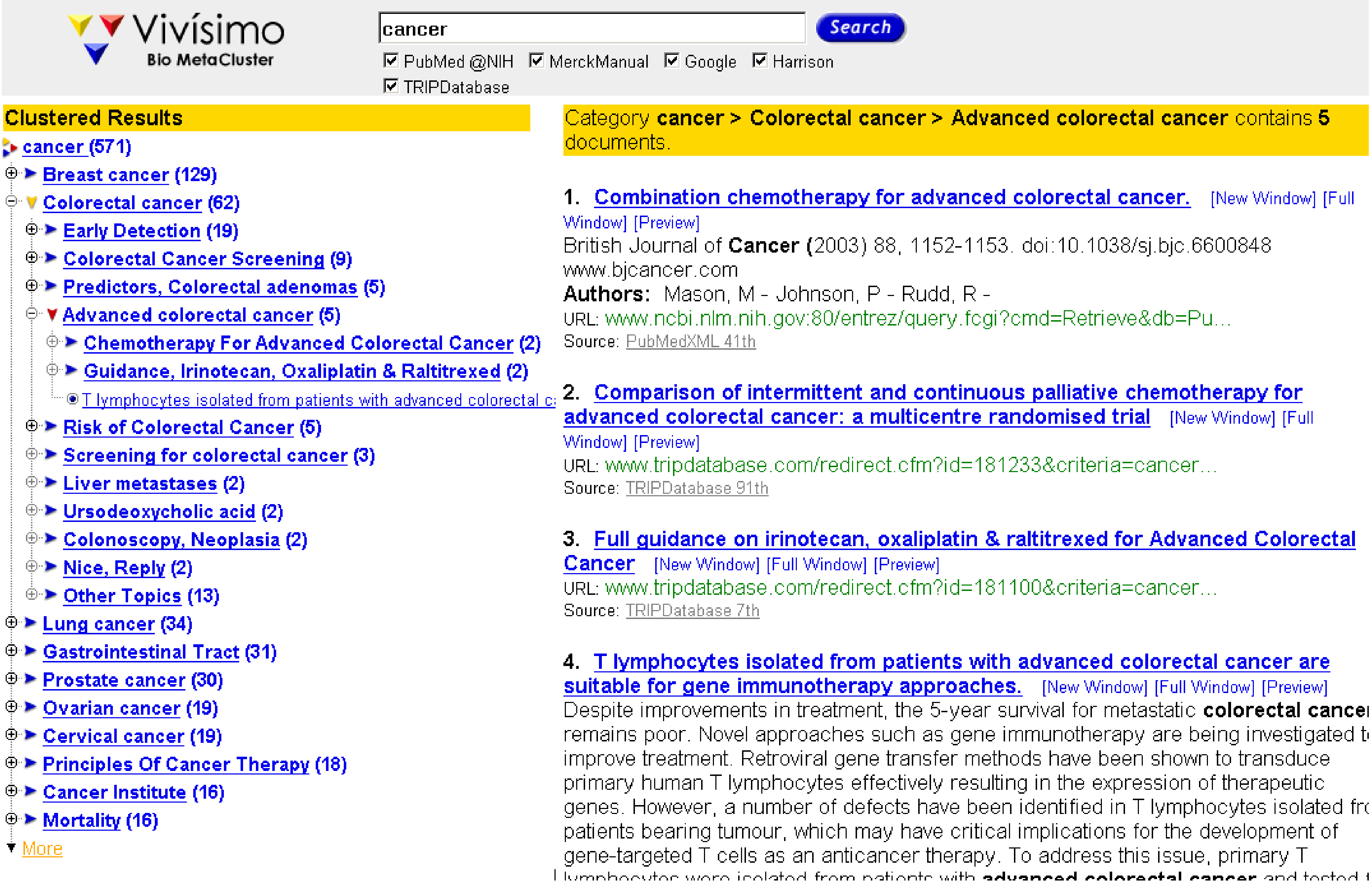
Display Summary Sort

Show: 20 Items 1-20 of 3444185

Search RESULTS You searched on cancer Total records matched = 8087 of 39336

Obesity – Problems and Interventions	2002
Brachytherapy for localised prostate cancer	2002
Tipping the balance?: lower risk of childhood bronchial asthma is a further reason to breast-feed	2002
Forty Fatal Years	2002
Key recommendations and overview of management of thyroid cancer	2002
Early Goal-Directed Therapy in the Treatment of Severe Sepsis and Septic Shock	2002
Complications of Femoral and Subclavian Venous Catheterization in Critically Ill Patients: A Randomized Controlled Trial	2002
Noninvasive ventilation in immunosuppressed patients with pulmonary infiltrates, fever, and acute respiratory failure	2002
Advanced Colorectal Cancer - irinotecan, oxaliplatin & raltitrexed	2002
Breast cancer - trastuzumab	2002
Ovarian cancer (advanced) - PLDH (Caelyx)	2002
Can initial prostate specific antigen determinations eliminate the need for bone scans in patients with newly diagnosed prostate carcinoma? A multicenter retrospective study in Japan	2002
Pill reduces risk of ovarian cancer	2002
Non-Hodgkin's lymphoma drug breakthrough	2002
Routine teaching of breast self examination does not save lives	2002
Risks of hormone replacement therapy	2002
Aspirin will cut tamoxifen risks"	2002
Tolestane in the treatment of advanced breast cancer	2002

What Vivisimo Makes Possible



Vivísimo
Bio MetaCluster

cancer

Search

☑ PubMed @NIH ☑ MerckManual ☑ Google ☑ Harrison
☑ TRIPDatabase

Clustered Results

- cancer (571)
  - Breast cancer (129)
  - Colorectal cancer (62)
    - Early Detection (19)
    - Colorectal Cancer Screening (9)
    - Predictors, Colorectal adenomas (5)
    - Advanced colorectal cancer (5)
      - Chemotherapy For Advanced Colorectal Cancer (2)
      - Guidance, Irinotecan, Oxaliplatin & Raltitrexed (2)
      - T lymphocytes isolated from patients with advanced colorectal c
    - Risk of Colorectal Cancer (5)
    - Screening for colorectal cancer (3)
    - Liver metastases (2)
    - Ursodeoxycholic acid (2)
    - Colonoscopy, Neoplasia (2)
    - Nice, Reply (2)
    - Other Topics (13)
  - Lung cancer (34)
  - Gastrointestinal Tract (31)
  - Prostate cancer (30)
  - Ovarian cancer (19)
  - Cervical cancer (19)
  - Principles Of Cancer Therapy (18)
  - Cancer Institute (16)
  - Mortality (16)
  - More

Category **cancer > Colorectal cancer > Advanced colorectal cancer** contains **5** documents.

1. **Combination chemotherapy for advanced colorectal cancer.** [New Window] [Full Window] [Preview]
British Journal of **Cancer** (2003) 88, 1152-1153. doi:10.1038/sj.bjc.6600848 www.bjcancer.com
Authors: Mason, M - Johnson, P - Rudd, R -
URL: www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=Pu...
Source: PubMedXML 41th

2. **Comparison of intermittent and continuous palliative chemotherapy for advanced colorectal cancer: a multicentre randomised trial** [New Window] [Full Window] [Preview]
URL: www.tripdatabase.com/redirect.cfm?id=181233&criteria=cancer...
Source: TRIPDatabase 91th

3. **Full guidance on irinotecan, oxaliplatin & raltitrexed for Advanced Colorectal Cancer** [New Window] [Full Window] [Preview]
URL: www.tripdatabase.com/redirect.cfm?id=181100&criteria=cancer...
Source: TRIPDatabase 7th

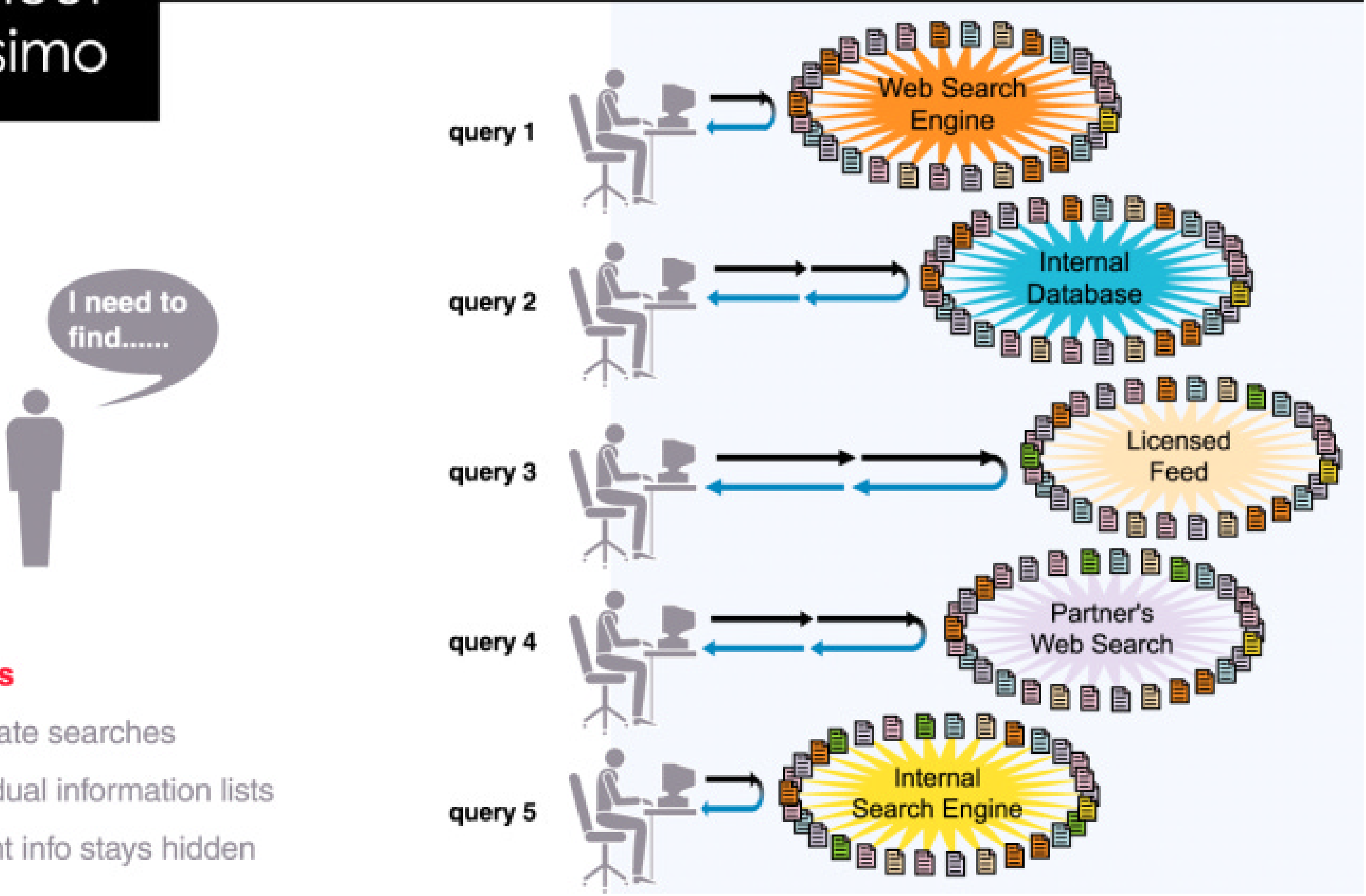
4. **T lymphocytes isolated from patients with advanced colorectal cancer are suitable for gene immunotherapy approaches.** [New Window] [Full Window] [Preview]
Despite improvements in treatment, the 5-year survival for metastatic **colorectal cancer** remains poor. Novel approaches such as gene immunotherapy are being investigated t improve treatment. Retroviral gene transfer methods have been shown to transduce primary human T lymphocytes effectively resulting in the expression of therapeutic genes. However, a number of defects have been identified in T lymphocytes isolated fro patients bearing tumour, which may have critical implications for the development of gene-targeted T cells as an anticancer therapy. To address this issue, primary T lymphocytes were isolated from patients with **advanced colorectal cancer** and tested t

Multiple Sources = Multiple Searches and Multiple Lists

Vivísimo

Without Vivísimo

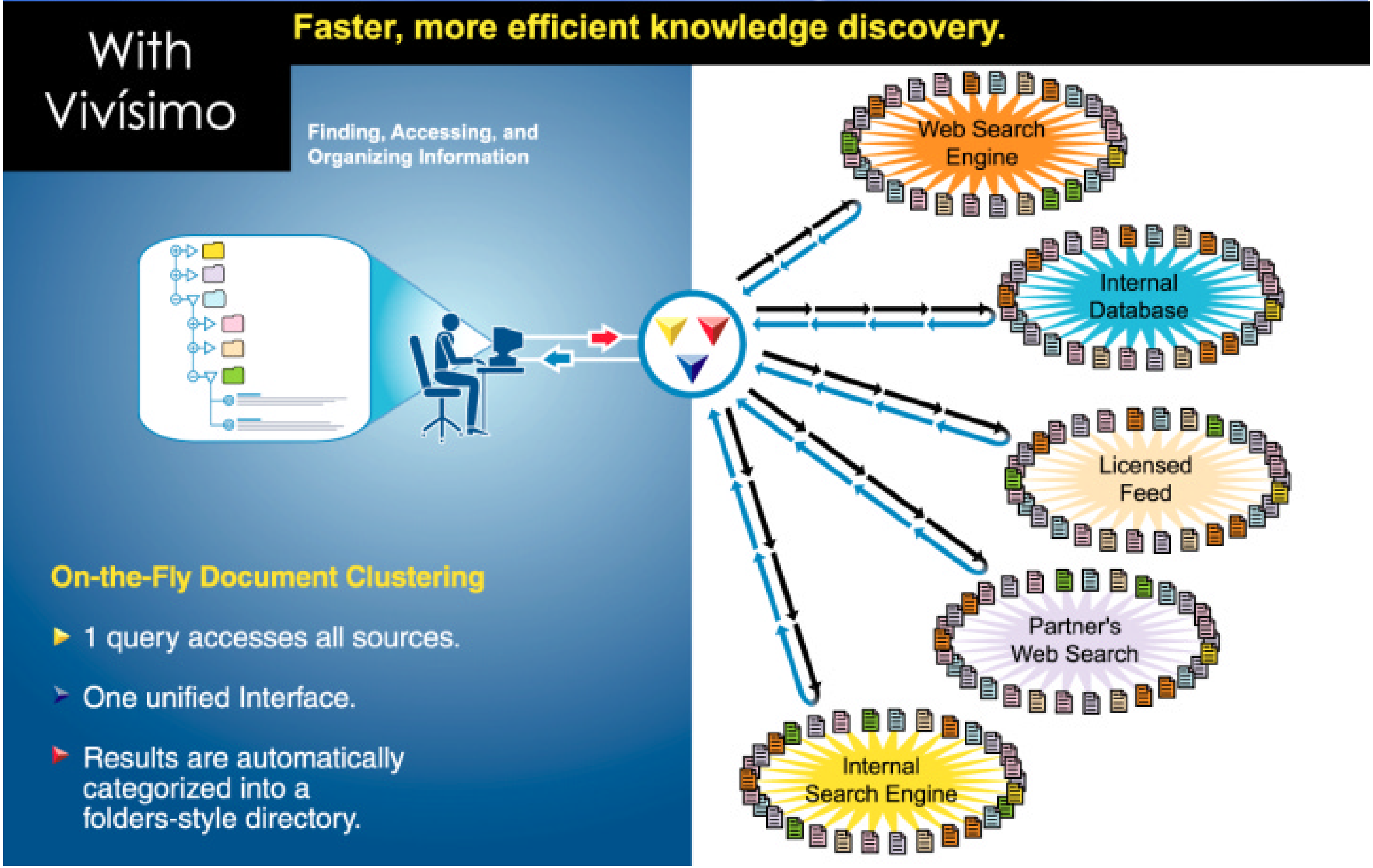
Wasted time. Killed productivity.



Problems

- 5 separate searches
- 5 individual information lists
- Relevant info stays hidden

A New Standard, A Better Way



With Vivísimo

Faster, more efficient knowledge discovery.

Finding, Accessing, and Organizing Information

On-the-Fly Document Clustering

- 1 query accesses all sources.
- One unified Interface.
- Results are automatically categorized into a folders-style directory.

Value Proposition of Clustering

End-users

- Easy access to useful but low-ranked results
- Learn at a glance the types of available information
- See results in context of similar results

Corporate

- More efficient employees/customer support operations
- Customers become more engaged with your content

IT Dept

- Quick installation on any search engine/database
- Overlays search; is non-invasive; no maintenance
- No need to train users

Conclusion

Vivísimo

- **Information Overlook imposes high opportunity costs**
- **Alleviate by creating and showing organized info**
- **Clustering into thematic folders is a natural approach**
  - Has proven hard to do well over many years of research
  - But is now largely a solved problem
- **Document clustering is Vivisimo's founding technology**
  - *"We help enterprises organize information from anywhere, any time, in any language, without the endless cost and complexity of building information taxonomies."*