



State-of-the-art search technology and future challenges

CTO Prof. Bjørn Olstad

Email: bjorn.olstad@fast.no

Fast Search and Transfer ASA

The Norwegian University of Science and Technology

Fast Search & Transfer (FAST)

Company

- Founded in '97
- Sold Internet BU to Overture/Yahoo
- > 1000 customers
- #2 growing technology company in Europe 1998-2002



Product

- Enterprise Search Platform
- Extreme capabilities in
 - Scalability
 - Accuracy
 - Analytics



FAST ESP™ is { "a leap into the future!" -IDC

FAST's Mission ...

The image displays four overlapping browser windows, each showing a different corporate website:

- IBM United States - Microsoft Internet Explorer:** Shows the IBM logo and a navigation menu with options like "Home / home office", "Small & medium business", "Large enterprise", "Government", "Education", "Developers", "IBM Business Partners", "Investors", and "Journalists".
- Careerbuilder.com - The smarter way to find jobs:** Features a "Quick Job Search" section with input fields for "Enter Keyword(s)", "Enter a City", "Select a State", and "Select a Category", along with a "Search" button.
- TIBCO Software Inc. - Microsoft Internet Explorer:** Promotes "ACHIEVE REAL-TIME BUSINESS" and "GET THE POWER OF NOW" with a "business solutions" section listing "CUSTOMER RELATIONSHIPS" and "FLEXIBILITY AND PREPAREDNESS".
- Scirus - for scientific information only:** Includes a search bar, a "Basic Search" section with checkboxes for "All journal sources", "All Web sources", and "Exact phrase", and a "Scirus Test Zone" link.

At the top of the collage is the **FIRSTGOV.gov** banner, "The U.S. Government's Official Web Portal", with a search bar and a "Go" button. Below the banner is a navigation menu with "Home", "About Us", "Site Map", "Help", "Español", and "Other Languages". A "Welcome from President Bush" message is visible on the right side of the banner area.

... Power the Most Challenging Information Retrieval Applications

FAST Research Strategy

– Strategic innovation

- Securing long term viability through leading industrial strength engine for aggregation, mining and information discovery in structured/unstructured data repositories/feeds

– Customer orientation

- Partnering with leading global companies to solve the biggest search challenges

– University partnerships

- Strategic deep relations to:
 - Cornell: Fred Schneider, Trustworthy Computing
 - Penn. State: Lee Giles, Niche/Meta searching
 - Munich: Franz Guenther, Linguistics
 - Trondheim/Tromsø: Algorithms/Architecture

– EU 6th Framework research projects

- Currently 3 funded projects: Analytical search, Integration of search & case based reasoning, and grid based search architectures

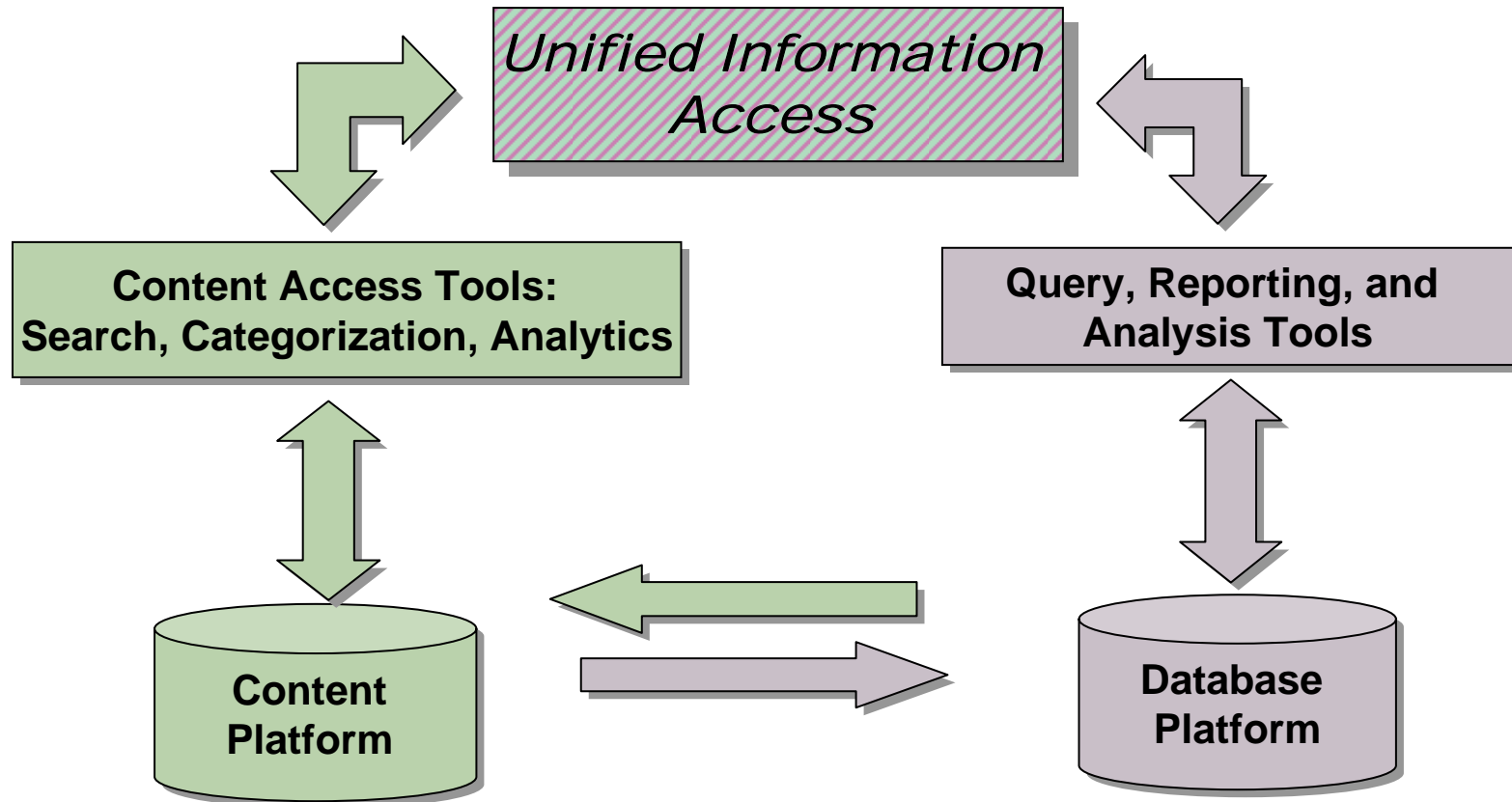


Gartner

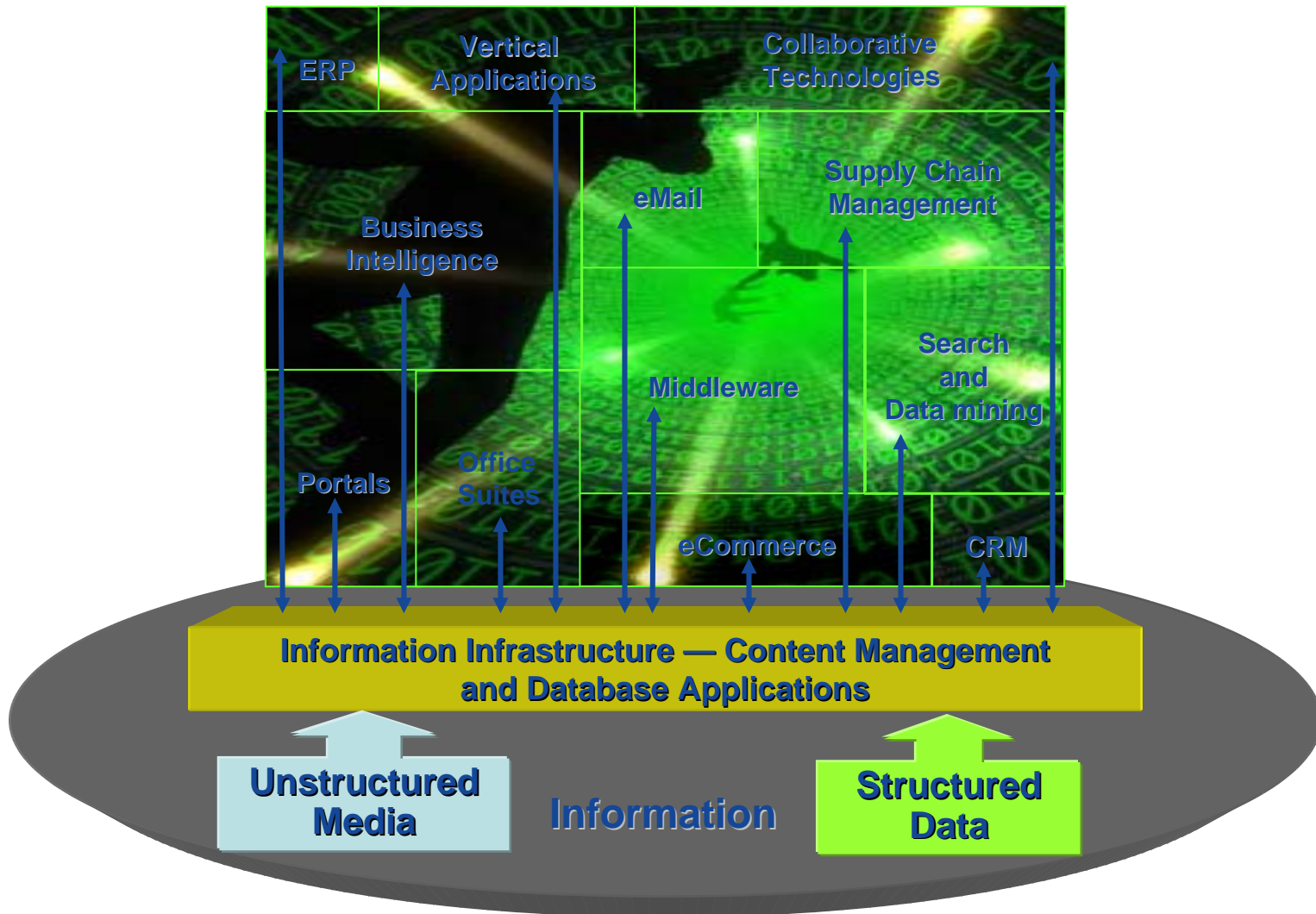
“Magic Quadrant: Most Visionary”



Merging access to content and data



Solving the Information Crisis





Search vs. Database Approach

SEARCH **DOESN'T** SUPPORT...

- Database transaction processing, rollback, ...
- Joins
- Extensive upfront schema modeling
- Pre-aggregation of values in data marts

... (therefore) SEARCH **DO** SUPPORT:

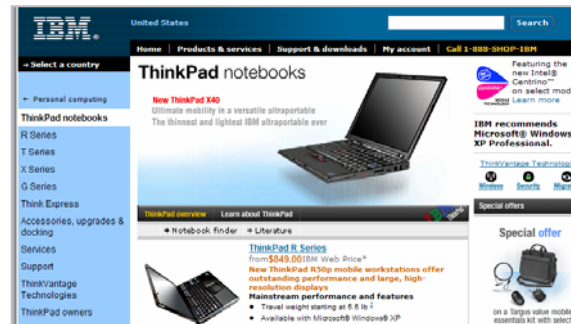
- Scalability:** • 10-100 times more cost efficient data aggregation
- Performance:** • 50-250 times lower search latencies
- Text:** • Both unstructured & structured data
- Intelligence:** • Ranking of results based on importance
- Analytics:** • On-the-fly mining of meta data properties

Where We're Going Now

- Applications that search can supercharge include:
 - Customer Relationship Management
 - Supply Chain Management
 - Business Intelligence
 - Market Intelligence
 - Research Support
 - Threat Detection
 - Anywhere data and unstructured text, speech or general volition collide

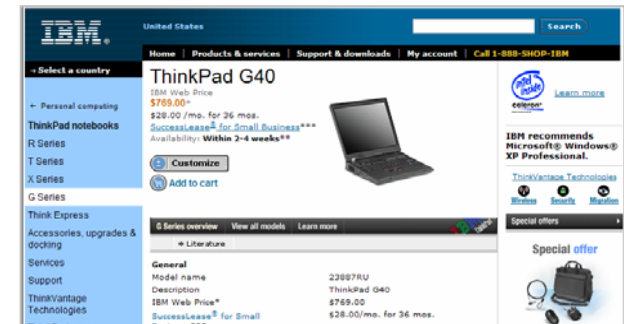
Web sites consist of two different kinds of pages

Navigation pages



ThinkPad Home Page

Destination pages



ThinkPad G40 Product Details

Purpose

Move to next page

Provide information

User question

Where do I go next?

Is this what I wanted?

Traffic

High

Lower

Searches

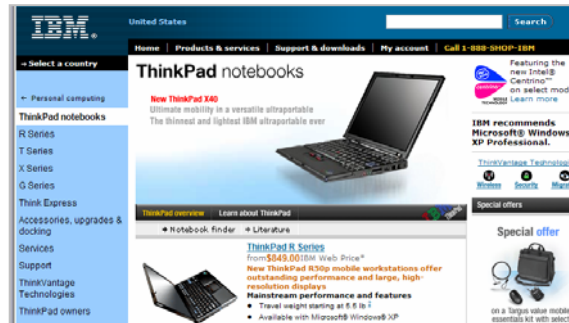
Broad queries

Specific queries

Page types defined in "Information Architecture for the World Wide Web" by Louis Rosenfeld and Peter Morville, p. 139

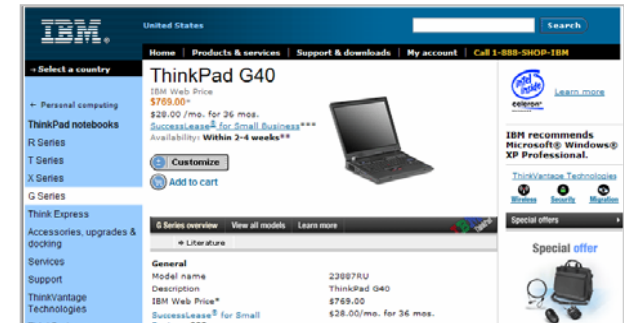
Different queries require different approaches

Broad queries



ThinkPad Home Page

Specific queries



ThinkPad G40 Product Details

Examples

notebook
laptop
thinkpad

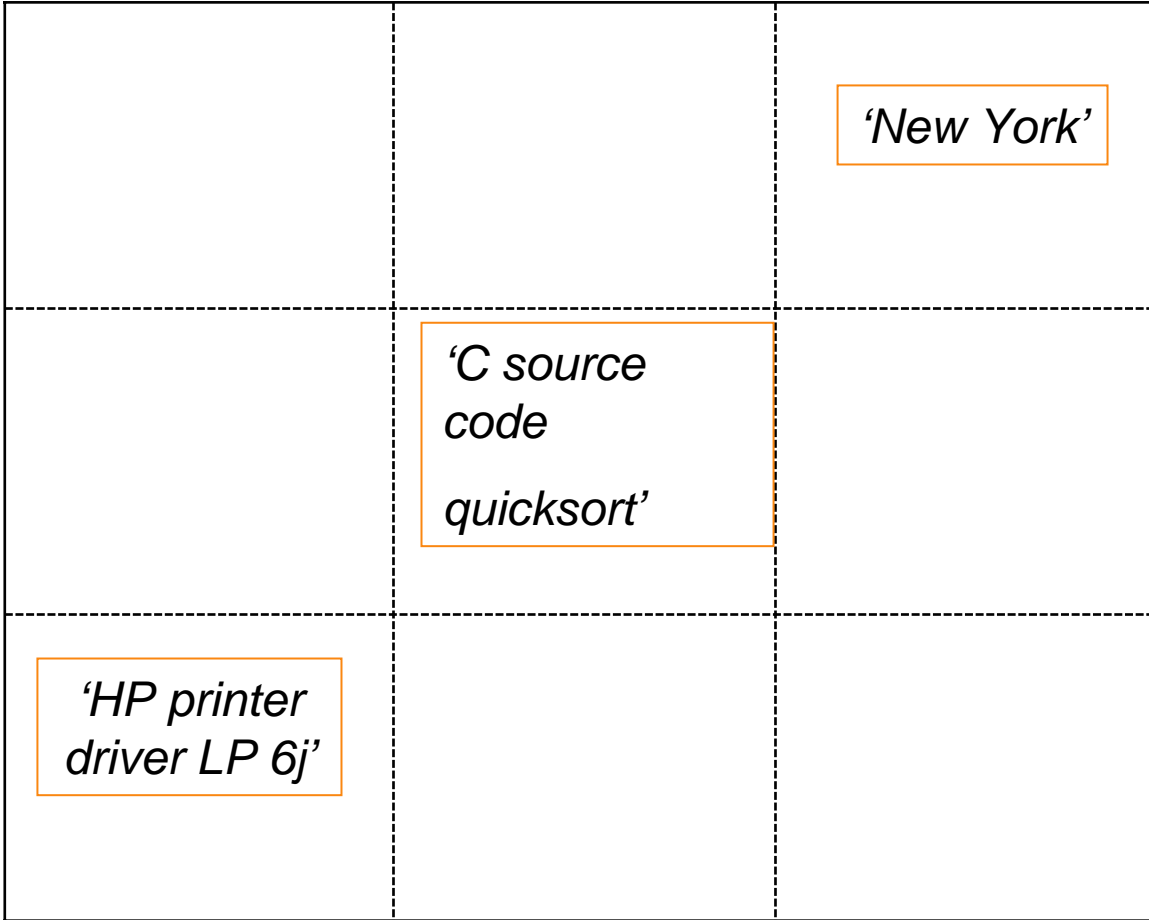
thinkpad g40
23887RU
2388-7RU

Approaches

Keywords on page
Inbound links
Anchor text

Keywords on page
Part numbers on page
(including variations)

The Query – Document Relationship



General Queries			<i>'New York'</i>
Problem Queries		<i>'C source code quicksort'</i>	
Specific Queries	<i>'HP printer driver LP 6j'</i>		
	Content	Format	Reference

Generating a TOC

Case: 12M Medline documents



This is the FAST Data Search 3.2
Lifescience Demo

Results found: 17410
Search time: 435ms
1-10 [11-20](#) [21-30](#) [31-40](#) [41-50](#) ▶

1 . The evolution of species-type specificity in the global DNA sequence organization of mitochondrial genomes.

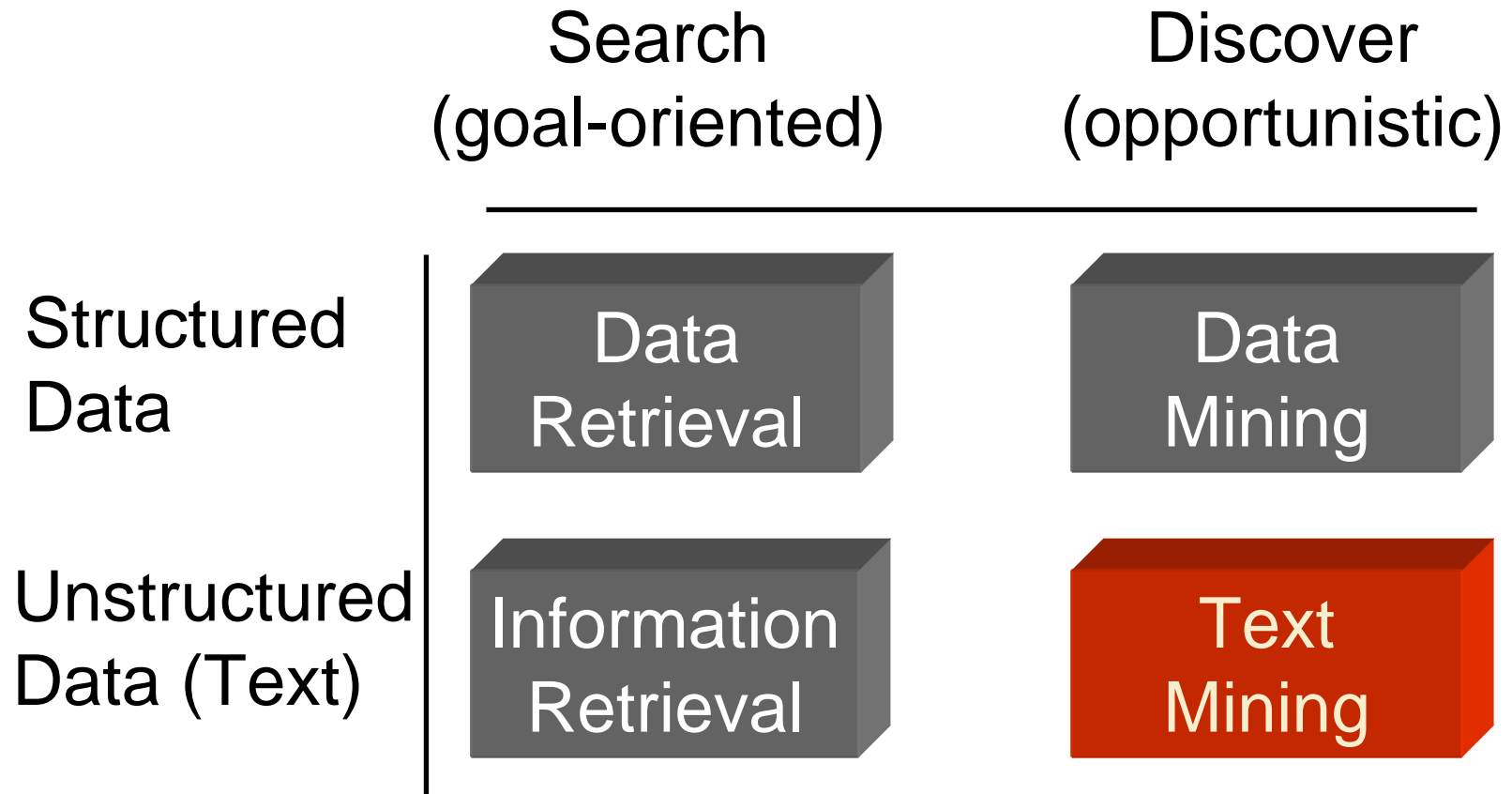
K Hill; S Singh;

Prokaryote genomes and nuclear genomes of eukaryotes have a global DNA sequence organization that is species type specific, determined primarily by nearest-neighbor nucleotide associations, and independent of gene function and sequence length. The determinants of such a global structure have remained largely uncharacterized. The monophyletic and endosymbiotic origin of mitochondria permit examination of the influence of evolutionary time and host species

#	MeSH term	#	chemical substance	#	publication	#	publication year	#	author
97	Human	8	DNA	54	Genome Res	29	1998	8	K Pruitt
58	Animal	8	Genetic Markers	25	Mamm Genome	22	2000	4	J Weissenbach
52	Genome, Human	5	DNA Primers	8	Law Hum Genome Rev	13	1999	4	B Chowdhary
44	Genome	4	DNA, Mitochondrial	3	Proc Natl Acad Sci U S A	12	1996	3	L Frönicke
40	Chromosome Mapping	4	DNA, Complementary	2	Genome	11	1997	3	H Scherthan
38	Support, Non-U.S. Govt	3	DNA Transposable Elements	2	J Virol	4	1995	2	J McPherson
26	Base Sequence	3	Proteins	1	J Mol Evol			2	I Gustavsson
26	Comparative Study	3	Plasmids	1	Virology			2	G Elqar
22	Support, U.S. Govt, P.H.S.	2	DNA-Binding Proteins	1	Int J Parasitol			2	E Green
22	Human Genome Project	2	RNA, Viral	1	FEMS Microbiol Rev			2	T Raudsepp
20	Mice	2	DNA Probes	1	Rom J Physiol			2	M Kirby
18	Molecular Sequence Data	2	DNA, Viral	1	Hum Mol Genet			2	C Fizames
18	Sequence Analysis, DNA	2	DNA, Satellite					2	A Billault
15	Databases, Factual	1	Muscle Proteins					2	G Gyapay
14	Chromosomes	1	Nerve Tissue Proteins					2	D Bud

Dynamic Drill-Down in Auto-Extracted Entities

Search & Discovery



Information Extraction



BC-dynegy-enron-offer-update5
 Dynegy May Offer at Least \$8 Bln to Acquire
 Enron (Update5)
 By **George Stein**
 SOURCEc.2001 Bloomberg News
 BODY

.....
 "Dynegy has to act fast," said **Roger Hamilton**, a money manager with **John Hancock Advisers Inc.**, which sold its Enron shares in recent weeks. "If **Enron** can't get financing and its bonds go to junk, they lose counterparties and their marvelous business vanishes."

Moody's Investors Service lowered its rating on Enron's bonds to "Baa2" and Standard & Poor's cut the debt to "BBB." in the past two weeks.

.....

Fact

Event

<Category>FINANCIAL</ Category >

<Author>George Stein</ Author >

<Company>Dynegy Inc</Company>

<Person>Roger Hamilton</Person>

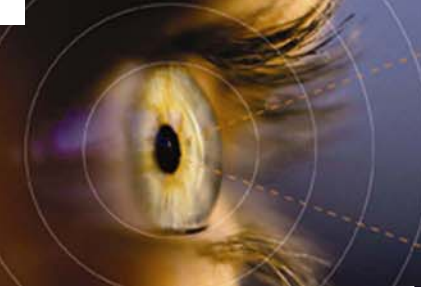
<Company>John Hancock Advisers Inc. </Company>

<PersonPositionCompany>
 <OFFLEN OFFSET="3576" LENGTH="63" />
 <Person>Roger Hamilton</Person>
 <Position>money manager</Position>
 <Company>John Hancock Advisers Inc.</Company>
 </PersonPositionCompany>

<Company>Enron Corp</Company>

<Company>Moody's Investors Service</Company>

<CreditRating>
 <OFFLEN OFFSET="3814" LENGTH="61" />
 <Company_Source>Moody's Investors Service</Company_Source>
 <Company_Rated>Enron Corp</Company_Rated>
 <Trend>downgraded</Trend> <Rank_New>Baa2</Rank_New>
 <__Type>bonds</__Type>
 </CreditRating>



Terminology extraction

Example: SCIRUS.com – 160 M Documents

All journal sources
 All Web sources
 Exact phrase

Searched for:	All of the words niels bohr
Found:	20,219 total 3,275 journal results 16,944 Web results
Sort by:	relevance date

- Refine your search using these keywords found in the results:
- [albert einstein](#)
 - [astronomical observatory](#)
 - [atomic bomb](#)
 - [atomic model](#)
 - [atomic nucleus](#)

- [atomic theory](#)
- [charged nucleus](#)
- [college park](#)
- [complementarity](#)
- [host galaxy](#)
- [nuclear physics](#)
- [orbits](#)
- [philosophy of science](#)
- [physicists](#)
- [quantum theory](#)
- [theoretical physics](#)

- Refine your search using these keywords found in the results:
- [albert einstein](#)
 - [angular momentum](#)
 - [atomic nuclei](#)
 - [atomic physics](#)
 - [atomic theory](#)
 - [cathode](#)
 - [neutron](#)
 - [nuclear model](#)
 - [nuclear physics](#)
 - [orbits](#)
 - [physicists](#)
 - [quantum physics](#)
 - [quantum theory](#)
 - [radioactivity](#)
 - [theoretical physics](#)
 - [uncertainty principle](#)

Niels Bohr – Biography



Niels Henrik David Bohr was born in Copenhagen on October 7, 1885, as the son of Christian Bohr, Professor of Physiology at Copenhagen University, and his wife Ellen, *née* Adler. Niels, together with his younger brother Harald (the future Professor in Mathematics), grew up in an atmosphere most favourable to the development of his genius - his father was an eminent physiologist and was largely responsible for awakening his interest in physics while still at

school, his mother came from a family distinguished in the field of education.

After matriculation at the Gammelholm Grammar School in 1903, he entered Copenhagen University where he came under the guidance of Professor C. Christiansen, a profoundly original and highly endowed physicist, and took his Master's degree in Physics in 1909 and his Doctor's degree in 1911.

The Nobel Prize in Physics 1922

Presentation Speech

- Niels Bohr**
- [Biography](#)
 - [Nobel Lecture](#)
 - [Banquet Speech](#)
 - [Swedish Nobel Stamps](#)
 - [Other Resources](#)

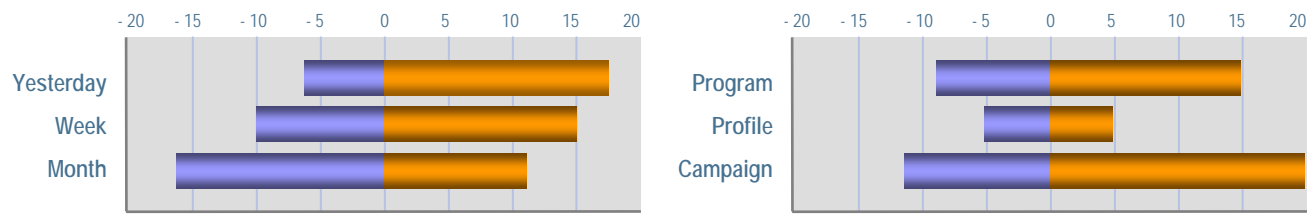
1921 1923

The 1922 Prize in:
 Physics
 Chemistry
 Physiology or Medicine
 Literature
 Peace

Find a Laureate:

Sentiment Analysis

FOX



▶ **Mark Shields: Stop-Dean Movement Stumbles**

Moreover | 11/10/2003 | 00:00
 (...) CNN.com SERVICES SEARCH Web CNN.com Mark Shields, nationally known columnist and commentator, is the moderator of CNN's The Capital Gang The Stop-Dean movement stumbles Story Tools WASHINGTON (Creators Syndicate)-- On February 21, 2003, former Vermont Gov. **Howard Dean**, then the darkest of dark horses for his party's (...) >>

▶ **Baghdad Hero Dismisses Ex-Nato Commander'S Presidential Bid**

Moreover | 11/10/2003 | 00:00
 (...) Libya, Somalia, Iran, and Sudan. In a weekend poll, congressman Dick Gephardt was found to be the front runner for the January 19 Iowa caucuses, a first step in the nomination of a Democratic party candidate. With 27% support he overtook former Vermont Governor **Howard Dean** (with 20%) as the (...) >>

▶ **Poll: 50 Percent Of Voters Against Bush**

Moreover | 11/10/2003 | 00:00
 (...) following positive news this week. Forty-four percent said they approve of the way Bush is handling the economy - up six points from the magazine's previous poll a month ago. Forty-eight percent said they disapprove. Among contenders for the Democratic presidential nomination, former Vermont Gov. **Howard Dean** edges out (...) >>

▶ **Kerry Aims Ads To Catch Up With Dean**

Moreover | 11/10/2003 | 00:00
 (...) a commercial Friday, an effort his advisers hope will help him narrow the gap with front-runner **Howard Dean**. The Massachusetts senator will run the same 30-second ad in New Hampshire that began airing in Iowa last week, said Jim Margolis, Kerry's media consultant. Campaign manager Jim Jordan said (...) >

▶ **Gore Denounces Bush On Civil Liberties**

Moreover | 11/10/2003 | 00:00
 (...) security - charging that there aren't sufficient protections in place for ports, nuclear facilities, chemical plants and other key infrastructure. His speech was sponsored by the liberal activist group Moveon.org, which earlier this year held an on-line presidential primary in which **Howard Dean** finished first. The second sponsor, the American (...) >>

▶ **Sen. John Kerry Fires Campaign Manager**

Moreover | 11/10/2003 | 00:00
 (...) Kerry told Jordan the reason he was removed was because changes were needed in the campaign. Kerry has been trailing considerably behind former Vermont Gov. **Howard Dean** in the polls. "From the bottom of my heart, I thank Jim Jordan for his leadership, extremely hard work, unsurpassed loyalty and devotion (...) >>

Related People

- ▶ Howard Dean
- ▶ Jim Jordan
- ▶ Sen. John Edwards
- ▶ President Bush
- ▶ Mark Shields
- ▶ Frank Sinatra
- ▶ Wesley Clark
- ▶ Bill Clinton
- ▶ Sen. John
- ▶ John Rocker

Related Companies

- ▶ Time Warner
- ▶ Associated Press
- ▶ Newsweek
- ▶ Digital Chicago
- ▶ Enron
- ▶ Chase Manhattan Mortgage
- ▶ Hary N Abrams
- ▶ State Farm Mutual Auto Ins
- ▶ Two Guys



Information discovery

Value



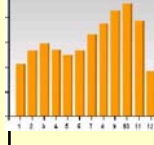
Viewed results

Noise

Documents

Meta-data

Autodetected Entities

	Price	Color	Abstract	Names	Brands	Geo	Topic
D_1							
D_2							
D_3							
D_4							
D_{10}							
...							
D_n							
Live Analytics™ 	Min Max Mean ...	Red 23 Green 17 Blue 5 Yellow 1 ...	Clustering	Arnold 7 George 4 ...	Sony 72 HP 45 ...	London 5 Oslo 4 ...	Sport 7 News 5 ...


Attributes



Analyzed results

Information Discovery

Example: Yellow Page

Gule Sider® | Telefonkatalogen™ | Search by number | Maps | SMS |  Log in

bmw

[Choose category](#) | [Choose location](#) | [Detailed search](#)

Garages (6 of 47 hits) [Show hits in map](#)

Maranello Motor AS    

Billingsstadsletta 83, 1396 Billingsstad 66 85 32 00





Jæger Automobil AS    

Kanalv. 111, 5068 Bergen 55 33 55 00

Furubakken Bilverksted A/S    

Furuv. 2, 1356 Bekkestua 67 59 12 37

Stabekk Biloppretting    

Gamle Drammensv. 20, 1369 Stabekk 67 53 81 00

Gulbrandsen Arne A/S    

Sverres g 3, 0652 Oslo 23 03 74 10

Martinsen Maskin og Automatqir    

3530 Røyse 32 15 73 40

Choose part of country:

- East Norway (126)
- Middle Norway (22)
- Northern Norway (21)
- Southern Norway (8)
- Western Norway (74)

Hits in business categories:

[Alphabetical order](#) | **Number of hits**

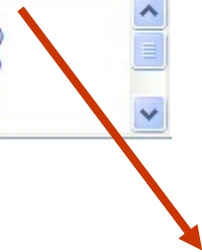
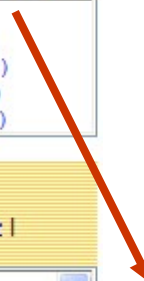
- Car dealers (54)
- Garages (47)
- Car parts and accessories
- Garages - Body work (10)
- Car breakers (8)
- Car parts and accessories
- Motorcycles and motor scoo
- Garages - Paint work (8)
- Driving schools (7)
- Car rental (5)
- Car dealers - import and wt
- Boat engines and repair (3)
- Auto glass (3)
- Car tyres and rims (3)
- Motorcycle and motor scoot
- Garages - Electrical (2)
- Caravans and mobile homes
- Boats (2)
- Car junkyards (1)
- Tour operators (1)
- Trailers and supplies (1)
- Rescue services (1)
- Financing services (1)
- Camping equipment (1)
- Boat slips (1)

Hits in country part: East Norway

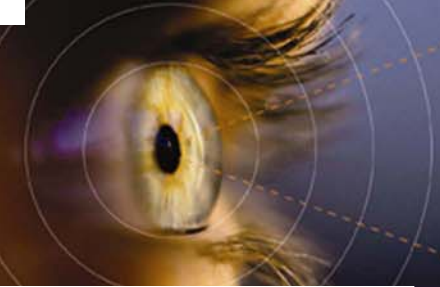
- Oslo (41)
- Akershus (25)
- Buskerud (17)
- Hedmark (10)
- Oppland (10)

Hits in municipality: Oslo

- Alna (7)
- Gamle Oslo (10)
- Grorud (3)
- Grünerløkka (7)
- Sagene (1)



Understanding content & users



Analyze Content

Analyze Query

Our research interests are focused on the mechanisms of RNA decay and their role in gene regulation. One of the biological systems employed in our investigations is RNAI, the antisense repressor of DNA replication of ColE1-type plasmids. Previously, we have shown that the rate of cleavage of RNAI by the endoribonuclease RNase E, an E. coli endoribonuclease essential for cell viability and decay of a variety of RNA species, is dependent on the sequence of the RNAI.

Unstructured Data

replication (Cell, 65: 1233-42, 1981).

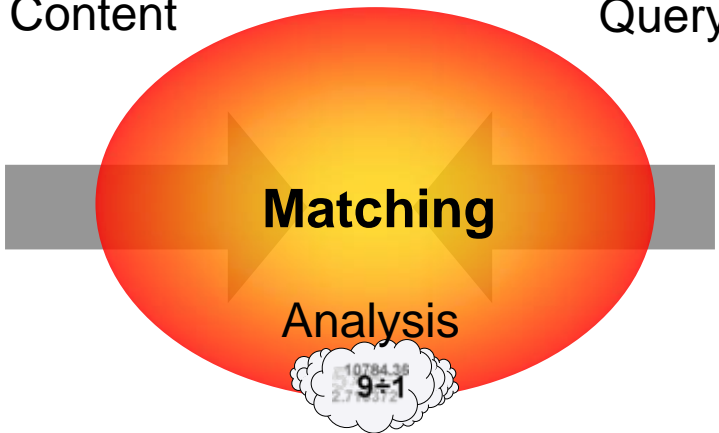
In subsequent studies, using genetic, molecular biological and biochemical approaches, we have defined RNase E-cleavage-specificities (J. Biol. Chem., 269, 10790-6, 1994; J. Biol. Chem., 269: 10797-803, 1994; Nature, 374: 287-90, 1995; J. Biol. Chem., 271: 13103-9, 1996). To further understand the mechanism of RNase E-mediated RNA decay, we had isolated a multi-component ribonucleolytic complex termed RNA degradosome, that contains RNase E and other associated proteins as well as RNA components (Fig.1; Proc. Natl. Acad. Sci. USA, 93: 3865-9, 1996).

We further identified that the RNA components are 23S rRNA, 16S rRNA, rne and malT mRNA fragments, 5S rRNA, 10Sa RNA and its decay intermediates, and demonstrated that rRNA degradation is carried out in the degradosome by RNase E cleavage of A+U-rich single-stranded regions of mature 23S and 16S rRNA (Fig. 2; Proc. Natl. Acad. Sci. USA, 95: 3157-61, 1998).

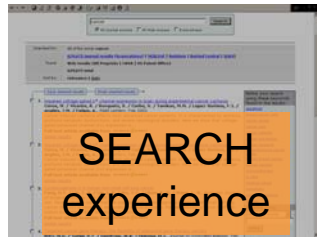
Our current investigations are aimed at the action of proteins in the degradosome on RNA degradation and processing. First, we identified locations of immuno-genic

#	Form Factor	
53	Notebook	+ - >
18	Tower (4x5)	+ - >
12	Desktop (4x4)	+ - >
11	Small Desktop (4x3)	+ - >
10	Small Desktop (4x3)	+ - >
4	Mini desktop (0x2)	+ - >

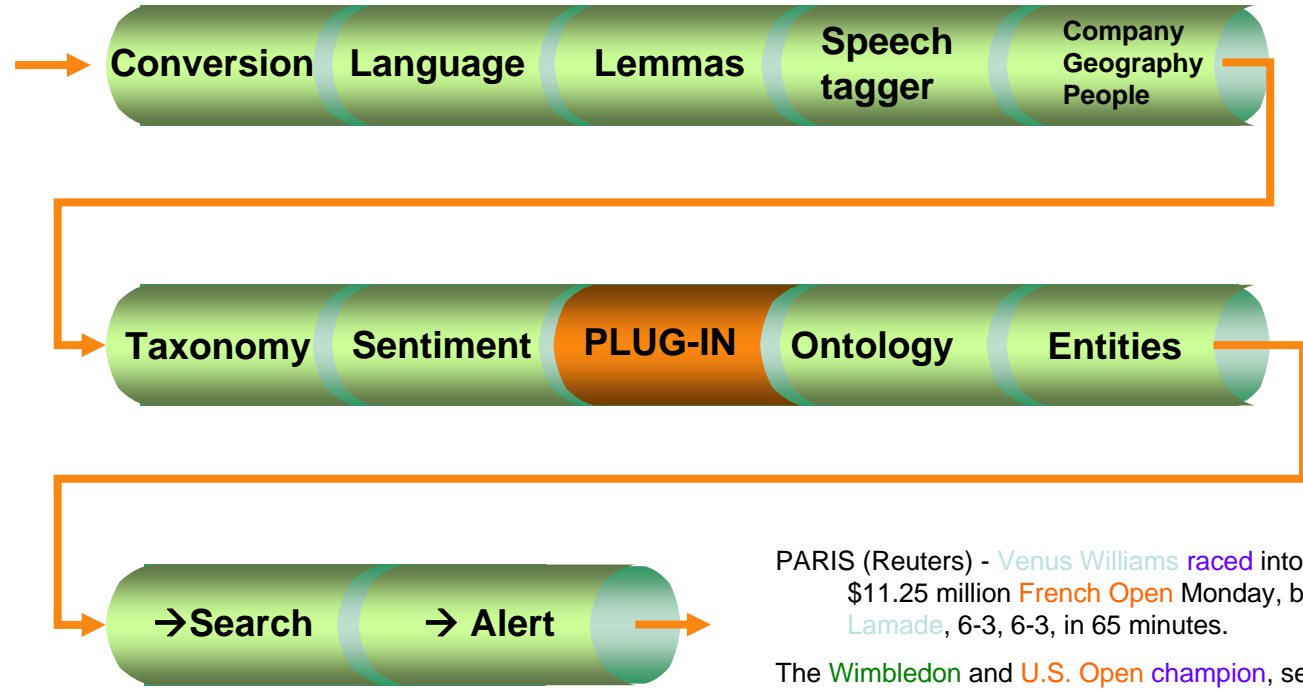
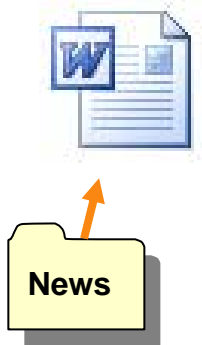
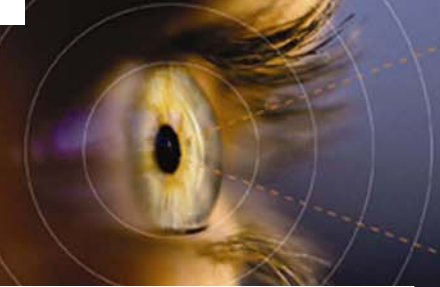
Structured



10784.35
27
9 ÷ 1



Real-Time Content Refinement



PARIS (Reuters) - Venus Williams raced into the second round of the \$11.25 million French Open Monday, brushing aside Bianka Lamade, 6-3, 6-3, in 65 minutes.

The Wimbledon and U.S. Open champion, seeded second, breezed past the German on a blustery center court to become the first seed to advance at Roland Garros. "I love being here, I love the French Open and more than anything I'd love to do well here," the American said.

A first round loser last year, Williams is hoping to progress beyond the quarter-finals for the first time in her career.



The *InPerspective* ranking model

Freshness

- How fresh is the document compared to the time of the query?

Completeness

- How well does the query match superior contexts like the title or the url?
- *Example: query="Mexico", Is "Mexico" or "University of New Mexico" best?*

Authority

- Is the document considered an authority for this query?
- *Examples: Web link cardinality, article references, product revenue, page impressions, ...*

Statistics

- How well does the contents of this document on overall match the query?
- *Examples: Proximity, context weights, tf-idf, degree of linguistic normalization, ++*

Quality

- What is the quality of the document?
- *Examples: Homepage?, Entry point to product group?, Press release?, ...*

Linguistic query analysis NLP

Query:
Do you have a
LCD monitor
under \$900?

~~Do you have a~~



Under \$900?
→
price < 900

LCD monitors
TFT monitor

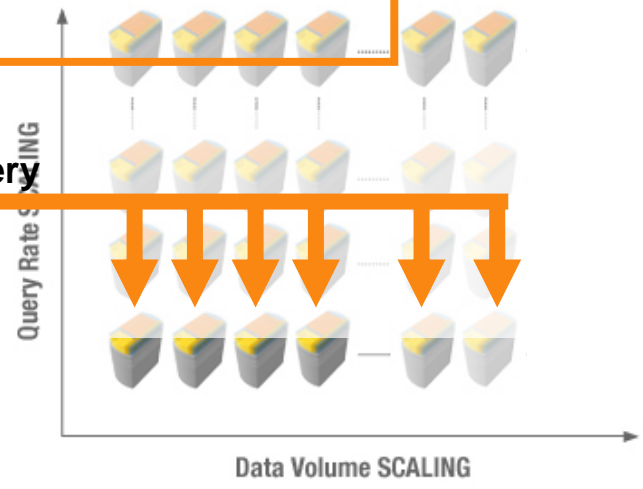
Flat TV
Plasma TV

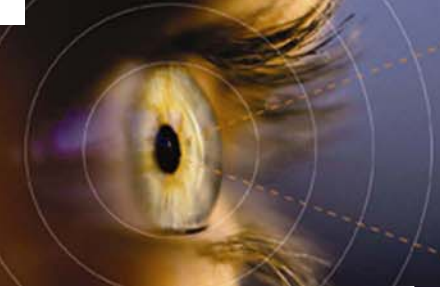
YES!
X = LCD monitor



Use "Product" collection
Rank profile = "Profit margin"

Modified query

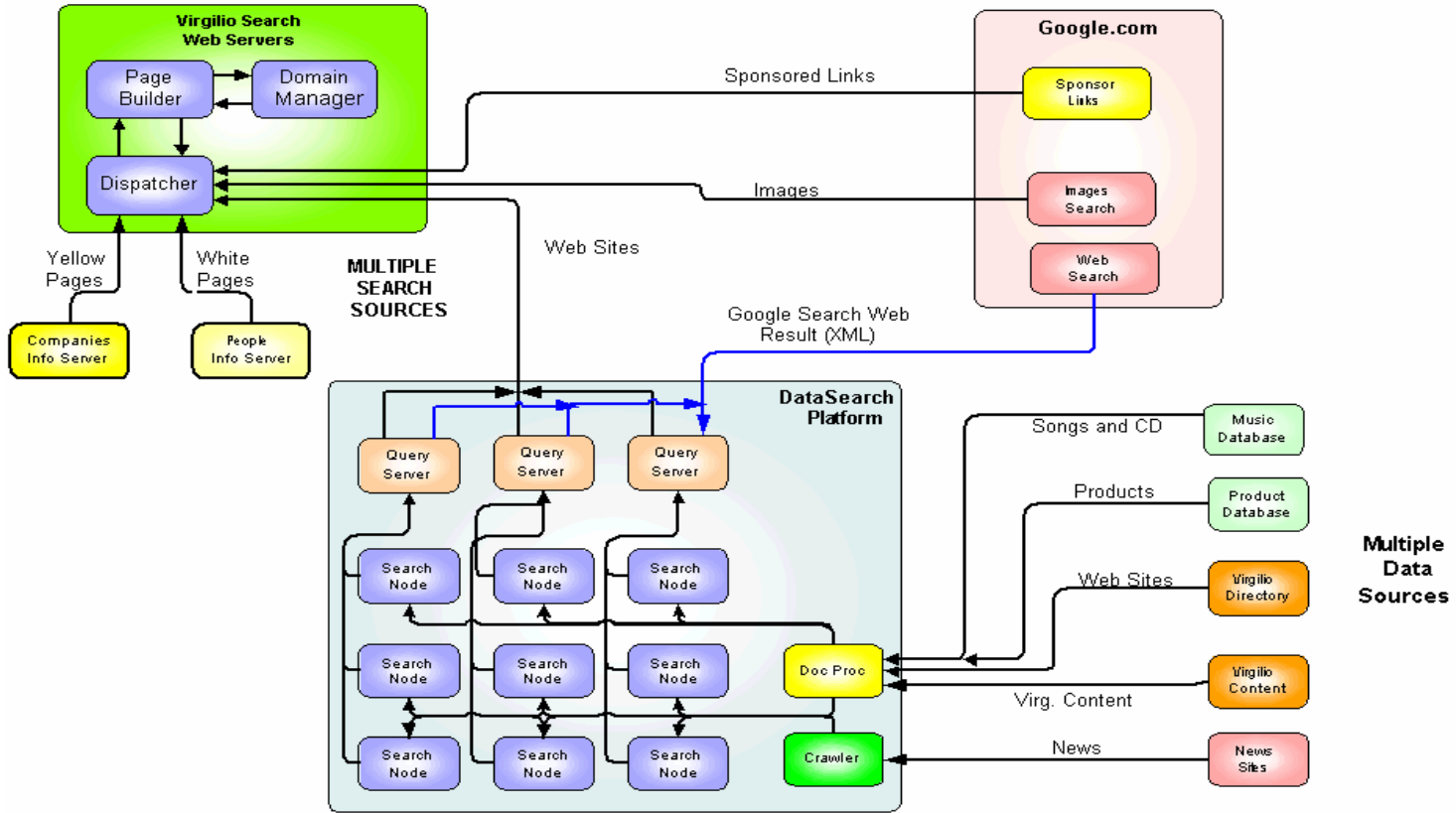




Federated Search

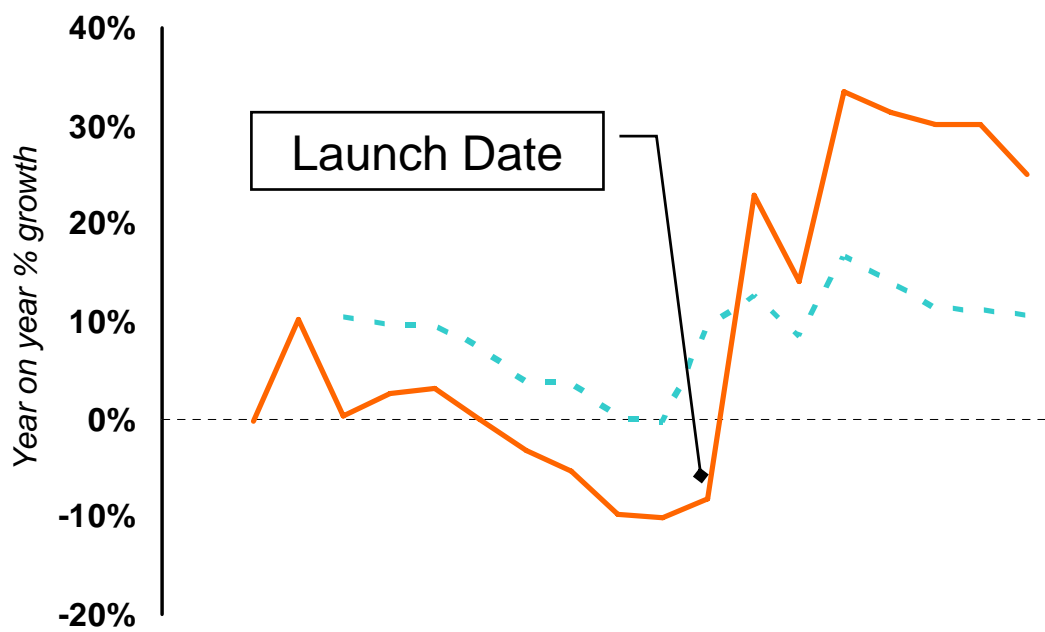
Use case: Virgilio – The largest Italian portal

“Federated Search” Architecture



Virgilio's Results

Results so far...



After New Search Launch:

- +27% Avg. traffic growth
- +12% Avg. users growth
- +12% Relevance Index vs Google
- Market leader in € value
- Sole competitor vs. usage leader

Summary

- Search engines can do more than just search...
 - Unified information access solution for digital libraries
 - Open, scalable and modular architecture: Allows for customization
 - Adapts to content and queries
 - Powerful data discovery, navigation, and visualization
- Many exciting technology developments to come
 - More advanced content and query analysis
 - Adaptive, personalized query- & content-sensitive matching
 - Dynamic result set presentation, navigation, discovery, visualization
 - Federation across external content applications