INF3800/INF4800
Søketeknologi

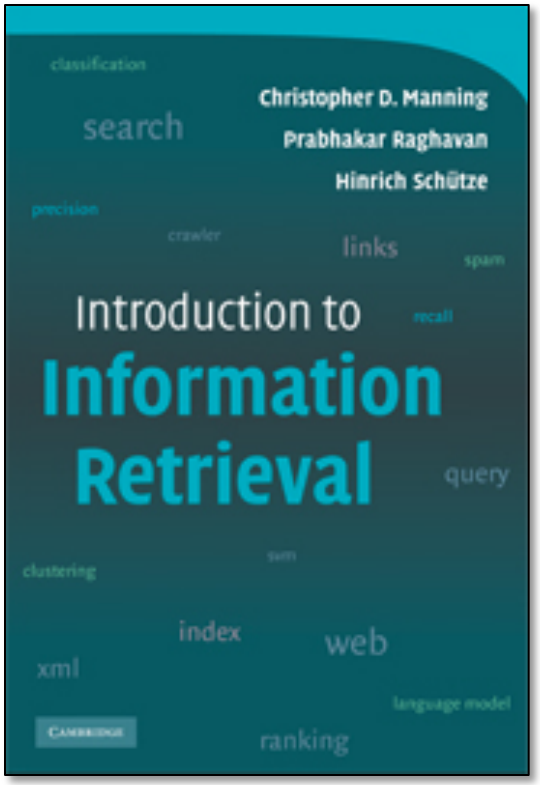
2013.01.16

Aleksander Øhrn
aleksaoh@ifi.uio.no

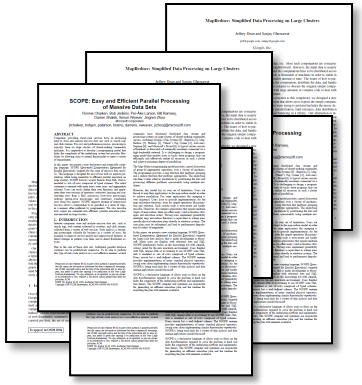
http://www.uio.no/studier/emner/matnat/ifi/INF3800/v13/

Pensum

http://nlp.stanford.edu/IR-book/information-retrieval-book.html



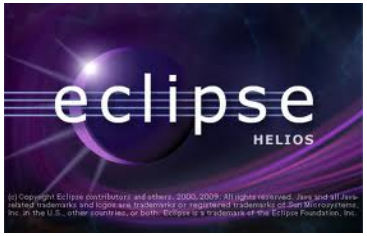
+



Øvinger

x

Al[illegible]gatori[illegible]e*

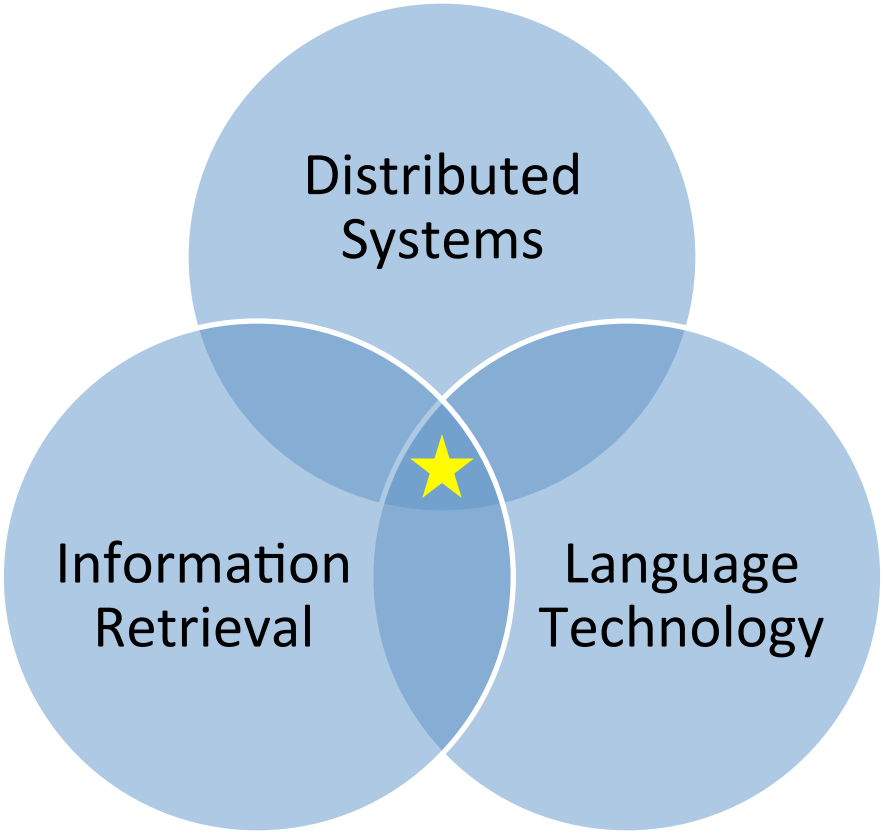


Eksamen

11. juni 2013
14:30
4 timer skriftlig eksamen

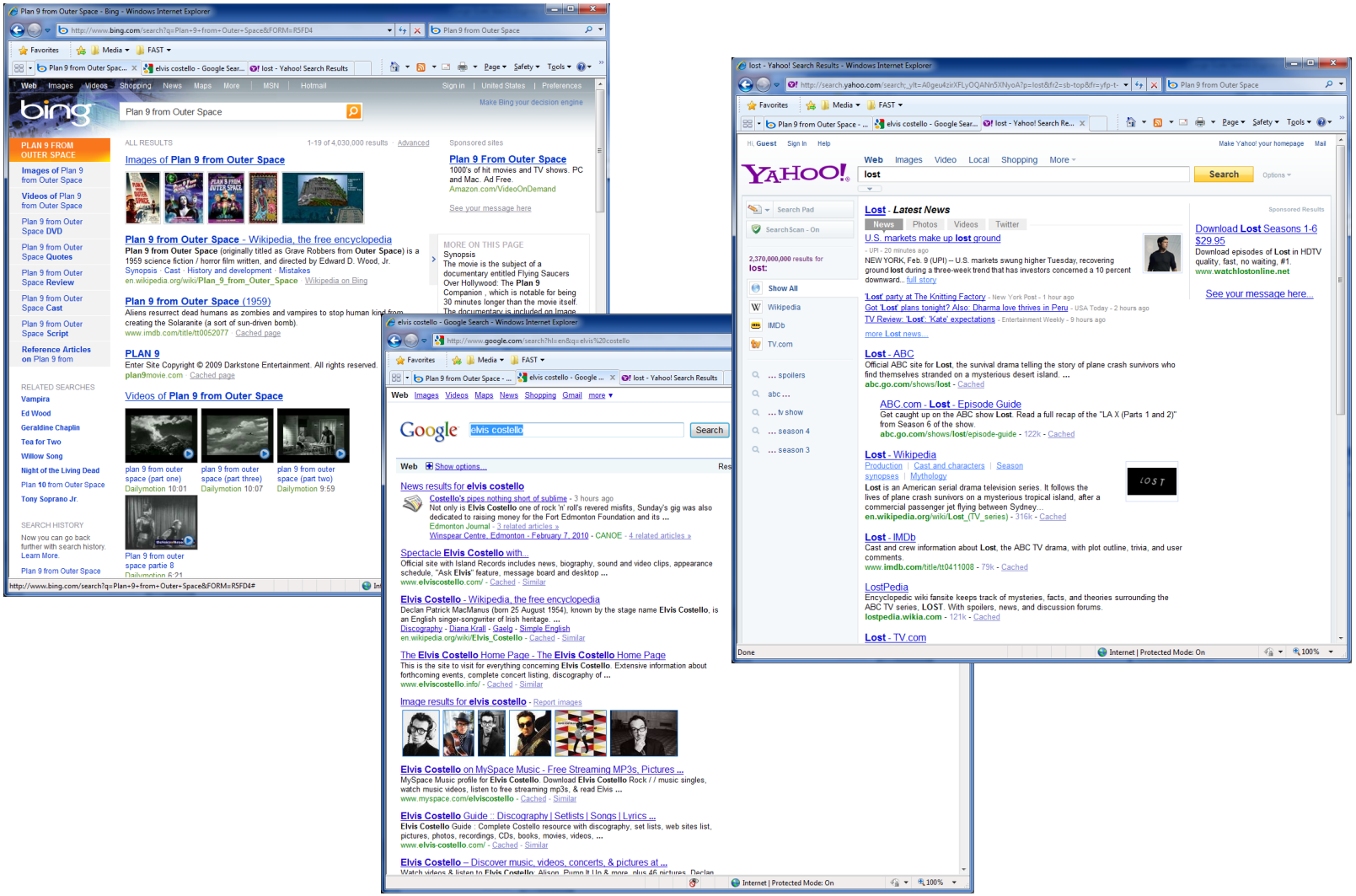
Introduksjon

The Sweetspot



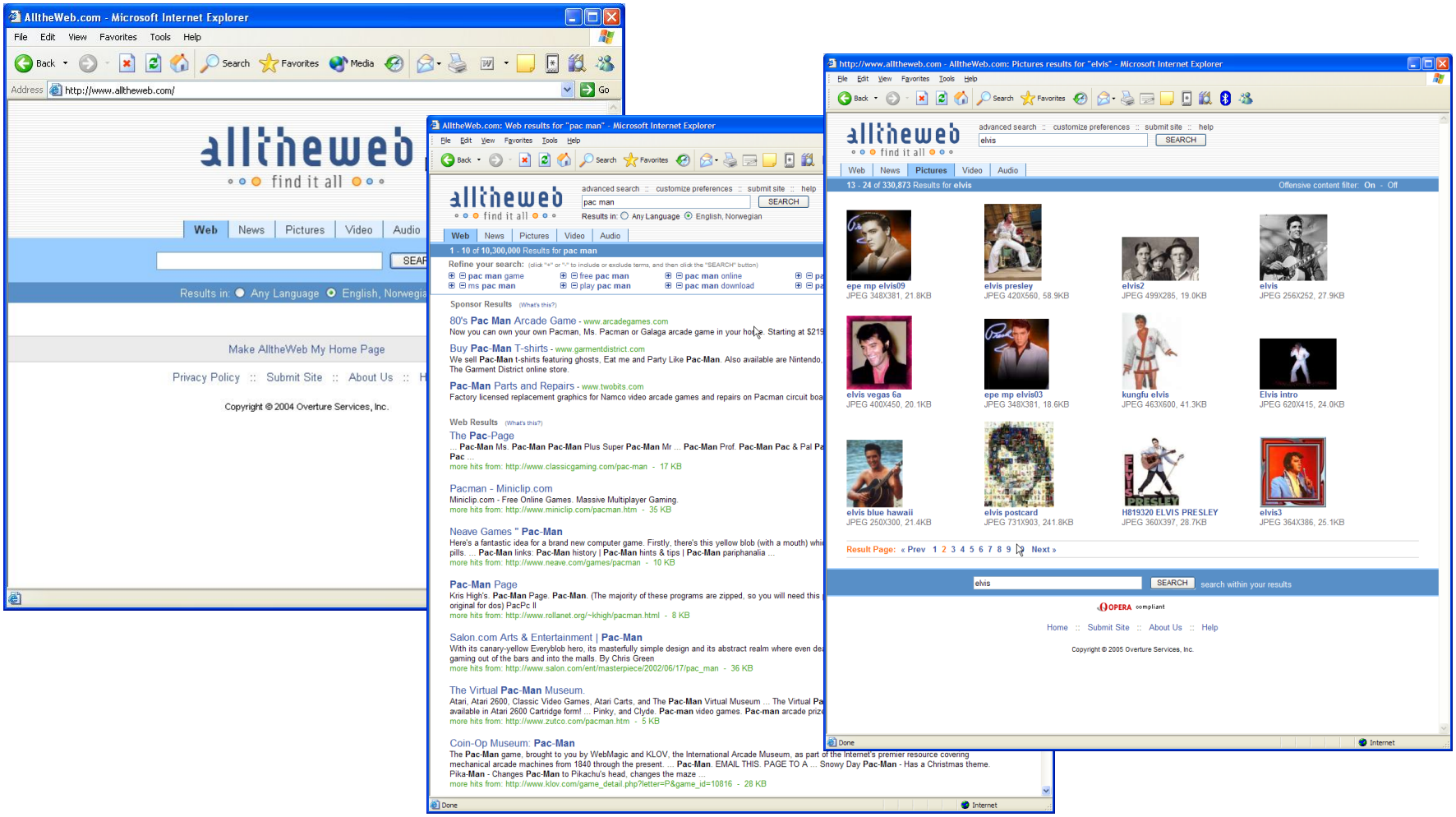
SECRET SAUCE FACTORY

Web Search


Plan 9 from Outer Space - Bing - Windows Internet Explorer
bing
Plan 9 from Outer Space
Images of Plan 9 from Outer Space
Plan 9 from Outer Space - Wikipedia, the free encyclopedia
Plan 9 from Outer Space (1959)
PLAN 9
Videos of Plan 9 from Outer Space
lost - Yahoo! Search Results - Windows Internet Explorer
YAHOO!
lost
Search
Lost - Latest News
Lost - ABC
Lost - Wikipedia
Lost - IMDb
LostPedia
Lost - TV.com
elvis costello - Google Search - Windows Internet Explorer
Google
elvis costello
Search
News results for elvis costello
Elvis Costello - Wikipedia, the free encyclopedia
Image results for elvis costello
Elvis Costello on MySpace Music - Free Streaming MP3s, Pictures ...

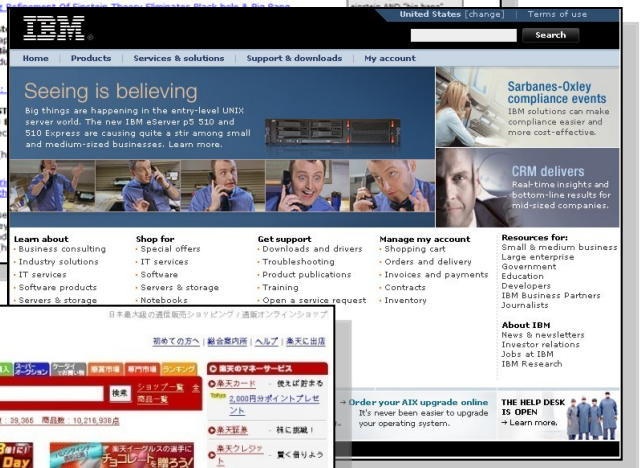
alltheweb.com

1999-2003



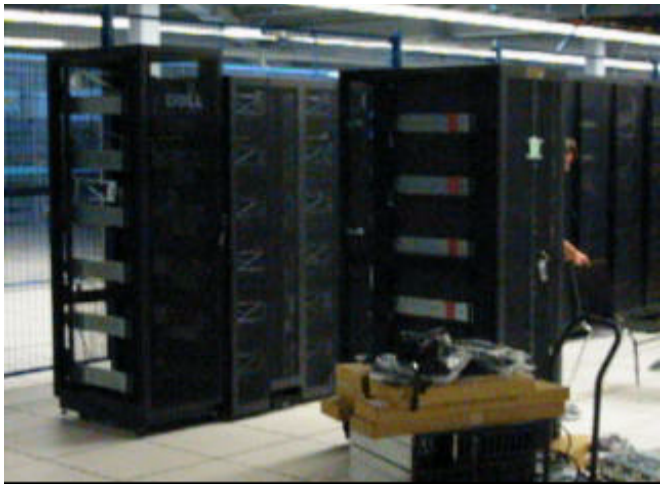
Enterprise Search

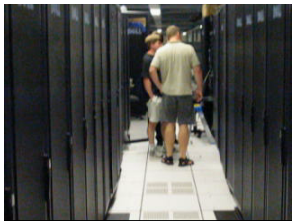
Much more than intranets

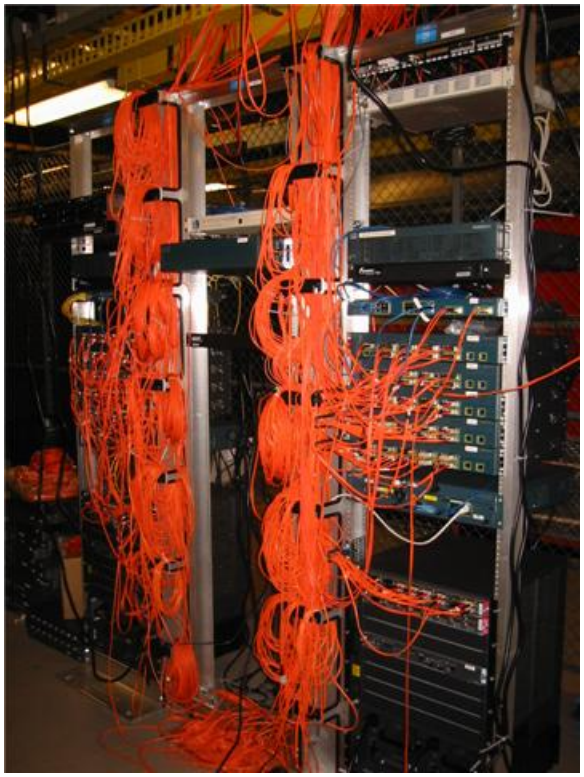


Data Centers

allltheweb.com 2000

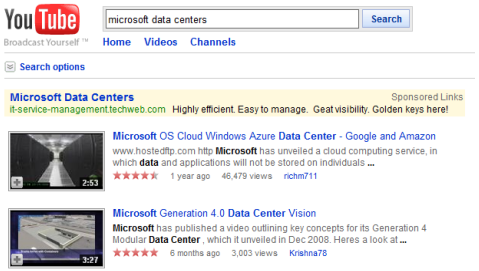






Data Centers

Microsoft 2010

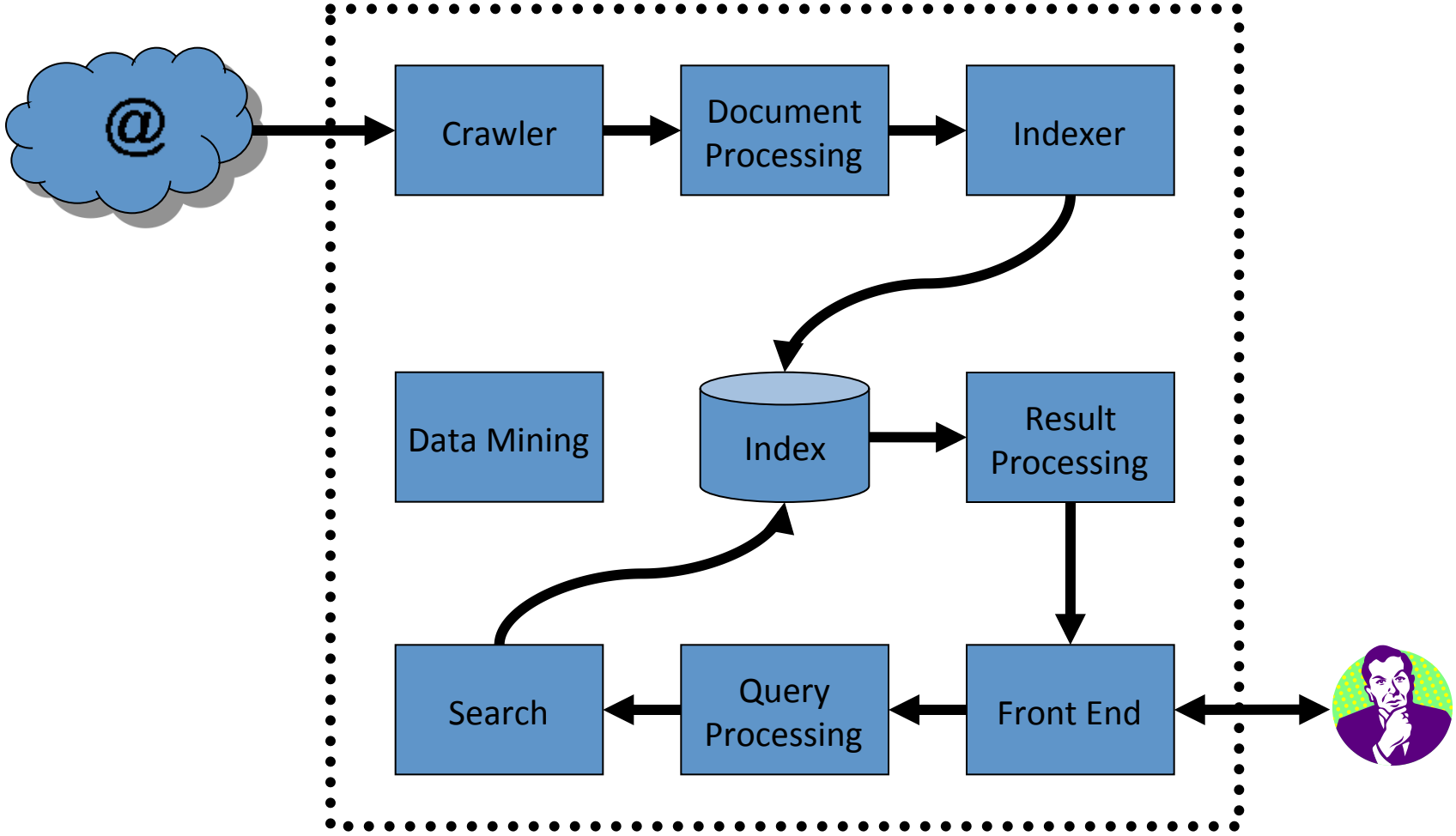


http://www.youtube.com/watch?v=K3b5Ca6lzqE

http://www.youtube.com/watch?v=PPnoKb9fTkA

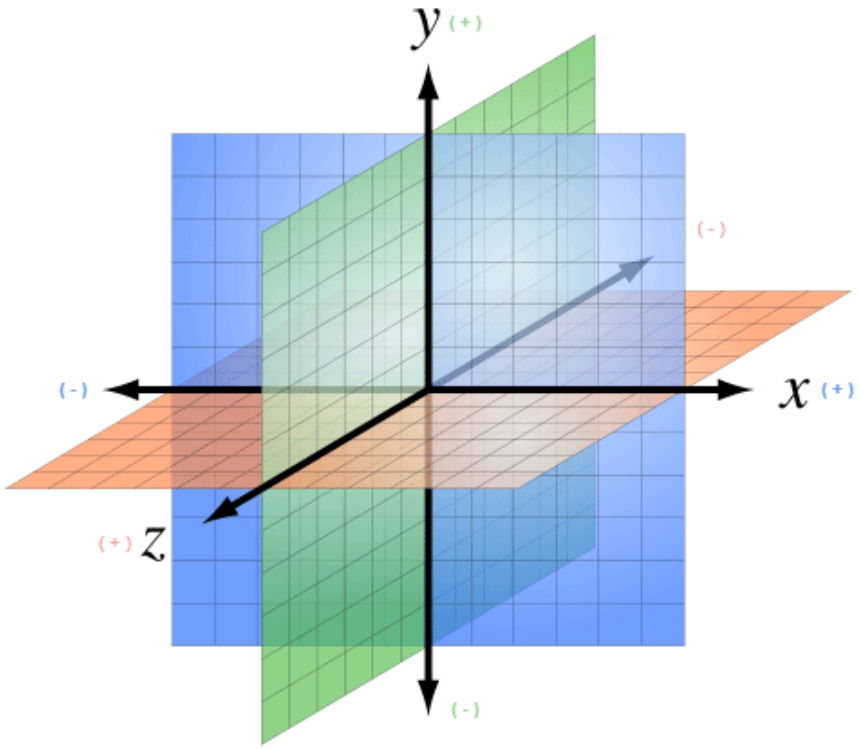
Search Platform Anatomy

The 50,000 Foot View

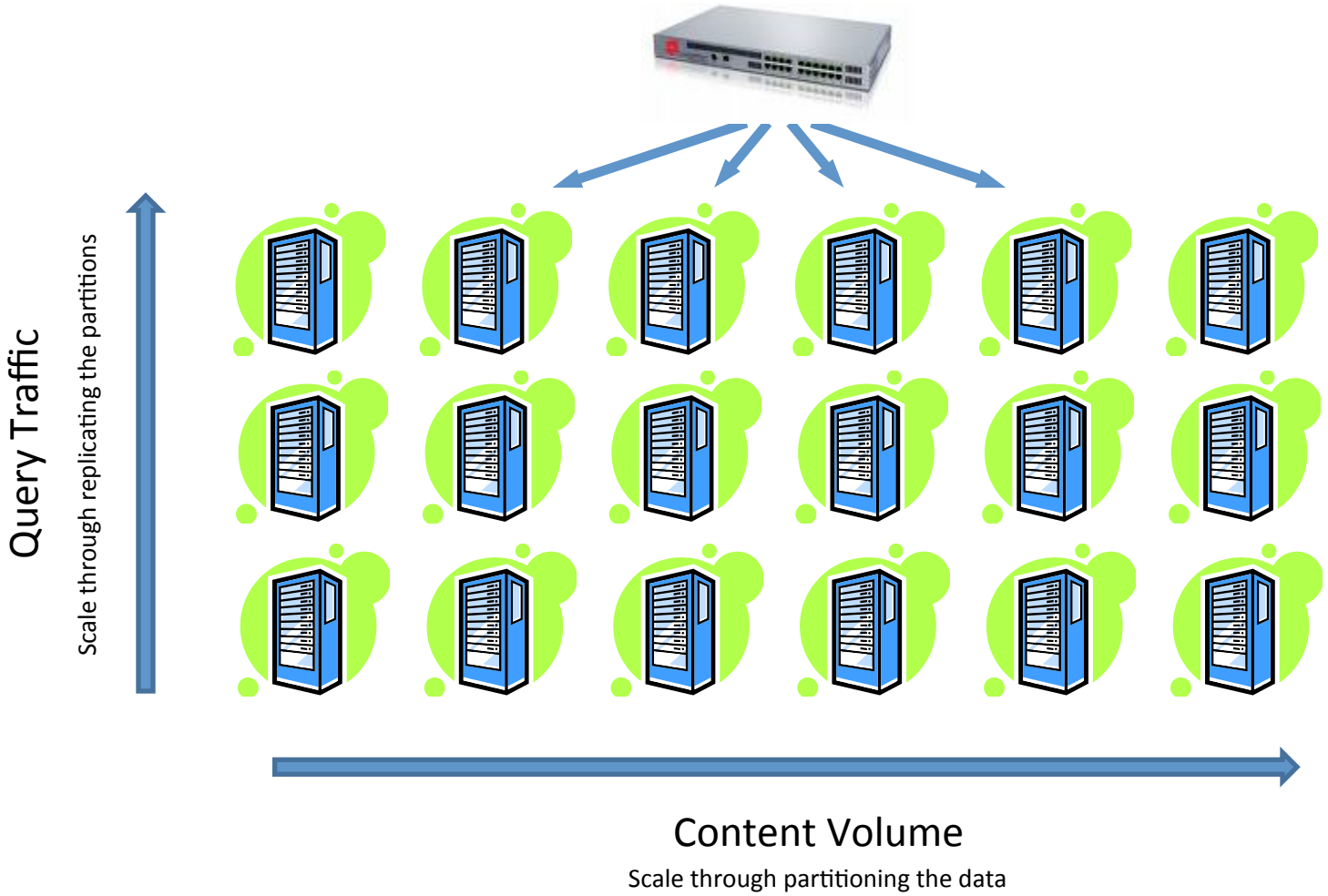


Scaling

- **Content Volume**
  - How many documents are there?
  - How large are the documents?
- **Content Complexity**
  - How many fields does each document have?
  - How complex are the field structures?
- **Query Traffic**
  - How many queries per second are there?
  - What is the latency per query?
- **Update Frequency**
  - How often does the content change?
- **Indexing Latency**
  - How quickly must new data become searchable?
- **Query Complexity**
  - How many query terms are there?
  - What is the type and structure of the query terms?

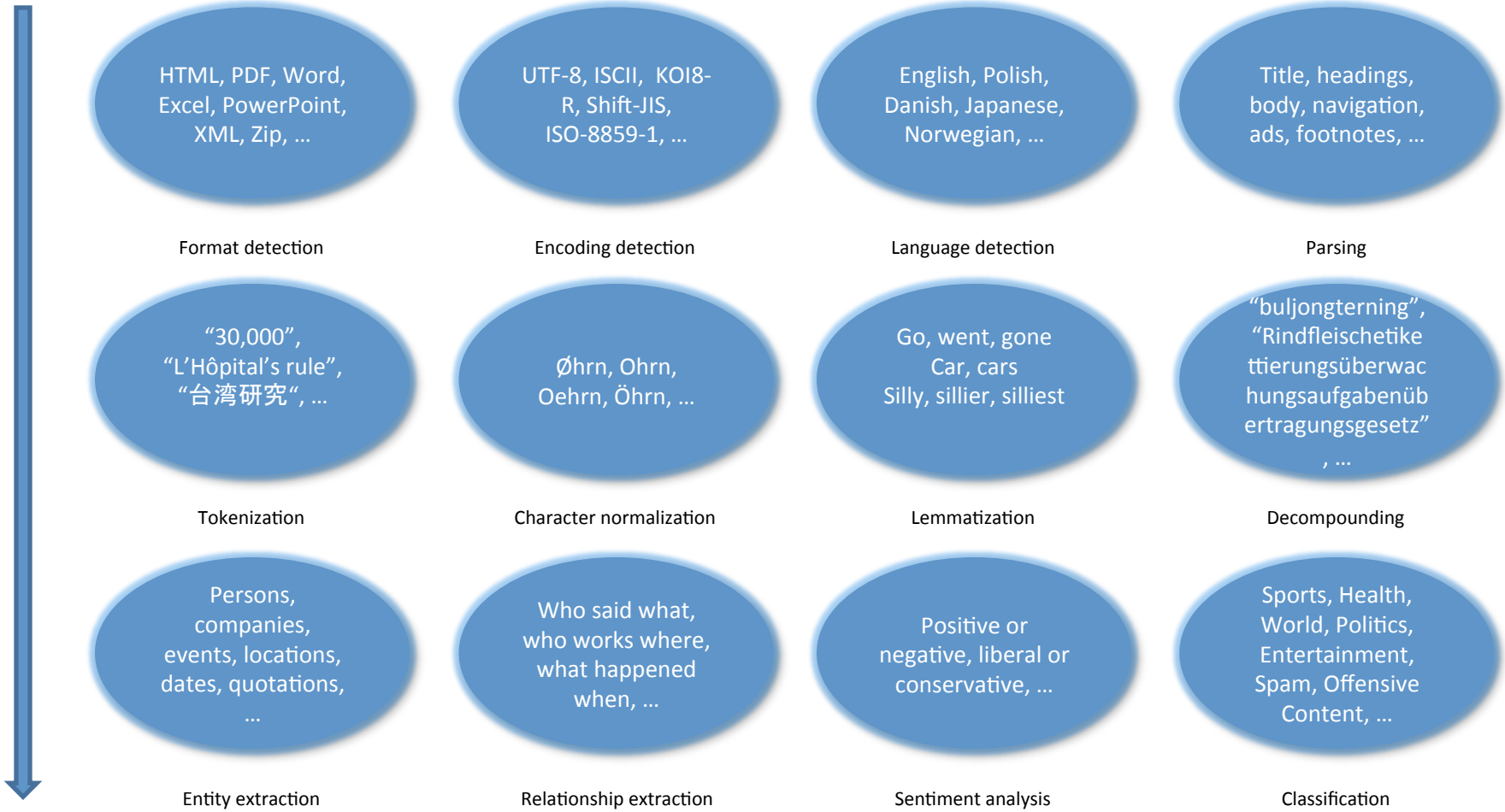


Scaling


Query Traffic
Scale through replicating the partitions
Content Volume
Scale through partitioning the data

Crawling The Web

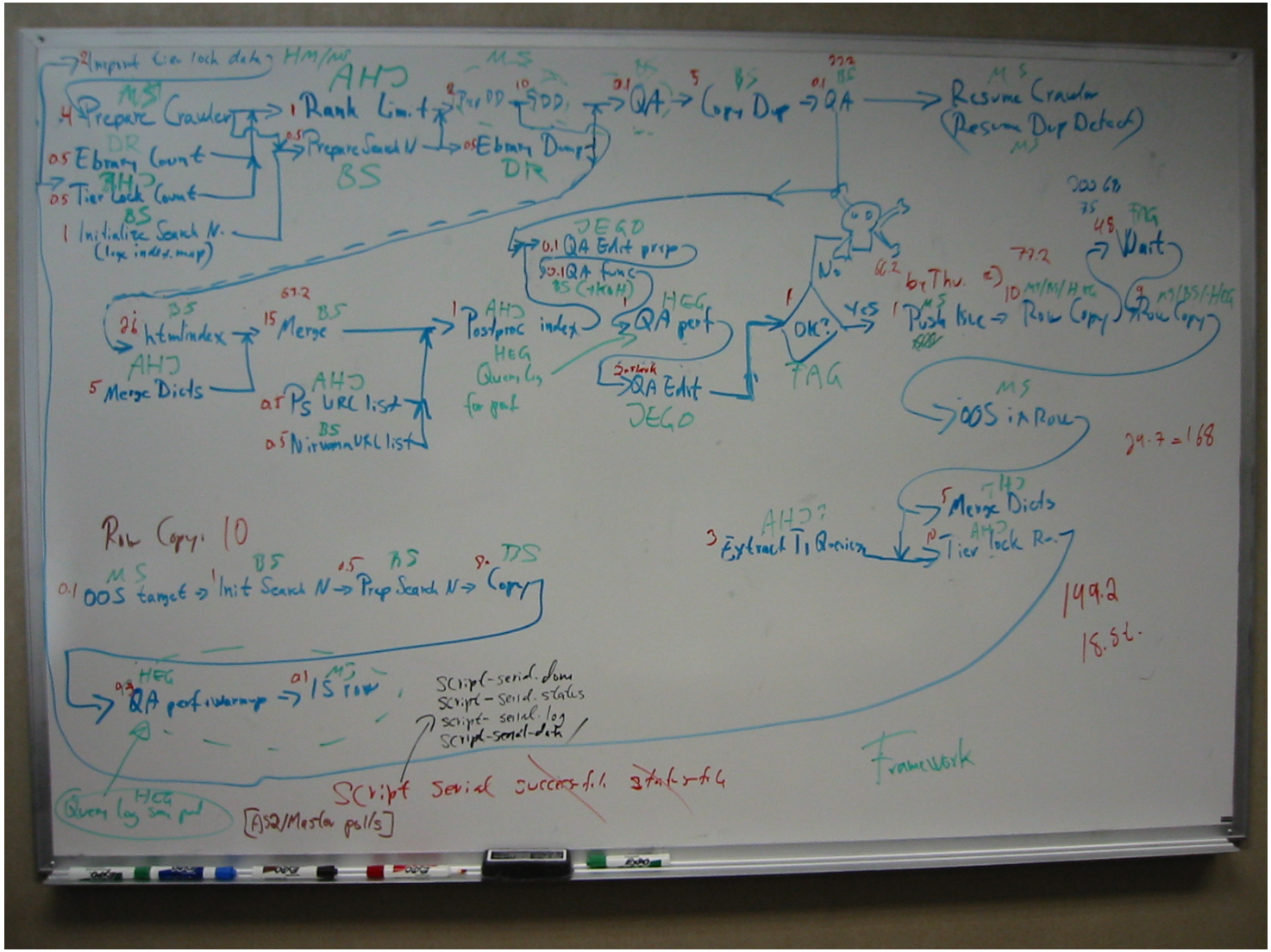
Processing The Content



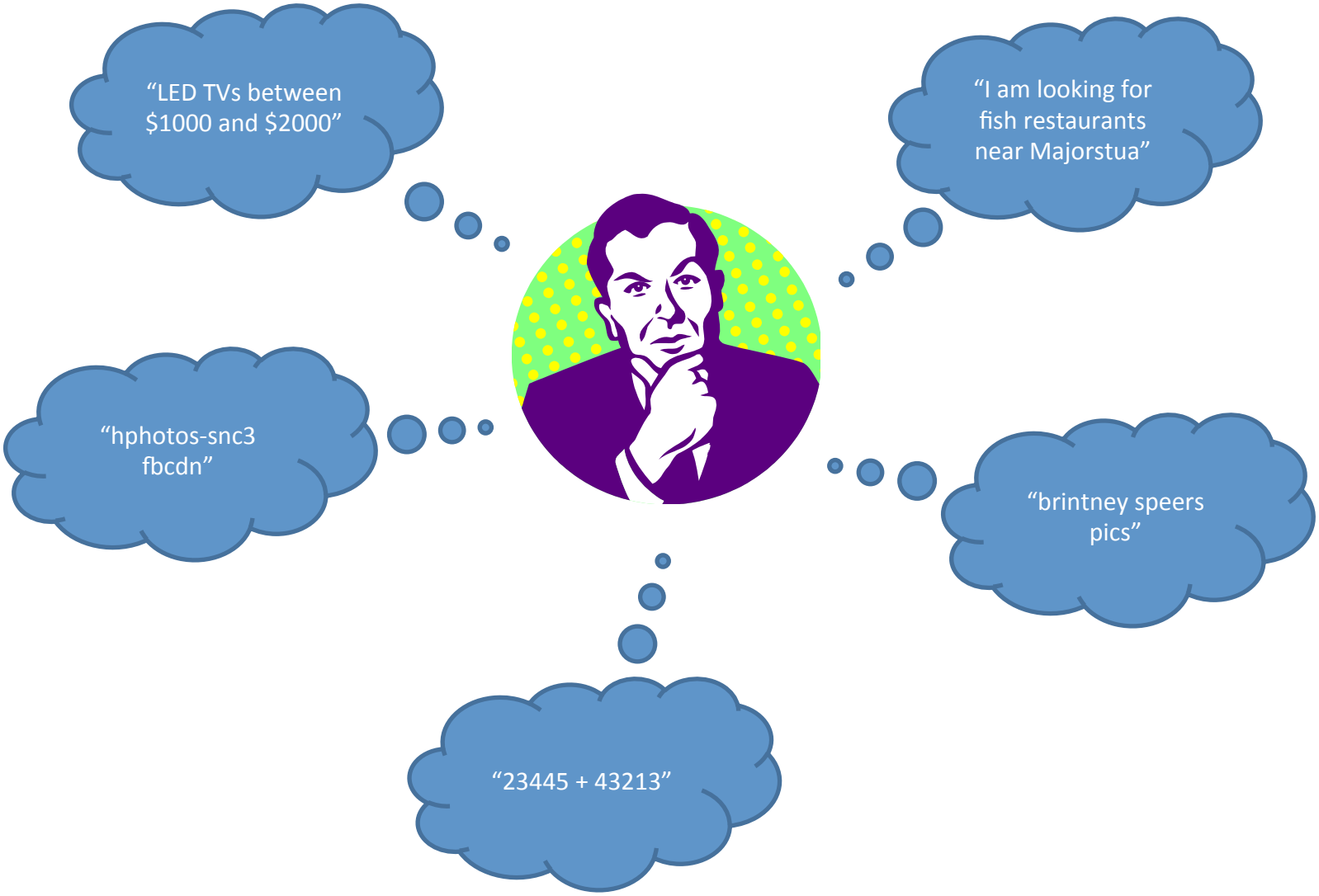
Creating The Index

Word	Document	Position
tea	4	22
	4	32
	4	76
	8	3
teacart	8	7
teach	2	102
	2	233
	8	77
teacher	2	57

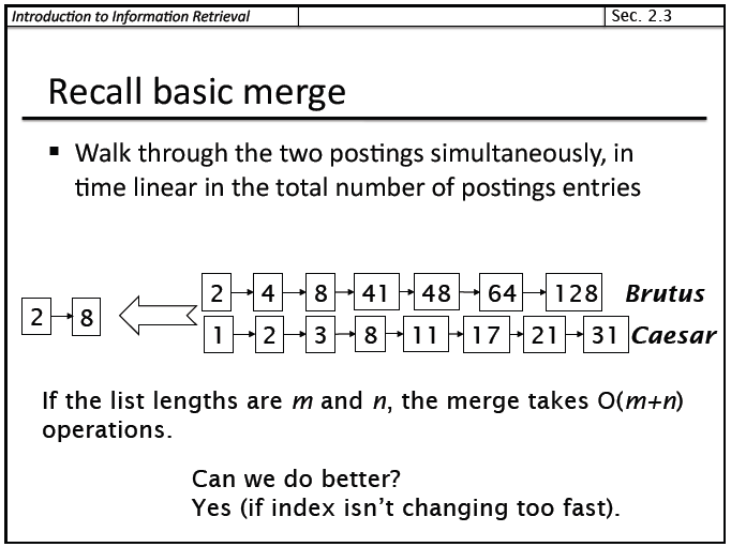
Deploying The Index

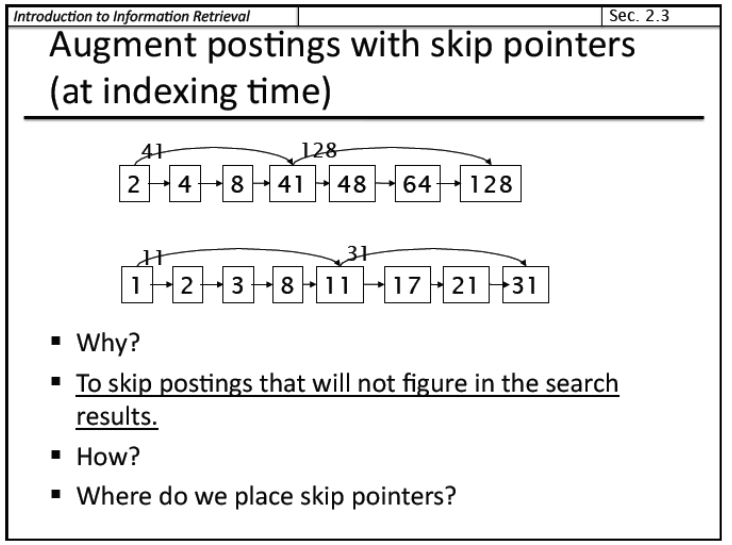


Processing The Query



Searching The Content





http://www.stanford.edu/class/cs276/handouts/lecture2-dictionary.pdf

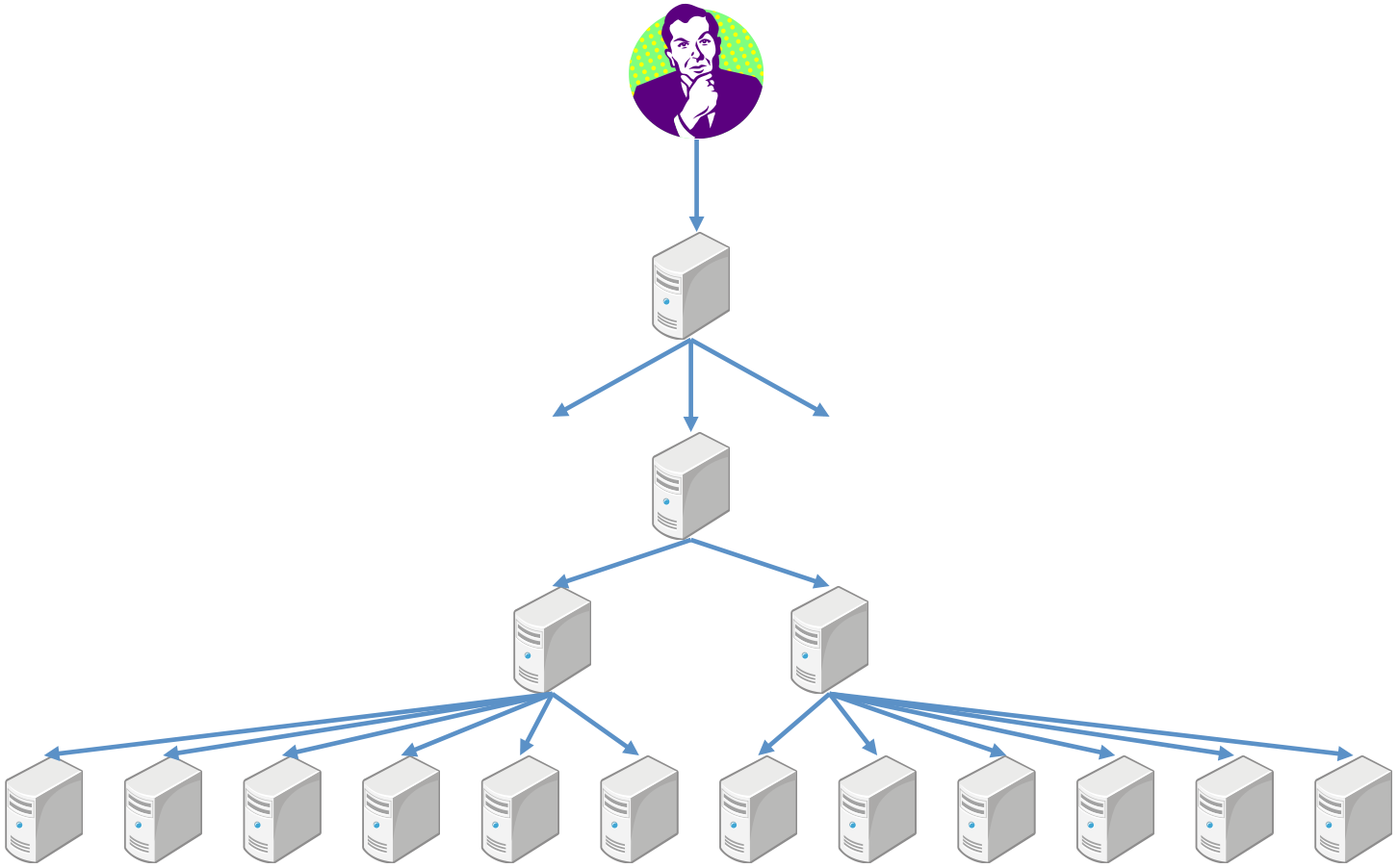
Assess relevancy as we go along

Searching The Content

Federation
Query processing
Result processing

Dispatching
Merging

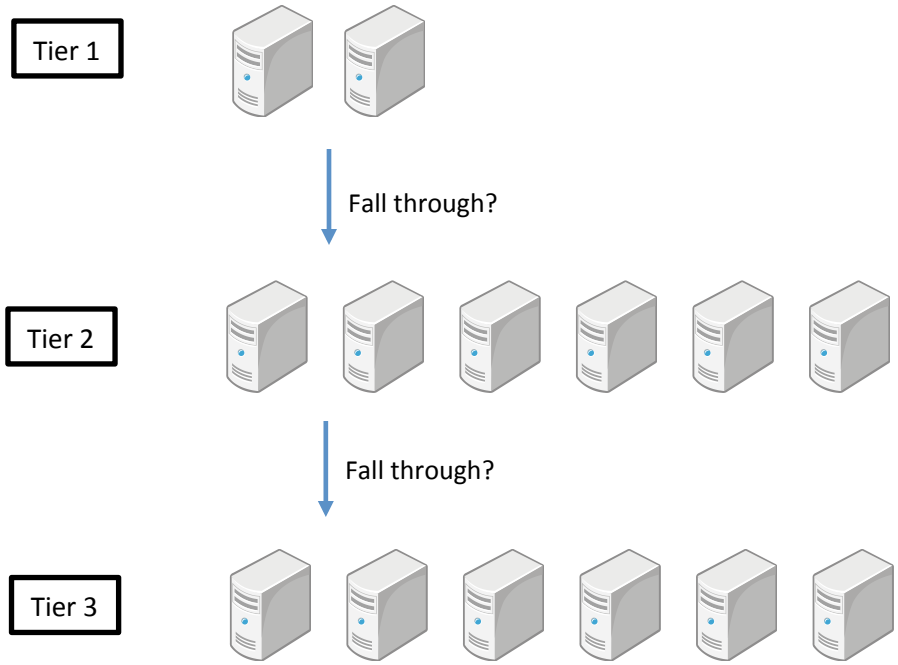
Searching
Caption generation



“Divide and conquer”

Searching The Content

Tiering

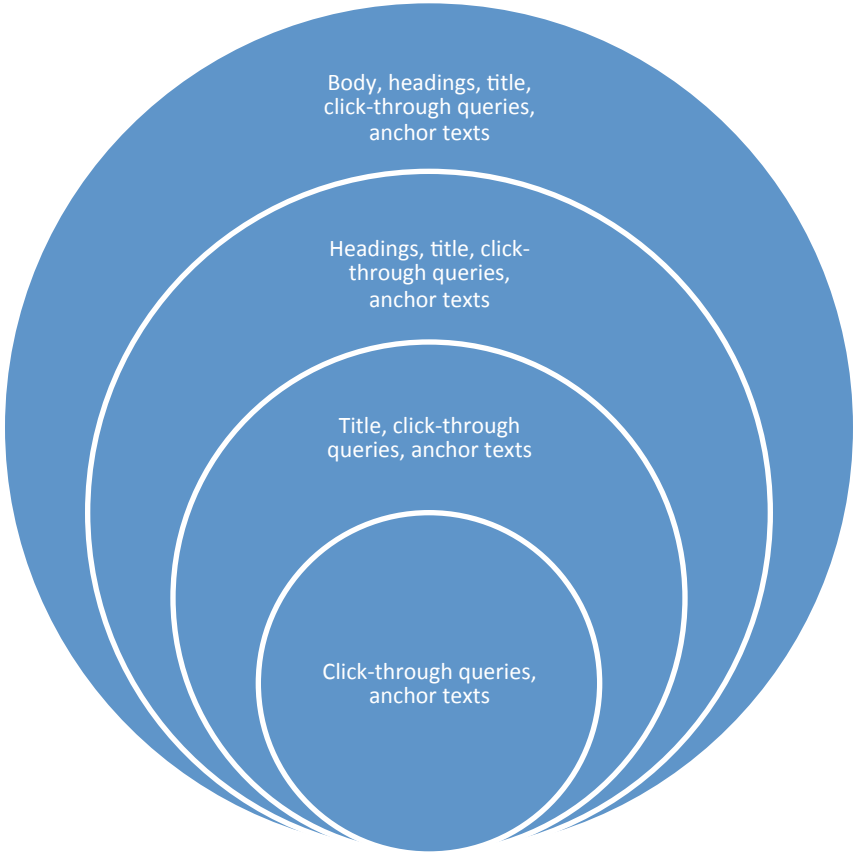


- Organize the search nodes in a row into multiple tiers
- Top tier nodes may have fewer documents and run on better hardware
- Keep the good stuff in the top tiers
- Only fall through to the lower tiers if not enough good hits are not found in the top tiers
- Analyze query logs to decide which documents that belong in which tiers

“All search nodes are equal, but some are more equal than others”

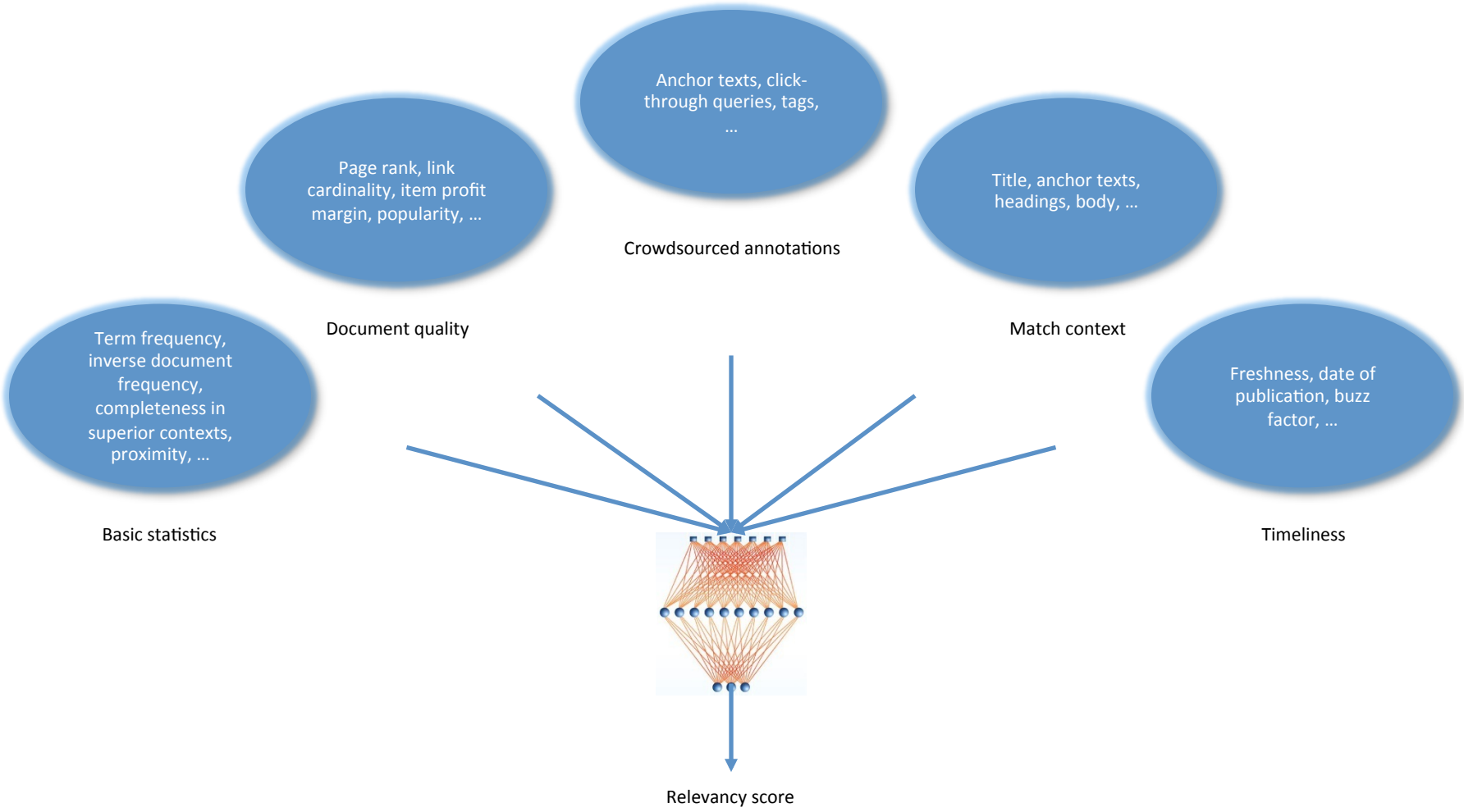
Searching The Content

Context Drilling



“If the result set is too large, only consider the superior contexts”

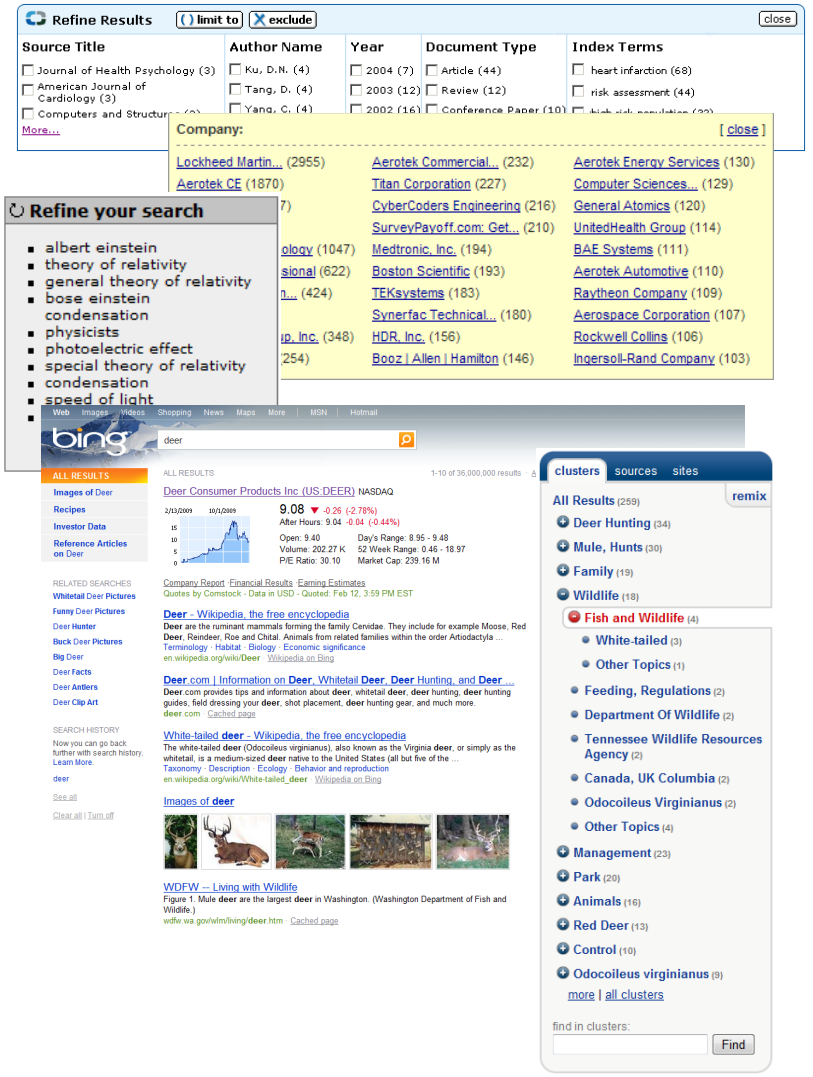
Relevancy



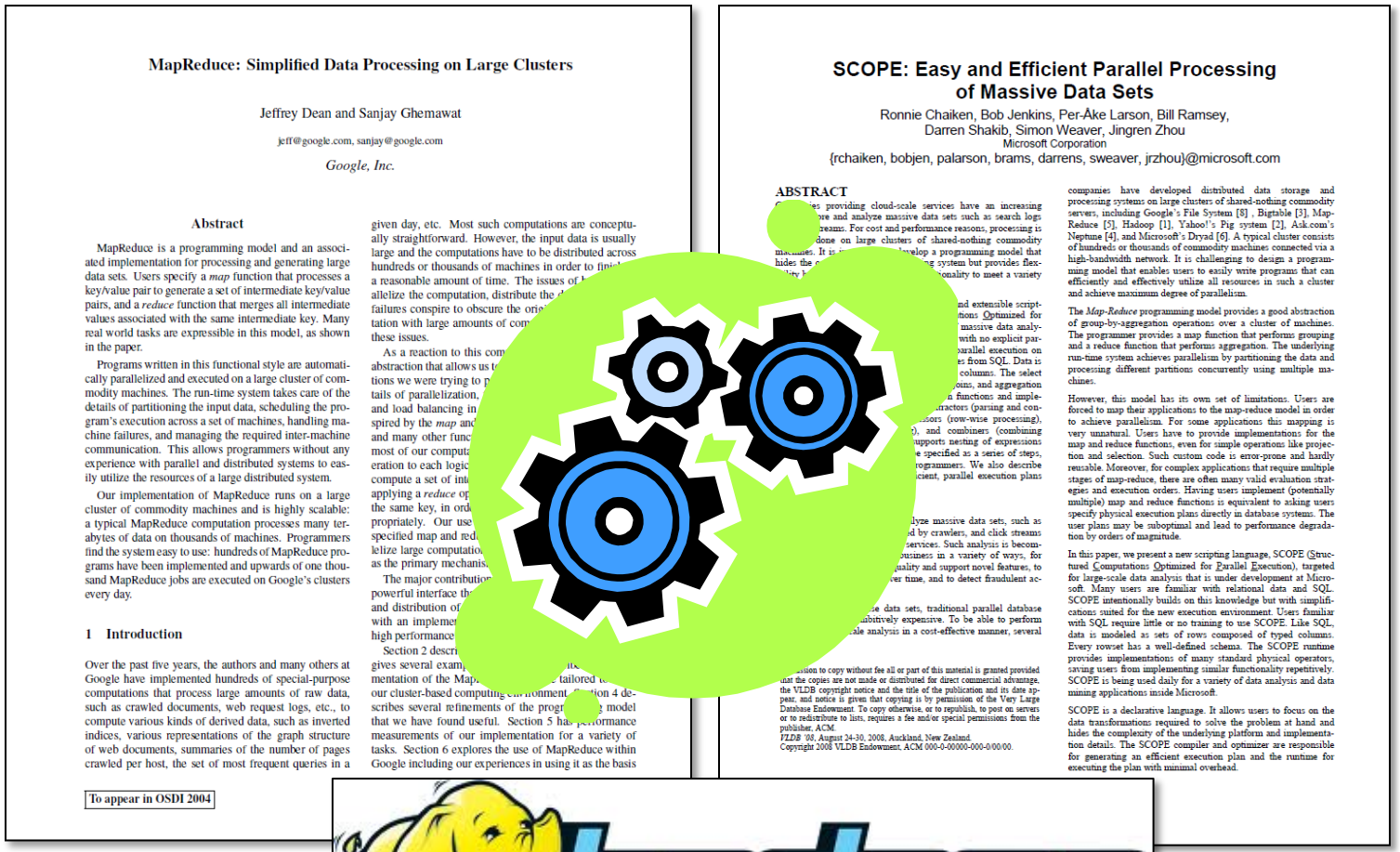
“Maximize the normalized discounted cumulative gain (NDCG)”

Processing The Results

- **Faceted browsing**
  - What are the distributions of data across the various document fields?
  - "Local" versus "global" meta data
- **Result arbitration**
  - Which results from which sources should be displayed in a federation setting?
  - How should the SERP layout be rendered?
- **Unsupervised clustering**
  - Can we automatically organize the results set by grouping similar items together?
- **Last-minute security trimming**
  - Does the user still have access to each result?



Data Mining



MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

Our implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable: a typical MapReduce computation processes many terabytes of data on thousands of machines. Programmers find the system easy to use: hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on Google's clusters every day.

1 Introduction

Over the past five years, the authors and many others at Google have implemented hundreds of special-purpose computations that process large amounts of raw data, such as crawled documents, web request logs, etc., to compute various kinds of derived data, such as inverted indices, various representations of the graph structure of web documents, summaries of the number of pages crawled per host, the set of most frequent queries in a given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this com… abstraction that allows us t… tions we were trying to p… tails of parallelization, … and load balancing in… spired by the *map* and… and many other func… most of our computa… eration to each logic… compute a set of int… applying a *reduce* op… the same key, in ord… propriately. Our use… specified map and red… lelize large computatio… as the primary mechanis…

The major contribution… powerful interface th… and distribution of… with an impleme… high performance …

Section 2 descri… gives several exam… mentation of the Map… tailored t… our cluster-based computing environment. Section 4 describes several refinements of the progr… model that we have found useful. Section 5 has performance measurements of our implementation for a variety of tasks. Section 6 explores the use of MapReduce within Google including our experiences in using it as the basis

To appear in OSDI 2004

SCOPE: Easy and Efficient Parallel Processing of Massive Data Sets

Ronnie Chaiken, Bob Jenkins, Per-Åke Larson, Bill Ramsey, Darren Shakib, Simon Weaver, Jingren Zhou

Microsoft Corporation

{rchaiken, bobjen, palarson, brams, darrens, sweaver, jrzhou}@microsoft.com

ABSTRACT

…es providing cloud-scale services have an increasing …re and analyze massive data sets such as search logs …reams. For cost and performance reasons, processing is …done on large clusters of shared-nothing commodity …develop a programming model that …g system but provides flex… …onality to meet a variety …

…nd extensible script… …tions Optimized for … massive data analy… … with no explicit par… …arallel execution on …es from SQL. Data is … columns. The select …oins, and aggregation … functions and imple… …ractors (parsing and con… …ssors (row-wise processing), …), and combiners (combining …supports nesting of expressions …e specified as a series of steps, …rogrammers. We also describe …cient, parallel execution plans …

…lyze massive data sets, such as …d by crawlers, and click streams …services. Such analysis is becom… …usiness in a variety of ways, for …uality and support novel features, to …er time, and to detect fraudulent ac… …

…e data sets, traditional parallel database …ibitively expensive. To be able to perform …ale analysis in a cost-effective manner, several …

…ission to copy without fee all or part of this material is granted provided …hat the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Database Endowment. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires a fee and/or special permissions from the publisher, ACM.

VLDB '08, August 24-30, 2008, Auckland, New Zealand

Copyright 2008 VLDB Endowment, ACM 000-0-00000-000-0/00/00.

companies have developed distributed data storage and processing systems on large clusters of shared-nothing commodity servers, including Google's File System [8] , Bigtable [3], MapReduce [5], Hadoop [1], Yahoo!'s Pig system [2], Ask.com's Neptune [4], and Microsoft's Dryad [6]. A typical cluster consists of hundreds or thousands of commodity machines connected via a high-bandwidth network. It is challenging to design a programming model that enables users to easily write programs that can efficiently and effectively utilize all resources in such a cluster and achieve maximum degree of parallelism.

The *Map-Reduce* programming model provides a good abstraction of group-by-aggregation operations over a cluster of machines. The programmer provides a map function that performs grouping and a reduce function that performs aggregation. The underlying run-time system achieves parallelism by partitioning the data and processing different partitions concurrently using multiple machines.

However, this model has its own set of limitations. Users are forced to map their applications to the map-reduce model in order to achieve parallelism. For some applications this mapping is very unnatural. Users have to provide implementations for the map and reduce functions, even for simple operations like projection and selection. Such custom code is error-prone and hardly reusable. Moreover, for complex applications that require multiple stages of map-reduce, there are often many valid evaluation strategies and execution orders. Having users implement (potentially multiple) map and reduce functions is equivalent to asking users specify physical execution plans directly in database systems. The user plans may be suboptimal and lead to performance degradation by orders of magnitude.

In this paper, we present a new scripting language, SCOPE (Structured Computations Optimized for Parallel Execution), targeted for large-scale data analysis that is under development at Microsoft. Many users are familiar with relational data and SQL. SCOPE intentionally builds on this knowledge but with simplifications suited for the new execution environment. Users familiar with SQL require little or no training to use SCOPE. Like SQL, data is modeled as sets of rows composed of typed columns. Every rowset has a well-defined schema. The SCOPE runtime provides implementations of many standard physical operators, saving users from implementing similar functionality repetitively. SCOPE is being used daily for a variety of data analysis and data mining applications inside Microsoft.

SCOPE is a declarative language. It allows users to focus on the data transformations required to solve the problem at hand and hides the complexity of the underlying platform and implementation details. The SCOPE compiler and optimizer are responsible for generating an efficient execution plan and the runtime for executing the plan with minimal overhead.

hadoop

Applications

Spellchecking

488941 britney spears
40134 brittany spears
36315 brittney spears
24342 britany spears
7331 britny spears
6633 briteny spears
2696 britteny spears
1807 briney spears
1635 brittny spears
1479 brintey spears
1479 britanny spears
1338 britiny spears
1211 britnet spears
1096 britiney spears
991 britaney spears
991 britnay spears
811 brithney spears
811 brtiney spears
664 birtney spears
664 brintney spears
664 briteney spears
601 bitney spears
601 brinty spears
544 brittaney spears
544 brittnay spears
364 britey spears
364 brittiny spears
329 brtney spears
269 bretney spears
269 britneys spears
244 britne spears
244 brytney spears
220 breatney spears
220 britiany spears
199 britnney spears
163 britnry spears
147 breatny spears
147 brittiney spears
147 britty spears
147 brotney spears
147 brutney spears
133 britteney spears
133 briyney spears
121 bittany spears
121 bridney spears
121 britainy spears
121 britmey spears
109 brietney spears
109 brithny spears
109 britni spears
109 brittant spears
98 bittney spears
98 brithey spears
98 brittiany spears
98 btitney spears
89 brietny spears
89 brinety spears
89 brintny spears
89 britnie spears
89 brittey spears

29 britent spears
29 brittnany spears
29 britttany spears
29 btiney spears
26 birttney spears
26 breitney spears
26 brinity spears
26 britenay spears
26 britneyt spears
26 brittan spears
26 brittne spears
26 btittany spears
24 beitney spears
24 birteny spears
24 brightney spears
24 brintiny spears
24 britanty spears
24 britenny spears
24 britini spears
24 britnwy spears
24 brittni spears
24 brittnie spears
21 biritney spears
21 birtany spears
21 biteny spears
21 bratney spears
21 britani spears
21 britanie spears
21 briteany spears
21 brittay spears
21 brittinay spears
21 brtany spears
21 brtiany spears
19 birney spears
19 brirtney spears
19 britnaey spears
19 britnee spears
19 britony spears
19 brittanty spears
19 britttney spears
17 birtny spears
17 brieny spears
17 brintty spears
17 brithy spears
17 brittanie spears
15 brinney spears
15 briten spears
15 briterney spears
15 britheny spears
15 britneny spears
15 brittamy spears
15 brittmey spears
15 brytnei spears
15 btirney spears
15 rittney spears
14 brinet spears
14 britneyy spears
14 britten spears
12 beritney spears
12 bretiny spears

9 brinttany spears
9 britanay spears
9 britinany spears
9 britn spears
9 britnew spears
9 britneyn spears
9 britrney spears
9 brtiny spears
9 brtittney spears
9 brtny spears
9 brytny spears
9 rbitney spears
8 birtiny spears
8 bithney spears
8 brattany spears
8 breitny spears
8 breteny spears
8 brightny spears
8 brintay spears
8 brinttey spears
8 briotney spears
8 britanys spears
8 britley spears
8 britneyb spears
8 britnrey spears
8 britnty spears
8 brittner spears
8 brottany spears
7 baritney spears
7 birntey spears
7 biteney spears
7 bitiny spears
7 breateny spears
7 brianty spears
7 brintye spears
7 britianny spears
7 britly spears
7 britnej spears
7 britneyu spears
7 britniey spears
7 britnnay spears
7 brittian spears
7 briyny spears
7 brrittany spears
7 brttiney spears
7 btiteny spears
7 btrittany spears
6 beritny spears
6 bhritney spears
6 birthney spears
6 breathney spears
6 breaty spears
6 bretany spears
6 briatany spears
6 brint spears
6 britenney spears
6 britian spears
6 britinty spears
6 brititney spears
6 britnsy spears

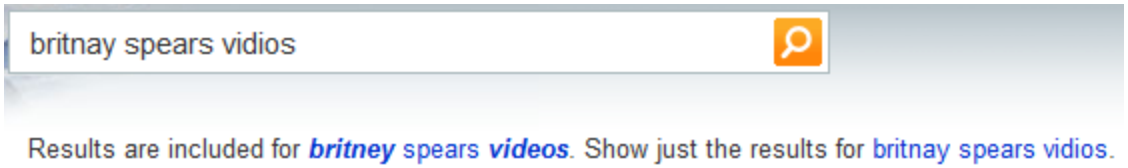
5 brney spears
5 broitney spears
5 brotny spears
5 bruteny spears
5 btiyney spears
5 btrittney spears
5 gritney spears
5 spritney spears
4 bittny spears
4 bnritney spears
4 brandy spears
4 brbritney spears
4 breatiny spears
4 breetney spears
4 bretiney spears
4 brfitney spears
4 briattany spears
4 brieteny spears
4 briety spears
4 briitny spears
4 briittany spears
4 brinie spears
4 brinteney spears
4 brintne spears
4 britaby spears
4 britaey spears
4 britainey spears
4 britinie spears
4 britinney spears
4 britmney spears
4 britnear spears
4 britnel spears
4 britneuy spears
4 britnewy spears
4 britnmey spears
4 brittaby spears
4 brittery spears
4 britthey spears
4 brittnaey spears
4 brittnat spears
4 brittneny spears
4 brittnye spears
4 brittteny spears
4 briutney spears
4 briyeny spears
4 brnity spears
4 brtteny spears
4 brttiany spears
4 bryney spears
4 brythney spears
4 brytne spears
4 brytni spears
4 brytnie spears
4 bvritney spears
4 dritney spears
4 priteny spears
3 beittany spears
3 bichney spears
3 biritny spears
3 birnety spears

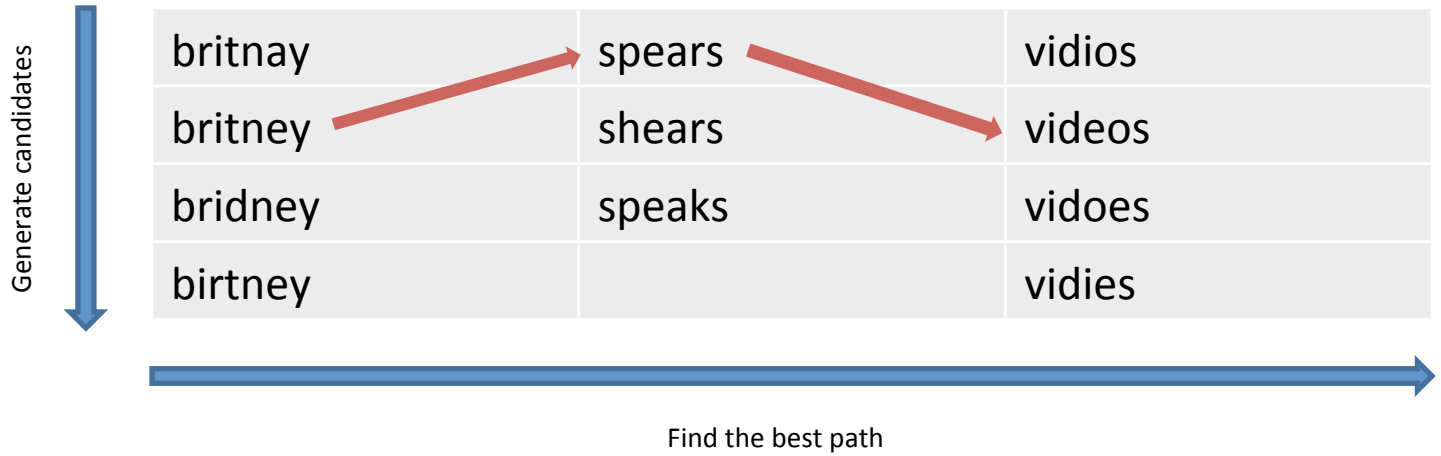
3 britiy spears
3 britmeny spears
3 britneeey spears
3 britnehy spears
3 britnely spears
3 britnesy spears
3 britnetty spears
3 britnex spears
3 britneyxxx spears
3 britnity spears
3 britntey spears
3 britnyey spears
3 britterny spears
3 brittneey spears
3 brittnney spears
3 brittnyey spears
3 brityen spears
3 briytney spears
3 brltney spears
3 broteny spears
3 brtaney spears
3 brtiiany spears
3 brtinay spears
3 brtinney spears
3 brtitany spears
3 brtiteny spears
3 brtnet spears
3 brytiny spears
3 btney spears
3 drittney spears
3 pretney spears
3 rbritney spears
2 barittany spears
2 bbbritney spears
2 bbitney spears
2 bbritny spears
2 bbrittany spears
2 beitany spears
2 beitny spears
2 bertney spears
2 bertny spears
2 betney spears
2 betny spears
2 bhriney spears
2 biney spears
2 bintey spears
2 biretny spears
2 biritany spears
2 birittany spears
2 birittny spears
2 birnty spears
2 birtey spears
2 birtheny spears
2 birtieny spears
2 birtnay spears
2 birtnet spears
2 bitnet spears
2 bitttany spears
2 bnrittany spears
2 bntney spears

2 brirreny spears
2 brirtany spears
2 brirttany spears
2 brirttney spears
2 britain spears
2 britane spears
2 britaneny spears
2 britania spears
2 britann spears
2 britanna spears
2 britannie spears
2 britannt spears
2 britannu spears
2 britanyl spears
2 britanyt spears
2 briteeny spears
2 britenany spears
2 britenet spears
2 briteniy spears
2 britenys spears
2 britianey spears
2 britin spears
2 britinary spears
2 britmy spears
2 britnaney spears
2 britnat spears
2 britnbey spears
2 britndy spears
2 britneh spears
2 britneney spears
2 britney6 spears
2 britneye spears
2 britneyh spears
2 britneym spears
2 britneyyy spears
2 britnhey spears
2 britnjey spears
2 britnne spears
2 britnu spears
2 britoney spears
2 britrany spears
2 britreny spears
2 britry spears
2 britsany spears
2 brittanay spears
2 brittang spears
2 brittans spears
2 brittanyh spears
2 brittanyn spears
2 brittany's spears
2 brittanyt spears
2 brittanyy spears
2 brittary spears
2 brittenie spears
2 brittenty spears
2 brittinney spears
2 brittley spears
2 brittn spears
2 brittnery spears
2 brittnety spears

http://www.google.com/jobs/britney.html

Spellchecking





1. Generate a set of candidates per query term using approximate matching techniques. Score each candidate according to, e.g., “distance” from the query term and usage frequency.

2. Find the best path in the lattice using the Viterbi algorithm. Use, e.g., candidate scores and bigram statistics to guide the search.

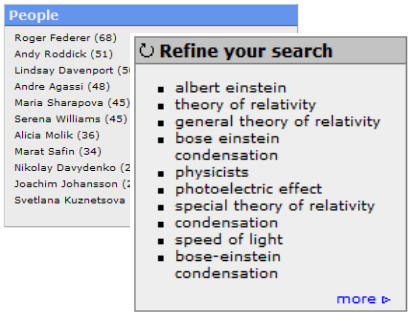
Entity Extraction

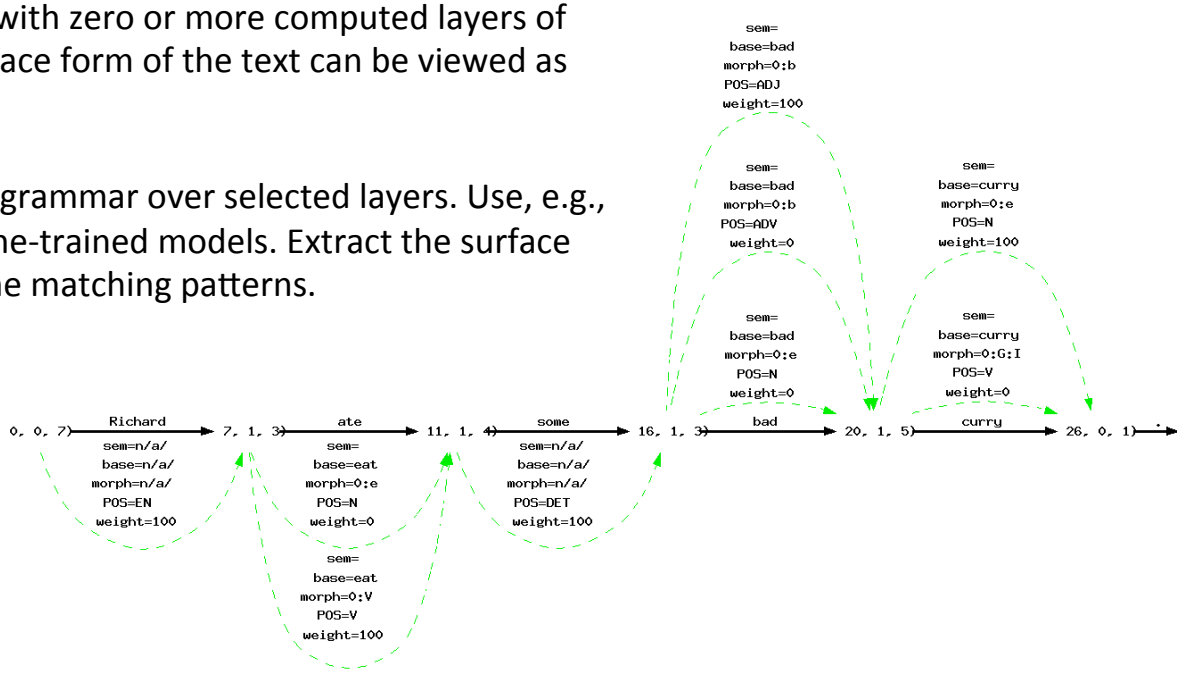
Levels of abstraction

…	…	…	…	…
MAN				FOOD
N/proper	V/past/eat	DET	ADJ	N/singular
Richard	ate	some	bad	curry

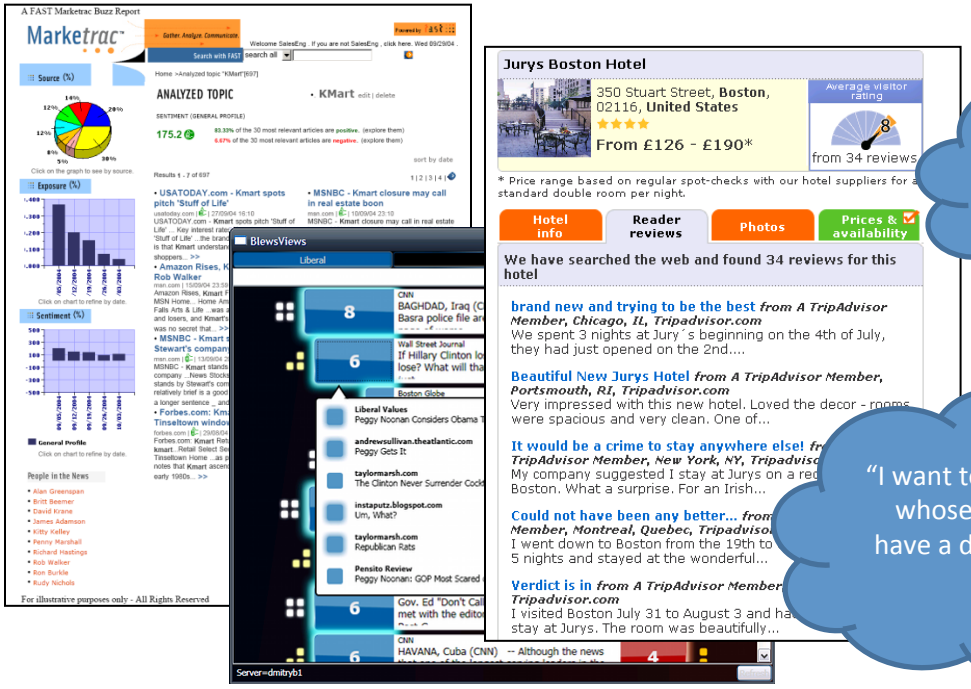
1. Logically annotate the text with zero or more computed layers of meta data. The original surface form of the text can be viewed as trivial meta data.

2. Apply a pattern matcher or grammar over selected layers. Use, e.g., handcrafted rules or machine-trained models. Extract the surface forms that correspond to the matching patterns.





Sentiment Analysis



"What is the current perception of my brand?"

"I want to stay at a hotel whose user reviews have a definite positive tone."

"What are the most emotionally charged issues in American politics right now?"

http://research.microsoft.com/en-us/projects/blews/

1. To construct a sentiment vocabulary, start by defining a small seed set of known polar opposites.

2. Expand the vocabulary by, e.g., looking at the context around the seeds in a training corpus.

3. Use the expanded vocabulary to build a classifier. Apply special heuristics to take care of, e.g., negations and irony.

Contextual Search

"Sentences where someone says something positive about Adidas."

xml:sentence:("adidas" and sentiment:@degree:>0)

"Dates and locations related to D-Day."

xml:sentence:("d-day" and (scope(date) or scope(location)))

"Paragraphs that discuss a company merger or acquisition."

xml:paragraph:(string("merger", linguistics="on") and scope(company) and scope(price))

"Paragraphs that contain quotations by Alan Greenspan, where he mentions a monetary amount."

xml:paragraph:quotation:(@speaker:"greenspan" and scope(price))

"Sentences where the acronym MIT is defined."

xml:sentence:acronym:(@base:"mit" and scope(@definition))

Persons that appear in **documents** that contain the word *{soccer}*

person@base
Jack Nicklaus (~10.0%)
Fred Davis (~10.0%)
Billie Jean King (~8.0%)
Richard Nixon (~8.0%)
John Wayne (~7.0%)
Margaret Smith (~7.0%)
Joe Frazier (~7.0%)
Irina Rodnina (~7.0%)
Mao Zedong (~6.0%)
Gordie Howe (~6.0%)
Richard M. Nixon (~6.0%)
More...

Example from Wikipedia

person@base
Diego Maradona (~4.0%)
David Beckham (~4.0%)
Alan Shearer (~3.0%)
Michelle Akers (~3.0%)
Mia Hamm (~3.0%)
Eric Wynalda (~3.0%)
Freddy Adu (~3.0%)
Michel Platini (~2.0%)
Stanley Matthews (~2.0%)
Oliver Neuville (~2.0%)
Bobby Moore (~2.0%)
More...

Persons that appear in **paragraphs** that contain the word *{soccer}*

Contextual Search

1. During content processing, identify structural and semantic regions of interest. Mark them up in context, possibly decorated with meta data.

D-Day is the name given to the landing of 160,000 Allied troops in Normandy, France, on June 6, 1944. The success of the invasion of Normandy was really the beginning of the end for Nazi Germany. The invasion, also called..

<sentence>D-Day is the name given to the landing of 160,000 Allied troops in <location country="france">Normandy</location>, <location type="country">France</location>, on <date base="1944-06-06">June 6, 1944</date>.</sentence><sentence>The success of the invasion of <location country="France">Normandy</location> was really the beginning of the end for Nazi Germany.</sentence><sentence>The invasion, also called..

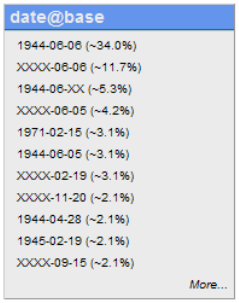
2. Make all the marked-up data fully searchable in a way that preserves context and where retrieval can be constrained on both structure and content. Possibly translate natural language queries into suitable system queries.

when was d-day?

xml:sentence:(and(string("d-day"), scope(date))

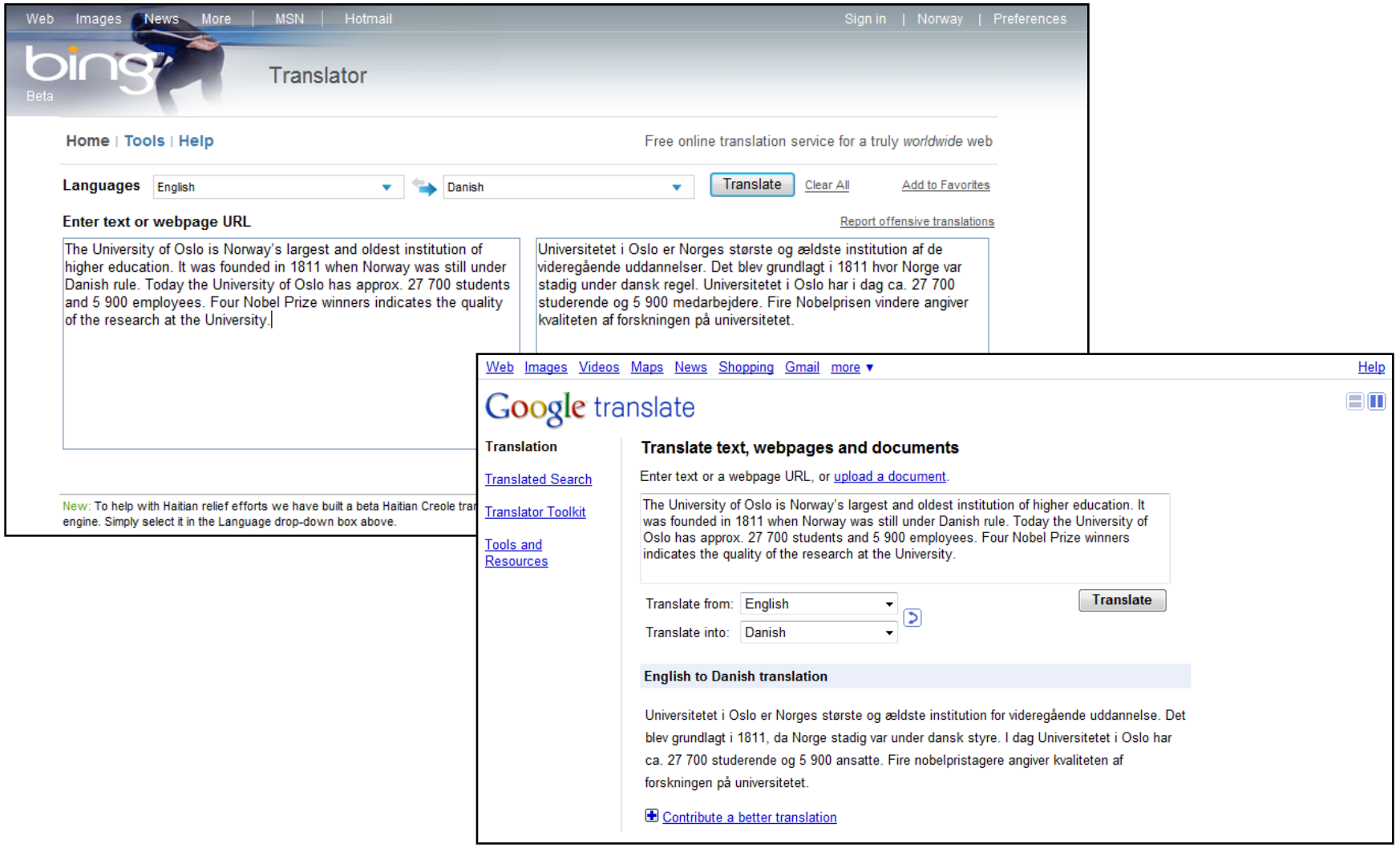
<matches>
<match id="43423" score="0.927">
<sentence>D-Day is the name given to the landing of 160,000 Allied troops in <location country="france">Normandy</location>, <location type="country">France</location>, on <date base="1944-06-06">June 6, 1944</date>.</sentence>
</match>
<match id="12245" score="0.831">
<sentence>The D-Day operation was pushed back to <date base="XXXX-06-06">June 6th</date>.</sentence>
</match>
<match id="54599" score="0.792">
<sentence>Following are key facts about D-Day, the Allied invasion of <location country="france">Normandy</location> on <date base="1944-06-06">June 6, 1944</date>.</sentence>
</match>
</matches>

3. Aggregate data over the matching fragments and enable faceted browsing on a contextual level.

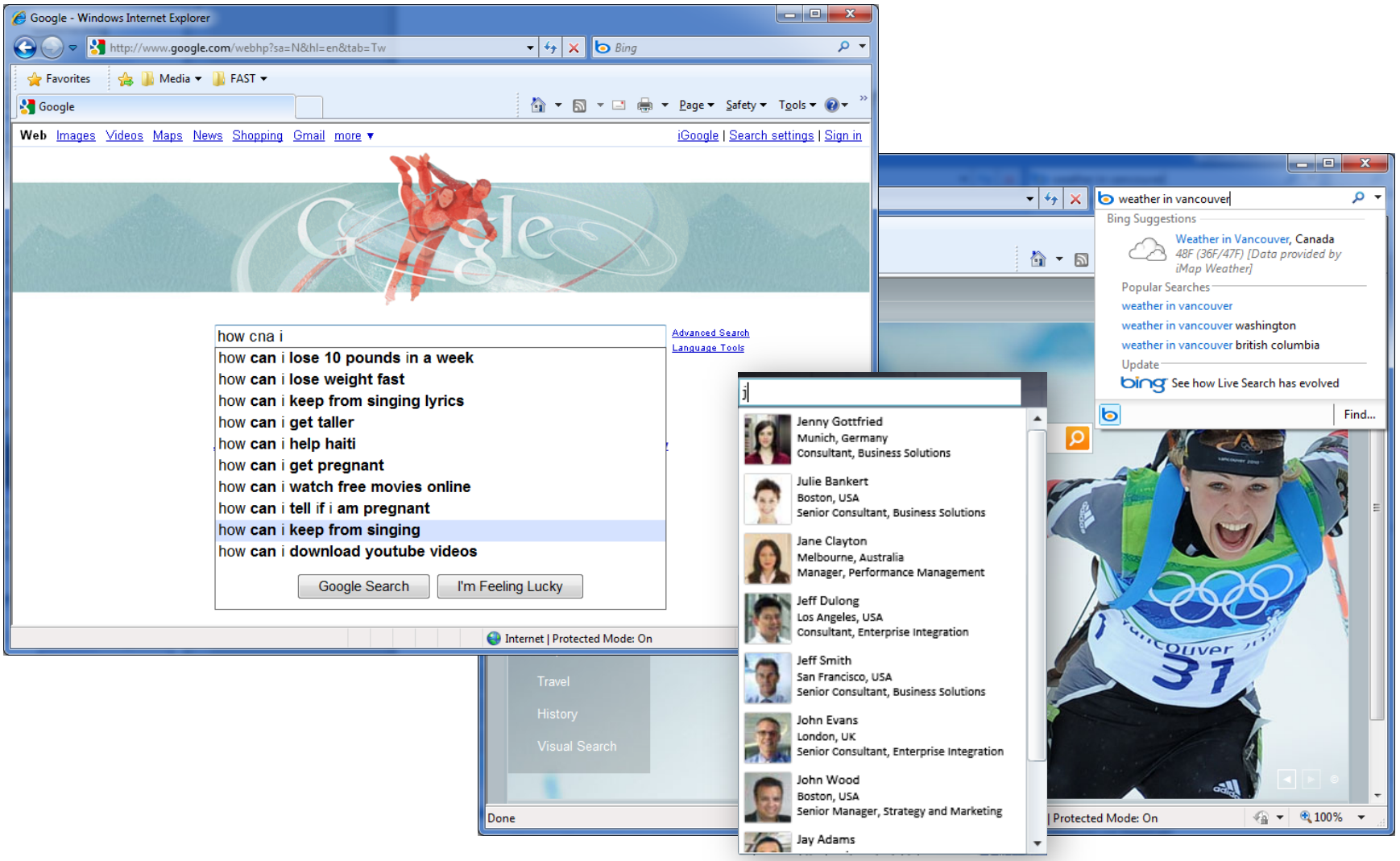


Example from Wikipedia

Machine Translation

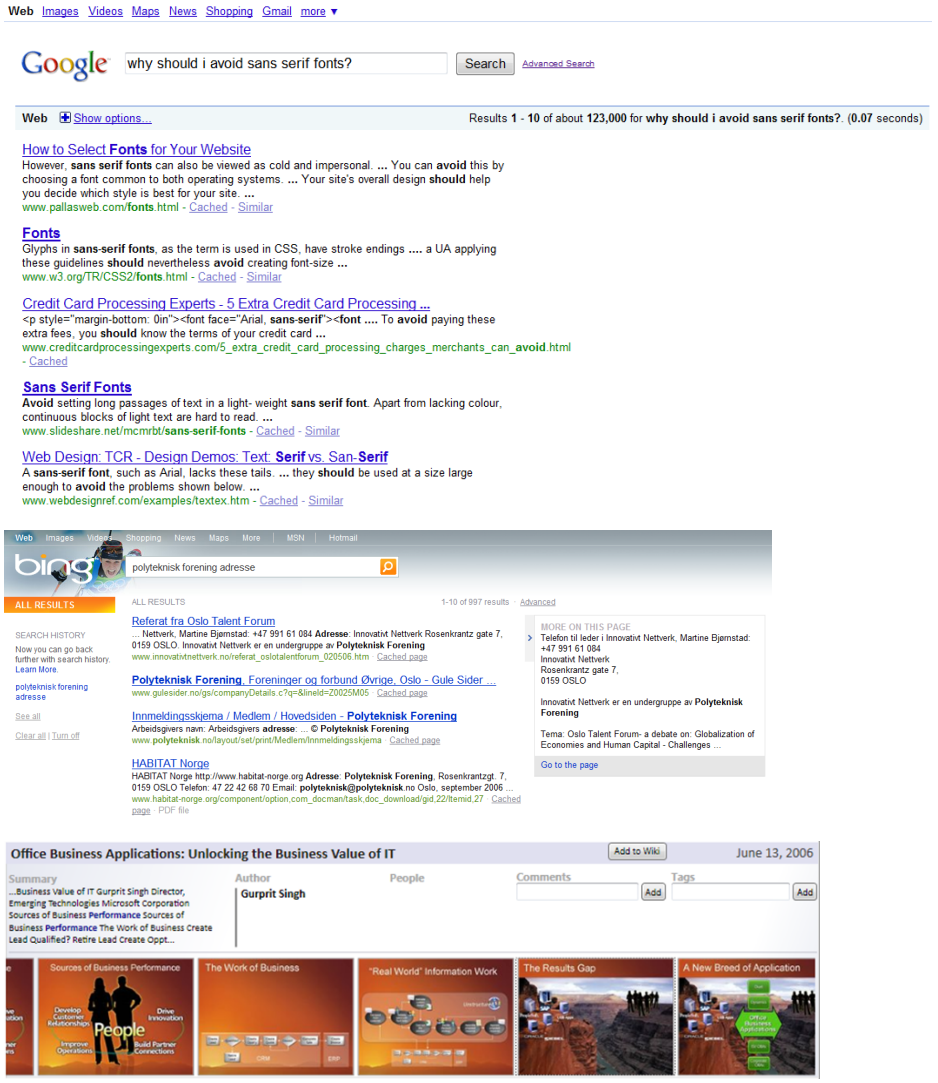


Query Completion



Caption Generation

- **Intra-document search**
  - Locate and rank relevant document fragments
  - But do it fast!
- **Perceived relevancy**
  - First impressions count
  - Can make or break a service
- **Trends towards richer captions**
  - Format-specific interactivity
  - Actionable elements



The Future