

The Continuing Metamorphosis Of the Web

Dr. Alfred Z. Spector
VP, Research and Special Initiatives
Google, Inc.
WWW 2009, Madrid, 24 April 2009



Abstract

The Continuing Metamorphosis of the Web

The invention of HTML and HTTP catalyzed a path of enormous innovation that was hard to foresee in the early 1990's. The Web's continuing metamorphosis has led to fantastically increased capabilities and economic value. It has catalyzed the creation of distributed systems orders of magnitude larger than any previously built, new programming and distribution models for computer applications, great advances in the fields of information retrieval, entirely new domains for theoretical computer science, and more.

This greatly enhanced web is changing the entire environment and enabling some early research promises to become a reality for most Internet users. In this presentation, I will discuss such examples, and in particular, what happens when speech, image processing, human language translation, and mobility are woven into all we do. I will also extrapolate from some current research and advanced web technologies to paint a picture of the web five-to-ten years out. This should have implications for the computer science community, as well as the vast community that is leveraging the web for ever greater goals.



Outline

1. Federation, Reach, and Evolution
2. Extraordinary Achievements of Note
3. The Evolutionary Path Forward
4. 3 Major Extraordinary Advances Brewing:
 - A. Totally Transparent Processing
 - B. The Rule of Distributed Computing
 - C. Hybrid, not Artificial, Intelligence
5. Some Research Challenges
6. Conclusions

3



Google's Mission and Google Research

**Organizing the world's information and
Making it universally accessible and useful.**



A research organization optimized for in situ work



Federation, Reach, and Evolution

- The simplicity of the early web standards were genius
 - Federated name space
 - Access (HTTP)
 - Simple data format (HTML)
 - Extensibility!
- Not over-architected in any dimension
- Brilliant omissions (or at least, mostly so 😊)
 - Security
 - Read-write data
 - Semantics
- *Interesting contrast to wide-area file systems work like AFS*

5



A Semi-Random Walk to Extraordinary Achievements

- The virtuous circle
 - Initial simplicity begat data and usage
 - Usage generated more data and transactions ←
 - Data modalities diversified
 - User experience blossomed →
- Architectural limitations were addressed as needed
- *A bottom-up architectural evolution repeatedly favoring local optimization has resulted in truly momentous results.*
 - The virtual Library of Alexandria
 - The search engine
 - The serving of the long tail
 - Vast changes in business models

6



Additional, Architectural Achievements

- Network
 - The Web became the catalyst for the rapid internet build-out
- The High Performance Cluster (old word, “multi-computer”)
 - The federated architecture, perhaps strangely, did not obviate the need for large processing complexes.
 - Some workloads require high throughput, low latency, massive data, albeit, embarrassingly parallel
 - The answer: parallel clusters of $O(10^5)$ CPUs & $O(10^{17})$ storage
 - Note “An order of magnitude is a qualitative difference!”
- Browser
 - With a few plug-ins, the application programming model of choice
 - Perhaps, the key client operating system functionality

7



The Evolutionary Path Forward to New Accomplishments

- Application mix will continue to grow in unpredictable ways:
 - Four areas in flux today: *publishing, education, healthcare and government*
- Systems will evolve: ubiquitous high performance networking, distributed computing, new end-user devices, ...
- Three truly big results brewing:
 - 1. Totally Transparent Processing**
 - 2. Ideal Distributed Computing**
 - 3. Hybrid, Not Artificial, Intelligence**

8



Totally Transparent Processing



Totally Transparent Processing

$$\forall d \in D, \forall l \in L, \forall m \in M, \forall c \in C$$

D: The set of all end-user access devices	L: The set of all human languages	M: The set of all modalities	C: The set of all corpora
Personal Computers	Current languages	Text	The normal web
Phone	Historical languages	Image	The deep web
Media Players/Readers	Other forms of human notation	Audio	Periodicals
Telematics	Possible language specialization	Video	Books
Set-top Boxes	Formal languages	Graphics	Catalogs
Appliances	...	Other sensor-based data	Blogs
Health devices			Geodata
...			Scientific datasets
			Health data
			...

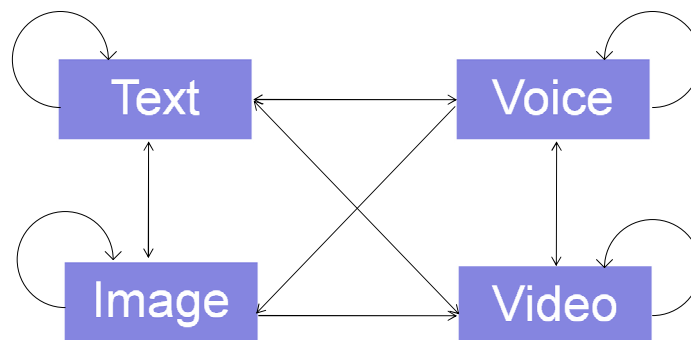


Types of Transparent Processing

- Search, of many forms
 - Navigation and Suggestion
 - Transformational Communication
 - Information Fusion
- Some Google Examples:
- Universal search
 - Voice Search
 - Find Similar, applied to images
 - Google Translate, particularly in mash-ups
 - Combining images and maps
 - Audio transcription
 - Images and 3d models



Fluidity Among the Modalities



Last two arrows are easily conceivable.



The New Frontier of Web Search – Better/Faster Queries

Query	Result Count
real	37.700.000 resultados
real madrid	29.000.000 resultados
real player	1.410.000 resultados
real academia española	12.100.000 resultados
realtek	2.420.000 resultados
real betis	2.710.000 resultados
real sociedad	1.830.000 resultados
real zaragoza	28.000.000 resultados
reale	1.070.000 resultados
real murcia	1.380.000 resultados
real valladolid	1.380.000 resultados

Query completion before: Used a fixed dictionary, e.g., in emacs, bash, T9, etc.

Query suggestion today: Model queries with query logs, serve them dynamically

Technical challenges:

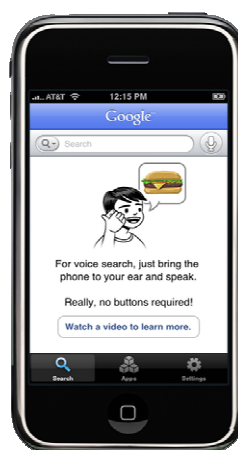
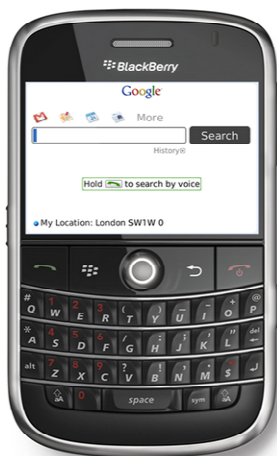
- response-time, coverage, freshness, corpus dependency (YouTube, image, mobile)
- domain dependent: rea -> real madrid good suggestion in Spain
- diversity (danger of popularity), filtering out duplicates, inappropriate results, etc.

Impact: Made possible by scale,

- the richer the query log corpus, the better
- the faster the response time, the better



Voice Search



Challenges and Rationale for Success

Technically this is very challenging:

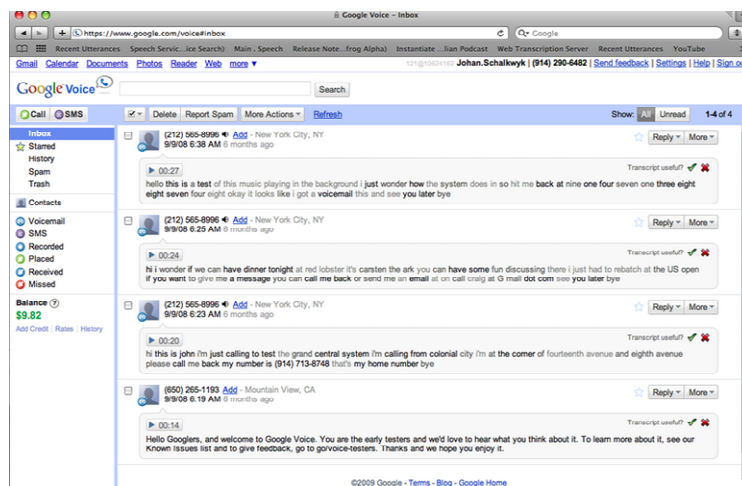
- Huge vocabulary
- Variability in accent
- Background noise

What makes this possible:

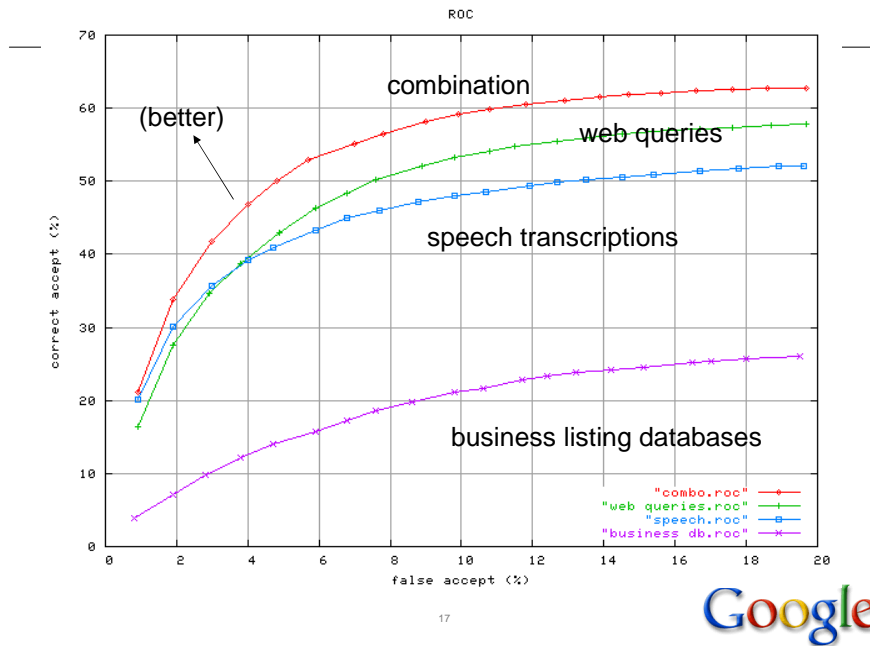
- Scalable technology
- Data inputs: Query logs, voice logs
- Compute power



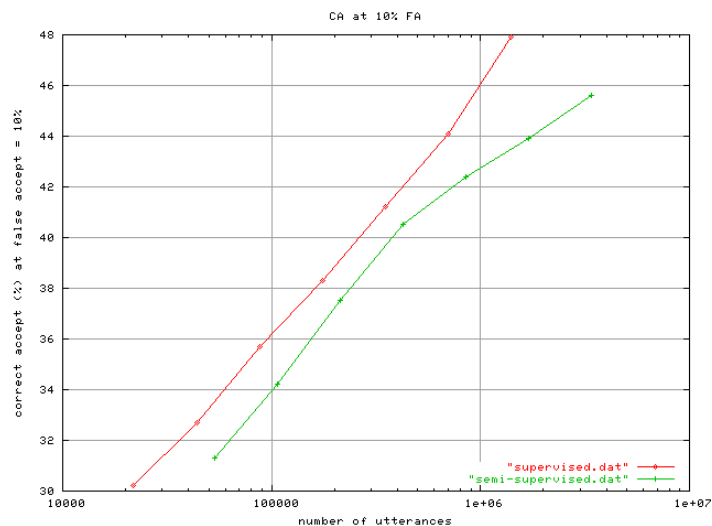
Transcriptions in Google Voice



Lots of data: utterance ROC curves: incl. rejection)



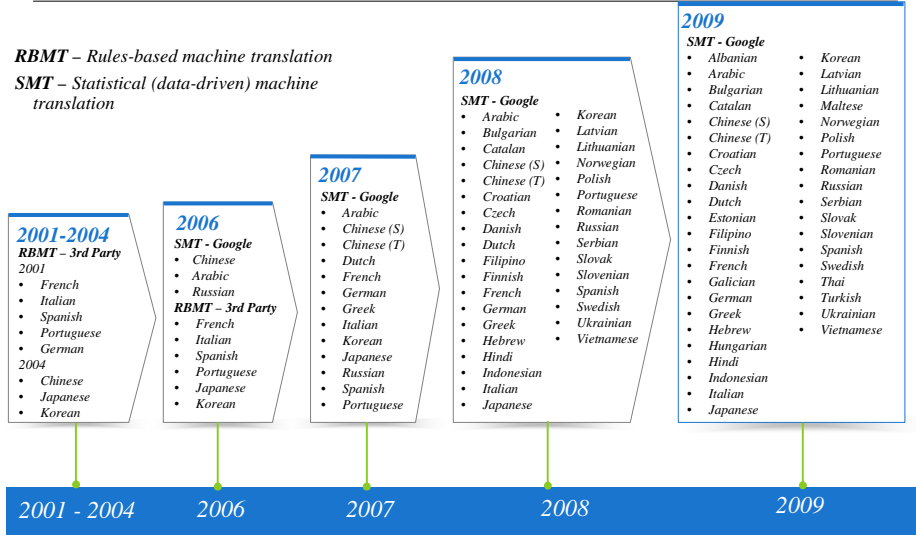
The Benefits of Unsupervised Training



Google Translate

RBMT – Rules-based machine translation

SMT – Statistical (data-driven) machine translation



19



Web Translation

This page was automatically translated from Japanese. [View original web page](#) or [mouse over text to view original language](#). [Back to Translate](#) [Remove frame](#)

価格.com カメラ

Digital SLR Camera

amazon.co.jp
10/13まで
最大90%OFF

Keyword: Digital SLR Camera

Related Features

- Latest! "Digital SLR Camera" choice
- Single-lens reflex camera even more enjoyable! Choose a replacement lens mount
- VAIO type C 83,800円税込

Ranking Digital SLR Camera

Position	Ranking	Ranking attention	Satisfaction Ranking
1st place	DANON DABURUZUMUKITTO EOS Kiss X2 Lows ¥ 83,000	DANON Body EOS 5D Mark II	Olimesa E-420 Body Lows ¥ 39,350

20



Web Translation Feedback

The screenshot shows a Google search result for "Digital SLR Camera". At the top, there is a notice: "This page was automatically translated from Japanese. View original web page or mouse over text to view original language." Below this, the search results are displayed in Japanese. A prominent banner from amazon.co.jp advertises a "最大90%OFF" (Maximum 90% OFF) sale on electronics, valid until 10/13. Below the banner, there is a list of "Manufacturers to choose from" including Canon, Sony, Olympus, Sigma, Nikon, Panasonic, Pentax, MAMIYA, Leica, and Fujifilm. A "Related Features" section highlights "Latest! Digital SLR Camera" and "Single-lens reflex camera even more enjoyable! Choose a replacement lens mount". A "Ranking" section is also visible, showing "Ranking Digital SLR Camera" with columns for Position, Ranking, Ranking attention, and Satisfaction Ranking. A small tooltip is visible over the search results, containing the text: "Original Japanese text: タイプ、容量ごとにお買い得なフラッシュメモリーをご紹介! Suggest a better translation: 買1得capacity of each type of flash memory with us! Contribute".

21



Cross-language search

The screenshot shows the Google Translate interface. At the top, there are tabs for "Text and Web", "Translated Search", "Dictionary", and "Tools". The "Translated Search" tab is selected. Below the tabs, there is a search bar with the text "digital slr camera". Below the search bar, there are two dropdown menus: "My language: English" and "Search pages written in: Japanese". Below these menus is a "Translate and Search" button.

How does this work?

1. Search for [Dubai tours](#) from English to Arabic.
2. We translate your query into "جولات دبي" and find Arabic web page results.
3. Finally, we translate the Arabic web page results back into English for you.

Other things you can search for

- [St. Petersburg restaurants](#) - (Санкт-Петербург рестораны) in Russian pages
- [Beijing apartments](#) - (北京公寓) in Chinese pages

22



Cross-language search

The screenshot shows a Google Translate search interface. The search term is "digital slr camera". The results are translated from Japanese web pages. The interface includes a search bar, language selection (English to Japanese), and a "Translate and Search" button. Below the search bar, there are three columns of results. The first column is titled "English translation" and contains three entries: "Prices .Com - Digital SLR Camera", "Prices .Com - Ranking Digital SLR Camera", and "Introduction to Digital SLR Camera". The second column is titled "Original Japanese" and contains three entries: "価格 .com - デジタル一眼レフカメラ", "価格 .com - デジタル一眼レフカメラ 売れ筋ランキング", and "デジタル一眼レフカメラ入門". Each entry includes a brief description and a link to the source page.

Translate My Page Gadget

The screenshot shows the MTA website (Metropolitan Transportation Authority) with a Google Translate gadget overlaid on the bottom left. The gadget is a small window with a search bar and a language selection dropdown menu. The dropdown menu is open, showing the text "言語を選択" (Select language). The gadget is circled in red. The MTA website content includes a navigation menu, a search bar, and several sections: "Schedules", "Maps", "Service Advisories", "MTA in Pictures - Oct 7, 2008", "Features", "Facts & Figures", and "Regional Travel". The Google logo is visible in the bottom right corner.

Google Desktop: America... Translated version of ht...
 http://www.translate.google.com/translate?prev=hp&hl=en&js=n&u=http://www.www2009.org%0D%0A%0D%0A&sl=en&tl=es
 Customized Links CorpMail Corp. Cal. Meme WSJ NYT MNR Pal W. Home Cal. google sites Bloomberg Google Docs iGoogle Fin Data. Other bookmarks

This page was automatically translated from English.
 View original web page or mouse over text to view original language. Back to Translate

WELCOME ATTENDING AUTHORS COMMITTEES PRESS ROOM SPONSORS STORE REGISTRATION

DAILY PROGRAM
PROCEEDINGS

Documentos y Presentaciones
 Twitter etiqueta # www2009
 Flickr etiqueta WWW2009MADRID
 Grupo de Facebook

DISFRUTAR DE ESPAÑA!
MAPA: LUGAR Y MONUMENTOS
CAPTURAS DE LOS HECHOS DE MADRID VIDA!
TOURS 1 DÍA PARA: TOLEDO, EL ESCORIAL O SEGOVIA
EXPERIENCIA BULLET viajes en tren a: SEVILLA O BARCELONA (2 1/2 horas)
Divertirse en SEVILLA Feria de Abril (4/28/2009 - 5/3/2009)
ENCONTRAR SU PREFERENCIA EN ONDA SURF MUNDAKA O Tarifa

Actualizaciones de Noticias Noticias Anteriores actualizaciones ...

"Abril 22 - CEREMONIA DE INAUGURACIÓN"
 La ceremonia se celebrará en el Auditorio A, justo después de la Web Grupo 20º aniversario.
 Con el fin de asistir a la ceremonia de inauguración, los participantes deben usar la Conferencia Insignia, de haber completado el proceso de inscripción en el registro de contador situado en el Palacio Municipal de Congresos del 20 de abril en adelante.
 Por razones de seguridad siguientes restricciones se aplicarán el 22 de abril:
 • Contra el registro el 22 de abril se abrirá a 07.45 hrs. A partir de las 10.00 a las 11.30 será cerrado
 • Lugar de acceso, le será prohibido de 10.15 a 10.30 hrs. (Incluso a los delegados el uso de una tarjeta de identificación WWW2009)
 • **2009/04/07** relacionado de datos disponibles para WWW2009 documentos | Navegar ponencias y pósters en EPrints
 • **2009/04/07** Mejor Papel / Cartel nominaciones anunciadas.

WWW 2009 MADRID/SPAIN

WWW2009 Programa de un vistazo

20 de abril Lunes	21 de abril Martes	22 de abril Miércoles	23 de abril Jueves	24 de abril Viernes
Tutoriales			Ponencias	
			Documentos de referencia de pista	
Talleres	BOF		W3C pista	
			Via web en Iberoamérica	
W4A			Pósters	
			Carteles Presentaciones de pista	

PASADO A SER PATROCINADOR O EXPOSITOR

PATROCINADORES PLATINO

PATROCINADORES DE ORO

PATROCINADORES PLATA

PATROCINADORES DE BRONCE

YouTube Caption Translation

YouTube Broadcast Yourself™
 Worldwide | English Si

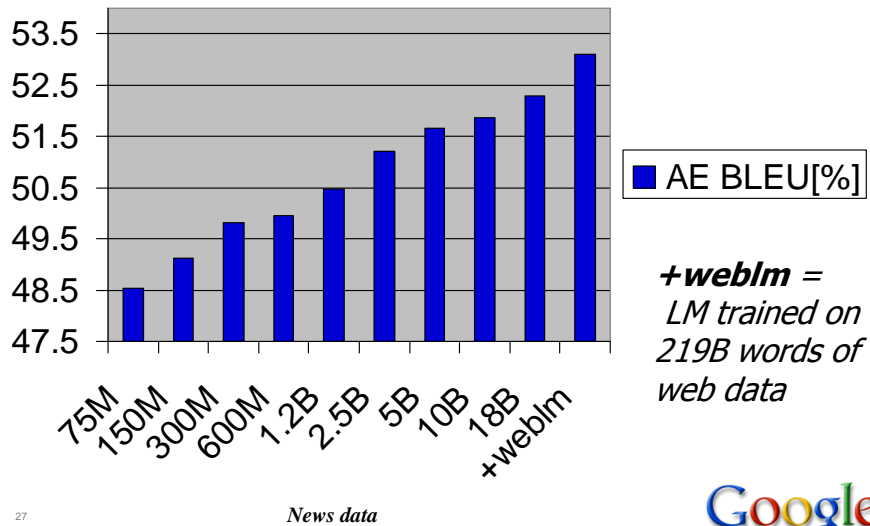
Home Videos Channels Community

2/28/09: Your Weekly Address

2:36 / 4:53 HD



Impact of data - More data is better ...



Challenges in Image processing

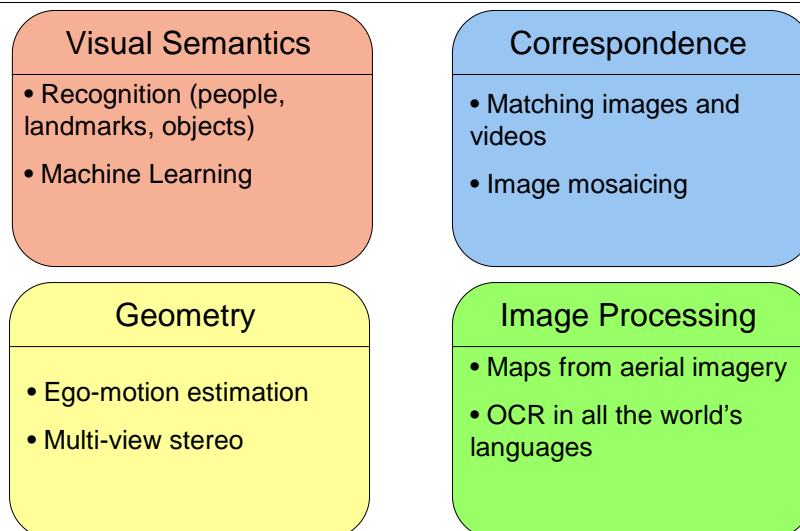


Image Analysis in Image Search

- Image Search helps users find the image they want quickly.
- Understanding the actual content of an image is critical.
- We've been focusing more and more on analyzing images
- This has been rolling out over the last year.
 - Both as user visible filters
 - Behind the scenes in our back-ends.
- Genre filters like clip art / line drawings / color are great examples
 - [\[flowers\]](#), [line drawings](#), [clip art](#), [photo](#), [face](#)
 - [\[porsche\]](#) , [red](#), [green](#), [yellow](#), [orange](#), ...



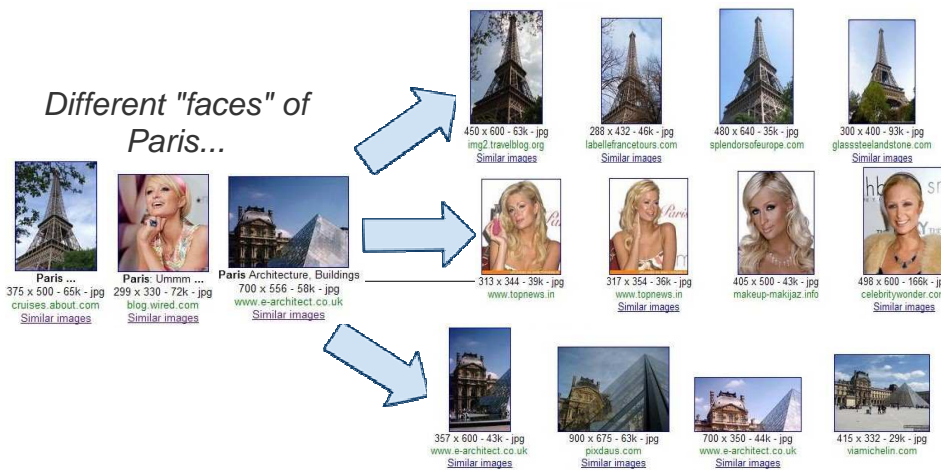
Similar Images in Image Search

- Google has just launched a "Similar Images" feature.
- Accessed by clicking on the similar images link under an image.
- It can also be accessed via preview thumbnails in the result frame.
- We think this will create a major shift in how to search for images.
- Searching for images can now become a navigational experience, where the text (or voice) query acts as a starting point



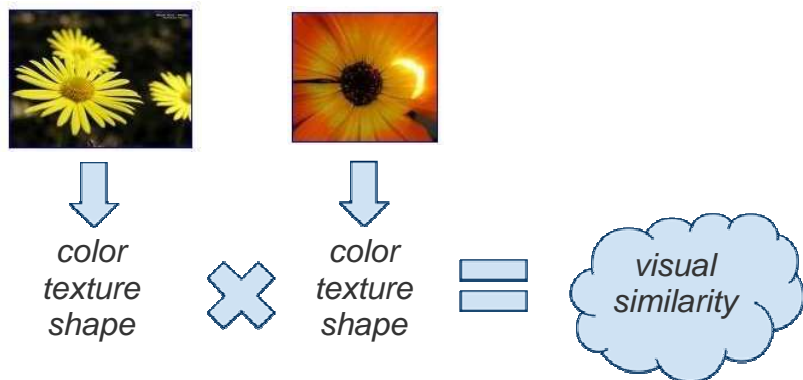
Similar Images in Image Search

Refine by the content of a specific image.



Similar Images in Image Search

- A variety of features are used to determine visual similarity.



Information Fusion across Modalities

*Get lists of people names from the web
by name detection techniques (NLP)*



*Scrape image search results using those
candidate names as queries*



*Detect faces in the images and generate
face signatures*



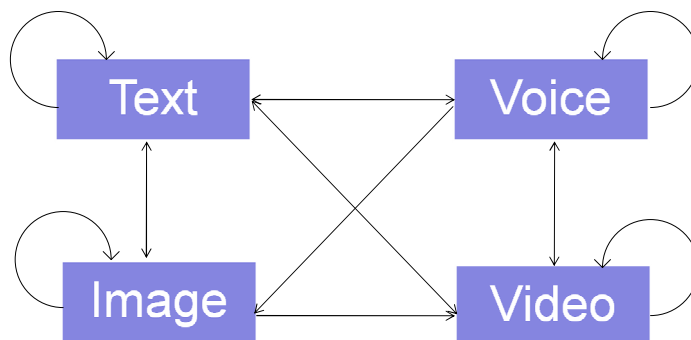
*Apply consistency learning to learn face
models*



Localize known people in Image/Video



Totally Transparent Processing In-Process...



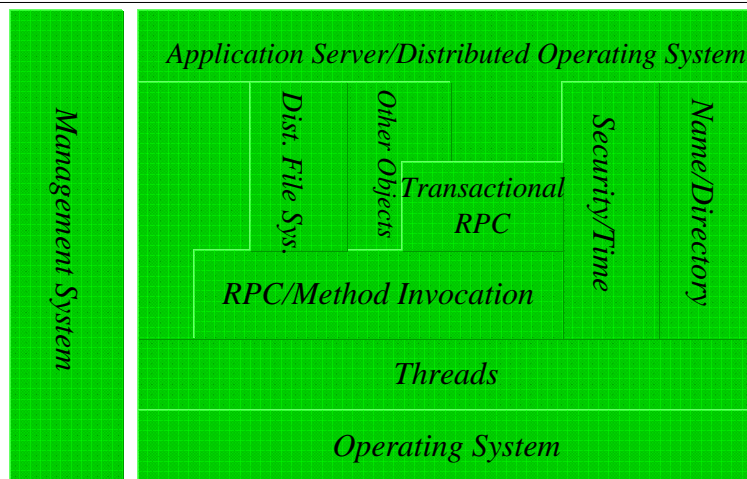
Last two arrows are easily conceivable.



Ideal Distributed Computing



Orthodox Architecture of 70's and 80's



From AZS Pres. To US National Research Council Study on Dependability, May 18, 2004, after a late 80's talk at Univ. of Michigan



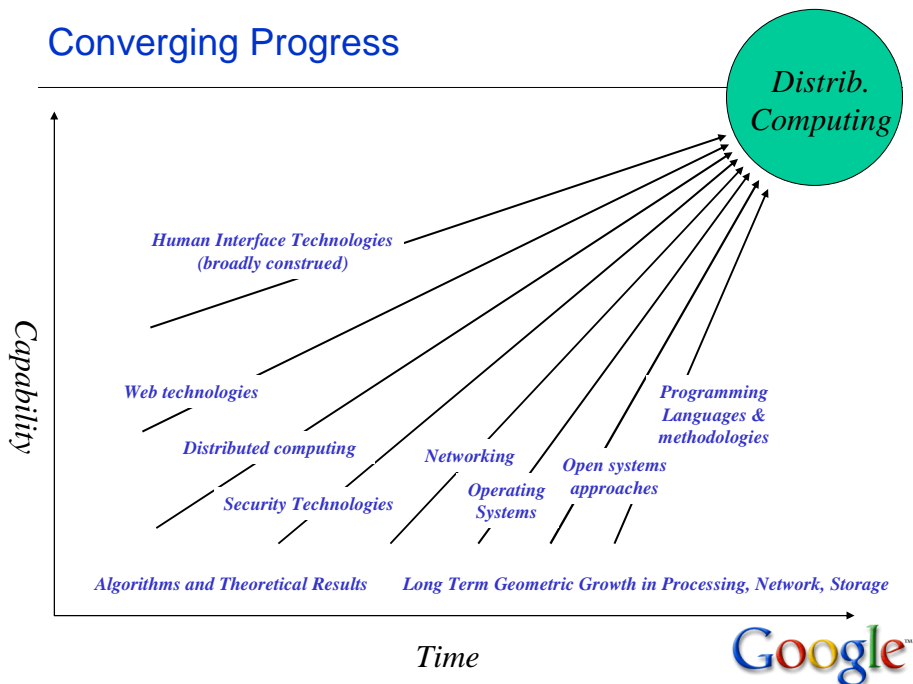
What Wasn't Internalized Very Well

- The application mix
- The true nature of global, open systems:
 - Implications on systems, applications, mix and match.
- The implications of *operations* at true scale
 - E.g., work on programming & runtimes predominated system mgmt.
- The complexity of the architecture that would result
 - We tend to assume, *if we can conceive it, it's okay*.
- The collection of further abstractions that would build on fundamentals then known
- In summary there was a limitation of understanding of (truly) large-scale, open integrated distributed systems

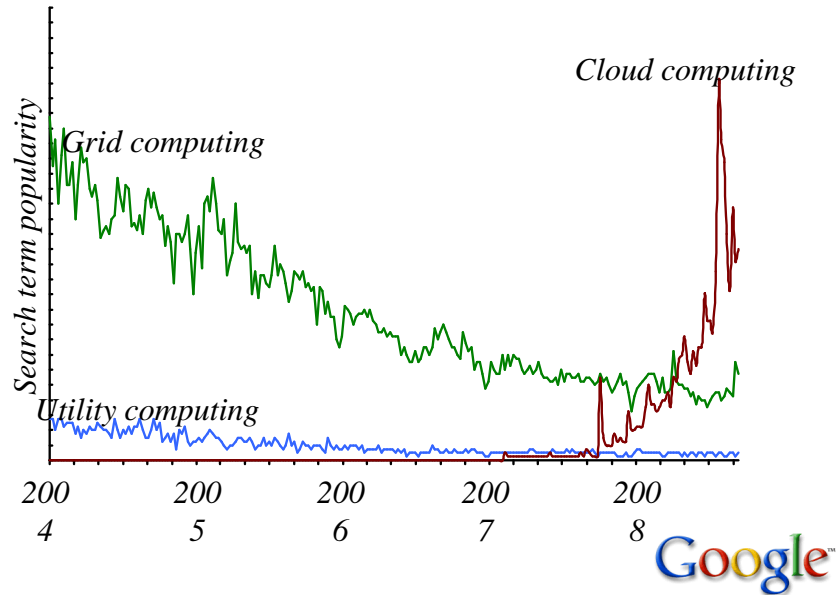
37



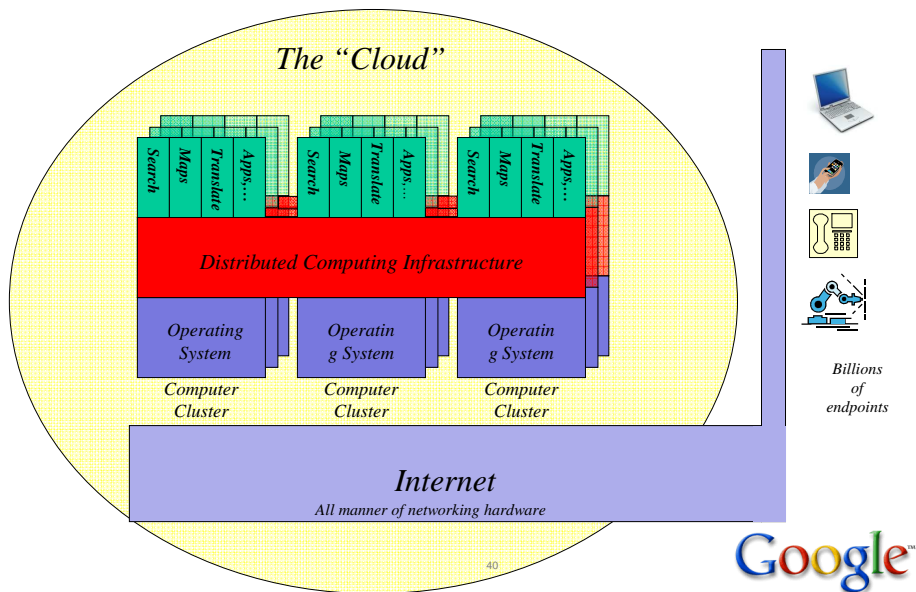
Converging Progress



Terminological Evolution



Cloud Computing Architecture



Excitement in Distributed Systems

- Size of user community
- Storage Scale (requiring various characteristics)
 - E.g., security, privacy, availability,
- Processing Scale
 - High performance batch processing
 - High throughput
 - Low latency
- Rapid dynamics
- Highly variable end-user devices
- Communication Scale
 - Bandwidth
 - Endpoints
- Efficiency
 - Equipment
 - Communication
 - Power
 - Management
- Extensibility
- Compliance
- *And more to come, no doubt*



41

Ideal Distributed Computing

Large networked clusters grow in a fully distributed world

- Arbitrarily high volume transactions
- And, various, partitionable batch process for learning, fusion, etc.
- Network
 - Response-time and bandwidth as needed
- Cluster Processing, or “Cloud Computing” growing ever larger
 - Massive parallelism to hit sweet spot of capital & operating efficiency
- Distributed computing
 - Data sharing, function shipping, as needed
 - Connected and disconnected operation, as seamless as possible
 - Auto balancing of loads between client device and cloud elements
 - Emphasis on manageability (newly, to handle consumers’ many endpoints)
- Significant efficiency gains

42



Hybrid, Not Artificial, Intelligence

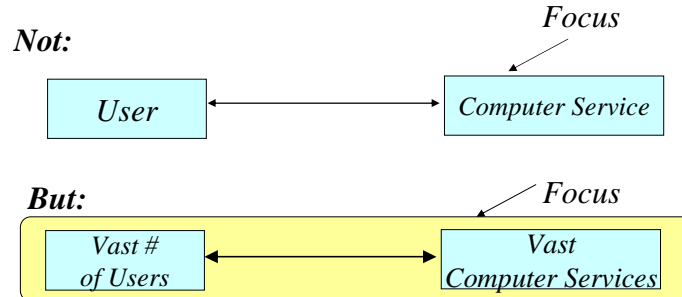


Hybrid, not Artificial, Intelligence

- “Artificial Intelligence” aimed at having computers as capable as people, often in very broad problem domains
- It has proven more useful for computers rather:
 - To extend the capability of people, not in isolation
 - To focus on more specific problem areas
- Aggregation of user responses has proven extremely valuable in learning
- Examples
 - Feedback in Information Retrieval; e.g., in ranking or spelling correction
 - Machine learning; e.g., image content analysis, speech recognition with semi-supervised learning
- Another example of bottom up successes



Key: Holistic Approach To Design



- **Implications**
 - Users and computers doing more than either could individually.
 - Virtuous circle from: *Data and Processing, Reach, Feedback in a virtuous circle.*

45



Empiricism - Let Measurement & Feedback Rule

108 milliseconds -0.54 2.7M 0.55060

\$4.78 RPM

-0.0000339 2,800,000,000 views

425,440.01 56.76% 17.35

9995.55 *1.3 searches per user* 108 seconds/search

480,000,000 total pageviews

1607.44 10,400 6.55

\$0.303 CPA \$7,660,400 108



My Long-held View on Semantics, Syntax, & Learning

- Large scale learning has proven surprisingly effective
- Learning is occurring over increasingly variegated features:
 - Both Semantic
 - And Syntactic, and generated in multiple ways
- In my WWW 2002 (Architecting Knowledge Middleware) and Semantic Web 2005 Keynotes, I referred to this as *The Combination Hypothesis*
- Today, I would refine this as the combination of approaches *and* learning from people.

47



Research Challenges



Challenges in Transparent Computing & Hybrid Intelligence

- Endless applications, with very new user interface implications
- Addressing limits to data
- Techniques to integrate user-feedback in acceptable fashions
- Approaches to new signal (e.g., annotations)
- Explanation, scale, and variance minimization in machine learning
- Information fusion/learning across diverse signals – The Combination Hypothesis, more generally
- Usability: devices and subpopulations

49



Research Challenges in Ideal Distributed Computing

- Alternative designs that would give better energy efficiency at lower utilization
- Server O.S. design aimed at many highly-connected machines in one building
- Unifying abstractions for exploiting parallelism beyond inter-transaction parallelism and map-reduce
- Latency reduction
- A general model of replication including consistency choices, explained and codified
- Machine learning techniques applied to monitoring/controlling such systems
- Automatic, dynamic world-wide placement of data & computation to minimize latency and/or cost, given constraints on
- Building retrieval systems that efficiently and useably deal with ACLs
- The user interface to the user's diverse processing and state

50



3 Interesting Challenges

- Security and Privacy Technologies and Policy
- Application of technologies to health
- Applications to Government

51



Conclusions

- The Web's brilliant initial design lead to a series of local optimizations with extraordinary results
- Evolutionary advances continuing
- They are aggregating into at least 3 major advances:
 - A. Totally Transparent Processing
 - B. The Rule of Distributed Computing
 - C. Hybrid, not Artificial, Intelligence
- Challenges for academics and industrial researchers/engineers abound

52



¡Muchas Gracias!

Thank you very much!

