

Mathematics at

Javier Tordable
Software Engineer



Index

1. How Google started
2. PageRank
3. Gallery of Mathematics
4. Questions

How Google started



Backrub

<http://www.google.es/intl/es/about/corporate/company/history.html>

- **1995:** Larry Page and Sergey Brin meet at Stanford. (Larry, 22, a U Michigan grad, is considering the school; Sergey, 21, is assigned to show him around.)
- **1996:** Larry and Sergey, now Stanford computer science grad students, begin collaborating on a search engine called BackRub. BackRub operates on Stanford servers for more than a year—eventually taking up too much bandwidth to suit the university
- **1997:** Larry and Sergey decide that the BackRub search engine needs a new name. After some brainstorming, they go with Google. The use of the term reflects their mission to organize a seemingly infinite amount of information on the web
- **1998:** In September, Google sets up workspace in Susan Wojcicki's garage. Google files for incorporation in California on September 4



Web search (1)

- Consider the web as a collection of documents. The standard search paradigm uses an index of terms
- An **index** is an inverted table. In this table we have, for each term, the list of documents that contain such term
- We can use this index to search for combinations of terms by intersecting the lists of documents for each term
- The problem consists in sorting this list of documents

Web search (2)

<http://www.one.com> (1)

document
number one

<http://www.two.com> (2)

document
number two

<http://www.three.com> (3)

yet another
page

document	1,2
number	1,2
one	1
two	2
yet	3
...	

Google downloads
documents and
builds an index

document

Google Search

<http://www.one.com>

<http://www.two.com>

When searching for a
word, Google shows the
corresponding entries in
the index table



Web search (3)

- Showing all documents retrieved is simply not possible
- Showing documents based on simple criteria, like date or number of occurrences of the term gives results of bad quality
- The idea of Larry Page and Sergey Brin was to use the links between documents as a signal to indicate the quality of a document. In a similar way to how citations between scientific papers work



PageRank

Google

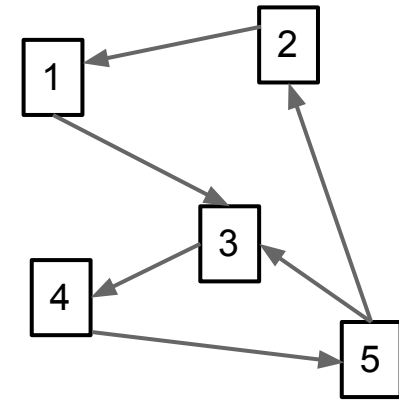
The Web as a graph

<http://one.com>

```
<h1>
Web page 1
</h1>
<p>
This is my web.
</p>
<p>
My other page is
<a href="http://two.com/">
here
</a>
<p>
```

<http://two.com/>

```
<h1>
Web page 2
</h1>
<p>
This is my other web.
</p>
<p>
My other page is
<a href="http://one.com/">
here
</a>
<p>
```

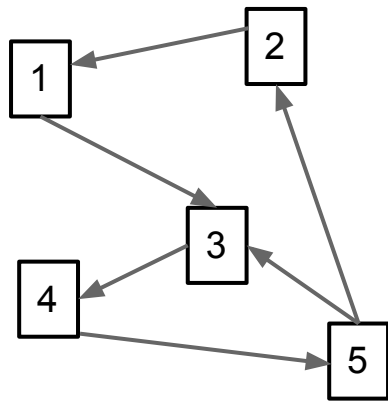


Iterative version of PageRank (1)

- PageRank is an approximation to the probability of reaching a page following links randomly
- For example: if a person is in page i with probability p_i , which has links to pages $\{j, k\}$ then the probability to reach j is $1/2 * p_i$ and the probability to reach k is also $1/2 * p_i$
- If a page doesn't have outgoing links we assume that it links to every other page
- Initially we assume that the probability of reaching all pages is the same



Iterative version of PageRank (2)



$$p1_k = p2_{k-1}$$

$$p2_k = 1/2 * p5_{k-1}$$

$$p3_k = p1_{k-1} + 1/2 * p5_{k-1}$$

$$p4_k = p3_{k-1}$$

$$p5_k = p4_{k-1}$$

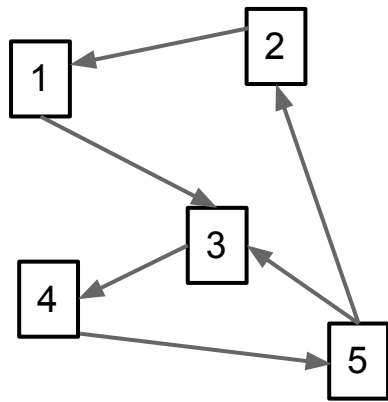
$$\begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}_k = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 \\ 1 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}_{k-1}$$

- In each phase, the probability (PageRank) is computed from the probability in the previous phase
- We can define a matrix A, which has in each position (i,j) a 0 if the page j does not link to page i, or 1/k if page j has k outgoing links and one of them is to page i
- In the first step we initialize the probabilities of all pages to the same value. Each subsequent step is computed according to $p_k = A * p_{k-1}$
- In general, after a reasonable number of iterations, we can obtain a reasonable approximation to PageRank

Algebraic version of PageRank (1)

- Consider web pages as nodes, links as edges, and the web as a directed graph
- PageRank is an estimation of the importance of each node in the graph
- If a page has k outgoing links to pages P_1, \dots, P_k , we can consider each link as a vote for page P_k
- The PageRank of page P_k , pr_k , is the sum of all the votes for this page. Each vote from a page P_i is weighted by the PageRank of P_i

Algebraic version of PageRank (2)



$$\begin{aligned} p_1 &= p_2 \\ p_2 &= 1/2 * p_5 \\ p_3 &= p_1 + 1/2 * p_5 \\ p_4 &= p_3 \\ p_5 &= p_4 \end{aligned}$$

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 \\ 1 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix}$$

- Taking $p = G \cdot p$, the PageRank vector is an eigenvector with eigenvalue 1
- G is a stochastic matrix. All elements are positive and the sum of the elements in each column is 1
- Column i contains $1/k$ for each one of the k outgoing links from node i
- If a node has no outgoing links, we assume that this node links to all other nodes. This is necessary for the matrix to be stochastic
- In this conditions the matrix will always have the eigenvalue 1

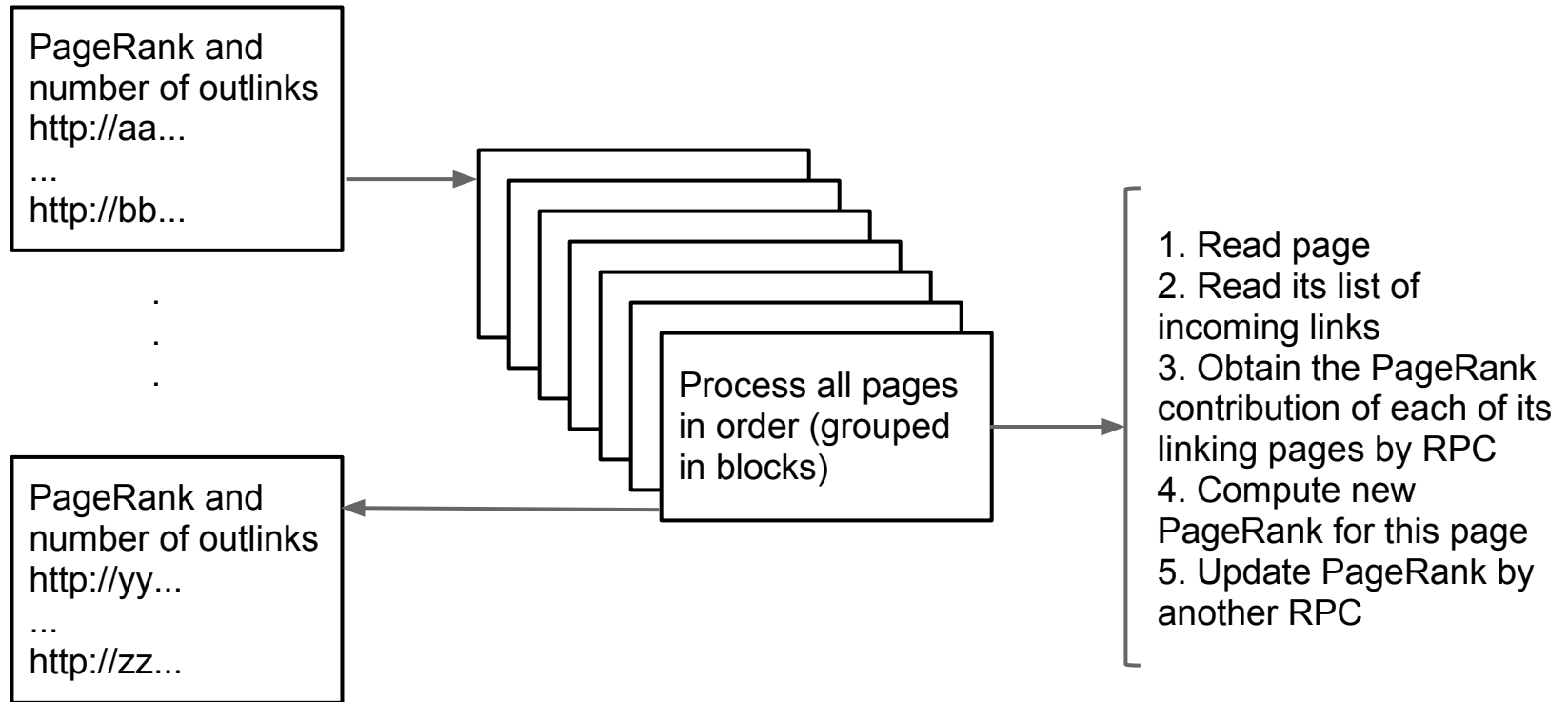
PageRank (1)

- The algorithms described before have problems when the graph is not connected. Either because it's not possible to reach a particular page by following links (in the iterative version) or because there are multiple eigenvectors for the eigenvalue 1 (in the algebraic version)
- The solution is to add a factor $\lambda * \frac{1}{n} * \mathbf{1}$, where $\mathbf{1}$ is a matrix with ones in all positions and n is the number of nodes (And normally $\lambda = 0.15$)

PageRank (2)

- The *Google* matrix is:
$$G = (1 - \lambda) A + \lambda \frac{1}{n} \mathbf{1}$$
- This matrix is also stochastic, and all elements are strictly positive
- From the Perron-Frobenius theorem, G has the eigenvalue 1 and the corresponding eigenvector has multiplicity 1
- Using the power iteration method with G it's possible to find this same eigenvalue in an iterative way

Sample implementation



Solving a linear system is difficult to parallelize. Even if the iterative method is slow, it's faster to reach an approximation

PageRank. Links

Original article about Google from Sergey Brin y Larry Page:

<http://infolab.stanford.edu/~backrub/google.html>

Presentation about PageRank at Cornell University:

<http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html>

PageRank:

<http://es.wikipedia.org/wiki/PageRank>

Perron-Frobenius theorem:

http://en.wikipedia.org/wiki/Perron%E2%80%93Frobenius_theorem

Power iteration:

http://en.wikipedia.org/wiki/Power_iteration




Gallery of Mathematics

Google

Gmail (1)

Gmail Calendar Documents Photos Reader Web more ▾ Hiking Fan ▾ ⚙

 **SEARCH MAIL** **SEARCH THE WEB** [Show search options](#)
[Create a filter](#)

Mail

COMPOSE MAIL

Inbox (3)

Starred ☆

Sent Mail

Drafts (2)

Hiking (3)

Urgent!

12 more ▾

Chat

Search, add, or invite

Hiking Fan

Set status here ▾

Call phone

Arielle

Emily


Jason

Michael


Paul

▾ Archive Report spam Delete Move to ▾ Labels ▾ More ▾ 1 - 15 of 15

- ☆ Jason Cornwell > **Please return my stapler** - Hi, You seem to have taken my stapler. Please, 1:10 pm
- ☆ Paul McDonald > **Fun Hike Yesterday!** - Thanks for the great hike yesterday, it was awesome 1:06 pm
- ☆ Arielle Reinstein > **July 4th weekend** - Hi there: I heard you'll be around this weekend and I'd lo Jun 28
- ☆ JS Bach > **Tonhalle concert Friday** - Hey man, there's a great concert this Friday evenir Jun 22
- ☆ Christine Chiu > **Hi Hiking, Looking for opinion on my diet/fitness app** - Hi Hiking, I bumped ir Jun 9
- ☆ Yan Tseytlin (2), **Draft** > **Hey there!** - I heard you found a great place to go hiking. Let me know when Mar 28
- ☆ Kenneth, me (2)
- ☆ Kenneth, me (2)
- ☆ Michael Bolognino
- ☆ Arielle Reinstein
- ☆ Jason Toff > **How are you?** - Hey there, We haven't spoken in a while. How are you? Wou Mar 24
- ☆ Jr Wikane > **VW Auction in Tacoma** - Hi, I was doing a search on Google for VW's in Tac Mar 6
- ☆ Google Voice > **New voicemail from (619) 819-5507 at 5:19 AM** - Voicemail from: (619) 819- Mar 6

 **Spammer** spammer@spammer.com Mar 12 ☆

to me ▾

 **Be careful with this message. It might contain a virus or a malicious link.** [Learn more](#)



Gmail (2)

- Spam detection is a classical example of classification using machine learning (the computer learns the algorithm from the data), in particular supervised learning (where we have previously classified data samples)
- In essence, machine learning has two phases, the training phase (when we build the classification model) and the classification phase (using the model to classify new instances)

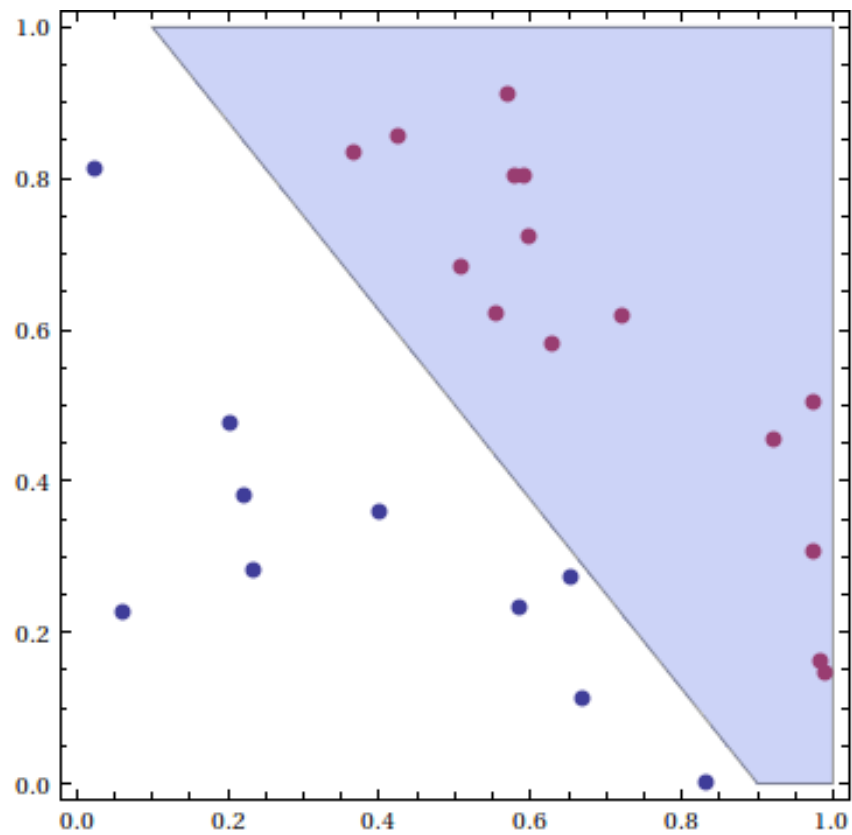
Gmail (3)

- The classification phase involves extracting the characteristics of the data instance, and then applying the model to the characteristics
- In general, the characteristics of an instance can be considered as elements in a vector of an n -dimensional euclidean space for a large n (100-1000 dimensions is normal, 1M-10M is not unheard of)
- The model is a subspace of dimension $n-1$ which divides the original space into two disjoint subspaces

Gmail (4)

- A simple example
- From an email we can extract characteristics such as: length of the email, number of capital characters, whether the sender is in the address book, etc.
- A simple classification model is a hyperplane in the space of characteristics. Data instances on one side of the hyperplane are classified as valid emails and instances on the other side are classified as spam

Gmail (5)



Gmail (6)

Slightly more complex examples:

- Decision trees (step functions)
- Neural networks (each node of the network is a composition of a function, normally a logistic function, with a linear combination of its inputs. A network is formed by multiple levels of nodes)
- Support vector machines with a kernel function (composition of a linear function with a nonlinear transform of the original space)

Gmail (7)

Links:

- The War Against Spam: A report from the front line
<http://research.google.com/pubs/pub36954.html>
- The Learning Behind Gmail Priority Inbox
research.google.com/pubs/archive/36955.pdf
- Publications by Googlers in Artificial Intelligence and Machine Learning
<http://research.google.com/pubs/ArtificialIntelligenceandMachineLearning.html>



Google trends (1)



Google trends (2)

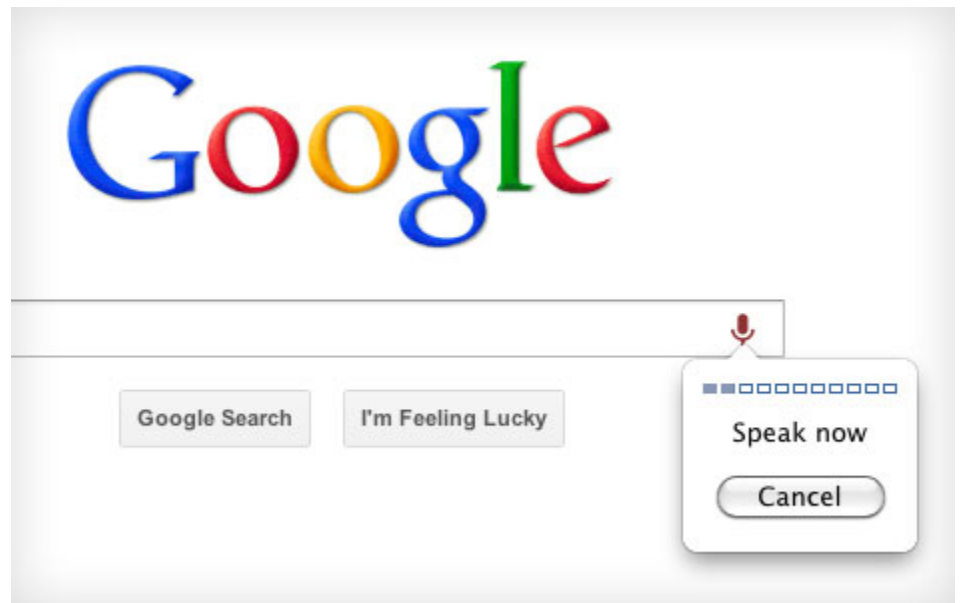
- Time series processing is one of the most common uses of applied mathematics. The techniques used range from regression to Fourier Analysis, hidden Markov models or self-correlation
- It is used to predict the number of search queries in a given day, number of users, income, etc. for a variety of products (thousands of daily analysis)

Large-Scale Parallel Statistical Forecasting Computations in R

<http://research.google.com/pubs/pub37483.html>



Voice search (1)



Voice search (2)

- Automated speech recognition (ASR) has two fundamental parts:
- First, the processing of the sound signal. Splitting it into smaller parts, applying the Fourier transform and extracting the most significant coefficients
- Second, modelling the speech using a hidden Markov model. In this model the states are the letters of the message and the sequence of events is the sound signal. The Viterbi algorithm can be used to obtain the sequence of states of maximum likelihood

Google Search by Voice: A case study

<http://research.google.com/pubs/archive/36340.pdf>



Google books (1)

Google

el quijote de la mancha



Books

Add to my library ▾

Write review

GET PRINT BOOK

No eBook available

[Amazon.com](#)

[Barnes&Noble.com](#)

[Books-A-Million](#)

[IndieBound](#)

[Find in a library](#)

[All sellers »](#)



New! [Shop for Books on Google Play](#)

Browse the world's largest eBookstore and start reading today on the web, tablet, phone, or ereader.

[Go to Google Play Now »](#)

[My library](#)

[My History](#)

[Books on Google Play](#)

El ingenioso hidalgo Don Quijote de la Mancha



👤 +1 -0

Alonso Fernández de Avellaneda

★★★★★

1 Review

Biblioteca Nueva, 2000 - [Literary Criticism](#) - 789 pages

From inside the book

el quijote de la mancha

Search

100 pages matching **el quijote de la mancha** in this book

Page 10

de Vega. Pero lo cierto es que Avellaneda nunca tuvo en su mira hacer lo mismo que ellos. De hecho, evitó mencionar intencionadamente la continuación más cercana en el tiempo, pero también la más problemática y próxima a la



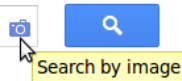
Google™

Google books (2)

- OCR techniques (optical character recognition) can be considered as a combination of image processing (obtaining individual characters images, with appropriate resolution, orientation and contrast levels) and machine learning (character classification)
- For example:
 - An Overview of the Tesseract OCR Engine
<http://research.google.com/pubs/archive/33418.pdf>
 - Low Cost Correction of OCR Errors Using Learning in a Multi-Engine Environment
<http://research.google.com/pubs/archive/35525.pdf>
 - Translation-Inspired OCR
<http://research.google.com/pubs/pub37260.html>



Image search (1)



at image results with new related search previews. [Learn more.](#)



Search by image

Search Google with an image instead of text.

[Paste image URL](#) | [Upload an image](#) ?

No file chosen



pic.jpg

scarlett johansson

Search

About 3,450 results (0.85 seconds)

Everything

Images

Maps

Videos

News

Shopping

More



Image size:
1600 × 1200

Find other sizes of this image:
[All sizes](#) - [Medium](#) - [Large](#)

Best guess for this image: [scarlett johansson](#)

[Scarlett Johansson - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Scarlett_Johansson

Scarlett Johansson (born November 22, 1984) is an American actress, singer. Johansson made her film debut in North (1994) and was later ...

[Demo](#)



Image search (2)

- Image search is an example of content based information retrieval (using colors, shapes, textures, etc.)
Content-based Multimedia Information Retrieval: State of the Art and Challenges
<http://www.liacs.nl/home/mlew/mir.survey16b.pdf>
- The key concept is the measure of similarity between images. For example the difference between the color histograms, or in general the difference between the characteristic vectors of the images
Tour the World: building a web-scale landmark recognition engine
<http://research.google.com/pubs/archive/35291.pdf>
(Image search) Web-scale Image Annotation
<http://research.google.com/pubs/archive/34669.pdf>



Picasa (1)

[Gmail](#) [Calendar](#) [Documents](#) **Photos** [Reader](#) [Web](#) [more -](#)

Picasa™ Web Albums [Home](#) [My Photos](#) [Explore](#) [Thomas K's Gallery](#)

Thomas K > Picnik

Slideshow Share Add photos [Download](#) Prints



Football!



Square crop, lots more



Portrait 4x6 crop and some



Happy St. Patty's

Unnamed people [Expand groups](#) [Show ignored faces](#)

Unnamed people in these photos:

-
-
-
-
-
-

Picasa (2)

- An image is basically a set of three integer valued matrices, one for each primary color
- Digital image processing, and in particular applying a filter consists in executing a convolution operation in these matrices

<http://lodev.org/cgtutor/filtering.html>

http://www.emt.jku.at/education/Inhalte/se_moderne_methoden/WS0304/Haim-Mathematics_in_Imaging.pdf

- One of the recent features in Picasa is automatic face recognition. In general face recognition is a complex problem in image processing and machine learning

Handbook of Face Recognition

<http://research.google.com/pubs/archive/36368.pdf>

Large-Scale Manifold Learning

<http://research.google.com/pubs/pub34395.html>



YouTube (1)

The image shows a screenshot of the YouTube homepage. At the top left is the YouTube logo. To its right is a search bar with a magnifying glass icon. Further right are navigation links: "Explorar", "Películas", and "Subir video". Below this is a dark navigation bar with categories: "Vídeos", "Música", "Películas", "Programas", "En directo", "Deportes", "Educación", and "Noticias". On the left side, there is a sidebar with "Todas las categorías" and a list of categories: "Recomendaciones personalizadas", "Motor", "Comedia", "Ocio", "Cine y animación", "Juegos", "Consejos y estilo", and "ONG y activismo". The main content area features a "Más vistos hoy" section with a star icon. It displays six video thumbnails in a 2x3 grid. Each thumbnail includes a video preview, a title, and view statistics.



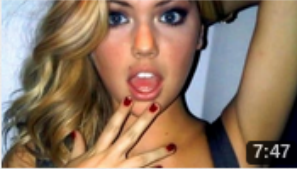



YouTube Explorar | Películas | Subir video

Vídeos | Música | Películas | Programas | En directo | Deportes | Educación | Noticias

Todas las categorías

- Recomendaciones personalizadas
- Motor
- Comedia
- Ocio
- Cine y animación
- Juegos
- Consejos y estilo
- ONG y activismo

★ Más vistos hoy »

 <p>Reveal Trailer - Official Call of Duty: Black Ops 2 Visto 14049349 veces hace 4 días CALLOFDUTY</p>	 <p>CHEESEBURGER! - Ray William Johnson Visto 4168215 veces hace 4 días RayWilliamJohnson</p>	 <p>KATE UPTON IS YO CAT DADDY!! Visto 2344191 veces hace 5 días sxephil</p>
 <p>Minecraft - Mind Control Hat - Explosives+ Mod Spotlight Visto 1433136 veces hace 5 días BlueXephos</p>	 <p>Black Ops 2 - Official Gameplay Reveal Trailer Visto 1102291 veces hace 4 días TmarTn</p>	 <p>Dear YouTube Visto 816481 veces hace 4 días RoosterTeeth</p>

YouTube (2)

- There are many mathematical applications in a complex product like YouTube. For example:
- YouTube video is compressed. (http://en.wikipedia.org/wiki/Data_compression). Compression algorithms fundamentals (http://en.wikipedia.org/wiki/Rate%E2%80%9393distortion_theory) come from information theory, coding theory, etc.
- Another problem is automatic event detection. For example to classify video, or to create snippets
YouTubeEvent: On Large-Scale Video Event Classification
<http://research.google.com/pubs/archive/37392.pdf>
YouTubeCat: Learning to Categorize Wild Web Videos
<http://research.google.com/pubs/archive/36387.pdf>

Google translate (1)

+Tú Búsqueda Imágenes Maps Play YouTube Noticias Gmail Docs Calendar Más ▾

Google Iniciar sesión

Traductor Del: inglés - detectado ▾ ↔ Al: español ▾ Traducir

inglés español francés español inglés francés

hello everybody ×

Hola a todos 🔊 ✓

¡Nuevo! Haz clic en las palabras anteriores para editar y ver traducciones alternativas. [Descartar](#)

Traductor de Google para empresas: [Google Translator Toolkit](#) [Traductor de sitios web](#) [Global Market Finder](#)



Google translate (2)

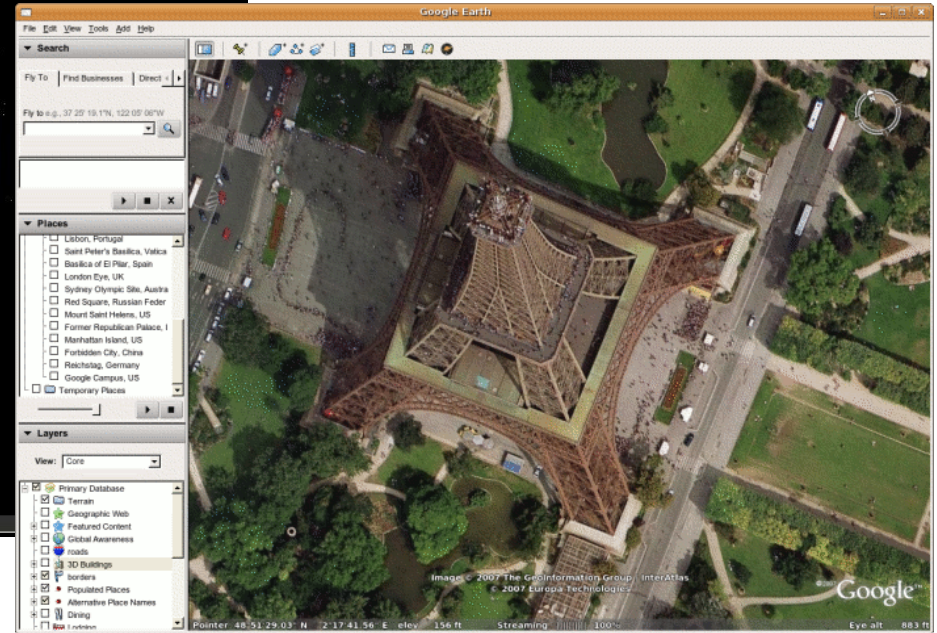
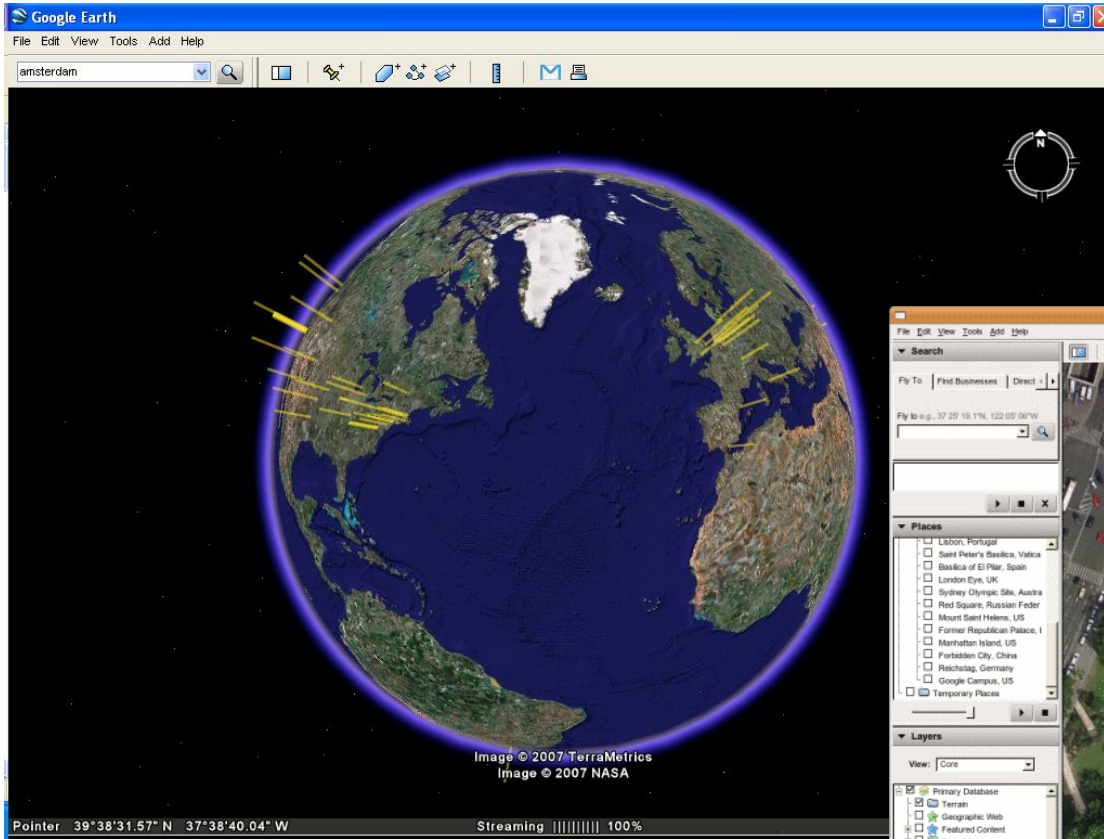
- There are multiple techniques for automatic translation. One of them consists in parsing the text into an abstract representation and then transforming this representation into the destination language. But this requires knowledge about the structure of language
- The method used at Google relies on an immense amount of data to build a statistical model of the translation

Large Language Models in Machine Translation

<http://research.google.com/pubs/archive/33278.pdf>



Google Earth (1)



Google Earth (2)

- The fundamentals are the 3D Euclidean geometry, topography and photogrammetry, fusion of 2D and 3D data, etc. All these are well understood areas
- The greatest contributions from Google are in the issues that come up with huge amounts of data, applying these techniques at Web scale

AdWords (1)

Google

best laptops 2012



Sign in

Search

About 199,000,000 results (0.19 seconds)



Everything

Images

Maps

Videos

News

Shopping

More

Seattle, WA

Change location

Any time

Past hour

Ad related to **best laptops 2012**

[Why this ad?](#)

[The new MacBook Air - The future of the notebook | apple.com](http://www.apple.com/macbookair)

www.apple.com/macbookair

Available now. Learn more.

211 people +1'd this page

↳ [Why you'll love a Mac - OS X Lion - Great Mac apps](#)

[Laptop Computer 2012 | Best Laptop Computers | Compare Laptop ...](#)

computers.toptenreviews.com > Computers

Compare the **best laptop** computers. Side-by-side comparisons of features and prices of top rated **laptops**. Easily see which **laptop** computer stands above the ...

↳ [Dell XPS 15z - Laptop Configurator - HP Pavilion dm4x - Sony S Series](#)

[Best & Worst Laptop Brands 2012](#)

www.laptopmag.com/mobile-life/best-brands-2012.aspx

Mar 14, 2012 – The biggest **laptop**-makers ranked from software to support. Find out which brand makes the **best laptops**.

↳ [Overall Scorecard - Acer/Gateway - ASUS - Dell/Alienware](#)

Ads - [Why these ads?](#)

[Windows Recommended PCs](#)

windows.microsoft.com/

Answer 5 Quick Questions & We'll Recommend the **Best** PCs For You.

[Best-Selling Laptops](#)

www.bestbuy.com/Laptops

Get Connected with a **Laptop**. Free Shipping + In-Store Pickup! 330 NE Northgate Way, Seattle, WA

[New 2012 Dell Laptops](#)

www.dell.com/Laptops

dell.com is rated ★★★★★

Find the Latest **2012**, Performance **Laptops** w/ Intel® Core™ at Dell! 1,724 people +1'd Dell

Google™

AdWords (2)

AdWords uses an auction algorithm. Each advertiser makes a bid for the ad inventory

Hal Varian. Online Ad Auctions:

<http://people.ischool.berkeley.edu/~hal/Papers/2009/online-ad-auctions.pdf>

- Auction theory studies different bidding strategies and their effectiveness. It's an applied branch of Game Theory
- In particular, AdWords uses a generalized second price auction

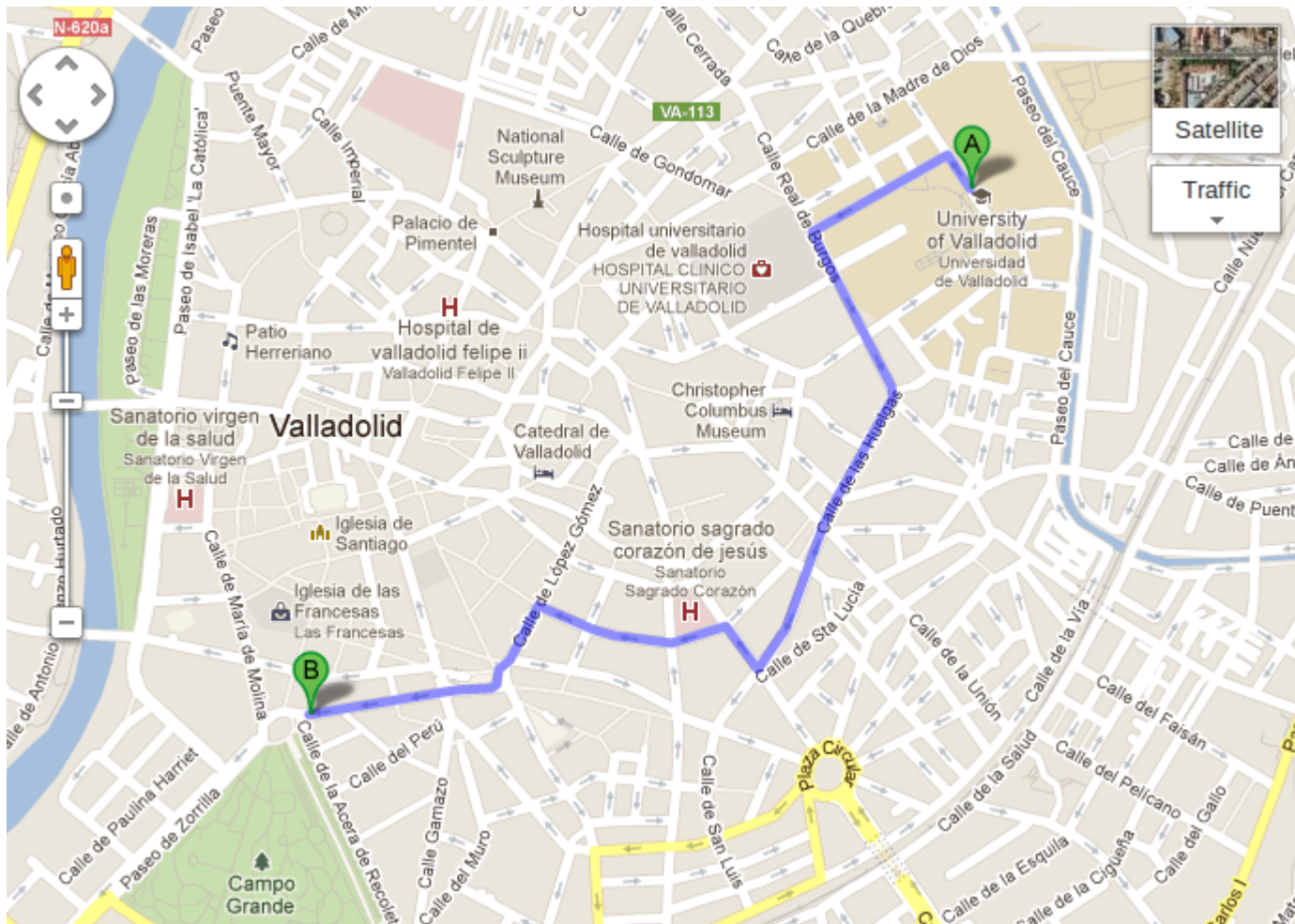
http://en.wikipedia.org/wiki/Generalized_second-price_auction

Adwords, An Algorithmic Perspective

<http://paul.rutgers.edu/~mangesh/cs514/notes/pres3.pdf>



Google Maps (1)



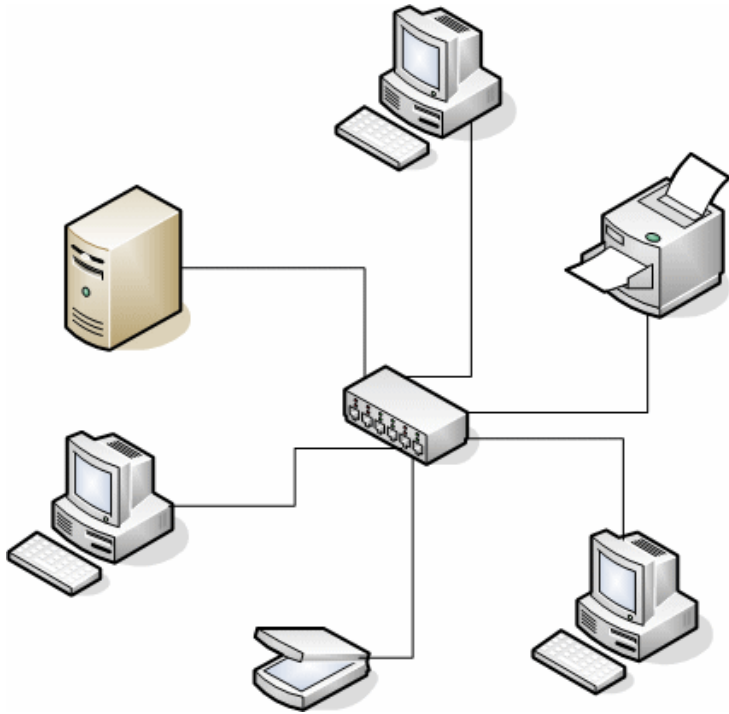
Google Maps (2)

- Google Maps uses many basic algorithms from Graph Theory. For example, to find the shortest path between two nodes in a graph (Dijkstra) in order to get driving directions
- One unique problem is that the graphs used in Google Maps contain millions of nodes, but the algorithms have to run in milliseconds. A technique used to improve performance is graph hierarchies

<http://algo2.iti.kit.edu/schultes/hwy/esaHwyHierarchies.pdf>



Distributed systems (1)



Distributed systems (2)

- There are many statistical techniques used to model availability of computing resources. This is similar to quality control techniques in other industries
- For example, hidden Markov models:
Availability in Globally Distributed Storage Systems
<http://research.google.com/pubs/pub36737.html>
Designs, Lessons and Advice from Building Large Distributed Systems
<http://www.cs.cornell.edu/projects/ladis2009/talks/dean-keynote-ladis2009.pdf>
- Queue Theory can be used to model the execution of batch jobs in a distributed system

Distributed systems (3)

- A classical example is the application of Graph Theory to network links between data centers or in general computer networks
- A network is modelled as a graph in which links can fail with some probability
- It's interesting to study which graph topologies provide the best fault tolerance, bandwidth (graph connectivity) or latency (diameter) for the lowest cost (less number of links)

Distributed systems (4)

- An especially interesting topic in Graph Theory with multiple applications to computing (not just at Google) is the concept of Ramanujan graph
- Ramanujan graphs are an example of expander graphs, which have topologies with properties that make them very useful
- They are also used to build sorting networks (AKS) which can sort n items in time $\log(n)$
- Ramanujan graphs are those that satisfy the equivalent of Riemann's Hypothesis for the Ihara zeta function
http://en.wikipedia.org/wiki/Ihara_zeta_function

Questions





More links

Publications by Googlers

<http://research.google.com/pubs/papers.html>

Course on the Web graph

<http://www.math.ryerson.ca/~abonato/webgraph.html>

Thanks!

