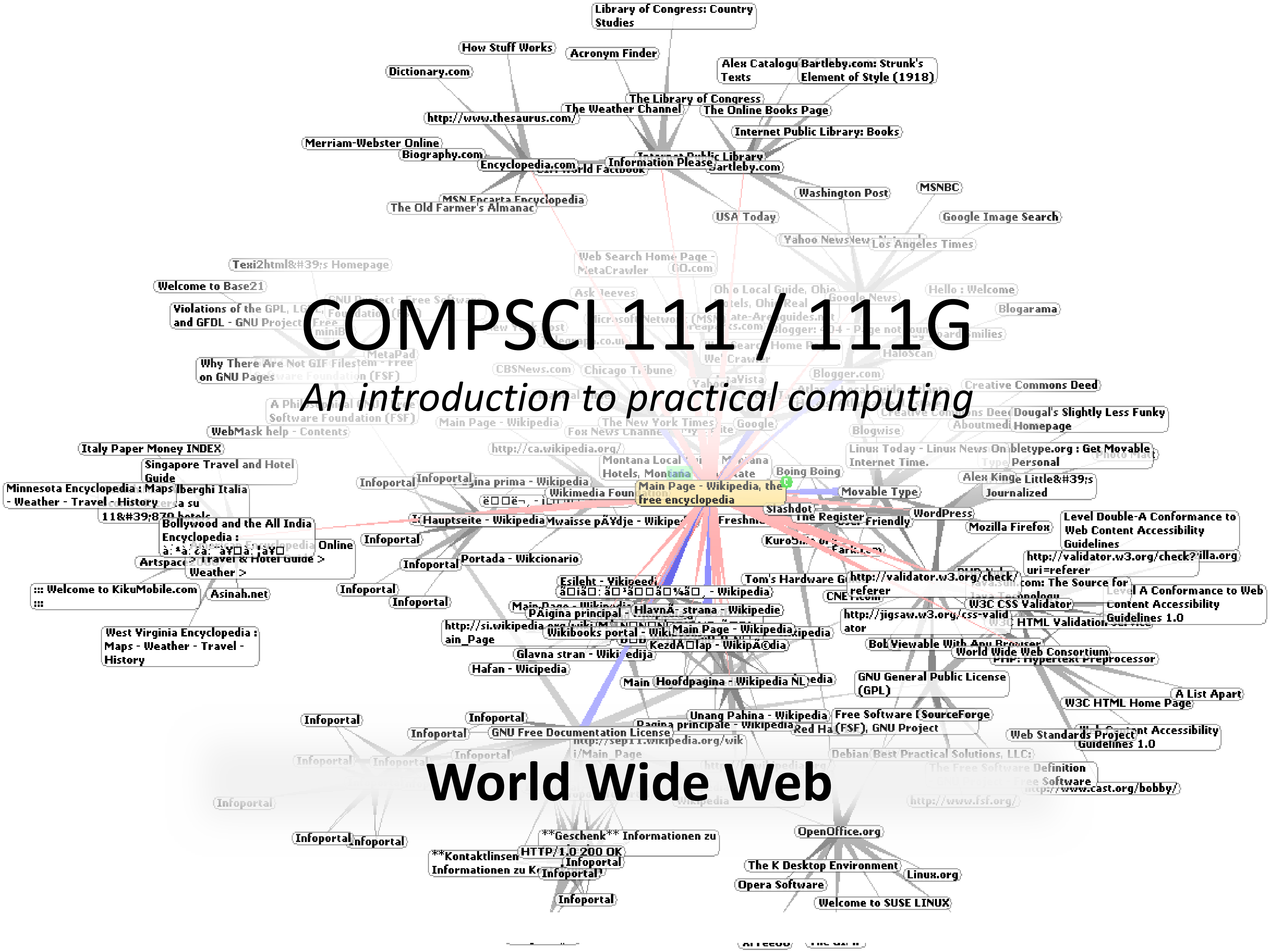
COMPSCI 111 / 111G

An introduction to practical computing



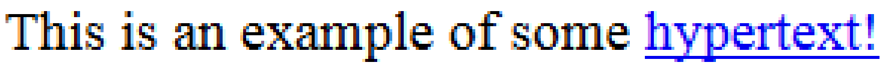
Library of Congress: Country Studies
How Stuff Works
Acronym Finder
Dictionary.com
Alex Catalogue Texts
Bartleby.com: Strunk's Element of Style (1918)
The Library of Congress
The Weather Channel
The Online Books Page
http://www.thesaurus.com/
Internet Public Library: Books
Merriam-Webster Online
Biography.com
Encyclopedia.com
Information Please
Bartleby.com
MSN Encarta Encyclopedia
The Old Farmer's Almanac
Washington Post
MSNBC
USA Today
Google Image Search
Yahoo News
Los Angeles Times
Texi2html&#39;s Homepage
Web Search Home Page - MetaCrawler
Welcome to Base21
Ask Jeeves
Hello : Welcome
Google News
Blogarama
Violations of the GPL, LGPL, and GFDL - GNU Project - Free Software Foundation (FSF)
MetaPad
HaloScan
Why There Are Not GIF Files on GNU Pages
CBSNews.com
Chicago Tribune
Blogger.com
Creative Commons Deed
Dougal's Slightly Less Funky Homepage
WebMask help - Contents
Main Page - Wikipedia
The New York Times
Fox News Channel
Google
Blogwise
Italy Paper Money INDEX
http://ca.wikipedia.org/
Linux Today - Linux News On Internet Time.
bletype.org : Get Movable Type Personal
Singapore Travel and Hotel Guide
Montana Local Guide, Hotels, Montana
Boing Boing
Alex King
Little&#39;s Journalized
Minnesota Encyclopedia : Maps - Weather - Travel - History
Infoportal
Pagina prima - Wikipedia
Main Page - Wikipedia, the free encyclopedia
Movable Type
Wikimedia Foundation
Slashdot
WordPress
The Register
Hauptseite - Wikipedia
Mozilla Firefox
Level Double-A Conformance to Web Content Accessibility Guidelines
Bollywood and the All India Encyclopedia
Online
Portada - Wikcionario
http://validator.w3.org/check?uri=referer
Artspace
Weather >
Esileht - Vikipeedia
Tom's Hardware Guide
Level A Conformance to Web Content Accessibility Guidelines 1.0
::: Welcome to KikuMobile.com :::
Asinah.net
W3C CSS Validator
HlavnÃ strana - Wikipedie
http://jigsaw.w3.org/css-validator
http://si.wikipedia.org/wiki/Main_Page
Wikibooks portal
Main Page - Wikipedia
W3C HTML Validation Service
West Virginia Encyclopedia : Maps - Weather - Travel - History
KezdÅ‘lap - WikipÃ©dia
Viewable With Any Browser
World Wide Web Consortium
PHP: Hypertext Preprocessor
Glavna stran - Wikipedija
Hafan - Wicipedia
Hoofdpagina - Wikipedia NL
GNU General Public License (GPL)
A List Apart
W3C HTML Home Page
Unang Pahina - Wikipedia
Free Software Foundation (FSF), GNU Project
SourceForge
Pagina principale - Wikipedia
Red Hat
GNU Free Documentation License
Web Standards Project
Web Content Accessibility Guidelines 1.0
Debian
Best Practical Solutions, LLC:
The Free Software Definition
http://www.cast.org/bobby/
http://www.fsf.org/
Geschenk Informationen zu
Kontaktlinsen Informationen zu
HTTP/1.0 200 OK
OpenOffice.org
The K Desktop Environment
Linux.org
Opera Software
Welcome to SUSE LINUX

World Wide Web

Hypertext

- **Hypertext**
  - Text with hyperlinks to other text.
  - Typically displayed on a computer screen or other electronic device.

- **Hyperlink**
  - Reference to data that the reader can follow via interaction.
  - Interaction is typically done using a mouse click, touching the screen or a keypress sequence.

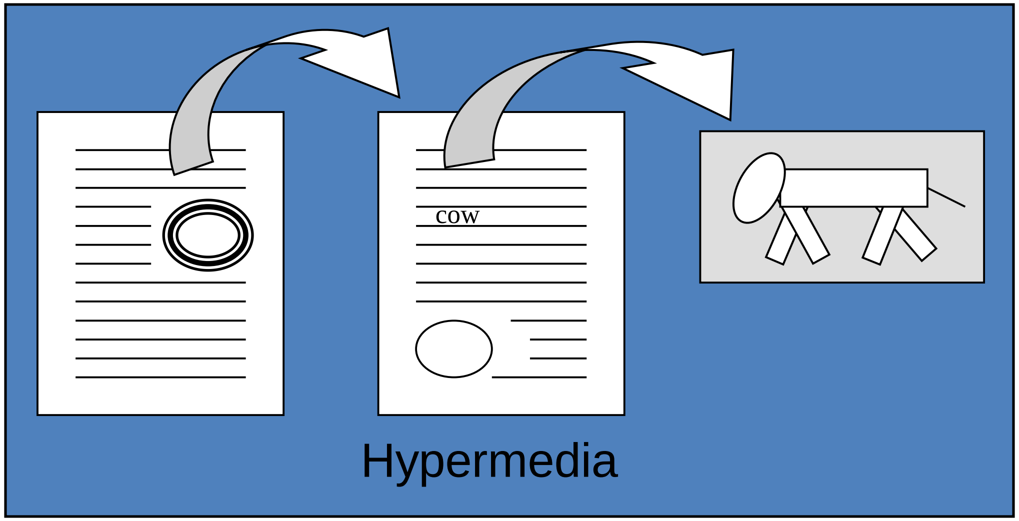


This is an example of some hypertext!

hyperlink

Multimedia and Hypermedia

- **Multimedia**
  - The integration of many forms of media
  - Text
  - Images
  - Sound
  - Animation



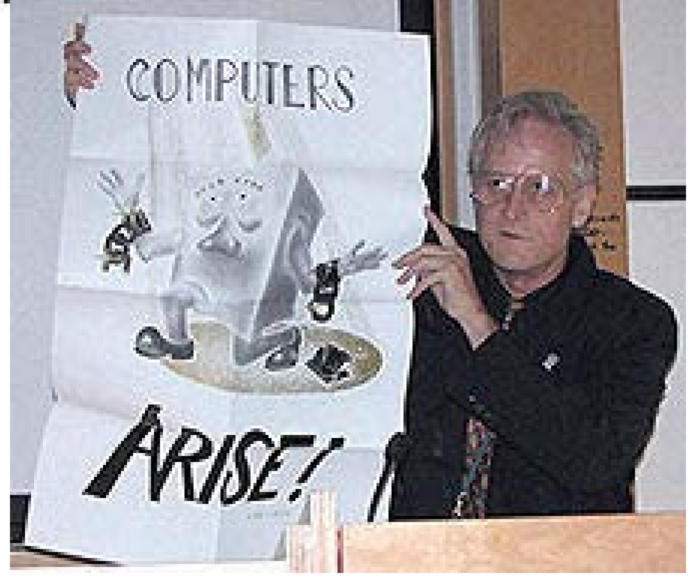
- **Hypermedia**
  - The combination of Hypertext and Multimedia
  - Hyperlinks are made between **any media**
  - Hypermedia and hypertext terms were coined by Ted Nelson

History

- **Vannevar Bush**
  - MEMEX system described in 1945 essay “As We May Think”.
  - Electromechanical device using microfilm for storage.
  - To be used to develop and read a large self-contained research library.

- **Ted Nelson**
  - Project Xanadu.
  - Envisioned as a “digital repository scheme for world-wide electronic publishing”.
  - First computer hypertext system.
  - First attempt at implementation began in 1960.
  - Incomplete implementation released in 1998

- **Tim Berners-Lee**
  - 1989 starts the WWW project at CERN



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

The WWW project

- **Background: CERN**
  - Many networks existed
  - Each network had many documents
- **Aims**
  - Access documents from any network in seamless manner
  - World-Wide (distributed)
  - Easy to add documents (dynamic)
- **Proposal**
  - Use Hypertext
  - No intention to support hypermedia
  - Research only

This image by Paul Clarke, http://www.flickr.com/people/34916866@N02, licensed under Creative Commons Attribution 2.0 Generic

Evolution of the web (1)

- **1989 Tim Berners-Lee begins work on the WWW project**
- **1991 WWW operational at CERN**
- **1992 WWW goes public**
- **1993 Mosaic created by Marc Andreessen (First GUI browser)**
- **1994 US Senate allow commerce on Internet**
- **1994 Netscape Communications formed, Yahoo! formed**

Evolution of the web (2)

- **1995 Microsoft Internet Explorer**

- **1998 Netscape became open-source, developed into Mozilla Google founded**

- **1997-2001 “Dot-com” boom and bust**

- **2002-on The web becomes ubiquitous**

Technical Details

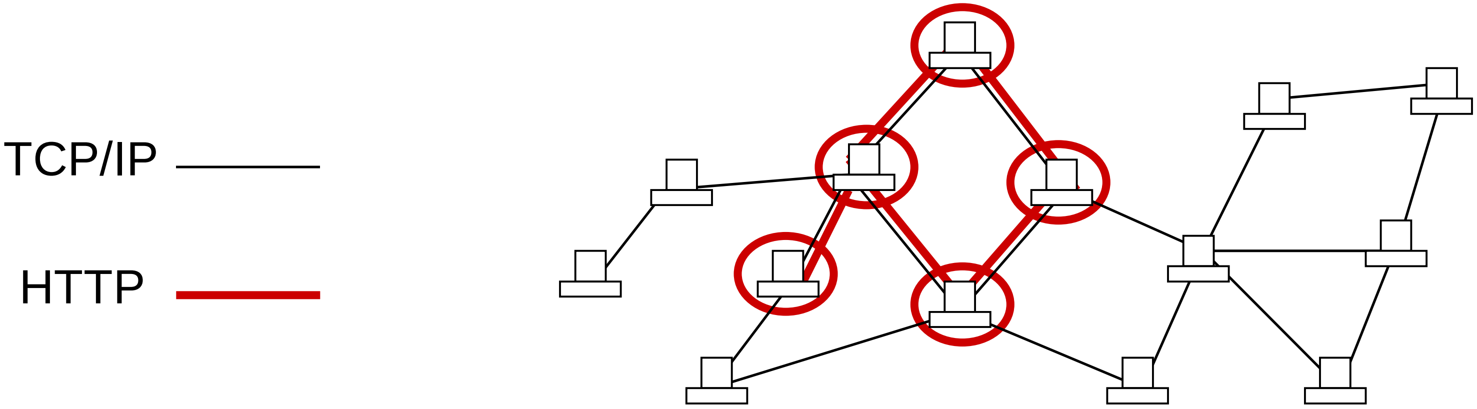
- **HTML**
  - Hypertext Markup Language
  - Language used to create Hypertext documents
  - Covered later on in course

- **HTTP**
  - Hypertext Transfer Protocol
  - Protocol used to transfer Hypertext documents
  - Client-Server Model

Technical Details

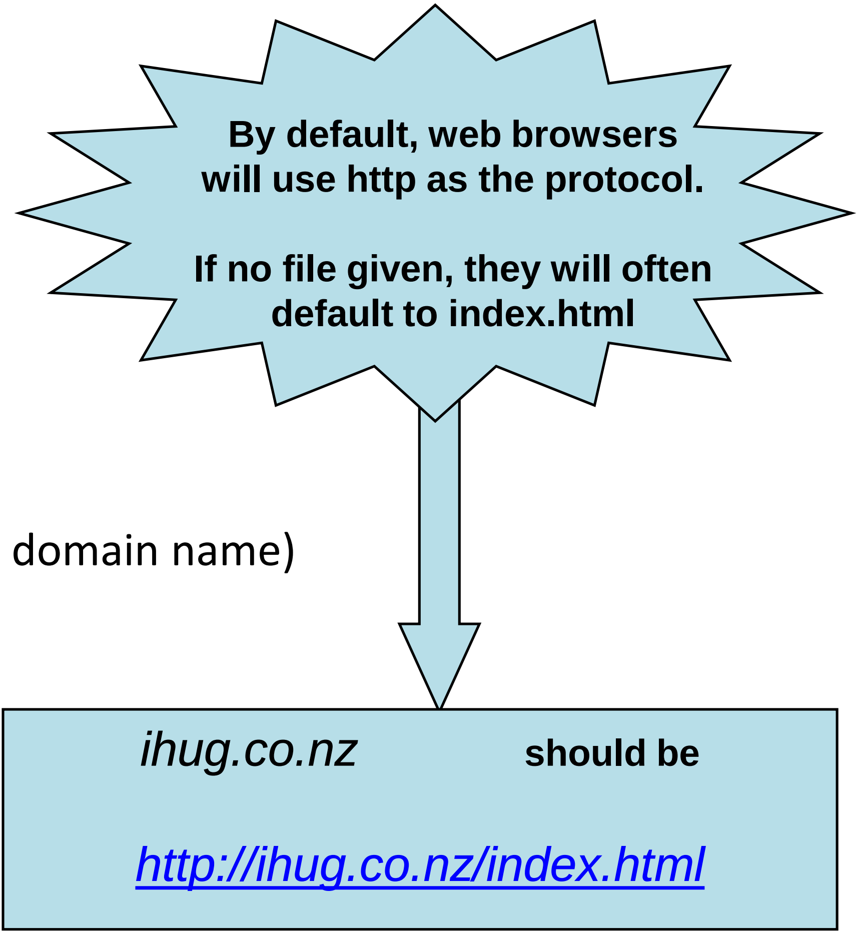
- **TCP/IP**
  - Ensures data is routed reliably (see lecture 4)

- **WWW**
  - Global body of information available using HTTP



Cyberspace Addresses

- **Uniform Resource Locators (URL)**
  - Address used for any web resource
- **Protocol**
  - Name of the protocol used
  - ftp:// http:// https://
- **Domain**
  - Name of a host computer (IP address or domain name)
  - www.cs.auckland.ac.nz
- **File/ Resource**
  - Path of the file
  - /Damir/LectureSlides.pdf



By default, web browsers will use http as the protocol.

If no file given, they will often default to index.html

ihug.co.nz **should be**

[*http://ihug.co.nz/index.html*](http://ihug.co.nz/index.html)

Terms

- **Web Site**
  - A collection of Web pages related to a single topic or theme. Normally designed and maintained by a single individual or organisation

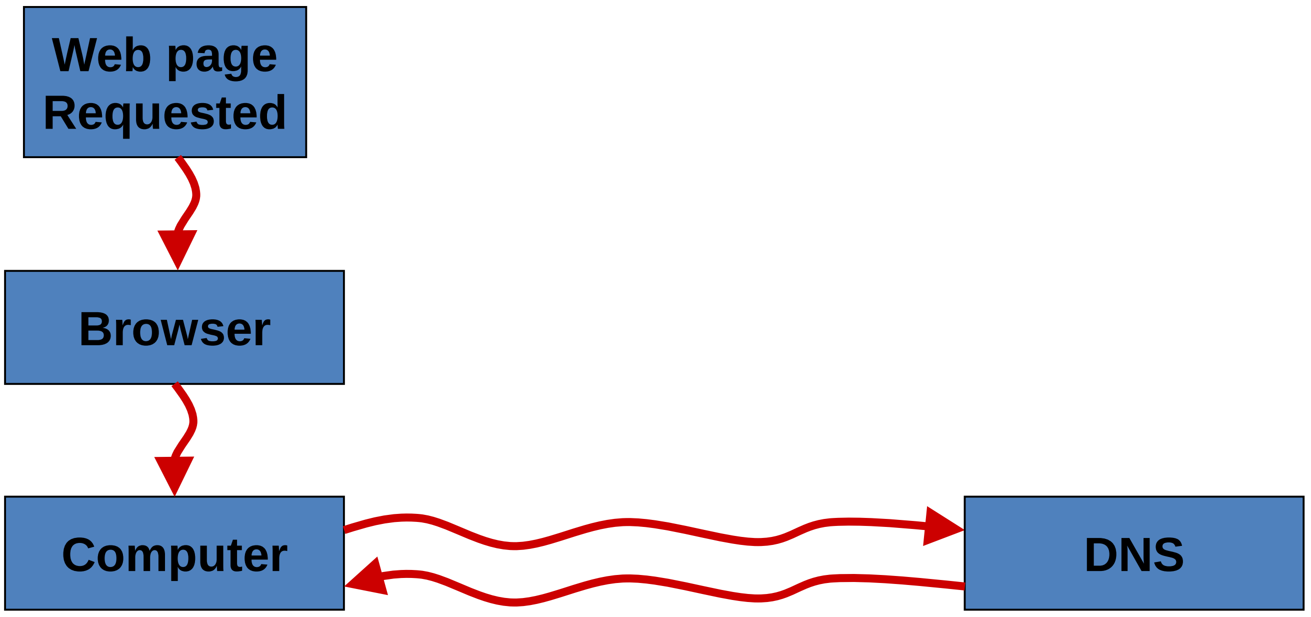
- **Web Page**
  - A hypermedia document designed for the WWW

- **Web Browser**
  - Software used to access information on the World Wide Web
  - Sends requests to a web server
  - Client

- **Web Server**
  - Software that makes local files available through the web
  - Fulfils requests from a web browser
  - Server

Accessing a web page (1)

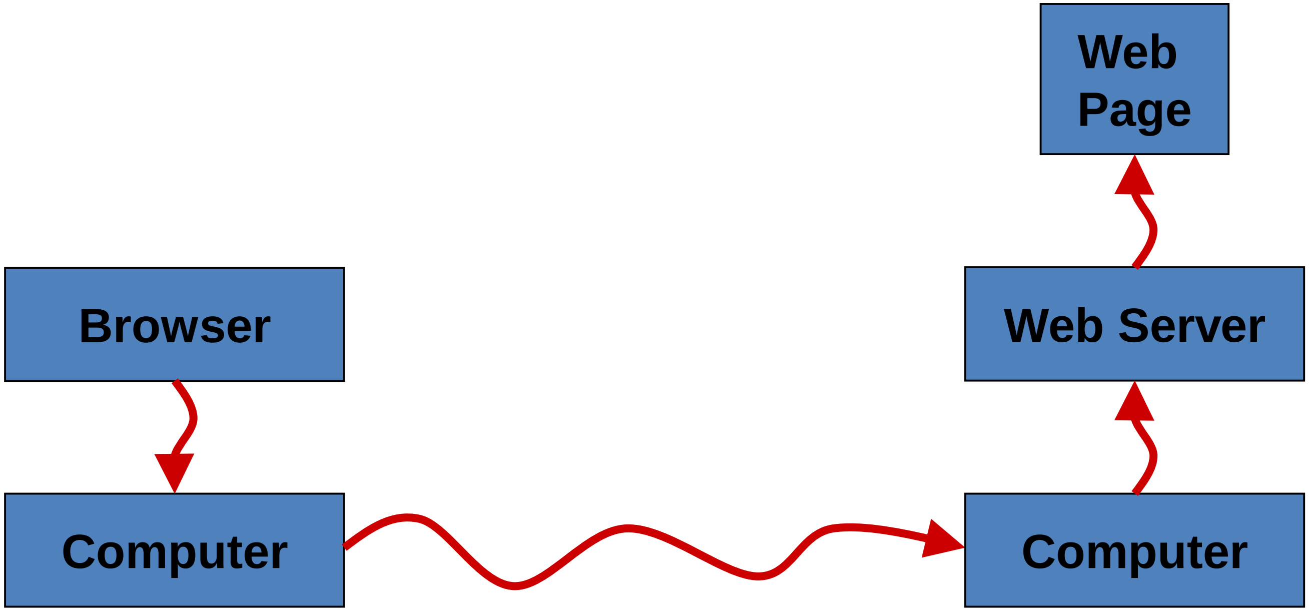
- **Client (Web Browser) runs on the local machine**
  - User requests a web page
  - Client contacts the DNS to find the IP address



DNS resolves the domain name

Accessing a web page (2)

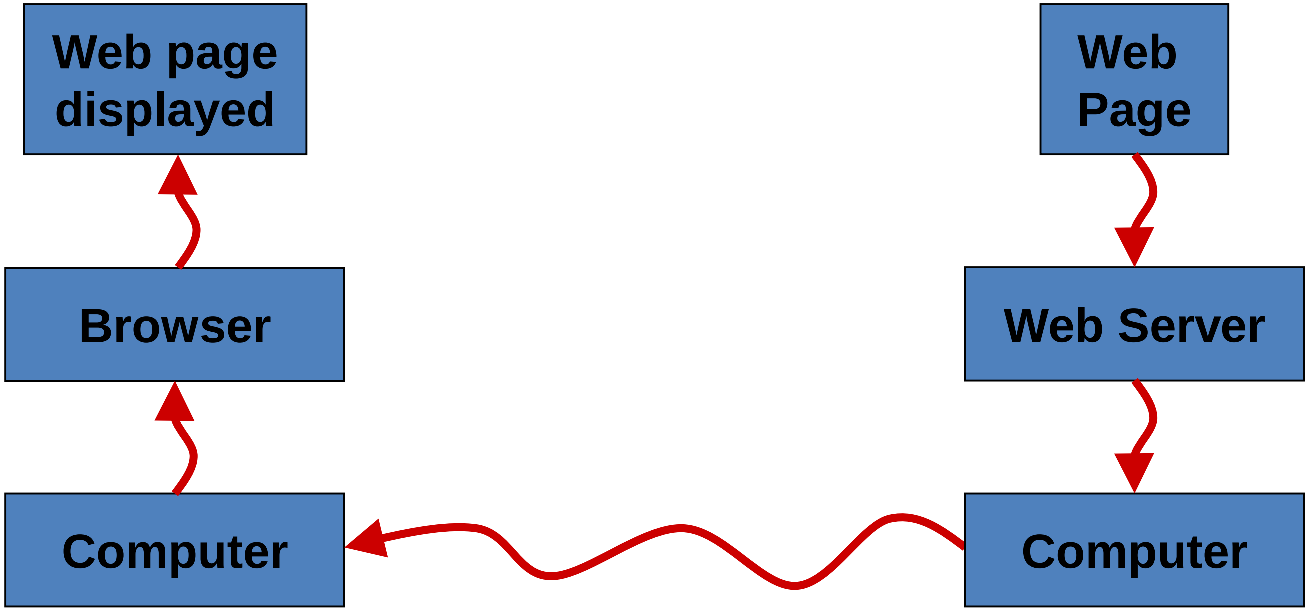
- **Web server runs on the destination machine**
  - Request sent to destination domain
  - Web server accepts the request and finds the web page



Web page requested from destination domain using HTTP

Accessing a web page (3)

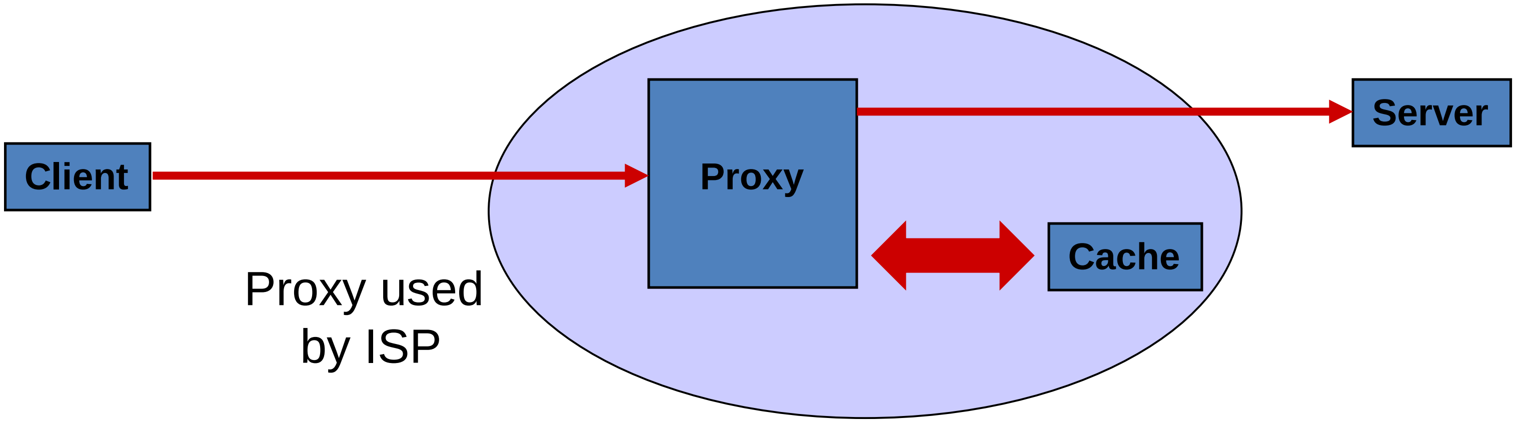
- **Web page is sent from the server to the client**
  - Client (web browser) displays the page



Web page sent from server to client using HTTP

More Terms

- **Proxy**
  - A computer which sits between the client and server, intercepts and processes requests
- **Cache**
  - Store of information for quick access
  - (e.g. caching may be used by proxy servers to speed web use)
- **Firewall**
  - Prevents unauthorised access to or from a private network



Logging web page access

- **Client keeps log**
  - History in web browser
- **Operating System keeps log**
  - Requests are logged by Windows on local machine
- **ISP keeps log**
  - Requests from "IP address" to "IP address" for "Page Name"
  - Some ISPs may have the logs available for users to check
- **Web server keeps log**
  - Gets requests from "IP address" for "Page Name”
- **Your viewing habits are being tracked!**

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

Navigating

- **Finding information**
  - Lots of users have problem finding new information
  - Lots of users have problems finding known information
  - Web is very large, rapidly changing

- **Search Engines**
  - Automated
  - Essential
  - Our gateway to information

Problems

Broken Links

- Pages which have been moved or deleted, but links are not updated.

No inherent security/ tracking/ accounting system

- Difficult to have layers of security
- Forces publishers to rely on advertising revenue

No inherent information indexing

- Much of the information is not accessed by search engines (e.g. encrypted, protected)
- Information created on-the-fly from databases
- Information in other formats (postscript, pdf, archived) may be missed

Search Engines

- **Companies (Worldwide Market Share 2014)**
  - Google (66.44%)
  - Baidu (China) (11.15%)
  - Microsoft Bing (10.29%)
  - Yahoo (uses Bing since 2009) (9.31%)
  - Specialised Alternatives
    - DuckDuckGo
    - Wolfram Alpha

- **Automatically search every web page**
  - Archive the contents
  - Index all the words
  - Try to determine the relevance of the page

http://en.wikipedia.org/wiki/Search_engines
http://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0

Can we trust the search engines?

- **Search Engines**
  - gateway to information
  - pages are rated (how?)
  - Since Ranking Algorithms are secret, we have to trust – but can we?

- **Publishers/ Advertisers**
  - Trick the search engines (repetition of words,)
  - Search engines tailor advertisements to searches
  - Pay for higher rating?
  - Who owns the search engines? How do they make their money?

- **Censorship**
  - 'Right to be forgotten' in the EU since June 2014
  - But even before that many requests for deletion (DMCA, local laws etc.)

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

Crawling the Web: Where do search engines get their information?

- **A ‘Web crawler’ is an internet bot that systematically browses the WWW and indexes encountered websites.**

- **Might store encountered websites for later processing.**

- **Start off with list of URLs and add any links encountered on these pages to their ‘To-Visit’ list.**

- **Follows a number of policies**
  - Selection: Only ‘important’ pages are indexed (2009: Large search engines index 40%-70% of indexable web, up from 16% in 1999)
  - Re-visit: When should the index for what page be updated, cost vs benefit.
  - Politeness: Crawlers are **really** good at getting lots of data quickly – they have to be careful not overload a website.
  - Parallelization: How do several crawlers split the task/web and recombine their results.

Searching

- **Searching Tips**
  - Learn how to use the advanced features of your search engine
  - If the first page is not promising, choose different key words
  - Some tips on searching with Google:

    http://www.otago.ac.nz/library/pdf/Google_searching.pdf

    http://www.google.co.nz/insidesearch/tipstricks/all.html

- **Finding useful sites**
  - Use specialist sites for specific searches
  - Build a list of useful resources:
    - Rotten Tomatoes
    - IMDB
    - IRD
    - Amazon

Google Top Trending 2014

NZ

- FIFA World Cup
- Robin Williams
- Commonwealth Games
- Malaysia Airlines
- iPhone 6
- Jennifer Lawrence
- Charlotte Dawson
- Flappy Bird
- Spark
- Ebola

Global

- Robin Williams
- World Cup
- Ebola
- Malaysia Airlines
- ALS Ice Bucket Challenge
- Flappy Bird
- Conchita Wurst
- ISIS
- Frozen
- Sochi Olympics

Google trends: topcharts NZ

Google Trends: topcharts Global

Google Top Trending 2015

NZ

- Agario
- Cricket World Cup
- Cyclone Pam
- Natalia Kills
- Jonah Lomu
- Google Classroom
- Lamar Odom
- Rugby World Cup
- Jerry Collins
- Caitlyn Jenner

Global

- Lamar Odom
- Charlie Hebdo
- Agar.io
- Jurassic World
- Paris
- Furious 7
- Fallout 4
- Ronda Rousey
- Caitlyn Jenner
- American SNiper

Google trends: topcharts NZ

Google Trends: topcharts Global

Google News

- **News aggregator, variation of the search engine**
- **Automatically searches thousands of publications and displays summaries, relevant parts. Examples:**

  Where in Ukraine is Viktor Yanukovych?

  - Yanukovych's exact whereabouts remained unknown
  - Yanukovych surfaced Saturday in the city of Kharkiv

  Robots will be smarter than us all by 2029, warns Google futurologists

  - computers will be able to understand our language, learn from experience
  - By 2029 they will outsmart even the most intelligent humans, according to Google's director of engineering Ray Kurzweil.

- **Many Publishers/News Agencies unhappy**
  - Google reuses (snippets of) their content
  - Shut down in Spain in December 2014, where new law requires payment for reuse

(Online) innovations

- **Voice over IP**
  - Google Hangouts, Skype, ISPs
  - Cheap/free voice communication
- **Peer to Peer networks**
  - BitTorrent
  - Swarming downloads
- **Wolfram**
  - WolframAlpha: searching = computing
  - Wolfram language: knowledge-based programming
- **Free Books**
  - http://digital.library.upenn.edu/books/
  - http://books.google.com
- **Internet for everybody anywhere**
  - Google Project Loon (http://www.google.com/loon/): using high-altitude balloons to create a wireless network that provides internet in rural and remote areas.
  - Outernet (https://www.outernet.is/en/): free internet anywhere in the world through satellites

Web-agents and other future directions

- **Computer programs that operate on your behalf**
  - Tracks all your browsing habits
  - Makes suggestions based on what you have read
  - Recommender Systems: Big, active research area, permeates many areas (shopping, video streaming, search)

- **TiVo Suggestions, Netflix recommendations**
  - Similar principle with television viewing
  - Netflix Prize (2006-2009), $1 Million: Improve Netflix' own algorithm for predicting user ratings for movies based on previous ratings by 10%.

- **The Internet is changing extremely rapidly**
  - Too fast for legislation to keep up
  - Too fast to predict the future

- **Some things coming soon**
  - Wearable PC's
  - Integrated Media (Interactive T.V.)
  - Household Appliance connections
  - And of course: new approaches to internet-related crime.