First SIAM International Conference on Data Mining
5 April 2001

Tutorial on E-commerce and Clickstream Mining

Jonathan Becher
VP, Product Strategy
Accrue Software, Inc.
jonbecher@yahoo.com

Ronny Kohavi
Director, Data Mining
Blue Martini Software
ronnyk@CS.Stanford.edu
http://www.Kohavi.com

Agenda

- Introduction (45 min)
- Architecture and Data Flow (45 min)
  - Collecting the data
  - Break (10 min)
  - Building the warehouse
  - Closing the loop
- Mining Web Data (75 min)
  - Transformations
  - Unofficial Break (10 min)
  - Reporting and OLAP
  - Mining
  - Visualization
- Teasers & Summary (15 min)

Introductions - Who are We?

- Ronny Kohavi
- Jon Becher
- Audience
  - How many from Academia vs. Vendor vs. Site?
  - How many analyzed clickstream data?
  - How many analyzed transactional data?
  - How many collect web-based data today?
- Logistics: bathroom is …
- Questions? Special requests?

Web Mining: Site Categories

- **Brochureware - simplest sites**
  - **Mostly static brochure content**
  - **About <company>**
  - **Examples: Exxon Mobil, Philip Morris**
- **Content Providers - dynamic content Communities, Portals, Aggregators**
  - **High conversion rates to members (over 50%) for repeat visitors †**
  - **Low ad revenue per visitor (less than $0.50)**
  - **Subscription revenues are rare**
  - **Examples: Yahoo!, CNN, Levi's, Wall Street Journal**

† Stats from E-performance:The path to rational exuberance, McKinsey Quarterly, 2001, No 1

Web Mining: Site Categories II

- **Transaction oriented sites**
  - **Sell items**
  - **Conversion rates (browsers to shoppers) around 2%**
  - **Revenue per customer around $150/month (high average includes travel sites) †**
  - **Visitor acquisition cost $1-$5 (=$50-$200 / customer)**
  - **Examples: Amazon, Dell**
- **Data Mining is most important for transaction sites and content providers**

† Stats from E-performance:The path to rational exuberance, McKinsey Quarterly, 2001, No 1

What is (not) Covered

- **Both Jon and Ronny have more experience in B2C (Business to Consumer) clients, although most principles apply to B2B**
- **We will not cover Information retrieval and network management.**
  Rajeev Rastogi and Minos Garofalakis will cover in tomorrow's tutorial
- **Disclaimer: we will mention books, products, and URLs that we found useful. This is not a comprehensive list**
- **Vendor slides are attached in the beginning**

Value Proposition

- **Why mine e-commerce and clickstream data?**
  - **Improve conversion rate through personalization**
  - **Optimize marketing campaigns (banners, email, other media) that bring visitors to your site by measuring return on investment (ROI).**
  - **Improve basket size through cross-sells and up-sells**
  - **Streamline navigation paths through the site**
  - **Avoid content delivery issues (poorly formatted for AOL, too rich for low bandwidth users, redundant or confusing content)**
  - **Identify customers segments that you can target offline**
  - **Experiment quickly. The Web is a laboratory. Understand what works quickly**

Web is DM's Killer Domain

- **Successful data mining benefits from:**
  - **Large amount of data (many records)**
  - **Rich data with many attributes (wide records)**
  - **Clean data collection (avoid GIGO)**
  - **Actionable domain (have real-world impact)**
  - **Measurable return-on-investment (did the recipe help?)**
- **Web mining has all the right ingredients**

Definitions

- **Hit** – any Web server request that generates a log file entry. A page has many elements (html, gifs), each generating a hit.
- **Page** – Web server file that is sent to client user agent, usually a browser. Typically HTML files, but not all HTML are considered pages (I.e., frame set). Can be static or dynamic
- **Session** – all actions (i.e. requests, resets) made in single visit, from entry until logout or time out (e.g., 20 minutes of no activity).
- **Visitor** – a user or bot/spider/crawler that makes requests at a site. Can be new, returning, registered, anonymous
- **Buyer** – visitor that purchases something
- **Customer** – a visitor that registers (sometimes defined as buyer)
- **Conversion** – rate at which visitors transition to desired state (buyers, customers, registered, started checkout)
- **Host** – remote machine, identified by IP address, used for visit.
- **Referrers** – page that provides a link to another page. Can be internal or external

Teaser - Page Definitions

A user visits Yahoo to find out what the weather in Chicago will be next week.

weather.yahoo.com

The weather map image for Chicago is dynamically loaded from another site, when needed.

www.weathernews.com

Clearly there was a page view at Yahoo, but was there also a page view at Weathernews? How about a hit? A visit?

Teaser - Conversion

- **Product conversions are computed as**

  rate = “Product quantity sold” / “Number of product views”

- **How can conversion rates be above 100%**

Case Study: KDD Cup 2000

- **Gazelle.com was a legcare and legwear retailer**
- **Data available for KDD Cup 2000**
- **Data enhanced with Acxiom demographics**
- **See** `http://www.ecn.purdue.edu/KDDCUP` **for details and access to data**

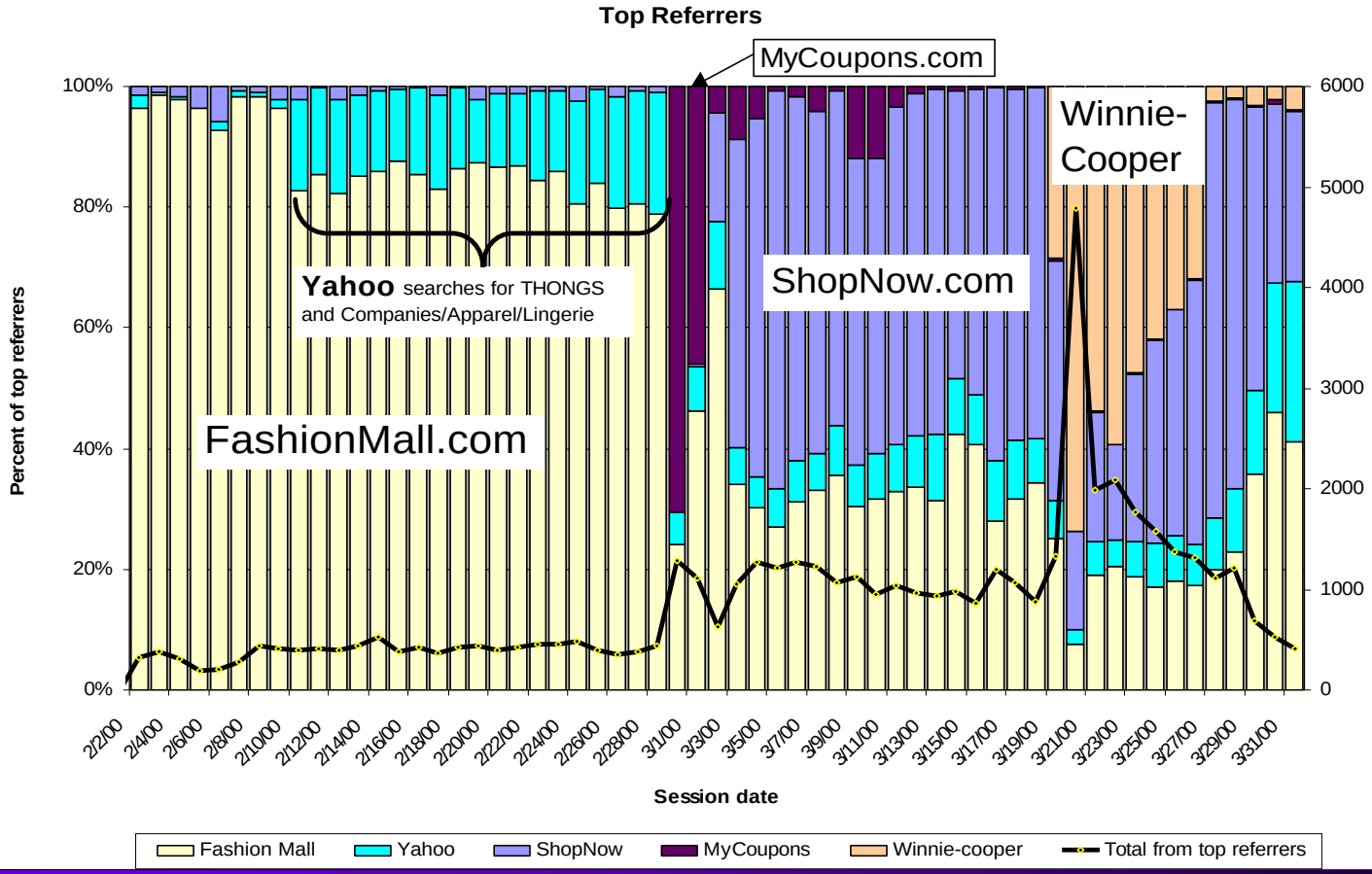
Heavy Purchasers

- **Factors correlating with heavy purchasers:**
  - **Not an AOL user (defined by browser) - browser window too small for layout (inappropriate site design)**
  - **Came to site from print-ad or news, not friends & family - broadcast ads versus viral marketing**
  - **Very high and very low income**
  - **Older customers (Acxiom)**
  - **High home market value, owners of luxury vehicles (Acxiom)**
  - **Geographic: Northeast U.S. states**
  - **Repeat visitors (four or more times)-loyalty, replenishment**
  - **Visits to areas of site - personalize differently**

Referring Traffic

Referring site traffic changed dramatically over time.

Graph of relative percentages of top 5 sites



Referrers - Ad Policy

- **Referrers - establish ad policy based on conversion rates, not clickthroughs!**
  - **Overall conversion rate: 0.8% (relatively low)**
  - **Mycoupons had 8.2% conversion rates, but low spenders**
  - **Fashionmall and ShopNow brought 35,000 visitors Only 23 purchased (0.07% conversion rate!)**
  - **What about Winnie-Cooper?**

Who is Winnie Cooper?

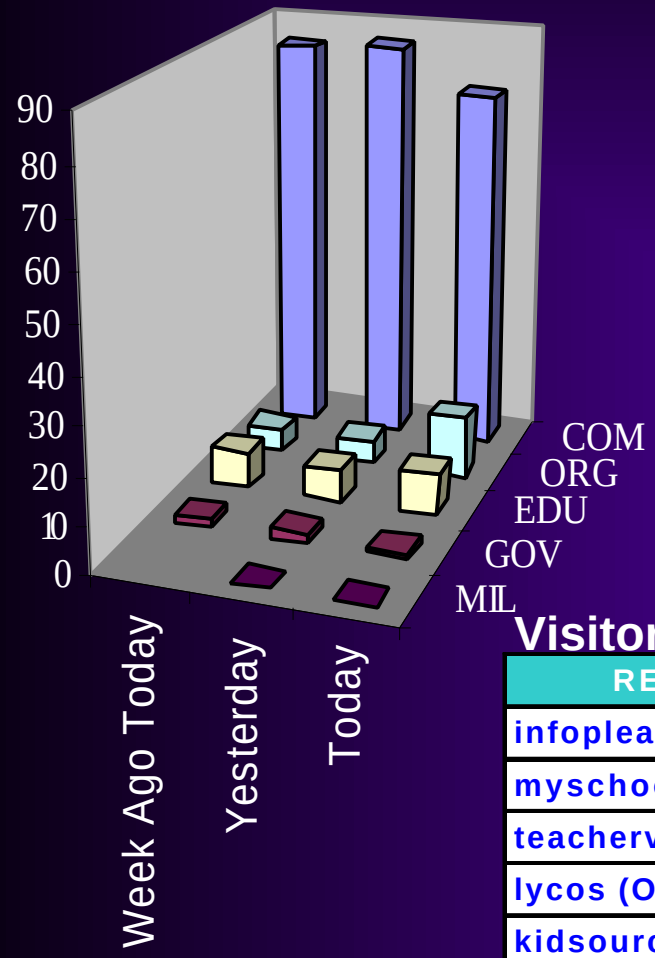
- **Winnie-cooper is a 31 year old guy who wears pantyhose**
- **He has a pantyhose site**
- **7000 visitors came from his site**
- **Actions:**

  - Make him a celebrity and interview him about how hard it is for a men to buy pantyhose in stores
  - Personalize for XL sizes

Case Study: On-line Newspaper

- **Regional newspaper focused on editorial content, classifieds, “yellow pages”, and syndicated content from third party providers**
- **Goals:**
  - **Increase traffic to increase advertising revenue (acquisition)**
  - **Increase percentage of registered users (conversion)**
  - **Increase pages/visits and visits/visitor (stickiness)**
  - **Deliver more targeted content to registered users**

The War Effect



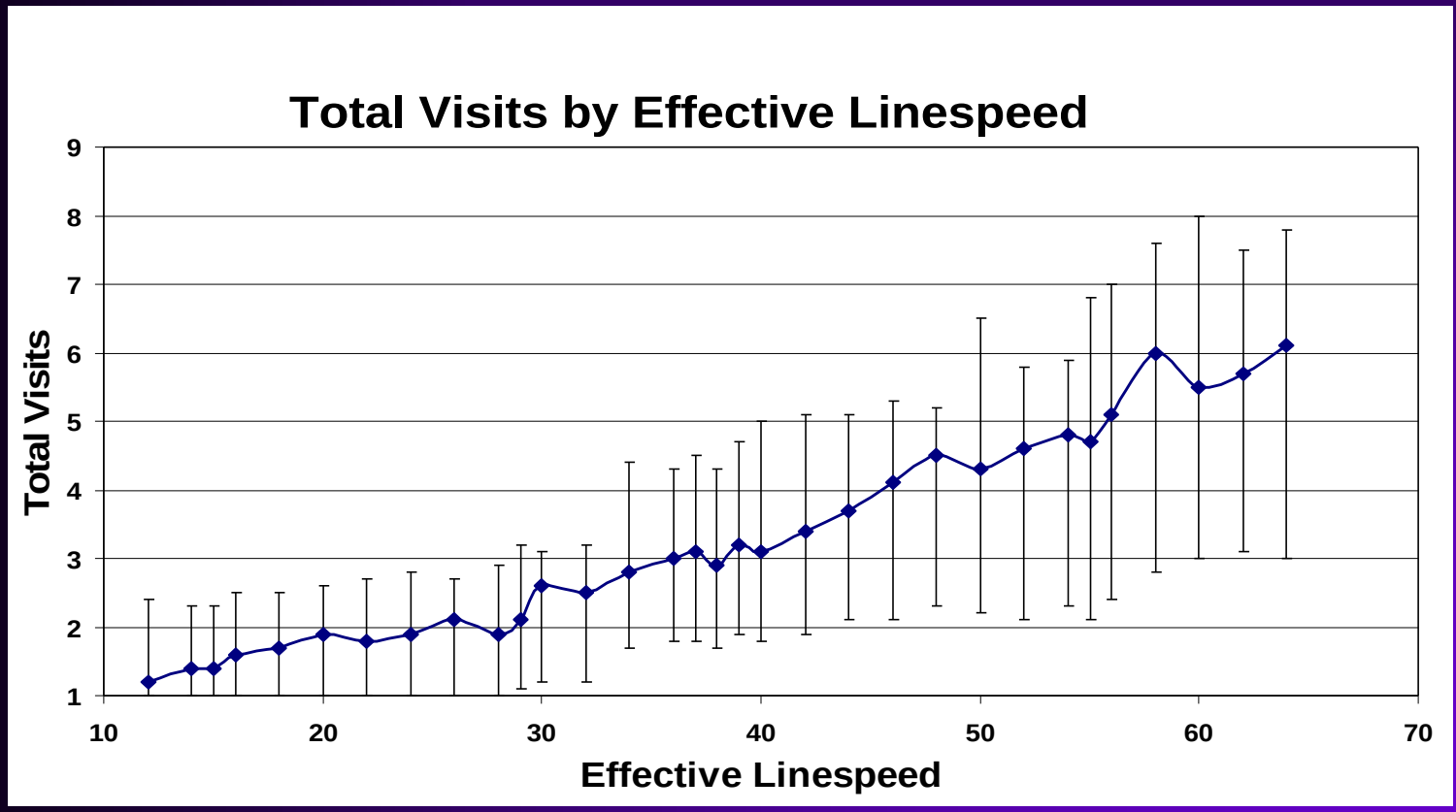
- When US launched its campaign in Serbia, site put up special section with links to past stories on Kosovo
- ⇐ Dramatic single day shift in mix of visitor domains to EDU and ORG
- ⇓ Biggest increase in referrers from education and teaching sites.
- **Conclusion**: outreach programs to classrooms based on special events

Visitor Sources: Biggest Increases

REFERRER	Today	Yesterday	Variance	Pct Variance
infoplease (ORG)	5,013.00	3,580.00	1433.00	40.03
myschoolonline (ORG)	21,719.00	20,933.00	786.00	3.75
teachervision (ORG)	2,066.00	1,765.00	301.00	17.05
lycos (ORG)	266.00	207.00	59.00	28.50
kidsource (ORG)	266.00	214.00	52.00	24.30
familyeducation (ORG)	616.00	575.00	41.00	7.13
awesomelibrary (ORG)	173.00	136.00	37.00	27.21

The Bandwidth Effect

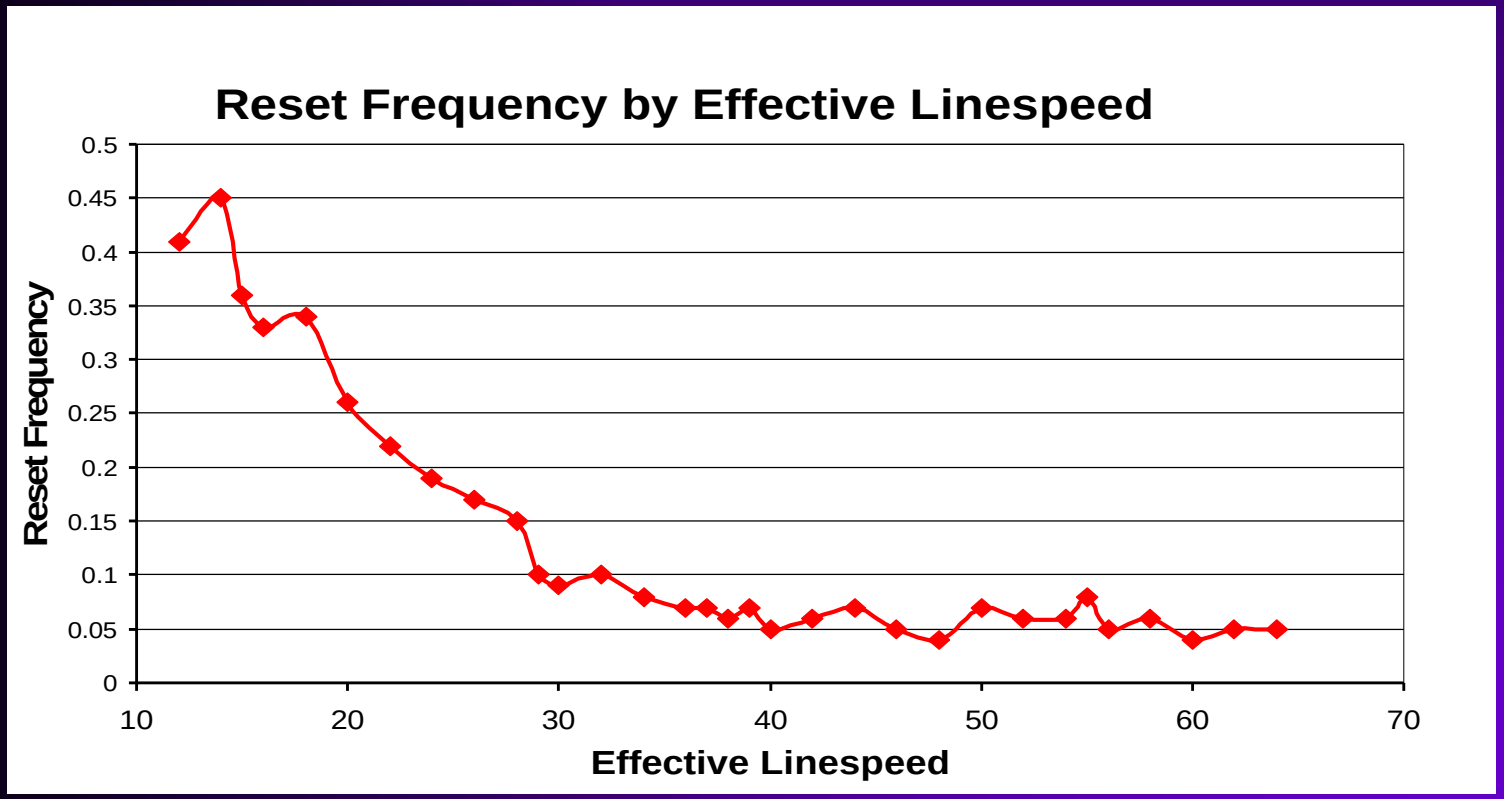
- **Users with high effective line speed are more likely to be return visitors**



Bars represent one standard deviation from average

The Bandwidth Effect II

- **Users with low effective linespeed connections are much more likely to give up on a page before it's done**
- **Conclusion: 1) two versions of the site, one with less rich graphics 2) use HTML instead of PDFs**



The Referrer Effect

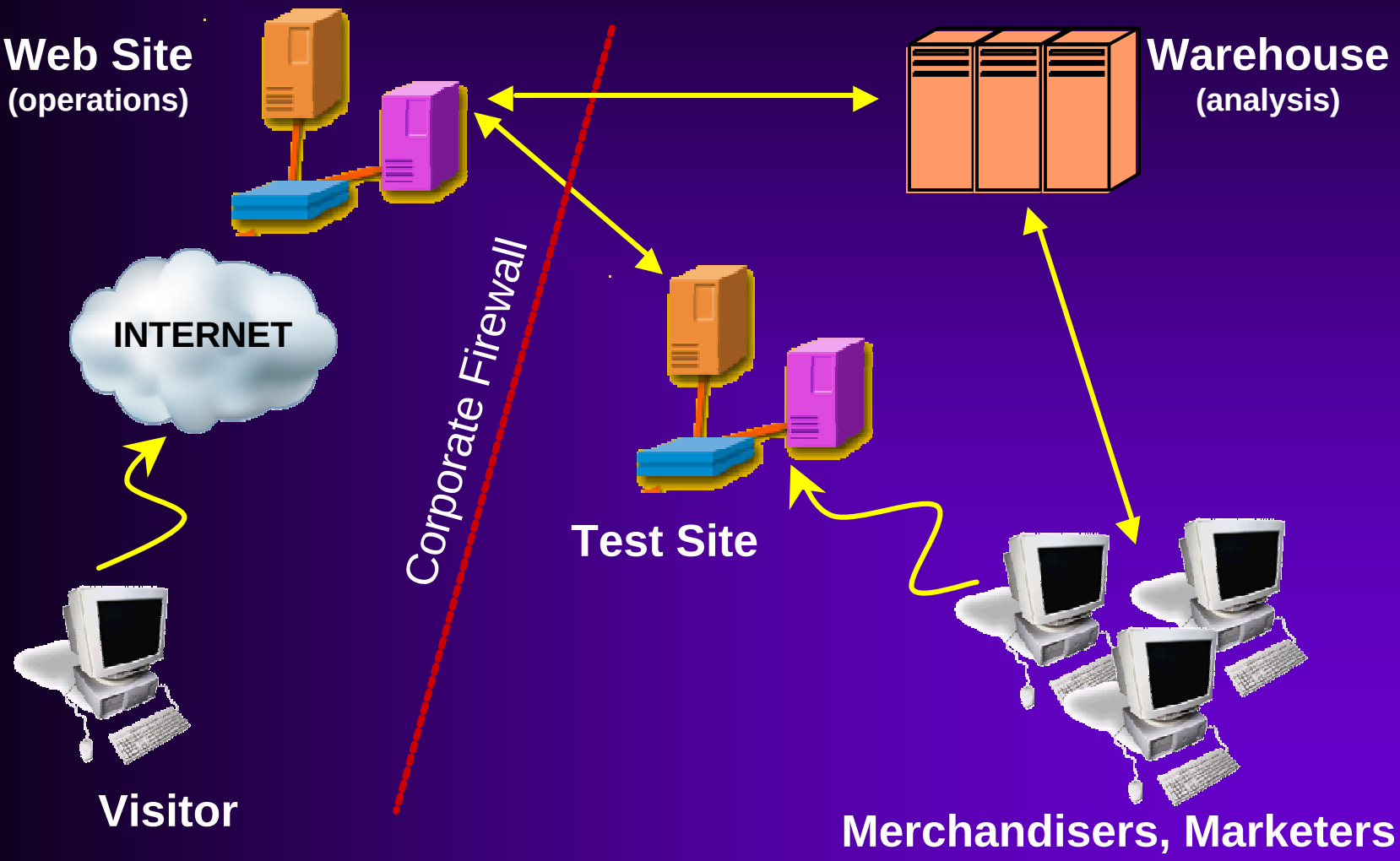
- Check on stickiness of the site based on the location of the referrer reveals visitors from banner ads, search engines, and portals have shallow visits
- Best results come from affiliates – content partners that share similar demographics
- Worst: banner advertising – almost no one looks at any pages beyond the initial redirect

	0 Pages	1 Page	2 Pages	3-5 Pages	6-10 Pages	11-25 Pages	26+ Pages
doubleclick (ORG)	210,175	6,941	1,422	1,217	1,170	804	479
yahoo (COM)	132,719	12,846	10,159	14,696	15,482	13,139	9,274
familyeducation (ORG)	103,942	116,371	11,357	13,252	16,352	19,749	26,396
myschoolonline (ORG)	97,066	10,225	10,575	9,842	14,228	19,839	22,343
lycos (COM)	38,967	3,628	476	289	265	149	77
google (COM)	11,623	3,265	615	387	257	145	114

Agenda

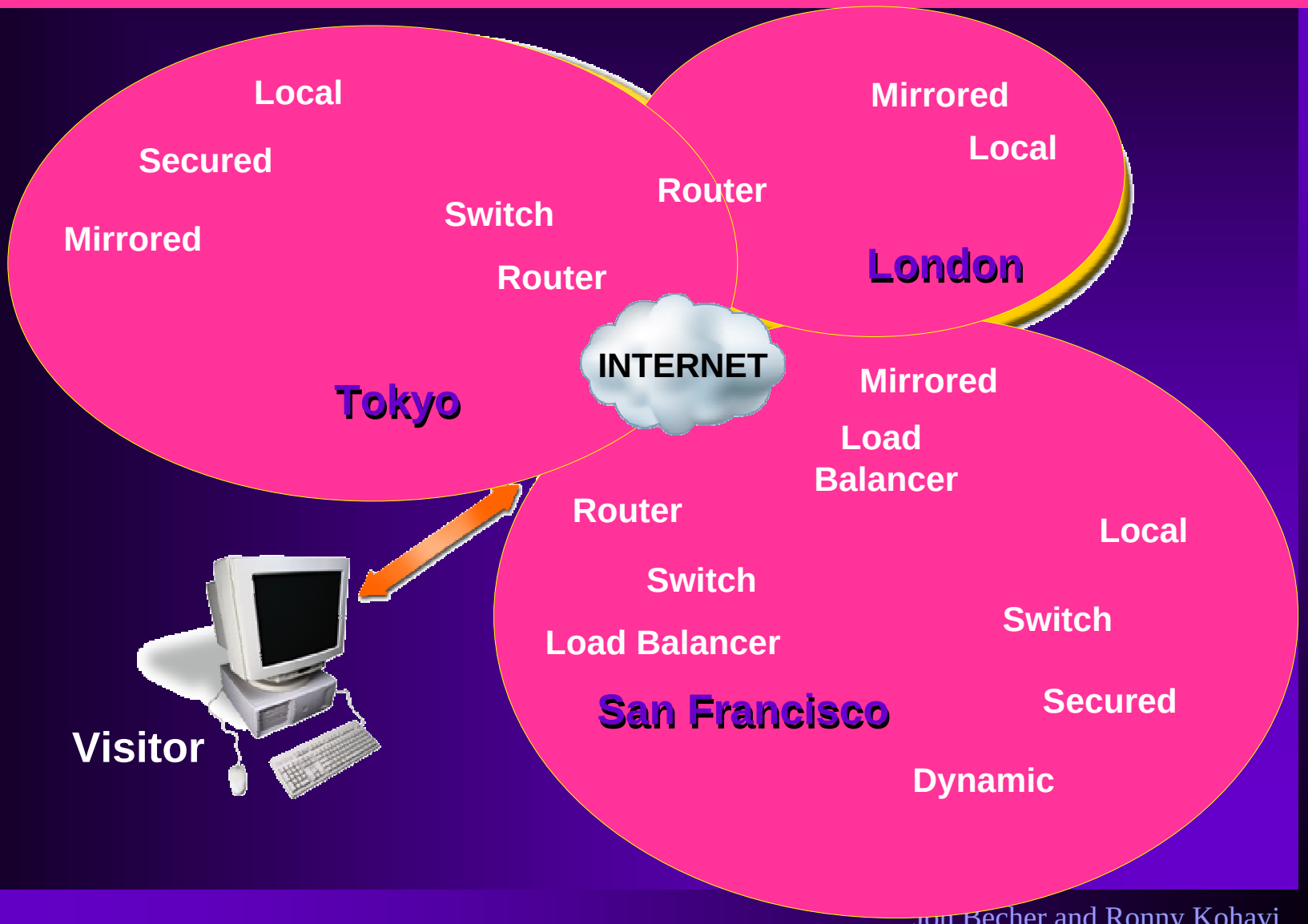
- Introduction (45 min)
- Architecture and Data Flow (45 min)
  - Collecting the data
  - Building the warehouse
  - Closing the loop
- Break (10 min)
- Mining Web Data (75 min)
  - Transformations
  - Reporting and OLAP
  - Mining
  - Visualization
- Summary (20 min)

Architecture



Web Site Topology

Need for scalability causes complexity of design



Data Collection

- **Visitor activity information**
  - **Web server log files**
  - **Web server instrumentation (plug-ins)**
  - **TCP/IP packet sniffing (network collection)**
  - **Application server instrumentation**
- **Other sources of data**
  - **Transactions**
  - **Marketing programs (banner ads, emails, etc)**
  - **Demographic (registration, third party overlay)**
  - **Call center (WISMO)**
  - **Supply chain (inventory and fulfillment)**

Collection: Server Log Files

- **Advantages**
  - **Everyone has got one**
  - **Useful for specialized data types (e.g. streaming media)**
- **Disadvantages**
  - **Multiple file formats (elf)**
  - **Designed for debugging Web servers, not for analysis**
  - **Multiple log files for multiple Web servers**
  - **Distributed sites make sessionizing more difficult**

Request date	Request time	Request IP address	Auth. user name	Server name	Computer name	Server IP address	Method	Requested file name	Query string	Error code	Bytes received	Bytes sent	Time taken	Version	Agent	Cookie	Referring page
7/19/98	16:07:10	193.237.55.144	-	W3SVC1	WEBSERVER	206.129.192.10	GET	/Default.htm	-	304	204	379	330	HTTP/1.0	Mozilla/4.04+[en]+(Win95;+I)	-	http://www.uu.se/Software/Ar
7/19/98	16:07:10	193.237.55.144	-	W3SVC1	WEBSERVER	206.129.192.10	GET	/navbar.htm	-	304	147	389	251	HTTP/1.0	Mozilla/4.04+[en]+(Win95;+I)	-	http://www.uu.se/Software/Ar
7/19/98	16:07:10	193.237.55.144	-	W3SVC1	WEBSERVER	206.129.192.10	GET	/frames-default.htm	-	304	147	396	290	HTTP/1.0	Mozilla/4.04+[en]+(Win95;+I)	-	http://www.uu.se/Software/Ar
7/19/98	16:07:10	193.237.55.144	-	W3SVC1	WEBSERVER	206.129.192.10	GET	/graphics/hproducts.gif	-	304	146	373	390	HTTP/1.0	Mozilla/4.04+[en]+(Win95;+I)	-	http://www.marketwave.com/r
7/19/98	16:07:10	193.237.55.144	-	W3SVC1	WEBSERVER	206.129.192.10	GET	/graphics/haboutus.gif	-	304	146	372	290	HTTP/1.0	Mozilla/4.04+[en]+(Win95;+I)	-	http://www.marketwave.com/r
7/19/98	16:07:10	193.237.55.144	-	W3SVC1	WEBSERVER	206.129.192.10	GET	/graphics/hordering.gif	-	304	146	373	310	HTTP/1.0	Mozilla/4.04+[en]+(Win95;+I)	-	http://www.marketwave.com/r
7/19/98	16:07:10	193.237.55.144	-	W3SVC1	WEBSERVER	206.129.192.10	GET	/graphics/hreseller.gif	-	304	146	373	271	HTTP/1.0	Mozilla/4.04+[en]+(Win95;+I)	-	http://www.marketwave.com/r

Collection: Server Plug-ins

- **Advantages**
  - **Allows for pre-processing of data before storage**
  - **Can automate scheduling of data to analysis server**
- **Disadvantages**
  - **No incremental data than available from log file**

Collection: Packet Sniffing

- **Advantages**
  - **Additional information available**
    - timing (server response, page download, packet roundtrip)
    - browser resets (stop button, move on before load)
  - **Any Web server can be supported**
  - **Data can be captured in real time**
  - **Multiple Web servers are handled as one**
  - **Reduces load on Web servers**
- **Disadvantages**
  - **Cannot handle encrypted traffic (SSL)**
  - **Does not capture sub URL information**

Collection: Application Servers

More e-commerce sites now employ application servers, which control logic and allow logging

- **Advantages**
  - **Can provide information sub page info (product shown, assortment if multiple products, promotion, ads, prices, etc.)**
  - **No issues sessionizing (app server controls sessions)**
  - **Can log events at higher levels than URLs**
    - completing a scenario (registration, checkout)
    - form information, such as search keywords
  - **Clickstream and purchase transactions share Ids**
  - **Robust to changes in URLs**
- **Disadvantages**
  - **Must work with an application server and design it properly**
  - **Does not capture network effects**

Collection: Other Sources

- **Advertising networks**
  - **Which banner ads on which sites cause the best traffic?**
  - **e.g., Angara, Doubleclick, Engage, Matchlogic, MediaPlex**
- **Campaign management products**
  - **Which marketing campaigns are bringing the most qualified visitors to your site?**
  - **e.g., Annuncio, Blue Martini, MarketFirst, Prime Respone, Unica, Xchange**
- **Commerce/transactional engines**
  - **Which products are most likely to be abandoned on the weekend?**
  - **e.g., ATG Dynamo Commerce, BEA Weblogic, Broadvision, IBM Websphere Commerce, OpenMarket Transact**
- **Overlay data providers**
  - **How do visitors' psychographic and demographic information correlate with their Web site browsing behavior?**
  - **e.g., Acxiom, Experian, InfoUSA, Nielson**

Advertising Analysis

How effective are banner ads?

Report: Ad by Content Preference						
Ad Name	Content Group	Visitors	Visitor Yield	New Visitors	Cost Per Visitor
Mustang	Financing	1179	44%	5.7%	$1.60
	Auto Ratings	533	20%	4.0%	$3.24
	Safety Info	433	16%	3.3%	$3.99
	Repair History	363	13%	6.7%	$4.76
	Used Cars	191	7%	21.6%	$9.05
	Total Ad	**2699**	**100%**	**5.7%**	**$3.26**
Corvette	Financing	1009	41%	3.1%	$1.71
	Auto Ratings	502	21%	1.8%	$3.44
	Safety Info	441	18%	3.3%	$3.01
	Repair History	291	12%	4.2%	$5.93
	Used Cars	191	8%	11.3%	$9.03
	Total Ad	**2434**	**100%**	**3.0%**	**$3.54**

Compare the effectiveness of ads at driving traffic to different areas of the site

Compare the effectiveness of ads at driving traffic from different external sites

Report: Impressions to Explorations						
Site Name	Ad Name	Impressions	Click-On Rate	Visitor Yield	Page Depth	Time (secs)
Portal 1	Mustang	341	6.2%	6.0%	1.2	256
	Sebring	346	4.0%	4.0%	1.7	314
	Corvette	921	3.4%	3.3%	2.9	563
	Intrigue	643	3.3%	3.3%	3.5	419
	Camaro	937	1.6%	1.5%	2.1	401
Portal 2	Mustang	98	3.1%	3.1%	6.4	772
	Corvette	106	2.0%	1.8%	4.3	456
Portal 3	Corvette	34	35.0%	22.0%	3.8	421
	Camaro	33	6.3%	4.3%	2.9	398
	Sebring	59	3.5%	2.9%	4.5	489

Agenda

- Introduction (45 min)
- Architecture and Data Flow (45 min)
  - Collecting the data
  - Building the warehouse
  - Closing the loop
- Mining Web Data (75 min)
  - Transformations
  - Unofficial break (10 min)
  - Reporting and Visualization
  - OLAP
  - Mining
- Summary (20 min)

Data Storage

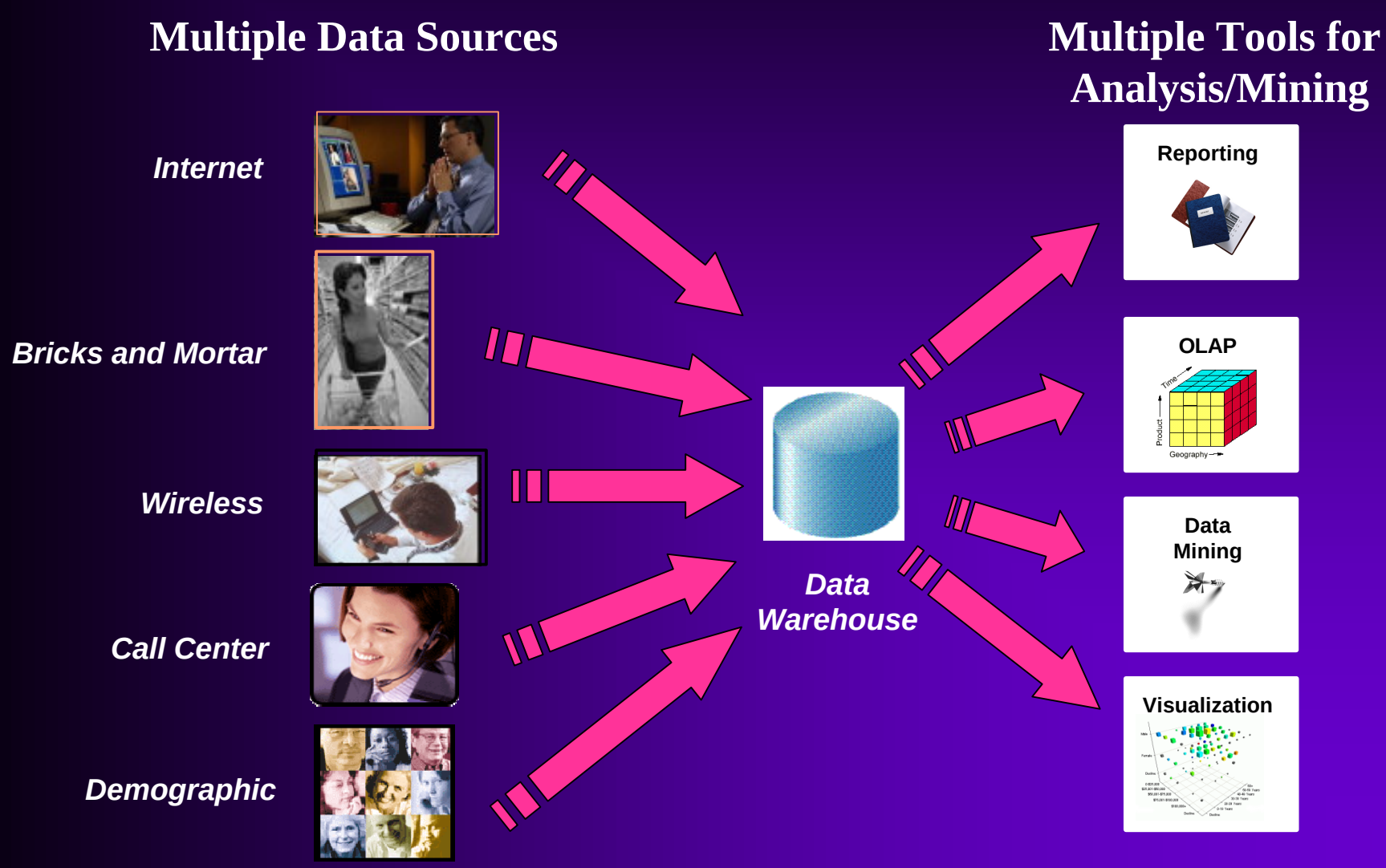
Operational vs. Analytical Storage

- **Decision support (data warehouse) has different needs than a transaction OLTP system**

Analytical	Operational
Few large transactions	Many small transactions
Customer centric	Session and product centric
Hard to parallelize	Easy to parallelize (multiple web/app servers)

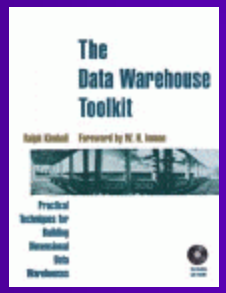
- **Tuning Oracle to perform well in a warehouse is not like tuning it for an operational system**

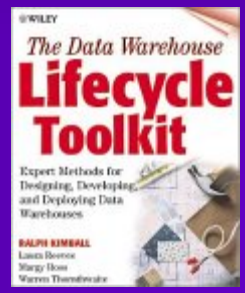
Building the Data Warehouse

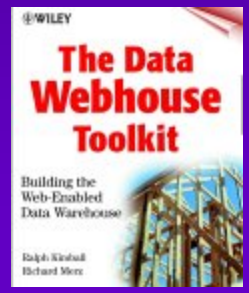


Extract Transform Load (ETL)

- **Building a data warehouse is a complex process involving data migration, consolidation, cleansing, transformations, and meta-data creation/transfer**
- **Use ETL tools such as Informatica, Data Junction, Sane's NetTracker for weblog data**
- **Resources:**
  - **Ralph Kimball's books**
  - `http://www.informatica.com`
  - `http://www.datajunction.com`
  - `http://www.sane.com`







Alternatives to Data Warehouse

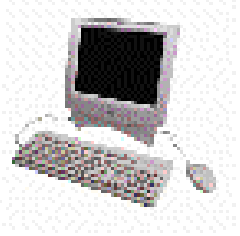
- **Simple models can be computed efficiently at the touchpoint (e.g., webstore)**
  - **Top items (easy to increment counters)**
  - **Item pair associations (people who bought this book also liked that book)**
  - **Incremental models (e.g., Perceptron, Naïve-Bayes)**
  - **Some lazy learning techniques (e.g., collaborative filtering) although these usually do not scale well without backend work**

Remember reasons for DW

- **Without a data warehouse**
  - **Only simple models can be implemented**
  - **Can't integrate external data easily nor go through data cleansing**
  - **Hard to use constructed features (e.g., number of purchases from category X paid by Amex)**
  - **Lacks human validation and insight to business**
  - **Many prediction problems show "leaks" exist in data that may not be discovered in time (e.g., heavy spenders pay more tax on purchases, so tax predicts purchase amount)**

Closing the Loop

Consumers and Businesses

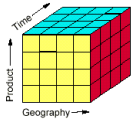


Closing the loop

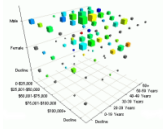
Analysis

Reporting

OLAP



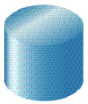
Visualization



Data Mining

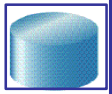
Touchpoints:
Web Store
Call Center
Campaigns

OLTP store



Data Warehouse

Other Data Sources



Syndicated data (e.g., Experian/ Acxiom)

Closing the Loop by Humans

- **Humans can close the loop**
  - **Analysis reveals comprehensible patterns**
  - **Humans generate hypotheses, test and validate**
  - **Humans take action and change interactions**
    - Offer new promotions
    - Offer new products (e.g., analyze failed searches)
    - Offer new cross-sells
    - Change advertising strategy based on segments
    - Execute e-mail and direct mail campaigns
  - **May result in strategic impact on business decisions**

Closing the Loop Automatically

- **Automated closing of the loop**
  - **Optimization of certain processes (e.g, cross-sell offers)**
  - **Faster cycle (no human involvement required), but requires tighter software integration of components and rarely results in interesting strategic insight**
  - **Can use opaque models (e.g., Neural Networks, Collaborative Filtering)**
  - **Legal issues (must not offer cigarettes to minors even though they correlate with chewing gum)**
- **Each method of closing the loop has its advantages/disadvantages**

Agenda

- Introduction (45 min)
- Architecture and Data Flow (45 min)
  - Collecting the data
  - Building the warehouse
  - Closing the loop
- Mining Web Data (75 min)
  - Transformations
  - Reporting and Visualization
  - OLAP
  - Mining
- Summary (20 min)

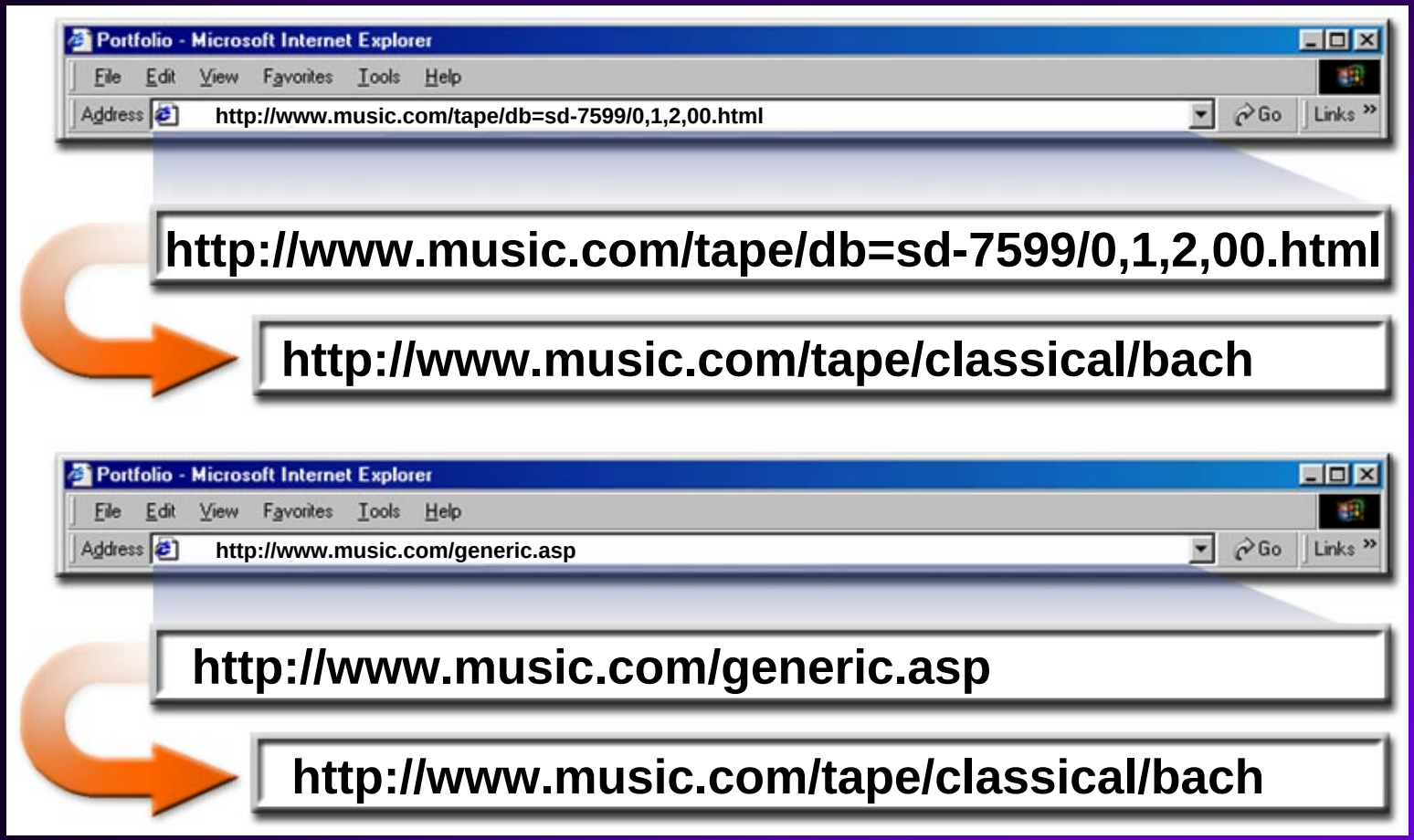
- **This slide intentionally (not) left blank**

Transformations

- **Creating a warehouse is not enough; you need to:**
  - **Make URLs more understandable (dynamic content, page titles)**
  - **Handle reverse DNS lookup (208.216.181.15 ➠ www.amazon.com)**
  - **Sessionize (decide which requests belong to same session if you are not using an application server). Commonly cookie-based**
  - **Identify crawlers/robots**
  - **Identify test users**
  - **Compute session-level attributes (number of pages, time spent, session milestones)**
  - **Create customer attributes (repeat visitor, frequent purchaser, high spender)**
  - **Use products and content attributes**
  - **Compute abstractions of existing attributes (e.g., product hierarchies, referrers, browsers, regions)**
  - **Calculate date/time attributes**

Dynamic Content

- Must rewrite the URL to increase understanding and facilitate analysis of served content



Crawler/Robots

- **Crawlers are programs that visit your site**
  - **Search crawlers** } **Good**
  - **Shopping bots** } **Good**
  - **IE5 offline viewer**
  - **Performance assessment (e.g., Keynote)**
  - **E-mail harvesters - Evil**
  - **Students learning Perl scripts**
- **For understanding your customers, it is very important to filter out crawlers**
- **They may account for 50% of sessions!**

Techniques to Identify Robots

- **Browser sends a USERAGENT strings (e.g., keynote, google). This requires large tables of USERAGENTs to be setup**
- **Bots commonly turn off images, have empty referrers**
- **Friendly bots will visit `robots.txt` file**
- **Page hit rate is too fast (although some crawl slowly to avoid hurting the sites)**
- **Pattern is a depth-first or breadth-first search of site**
- **Bots never purchase (helps identify USERAGENT strings)**
- **Eliminate very long paths and unique path sequences**
- **Setup trap (hidden link) and see who follows it**
- **Resource:** `http://bots.internet.com/search`

Test Users

- Every respectable site has a QA department
- Their users hit the site with different patterns
  - Their goal is to break the site, not to purchase
  - They'll change URLs
  - They'll surf quickly
  - They'll click on random links
- Purchases by the QA team are recognized and ignored by fulfillment center
- Must identify them
  - Requests from specific IP addresses
  - Use of special credit card numbers

Session-level Attributes

- **Pages**
  - Page views per session (deep vs. shallow)
  - Unique pages per session
  - Promotional vs. standard entry
- **Time**
  - Time spent per session
  - Average time per page
  - Fast vs. slow connection
- **Session Milestones**
  - Did they go through registration, when?
  - Did they look at the privacy statement?
  - Did they use search?
  - Did they start and/or complete checkout?

Customer Attributes

- **Some attributes based on customer history**
  - **Initial vs. Repeat visitor/purchaser**
  - **Recent visitor/purchaser**
  - **Frequent visitor/purchaser**
  - **Readers vs. browsers (time per page)**
  - **Heavy spender**
  - **Original referrer**

Recency
Frequency
Monetary / Duration

- **Other attributes are created as hypotheses**
  - **Heavy purchaser of children's products**
  - **Lunchtime visitor**

Product and Content Attributes

- Generalization often has to happen at higher levels than individual content URLs and product ids
- Products
  - Common attributes are color, size, and weight
  - Specific attributes for category (power consumption for electrical appliances, inseam size for pants)
- Content
  - Common attributes are topic, version, and author
  - Specific attributes for content types (story and event for news articles, photographer and length for videos)
- Harder problem: assign attributes to pages showing collections of products (assortments) or multiple content sets (portals)

Abstract Attributes

- **Many attributes have too many values**
  - **There are over 100 colors for Jeans**
  - **There are hundreds of area codes and zip codes**
  - **There are hundreds of referring sites**
- **Higher-level abstractions must be created**
- **One common abstraction is to use the hierarchy**
  - **Organizations naturally organize products in a hierarchy**
  - **Products: jeans, Men's Jeans, Levi's, 505, button fly, …**
  - **Content: classified, auto classified, SUV auto classified, Pathfinders**

Date/Time Attributes

- There are many date/time attributes
  - First session time
  - Registration time
  - Delivery time
- Most tools are poor at handling date/time
- Abstract attributes can be created
  - Day of week or month
    (people get paid on Fridays or on the 1st and 15th)
  - Hour of day
    (behavior is different in the morning than at night)
  - Weekend vs. Weekdays
  - Seasons
- Differences between dates are important for showing trends

Tracking Visitors

Within one session

- Referring URLs
  - When traffic is due to a specific reason (search, ad, affiliate)
- Special URLs
  - www.kodak.com/go/freestuff
- Query Strings at the end of URLs
  - www.kodak.com?AdName=freestuff

Across sessions

- Host IP + Browser String
  - Proxies limit accuracy (e.g., AOL, WebTV)
    http://webusagemining.com/sys-itmpl/webdataminingworkshop/
- Cookies
  - Stored on visitor's browser on first visit to site
- Registration
  - Require login for every visit

Agenda

- Introduction (45 min)
- Architecture and Data Flow (45 min)
  - Collecting the data
  - Building the warehouse
  - Closing the loop
- Break (10 min)
- Mining Web Data (75 min)
  - Transformations
  - Reporting and Visualization
  - OLAP
  - Mining
- Summary (20 min)

Reports

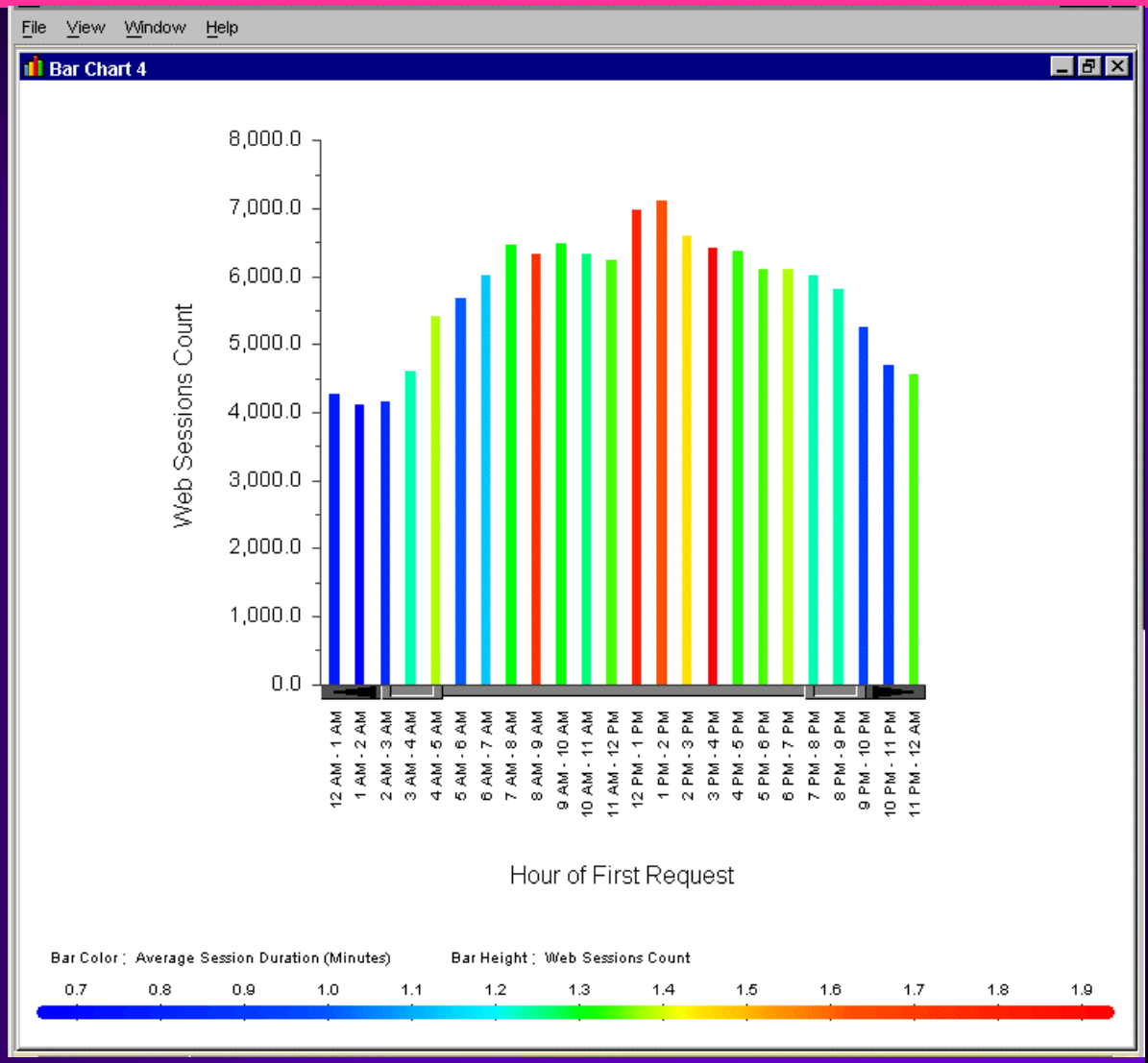
- **Traditional representation of data as tables**
  - **Elements may be changed by user (which columns appear)**
  - **Format may be change by user (order of columns, color, etc.)**
  - **Once report has been generated, user typically cannot change it or ask questions of it, without regenerating the report**
- **The most important tool for business users**
  The most unappreciated tool by companies
  - **Many companies provide great analytics but miss basic reporting**
  - **WebTrends has simple log analysis but very clear and nice reports**
- **Examples: Actuate, AlphaBlox, Brio, Business Objects, Crystal Decisions (Seagate), Microsoft Excel**

Visualizations

- **Tabular data can be hard to interpret**
  - **Provide simple bar charts and scatter plots**
- **Business users need to quickly see trends**
  - **Provide time-series graphs**
- **Avoid creating state-of-the visualizations that only the creators can understand**

Simple Bar Charts

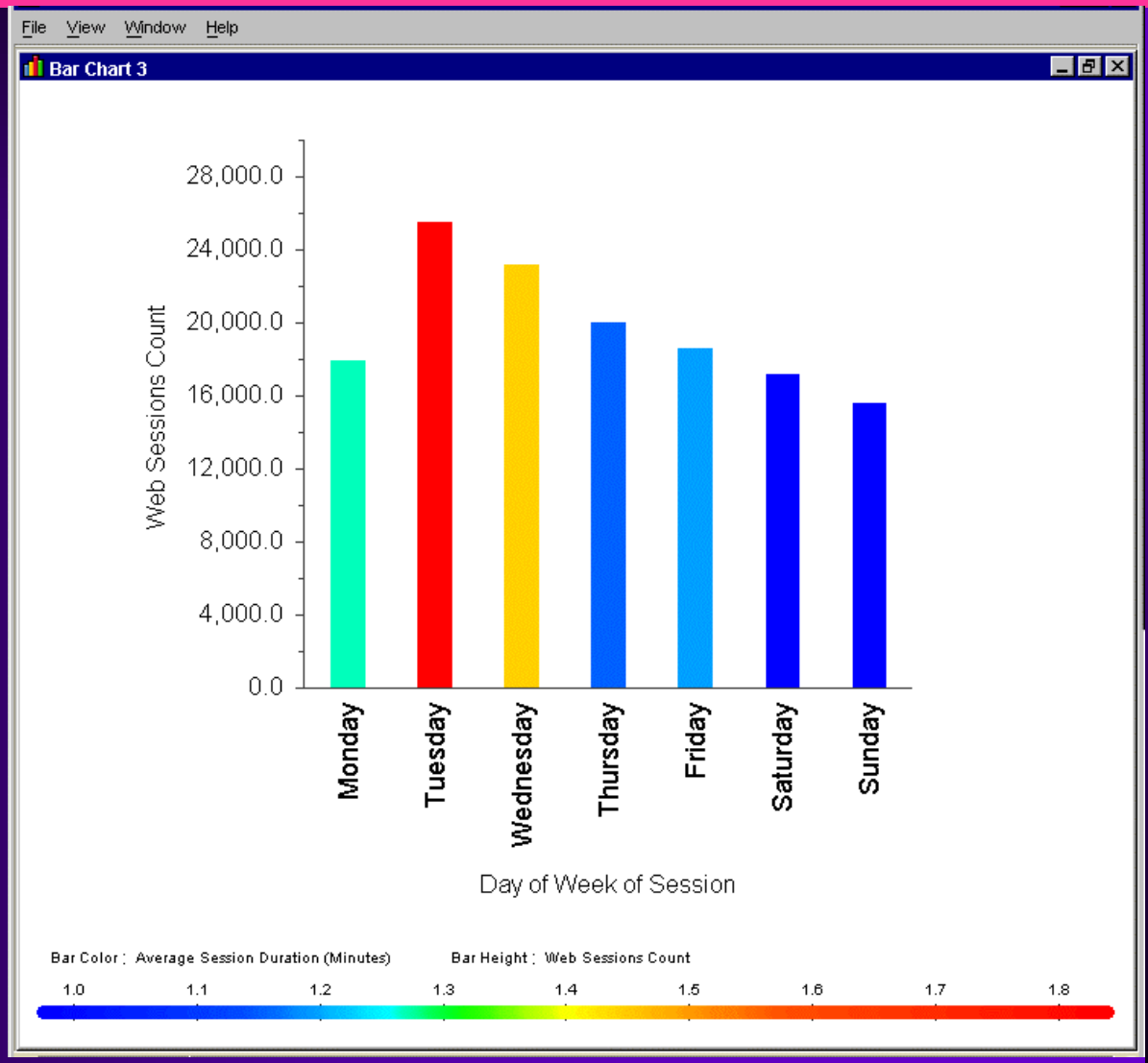
Example of real data.
Height = session count
Color = duration
(cold to hot)



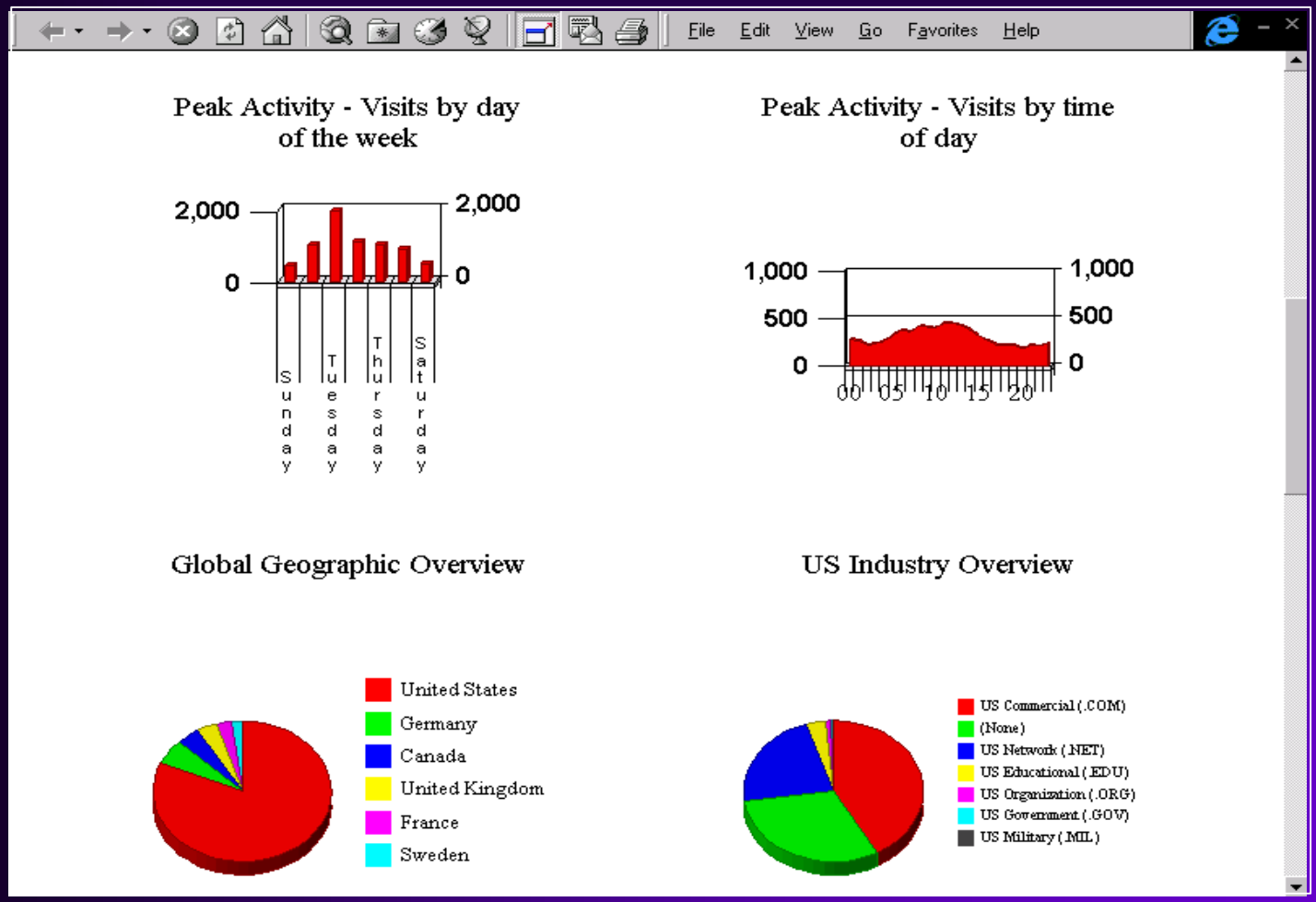
Simple Bar Chart II

Example of real data.
Height = session count
Color = duration
(cold to hot)

Tuesday and Wednesday are special.
What happened?



Common Web Reports

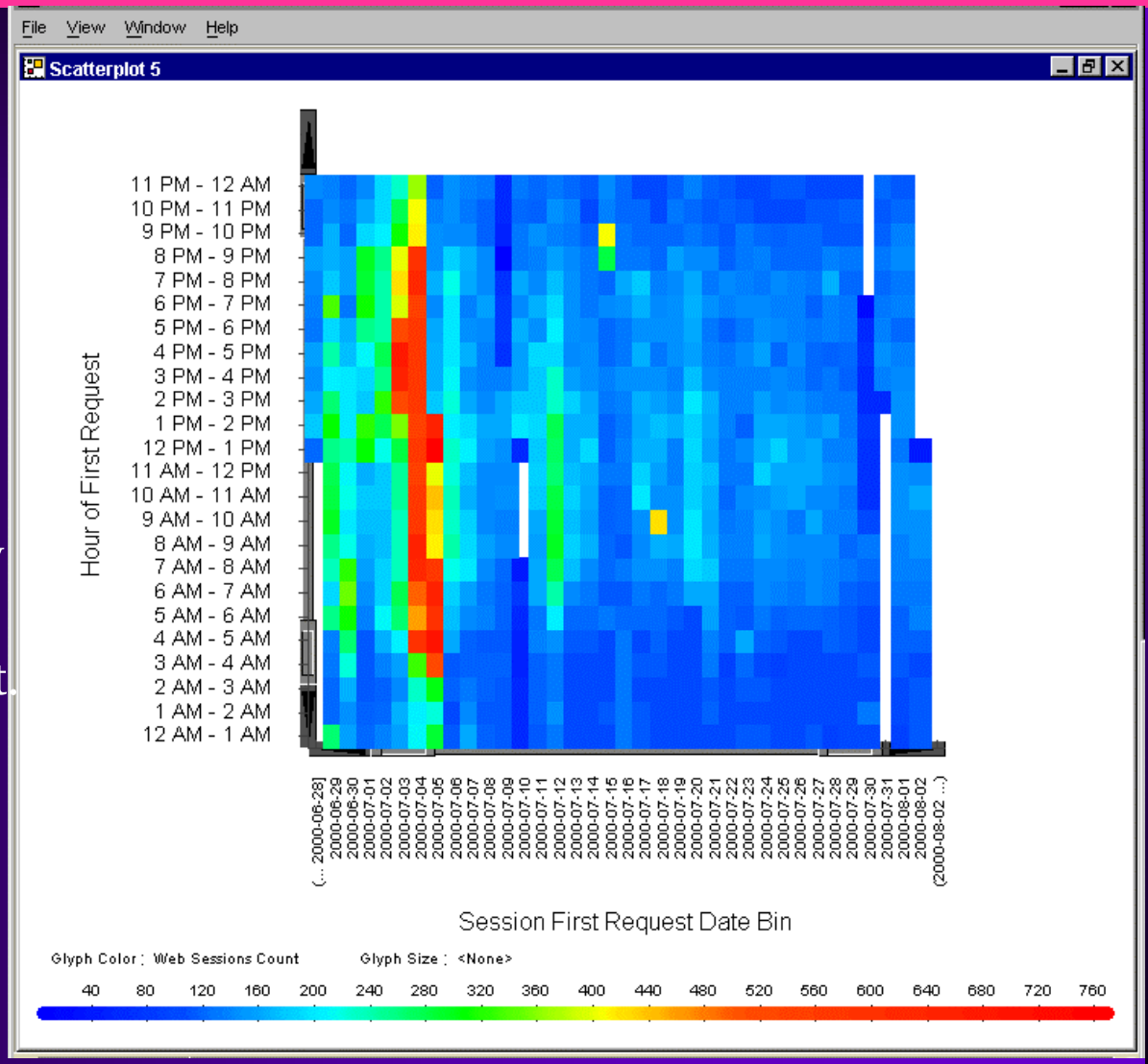


Heat Map Visualization

Example of real data.
Plot of every hour over several weeks
Color = session count (cold to hot)

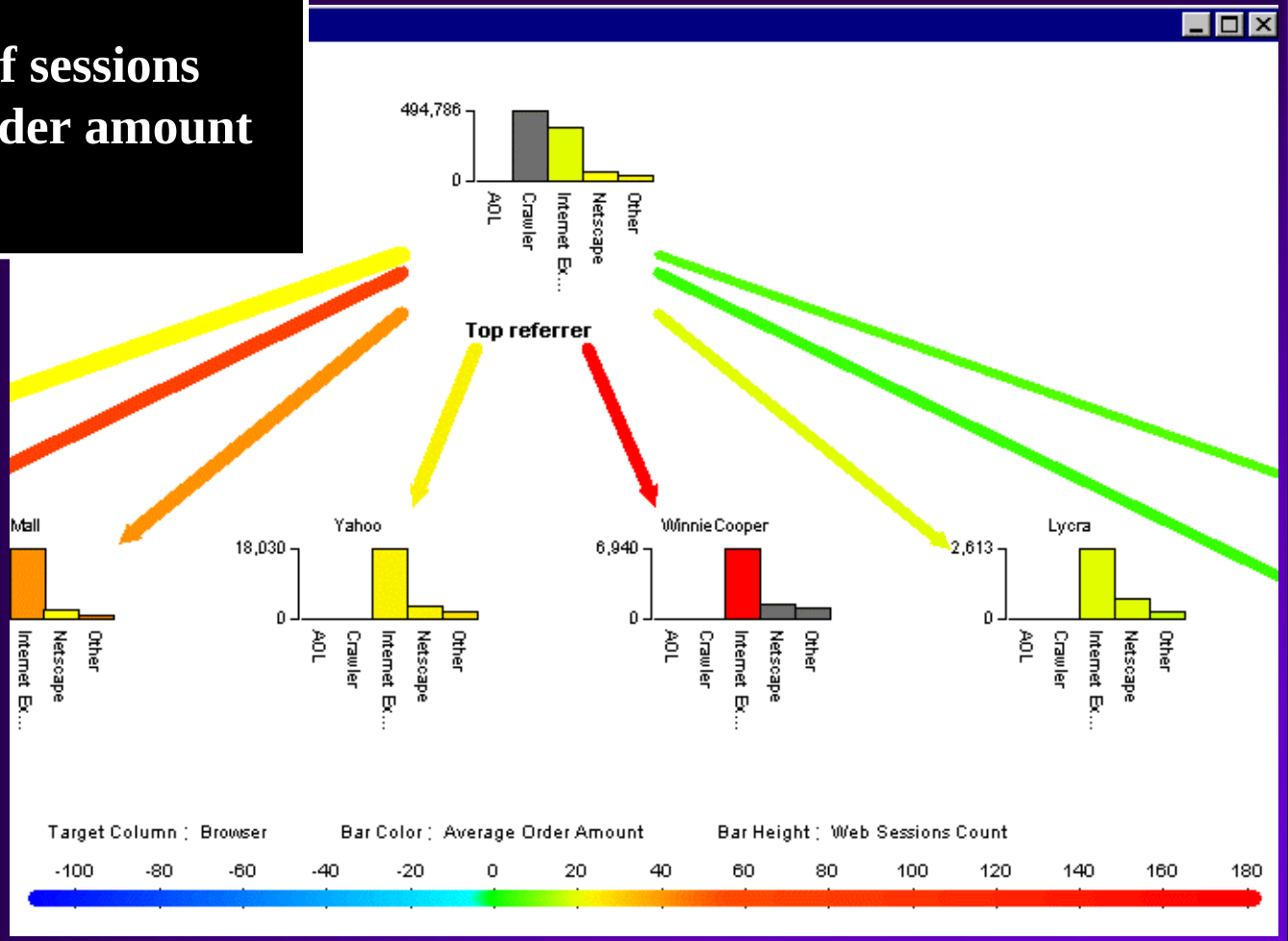
Tue/Wed are not generally high, but holiday and promotion made an impact.

Also note white downtime



Hierarchical Decomposition

Every node shows browser type on X-axis
Height = number of sessions
Color = average order amount



On-Line Analytical Processing

- **Transforms raw data to reflect dimensionality "How much did we spend on health benefits, by month; in our largest three divisions, in each state, compared with plan?"**
- **Very fast flexible operations (e.g., sum, average) on large amounts of data**
- **Two primary variations**
  - **Relational OLAP (ROLAP)**
  - **Multidimensional OLAP (MOLAP)**
  - **Hybrid OLAP solutions are emerging**
- **Resources:**
  www.olapreport.com
  www.olapcouncil.org/whtpap.html

Relational vs. Multi-dimensional

Relational tables have records with fields

Customer Name	Customer #	Amount	Address	Region
Jack's Hardware	10456	103.2	40 Main St.	West
Value Stores	10114	97.2	18 Elm St.	Central
Housewares Inc.	11104	233.22	17 Main St.	East
Walter Lock	11230	57.2	6 Charles St.	West

A two-dimensional matrix with customer name going down and a dimension (e.g., region) going across with a measure (e.g., amount spent) in the intersection is sparsely populated

Customer Dimension	Dimension →		
	West	Central	East
Jack's Hardware	103.2		
Value Stores		97.2	
Housewares Inc.			233.22
Walter Lock	57.2		

Relational vs. Multi-Dimensional II

Product	Region	Sales
Nuts	East	50
Nuts	West	60
Nuts	Central	100
Nuts	**Total**	**210**
Screws	East	40
Screws	West	70
Screws	Central	80
Screws	**Total**	**190**
Bolts	East	90
Bolts	West	120
Bolts	Central	140
Bolts	**Total**	**350**
Washers	East	20
Washers	West	10
Washers	Central	30
Washers	**Total**	**60**
Total	**East**	**200**
Total	**West**	**260**
Total	**Central**	**350**
Total	**Total**	**810**

This relational table has more than one product per region and more than one region per product. It lends itself to a multidimensional representation with products and regions.

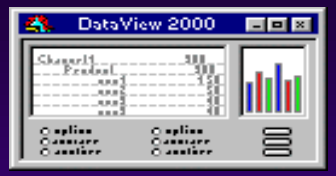
	East	West	Central	Total
Nuts	50	60	100	**210**
Screws	40	70	80	**190**
Bolts	90	120	140	**350**
Washers	20	10	30	**60**
Total	**200**	**260**	**350**	**810**

OLAP

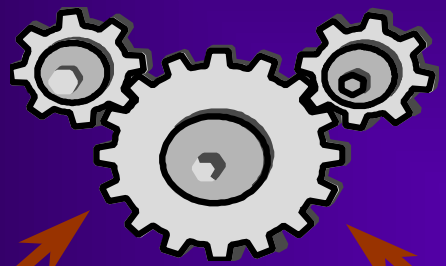
- **Relational OLAP (ROLAP)**
  - **Query data directly from relational structure**
  - **Typically requires multi-way joins**
  - **Performance suffers with complexity of questions**
  - **Verdict: very flexible but doesn't scale well**
  - **Examples: Business Objects, Cognos, MicroStrategy**
- **Multi-dimensional OLAP (MOLAP)**
  - **Built n-dimensional cubes from source data**
  - **Data access is n-dimensional lookup**
  - **Building cubes can be time intensive**
  - **Verdict: very fast but not very flexible**
  - **Examples: Hyperion, Microsoft , Oracle Express**

Hybrid OLAP (HOLAP)

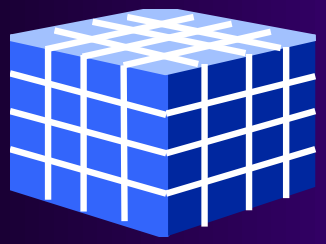
User Interface



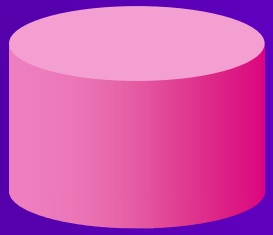
Analysis Engine



MD Views
Cross Tabulations
Time Intelligence
Slice & Dice
Filtering
Sorting
Calculation
Consolidation



MDB



RDBMS

Tree Drill-Down

- **Front-ends to MDDB (multi-dimensional databases) provide easy access to data**

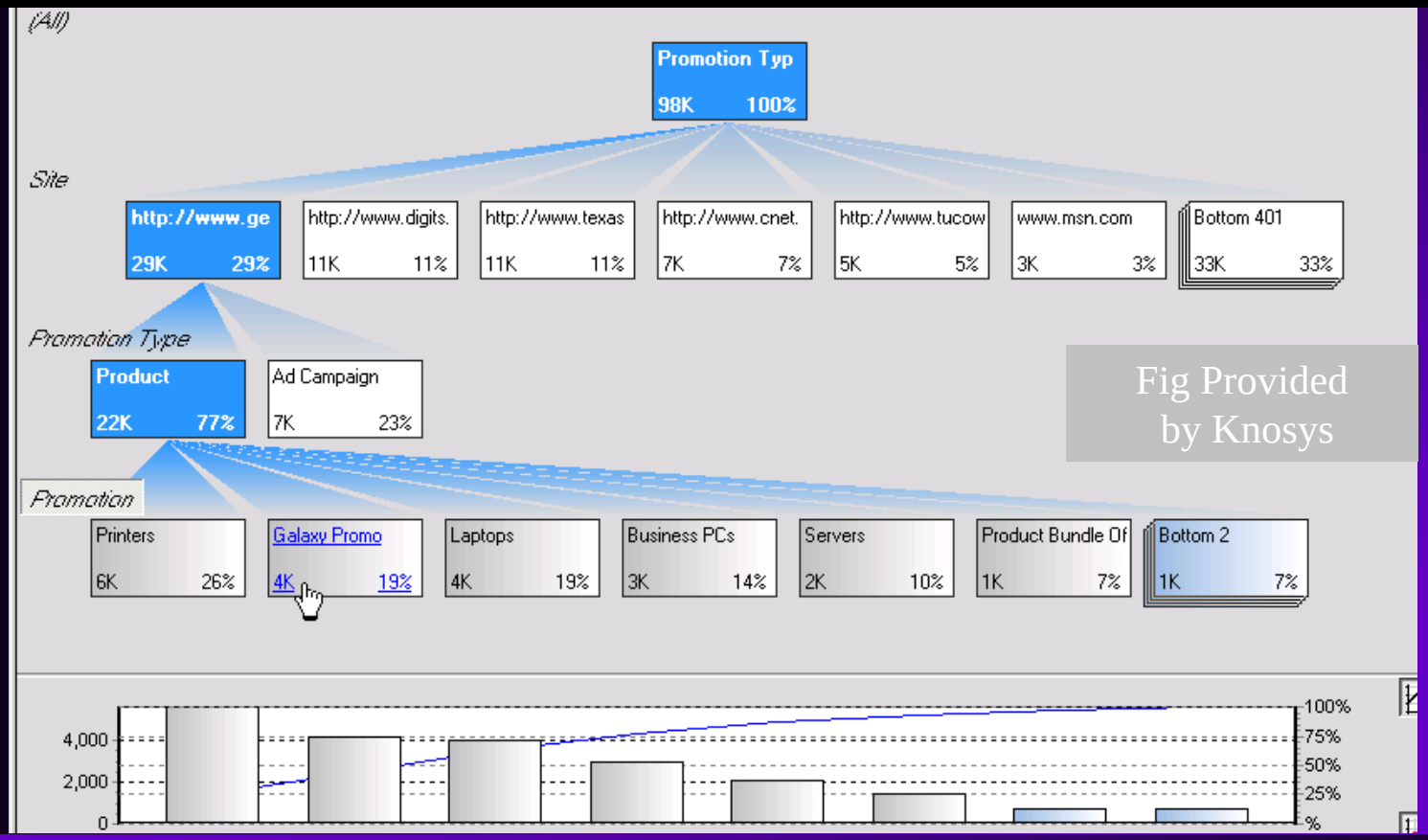


Fig Provided by Knosys

OLAP Visualizations

- **Front ends now provide powerful visualizations that are very fast and easy to manipulate**

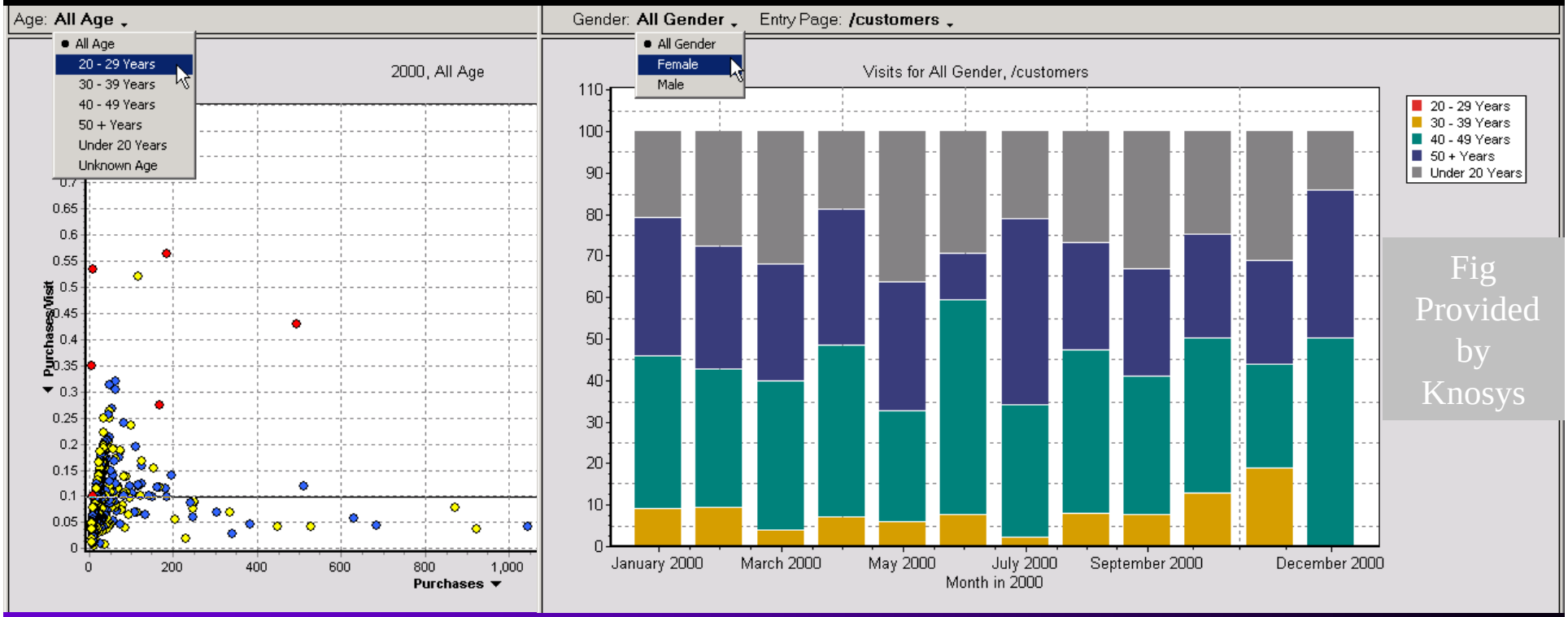


Fig Provided by Knosys

OLAP Example

Case Study: How does visitor preferences vary by content?

Report: All Content Stats

Site	Profile Content Group (Recent) ID	Unique Visitors	Avg Pages / Visit (Recent)	Avg Time/ Visit (Recent)
washington.com	Politics	64,217	6	2,089
	Style	64,067	7	2,109
	Politics.Special	54,351	7	2,222
	Real Media	39,256	8	2,556
	Classifieds	22,250	9	3,451
	Business	18,786	9	3,267
	Sports.NFL	9,319	11	4,238
	Sports Search	8,009	11	5,143
	Comics	7,796	12	2,744
	Frompost	6,229	14	4,121
	Horoscope	5,792	12	2,691
	Travel	4,591	15	5,202
	Business.Daily	4,325	15	3,944
	Restaurants	3,550	16	4,623
	Sports.NBA	3,011	17	4,894
	Cooking Discussion	1,846	16	3,262
	Redskins	670	19	8,983

- **Why is *pages/ visit* for politics relatively low?**
- **Theory: politics readers are high frequency and low *pages/visit***
- **Let's test theory: drill down on politics, show frequency**

OLAP Example

Drilldown on "Politics"

Report: All Content Stats

Site	Total Visit Category ID	Unique Visitors	Pages/Visit (Recent)	Time/ Visit (min, Recent)
	Measures			
washington.com	01 Visit	28,110	8	649
	02 - 05 Visits	14,476	6	520
	06 - 10 Visits	5,707	5	511
	11 - 25 Visits	5,018	6	1,439
	26+ Visits	4,346	7	3,113

- **Answer:** ***time/visit*** **increases dramatically at high frequency**
- **Politics readers** ***read*** **instead of browse!**
- **From here, we could continue to drill down or drill back up.**

Mining – Induction

- **Analysis Type**
  - **Prediction, or business rules created by a person**
- **Sample Applications**
  - **Which product or banner should be displayed?**
  - **Which person is most likely to respond to an outbound email?**
  - **How likely is a visitor to return to the Web site?**
  - **Which customers are the heaviest spenders?**
- **Objections**
  - **Dynamic nature of Web data is difficult to model**
  - **Algorithms are not well understood by business users**
- **Example Companies**
  - **Accrue, Angoss, Broadbase, Blue Martini, E.piphany, Microsoft analytical services, SAS**

Mining – Segmentation

- **Analysis Type**
  - Cluster to discover groups of similar behavior or a similar profile
- **Sample Applications**
  - Find customer segments
  - Generate small number of different web sites or stores
  - Discover communities of visitors with similar interests
  - Identify substitute or cannibal products
- **Objections**
  - How well do customers fit in a particular group?
  - Hard to understand high-dimensional segments
- **Example Companies**
  - Accrue, ATG Scenario Server, Blue Martini

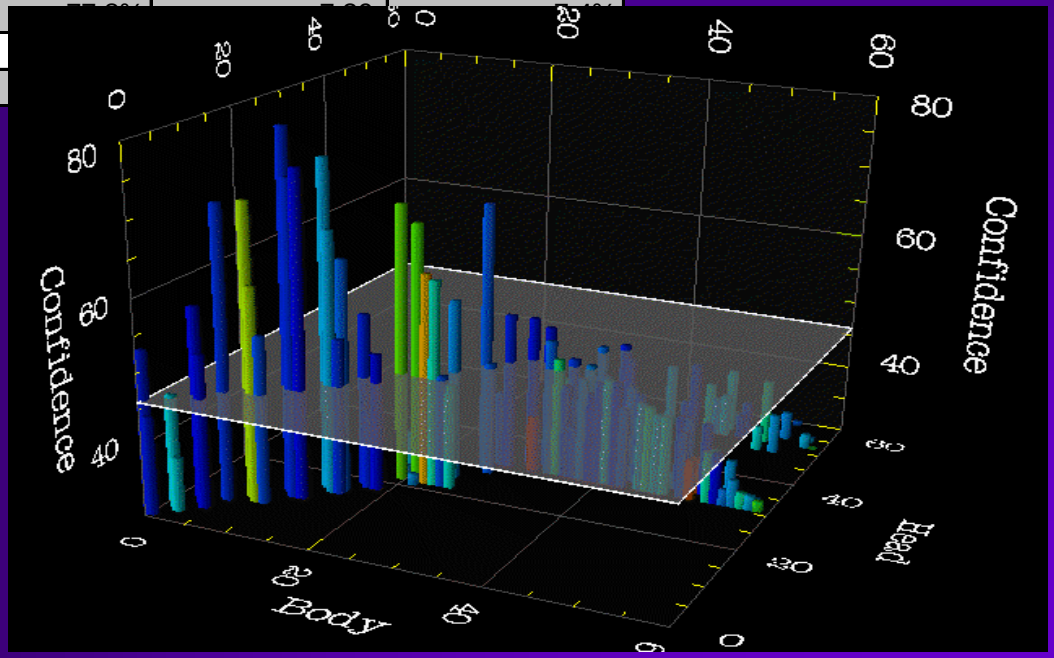
Mining – Associations

- **Analysis Type**
  - Link analysis for associations or time-based sequences
- **Sample Applications**
  - Shopping cart analysis
  - Up-sell and cross-sell
  - Path analysis
- **Objections**
  - Shear number of rules makes interpretation difficult
  - With no holdout testing, difficult to know whether results will stand up over time
- **Example Companies**
  - Accrue, IBM, SGI, Vignette

Association Example

Recommend potential purchases based on basket contents

Driver Item	Recommendation	Confidence	Lift	Support
Arugula	Dill	57.1%	7.76	4.2%
	Basil	44.4%	5.43	3.1%
Basil	Parsley	70.0%	7.39	7.4%
Colombian	Jamaica	50.0%	6.79	7.8%
Cool Breezer	Grape	75.0%	11.88	3.2%
Dill	Arugula	57.1%	7.76	4.2%
	Basil	67.4%	5.43	3.8%
Pineapple	Grape			
Yellow Pepper	Jalapeno			
	Granny Smith			



Mining – Path Analysis

- **Analysis Type**
  - Explore, understand, or predict visitors navigation patterns through Web site
  - Multiple analytic techniques: statistics, sequences, induction, clustering, compression
- **Sample Applications**
  - Designing a more efficient or user friendly site
  - Discovering misleading, duplicative, or overlapping content
  - Understanding the effectiveness of referring links
- **Objections**
  - Most path analysis provides only simple reporting
- **Example Companies**
  - Nearly everyone

Most Frequent Path Report

Top Paths Through Site by Visits			
Start Page	**Paths from Start**	**Visits**	**%**
Products	1.**Products** http://www.businesscomputing.com/products/	837	9.28%
	1.**Products** http://www.businesscomputing.com/products/ 2.**110 Desktop Computer Specs** http://www.businesscomputing.com/products/pc110/	111	1.23%
	1.**Products** http://www.businesscomputing.com/products/ 2.**330 XL Desktop Computer** http://www.businesscomputing.com/products/pc330xl/	67	0.74%
	1.**Products** http://www.businesscomputing.com/products/ 2.***Page Has No Title*** http://www.businesscomputing.com/shoppingcart.htm	60	0.66%
	1.**Products** http://www.businesscomputing.com/products/ 2.**110 Desktop Computer Specs** http://www.businesscomputing.com/products/pc110/ 3.**110 Desktop Computer** http://www.businesscomputing.com/products/pc110/intro.htm	47	0.52%

Collaborative Filtering

- **Analysis Type**
  - Recommend small # of products out of 1,000's
- **Benefits**
  - No need for a training set; algorithm bootstraps itself
  - Can be used directly against operational data store
  - Learning is incremental and should improve over time
- **Objections**
  - Tie lag to gather data before recommendations valid
  - Black box perception: Why is a recommendation made?
  - Difficult to produce a confidence interval in prediction.
  - In practice, few examples leads to sparse data such that the recommendations are weak
- **Example Companies**
  - Like Minds, Net Perceptions

Teaser - Birth Dates

A bank discovered that almost 5% of their customers were born on the exact same date

How can that be explained?

Teaser - Gender Mystery

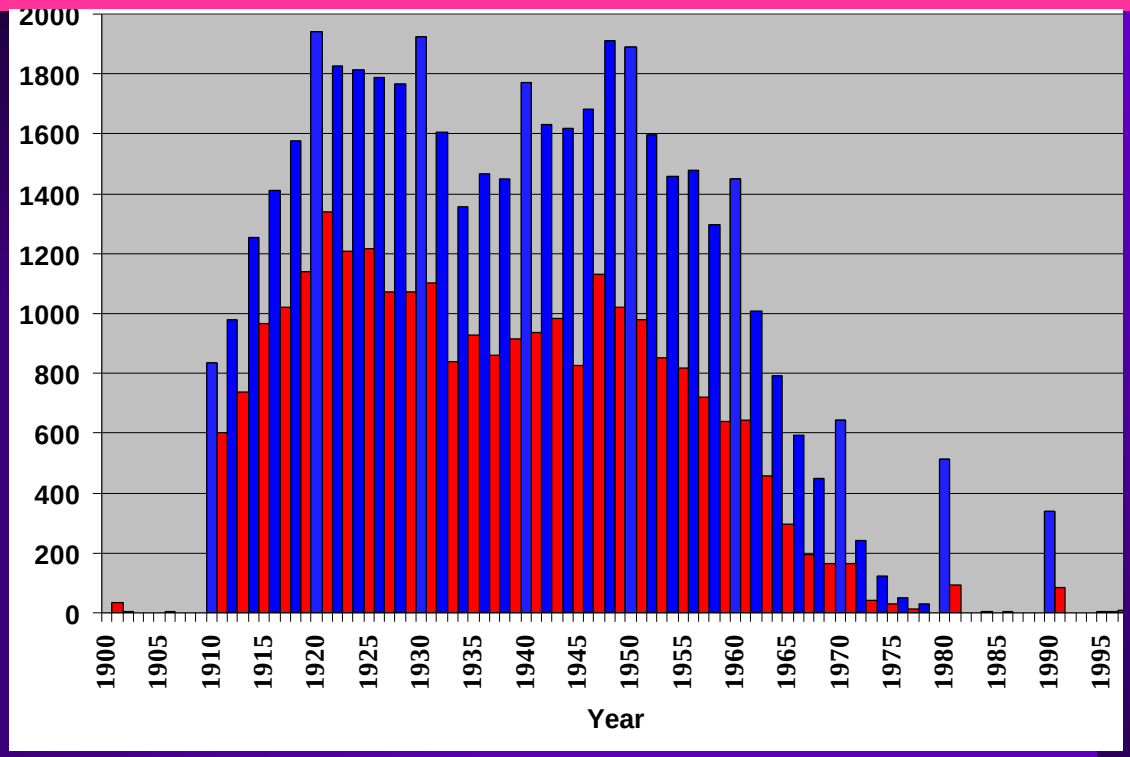
- A site has gender on the registration form
- Acxiom, a syndicated data provider, also provides gender
- A very large discrepancy found between
  - **Males according to registration form and**
  - **Acxiom provided data**

Why?

Hint: Acxiom only conflicted with females, claiming some females are males. Never in the other direction

Teaser - Mysterious Birth Years

The KDD CUP 98 data contained anomalies for date of birth
[Georges and Milley, SIGKDD Explorations 2000]



- Spikes on years ending in zero (white dots on blue)
- Few individuals born prior to 1910
- Many more individuals who were born on even years (blue) as on odd years (red)

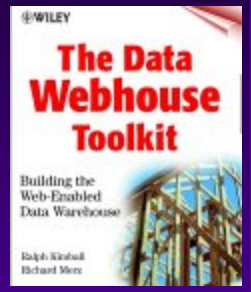
Why?

Summary

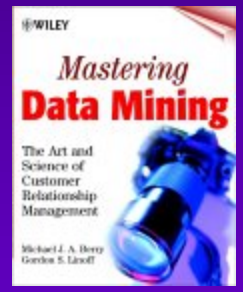
- **Significant Return On Investment from analyzing e-commerce data. Killer domain**
- **Data collection is important**
  Design the site with analysis in mind
- **Build a data warehouse (ETL, construct attributes, deal with bots)**
- **Analyze (reports, OLAP, visualization, algorithms)**
- **Close the loop. Experiment and improve.**

Resources (I)

- **The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse by Ralph Kimball, Richard Merz. ISBN: 0471376809 (Jan 2000)**



- **Mastering Data Mining: The Art and Science of Customer Relationship Management by Michael J. A. Berry, Gordon Linoff. ISBN: 0471331236**



- **KDNuggets, Software for Web Mining**
  http://www.kdnuggets.com/software/web.html
- **WEBKDD - Workshops in Web Mining**
  http://robotics.Stanford.EDU/~ronnyk/WEBKDD2000/index.html
  http://robotics.Stanford.EDU/~ronnyk/WEBKDD2001/index.html

Resources (II)

- **Web Mining Research: A Survey**
  http://www.acm.org/sigs/sigkdd/explorations/issue2-1/contents.htm#Kosala
- **Web Data Mining course at DePaul University by Bamshad Mobasher**
  http://maya.cs.depaul.edu/~classes/cs589/lecture.html
- **Integrating E-commerce and Data Mining: Architecture and Challenges, WEBKDD'2000**
  http://robotics.Stanford.EDU/~ronnyk/ronnyk-bib.html
- **Drinking from the Firehose: Converting Raw Web Traffic and E-Commerce Data Streams for Data Mining and Marketing Analysis by Rob Cooley**
  http://www.webusagemining.com/sys-tmpl/webdataminingworkshop/

Resources (III)

- **An Ideal E-Commerce Architecture for Building Web Sites Supporting Analysis and Personalization**
  http://robotics.Stanford.EDU/~ronnyk/ronnyk-bib.html
- **Analyzing Web Site Traffic, Sane Solutions**
  http://www.sane.com/products/NetTracker/whitepaper.pdf
- **Web Mining, Accrue Software**
  http://www.accrue.com/forms/webmining.html