KDD 2001

8/28/2001

Mining E-commerce Data
The Good, the Bad, and the Ugly

Ronny Kohavi, Ph.D.

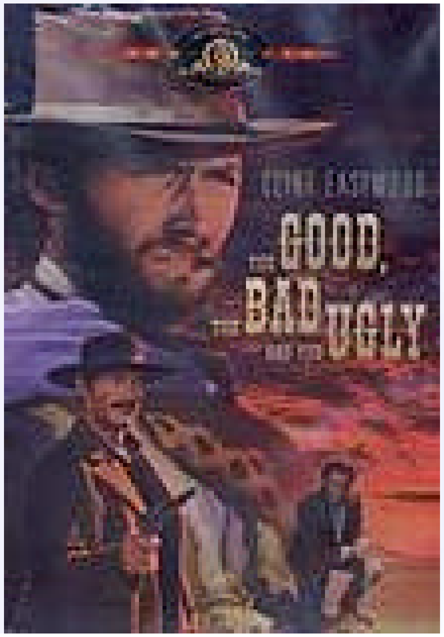
Senior Director of Data Mining

Blue Martini Software

ronnyk@bluemartini.com

http://www.bluemartini.com

http://www.Kohavi.com



Overview

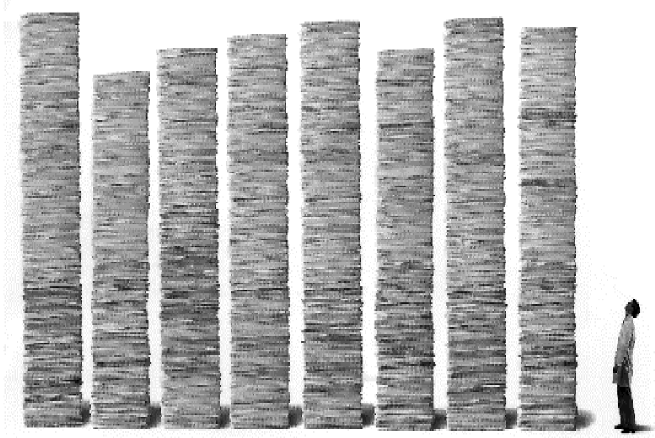
- **The Good**
  E-commerce is the killer domain for data mining
- **The Bad**
  You need more than web logs and you must conflate many data sources
- **The Ugly**
  Pre-processing and post-processing are hard
- **E-metrics**
- **Stories from mining real data**
  "Peeling the onion"
- **Summary**

The Killer Domain

Successful data mining benefits from:

- **Large amount of data (many records)**
  If you sell five items an hour on average, you'll have about 1.8M clicks in the first month. Yahoo! has 1B clicks, or 10GB/hour
- **Rich data with many attributes (wide records)**
  With proper design, many attributes are available
- **Clean data collection (avoid GIGO)**
  Electronically collected, no legacy systems
- **Actionable domain (have real-world impact**
  Easy to change sites, ads, targeted campaigns, cross-sells
- **Measurable return-on-investment**

E-commerce has all the right ingredients



Many Records

- **Clickstreams generate huge amounts of data**
- **Yahoo! has 1 billion page views a day.**
  Web log data for page views is 10GB per hour!
- **New e-commerce sites, even if small, generate sufficient data for effective mining quickly**
  If you sell five items an hour on average, that's

  5 items * 24 hours * 30 days / 2% conversion * 9 clicks-in-session >

  1.6 million page views

Rich Records

Effective site design can log many attributes about what was shown or purchased:

- **Product and product attributes**
- **Assortment attributes (when multiple products are shown)**
- **Promotions shown**
- **Visit attributes (e.g., visit count)**
- **Customer attributes (when known through login/registration)**

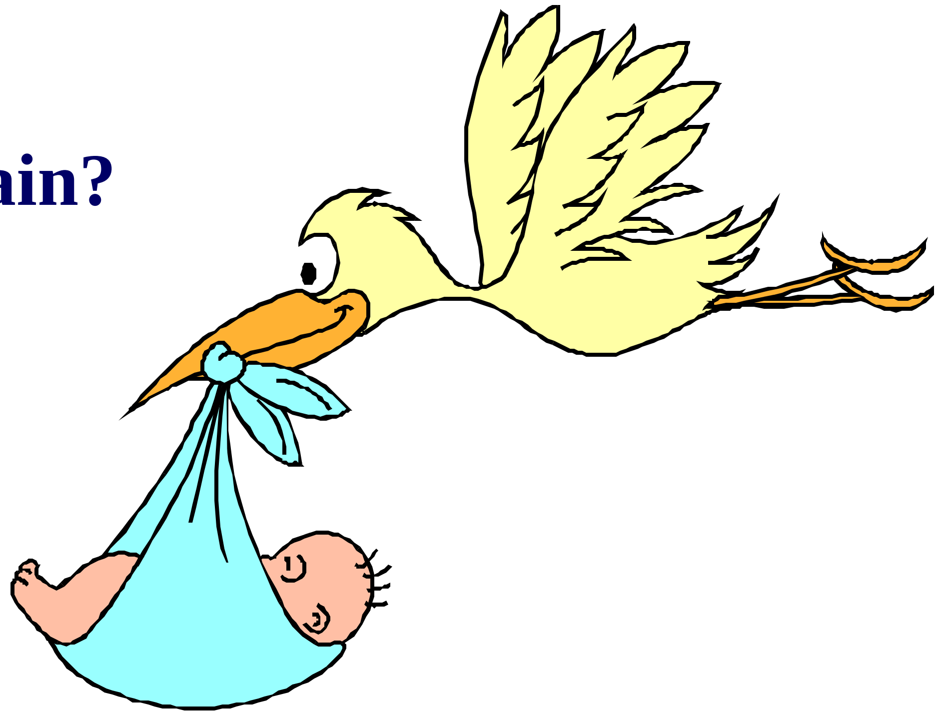
Clean Data

- **Collect data directly at webstore**
  No legacy systems
- **Collect what is needed by design**
  Not as an afterthought
- **Collect electronically - reliable data**
  No humans typing survey data from forms
- **Sample at the right granularity level**
  Architecture design principle: sample at the customer or session level,
  ***never* at page view level**

Teaser - Birth Dates

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

Can you explain?

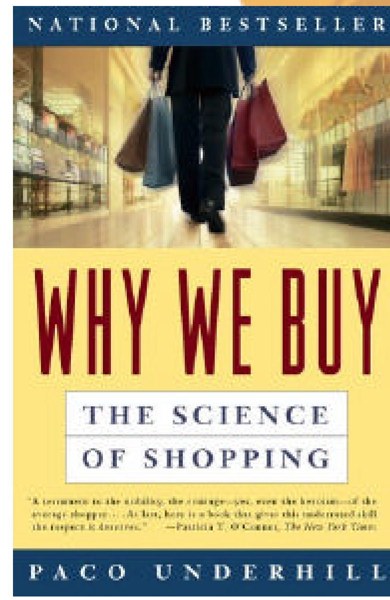


Actionable Domain

- **Few data mining discoveries had a real impact on businesses.**
  Taking action requires changing complex systems, procedures, and human habits - HARD in general
  Easier in the electronics world
- **In e-commerce, many discoveries can be made actionable by**
  - **Changing web sites (e.g., personalization)**
  - **Targeted campaigns**
  - **Changing advertising strategies based on ROI**
- **Easy to offer cross-sells or up-sells**
  Contrast with changing actual store layouts

Measure ROI

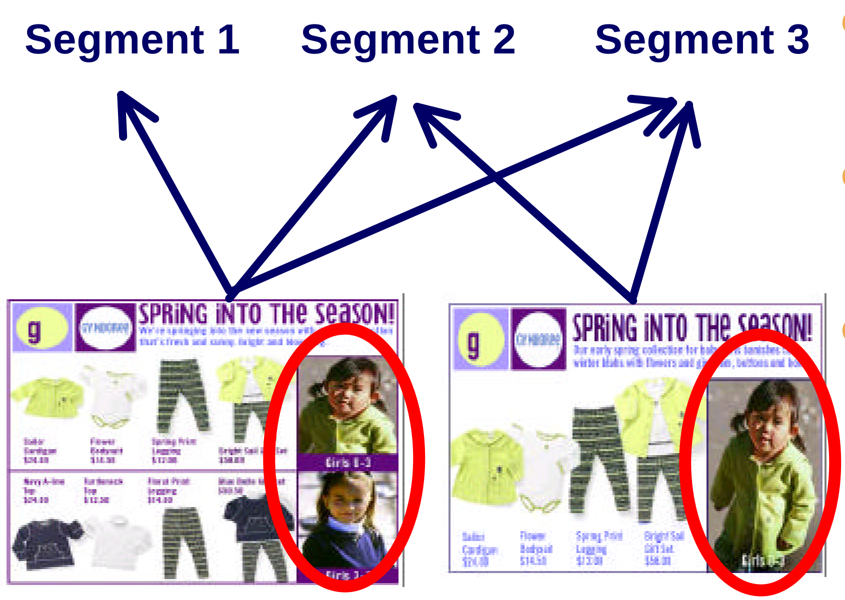
- **In e-commerce, it is easy to evaluate metrics, unlike in brick-and-mortar stores.**
  See *Why We Buy: the Science of Shopping* by Paco Underhill



- **In e-commerce it is easy to measure the *effect* of changes.**
  One can easily set control groups on a web site
- **Response to e-mails and surveys is days, not weeks and months**
- **The web is an experimental laboratory**
  It is easy to change and measure the effect

E-mail campaigns: Immediate ROI

One of of our customers, Gymboree, sent e-mail campaign based on analysis of website data of registered users: 7 email designs to 4 segments



Results:

- **Very high clickthrough rate of 22% (normal is 10%)**
- **Average order size was 36% higher than normal**
- **Email with two age groups of the same gender outperformed that with single age (medium targeting)**
- **Lifestyle images better than products**

GYMBOREE

The Bad

Firms need web intelligence, not log analysis
-- Forrester Report, Nov 1999

Web logs provide little data, even in the Extended Common Log Format (ECLF)

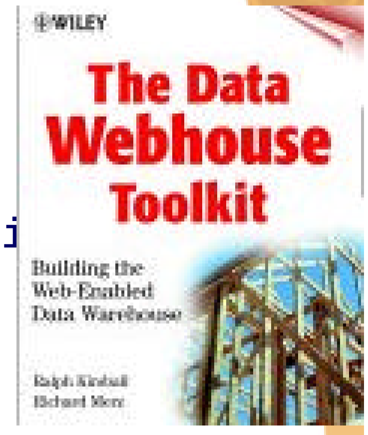
- **Host**
- **Time**
- **Request, e.g., an html page**
- **Referrer**
- **User agent (browser identifier)**
- **IP address**
- **Cookie**
- **Bytes, status, ...**

What is on the Web Page?

- **Weblogs designed for analyzing web servers, not for mining e-commerce transactions and clickstreams**
- **Given a URL, what was displayed?**
  - **Reverse URL mapping. Very brittle.**

    `http://www.amazon.com/exec/obidos/ASIN/0471376809/qi`
    `41580/105-9856660-9155942`

    is *The Data WebHouse Toolkit*



- **Hard to derive attributes of the product, such as soft cover, author, edition, year?**

Dynamic Content is Harder

Dynamic content, which is becoming more common makes web log analysis harder

- **The same URL will display different items**
- **URLs are amazingly long in dynamic sites and information is in the application server session:**

  http://www.im.aa.com/American?BV_EngineID=dealikcjfekgbfdmcflmcfkhdgfh.7&BV_Operation=Dyn_RawSmartLink&BV_SessionID=%40%40%40%4008226 17159.0968100982%40%40%40%40&form%25destination=index-member.tmpl&BV_ServiceName=American

- **Personalized content (e.g., recommended cross-sell) is practically impossible to reconstruct from web logs**

Sessionizing is Heuristic-Based

- **HTTP is stateless**
- **Sessionizing is still a research topic**
  See Measuring the Accuracy of Sessionizers for Web Usage Analysis Berent, Mobasher, Spiliopoulou, and Wiltshire, in Proceedings of the Web Mining Workshop at the First SIAM International Conference on Data Mining, 2001
- **Recreating user sessions is heuristic based:**
  - **IP addresses**
  - **Cookies**
  - **Browser type**

Business Events

Some events cannot be determined from weblogs:

- **Add to shopping cart - needed to compute value of abandoned shopping carts**
- **Change quantity of item in cart**
- **Promotion offered on page**
- **Out of stock shown on the page**
- **Form information**
- **Search - common keywords or keywords that were not found**
  (an important warning to an e-commerce site)

Example - Search Keywords

- On one of our sport-related sites, the top searched keywords were:
  - **Baseball**
  - **Video**
  - **Softball**
  - **Volleyball**
  - **Pins**
  - **Equestrian**
  - **Videos**
  - **Posters**
  - **Music**
  - **Poster**

What is common to the words in red?

Example - Search Keywords

- On one of our sports-related sites, the top searched keywords (in order) were:
  - Baseball
  - Video
  - Softball
  - Volleyball
  - Pins
  - Equestrian
  - Videos
  - Posters
  - Music
  - Poster

Searches for red words yielded zero results!

• Some words just need a synonym

• Some words should send a strong message about items the store should carry!

Weblogs do not typically contain sufficient information to extract failed searches.

This isn't fancy analytics, but it's crucial.

About 11% of searches fail

Matching Web Logs to DB

- **Given a request, how do you**
  - **Match it to the customer in your database that filled a registration form?**
  - **Determine if this is the customer’s second visit or the 100th visit?**
  - **Determine if the customer previously purchased?**
- **These common requests are very hard to implement as an afterthought**
- **They are even harder when you try to find “scenarios” that match multiple events**

Conversion Rates

Using hits and page views to judge site success is like evaluating a musical performance by its volume
-- Forrester Report, 1999

- **Most often-requested measures relate to conversion rates (buyers to browsers)**
- **Especially useful by referrer (e.g., ad)**
- **Given an HTTP request that has one of your ads as the referrer field, how can you tell if it resulted in a sale?**

A Real-World Referrer Example

- The KDD Cup 2000 data shows the following in their initial rampup period

Referrer	# Sessions	% of traffic	# Sales	Conv rate
ShopNow	16,178	6.9%	6	0.04%
FashionMall	19,685	8.4%	17	0.09%
MyCoupons	2,052	0.9%	170	8.28%

- Conversion rates differ by a factor of over 200!
- Knowing the likelihood of purchase dramatically changes the message to present

“Bad” Is Not So Bad

- **Ignore web logs**
  They are at the wrong granularity level to be useful
- **Log the information yourself at the application layer**
- **The application knows what is on the page**
  - **The app controls sessions**
  - **The app can log business events**
  - **The app can tie a visitor to their customer information upon login**
- **Also see *Structure and Content Preprocessing* by Rob Cooley for more information**

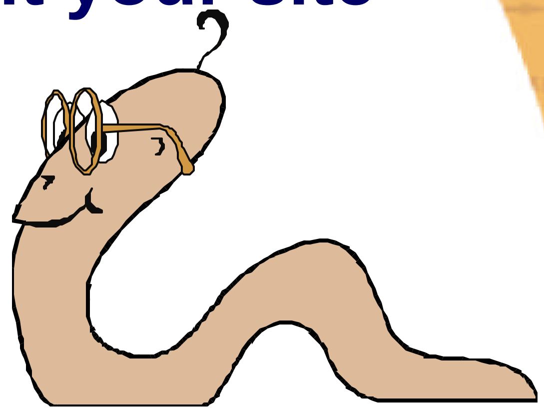
The "Ugly"

- **There are several hard problems:**
  - **Crawlers**
  - **Handling large amounts of data**
  - **Data transformations for analysis**
  - **Marketing-level insight**
- **These are excellent research topics**

Crawlers

- Crawlers are programs that visit your site
  - Search crawlers
  - Shopping bots
  - IE5 offline viewer
  - E-mail harvesters - Evil
  - Students learning Perl scripts

Good



- For understanding your customers, it is very important to filter out crawlers
- 30% of sessions come from bots/crawlers (most are measure of service bots such as Keynote)
- Fairly hard problem
  Some try to hide themselves

Identifying Bots/Crawlers

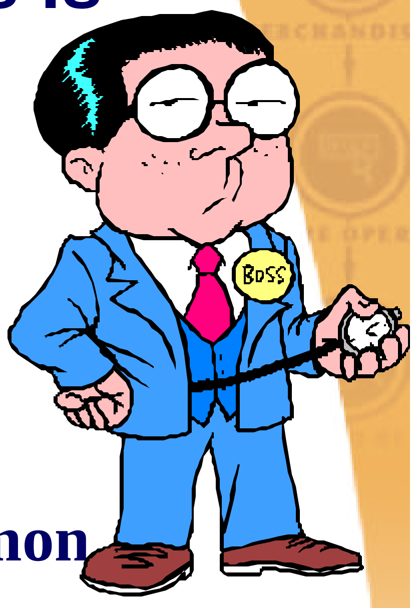
- **Browsers identify themselves on every request in a field called *Useragent***
- **Which useragents are bots?**

```
Mozilla/4.75 [en] (WinNT; U)
Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 4.0)
Mozilla/3.01 [en] (Win95; I)
Mozilla/4.0 (compatible; MSIE 5.0; Windows 98; DigExt; MSIECrawler)
Googlebot/2.1 (+http://www.googlebot.com/bot.html)
Mozilla/4.0 (compatible; MSIE 5.0; Win32)
```

- All but the top two
- **Criteria**
  - **Known robot user agents**
  - **User agents which never log in/never purchase**
  - **User agents with only 1-click or more than 100-click sessions**
  - **User agents with no referrer on all requests**

Data Transformations

- 80% of the time spent in data analysis is typically spent transforming data
- What can be done today:
  - Automate transfer of data from webstore environment to data warehouse
  - Provide data transformation UI
  - Provided "canned" transformations for common business problems



- Business users without "data" or "analyst" in their title cannot spend the time to learn how to transform data

Business Level Insight

- **Everything is a GO!**
  - **You collected data correctly**
  - **You built a data warehouse**
  - **You transformed the data**
  - **You ran a simple Perceptron (1-layer) neural network that predicts the target well**

- **The business user asks:**

  What does the 237-dimensional hyperplane represent?

- **Insight must be comprehensible to biz users**

  Sometimes required for legal reasons (e.g., no discrimination)

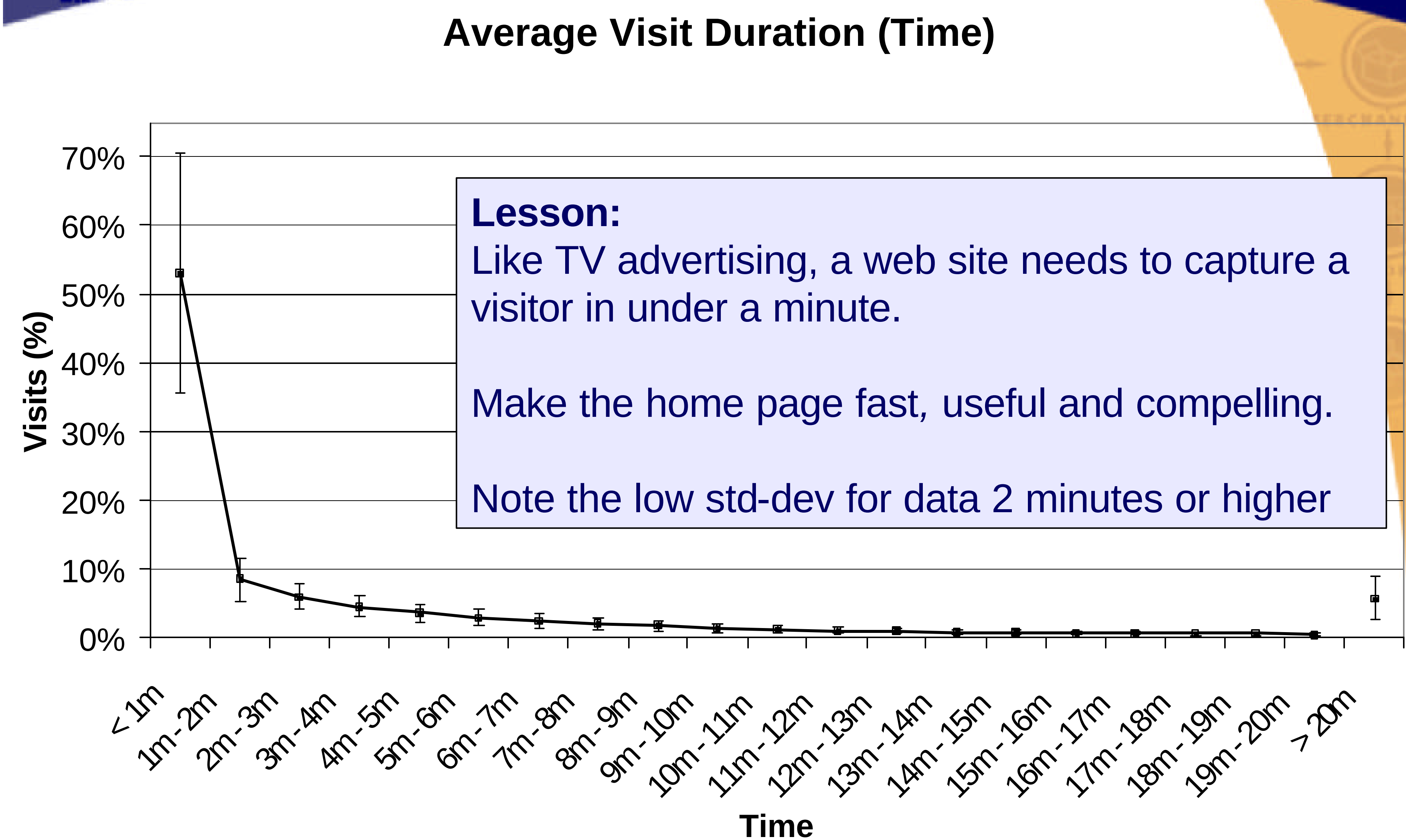
E-Metrics

- Goal: Identify commonalities and differences across sites
- Data collected over the Christmas 2000 period
- Five e-commerce web sites running Blue Martini's software
- Both US (three sites) and Europe (two sites)
- Removed bots/crawlers
- Study available at developer.bluemartini.com under Data Mining and Analytics articles

Visit Statistics

- **An average visitor**
  - **Views 8-10 pages**
  - **Spends 5 minutes on the site**
  - **Spends 35 seconds between pages**
- **An average *purchasing* visitor**
  - **Views 50 pages**
  - **Spends 30 minutes on the site**
- **These numbers are *Extremely* consistent across all of the sites in the study.**

Visit Duration - Time



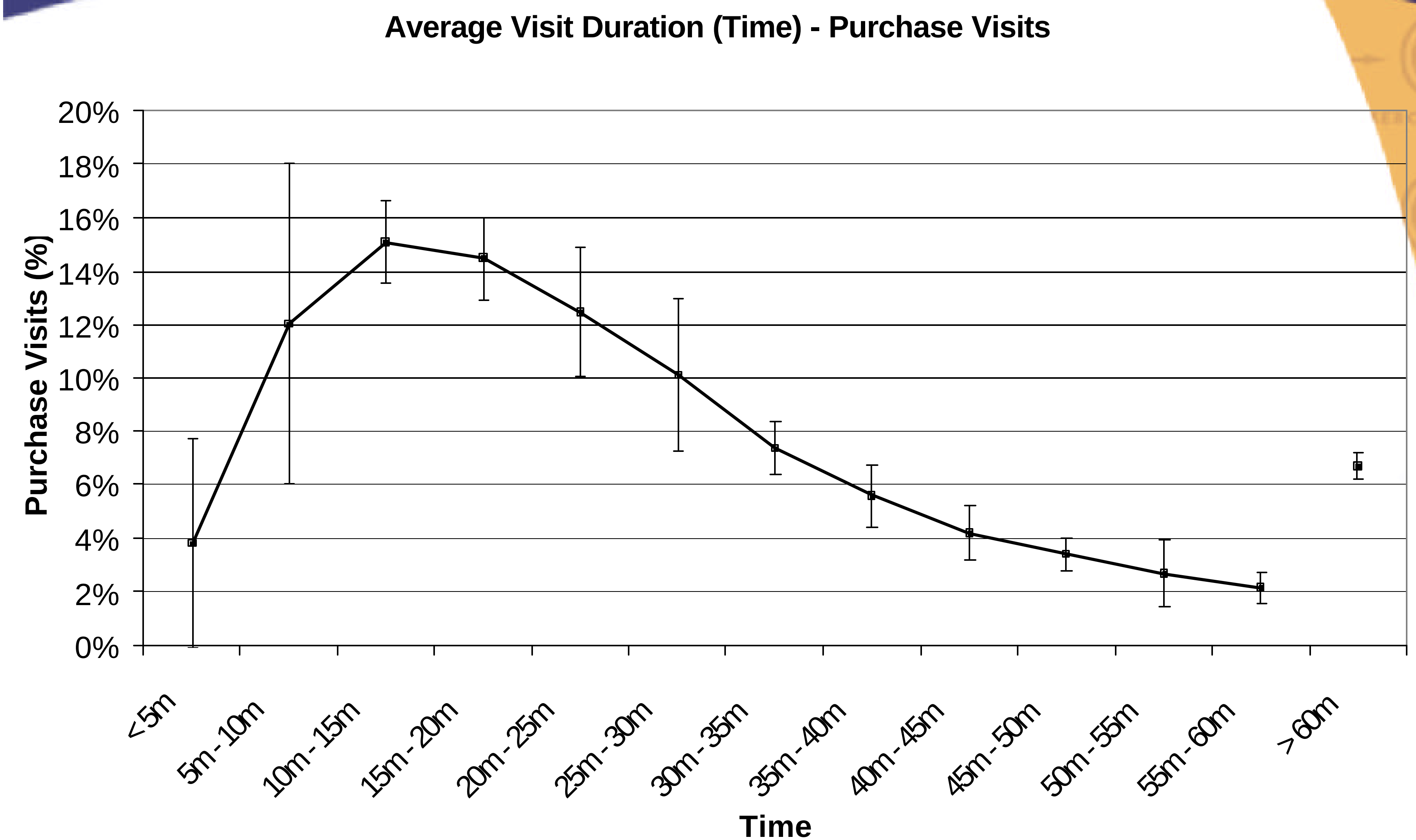
Lesson:

Like TV advertising, a web site needs to capture a visitor in under a minute.

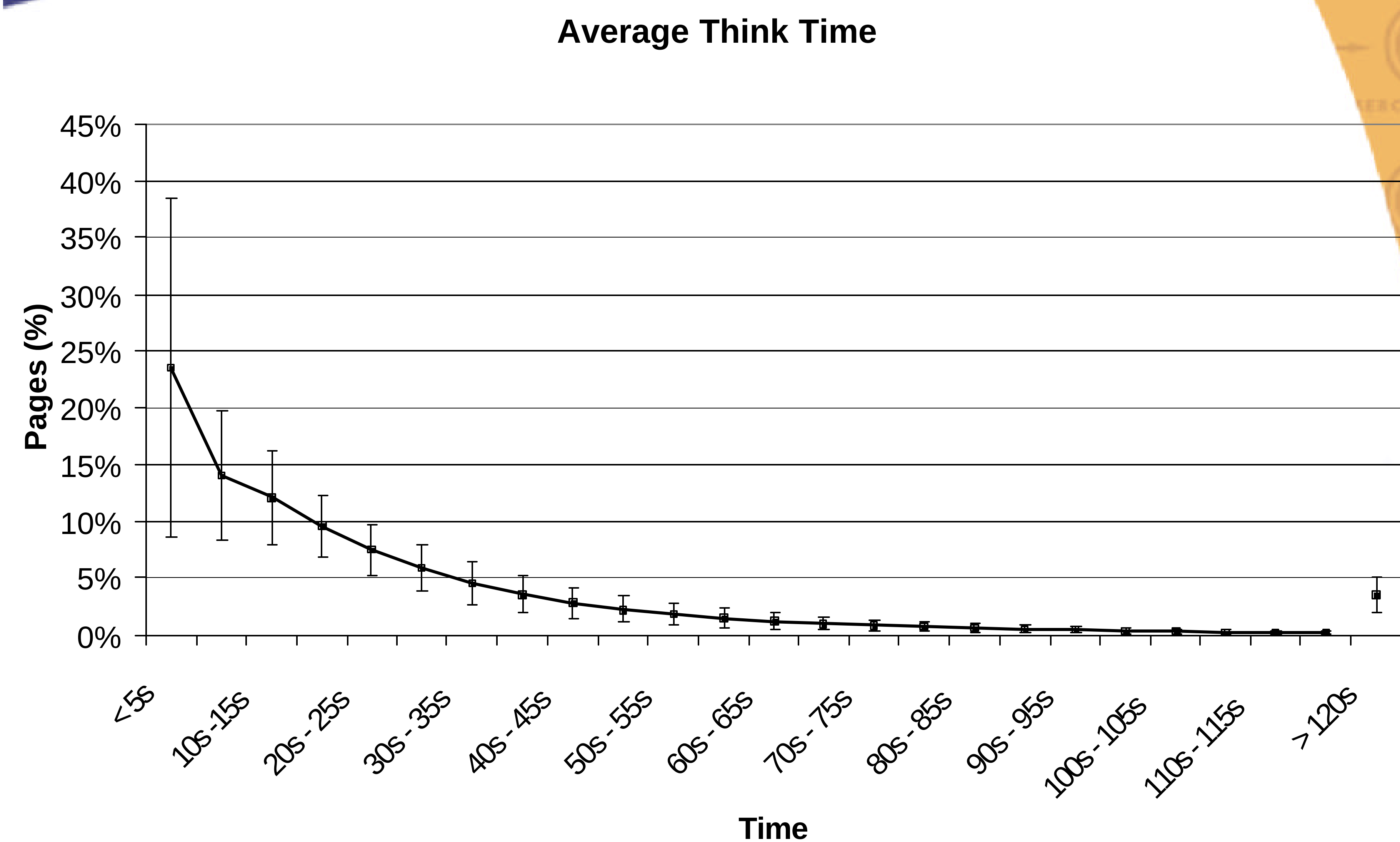
Make the home page fast, useful and compelling.

Note the low std-dev for data 2 minutes or higher

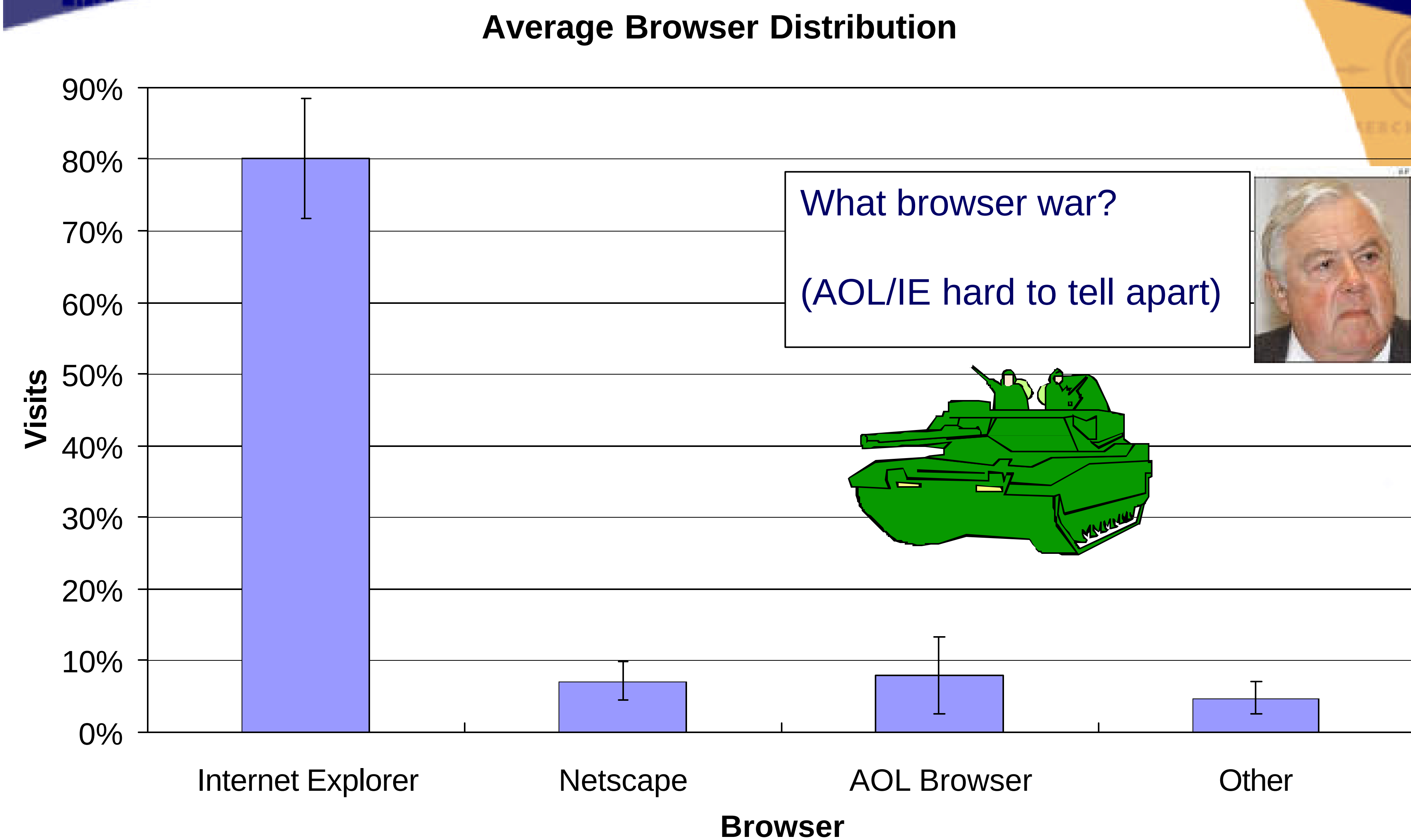
Purchase Visit Duration - Time



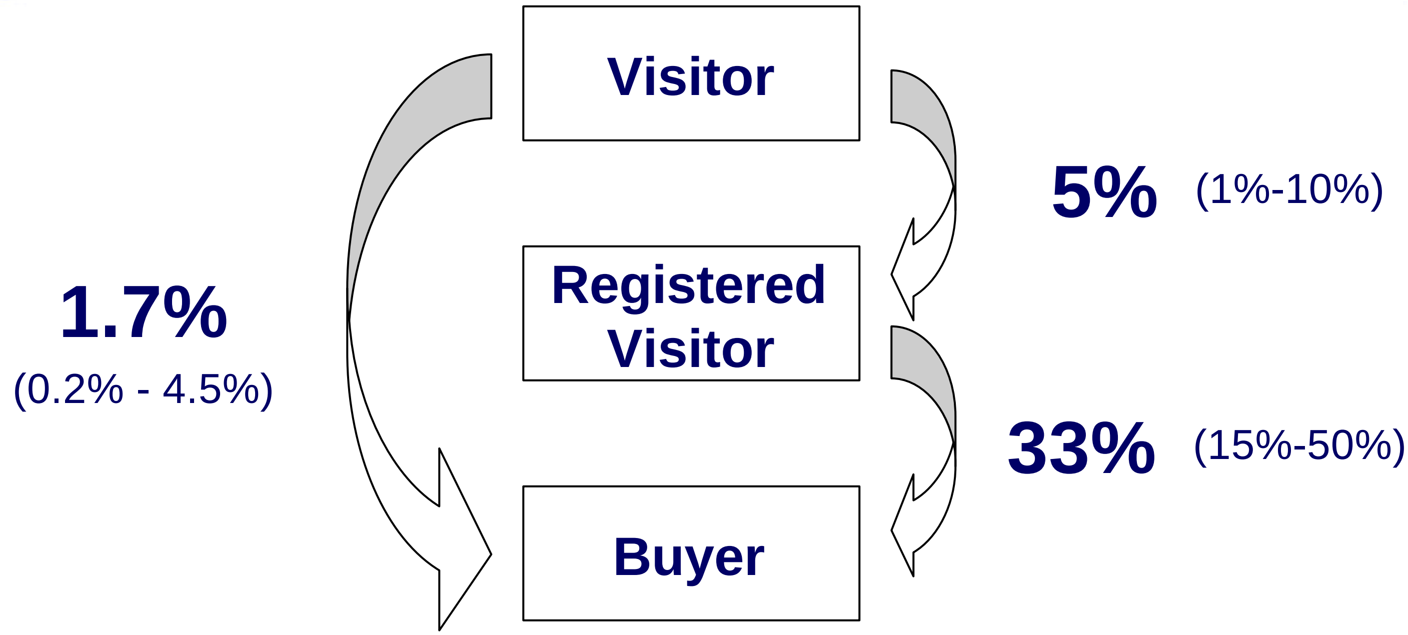
Think Time


Average Think Time
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Pages (%)
<5s
10s-15s
20s - 25s
30s - 35s
40s - 45s
50s - 55s
60s - 65s
70s - 75s
80s - 85s
90s - 95s
100s - 105s
110s - 115s
>120s
Time

Browser Distribution



Conversion Rates



Lesson:
Few visitors bother to register.
Make it easy, fast and give incentives.

Lesson:
Many people who bother to register don't purchase.
You know who they are, so you can target them.

Business Events

- Shopping cart abandonment rate

  34% (15% - 50%)

Lesson:
You know what they want, and sometimes know who they are.
You can target them.

- Search

  6% of visits use search (1%-10%)

  10% of searches fail to find any results

Lesson:
You can find out what they were looking for.
You can correct the problem.

Teaser - Privacy

- **92% of Americans are concerned (67% very concerned) about the misuse of their personal information on the Internet.**

  - FTC Report, May 2000

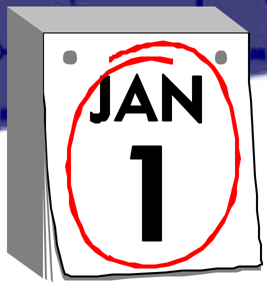
- **86% of executives don't know how many customers view their privacy policies.**

  - Forrester Report, November 2000

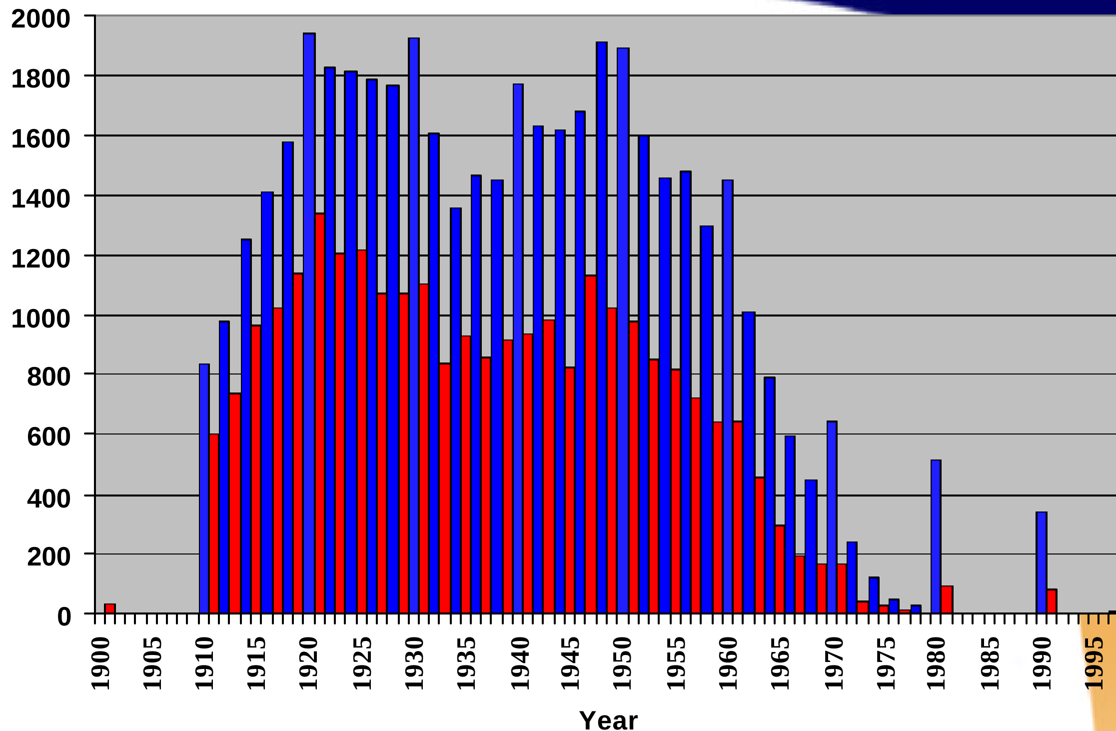
- **Q: What percentage of visitors read the privacy statement?**

  - A: Less than 0.3%

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?

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

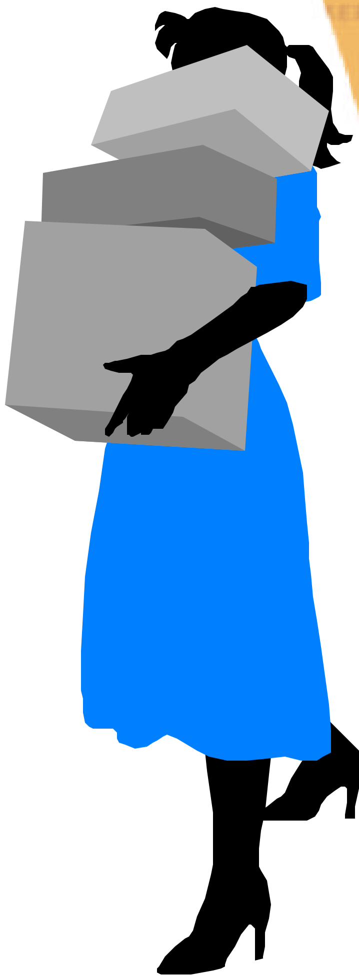
Why?

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

Teaser - Low Conversion Rates

- **Recall that Conversion Rate is the ratio of buyers to browsers.**
- **High conversion rates are desired**
- **Reports showed some products have really low conversion rates?**

Why?



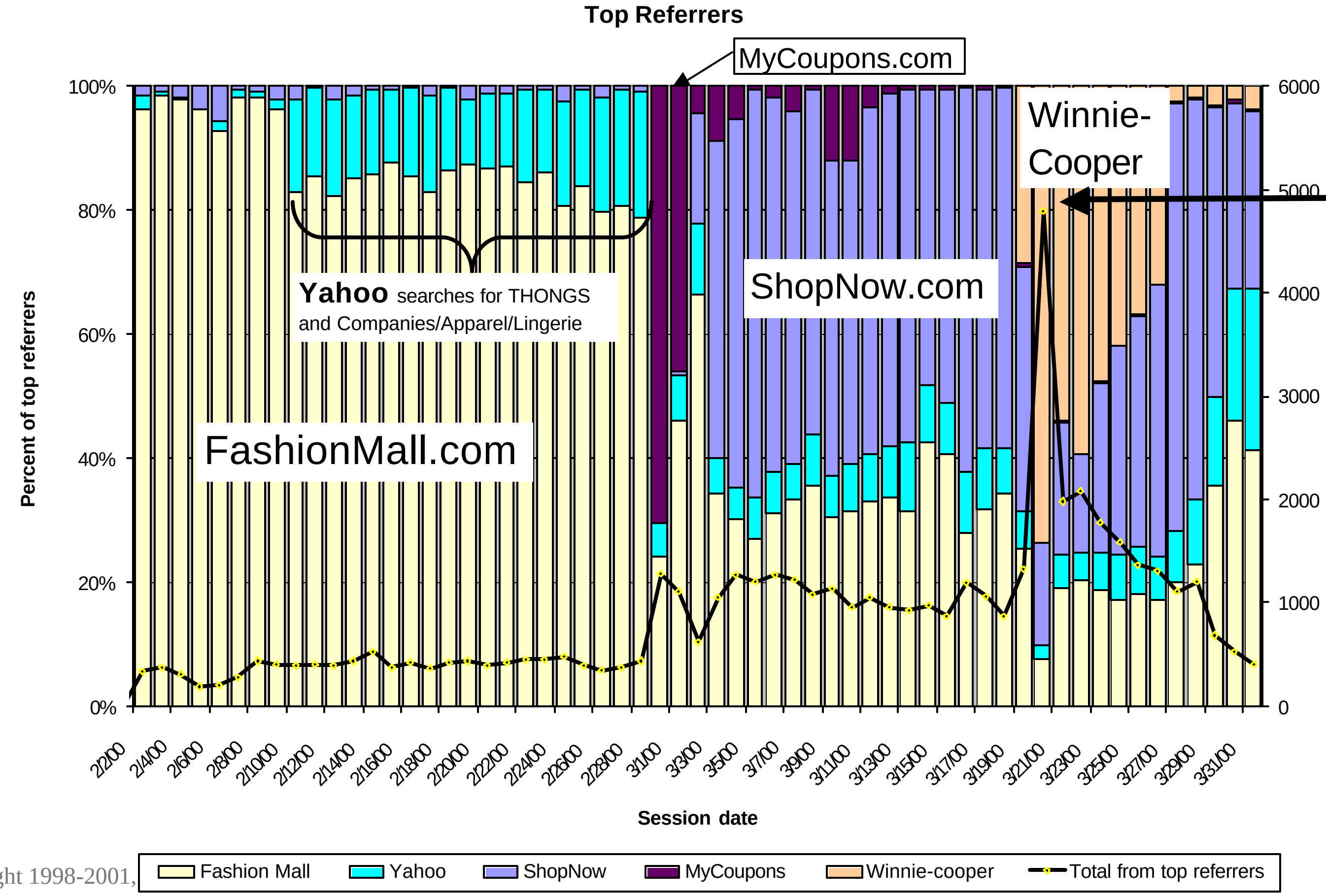
Teaser - High Conversion Rates

- **Product Conversion Rate is the ratio of product purchases to product views**
- **High can conversion rates be over 100%**

Teaser - Who is Winnie?

Referring site traffic for Gazelle.com, a leg-wear and leg-care web retailer. From KDD Cup 2000

Who is Winnie Cooper? What can you do about it?



Note spike in traffic

Answer to Teaser

- Winnie-cooper is a 31 year old guy who wears pantyhose
- He has a pantyhose site
- 8,700 visitors came from his site in a few days (!)
- 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

Resources (I)

- 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
- WEB Mining Tutorials
  - **E-commerce and Clickstream Mining, Jon Becher and Ron Kohavi, First SIAM International Conference on Data Mining, 2001**
    http://robotics.Stanford.EDU/~ronnyk/miningTutorialSlides.pdf
  - **Web Mining for E-Commerce, Jaideep Srivastava, The Fifth Pacific Asia Conference on Knowledge Discovery and Data Mining, 2001**

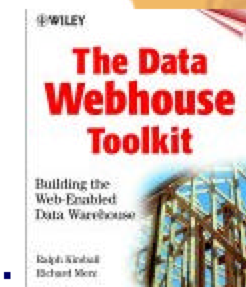
Resources (II)

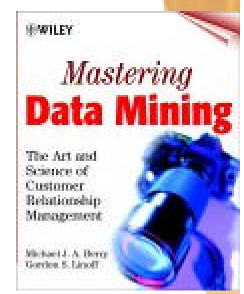
- The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse by Ralph Kimball, Richard Merz. ISBN: 0471376809 (Jan 2000)
- Mastering Data Mining: The Art and Science of Customer Relationship Management by Michael J. Berry, Gordon Linoff. ISBN: 0471331236
- The Data Mining and Knowledge Discovery special issue on Application of Data Mining to Electronic Commerce (volume 5, 1/2) January/April 2001. Special issue:

  http://www.wkap.nl/issuetoc.htm/1384-5810+5+1/2+2001

  Book ISBN: 0792373030

  http://www.amazon.com/exec/obidos/ASIN/0792373030





Resources (III)

- 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 (IV)

- 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

Take Home Messages (I of III)

- **Direct effect of web on established retailers may not be large, but lessons learned will affect other channels, so additional ROI comes from improvements to other channels**
- **The webstore provides an experimental laboratory and a trend-discovery system**
  - **Which cross-sells work?**
  - **Which ads are effective?**
  - **What are people looking for (failed searches for pokédex)**

Take Home Messages (II of III)

- **Good:** **E-commerce is the killer-domain for data mining with all the right ingredients**
- **Bad:** **Good data collection is hard**
  - **Web logs are information poor**
  - **New sites should log clickstream and events in the app**
- **Ugly:**
  - **Data transformations take longer than you expect.**
  - **You must "peel the onion" for interesting insight**
    See KDD CUP 2000 at
    http://www.ecn.purdue.edu/KDDCUP

Take Home Messages (III)

- **Always involve the business user**

  Many "interesting" discoveries turn out to be a result of some intentional activity. "Peel the onion."

- **Business users want simple, comprehensible results**
  - **Reports are not glamorous but most often needed**
  - **Simple algorithms are most useful especially if coupled with good visualizations**
- **The web is a measurement and experiments lab**
  - **Half the discoveries will carry over to the "real world"**

Acknowledgement

- **Many thanks to the data mining team members at Blue Martini Software**
- **Special thanks to Llew, Zijian, Brian, and Eric, who worked on the e-metrics project**
- **A copy of this talk is available at** www.kohavi.com
- The reference for the talk/paper is
  Ron Kohavi. Mining e-commerce data: The good, the bad, and the ugly. In Foster Provost and Ramakrishnan Srikant, editors, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2001.
  http://robotics.Stanford.EDU/users/ronnyk/goodBadUglyKDDItrack.pdf