BLUE MARTINI SOFTWARE

SAS M2000 Conference
Oct 3-4, 2000

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

Ronny Kohavi, Ph.D.

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
- **Stories from mining real data**
  "Peeling the onion" on observations to yield insight
- **Summary**

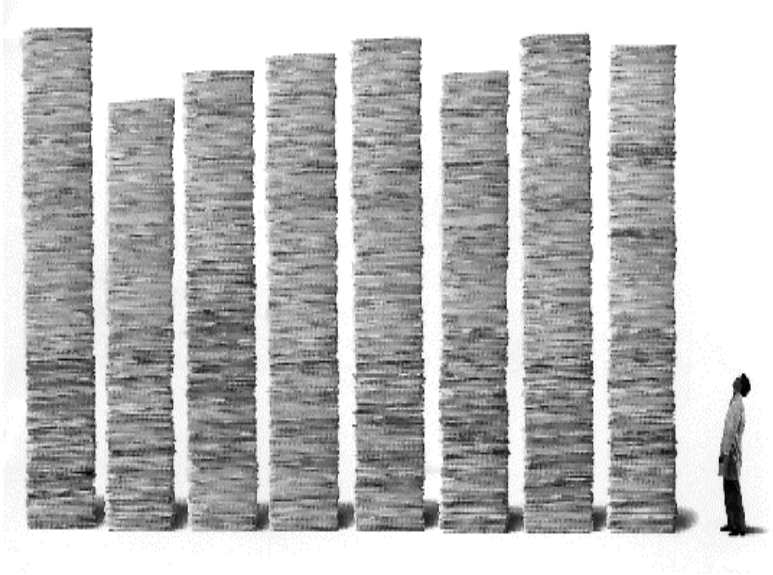
The Killer Domain

BLUE MARTINI SOFTWARE

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)

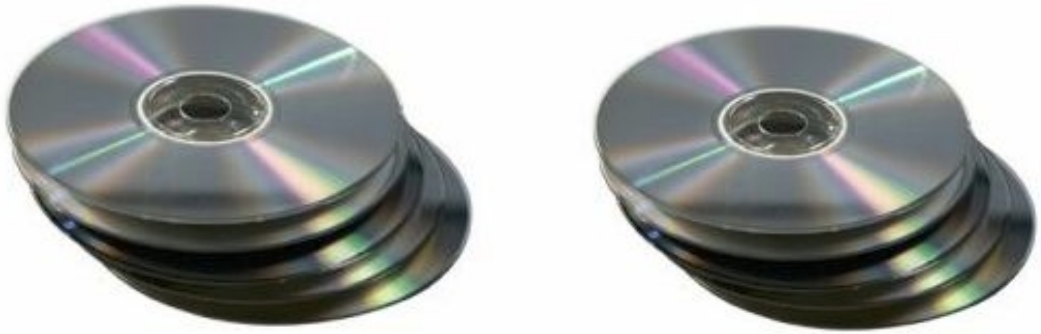
E-commerce has all the right ingredients



Many Records

BLUE MARTINI SOFTWARE

- Clickstreams generate huge amounts of data
- New e-commerce sites, even if small, generate sufficient data for effective mining quickly
- Yahoo! serves 680 million page views a day. Web log data for page views is **6GB per hour!**



Rich Records

BLUE MARTINI SOFTWARE

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

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

Clean Data

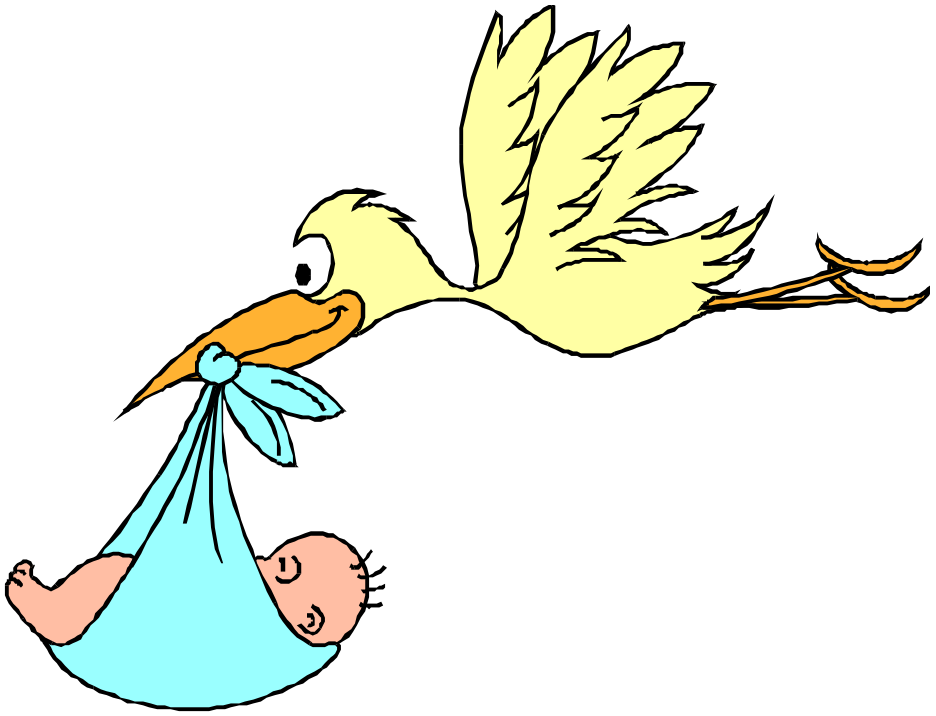
- **Collect data directly at webstore**
  No legacy systems
- **Collect what is needed by design**
  Not as an afterthought
- **Collect electronically - reliable data**
  No humans typing survey data from forms
- **Sample at the right granularity level**
  sample at the customer or session,
  never at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

Teaser - Birth Dates

BLUE MARTINI SOFTWARE

A bank discovered that almost 5% of their customers were born on the exact same date: 11 Nov 1911

Why?



Actionable Domain

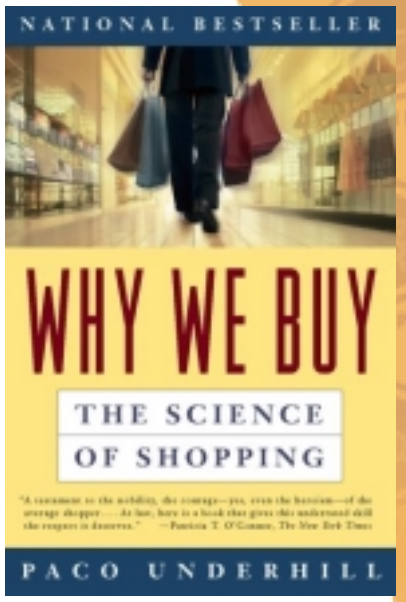
BLUE MARTINI SOFTWARE

- **Few data mining discoveries had a real impact on businesses.**
  Taking action requires changing complex systems, procedures, and human habits - HARD
- **In e-commerce, many discoveries can be made actionable by**
  - **Changing web sites**
  - **Personalizing web sites**
  - **Changing advertising strategies**
- **Easy to offer cross-sells or up-sells**
  Contrast with changing actual store layouts

Measure ROI

BLUE MARTINI SOFTWARE

- **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 Bad

BLUE MARTINI SOFTWARE

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?

BLUE MARTINI SOFTWARE

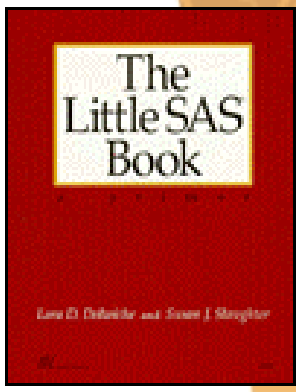
Given a URL, what was displayed?

- **Reverse URL mapping. Very brittle.**

  http://www.amazon.com/exec/obidos/ASIN/1580252397/o/qid=967793582/sr=2-3/103-0457188-1595821

  is *The Little SAS Book: A Primer*

  How can you derive attributes of the book?



- **Reverse content mapping**

  Sniff the packets and try to map content to attributes

- **Don't use web logs - let the application server log**

Dynamic Content is Harder

BLUE MARTINI SOFTWARE

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%400082261715 9.0968100982%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

BLUE MARTINI SOFTWARE

- HTTP is stateless
- Sessionizing is still a research topic
- Recreating user sessions is heuristic based:
  - IP addresses
  - Cookies
  - Browser type

Business Events

BLUE MARTINI SOFTWARE

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**
- **Dynamically constructed media (e.g., Flash)**
- **Search - common keywords or keywords that were not found**
  (an important warning to an e-commerce site)

Matching Web Logs to DB

BLUE MARTINI SOFTWARE

- 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

BLUE MARTINI SOFTWARE

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

BLUE MARTINI SOFTWARE

- **On one of our sites, we saw 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

BLUE MARTINI SOFTWARE

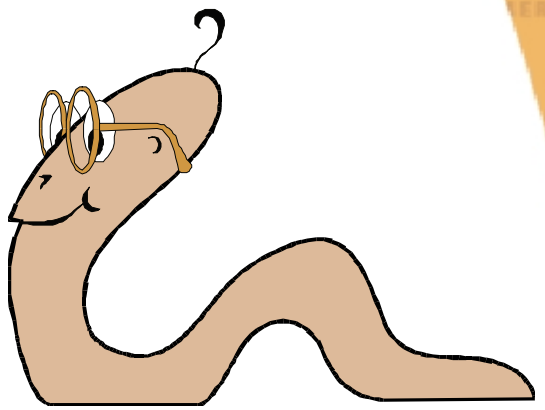
- **Ignore web logs**
  They are at the wrong granularity level to be useful
- **Log the information yourself at the application layer (e.g., Blue Martini's solution)**
  - **The application knows what is on the page**
  - **The app controls sessions**
  - **The app can log business events**
  - **The 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 (previously mentioned)**
  - **Data transformations for analysis**
  - **Marketing-level insight**

Crawlers

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

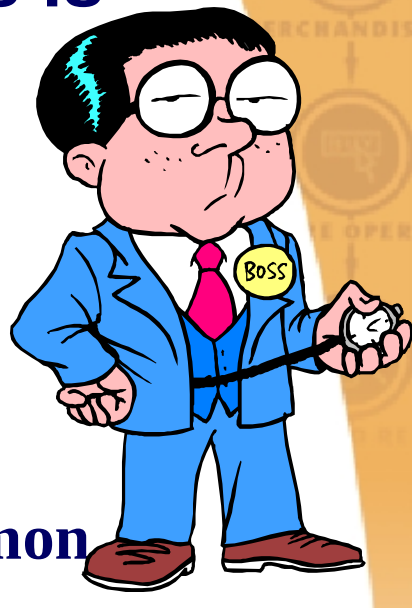


- **For understanding your customers, it is very important to filter out crawlers. Fairly hard!**

Data Transformations

BLUE MARTINI SOFTWARE

- 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

BLUE MARTINI SOFTWARE

- **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?

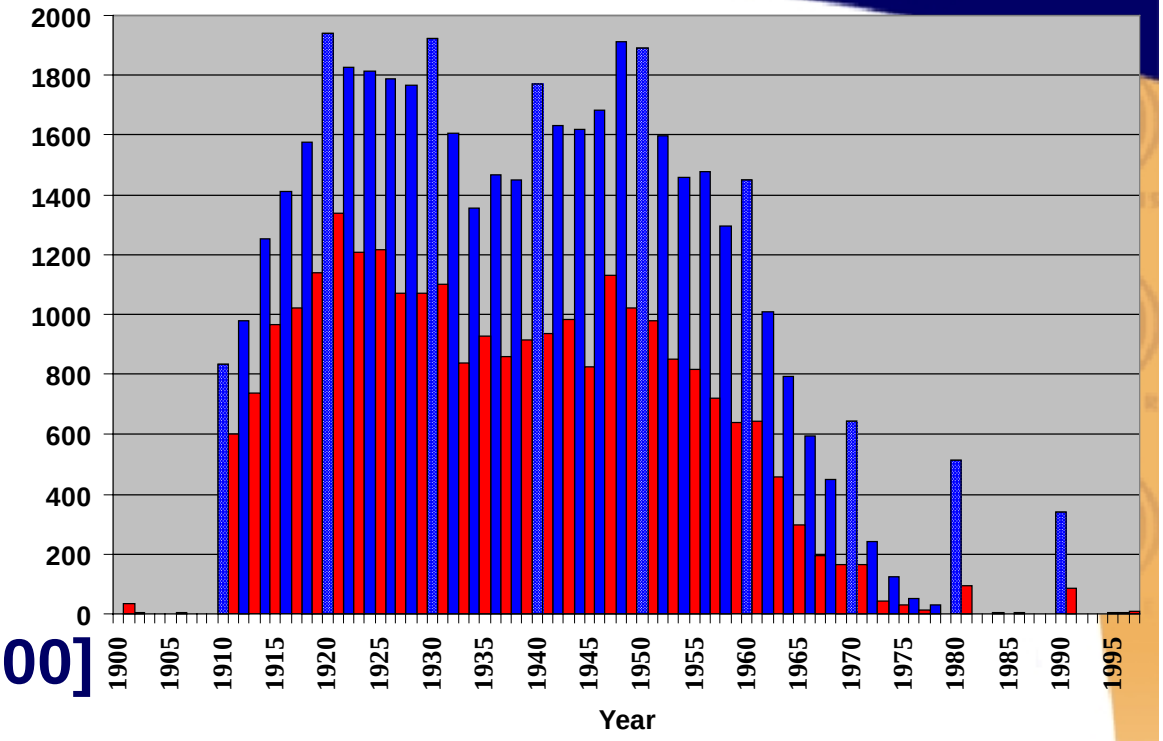
- **Insight must be comprehensible to biz users**

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

Teaser - Mysterious Birth Years

BLUE MARTINI SOFTWARE

The KDD-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 II - Gender Mystery

BLUE MARTINI SOFTWARE

- 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?

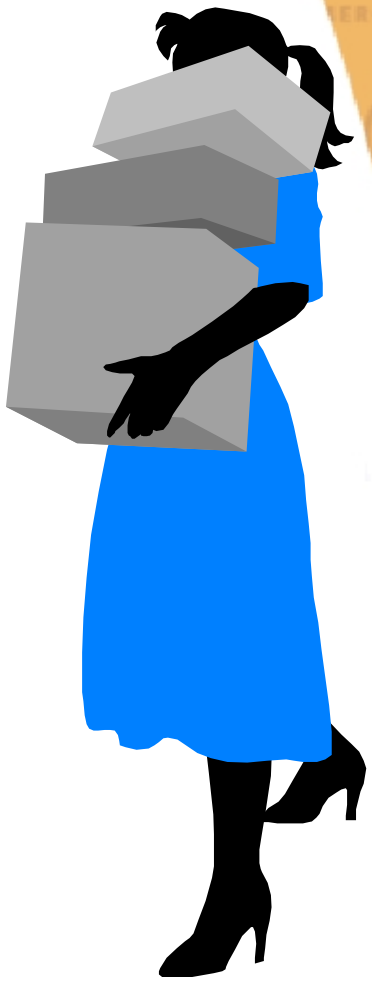
Hint: Acxiom only conflicted with females, claiming some females are males.
Never in the other direction

Teaser III - Low Conversion Rates

BLUE MARTINI SOFTWARE

- **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?

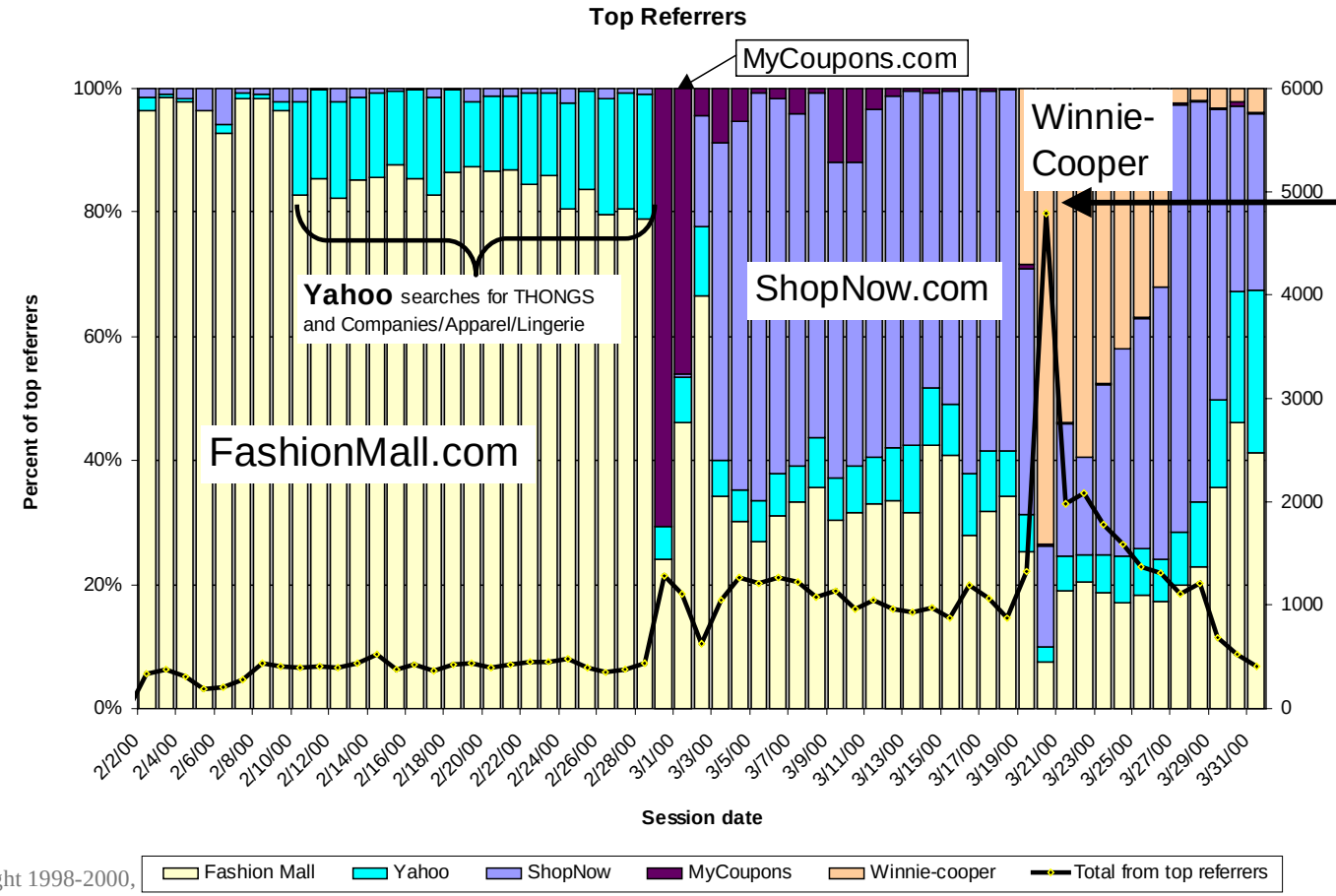


Teaser IV - Who is Winnie?

BLUE MARTINI SOFTWARE

Referring site traffic for Gazelle.com, a webleg and webcare web retailer. From KDD Cup 2000

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



Note spike in traffic

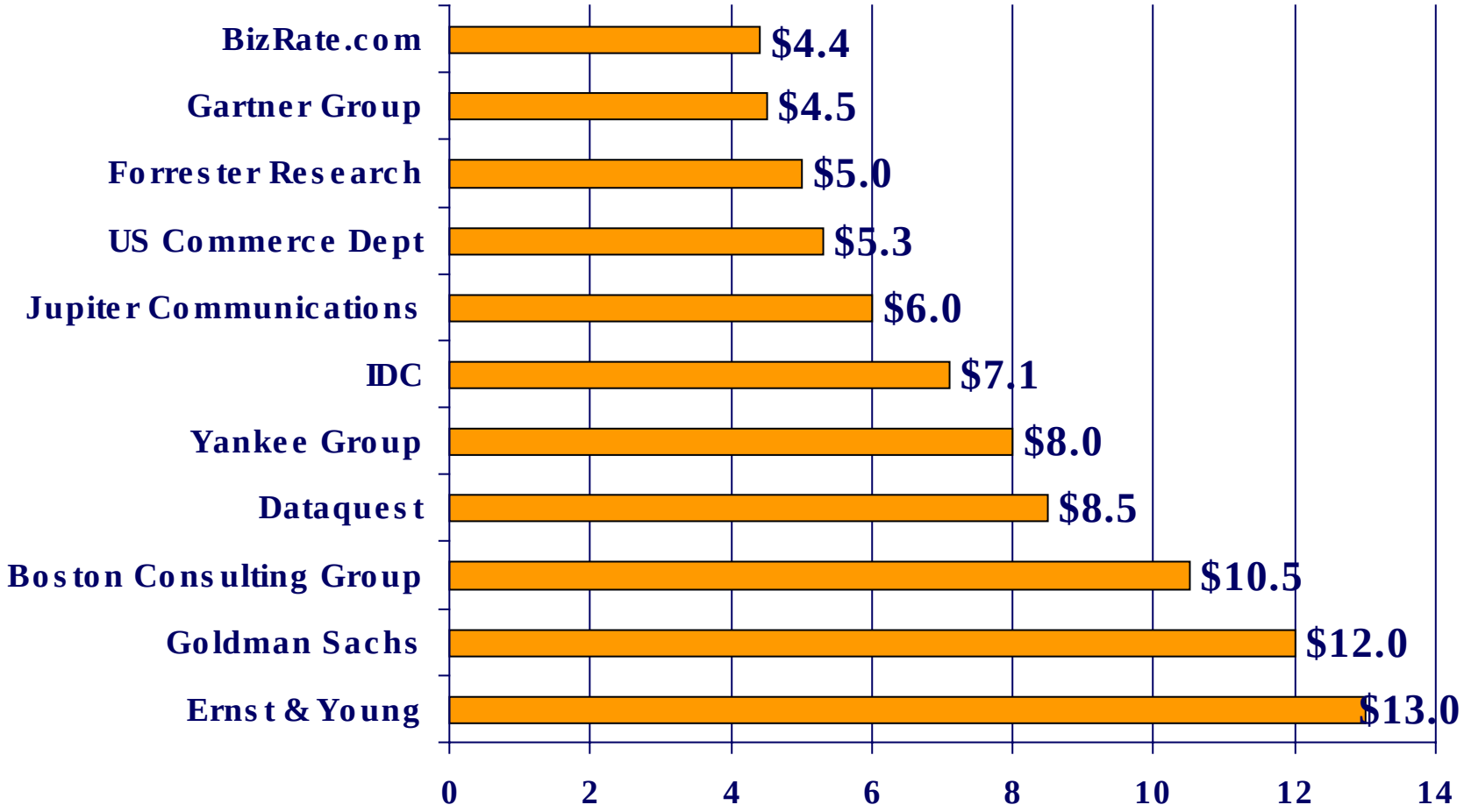
Answer to Teaser IV

- 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

Teaser V - Analyst numbers

BLUE MARTINI SOFTWARE

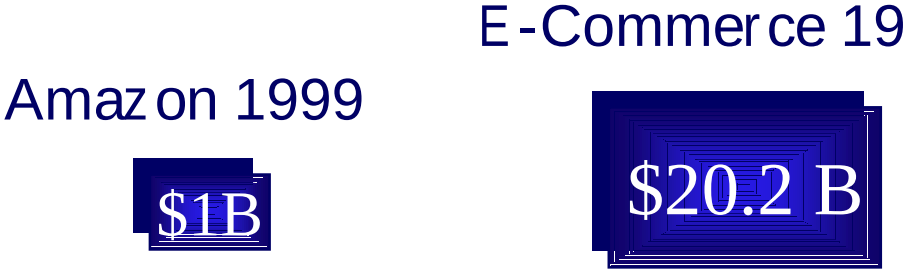
4Q 1999 Online Holiday Shopping Revenues (Billions). Why don't the numbers agree?



The Webstore is an Experimental Laboratory

BLUE MARTINI SOFTWARE

- **Effect of web on established retailers is small**
- **However, lessons learned will affect 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)**



Wal-Mart
1999 revenues:
$162.8 B

Take Home Messages (I of II)

BLUE MARTINI SOFTWARE

- 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**
  - **Existing sites should extract data from HTML traffic (e.g., sniffer packages). Plan to upgrade to a better architecture**
- Ugly:
  - **Data transformations take longer than you expect.**
  - **You must "peel the onion" for interesting insight**
    (see KDD CUP 2000 http://www.ecn.purdue.edu/KDD

Take Home Messages (II)

BLUE MARTINI SOFTWARE

- **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"**