KDD-Cup 2000
Peeling the Onion

Carla Brodley, Purdue University

Ronny Kohavi, Blue Martini Software

Co-Chairs

Special thanks to Brian Frasca, Llew Mason, and Zijian Zheng from Blue Martini engineering; Catharine Harding and Vahe Catros, our retail experts; Sean MacArthur from Purdue University; Gazelle.com, the data provider; and Acxiom Corporation, the syndicated data provider.

I See Dead People

What is wrong with this statement?

Everyone who ate pickles in the year 1743 is now dead.

Therefore, pickles are fatal.

Correlation does not imply causality

Harder Example

BLUE MARTINI SOFTWARE

True statement (but not well known):

Palm size correlates with your life expectancy

The larger your palm, the less you will live, on average.

Try it out - look at your neighbors and you'll see who is expected to live longer.

Why?

Women have smaller palms

and live 6 years longer on average

Peeling the Onion

The #1 lesson from the KDD Cup 2000

Peel the Onion:

Don’t stop at the first correlation.
Ask yourself (and the data) WHY?

Most of the entries did not identify the fundamental reasons behind the correlations found

Overview

- Data Preparation
  - The Gazelle site
  - Data collection
  - Data pre-processing
  - The legalese
- Statistics
- The five tasks & highlights from each
- Winners talk (5x5 minutes)

Detailed poster by winners and organizers tomorrow, Monday, 6 - 7:30PM

The Gazelle Site

- Gazelle.com was a legwear and legcare web retailer.
- Soft-launch: Jan 30, 2000
- Hard-launch: Feb 29, 2000
  with an Ally McBeal TV ad on 28th
  and strong $10 off promotion
- Training set: 2 months
- Test sets: one month
  (split into two test sets)

Data Collection

- Site was running Blue Martini's Customer Interaction System version 2.0
- Data collected includes:
  - Clickstreams
    - Session: date/time, cookie, browser, visit count, referrer
    - Page views: URL, processing time, product, assortment (assortment is a collection of products, such as back to school)
  - Order information
    - Order header: customer, date/time, discount, tax, shipping.
    - Order line: quantity, price, assortment
  - Registration form: questionnaire responses

Data Pre-Processing

- Acxiom enhancements: age, gender, marital status, vehicle lifestyle, own/rent, etc.
- Keynote records (about 250,000) removed.
  They hit the home page 3 times a minute, 24 hours.
- Personal information was removed, including:
  Names, addresses, login, credit card, phones, host name/IP, verification question/answer.
  Cookie, e-mail were obfuscated.
- Test users were removed based on multiple criteria (e.g., credit card number) not available to participants
- Original data and aggregated data (to session level) were provided

Legalese

- Concern from both the Gazelle and Blue Martini about legal exposure
- Created NDA (non-disclosure agreement), which was designed to be simple - half page. We used efax to get faxes of signed signatures
  - One large company sent us back a 4-page legal agreement on watermark paper describing details such as stock ownership of Blue Martini subsidiaries.
    Others from that company signed anyway
  - One person asked to void his signature after two weeks because he is not a “functional manager”

KDD Cup Cruise?

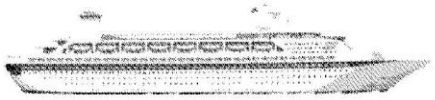
Received: 07.Jul.00 07:45 AM From: UnknownSender To: 6034158866 Powered by eFax.com Page: 1 of 1

07/07/00 05:41:59 -> (603)415-8866 <-1 603 415-8866-> Page 001

ENJOY 7 DAYS AND 6 NIGHTS
FLORIDA AND ALL INCLUSIVE BAHAMAS CRUISE PLUS
CANCUN OR HAWAII GETAWAY
WITH FREE ROUND TRIP AIRLINE TICKETS!

TO: All Employees
FROM: Sheryl Stein, Corporate Department
RE: Very limited supply of tickets (1st come first served basis)
MEMO: Because of the overwhelming amount of inquiries from employees we have decided to run this promotion. Please distribute this memo freely.

CALL TODAY !!!
1-800-834-1759



CALL TODAY !!!
1-800-834-1759

$299.00

* Per person plus applicable port & service charges

ENJOY 7 DAYS & 6 NIGHTS:
2 NIGHT Unforgettable Cruise to Bahamas on the S.S. Dolphin IV or Oceanbreeze
2 NIGHTS - In Cocoa Beach, Florida (Kennedy Space Center and *Beautiful* white sand beaches)
2 NIGHTS - In Orlando, Florida - The Vacation Capital of the world
(minutes away from All the attractions)
Ask your Travel Coordinator about Complimentary Rental Car for One Week!

BONUS:
By making Reservations Today You Will Also Receive Free Round Trip Airline Tickets
for "TWO PEOPLE" TO:
CANCUN, MEXICO
OR HONOLULU, HAWAII

- Price is $299.00 per person
- This price includes ALL VACATIONS and bonuses above.
- Tickets are good for 12 Full Months and are extendible and fully transferable.
- The airline tickets cannot be used in conjunction with the Florida vacation.
- Double occupancy is required.

TO BOOK RESERVATIONS CALL @ 1-800-834-1759
If busy, keep trying. Hours of Operation 8AM - 8PM M-F and 10AM - 5PM Saturday (E.S.T.)

If you received this fax in error and would like to have your number removed from our database, call toll-free at 800-992-5329.

Received: 10.Jul.00 08:37 AM From: UnknownSender To: 6034158866 Powered by eFax.com Page: 1 of 1

07/10/2000 09:37 16034158866 001

8 DAY 7 NIGHT VACATION

To: All Corporate Employees
From: Corporate Travel

Our wholesale travel department has been asked to forward this travel information to you and your employees. Please distribute this memo to all interested persons. Now on a first-come first-serve basis. You must act quickly to take advantage of this offer with a value of over $3000.00 on a published retail price.

ORLANDO, CANCUN, HAWAII
OR
Choose from Hundreds of Fabulous Resort destinations all over the world!
This vacation can be taken at any time, and is 100% transferable.
This offer is only available during this corporate sellout, for your employees, their families and associates.

$349.00

Per Person
Double Occupancy

+ Tax & Service Fees

BONUS: FOR RESERVING TODAY

COMPLIMENTARY:
4 DAY / 3 NIGHT CARNIVAL CRUISE FOR 2

For Reservations Call Toll Free
1 (800) 423-9242 (ext. 202)

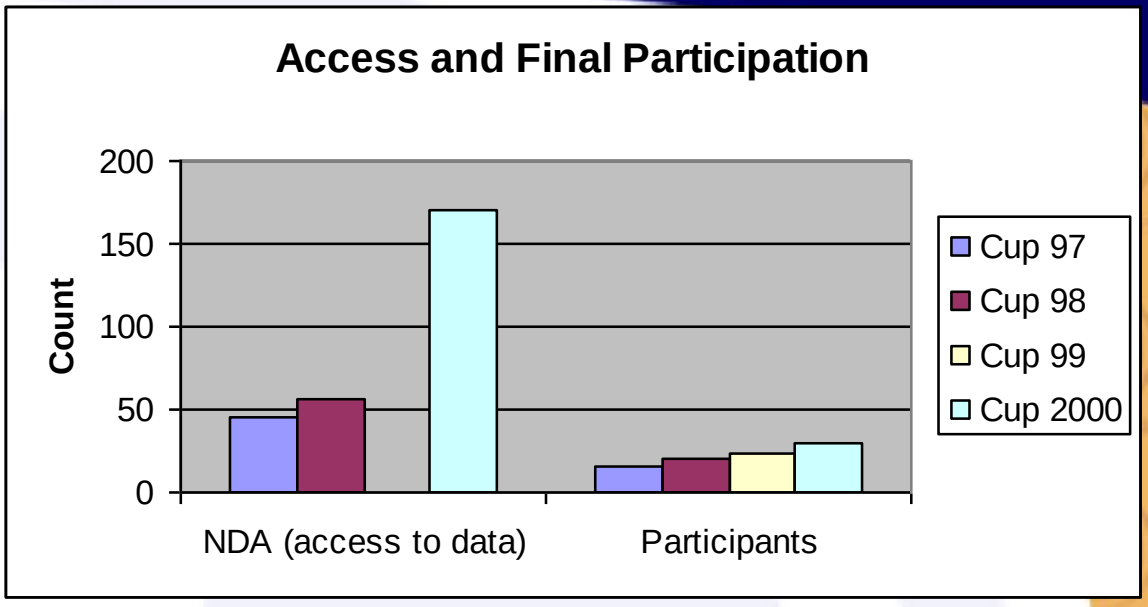
If lines are busy please keep trying!

IF YOU RECEIVED THIS FAX IN ERROR AND WOULD LIKE TO HAVE YOUR FAX NUMBER REMOVED FROM OUR DATABASE, PLEASE CALL US TOLL-FREE AT 1-877-453-8906. FOR ADDITIONAL INFORMATION ON BEING REMOVED, CONTACT THE NATIONAL NO FAX DATABASE AT WWW.NOFAX.COM OR FAX BACK YOUR INFO TO 1-888-499-0739.

And we also got faxes for cheap cruises :-)

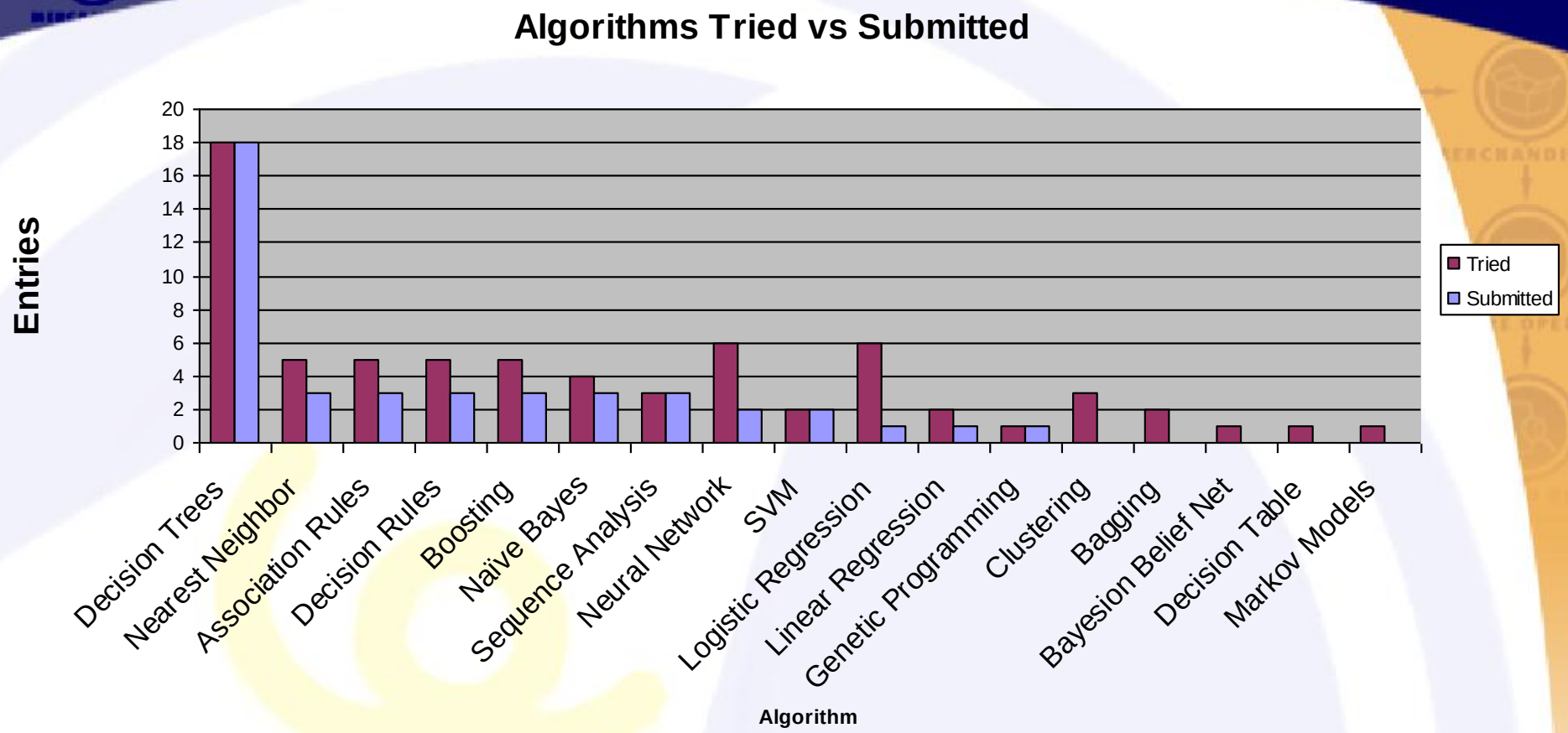
Statistics

- KDD Cup 2000 grew significantly over previous years, especially requests to access the data



- Total person-hours spent by 30 submitters: 6,129
- Average person-hours per submission: 204
  Max person-hours per submission: 910
- Commercial/proprietary software grew from 44% (cup 97) to 52% (cup 98) to 77% (cup 2000)

Statistics II



Decision trees most widely tried and by far the most commonly submitted

Note: statistics from final submitters only

Evaluation Criteria

- Accuracy/score was measured for the two questions with test sets
- Insight questions judged with help of retail experts from Gazelle and Blue Martini
- Created a list of insights from all participants
  - Each insight was given a weight
  - Each participant was scored on all insights
  - Additional factors:
    - Presentation quality
    - Correctness
- Details, weights, insights on the KDD-Cup web page and at the poster session

Question: "Heavy" Spenders

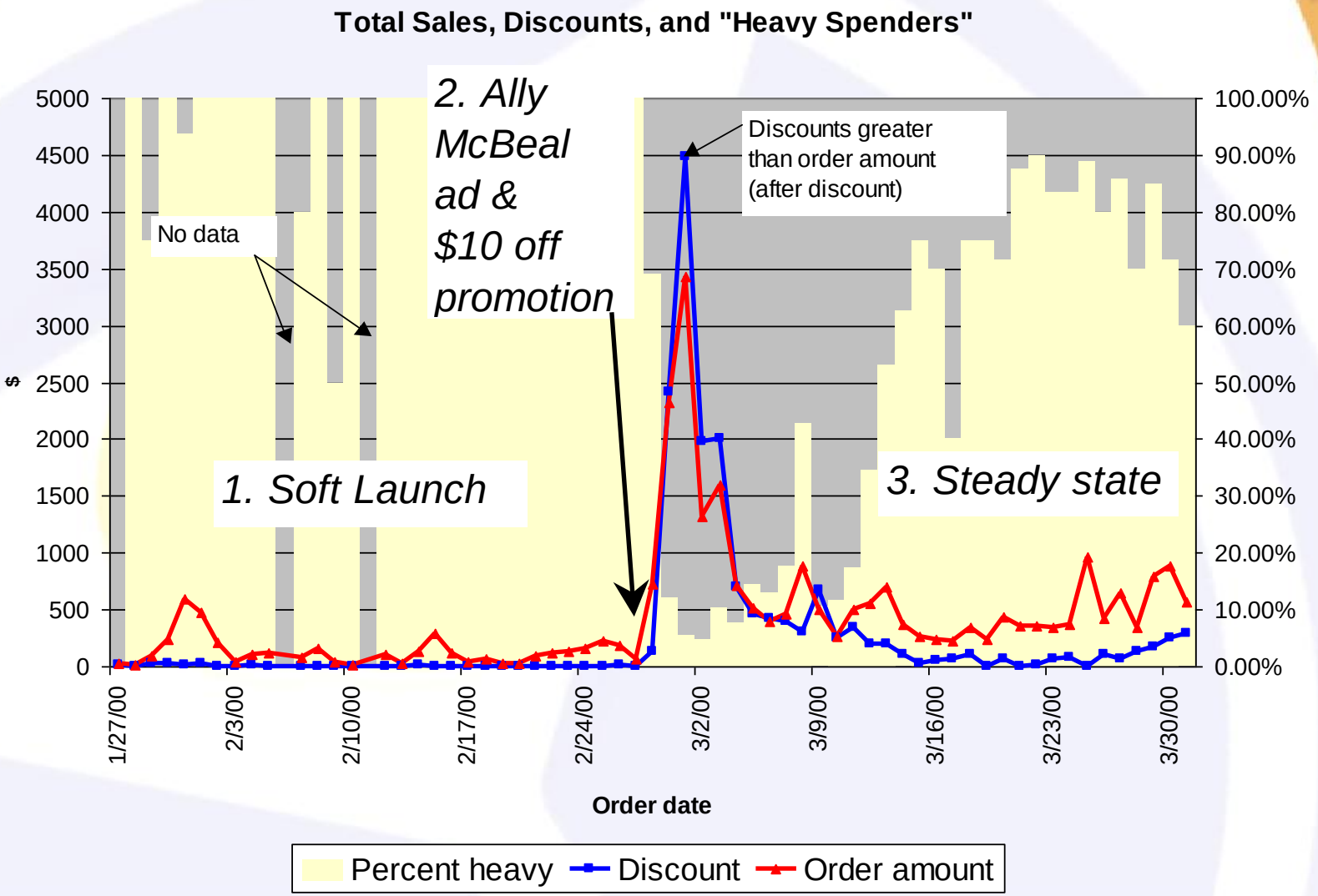
- Characterize visitors who spend more than $12 on an average order at the site
- Small dataset of 3,465 purchases
  1,831 customers
- Insight question - no test set
- Submission requirement:
  - Report of up to 1,000 words and 10 graphs
  - Business users should be able to understand report
  - Observations should be correct and *interesting*

    average order tax > $2 implies heavy spender

    is not interesting nor actionable

Good Insights

Time is a major factor



Good Insight (II)

- 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
    - lifestyle assortments
    - leg-care details (as opposed to leg-ware)

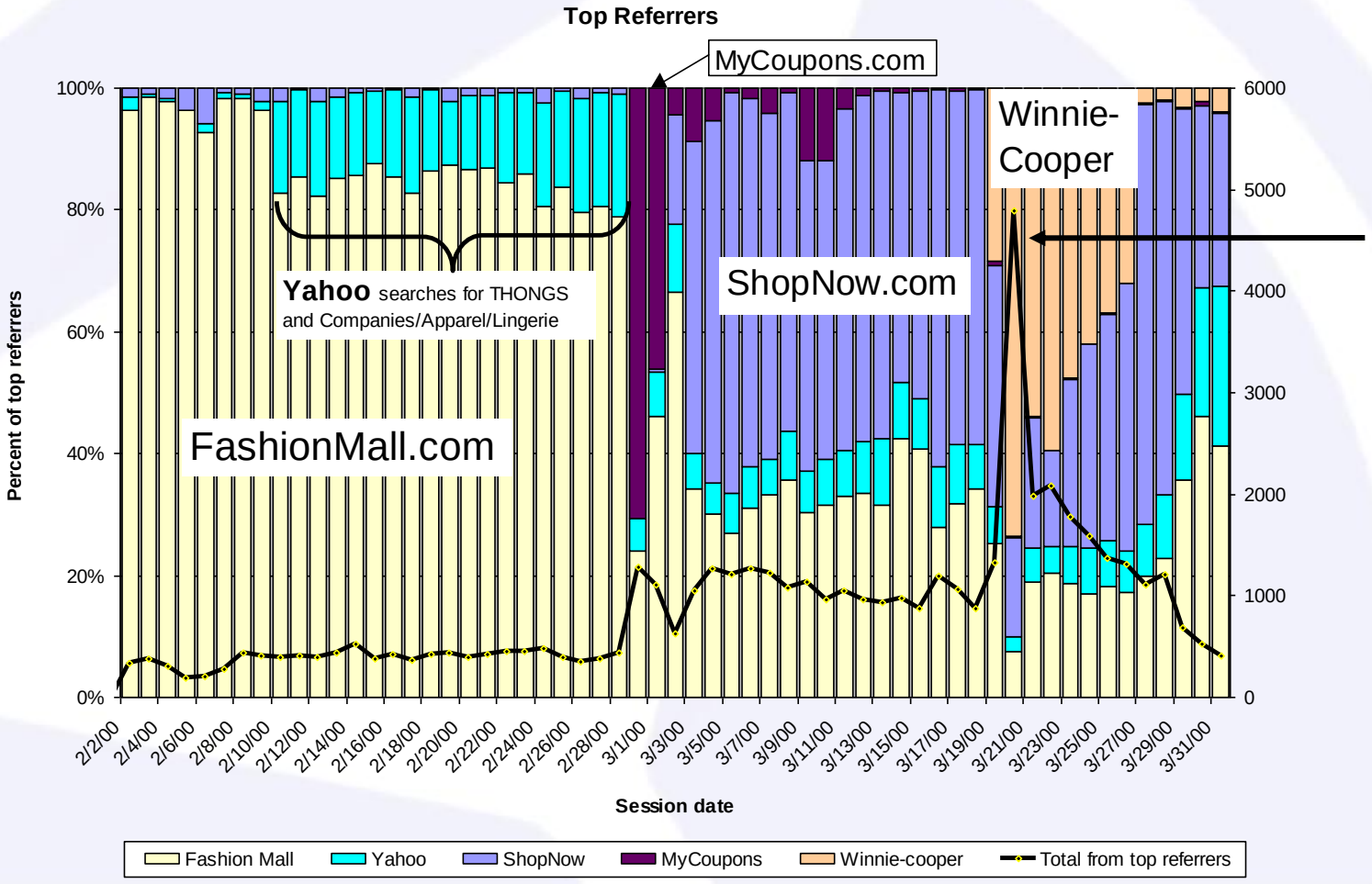
Target segment

Good Insights (III)

BLUE MARTINI SOFTWARE

Referring site traffic changed dramatically over time.

Graph of relative percentages of top 5 sites



Note spike in traffic

Good Insights (IV)

- 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?
    Winnie-cooper is a 31 year old guy who wears pantyhose and has a pantyhose site. 8,700 visitors came from his site (!)

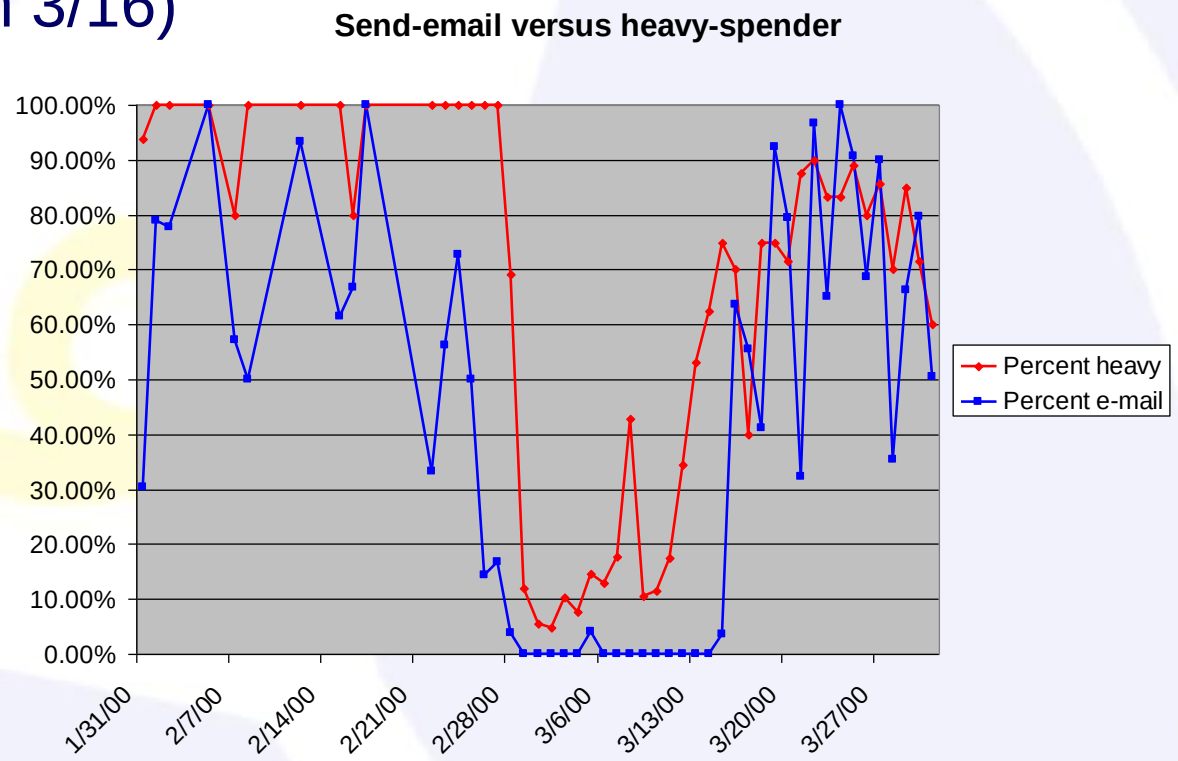
    Actions:
    - Make him a celebrity and interview him about how hard it is for a men to buy in stores
    - Personalize for XL sizes

Common Mistakes

- Insights need support.
  Rules with high confidence are meaningless when they apply to 4 people
- Not peeling the onion.
  Many “interesting” insights with really interesting explanations were simply identifying periods of the site. For example:
  - “93% of people who responded that they are purchasing for others are heavy purchasers”
    True, but simply identifying people that registered prior to 2/28 before the form was changed. All others have null value
  - Similarly, “presence of children" (registration form) implies heavy spender.

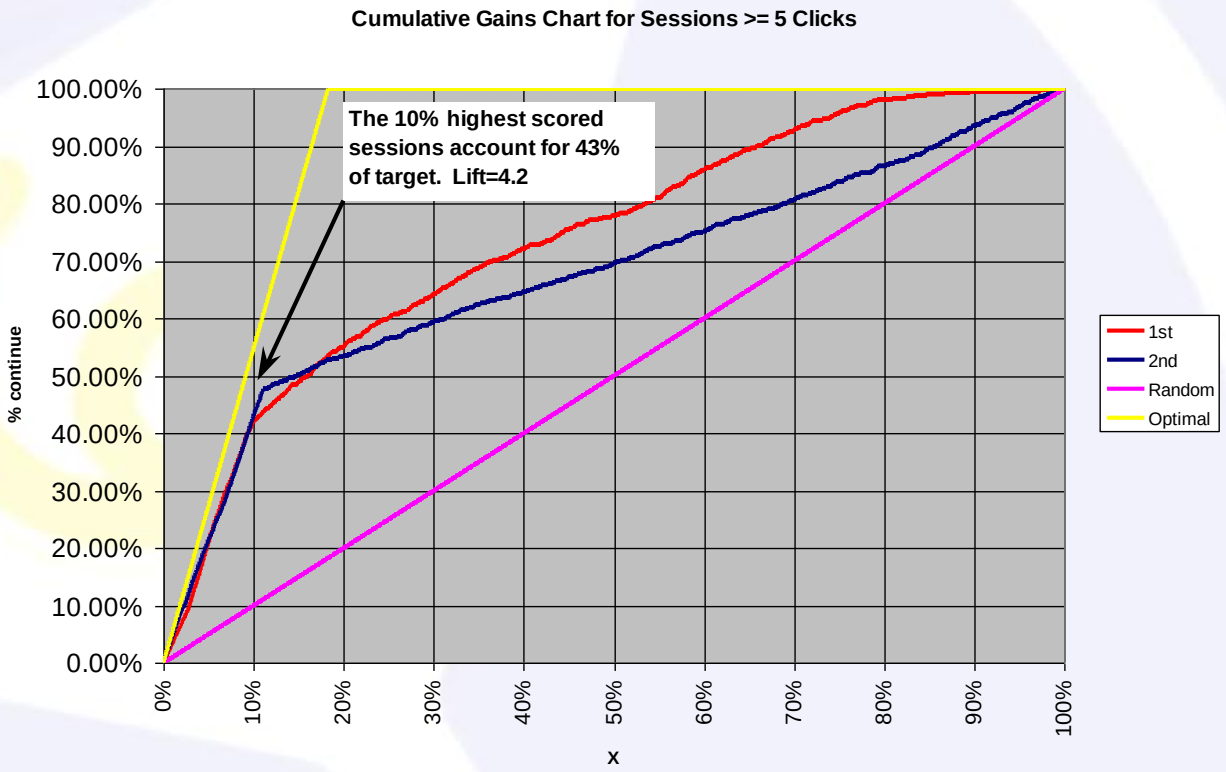
Outer-onion observation

- Agreed to get e-mail in their registration was claimed to be predictive of heavy spender
- It was mostly an indirect predictor of time (Gazelle changed the default for this on 2/28 and back on 3/16)



Question: Who Will Leave

- Given a set of page views, will the visitor view another page on the site or will the visitor leave?
  Very hard prediction task because most sessions are of length 1.
  Gains chart for sessions >=5 is excellent!

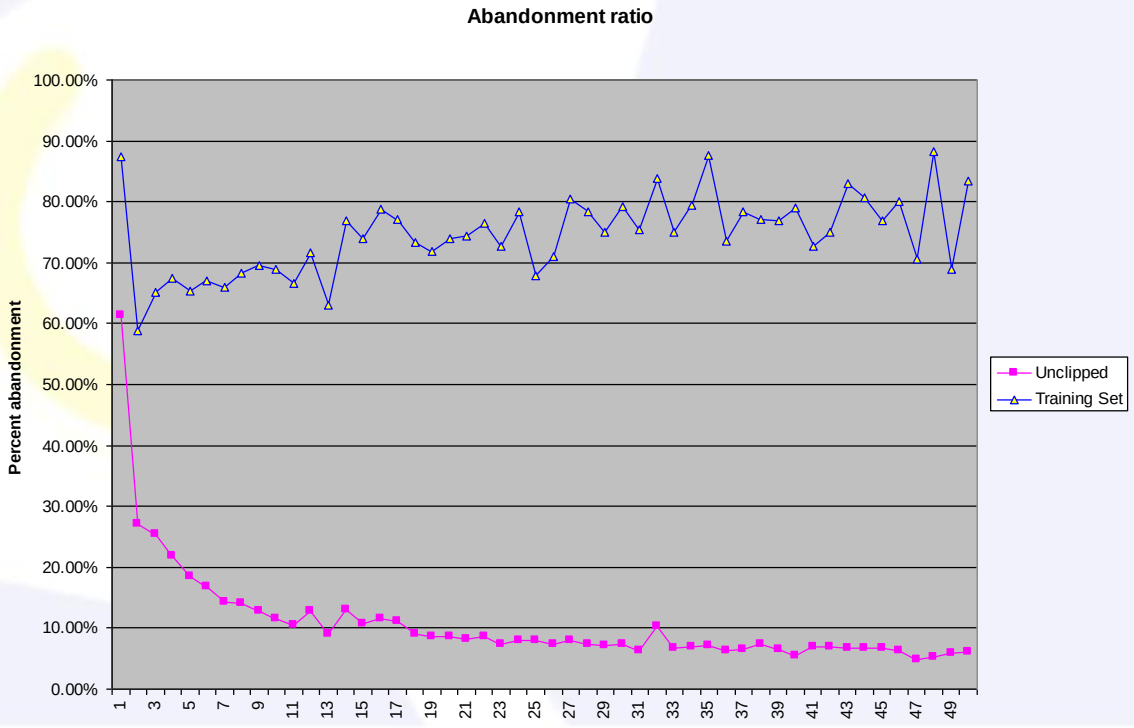


Insight: Who Leaves?

- Crawlers, bots, and Gazelle testers
  Crawlers that came for single pages accounted for 16% of sessions - major issue for web mining!
  Mozilla/5.0 (compatible; MSIE 5.0) had 6,982 sessions of length 1 (there is no IE compatible with Mozilla 5.0)
  Gazelle testers had very distinct patterns and referrer file://c:\...
- Referring sites: mycoupons have long sessions, shopnow.com are prone to exit quickly
- Returning visitors' prob of continuing is double
- View of specific products (Oroblue,Levante) cause abandonment - Actionable!
- Replenishment pages discourage customers. 32% leave the site after viewing it - Actionable!

Insight: Who Leaves (II)

- Probability of leaving decreases with page views
  Many many many “discoveries” are simply explained by this. For example, “viewing three different product implies low abandonement” (need to view multiple pages to satisfy criteria).
- Aggregated training set contained clipped sessions
  Many competitors computed incorrect statistics



Insight: Who Leaves (III)

- People who register see 22.2 pages on average compared to 3.3 (3.7 without crawlers)
- Free Gift and Welcome templates on first three pages encouraged visitors to stay at site
- Long processing time (> 12 seconds) implies high abandonment - Actionable
- Users who spend less time on the first few pages (session time) tend to have longer session lengths

Question: Brand View

- Given a set of page views, which product brand (Hanes, Donna Karan, American Essentials, on none) will the visitor view in the remainder of the session?
- Good gains/lift curves for long sessions (lift of 3.9, 3.4, and 1.3 for three brands at 10% of data).
- Referrer URL is great predictor:
  - Fashionmall.com and winnie-cooper are referrers for Hanes and Donna Karan - different population segments reach these sites
  - mycoupons.com, tripod, deal-finder are referrers for American Essentials - AE contains socks, which are excellent for coupon users
- Previous views of a product imply later views
- Few competitors realized Donna Karan was only available starting Feb 26

Summary (I of II)

- Data mining requires peeling the onion
  - Don't expect to press a button and get enlightenment
    Competitors spent over 200 hours on average.
    Organizers did significant data preparation and aggregation
  - Many discoveries are not causal (pickles example, send-email registration question)
  - Background knowledge and access to business users is a must (TV ads, promotions, change in registration form)
  - Comprehensibility is key - be careful of black-boxes
- Web Mining is challenging: crawlers/bots, frequent site changes

Summary (II of II)

- You can't always predict well, but you can predict when the confidence is high
  (very good gains charts and lifts)
- Many important actionable insights
  - Identifiable Heavy-Spender segments
  - Referrers - change your advertising strategy
    Discover the Winnie-Coopers and mycoupons.com and personalize for them
  - Pages and areas of the site causing abandonment
    (e.g., replenishment page exits should raise a red flag)
  - Site not properly designed for AOL browser
- KDD Cup data will be available for research and education

Next talk

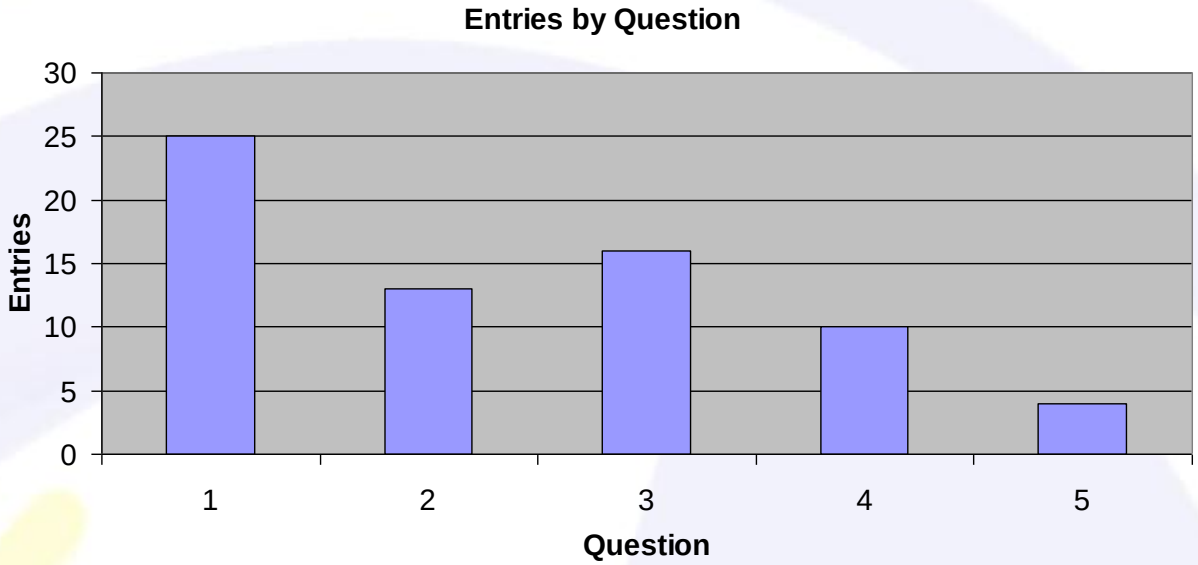
EXTRA SLIDES

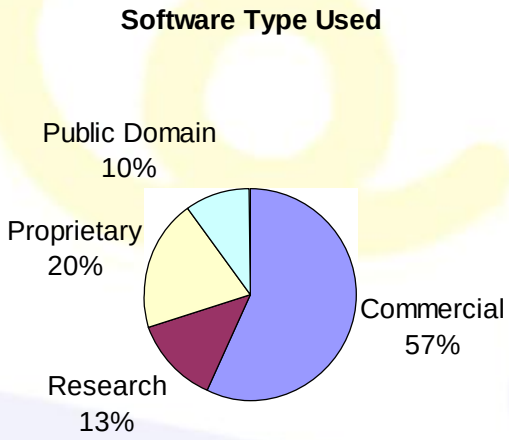
- EXTRA SLIDES
- EXTRA SLIDES
- EXTRA SLIDES
- EXTRA SLIDES
- EXTRA SLIDES
- EXTRA SLIDES
- EXTRA SLIDES
- EXTRA SLIDES
- EXTRA SLIDES

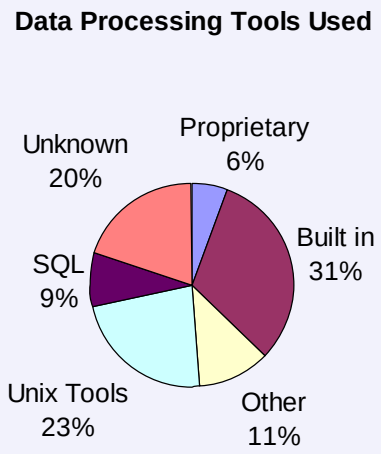
More Statistics

- Total hours spent by organizers: 800 person hours
- Ronny's e-mail for KDDCup (1060 e-mails)
- Max CPU time to generate model: 1000 hours

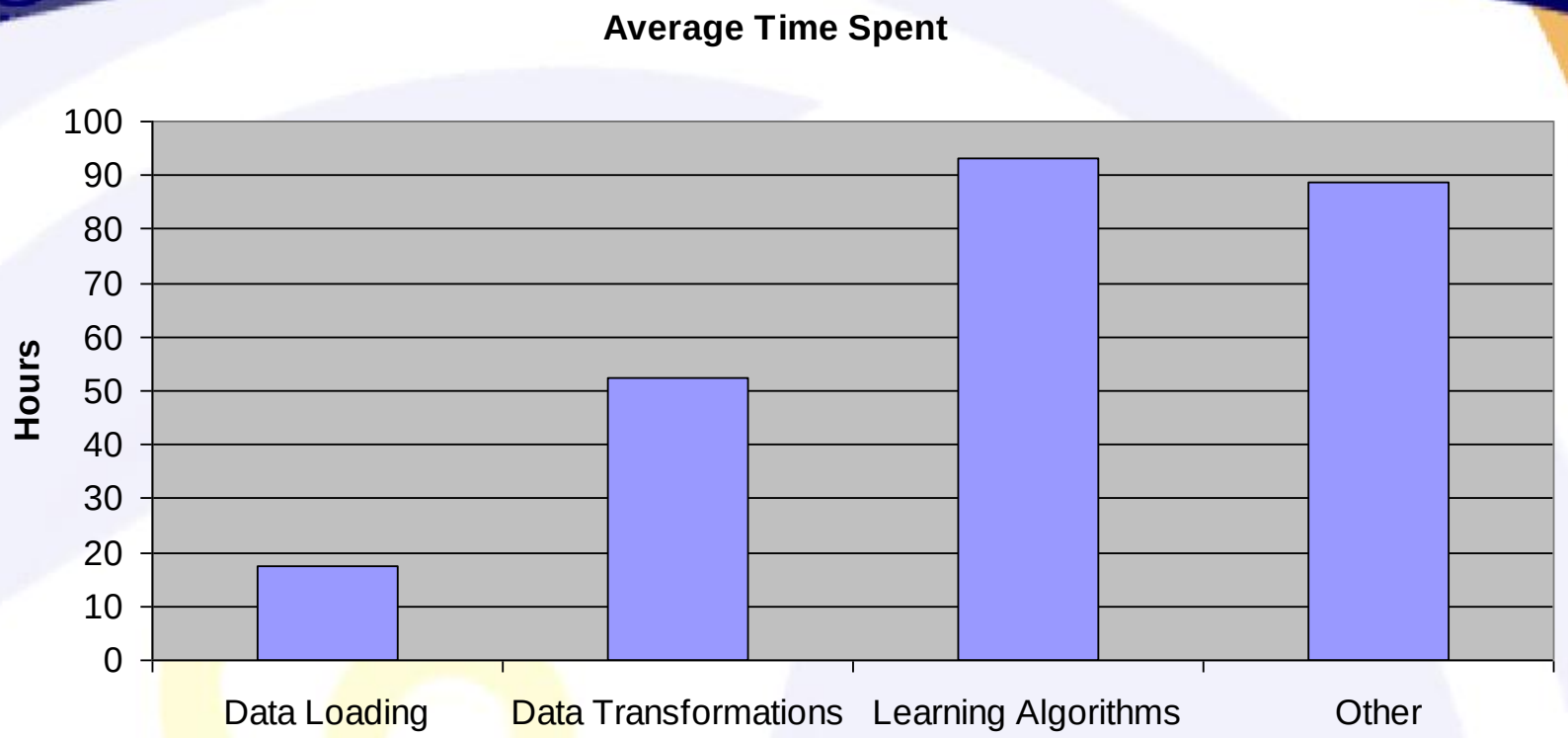
Statistics II







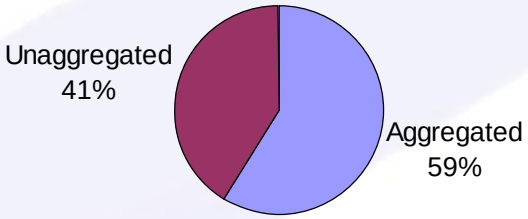
Statistics III



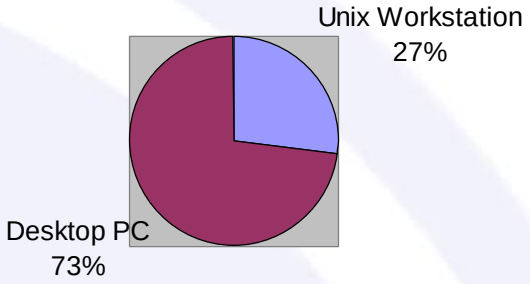
- 32% used database, 68% flat files
- 41% used unaggregated data, 59% used the aggregated
- Operating systems: Windows (54%), Unix (30%), Linux (16%)

Statistics IV

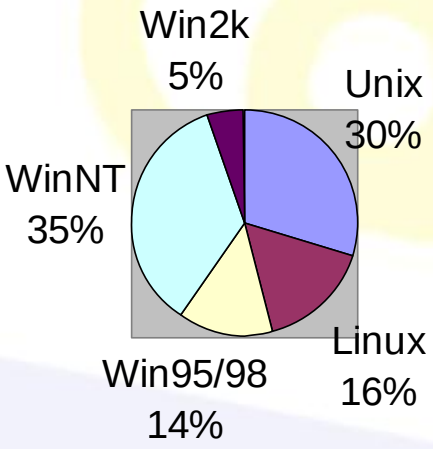
Aggregated vs Unaggregated Data



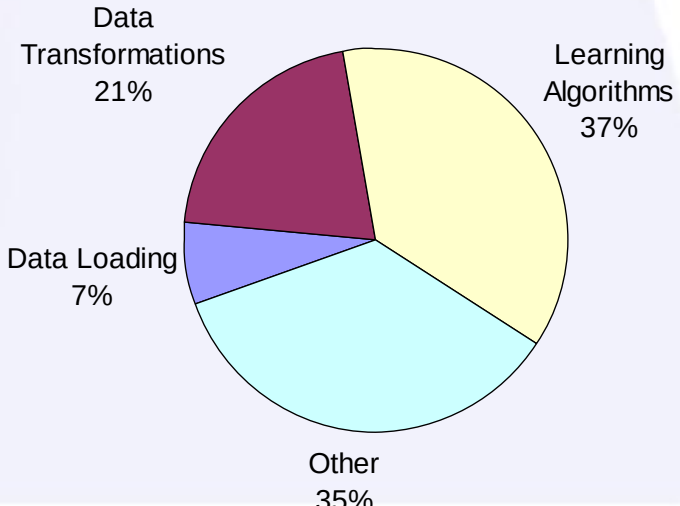
Hardware Used



Operating Systems Used



Average Time Spent Relative



More Insight

- Coupon users ($10 off) were buying less even ignoring the discount!

Clipping

Given a set of page views, will the visitor view another page on the site or will the visitor leave?

- To simulate a user who is in mid session (continuing), we clipped the test set sessions
- In the training set, we marked clipping points but released the whole dataset
- Since the data contains multiple records per session and most packages can't handle that, we provided an aggregated version with one record per session (59% of the participants used the aggregated version)