L4: Outline: CoAd Lectures

- **Introduction**

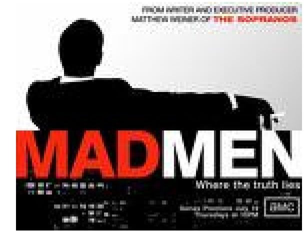
L1
- **Online advertising background**
- **Business models, Campaigns**

L2
- **Technology and Economics**
  - Forward Markets
    - Gradient Descent, Operations research, LP, QP

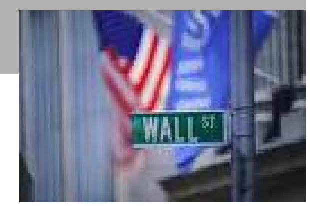
L3
  - Auction Theory and Game Theory
  - Spot Markets
    - ML, Ad quality, Ranking, Budgeting

L4
- **New Directions**
- **Challenges in online advertising**
- **Summary**

Business, Gold rush



Tech



Hot Areas

CoAd Lectures		
Friday	9/11/2009	10:30-12:00
Saturday	9/12/2009	8:30-10:00
Sunday	9/13/2009	8:30-10:00
Monday	9/14/2009	8:30-10:00

Course philosophy

- **Socratic Method (more inspiration than information)**
  - participation strongly encouraged (please state your name and affiliation)
- **Highly interactive and adaptable**
  - Questions welcome!!
- **Lectures emphasize intuition, less rigor and detail**
  - Build on lectures from other faculty
  - Background reading will provide more rigor & detail
- **Action Items**
  - Read suggested books first (and then papers), read/**write** Wikipedia, watch/**make** YouTube videos, take courses, participate in competitions, do internships, network
  - Prototype, simulate , publish, participate
  - Classic (core) versus trendy (applications)

Lecture 2: Homework

- **Email solutions to James.Shanahan__AT__gmail.com**
- **Exercises**
  - Find a local minimum of the function $f(x)=6x^5-8x^2+6$
  - Implement gradient descent version of Perceptron
  - Implement gradient descent version of OLS; show evolution of weight vector during training

Lecture 3: Homework

- **Email solutions to James.Shanahan__AT__gmail.com**
- **Exercises**
  - What is the mixed strategy equilibrium for Rocks Scissors Paper game?
  - What are the three equilibria for the following game?

Player 2			
Player 1		Up	Down
	Up	(2, 2)	(0, 0)
	Down	(0, 0)	(1, 1)

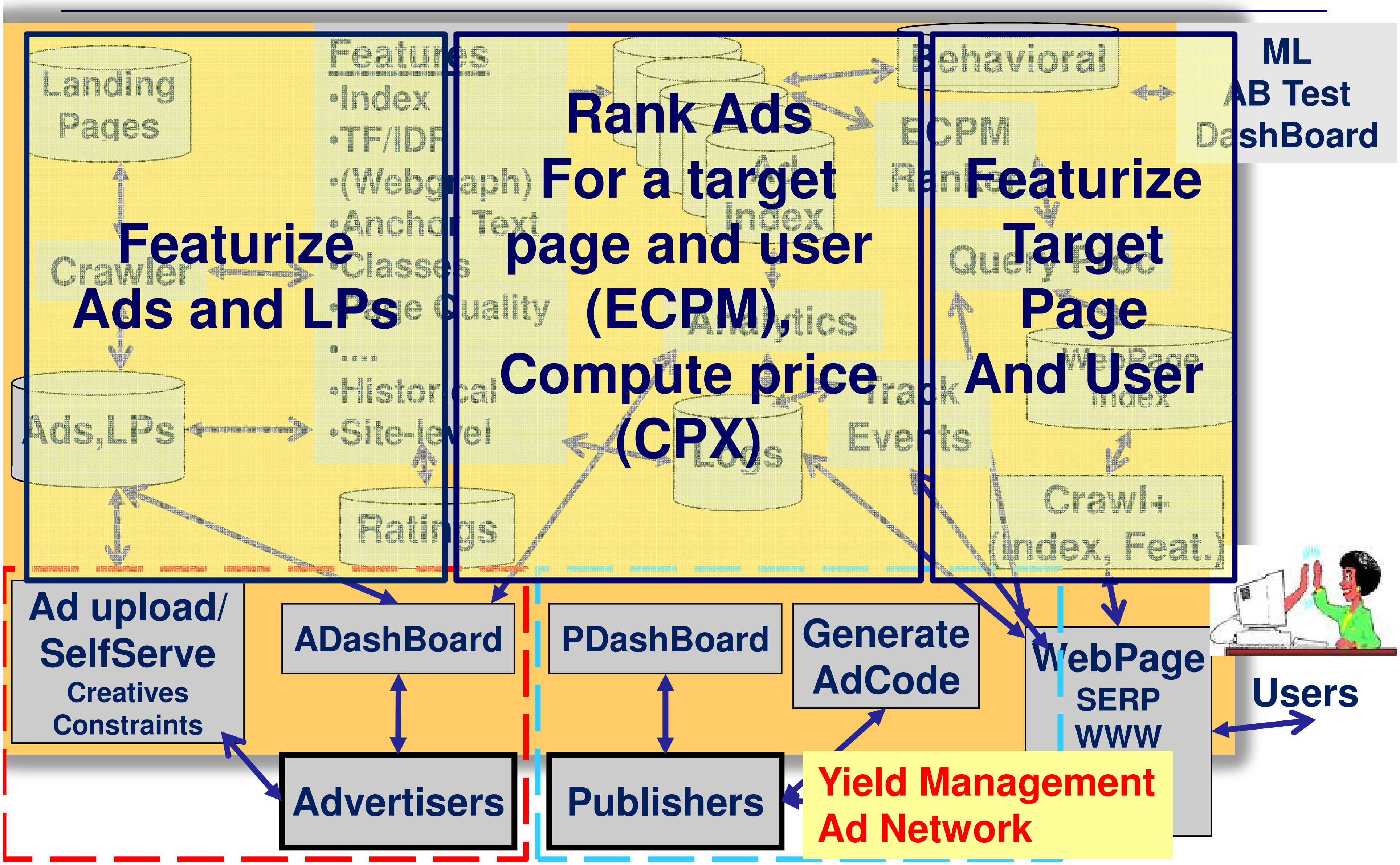
Outline

- **Introduction**
- **Online advertising background**
- **Business models**
- **Creating an online ad campaign**
- **Technology and Economics**
  - Advertisers (optimizing ROI thru ads and ad placement)
  - Publishers (optimizing revenue and consumer satisfaction)
    - Forward Markets
    - Auction Systems, Game Theory
    - Spot Markets
      - Background
      - Ad Quality
      - Budgeting
- **New Directions**
- **Challenges in online advertising**
- **Summary**

Ad Quality Outline

- Motivation
- Machine Learning Overview
- Metrics
- IR as a means of Ranking
- Learning to Rank (LETOR)
- Online Learning
- Open issues

Ad Network Architecture: Spot Market



Rank by ECPM; CPC Calculation

Payoff = Value – Price
Payoff = ValuePerClick – CPC
ECPM = Bid * QualityScore

$$ECPM_{Ad} = Quality_{Ad} * Bid_{Ad} * 1000$$

$$Bid_1 \times DQ_1 > Bid_2 \times DQ_2$$

For ad_1 to maintain it's current rank then Bid_1 needs to be at least:

$$Bid_1 \geq \frac{Bid_2 \times DQ_2}{DQ_1}$$

CPC for click at rank1

Ad Id	1. Receive Bid	2. Assess Quality	3. Calculate Rank	4. Set CPC Price
123	$5.80	10	$58.00	$1.71
ABC	$4.25	4	$17.00	$3.01
NOP	$2.00	6	$12.00	$0.51
TUV	$3.00	1	$3.00	$1.66
XYZ	$0.55	3	$1.65	Reserve Bid

Online Ad Targeting Systems (OATS)

- **OAT systems has many similarities with "Direct Mail" systems**
  - Can borrow a lot from these experiences and ideas
  - Many studies, e.g.,
    - Network-based Marketing: Identifying likely adopters via consumer networks. S. Hill, F. Provost, and C. Volinsky. ***Statistical Science*** 21 (2) 256–276, 2006.
    - Data Mining for CRM
      - See http://www.dmreview.com/article_sub.cfm?articleId=1046025
    - Daniel T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, ISBN: 0471666572, John Wiley, 2004 (see also companion site for Larose book).
- **Recently many OATS…..**

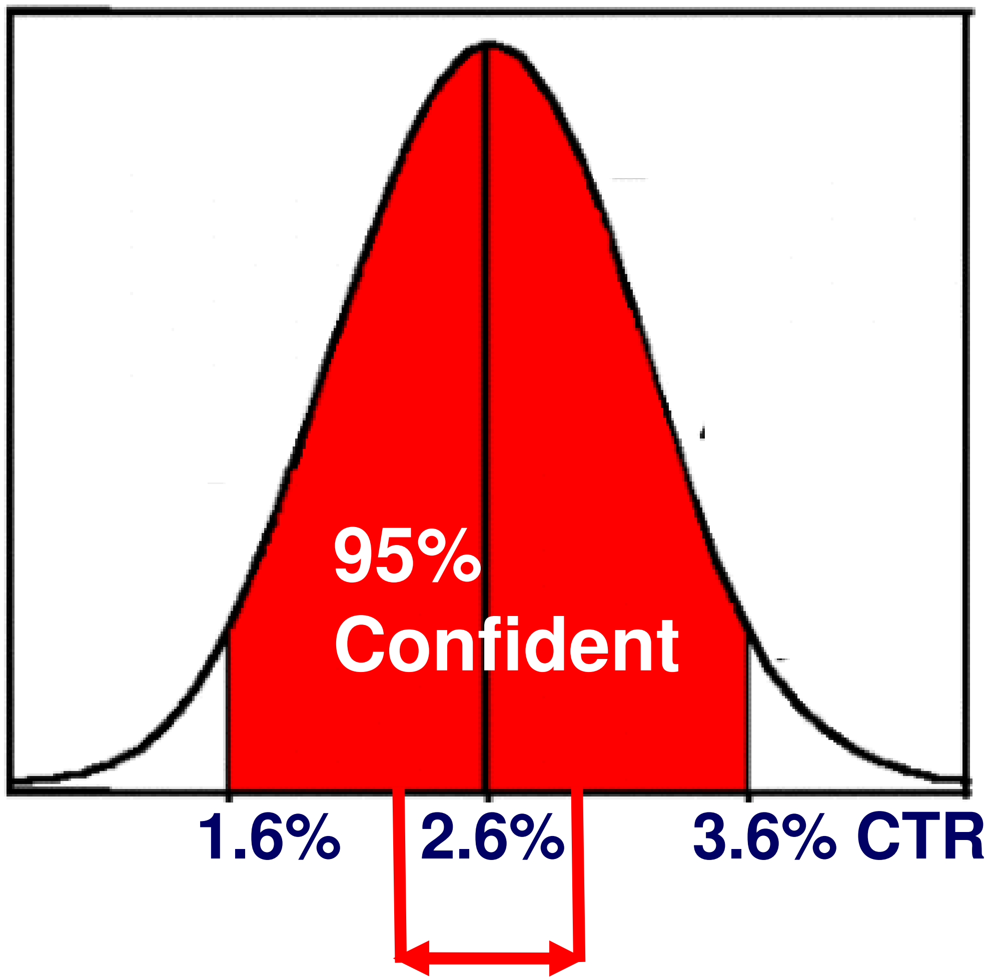
Accurate CTR Estimates are Crucial

$$ECPM_{Ad} = \boxed{CTR_{Ad}} * Bid_{Ad} * 1000$$

- **Very important to have accurate estimates of CTR_{Ad} for a keyword or publisher page**
  - for ranking and for revenue purposes
- **E.g., A true CTR for an Ad is 2.6% must be shown 1,000 times before we are 95% confident that this estimate is within 1% of the true CTR**

- **Curiously, average CTR and CPC**
  - 2.6% CTR for ads in sponsored search advertising
  - Average CPC (cost-per-click) on Google was $1.60
  - [MarketingSherpa, 9/2005]

Estimating CTR (and later AR)



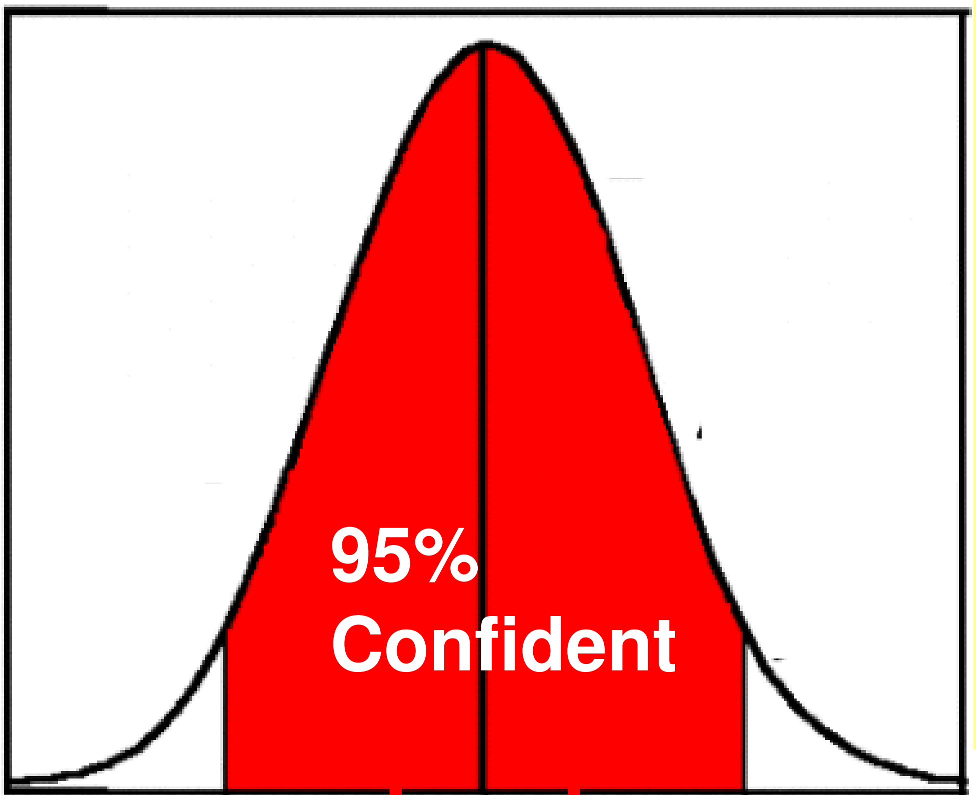
Estimate using Binomial
MLE Estimates
I.e., #Clicks/#Impression

$40/1,000 @CPC of $1.60
$400/10,000

CTR (after 1,000 impressions)

(after 10,000 impressions)

Estimating CTR (and later AR)



1.6% 2.6% 3.6% CTR (after 1,000 impressions)

2.3% 2.9% (after 10,000 impressions)

For a network of
~10^9 target pages,
~10^6 ads
~10^7 users
……
- Cannot afford this evaluation/auditioning
- Borrow strength, marginalize
- CoD (curse of $\text{dim}^{\text{ality}}$)

Accurate CTR Estimates are Crucial

$$ECPM_{Ad} = \boxed{CTR_{Ad}} * Bid_{Ad} * 1000$$

- **Very important to have accurate estimates of CTR_{Ad} for a keyword or publisher page**
  - for ranking and for revenue purposes
  - CTR drop exponentially with position [enquiro.com] ; NDCG Metric
- **E.g., A true CTR for an Ad is 2.6% must be shown 1,000 times before we are 95% confident that this estimate is within 1% of the true CTR, i.e., [1.6, 3.6]**
  - Very noisy!!

Ad Quality Outline

- **Motivation**
- **Machine Learning Overview**
  - Background
  - Gradient Descent
  - From Perceptrons to SVMs
  - Probabilistic Models
  - Decision Trees
  - Bagging/Boosting
- **Metrics**
- **IR as a means of Ranking**
- **Learning to Rank (LETOR)**
- **Online Learning**
- **Open issues**

Machine Learning is about…

- **Models**
- **Loss functions**
  - Maximum likelihood, mean squared errors, accuracy
- **Optimization**
  - gradient descent, Newton, linear/quadratic programming, etc.
- **Data**
  - Training data, Validation data, Test data, crossfold validation
- **Metrics**
  - Precision, recall, DCG, Revenue etc.
- **And more!**

Machine Learning Background

Machine Learning (ML):"*a computer program that improves its performance at some task through experience*" [Mitchell 1997]

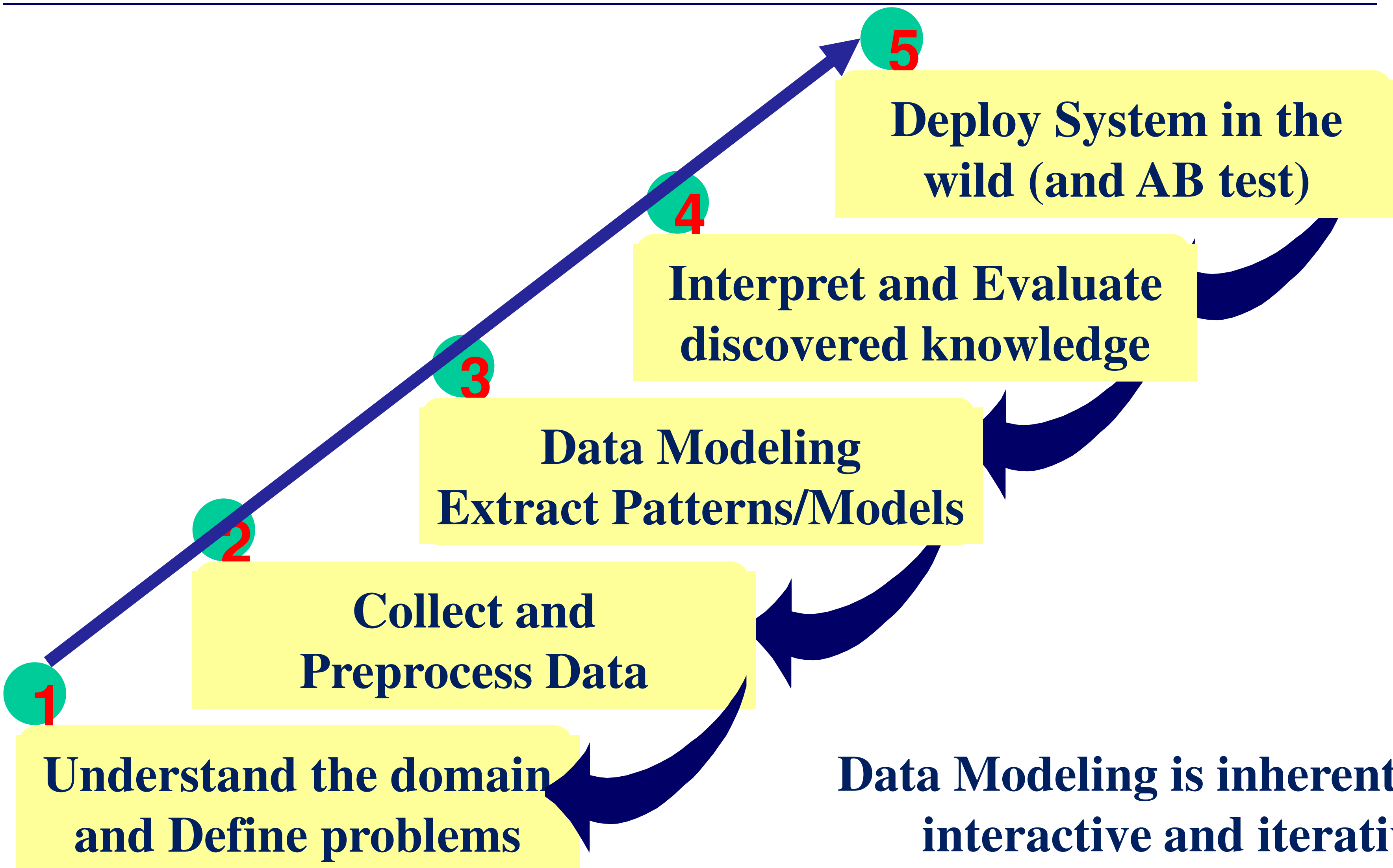
GIVEN: Input data is a table of attribute values and associated class values (in the case of supervised learning)
GOAL: Approximate $f(x_1,\ldots,x_n) \rightarrow y$

Instance\Attr	x_1	x_2	…	x_n	y
1	3	0	..	7	-1
2	[illegible]	[illegible]	[illegible]	[illegible]	+1
…	…	…	…	…	…
L	0	4	…	8	-1

What is Machine Learning?

- **Automatically programming computers to perform tasks that humans perform well but find difficult to specify algorithmically**
- **Principled way of building high performance information processing systems**
  - **search engines, information retrieval**
  - **Natural language processing**
  - **Finance**
  - **Engineering**

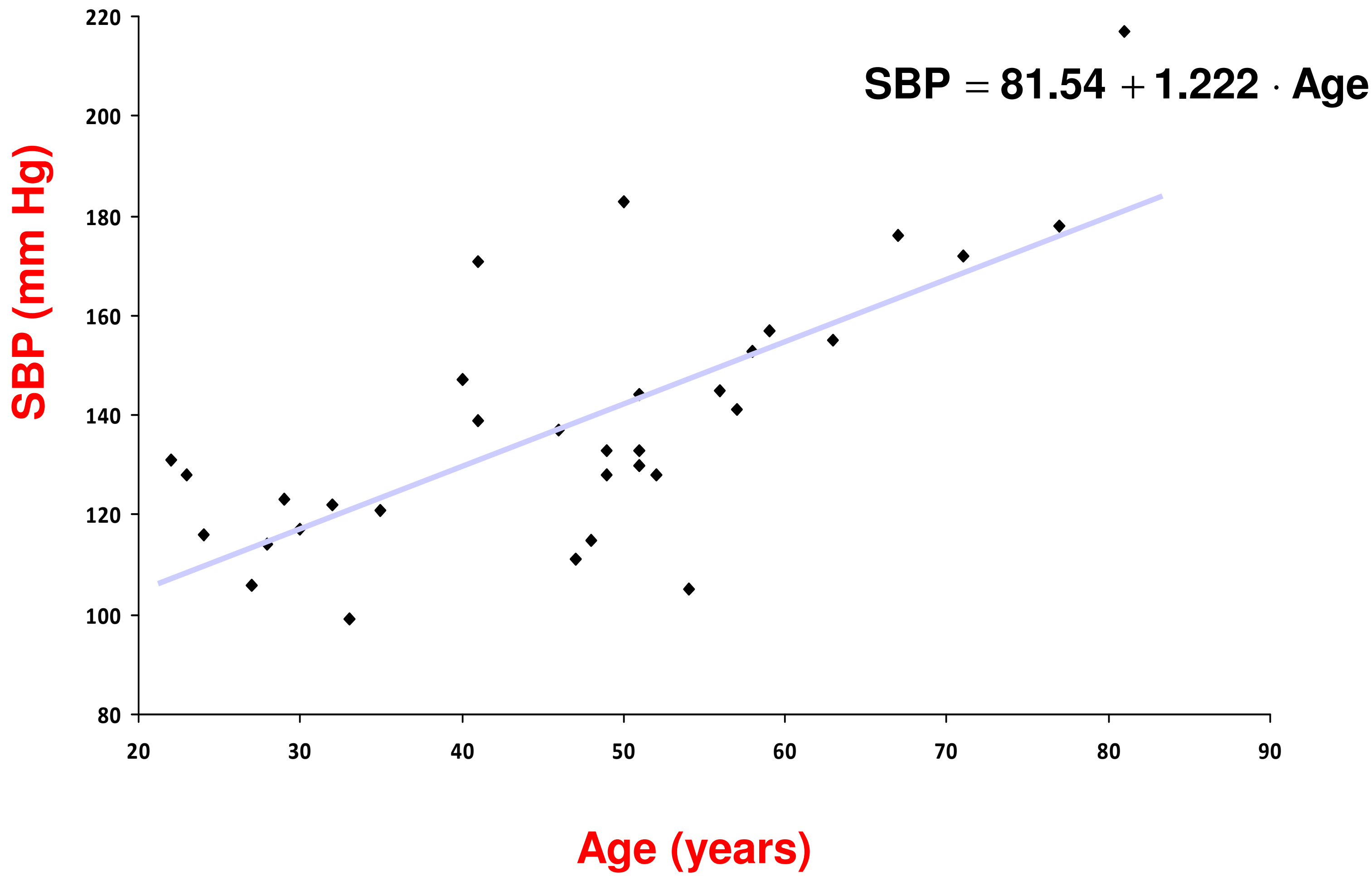
Data Modeling Process



Data Modeling is inherently interactive and iterative

Linear Regression

Predict systolic blood pressure (SBP) from Age using 33 adult women examples



adapted from Colton T. Statistics in Medicine. Boston: Little Brown, 1974

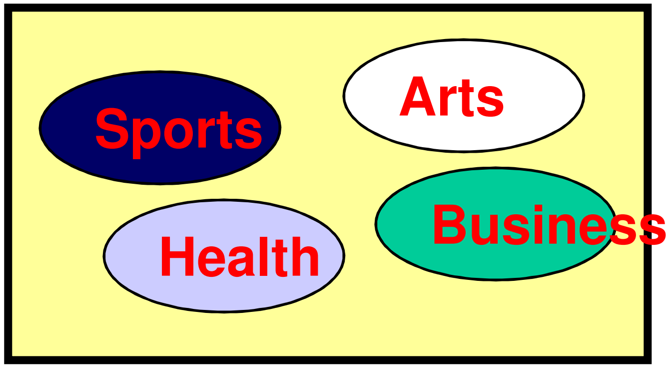
Types Machine Learning

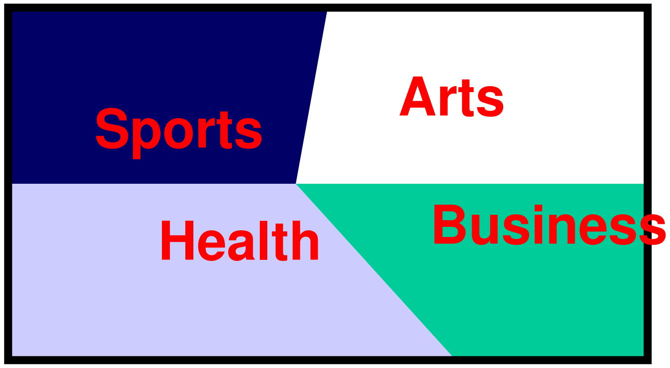
- **Design of a learning algorithm is affected by**
  - Task to be learned?
  - What feedback is available to from?
  - What representation is used for the learnt model?

- **Type of feedback:**
  - Supervised learning: correct answers for each example
  - Unsupervised learning: correct answers not given
  - Reinforcement learning: occasional rewards
  - Co-learning, Transductive Learning, etc.

Families of Supervised Learning

- **Generative Classifier (Bottom-up learning)**
  - Build model of each class
  - Assume the underlying form of the classes and estimate their parameters (e.g., a Gaussian)
- **Discriminative Classifier (Top down)**
  - Build model of boundary between classes
  - Assume the underlying form of the discriminant and estimate its parameters (e.g., a hyperplane)





Generative vs. Discriminative

- **Discriminative learning (e.g., ANN, SVM, logistic regression, Conditional Random Fields (CRF)) typically more accurate**
  - Better with small datasets
  - Faster to train
- **Generative learning (e.g., Bayesian Networks, HMM, Naïve Bayes, Mixtures of Gaussians) typically more flexible**
  - More complex problems
  - More flexible predictions

Ad Quality Outline

- **Motivation**
- **Machine Learning Overview**
  - Background
  - Gradient Descent (covered in Lecture 2)
  - From Perceptrons to SVMs (covered partially in lecture and in notes)
  - Probabilistic Models (in notes and see other lectures)
  - Decision Trees
  - Bagging/Boosting
- **IR as a means of Ranking**
- **Learning to Rank (LETOR)**
- **Online Learning**
- **Open issues**

Ad Quality Outline

- **Motivation**
- **Machine Learning Overview**
  - Background
  - Gradient Descent
  - From Perceptrons to SVMs
  - Probabilistic Models (in notes and see other lectures)
  - Decision Trees
  - Bagging/Boosting
- **IR as a means of Ranking**
- **Learning to Rank (LETOR)**
- **Online Learning**
- **Open issues**

Maximum Margin Classifier: take 3

- **The classifier that produces the maximum margin (over the training data)**
- **I.e., the hyperplane that is furthest from the data**

Better!
But the objective is non-convex

$$\underset{\gamma, W, b}{\text{Maximize}} \frac{\gamma_{fun}}{\|W\|}$$

$$\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq \gamma_{fun} \ \forall \ i = 1, \ldots, L$$

Versus

Best!
Convex objective

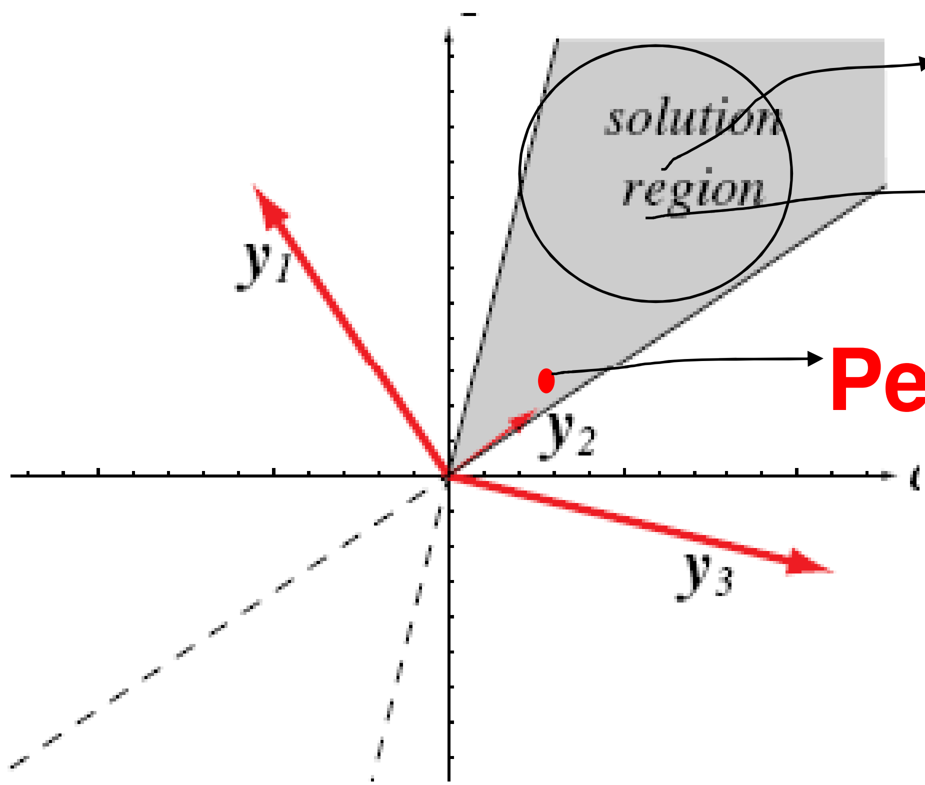
$$\underset{\gamma, W, b}{\text{Maximize}} \frac{1}{\|W\|^2}$$

$$\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \ \forall \ i = 1, \ldots, L$$

$$\gamma_{Fun} = 1$$

$$\text{Scaling constraint}: \gamma_{Fun} \text{ of } (\mathrm{W}, \mathrm{b}) \text{ WRT } S \text{ is set to } 1$$

Learning Algorithms in Version Space



SVM

Bayes Point Machine

Perceptron

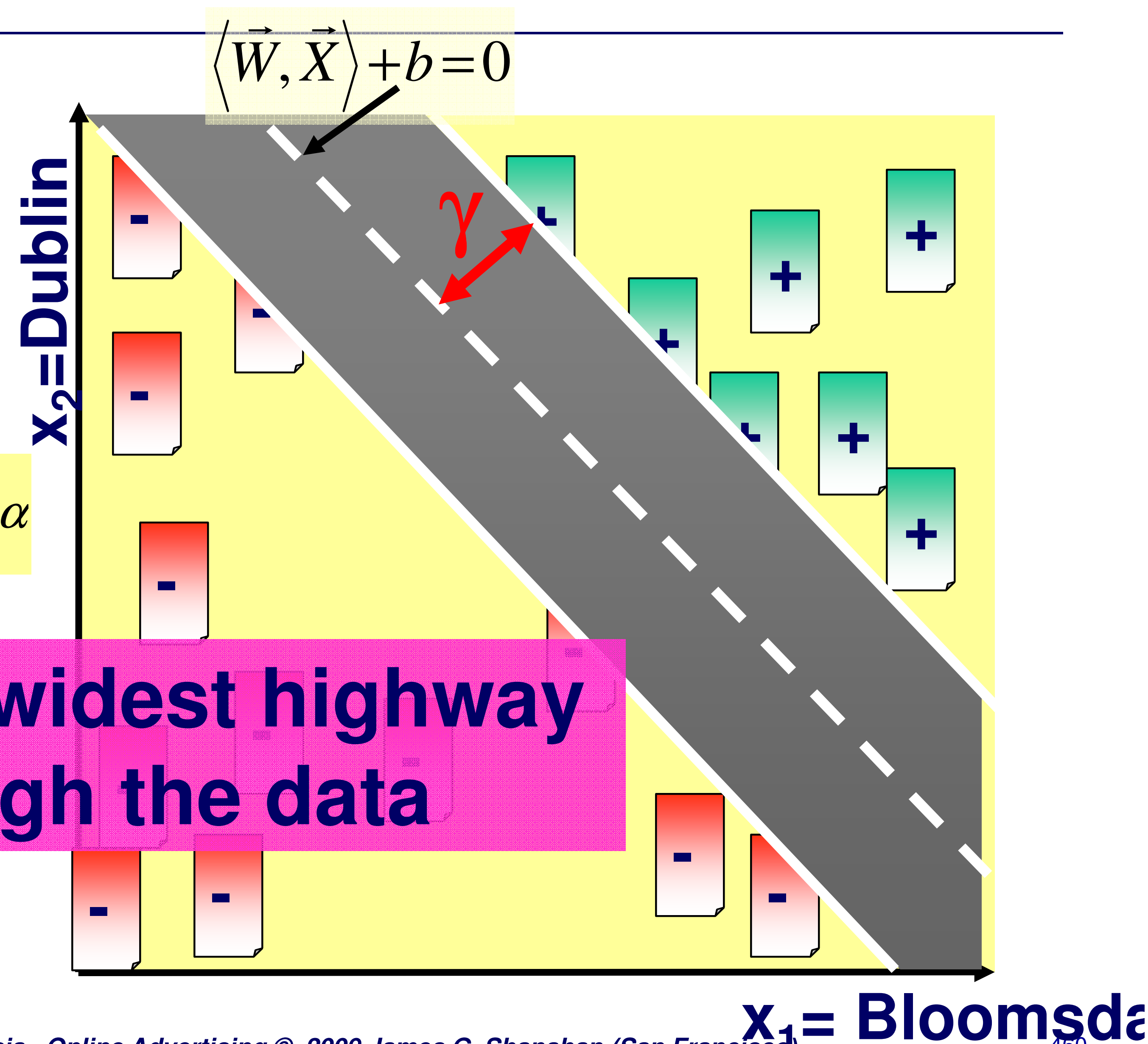
SVMs find the center of the largest radius hypersphere whose center can be placed in version space and whose surface does not intersect with the hyperplanes corresponding to the labeled instances.

SVM

$$\text{Min}\ \frac{\|\mathbf{w}\|^2}{2}$$

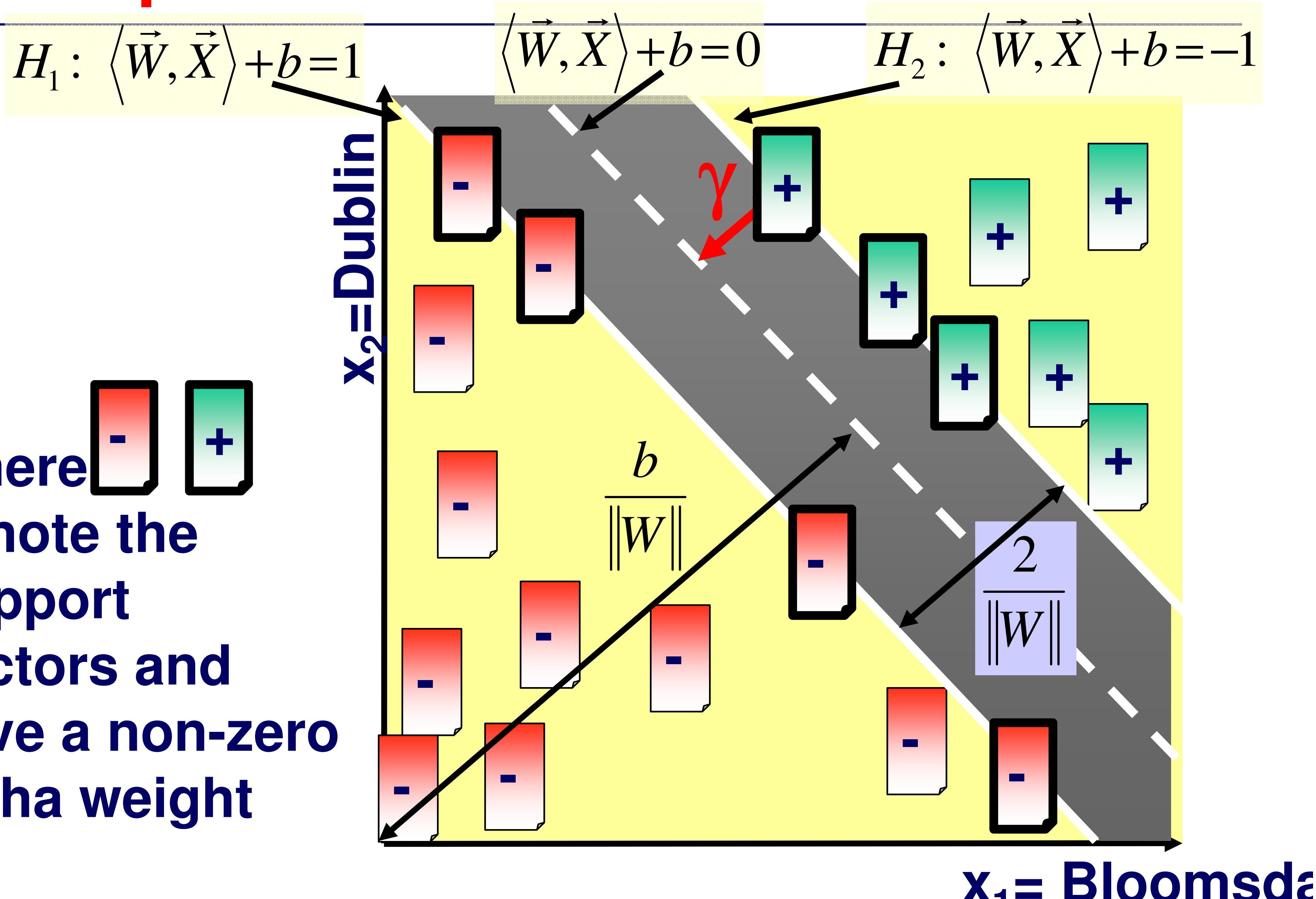
$$Max\ \frac{2}{\|\mathbf{w}\|^2}$$

$$\max_{\alpha} L(\alpha) = \max_{\alpha} I^T \alpha - \frac{1}{2} \alpha^T Q \alpha$$



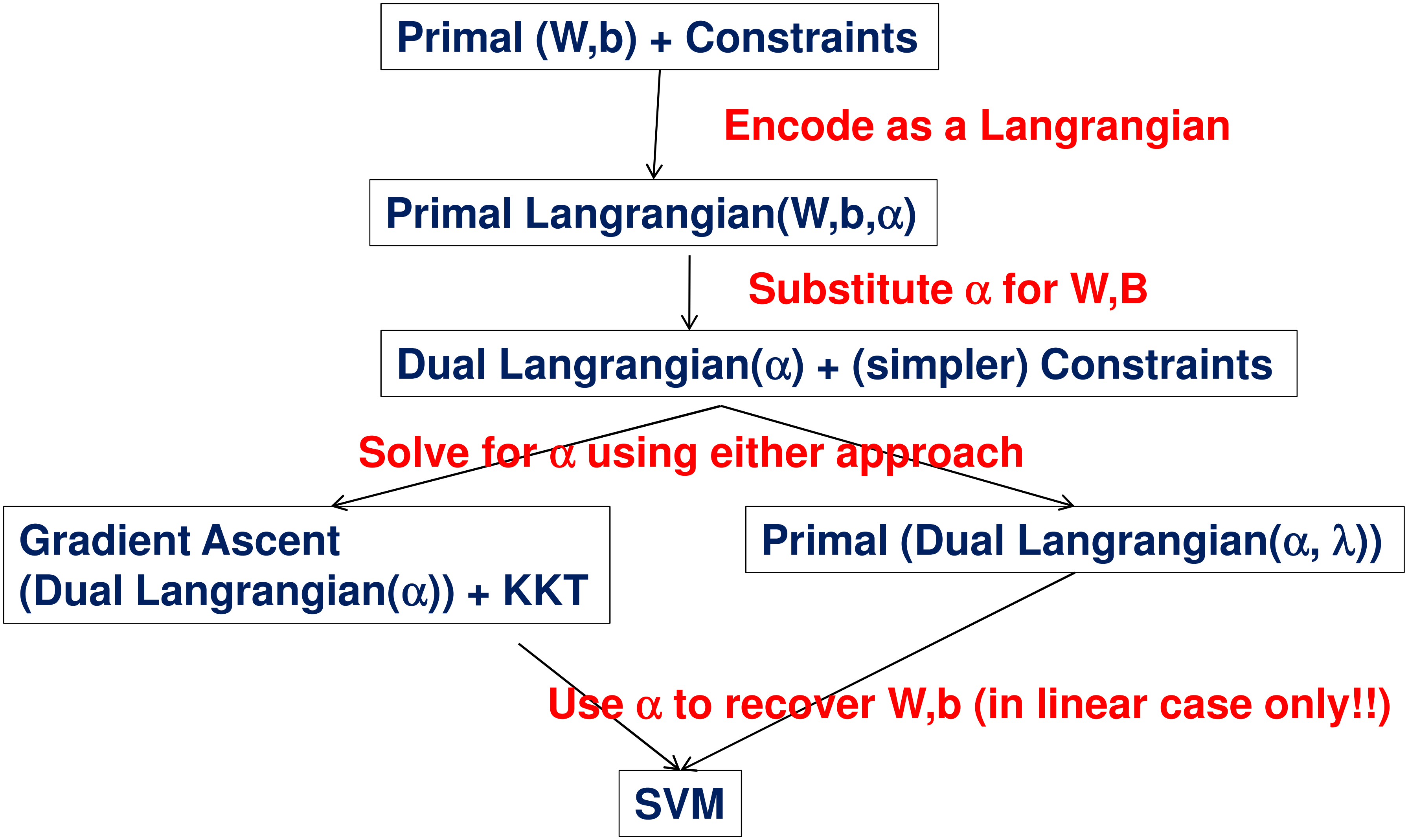
Find widest highway through the data

Topic = Product Watch Filter for K2



Where [-] [+] denote the Support Vectors and have a non-zero alpha weight

SVM Learning Algorithms



SVM Learning Algorithms

- **Gradient Descent**
  - Perceptron Ensemble
  - Perceptron
  - Kernel-Adatron
  - PAUM
- **Naïve Gradient Descent in dual mode**
- **Population-based search**
  - E.g., Evolutionary algorithms
- **Quadratic/linear programming based (QP/LP) Algorithms**
  - Interior point algs; Fixed or variable sized Chunking
- **Non-QP-based, with closed form**
  - SMO, SMOK1, SMOK2

Gradient Ascent Method

$$\max_{\alpha} W(\alpha) = \max_{\alpha} \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{j=1}^{\ell} y_i y_j \langle X_i, X_j \rangle \alpha_i \alpha_j,$$

Given Training set S and learning rate, and α=0

Repeat

for all train set *i = 1 to l*

update

$$\alpha_i \leftarrow \alpha_i + \frac{\omega}{K(x_i, x_i)} \left(1 - y_i \sum_{j=1}^{l} \alpha_j y_j K(x_i, x_j) \right)$$

$$where \quad \omega \in (0,2)$$

Ensure α are within range [0, C]

$$if \quad \alpha < 0, \quad then \quad \alpha \leftarrow 0$$

$$if \quad \alpha > C, \quad then \quad \alpha \leftarrow C$$

End for

Until stop criterion satisfied (KKT conditions)

Return α

Learning Algorithms

SVM Learning Strategy

- **Non-QP-based are more desirable**
  - More efficient in terms of CPU and memory
  - Easier to program
  - Less intellectual property
- **Implemented non-QP-based algorithms**
  - Kernel-Adatron
  - SMO (patented by Microsoft)
  - Two modifications SMOK1, SMOK2 (IP free)

SVMs Basic Ideas: Learning

SVM Parameters

Training Set

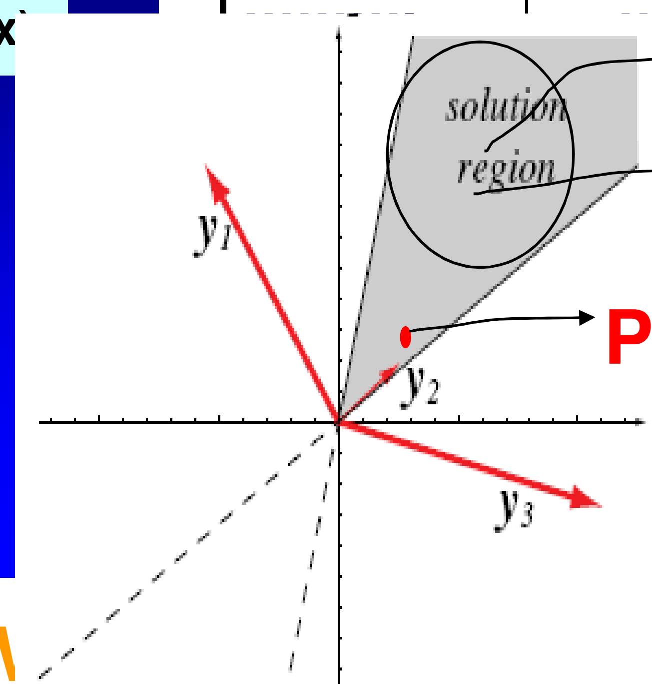
Kernel Computation (Doc Similarity matrix)

Find Hyperplane in dual space

Convert Hyperplane to primal space

Hyperplane (W, b)

SVM Filter	
Term	Weight
T_1	Wgt_1
T_2	Wgt_2
	gt_n
	b



SVM

Bayes Pc

Mac

Perceptron

> 0

W

o

o

$f(X) < 0$

x_1

Conclusion

- **Support vector machines (SVM) are a general purpose suite of machine learning algorithms for classification and regression. Generic support vector machines (SVMs) provide excellent performance on a variety of learning problems ranging from text to genomic data.**
- **There has been an explosion of new learning algorithms, kernels and applications**
- **Tuning SVMs remains a black art: selecting a specific kernel and parameters is usually done in a try-and-see manner.**
- **Many public domain packages available on the WWW**
- **Building kernels is a very hot of research area**

- **Anyone involved in text processing should consider using support vector machines but *astutely***

R code

- library (e1071)
- iris .svm <- svm(Species ~ ., data =iris , probability = TRUE)
- plot (iris .svm , iris , Petal . Width ~ Petal .Length,
- slice = list (Sepal . Width = 3, Sepal . Length = 4))
- pred <- predict (iris .svm , iris , probability = TRUE)
- attr (pred , " prob ") # to get the probabilities

Software

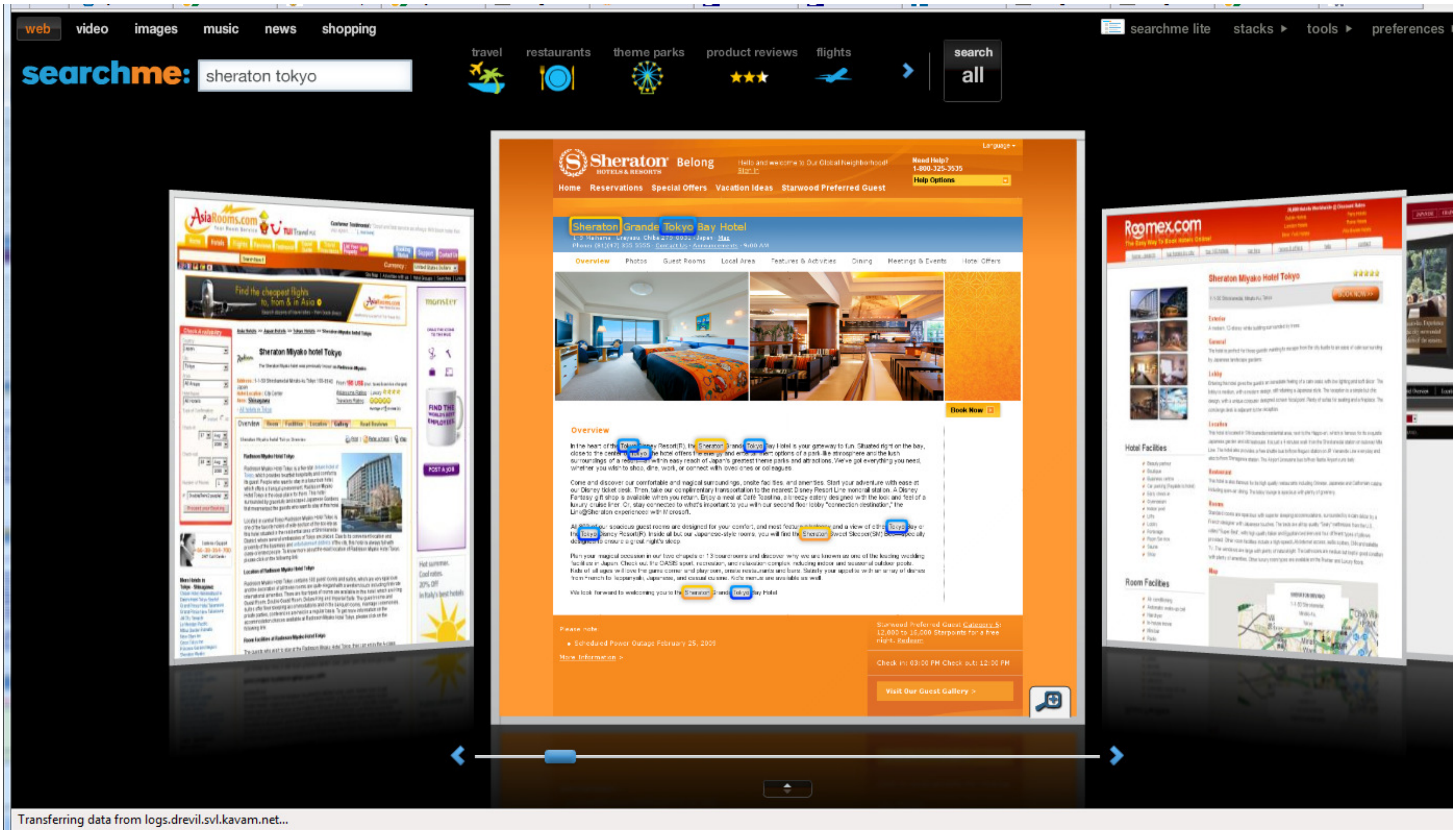
- A list of SVM implementations can be found at http://www.kernel-machines.org/software.html
- LIBSVM (Java, and C++) can handle multi-class classification
  - http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- Weka is open source machine learning toolkit (Java-based) available from
  - http://www.cs.waikato.ac.nz/~ml/weka/
- SVMLight is among one of the earliest implementations of SVMs (written in C)
- Several high level toolboxes for SVMs are also available
  - Matlab, R

Background Material

- **Background information about SVMs can be found in:**
  - SVMs in R
    - http://www.jstatsoft.org/v15/i09/paper
  - Burges' tutorial (http://svm.research.bell-labs.com/SVMdoc.html)
  - Cristianini, N., Shawe-Taylor, J., An Introduction to Support Vector Machines, Cambridge University Press, (2000).
  - Schölkopf, S., Burges, C. J. C., Smola, A. J., Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, (1999).
  - Vapnik, V. *Statistical Learning Theory.* Wiley-Interscience, New York, (1998).
  - Publication list at kernel-machines.org
  - References in An Introduction to Support Vector Machines

Potential Applications of SVMs

- **Page Quality, Page Category, Webspam, LETOR**



James.Shanahan_AT_gmail_DOT_com

Ad Quality Outline

- **Motivation**
- **Machine Learning Overview**
  - Background
  - Gradient Descent
  - From Perceptrons to SVMs
  - Probabilistic Models
  - Decision Trees
  - Bagging/Boosting
- **IR as a means of Ranking**
- **Learning to Rank (LETOR)**
- **Online Learning**
- **Open issues**

Fruit Classifier: CART DT

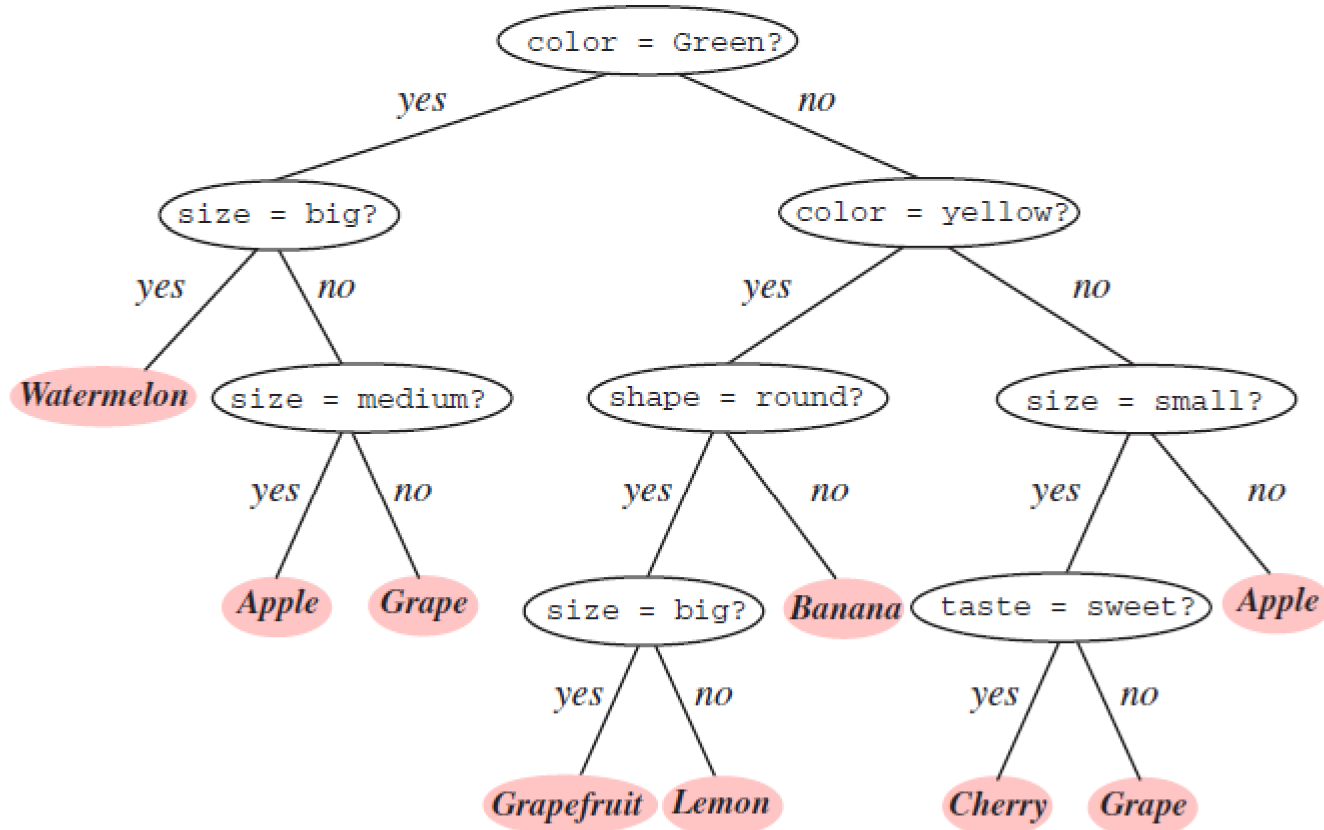


FIGURE 8.2. A tree with arbitrary branching factor at different nodes can always be represented by a functionally equivalent binary tree—that is, one having branching factor $B = 2$ throughout, as shown here. By convention the "yes" branch is on the left, the "no" branch on the right. This binary tree contains the same information and implements the same classification as that in Fig. 8.1. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

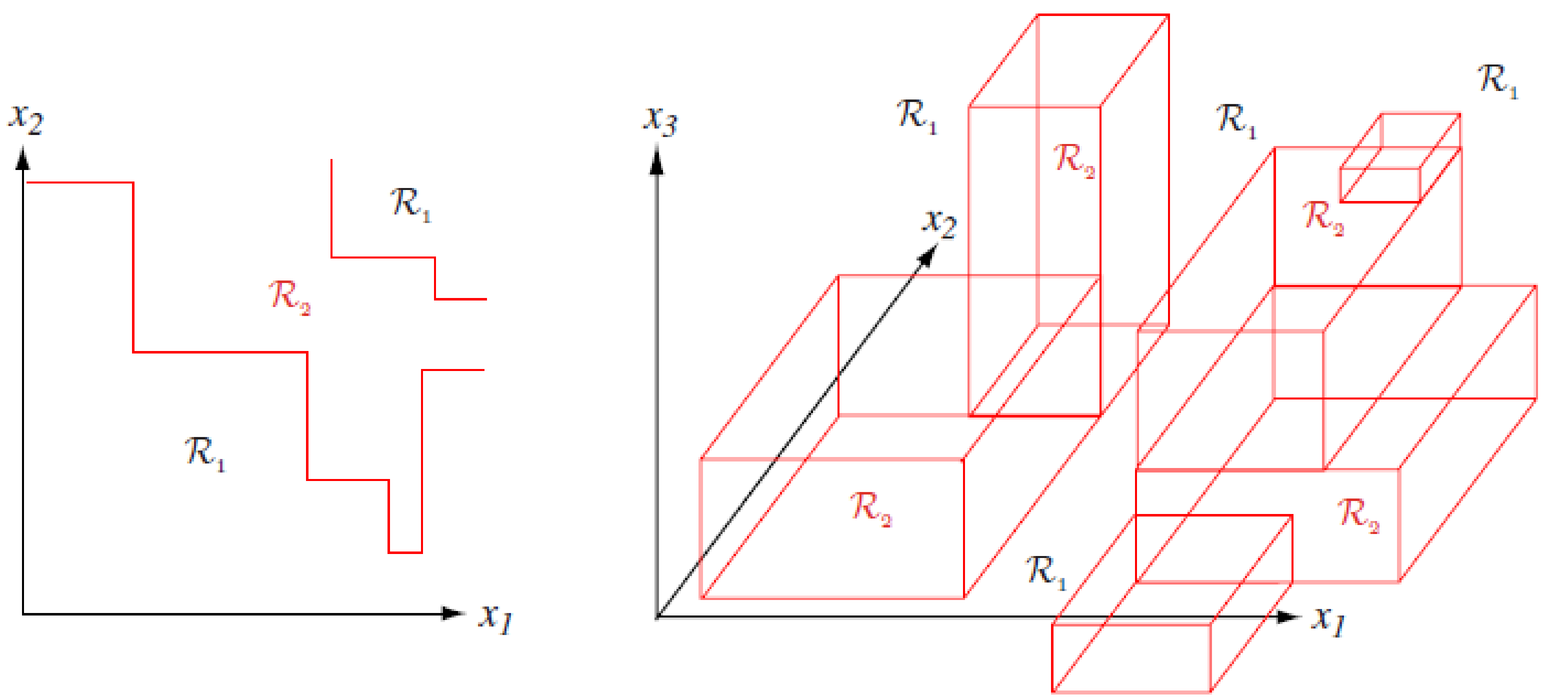


FIGURE 8.3. Monothetic decision trees create decision boundaries with portions perpendicular to the feature axes. The decision regions are marked $\mathcal{R}_1$ and $\mathcal{R}_2$ in these two-dimensional and three-dimensional two-category examples. With a sufficiently large tree, any decision boundary can be approximated arbitrarily well in this way. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Variable split? And node prediction?

- **For Classification**
  - Select variable split using Impurity measures and reduction impurity
    - Entropy, GINI, Misclassifcation Error
  - Node prediction: dominant class, or conditional probability of dominant class
- **For Regression**
  - Select variable split using squared prediction error
  - Node prediction: mean or median value of the response values of the training examples that make it to that node

Using information theory

- To implement `Choose-Attribute` in the DTL algorithm
- Information Content (Entropy):

  $$I(P(v_1), \ldots , P(v_n)) = \Sigma_{i=1} -P(v_i) \log_2 P(v_i)$$

- For a training set containing *p* positive examples and *n* negative examples:

$$I(\frac{p}{p+n}, \frac{n}{p+n}) = -\frac{p}{p+n}\log_2 \frac{p}{p+n} - \frac{n}{p+n}\log_2 \frac{n}{p+n}$$

[Adapted from Russell and Norvig 1995]

Information gain

- **A chosen attribute *A* divides the training set *E* into subsets $E_1, \ldots, E_v$ according to their values for *A*, where *A* has v distinct values.**

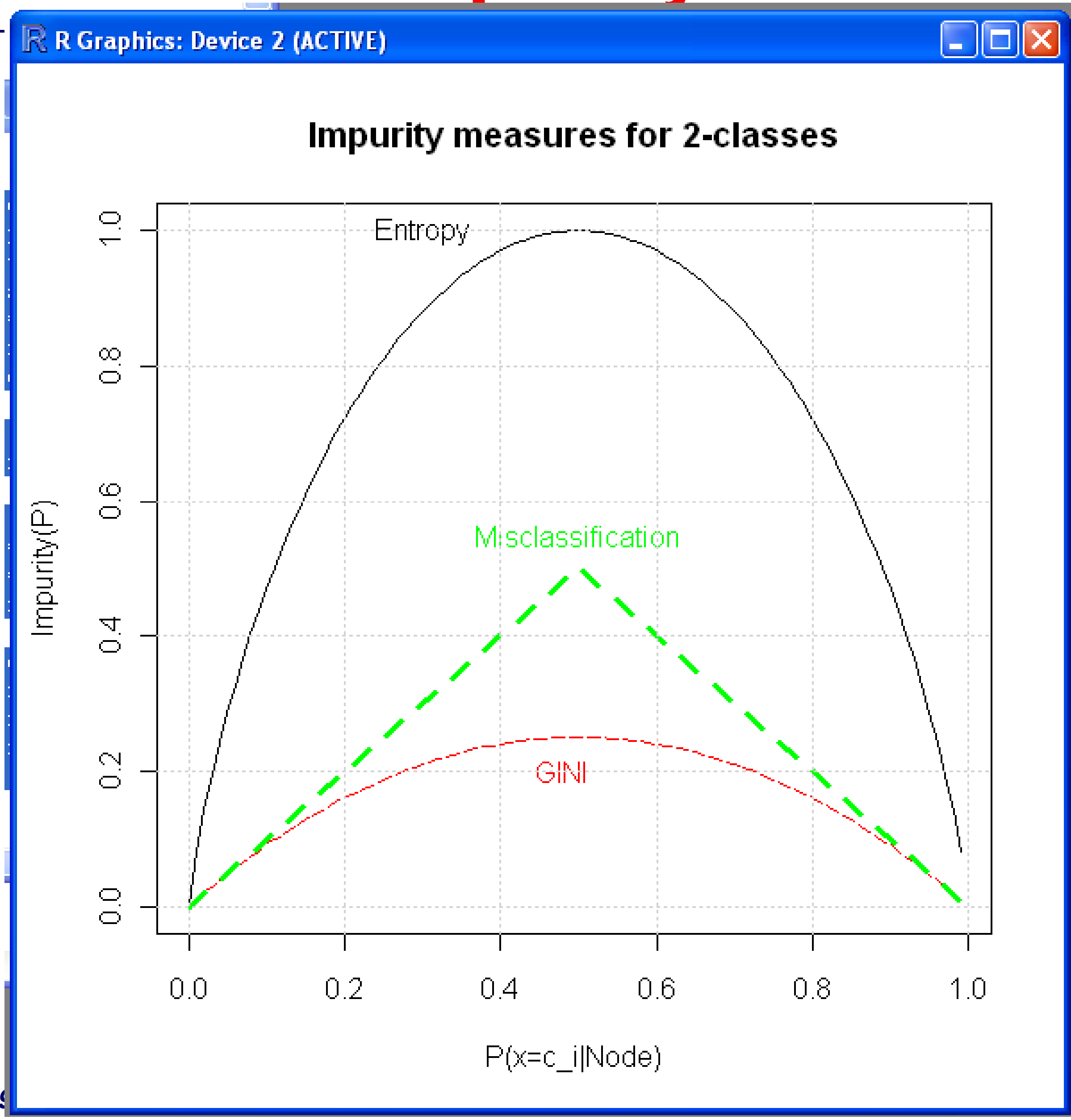
$$remainder(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I\left(\frac{p_i}{p_i+n_i}, \frac{n_i}{p_i+n_i}\right)$$

- **Information Gain (IG) or reduction in entropy from the attribute test:**

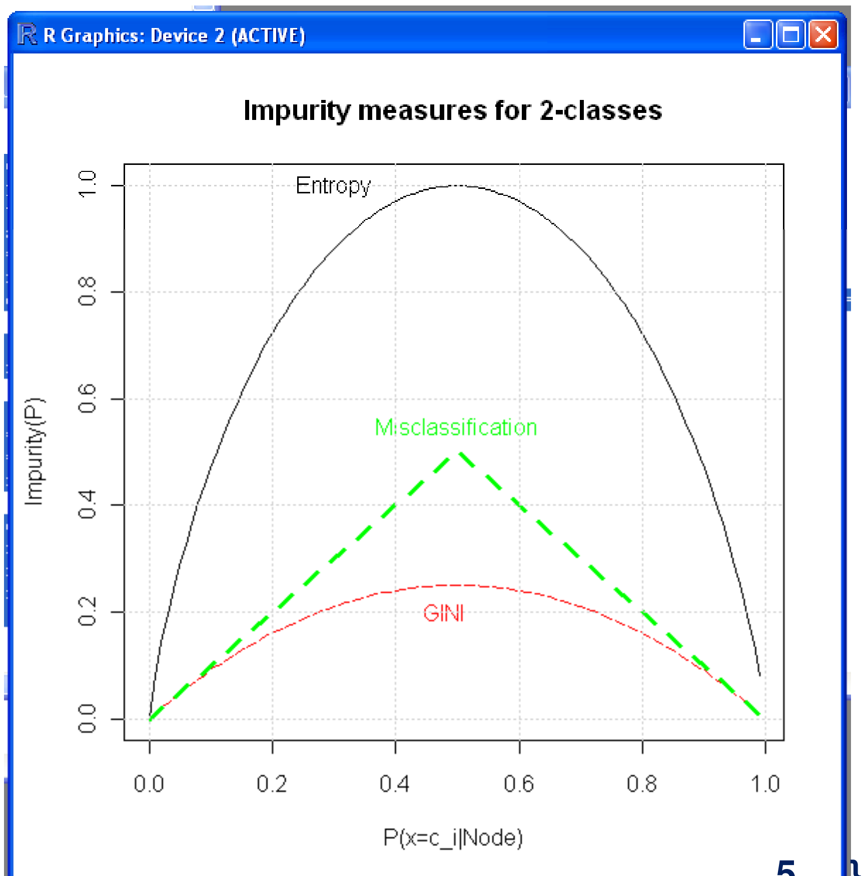
$$IG(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - remainder(A)$$

- **Choose the attribute with the largest IG**

Node Impurity for Binary Class



Node Impurity for Binary Class



1. #
2. #plot binary impurity measures
3. # plotImpurity()
4. plotImpurity = function () {
   1. x=seq(0.00001, 0.99999, by=0.01)
   2. y =numeric(length(x))
   3. for (i in 1:length(x)) y[i] = -x[i]*log(x[i],2)-(1-x[i])*log(1-x[i],2)
   4. plot(x,y, type="l", lty=1, main="Impurity measures for 2-classes", xlab="P(x=c_i|Node)", ylab="Impurity(P)")
   5. #plot GINI impurity measure
   6. lines(x, x*(1-x), col="red", lty=5, type="l")
   7.
   8. #plot misclassification error
   9. x1=c(0, 0.5, 1)
   10. y1=c(0, 0.5, 0)
   11. lines(x1, y1, col="green", lwd=3, type="l", lty="dashed")
   12. #some labels
   13. text(0.3,1.0,"Entropy", col="black")
   14. text(0.48,0.2,"GINI", col="red")
   15. text(0.5,0.55,"Misclassification", col="green")
   16. grid() #make it easy to read!
5. }

Decision tree learning

- **Aim: find a small tree consistent with the training examples**
- **Idea: (recursively) choose "most significant" attribute as root of (sub)tree**
- **Algorithms**
  - CART, ID3, C4.5, C5.0

function DTL(*examples*, *attributes*, *default*) **returns** a decision tree

    if *examples* is empty **then return** *default*
    else if all *examples* have the same classification **then return** the classification
    else if *attributes* is empty **then return** MODE(*examples*)
    else
        $best \leftarrow$ CHOOSE-ATTRIBUTE(*attributes*, *examples*)
        $tree \leftarrow$ a new decision tree with root test *best*
        for each value v_i of *best* **do**
            $examples_i \leftarrow \{$elements of *examples* with $best = v_i\}$
            $subtree \leftarrow \mathrm{DTL}(examples_i, attributes - best, \mathrm{MODE}(examples))$
            add a branch to *tree* with label v_i and subtree *subtree*
        return *tree*

[Adapted from Russell and Norvig 1995]

Decision Trees: Strengths

- **Decision trees are able to generate understandable rules (sometimes).**
- **Can be cheap to compute and in memory footprint.**
- **Decision trees are able to handle both continuous and categorical variables.**
- **Missing data is handled naturally**
- **Variable importance**
  - Decision trees provide a clear indication of which features are most important for prediction or classification.

- **Software: in R, use rpart, tree, gbm packages**

Decision Trees: Weaknesses

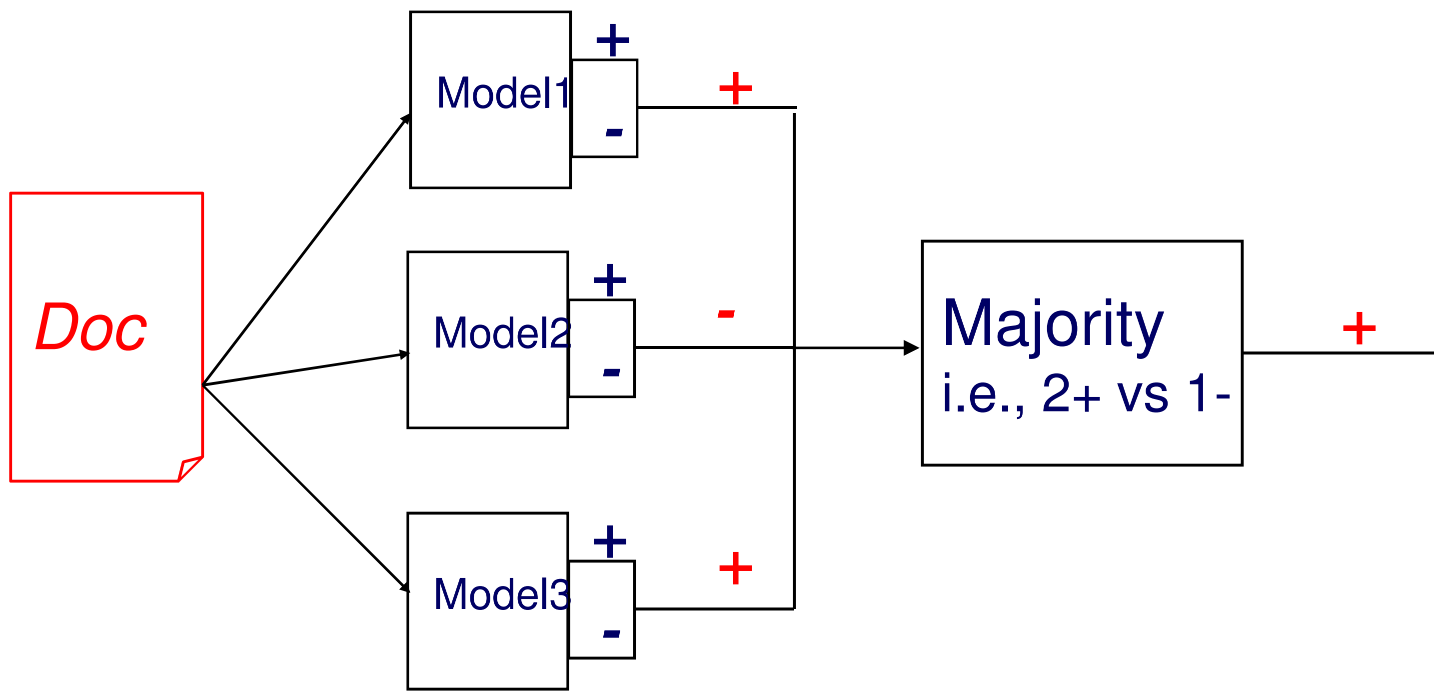
- **Decision tree can be computationally expensive to train.**
  - Need to compare all possible splits
  - Pruning is also expensive
- **Most decision-tree algorithms only examine a single field at a time.**
  - This leads to rectangular classification boxes that may not correspond well with the actual distribution of records in the decision space.
  - Each node could be a linear regression of all variables

Ad Quality Outline

- **Motivation**
- **Machine Learning Overview**
  – Background
  – Gradient Descent
  – From Perceptrons to SVMs
  – Probabilistic Models
  – Decision Trees
  – Bagging/Boosting
- **IR as a means of Ranking**
- **Learning to Rank (LETOR)**
- **Online Learning**
- **Open issues**

Ensemble Models

An ensemble model is composed of multiple individual (weak) models that are combined in some way (e.g., by voting) to classify new examples



Popular approaches include bagging [Breiman 1996] and boosting [Schapire 1990]

Bagging(bootstrap aggregation)

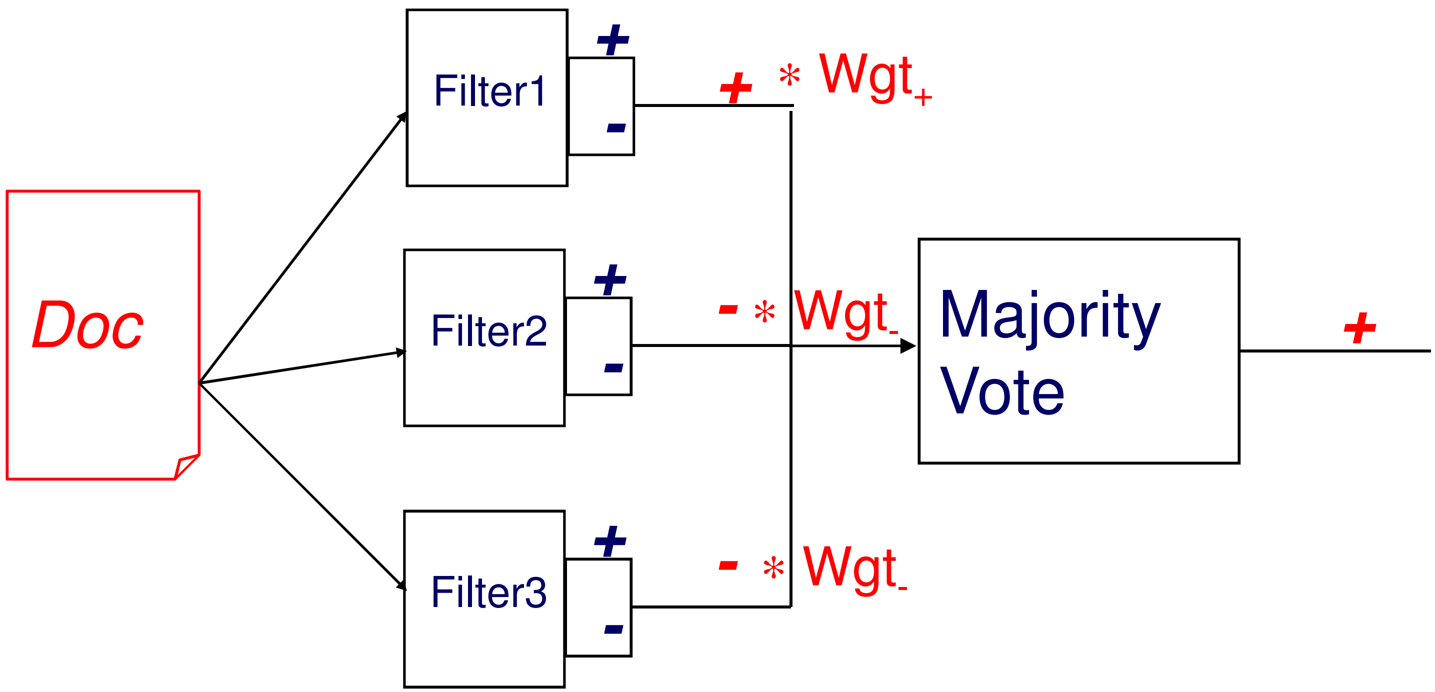
TRAINING

- **Let m be the number of instances in the training data**
- **For each of t iterations:**
  - Sample m instances with replacement from original training dataset
  - Apply the learning algorithm to the sample
  - Store the resulting model in Models[t]

CLASSIFICATION

- **For each of the t models:**
  - Predict class of the test instance using Models[t]
- **Return class that has been predicted most often**

A Bagging Filter



Here, Doc has been classified as + by Filter1, and as – by Filter2, and Filter3. Since the positive class has the majority, Doc is classified as + using this ensemble

NOTE: Here weight Wgt_+ and Wgt_- are uniform

Boosting

TRAINING (m =number of instances in the training data)

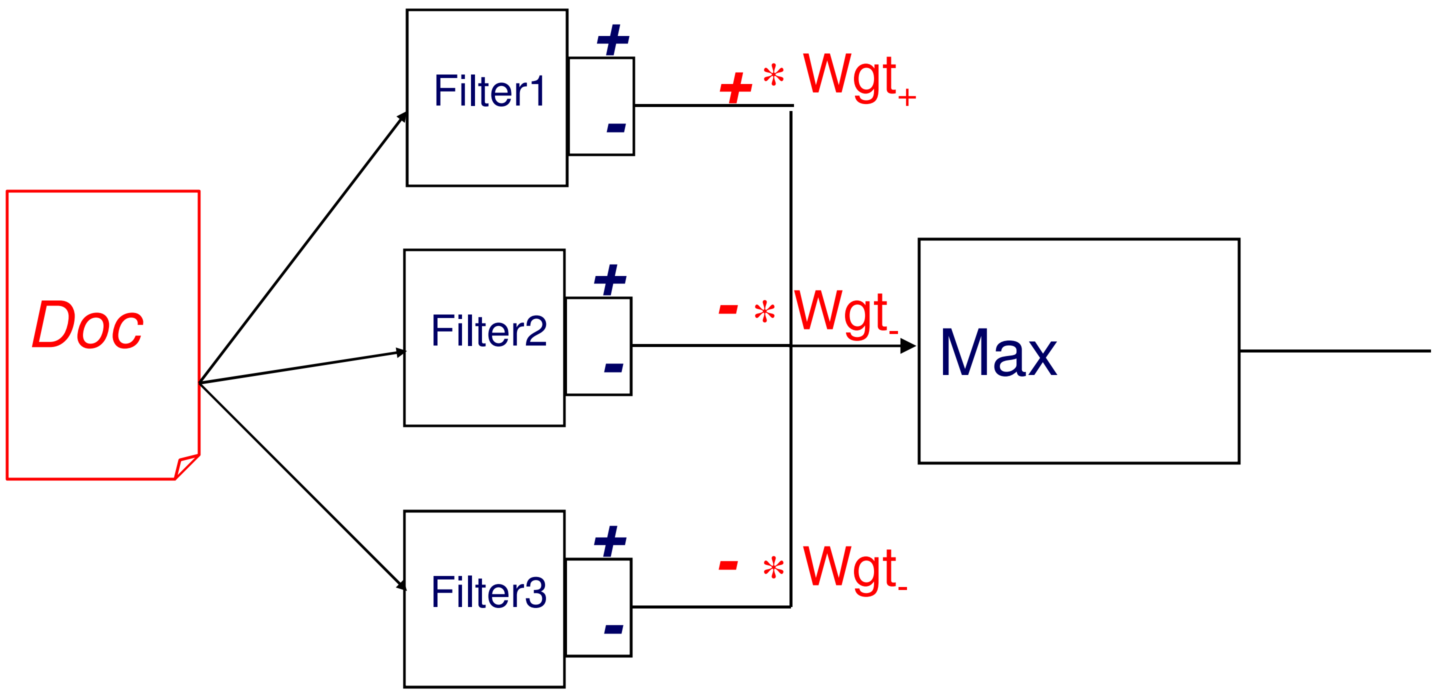
- **Assign equal weight to each training instance**
- **For each of t iterations:**
  - Sample m instances with replacement from weighted training dataset
  - Apply the learning algorithm to the sample
  - Compute error e of model on training dataset
  - If e == 0 or e > 0.5 Terminate
  - Foreach instance in training dataset:
    - If instance classified correctly by model then
      - Multiply weight of instance by e/(1-e)
  - Normalise weights of all instances
  - Set weight of model to –log(e/(1-e))
  - Store the resulting model in Models[t] and weight

Boosting

CLASSIFICATION

- **Assign weight of zero to all classes**
- **For each of the t models:**
  - Predict class of the test instance using Models[t]
  - Add Weight[t] to weight of class predicted by model[t]
- **Return class with highest weight**

A Boosting based Filter



NOTE: Here weight Wgt_+ and Wgt_- are not uniform

Gradient Boosted Decision Trees

Algorithm 2: Stochastic Gradient_TreeBoost

$F_0(\mathbf{x}) = \arg\min_\gamma \sum_{i=1}^{N} \Psi\left(y_i, \gamma\right)$

For $m = 1$ to M do:

$\quad \{\pi(i)\}_1^N = \text{rand\_perm}\ \{i\}_1^N$

$\quad \tilde{y}_{\pi(i)m} = -\left[\frac{\partial \Psi(y_{\pi(i)}, F(\mathbf{x}_{\pi(i)}))}{\partial F(\mathbf{x}_{\pi(i)})}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})},\ i = 1, \tilde{N}$

$\quad \{R_{lm}\}_1^L = L - \text{terminal node } tree(\{\tilde{y}_{\pi(i)m}, \mathbf{x}_{\pi(i)}\}_1^{\tilde{N}})$

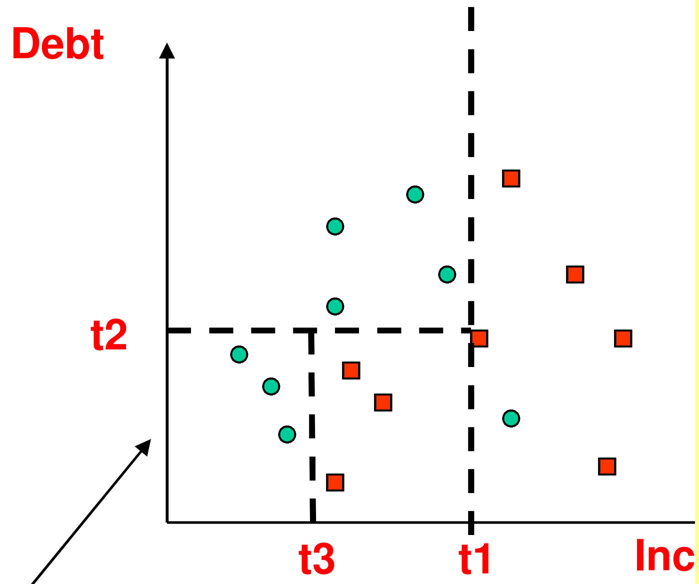
$\quad \gamma_{lm} = \arg\min_\gamma \sum_{\mathbf{x}_{\pi(i)} \in R_{lm}} \Psi\left(y_{\pi(i)}, F_{m-1}(\mathbf{x}_{\pi(i)}) + \gamma\right)$

$\quad F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} 1(\mathbf{x} \in R_{lm})$

endFor

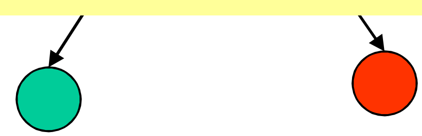
[Friedman 1999]

Gradient Boosted Decision Trees



Gradient Boosted Decision Trees are an additive classification or regression model consisting of an ensemble of trees, fitted to current residuals, gradients of the loss function, in a forward step-wise manner. In the traditional boosting framework, the weak learners are generally shallow decision trees consisting of a few leaf nodes.
GBDT ensembles are found to work well when there are hundreds of such decision trees. Gradient Boosted Decision Trees was introduced by Jerome Friedman in 1999.

Tree boundaries piece-wise linear linear and axis-parallel



Gradient Descent Boosting
Learn a committee of trees
Model the residual at each iteration
[Friedman, 2000]

R code

- > load (" wine . RData ")
- > ds <- wine
- > ds$ Type <- as. numeric (ds$ Type)
- > ds$ Type [ds$Type >1] <- 0
- > ds$ Type
- [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
- > ds.gbm <- gbm(Type ~ Alcohol + Malic + Ash + Alcalinity + Magnesium Phenols + Flavanoids + Nonflavanoids + Proanthocyanins Color + Hue + Dilution + Proline ,
- data =ds , distribution =" adaboost ", n. trees =100)

Relative Importance of Variables

```
> summary(ds.gbm)
               var   rel.inf
1          Proline  91.82978
2       Flavanoids   8.17022
3          Alcohol   0.00000
4            Malic   0.00000
5              Ash   0.00000
6       Alcalinity   0.00000
7        Magnesium   0.00000
8          Phenols   0.00000
9    Nonflavanoids   0.00000
10 Proanthocyanins   0.00000
11           Color   0.00000
12             Hue   0.00000
13        Dilution   0.00000
> pretty.gbm.tree(ds.gbm)
  SplitVar SplitCodePred LeftNode RightNo
0       12  8.675000e+02        1
3       65.36408       89
1       -1 -8.139656e-04       -1          -1
-1        0.00000       62
2       -1  9.236987e-04       -1          -1
-1        0.00000       27
3       -1 -2.868090e-04       -1          -1
-1        0.00000       89
    Prediction
0 -0.0002868090
1 -0.0008139656
2  0.0009236987
3 -0.0002868090
```

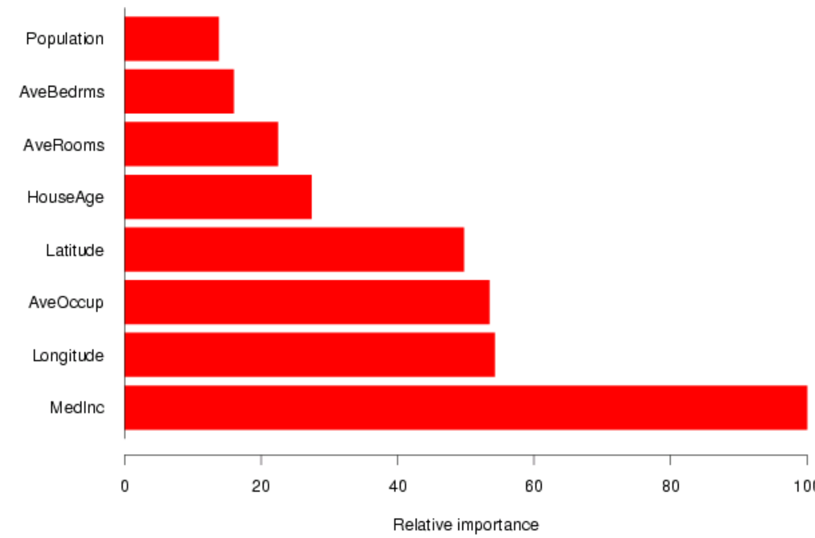


Figure 10.14: *Relative importance of the predictors for the California housing data.*

$$I_l^2(T) = \sum_{\text{over all node using } x_1 \text{ for partition}} (f(x_i) - y_i)^2 = \sum_{\text{over all node using } x_1 \text{ for partition}} (\mu(x_i) - y_i)^2$$

$$I_l^2 = \frac{1}{M}\sum_{m=1}^{M} I_l^2(T_m)$$

Gradient Boosted Trees

```
> gbm.show.rules(ds.gbm)
Number of models: 100

Tree 1: Weight XXXX
  Proline < 867.50 : 0 (XXXX/XXXX)
  Proline >= 867.50 : 1 (XXXX/XXXX)
  Proline missing : 0 (XXXX/XXXX)
[...]
Tree 100: Weight XXXX
  Proline < 755.00 : 0 (XXXX/XXXX)
  Proline >= 755.00 : 1 (XXXX/XXXX)
  Proline missing : 0 (XXXX/XXXX)
```

Relative Importance of Variables

– For a single tree, define the importance of x_l as

$$\mathrm{I}_l^2(T) = \sum_{\text{over all node using } x_l \text{ for partition}} \text{improve in square error risk over for a constant fit over the region}$$

– For additive tree, define the importance of x_l as

$$\mathrm{I}_l^2 = \frac{1}{M}\sum_{m=1}^{M} \mathrm{I}_l^2(T_m)$$

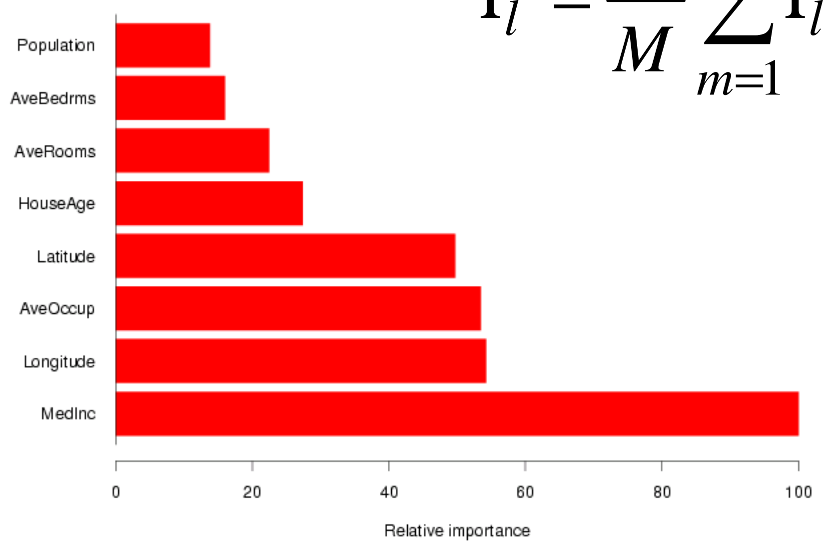


Figure 10.14: *Relative importance of the predictors for the California housing data.*

James G. Shanahan (San Francisco)
James.Shanahan_AT_gmail_DOT_com

Ad Quality Outline

- **Motivation**
- **Machine Learning Overview**
  - Background
  - Gradient Descent
  - From Perceptrons to SVMs
  - Probabilistic Models
  - Decision Trees
  - Bagging/Boosting
- **Metrics**
- **IR as a means of Ranking**
- **Learning to Rank (LETOR)**
- **Online Learning**
- **Open issues**

Metrics (some examples)

- **ML Metrics**
  - Mean squared error , Accuracy, Maximum Likelihood, Deviance, Cross entropy, KL Divergence, etc.
- **IR metrics**
  - Precision, Recall, MRR, DCG, pRank etc.
  - CIKM 2009 Paper on a new flavor of DCG!!
- **Advertising Metrics**
  - CTRs, Revenue, ECPM

Possible Document Classifications

Document Type	Retrieved	Not Retrieved
Relevant	R^+	R^- Type II Error (False Negative)
Not Relevant	N^+ Type I Error (False Positive)	N^-

$$accuracy = \frac{R^+ + N^-}{R^+ + R^- + N^+ + N^-} \quad recall = \frac{R^+}{R^+ + R^-} \quad precision = \frac{R^+}{R^+ + N^+}$$

$$F_{0.5} = \frac{1.25 * R^+}{0.25 * R^- + N^+ + 1.25 * R^+}.$$

User Models

- **User satisfaction/expectation can be expressed as a utility measure**
- **Linear utility provides an empirically proven evaluation metric**
  - TREC (Text Retrieval Conference)
  - *Linear Utility= $2R^{+}$ - N^{+}*
- **Incorporate user models into SVM threshold relaxation operation**

Document Type	Retrieved	Not Retrieved
Relevant	R^{+}	R^{-} Type II (FN)
Not Relevant	N^{+} Type I (FP)	N^{-}

Metrics for Web Search

- **Existing metrics limited**
  - Not always clear-cut binary decision: relevant vs. not relevant
  - Not position sensitive:
    d: *definitely relevant*, p: *partially relevant*, n: *not relevant*
    ranking 1: p d p n n n n
    ranking 2: d p n p n n n
- **How do you measure recall over the whole web?**
  - How many of the potentially billions results will get looked at? Which ones actually need to be good?
- **Normalized Discounted Cumulated Gain (NDCG)**
  - K. Jaervelin and J. Kekaelaeinen (TOIS 2002)
  - Gain: relevance of a document is no longer binary
  - Sensitive to the position of highest rated documents
    - Log-discounting of gains according to the positions
  - Normalize the DCG with the “ideal set” DCG (NDCG).

Normalized Discounted Cumulative Gain (NDCG)

- **Judgments**
  - E.g., Very relevant > Somewhat relevant > Not relevant
- **Want most relevant documents to be ranked first**

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- **NDCG = DCG / ideal ordering DCG**
  - NDCG Ranges from 0 to 1

NDCG Example

- **Proposed document ordering:**

4 2 0 1

- **DCG = 4 + 2/log(2) + 0/log(3) + 1/log(4)**
  - = 6.5

- **IDCG = 4 + 2/log(2) + 1/log(3) + 0/log(4)**
  - = 6.63

- **NDCG = 6.5/6.63 = .98**

Normalized Discounted Cumulative Gain (NDCG) evaluation measure

- Query: q_i
- DCG at position *m*: $N_i = Z_i \sum_{j=1}^{m} (2^{r(j)} - 1)/\log(1+j)$
- NDCG at position *m*: average over queries
- Example
  - (3, 3, 2, 2, 1, 1, 1) **rank *r***
  - (7, 7, 3, 3, 1, 1, 1) **gain** $2^{r(j)} - 1$
  - (1, 0.63, 0.5, 0.43, 0.39, 0.36, 0.33) **discount** $1/\log(1+j)$
  - (7, 11.41, 12.91, 14.2, 14.59, 14.95, 15.28) $\sum_{j=1}^{m} (2^{r(j)} - 1)/\log(1+j)$
- Z_i normalizes against best possible result for query, the above, versus lower scores for other rankings
  - Necessarily: High ranking number is good (more relevant)

[IIR secs 6.1.2–3 and 15.4]

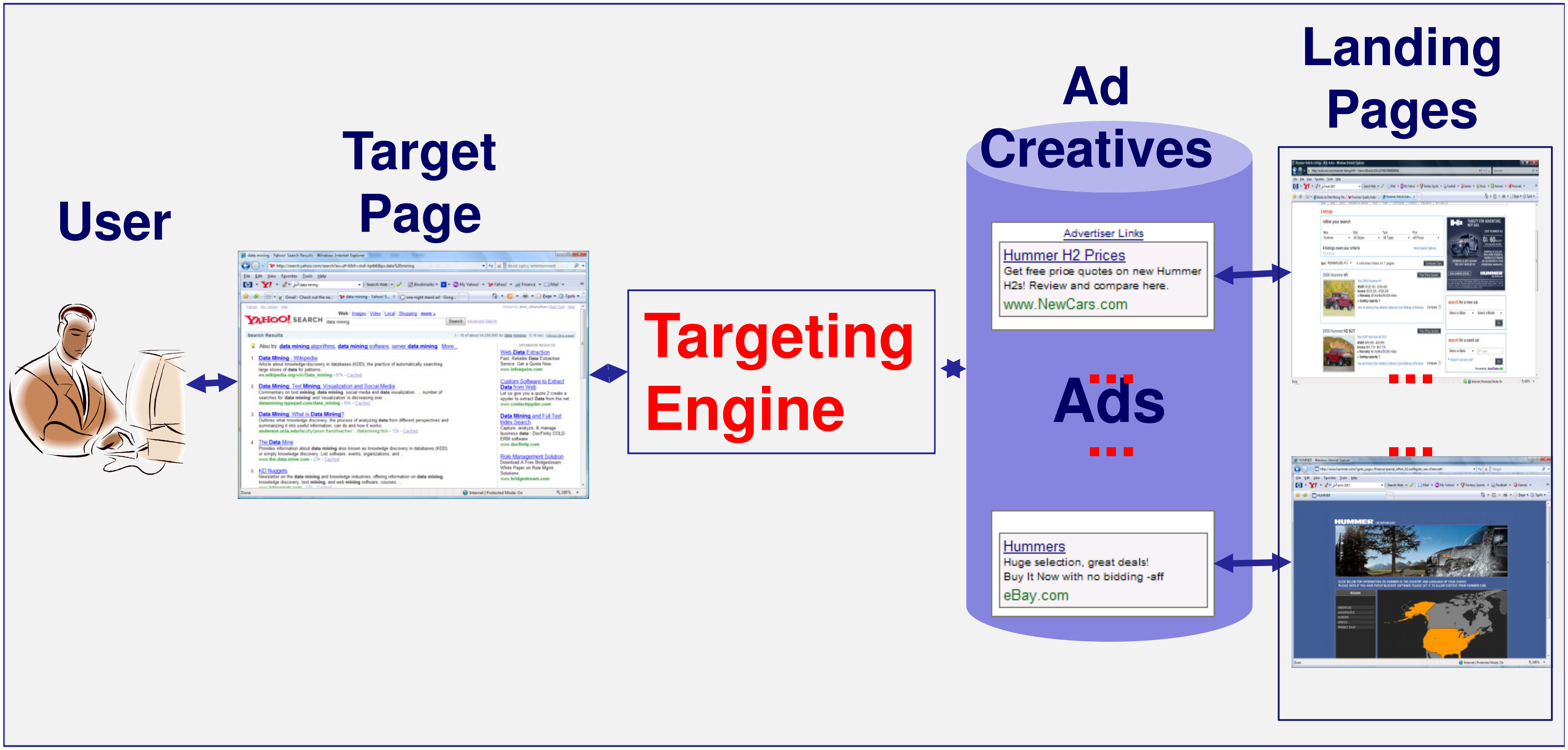
Ad Quality Outline

- Motivation
- Machine Learning Overview
- Metrics
- IR as a means of Ranking Ads
- Learning to Rank (LETOR)
- Online Learning
- Open issues

Ad Quality Outline

- Motivation
- Machine Learning Overview
- Metrics
- IR as a means of Ranking Ads
- Learning to Rank (LETOR)
- Online Learning
- Open issues

Ranking Ads using IR



Features

View TP as query

IR Engine

View Ad as document with different sections

- Keywords
- Title + Description +URL
- Landing Page

Ranking ads using IR

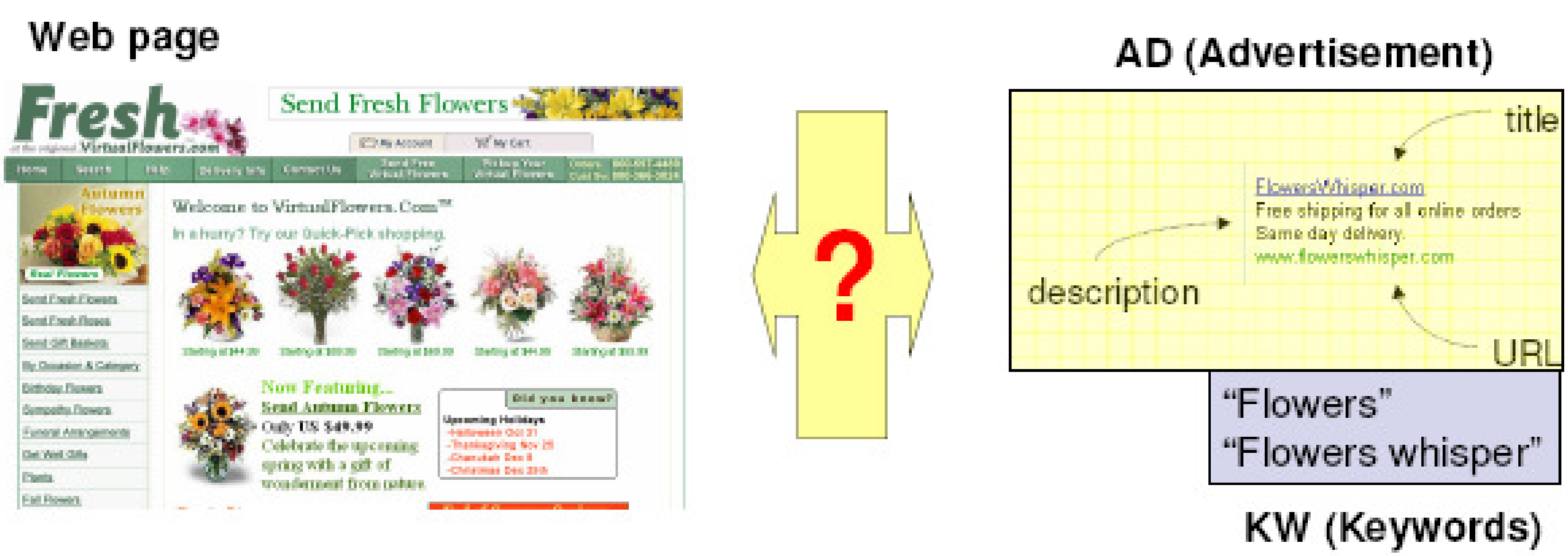
- **Contextual advertising study**
- **View target page as a query and use IR techniques**
- **Features**
  - Rank ads for target page
    - by matching target page words with ad keywords and ad creative words
  - Variations
    - Query Expansion; expand keywords of target page
    - Use words from landing page

$$sim(q, d_j) = \frac{\vec{q} \bullet \vec{d_j}}{|\vec{q}| \times |\vec{d_j}|} = \frac{\sum_{i=1}^{n} w_{iq} \cdot w_{ij}}{\sqrt{\sum_{i=1}^{n} w_{iq}^2}\sqrt{\sum_{i=1}^{n} w_{ij}^2}}$$

See [Berthier A. Ribeiro-Neto, Marco Cristo, Paulo Braz Golgher, Edleno Silva de Moura: Impedance coupling in content-targeted advertising. SIGIR 2005: 496-503]

Content-Targeted Advertising
Matching Strategies

- *Match the ad and its keywords to the Web page*



[Ribeiro-Neto et al., SIGIR 2005]

Content-Targeted Advertising
Experimental Evaluation

- Ad Sample Collection
  - 1,744 announcers **(Advertiser)**
  - 93,972 ads in 2,029 campaigns
  - 68,238 keywords
- Test Collection **(100 Queries)**
  - 100 pages of a Brazilian newspaper
  - Topics include economy, sports, culture, and politics
- Matching Function
  - Vector based with tf-idf weights

[Ribeiro-Neto et al., SIGIR 2005]

Test Set Creation

15 Users manually labeled up to 30 ads per target page (on average, 5 from 15 we... judged relevant)

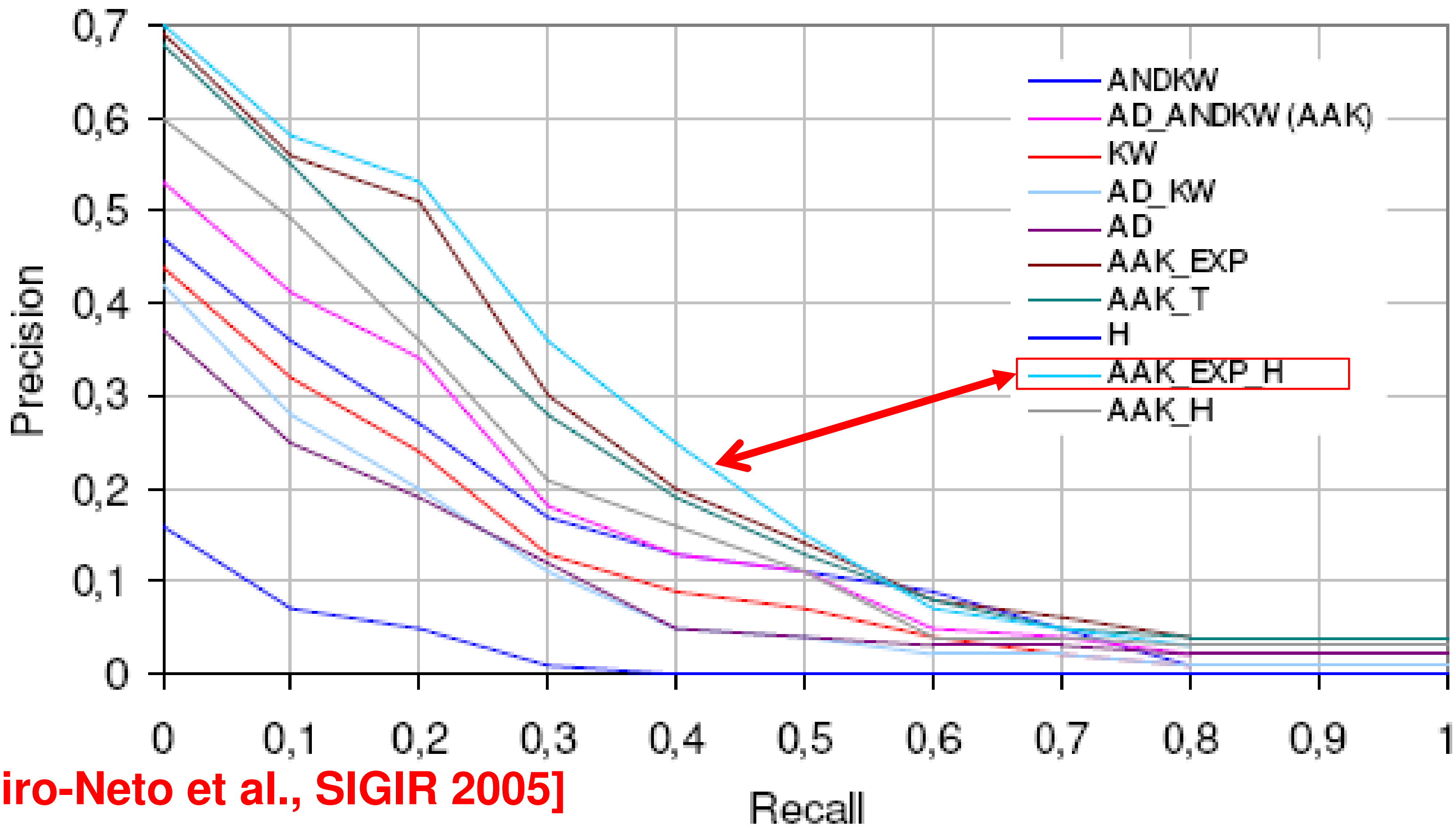
For each of our 100 target p... e
ranked ads provided by ... our ...
gies. Thus, for eac... pag...
ads. These t... were ... that
target p... ... adver-
tisem... All ... were submitted
to a man... users. The average
num... per page pool was 5.15.
N... same pooling method used to
eva... -based collection [6].

What type of IR approaches could be used?

MT approach: [Rion Snow, Brendan O'Connor, Daniel Jurafsky and Andrew Y. Ng. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. To appear in Proceedings of EMNLP 2008]

Content-Targeted Advertising
Comparison among All Methods

AAK_EXP_H = TP Expansion, Keywords, Landing Page



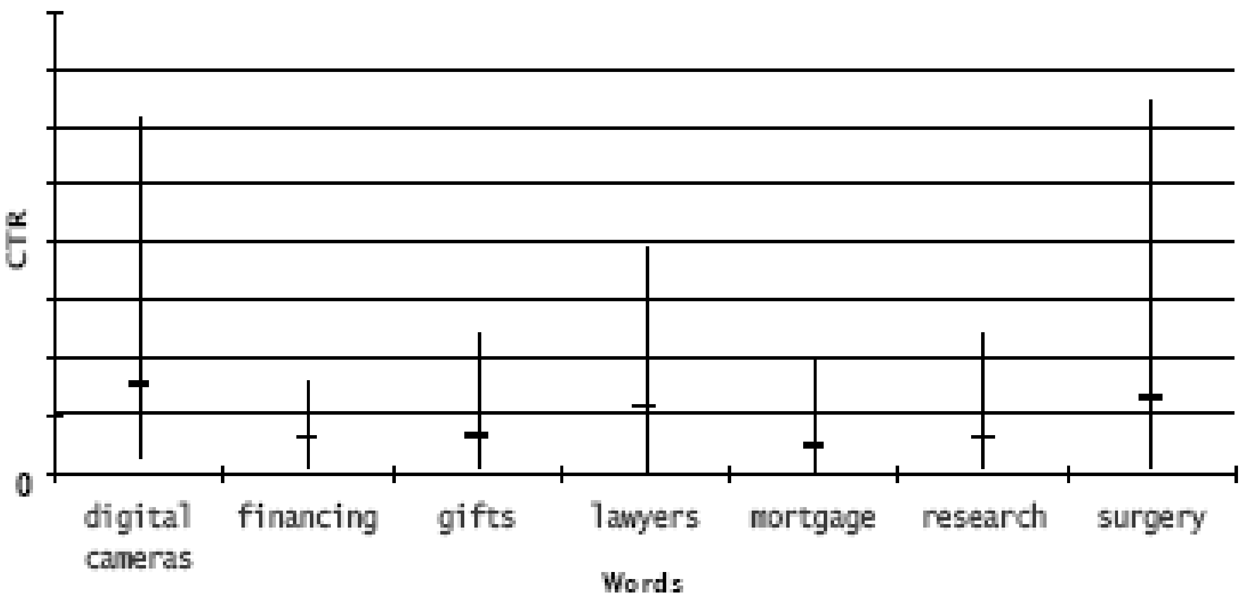
[Ribeiro-Neto et al., SIGIR 2005]

Ranking Ads using IR

- **Ribeiro et al. study was limited**
  - Evaluated ranking of ads but not CTR estimation; this is a much easier problem;
  - System would not yield accurate CTR estimates in practice
  - Study highlights gap between academia and industry and a way of bridging it
- **Jones, R., Rey, B., Madani, O., and Greiner, W. 2006. Generating query substitutions.**
  - In *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, NY, 387-396.
- **This could be seen as the first step in ad serving**
  - i.e., select a subset of ads and then do ECPM-based ranking of the this subset

Estimating CTRs using KW Clustering

- **Estimate the CTR of new ads by using the CTRs of existing ads with the same bid terms or topic clusters**
  – See [Regelson and Fain, 2006]
  – However, keywords have highly variable CTRs (e.g., max of ***"Lawyers"*** is 300% that of the average CTR)



Min, Avg, Max CTR across all ads for a KW_i

[Richardson et al, 2007]

Other IR approaches to OAT

- **Leverage a taxonomy distance and traditional text match**
  – Classify query and ad into a taxonomy (node represents a set of queries)
  – Weighted sum of taxonomy/topic similarity score and traditional IR score.
  – [A Semantic Approach to Contextual Adv.,Broder, et al., SIGIR 2007]
- **Keyword harvesting from webpages**
  – Treat target page as a query but can be long. Suggest keywords. Which terms to extract? Train a LogReg classifier on hand-picked keywords and their features. Predict if a word is keyword or not. Study performed at Microsoft Research
  – [Yih et al., WWW 2006, "Finding advertising keywords on web pages"]
- **Ad Matching for Rare Queries (enhancing broad match)**
  – Use classes of top ranked pages (for a query) to suggest ads
  – [Robust Classification of Rare Queries…, Broder, et al., SIGIR 2007]
- **All the problems in traditional IR + more (see WWW, SIGIR CIKM)**
  – Synonym detection; entity extraction from queries; query disambiguation

Ad Quality Outline

- Motivation
- Machine Learning Overview
- Metrics
- IR as a means of Ranking
- Learning to Rank (LETOR)
- Online Learning
- Open issues

Learning to Rank

- **Learning to rank documents (main approaches)**
  - Point-based
  - Pair-wise-based
  - List-based
- **Very active area of research**
  - LETOR is a package of benchmark data sets for LEarning TO Rank, released by Microsoft Research Asia.
  - Extracted features for each query-document pair in the OHSUMED and TREC collections
  - Benchmarked several state-of-the-arts ranking models with these features and provide baseline results for future studies
- **Similar principles apply to ranking of ads**

Table 3. All the features for the TREC datasets

Feature	Descriptions	References
1	BM25	[27]
2	document length (dl) of body	[1]
3	dl of anchor	
4	dl of title	
5	dl of URL	
6	HITS authority	[21]
7	HITS hub	
8	HostRank (SIGIR feature)	[34]
9	Inverse document frequency (idf) of body	[1]
10	idf of anchor	
11	idf of title	
12	idf of URL	
13	Sitemap based score propagation (SIGIR feature)	[28]
14	PageRank	[25]
15	LMIR.ABS of anchor	[35],
16	BM25 of anchor	[35]
17	LMIR.DIR of anchor	
18	LMIR.JM of anchor	
19	LMIR.ABS of extracted title (SIGIR feature)	
20	BM25 of extracted title (SIGIR feature)	[29]
21	LMIR.DIR of extracted title (SIGIR feature)	[35]
22	LMIR.JM of extracted title (SIGIR feature)	
23	LMIR.ABS of title	
24	BM25 of title	[29]
25	LMIR.DIR of title	[35]
26	LMIR.JM of title	
27	Sitemap based feature propagation (SIGIR feature)	[28]
28	tf of body	[1]
29	tf of anchor	
30	tf of title	
31	tf of URL	
32	tf*idf of body	
33	tf*idf of anchor	
34	tf*idf of title	
35	tf*idf of URL	
36	Topical PageRank (SIGIR feature)	[24]
37	Topical HITS authority (SIGIR feature)	
38	Topical HITS hub (SIGIR feature)	
39	Hyperlink base score propagation: weighted in-link (SIGIR feature)	[28] [30]
40	Hyperlink base score propagation: weighted out-link (SIGIR feature)	
41	Hyperlink base score propagation: uniform out-link (SIGIR feature)	
42	Hyperlink base feature propagation: weighted in-link (SIGIR feature)	
43	Hyperlink base feature propagation: weighted out-link (SIGIR feature)	
44	Hyperlink base feature propagation: uniform out-link (SIGIR feature)	

Text, Anchor Text, Web graph

Great start but be aware….

Missing Features

-Log features

-User features

-……and others

Current LETOR baselines

- **LETOR Algorithms**
  - Ranking SVM
  - RankBoost
  - AdaRank
  - Multiple hyperline ranker
  - FRank
  - ListNet
- **LETOR Datasets**
- **LETOR metrics such as DCG, NGCG**

(b) Mean average precision

Algorithms	MAP
RankBoost	0.383514
Ranking SVM	0.350459

LETOR test collection

- **From Microsoft Research Asia**
- **An openly available standard test collection with pregenerated features, baselines, and research results for learning to rank**
- **It's availability has really driven research in this area**
- **OHSUMED, MEDLINE subcollection for IR**
  - 350,000 articles
  - 106 queries
  - 16,140 query-document pairs
  - 3 class judgments: Definitely relevant (DR), Partially Relevant (PR), Non-Relevant (NR)
- **TREC GOV collection (predecessor of GOV2, cf. *IIR* p. 142)**
  - 1 million web pages
  - 125 queries

Dataset Partitioning

Table 4. Data Partitioning for 5-fold Cross Validation

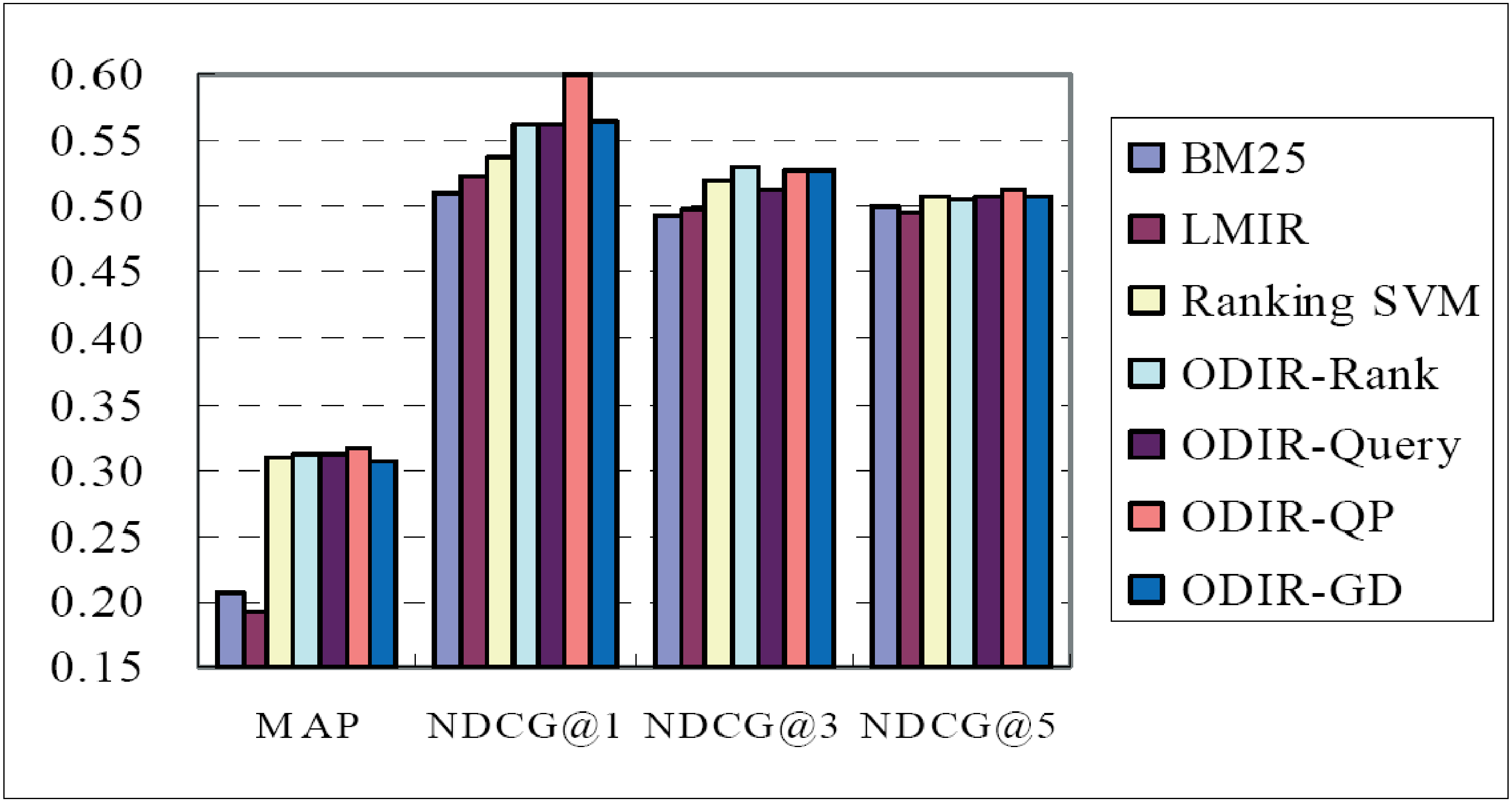
Folds	Training set	Validation set	Test set
Fold1	{S1, S2, S3}	S4	S5
Fold2	{S2, S3, S4}	S5	S1
Fold3	{S3, S4, S5}	S1	S2
Fold4	{S4, S5, S1}	S2	S3
Fold5	{S5, S1, S2}	S3	S4

- **five-fold cross validation**
  - The training set is used to learn the ranking model.
  - The validation set is used to tune the parameters of the ranking model.
  - The test set is used to report the ranking performance of the model.

Homework

- Download LETOR OHSUMED
- Model ranking function using linear regression (in R) and compare with a standard SVM, ranking SVM
- Compute DCG, NDCG scores and discuss your findings. Base results on five-fold cross validation

Experimental Results (OHSUMED)



Resources

- ***IIR*** **secs 6.1.2–3 and 15.4**
- **LETOR benchmark datasets**
  - Website with data, links to papers, benchmarks, etc.
  - http://research.microsoft.com/users/LETOR/
  - Everything you need to start research in this area!
- **SIGIR 2007-9 workshops**
- **Nallapati, R. Discriminative models for information retrieval. *SIGIR 2004.***
- **Cao, Y., Xu, J. Liu, T.-Y., Li, H., Huang, Y. and Hon, H.-W. Adapting Ranking SVM to Document Retrieval, *SIGIR 2006.***
- **Y. Yue, T. Finley, F. Radlinski, T. Joachims. A Support Vector Method for Optimizing Average Precision. *SIGIR 2007*.**

Learning to Rank Ads from Clicks

- **LETOR Ads in the context of sponsored search**
- **More relevant versus absolutely relevant (preference judgements)**
  - Infer pairwise preference of ads for a query using clicks based on the block approach proposed by [Joachims, KDD 2002]
  - Dropped clicked ads at rank 1; queries of 3 terms or more
  - 123,798 blocks [each query could be made up multiple blocks]
- **Evaluated linear and nonlinear learning algorithms**
  - Linear perceptron in classification mode {-1, +1}
  - Linear preceptron in pairwise learning mode [Ranking SVM]
  - Multilayer perceptron (Neural network) in classification mode {-1, +1}

[Ciaramita M., Murdock V., Plachouras, Online Learning from Click Data for Sponsored Search, WWW 2008]

Getting labeled data from clicks

Ad Id	Clicked or not
14	Click
2	----
3	Clicked by user1
1	----
6	----
7	Clicked by user1

Delete Clicks at rank 1

Block 1: ad3>>ad2)

Block 2: ad7 >> ad6
ad7 >> ad1

singles ad3 labeled as +1 and ad2 as -1

pairwise
ad3>>ad2 is labeled as +1
ad2 << ad3 is labeled -1

Ranking SVM: learning pairwise

Herbrich et al (2000)

- Input space: X
- Ranking function $f : X \rightarrow R$
- Ranking: $x_i \succ x_j \Leftrightarrow f(x_i; w) > f(x_j; w)$
- Linear ranking function: $f(x; w) = \langle w, x \rangle$

$$\langle w, x^{(1)} - x^{(2)} \rangle > 0 \quad \Leftrightarrow \quad f(x^{(1)}; w) > f(x^{(2)}; w)$$

- Transforming to binary classification:

$$(\vec{x}^{(1)} - \vec{x}^{(2)}, z), \; z = \begin{cases} +1 & x^{(1)} \succ x^{(2)} \\ -1 & x^{(2)} \succ x^{(1)} \end{cases}$$

Shen and Joshi adapted this approach to the perceptron; [Shen and Joshi 2005]

Features

Feature Name	Abbrev.	Description
		Word Overlap Features
NoKey	O	1 if no query term is present in the ad materials; 0 otherwise
SomeKey	O	1 if at least one query term is present in the ad materials; 0 otherwise
AllKey	O	1 if every query term is present in the ad materials; 0 otherwise
PercentKey	O	The number of query terms present in the ad materials divided by the number of query terms
		Cosine Similarity Features
Ad	B	The cosine similarity between the query and the ad materials (baseline)
Title	F	The cosine similarity between the query and the ad title
Description	F	The cosine similarity between the query and the ad description
Bidterm	F	The cosine similarity between the query and the bidded terms
		Correlation Features
AvePMI	P	The average pointwise mutual information between terms in the query and terms in the ad
MaxPMI	P	The maximum pointwise mutual information between terms in the query and terms in the ad
CSQ	C	Number of query-ad term pairs that have a χ^2 statistic in the top 5% of computed χ^2 values.

Table 1. Summary of features. The column "Abbrev." provides an abbreviated name for one or more features, as they will be used in the experiments.

MLP seems to give best results

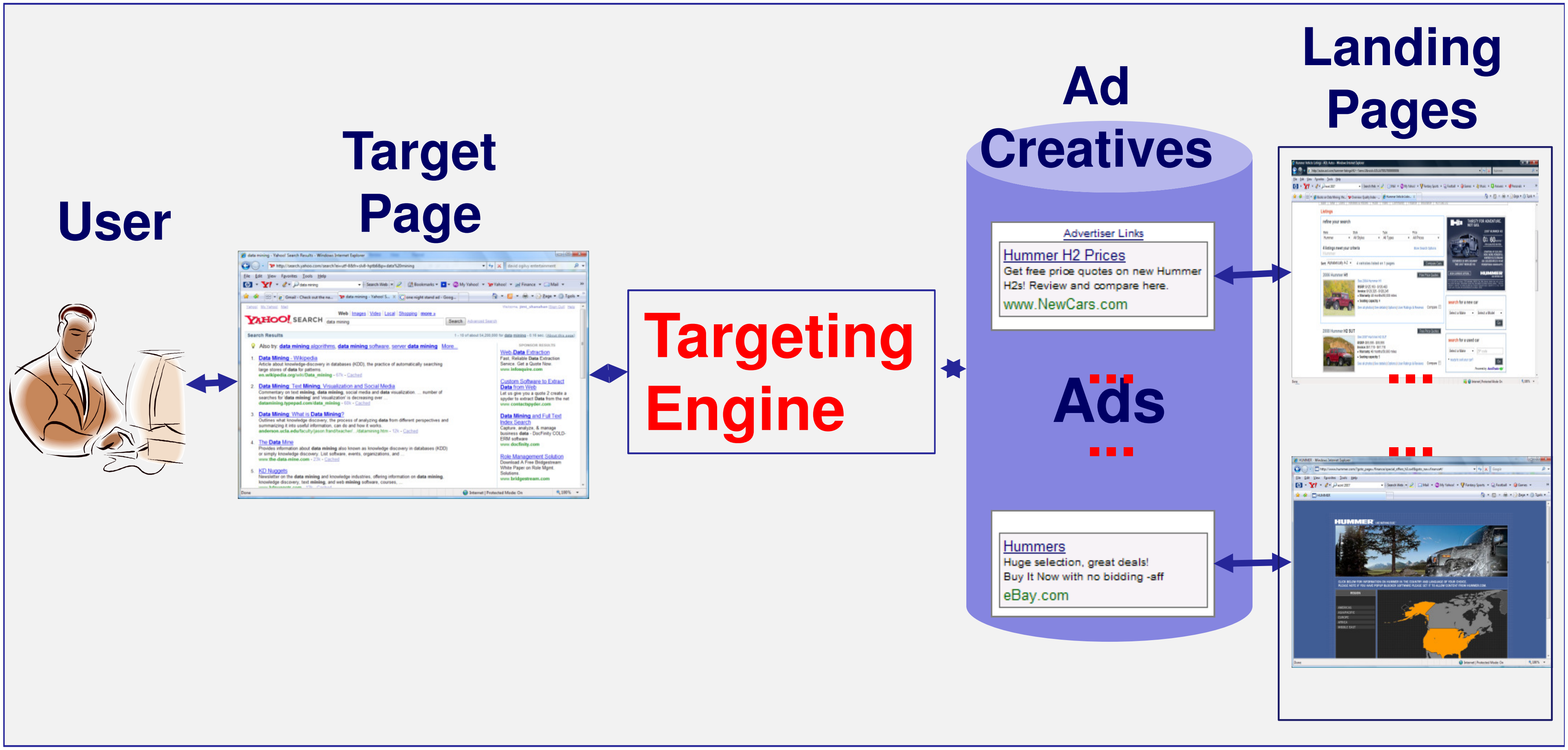
- ….though not sure about the significance compared to other learnt models
- Limited to ranking of ads

RR=1/RankOfFirstRel

	Classification		Ranking		Regression	
Feature set	Prec at 1	MRR	Prec at 1	MRR	Prec at 1	MRR
B	0.322	0.582 ±0.306	0.333	0.590 ±0.307	0.328	0.585 ±0.307
BO	0.319	0.578⋆ ±0.306	0.352	0.602⋆ ±0.310	0.343	0.596⋆ ±0.309
BF	0.341	0.593⋆ ±0.309	0.347	0.597⋆ ±0.310	0.374	0.615⋆ ±0.314
BFO	0.357	0.605⋆ ±0.311	0.357	0.605⋆ ±0.311	0.371	0.614⋆ ±0.313
BFOP	0.357	0.604⋆ ±0.311	0.359	0.606⋆ ±0.311	0.374	0.617⋆ ±0.313
BFOC	0.351	0.601⋆† ±0.310	**0.364**	**0.610⋆† ±0.311**	0.381	0.619⋆† ±0.315
BFOCP	**0.360**	**0.606⋆ ±0.311**	0.363	0.609⋆ ±0.311	**0.388**	**0.624⋆† ±0.315**

Table 4. The results for classification, ranking and regression, computed over all trials. The best result is indicated in bold. Results that are statistically significant with respect to the baseline are indicated with a star. Results indicated with a dagger are statistically significant with respect to the features B + F + O. The results for precision at one were not tested for statistical significance.

Ranking Ads using IR



Zip Code
BrowserType
Etc

Terms
etc.

ML
Model

Keywords
Title + Description +URL
Landing Page
Appearance

Estimating CTRs using ML

- **Estimate CTR using Pr_{Ad}(Click|Keyword)**
- **Frame as machine learning problem**
  - E.g., Matthew Richardson, Ewa Dominowska, Robert Ragno: Predicting clicks: estimating the click-through rate for new ads. WWW 2007 pages 521-530
    - Model using Logistic Regression and MART (B… decision trees using stochastic gradient de… [Friedman 2000])
  - Esteban Feuerstein, Pablo Heiber, Jav… rtinez-Viademonte and Ricardo Baeza-Y… New Stochastic Algorithms for Placing Ads in S… ed Search. LA-Web, Santiago, Chile 2007

What features could be used?

ML Features 1/2

Features(KW,AD, LP)->CTR
X_i ->CTR_i

- **Historical data**
  – CTR of KW based on other ads with this KW
  – Related terms CTRs

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}.$$

- **Appearance**
  – #words in title/body; capitalization; punctuation; word length
- **Attention Capture**
  – Title/body contain action words, e.g., buy/join/etc
- **Reputation**
  – .com/.net/etc, length of URL, #segments in URL, numbers in URL
- **Landing page quality**
  – Contains flash? Fraction of page in images? W3C compliant
- **Text Relevance**
  – keyword match with ad title/body; fraction of match

[Richardson et al., 2007]

ML Features 2/2

- Historical data
- Related terms CTRs
- Appearance
- Attention Capture
- Reputation
- Landing page quality
- **Text Relevance**
  – keyword match with ad title/body; fraction of match
- **10K unigrams (appearing in Ad title and Ad body); bi/trigrams did not bring significant improvement;**
  – Binary feature; 1 if term occurs in ad 0 otherwise
- **Freq of term on web; in query logs**
- **Many others could be used!!!** **[Richardson et al.]**

Learning Setup

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}.$$

- **Logistic Regression**
  - Used a cross entropy loss function
  - Standardized all features using training data
    - (mean and variance, of 0 and 1)
  - Thresholded data beyond 5 std deviations
  - Added derived features
    - (i.e., foreach feature f, log(f +1) and f^2)
- **Baseline**
  - Predict the average CTR of the training dataset
- **MART (Boosted decision trees using stochastic gradient descent [Friedman 2000])**
  - Experiments did not show significant improvement over LR
  - LR is a more transparent model

For LR see:

1. http://en.wikipedia.org/wiki/Logistic_regression
2. http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_4.html

Learning Setup

- **Error measures**
  - Mean Squared error between predicted CTR and true CTR
  - KL Divergence between the predicted CTR and true CTR (in both cases lower is better; 0 is best)
- **Issues?**
  - Weighted?
  - ??

Dataset

- **10,000 Advertisers**
- **1 Million examples of <Keyword, Ad> -> CTR**
  - (view <Keyword, Ad> as <TP, Ad>)
- **Keywords are both exact and broad match**
- **100,000 unique ad texts**
- **Required that each example had more than 100 views**
- **70-10-20 data split (train, validation, test)**

[Richardson et al.]

Results

Table 7: Comparison of results for a model trained and tested on ads with over 100 views vs. over 1000 views.

	%Imprv	
Features	>100 views	>1000 views
Baseline ($\overline{CTR}$)	-	-
+Term CTR	13.28	25.22
+Related CTR	19.67	32.92
+Ad Quality	23.45	33.90
+Order Specificity	28.97	40.51
+Search Data	29.47	41.88

Transparency of Results

Table 5: Non-unigram features with highest (lowest) weight

Top ten features	*Bottom ten features*
log(#chars in term)	log(# terms in order)
v_{12}	log(v_{0*})
v_{22}	sqr(p_{00})
log(order category entropy)	sqr(order category entropy)
log(#most common word)	log(#chars in landing page)
sqr(#segments in displayurl)	log(a_{01})
sqr(#action words in body)	a_{13}
p_{10}	sqr(p_{0*})
p_{**}	log(#chars in body)
log(v_{00})	sqr(#chars in term)

Table 6: Unigrams with highest (and lowest) weight.

Top ten unigrams		*Bottom ten unigrams*	
official	body	quotes	title
download	title	hotels	title
photos	body	trial	body
maps	body	deals	body
official	title	gift	body
direct	body	have	text
costumes	title	software	title
latest	body	engine	body
version	body	compare	title
complete	body	secure	body

CTR Evolution

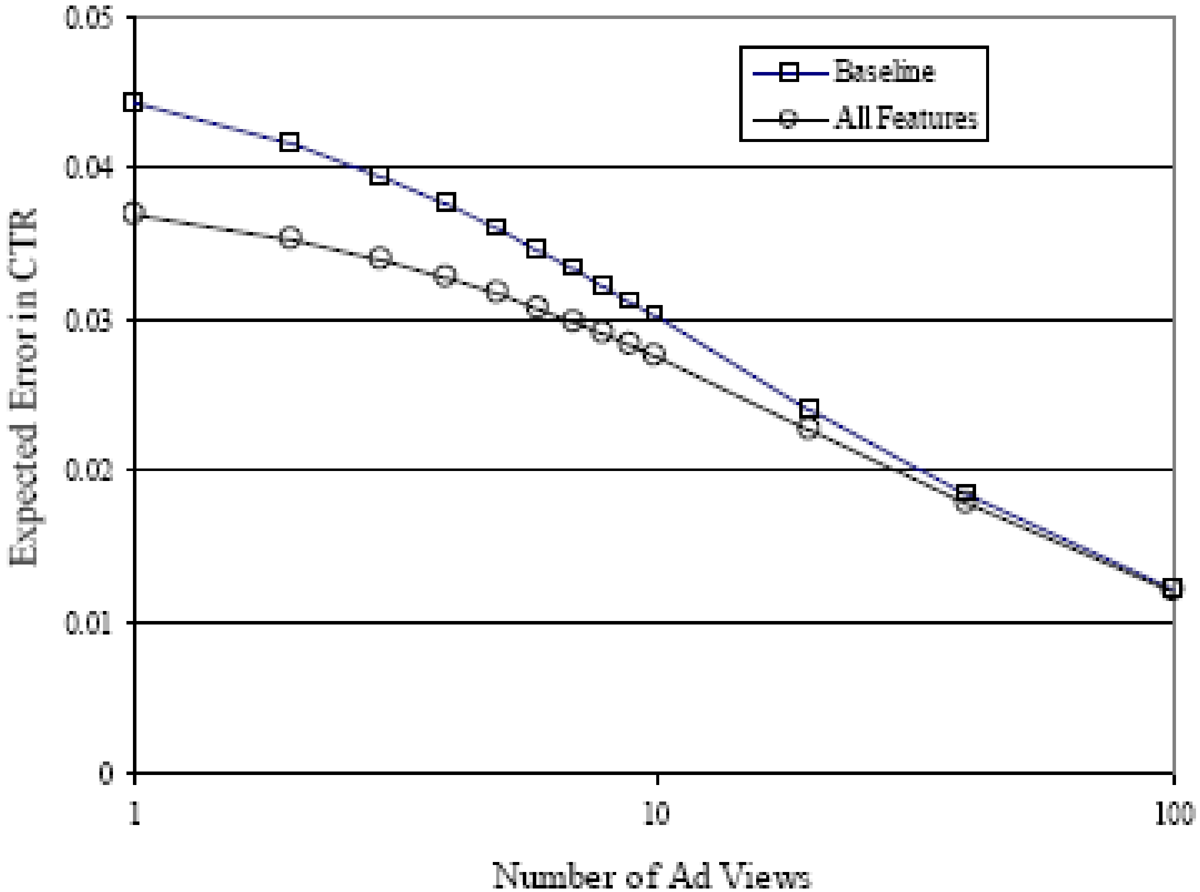


Figure 6: Expected mean absolute error in CTR as a function of the number of times an ad is viewed.

$$\hat{p} = \frac{\alpha p_0 + clicks}{\alpha + views}$$

Model predicted CTR using a Beta distribution where p_0 is the predicted (and baseline) CTR

α: strength of prior as measured in terms of impressions

Estimating CTRs using ML

Intermediate Conclusions

- **Richardson et al. report a very interesting approach and case study**
  - Despite realistic problem setting results are preliminary
- **Transparency of model**
- **Using many features helps insulate from adversarial attacks (can be useful in adversarial detection)**
- **Applied to new ads but could be extended to deal with existing ads, display/graphical ads**
  - Homework!!
- **But many issues remain!!**

Ad Quality Outline

- Motivation
- Machine Learning Overview
- IR as a means of Ranking
- Learning to Rank (LETOR)
- Online Learning
- Open issues

Modeling CTR Challenges

- **Extremely rare events (Typical CTRs < 1% for contextual)**
- **Biased dataset (the rich get richer; suboptimal locking)**
- **Very sparse (only a small percentage of <TP, Ad> get impressions; can impede generalization)**
  - Missed opportunities
- **Accuracy of estimates**
  - ML approaches are hugely biased; bias correction [see Provost and Domingos; Platt]
- **Scale and Speed**
- **Non-Stationary, new ads, changes in network**
- **Marginalization versus segmentation (resolution vs. sufficient data)**
- **….**

Other Challenges in the CPC world

- **Define a portfolio of keywords and bid prices**
  - Number of keywords per campaign per month is increasing
  - 9,100 keywords in 9/2004 to 14,700 in 3/2005
- **Manage keyword portfolio**
- **Impression Fraud, Click Fraud**

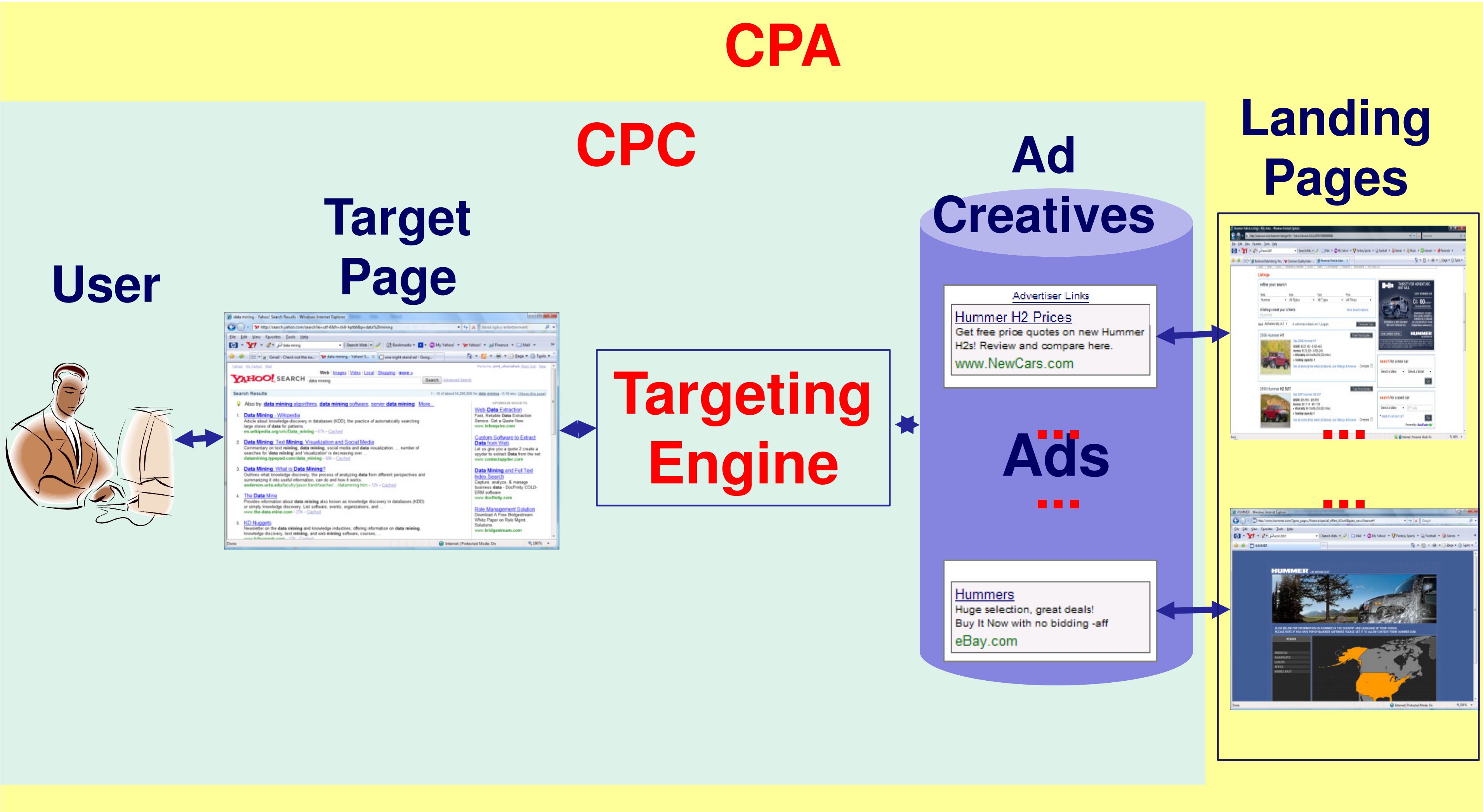
CPA versus CPC

$$CPC\_ECPM_{Ad} = CTR_{Ad} * Bid_{Ad}$$

$$CPA\_ECPM_{Ad} = CTR_{Ad} * \boxed{AR_{Ad}} * Bid_{Ad}$$

- **Similar estimation challenges (but on even rarer data)**
- **Non-keyword –based networks (term extraction)**
- ***www.Turn.com*** **is a trail blazer in automatically targeted CPA marketplaces**
- **Google is currently offering CPA in beta-mode (targeting done by publisher or automatically)**

CPC versus CPA (e.g., Turn Inc)



Outline

- **Introduction**
- **Online advertising background**
- **Business models**
- **Creating an online ad campaign**
- **Technology and Economics**
  - Advertisers (optimizing ROI thru ads and ad placement)
  - Publishers (optimizing revenue and consumer satisfaction)
    - Forward Markets
    - Spot Markets
      - Background
      - Auction Systems, Game Theory
      - Ad Quality
      - Budgeting
- **New Directions**
- **Challenges in online advertising**
- **Summary**

ECPM with Budget Constraints

- **Advertises impose budgets on money spent**
- **Allocate ads greedily (consume budget of highest bidder)**
  - ***Not optimal***
  - Worst case scenario: 50% of Optimal
- **Introduce a throttle factor to slow down budget exhaustion**

$$CPC\_ECPM_{Ad} = CTR_{Ad} * Bid_{Ad} * ThrottleFactor$$

[AdWords and Generalized Online Matching, Aranyak Mehta, Amin Saberi, Umesh Vazirani, Vijay Vazirani, Journal of the ACM, vol. 54, no. 5 (2007)]

ECPM with Budget Constraints

- **Advertises impose budgets on money spent**
- **Allocate ads greedily (consume budget of highest bidder)**
  - ***Not optimal***
  - Worst case scenario: 50% of Optimal
- **Introduce a throttle factor to slow down budget exhaustion**
  - Throttle algorithm: 1 – 1/e ≈ 0.63 competitive (63% competitive); theoretical result [Mehta et al.]

$$CPC\_ECPM_{Ad} = CTR_{Ad} * Bid_{Ad} * ThrottleFactor$$

$$ThrottleFactor = 1 - e^{-(1 - fractionOfBudgetSpent)}$$

[AdWords and Generalized Online Matching, Aranyak Mehta, Amin Saberi, Umesh Vazirani, Vijay Vazirani, Journal of the ACM, vol. 54, no. 5 (2007)]

Technology + Economics Summary

- **Forward Markets vs Spot Markets**
- **Allocations via operations research**
- **Advertising is not search!**
  - Relevance == Topicality|Usefulness? (hint $$)
  - Ranking versus prediction
  - Intellectually diverse discipline (Economics, IR, ML, Stats, Social)
- **Key directions**
  - Accurately estimating CTR/AR; auction systems; new learning algorithms; targeting graphical ads
  - Auction mechanisms
  - Global optimization algorithms

Technology + Economics Summary

- **Forward Markets vs Spot Markets**
- **Allocations via operations research**
- **Advertising is not search!**
  - Relevance == Topicality|Usefulness? (hint $$)
  - Ranking versus prediction
  - Intellectually diverse discipline (Economics, IR, ML, Stats, Social)
- **Key directions**
  - Accurately estimating CTR/AR; auction systems; new learning algorithms; targeting graphical ads
  - Auction mechanisms
  - Global optimization algorithms

Outline

- **Introduction**
- **Online advertising background**
- **Business models**
- **Creating an online ad campaign**
- **Technology and Economics**
- **New Directions**
  - Behavioral Targeting
  - Mobile
  - Web 2.0
  - Social Advertising
  - Data Exchanges
- **Challenges in online advertising**
- **Summary**

New Directions

- **Behavioral Targeting**
  - CTRs >> E.g., 11% to 840% higher than average
  - $8 Billion US Display ad market
- **Mobile**
  - By 2012, the number of devices accessing the Internet will double to more than three billion, half of which will be mobile devices [IDC, 2008]
- **Web 2.0**
  - E.g., Video: Research firm eMarketer predicts online video advertising will grow 45% this year to $850 million.
  - Freeium models are more actively being explored
- **Social Advertising**
  - leverages historically "offline" dynamics, such as peer-pressure, recommendations, and other forms of social influence
- **Data Exchanges**
  - Mining and auctioning audiences from user transactions and browsing at partner sites (ecommerce or otherwise)

Behavioral Targeting: Modeling The User

- **Target ads based on user's browsing behavior**
  - Behavioral targeting uses information collected on an individual's web-browsing behavior, such as the pages they have visited or the searches they have made, to select which advertisements to display to that individual.
  - Helps them deliver their online advertisements to the users who are most likely to be influenced by them.
- **Commonly used by:**
  - Website owners (e.g., e-commerce websites)
  - Ad networks
- **Key players include:**
  - E-commerce websites such as Amazon
  - Blue Lithium (acq by Yahoo!, $300M), Tacoda (acq by AOL, $275M), Burst, Phorm and Revenue Science, Turn.com, and others…

[For more background see: http://en.wikipedia.org/wiki/Behavioral_targeting]

Ad Network Behavioral Targeting

- **Selling Audiences (and not sites)**
  - Traditionally did this based on panels (user surveys or using Comscore/NetRatings); very broad and not very accurate
  - Through a combination of cookies and log analysis BT enables very specific segmentation
- **BT based on:**
  - Website/web page visited
    - E.g., Users who frequent Orbitz.com and Expedia.com or the travel section of USAToday.com would become part of the “Travel Shoppers” segment. Then, these users are re-targeted when they are found on other more general content type sites
  - Keyword profile (using recent searches or content that was read)
  - Retarget past visitors to your website
    - The goal is to use BT to let you send the appropriate message to each user based on where they are in your product sales cycle.

BT Extends Targeting Variables

- **Behavioral variables, e.g., :**
  - Client site visits/conversions
  - Email/newsletter list membership
  - User interest category
- **Demographic/psychographic variables:**
  - Geography
  - Prizm™ Cluster
  - Day part/day of week
  - Gender
  - At work/at home
- **Other variables:**
  - Connection speed
  - Browser/OS
  - ISP/domain”
  - *And many others (that we have looked at already)*

BT Technology

- **Segmenting users?**
  - Clustering, data mining, classification
  - Rule-based system, hybrid systems
  - Segmenting publisher real estate into to categories
- **Collaborative filtering**
  - People who bought this also bought X…
- **Collaborative filtering algorithms**
  - Yehuda Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model. KDD 2008 426-434
  - Remember NetFlix problem??
  - Gradient descent etc…
- **Very active area!!!**

BT is a happening place

- **$11B M&A of real estate and behavioral targeting**
  - 2007 Yahoo acquires RightMedia and BlueLithium (1$B)
  - Google acquires DoubleClick ($3B)
  - Microsoft Acquires Aquantive ($6B)
  - AOL acquires Quigo ($340M), Tacoda (~$300M)
- **Google's Interest-based Advertising announced [3/2009]**
- **Yahoo! SmartAds (dynamically assemble ads)**
  - Behavioral, demographic and geographic targeting
  - For example, if a user is browsing **for hybrid cars in Yahoo! Autos** and has selected **San Francisco as their default location in Yahoo! Weather**, Yahoo!'s SmartAds platform can assemble and deliver a display ad in real time that showcases a hybrid vehicle from a major auto brand, as well as local dealer information and current lease rates.

WWW 2009 Paper: Microsoft BT Study

- **Behavioral Targeting (BT) attempts to deliver the most relevant advertisements to the most interested audiences, and is playing an increasingly important role in online advertising market. However, there have been not any public works investigating on how much the BT can truly help online advertising in commercial search engines? To answer this question, in this paper we provide an empirical study on the ads click-through log collected from a commercial search engine. From the comprehensively experimental results on the sponsored search log of a commercial search engine over a period of seven days, we can draw three**
- **Conclusions:**
  - Users who clicked the same ad will truly have similar behaviors on the Web;
  - The Click-Through Rate (CTR) of an ad can be averagely improved as high as 670% by properly segmenting users for behavioral targeted advertising;
  - Using the short term user behaviors to represent users is more effective than using the long term user behaviors for BT.
  - Represented users in terms of a search keyword profile ad URL profile over one day or one week

Good disciplined study

[How much the Behavioral Targeting can Help Online Advertising? Jun Yan, et al., WWW 2009]

Behavioral KDD 2009

- **Large-Scale Behavioral Targeting - Ye Chen* Yahoo! Labs; Dmitry Pavlov Yahoo! Labs; John Canny Computer Science Division University of California Berkeley**
  - **Best Application Paper Award Winner**
  - Describe a highly scalable and efficient solution to BT using Hadoop MapReduce framework.
  - Parallel algorithm that builds 450 BT-category models from the entire Yahoo's user base
  - Describe approach has yielded 20% CTR lift over the existing production system by leveraging a well-grounded probabilistic model fitted from a much larger training dataset.

NebuAd suffers a public outcry

- **NebuAd (Silicon Valley) was one of several companies developing behavioral targeting advertising systems, seeking deals with ISPs to enable them to analyse customer's websurfing habits in order to provide them with more relevant, micro-targeted advertising.**
- **Due to fallout following (U.S.) public and Congressional concern, NebuAd's largest customers have all pulled out.**

Privacy

- Behavioral targeting agencies claim that their actions are legal, due to the privacy policy information located on websites. Most of them contain links that allow consumers to opt out of tracking.
- BlueKai lets consumers opt out of having their data traded online, as well as letting you "see exactly what data is known about you by marketers" through its BlueKai Registry program. Other networks allow similar capabilities
- A common complaint is the lack of or not-so-easy-to-access information

[See http://digg.com/d1o46k]

In the US

- http://digg.com/d1mVck (Techrepublic.com)
- Self-regulation
- The FTC's (governing body) response (somewhat aligned with Eric Schmidt's) is to let the market regulate itself, but with new and stronger guide lines. I penned an article titled, "Behavioral targeting: FTC still prefers self-regulation" that explains the new guidelines.
- http://www.ftc.gov/privacy/ (Federal Trade Commission)

Ad Matching Opt-Out

When possible, we try to match the ads that we show you to your interests. We call this "ad matching," (sometimes also called ad customization) and we do this in order to make the ads you see more relevant and useful for you. Ad matching uses data about your visits to both Yahoo! and our partner sites and about the ads you view and click.

We offer you the choice to opt-out of ad matching. If you opt-out the ads we show you will be more general and less relevant to you. Opting-out will not turn off ads.

We also use web beacons for a number of purposes including research and reporting for some of our clients. If you opt-out out of ad matching, Yahoo! also will not use the information gathered through our web beacons for research and reporting purposes.

You are currently not opted out

Opt-Out

If your status has not changed after selecting Opt-Out, your browser may be set to block Yahoo! Cookies. Please unblock Yahoo! Cookies and try again.

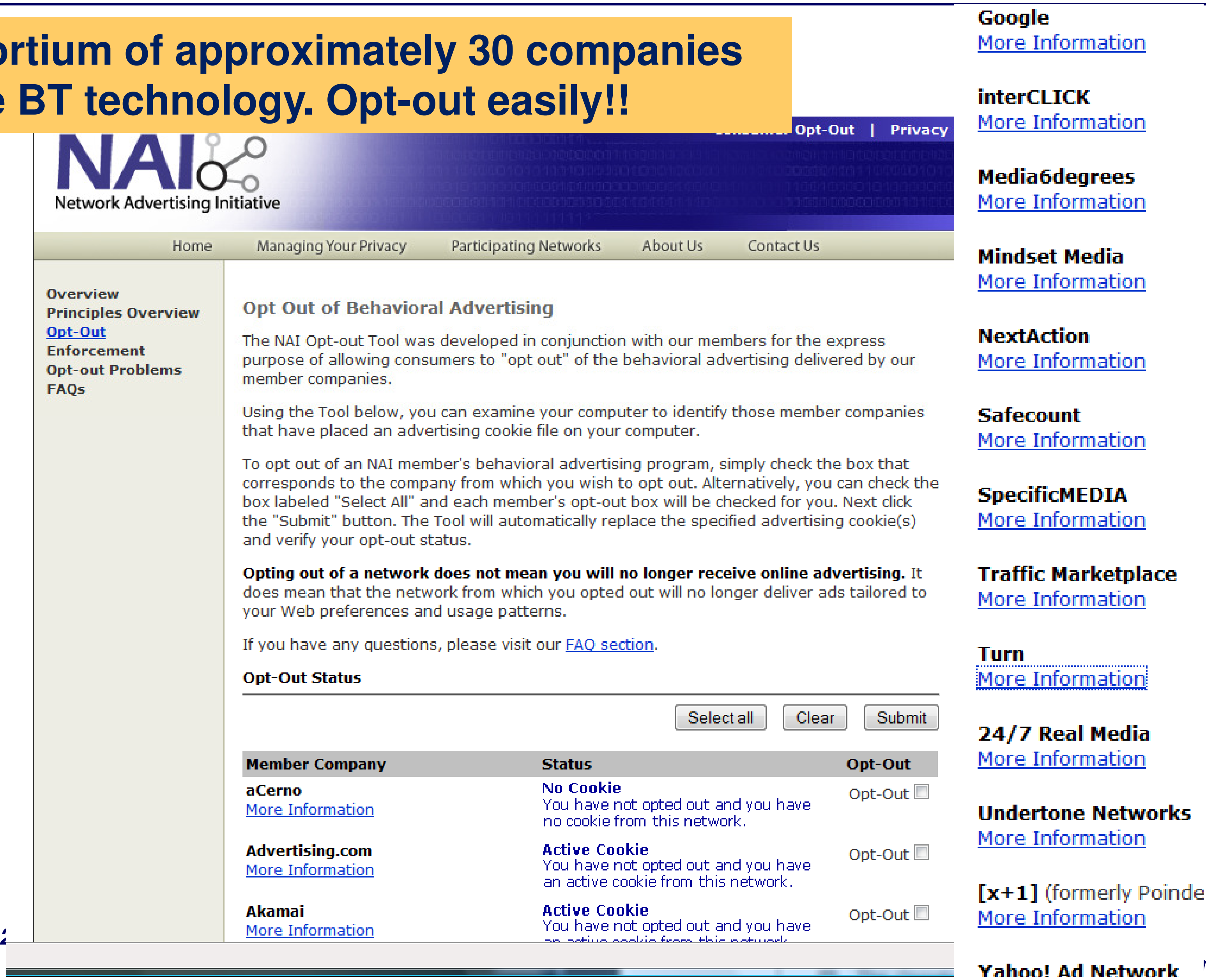
Note: This opt-out applies to a specific browser rather than a specific user. Therefore you will have to opt-out separately from each computer or browser that you use.

Yahoo! is a participating member of the Network Advertising Initiative. You can exercise the same Yahoo! opt-out on the NAI site and also manage your other ad network opt-outs.

To see which ad networks serve ads on Yahoo!, you can visit our page about Third Party and Affiliate Cookies on Yahoo!.

The Network Advertising Initiative (NAI)

A consortium of approximately 30 companies that use BT technology. Opt-out easily!!



Privacy Issues

PLATFORM

Home :: Advertiser Solutions :: Audience Targeting :: Behavioral Targeting :: Consumer Notice

ADVERTISER SOLUTIONS

- Affiliate
- Audience Targeting
  - Platform Targeting
  - Behavioral Targeting
    - Audience Behaviors
    - Consumer Notice
    - Privacy Policy
  - Custom Targeting
  - Consumer Profiles
- Behavioral Targeting
- Creative Services
- Display
- Mobile
- Promotions
- Research
- Rich Media
- Search & Feeds
- Self Service
- Sponsored Listings
- Video
- Widgets

Consumer Notice

Making advertising relevant, while respecting your choice.

TACODA works with hundreds of existing websites to ensure that the online advertising you see is relevant and useful.

We anonymously categorize web surfing interests using a small text file in the browser called a cookie to deliver targeted advertising. You may opt-out of these targeted ads at any time.

For example:

With TACODA-enabled ads you will be more likely to receive advertising that is relevant to you. For example, if you are researching a new car you will see:

If you opt-out of TACODA, you are more likely to see advertising that is not relevant to you:

TACODA's Use of Cookies Statement

- How Does TACODA Improve Online Advertising?
  For many, the term "online advertising" conjures up images of web pages cluttered with irrelevant ads and annoying offers. TACODA believes your online experience can be better. We deliver ads that are likely to offer greater relevance by working with hundreds of name brand websites and advertisers to deliver ads based on your previous web viewing interests.
- What is a Cookie?

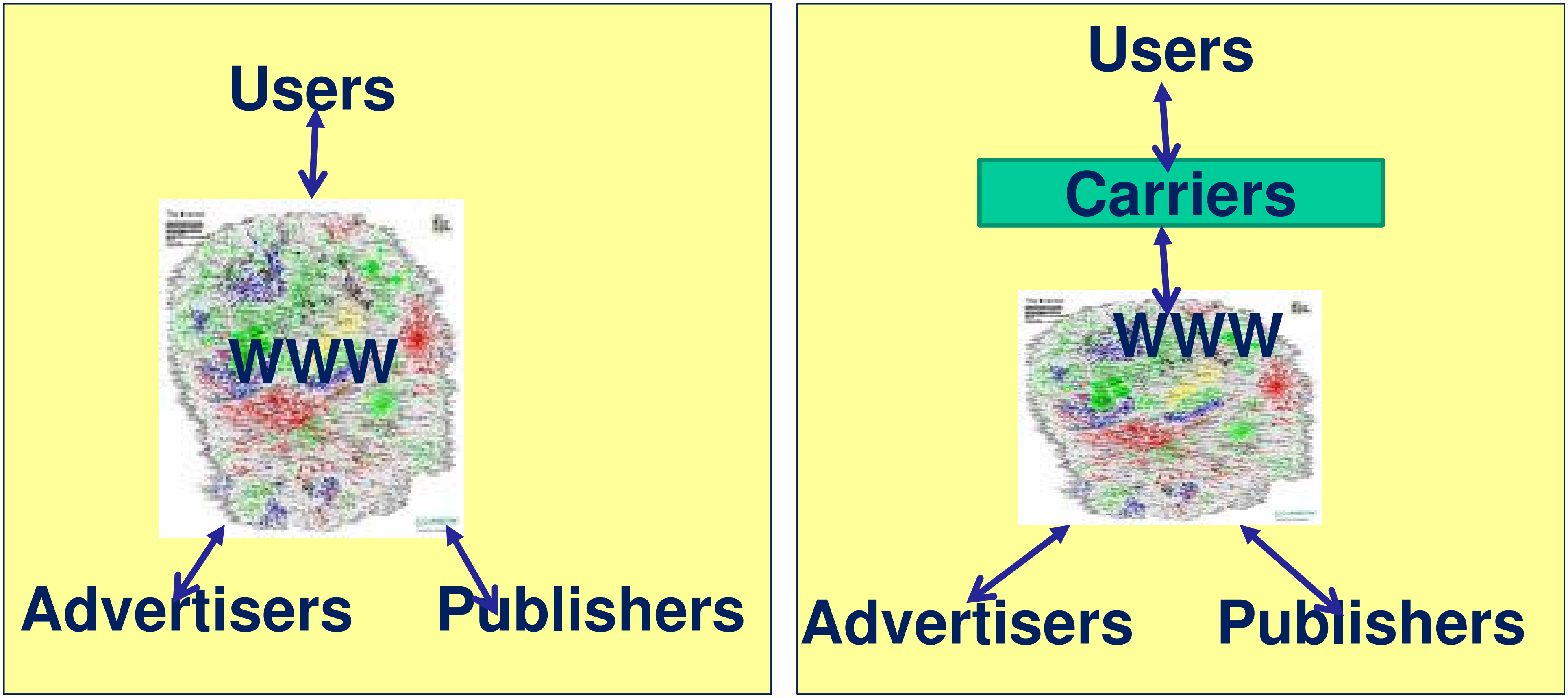
Behavioural Targeting

- Holds huge promise to monetize longtail inventory (such as Web2.0)
- Privacy concerns need to be addressed before consumers, advertisers and publishers embrace this direction in a big way

Outline

- **Introduction**
- **Online advertising background**
- **Business models**
- **Creating an online ad campaign**
- **Technology and Economics**
- **New Directions**
  – Behavioral Targeting
  – Mobile
  – Web 2.0
  – Social Advertising
  – Data Exchanges
- **Challenges in online advertising**
- **Summary**

Mobile Web vs. Internet



Online

Mobile

Tipping Points

- The true turning point for the industry was the introduction of the smartphone, heightened by Apple's iPhone launch in mid-2007.
- Tearing down the carrier wall
- The development of third-generation (3G) mobile phones led to better connection speeds
- Wi-Fi connectivity and the rise of mobile Internet browsing.
- Richer Content: Open Internet
- In developing markets, mobile is the primary gateway to the Internet not PC!

Mobile Advertising

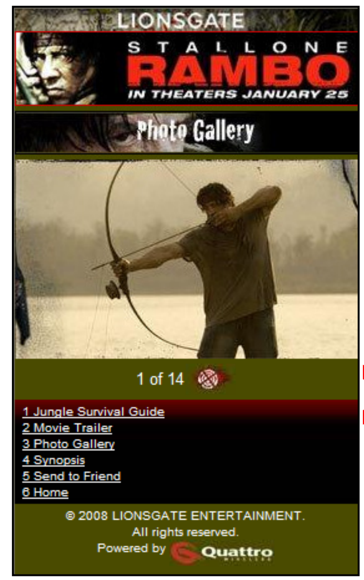
[John Gantz, IDC 2008]

- While the PC is currently the dominant means of gaining access to the Internet, IDC expects the number of mobile devices accessing the Internet will surpass the number of online PCs by 2012.
- Roughly 40 per cent of all Internet users worldwide currently have mobile Internet access. The number of mobile Internet users will reach 546 million in 2008, almost twice as many as in 2006, and is forecast to surpass 1.5 billion worldwide in 2012.
- The latest wave of post-iPhone smart phones have put mobile social networking, search, location services and even gaming and video into the fast lanes of user adoption.
- Data usage is up 39% in a year.

Branded Landing Page & Click-to-Video

Gallery/Image

Movie Trailer



Versus

CPM Banner Ad

Branded Landing Page

Mobile Monetization

- **All the internet monetization models are applicable to mobile.**
- **New monetization models that leverage unique mobile features e.g. pay per call.**

	Search	Contextual	Display
CPM			x
CPC	x	x	x
CPA	x	x	x
PPC	**x**	**x**	**x**

Select Best Ads to Display

- **Requires:**
  - Ads to select from (i.e., ads with budget remaining)
  - Ability to determine best ad to show (see below);
  - Serve ad or not (social responsibility)

- **Selecting Best Ad(s)**
  - Estimate the CTR of each ad using available features.
  - A number of techniques for estimating CTR:
    - Regularized Regression models .
    - Multi-armed bandit models
    - Non parametric nearest neighbor methods.
  - Features for estimating CTR:
    - Site characteristics
    - Ad Characteristics
    - User Characteristics

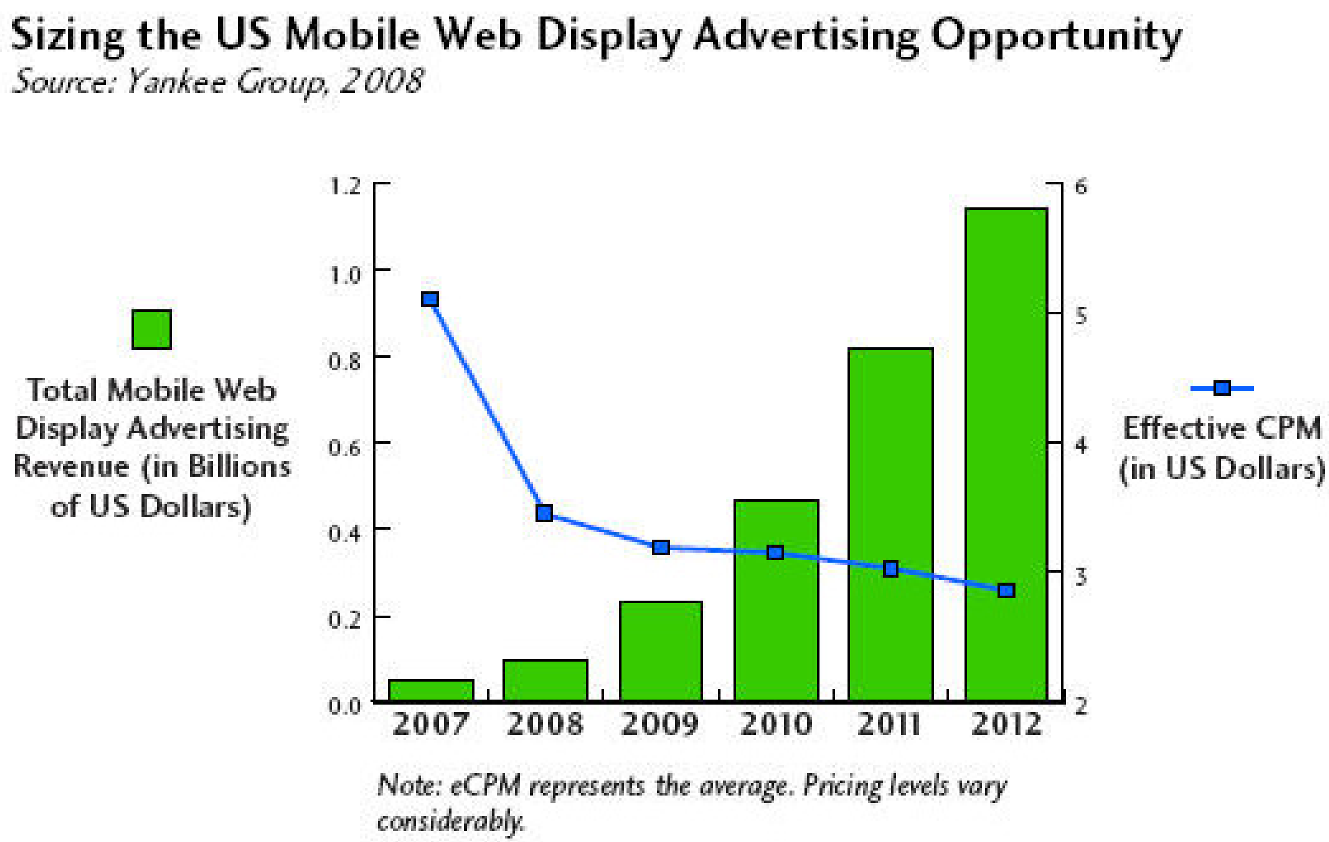
Targeting for Mobile versus Online

How to predict CTR and allocate based on the product of CTR and CPC (ECPM)?

Do not have the features available to search engines for ranking sponsored search.

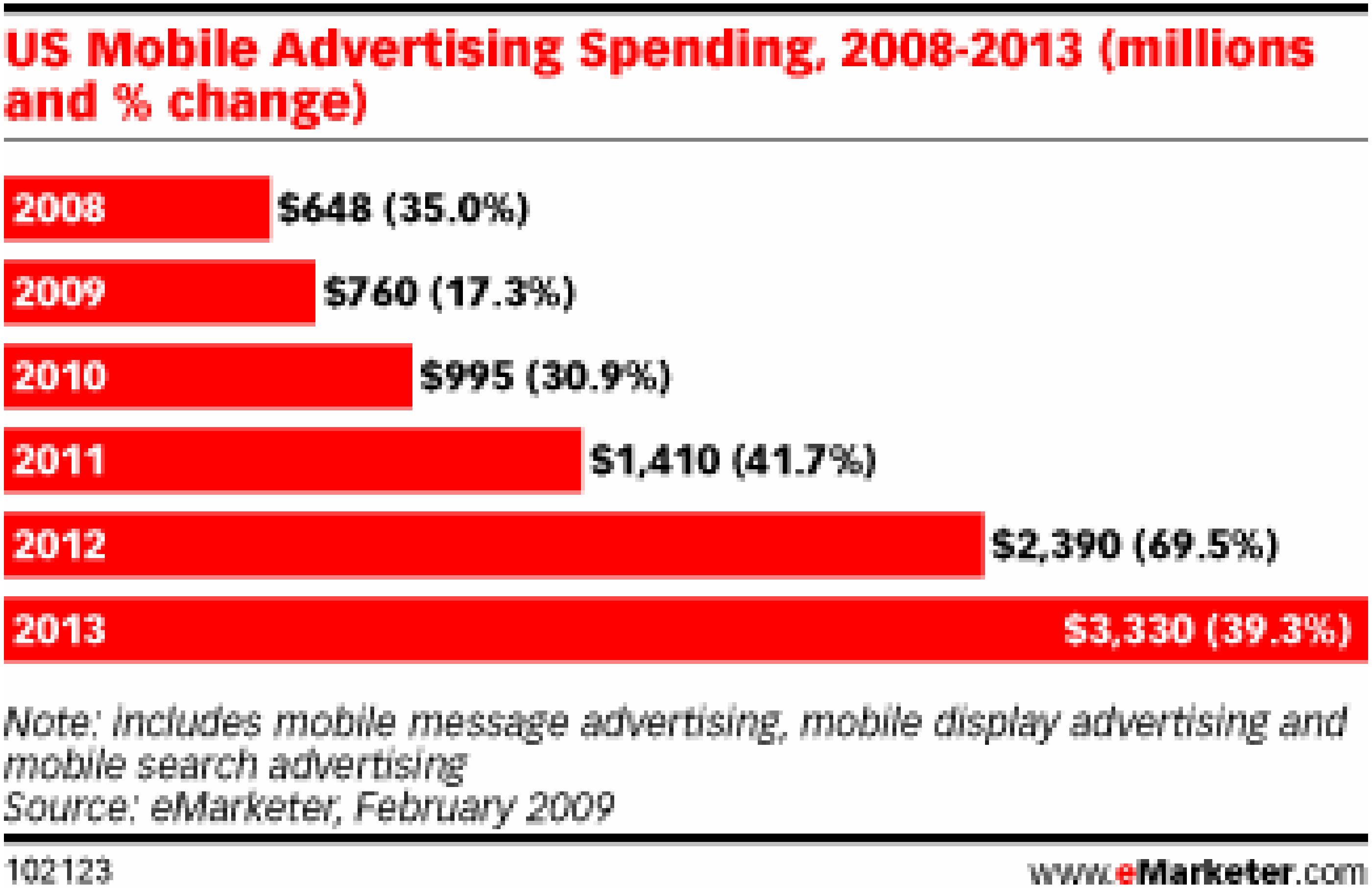
Contextual information is often limited because of lack of content.

Large Percentage of the traffic is user generated content which is difficult to monetize.
Ads are often targeted down to the carrier and device level.
e,g, Phones with video capability on Sprint!
-----But have location based info….. Treat in a e.g., collaborative way



Sizing the US Mobile Web Display Advertising Opportunity
Source: Yankee Group, 2008
Total Mobile Web Display Advertising Revenue (in Billions of US Dollars)
Effective CPM (in US Dollars)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
6
5
4
3
2
2007
2008
2009
2010
2011
2012
Note: eCPM represents the average. Pricing levels vary considerably.

Ad spend will quadruple in 4 years



Outline

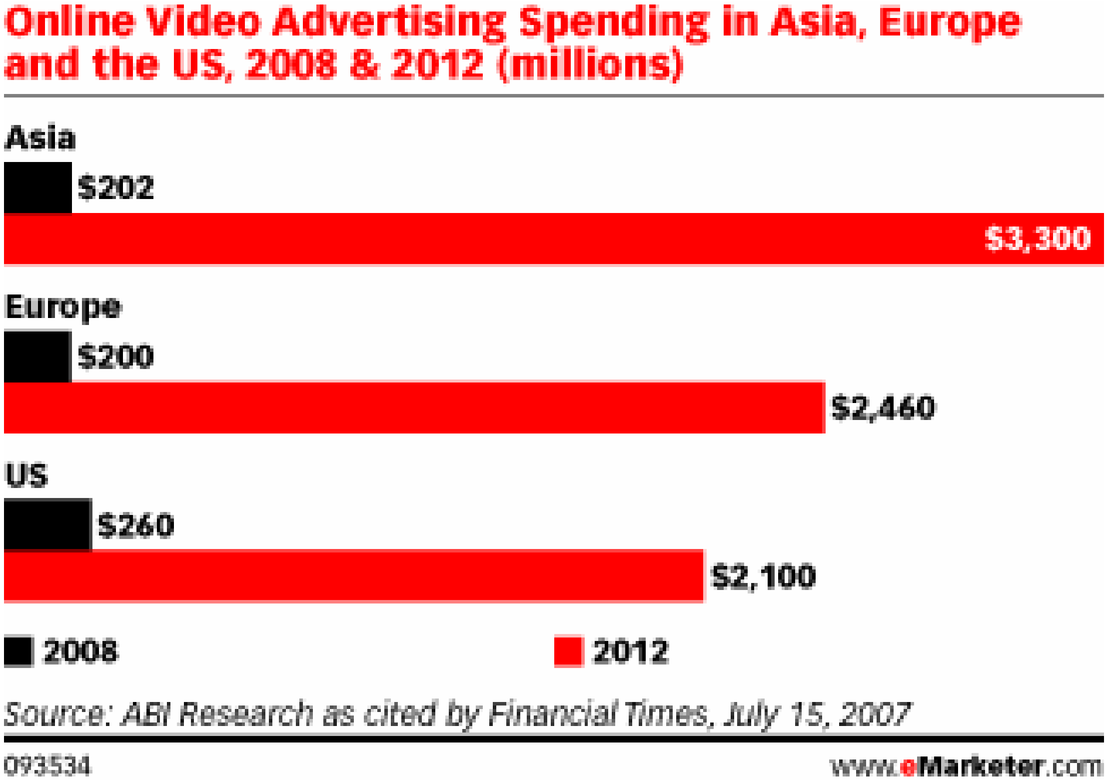
- **Introduction**
- **Online advertising background**
- **Business models**
- **Creating an online ad campaign**
- **Technology and Economics**
- **New Directions**
  - Behavioral Targeting
  - Mobile
  - Web 2.0
  - Social Advertising
  - Data Exchanges
- **Challenges in online advertising**
- **Summary**

Web 2.0

- Online reviews:
- Social networking:
- Social media sites:
- Tagging/folksonomies systems:
- Online encyclopendia:
- Online blogs:
- Online question-answer forums:
- CrowdSourcing forums:
- Cloud Computing
- Wikis
- Online Gaming

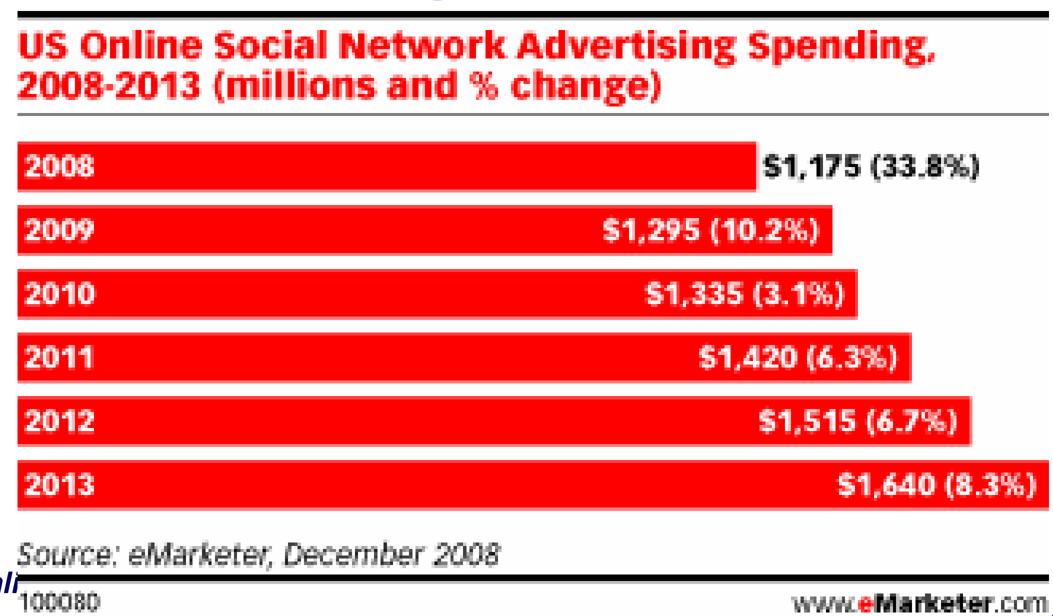
Online Video Advertising

- **Online Video Advertising report examines this nascent but rapidly growing ad channel**
  – YouTube signs deal with Hollywood for TV shows and Movies [4/2009]; Ads and possibly payment for premium content
- **YouTube will loose $470M in 2009 [Credit Suisse 2009]**



Social Network Ad Spend Cools Off

- **With the economy taking a toll on even social-networking sites, eMarketer reduced its global forecast for advertising in the sector this year to $2.3 billion, according to a WSJ (3/2009).**
  - Facebook and MySpace claim $900M revenue in US and over 50% of the world wide spend (WSJ)
  - http://blogs.wsj.com/digits/2009/03/17/social-networking-sites-to-see-slower-ad-growth/?mod=rss_WSJBlog?mod=



Web 2.0-> Web 3.0

- **Very low CPMs**
- **Taking Web 2.0 to Web 3.0?**
  - Behavioral Targeting
  - Social search or a social search engines
    - a type of web search method that determines the relevance of results by considering the interactions or contributions of users
    - How to leverage community and collaboration?
  - Social behavioral targeting?
  - Are new business models required?
  - ??

Web 2.0 Conclusions

- **Social Networks**
  - 79.5 million people—41% of the US Internet user population—visited social network sites at least once a month, an 11% increase over the previous year. By 2013, an estimated 52% of Internet users will be regular social network visitors.
  - Growth in spending on social network advertising is slowing around the world, a result of poor economic conditions and a lack of proven ad models
  - Marketers continue to show interest in social networks but increasingly are looking outside buying ads!!
    - Building fan communities on social network sites.
    - Subscriber –based; pay for premium content
    - http://www.shutterstock.com/
- **Online video -> profitability**
- **Microblogging is growing rapidly; Twitter; Mixi (Japan)**

Outline

- **Introduction**
- **Online advertising background**
- **Business models**
- **Creating an online ad campaign**
- **Technology and Economics**
- **New Directions**
  - Behavioral Targeting
  - Mobile
  - Web 2.0
  - Social Advertising
  - Data Exchanges
- **Challenges in online advertising**
- **Summary**

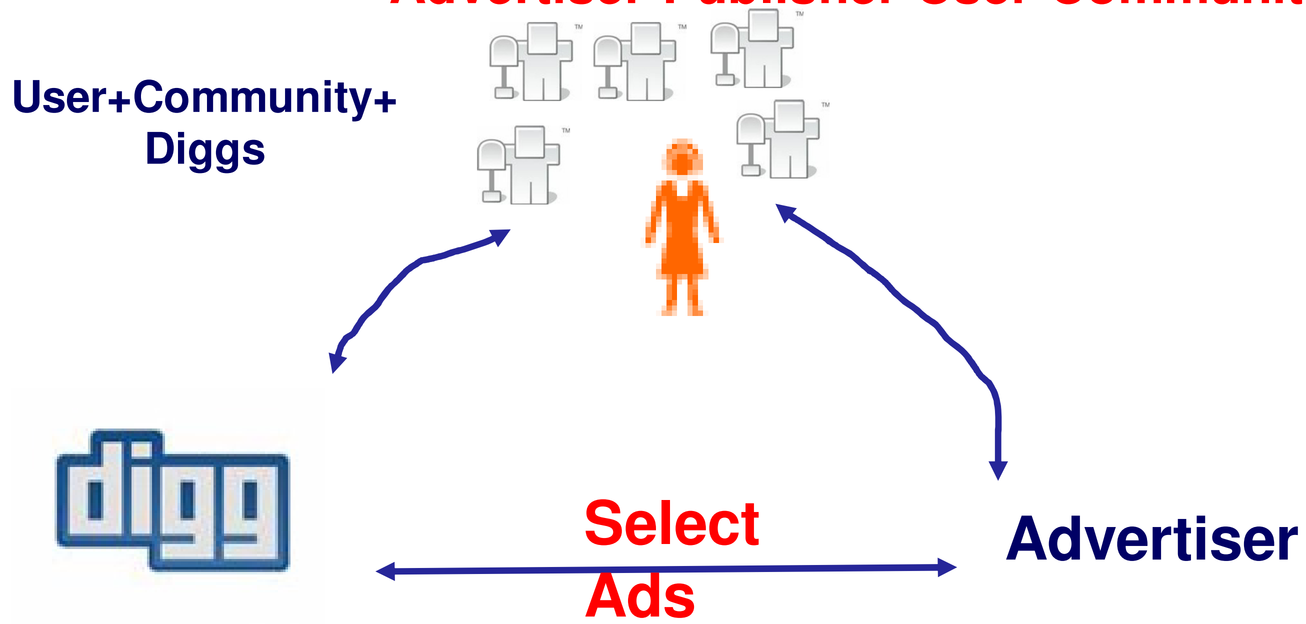
Social Advertising

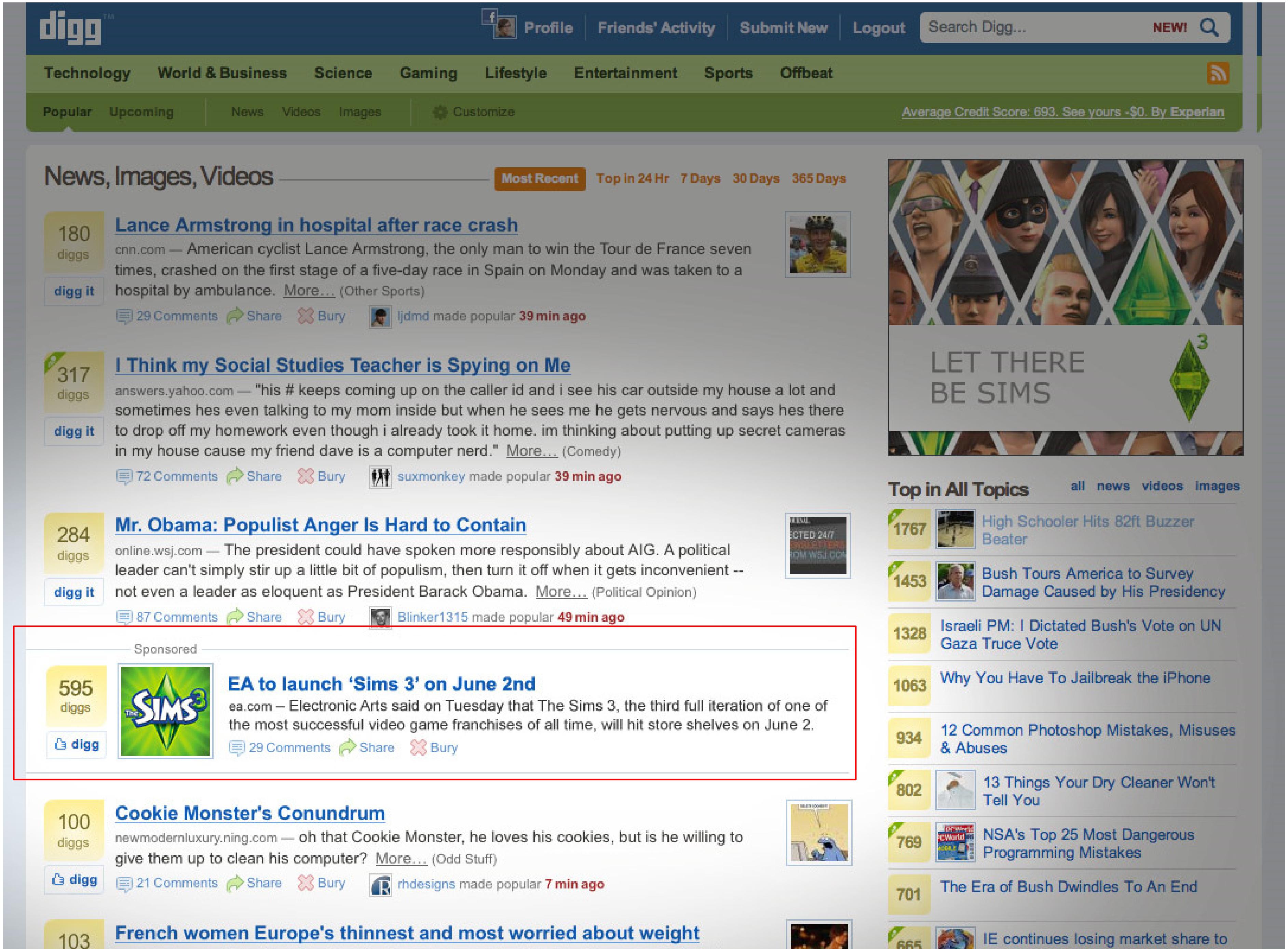
- *Social advertising* was introduced in the context of community-oriented websites such as Facebook and Digg.
- This form of advertising systematically leverages historically "offline" dynamics, such as peer-pressure, recommendations, and other forms of social influence.
- In the case of Digg, this translates to enabling users to determine what ads appear on the website news streams by voting up or "burying" ads, much the way they can digg or "bury" organic news items.

Social Advertising (at Digg)

Social Online Advertising

Advertiser-Publisher-User-Community





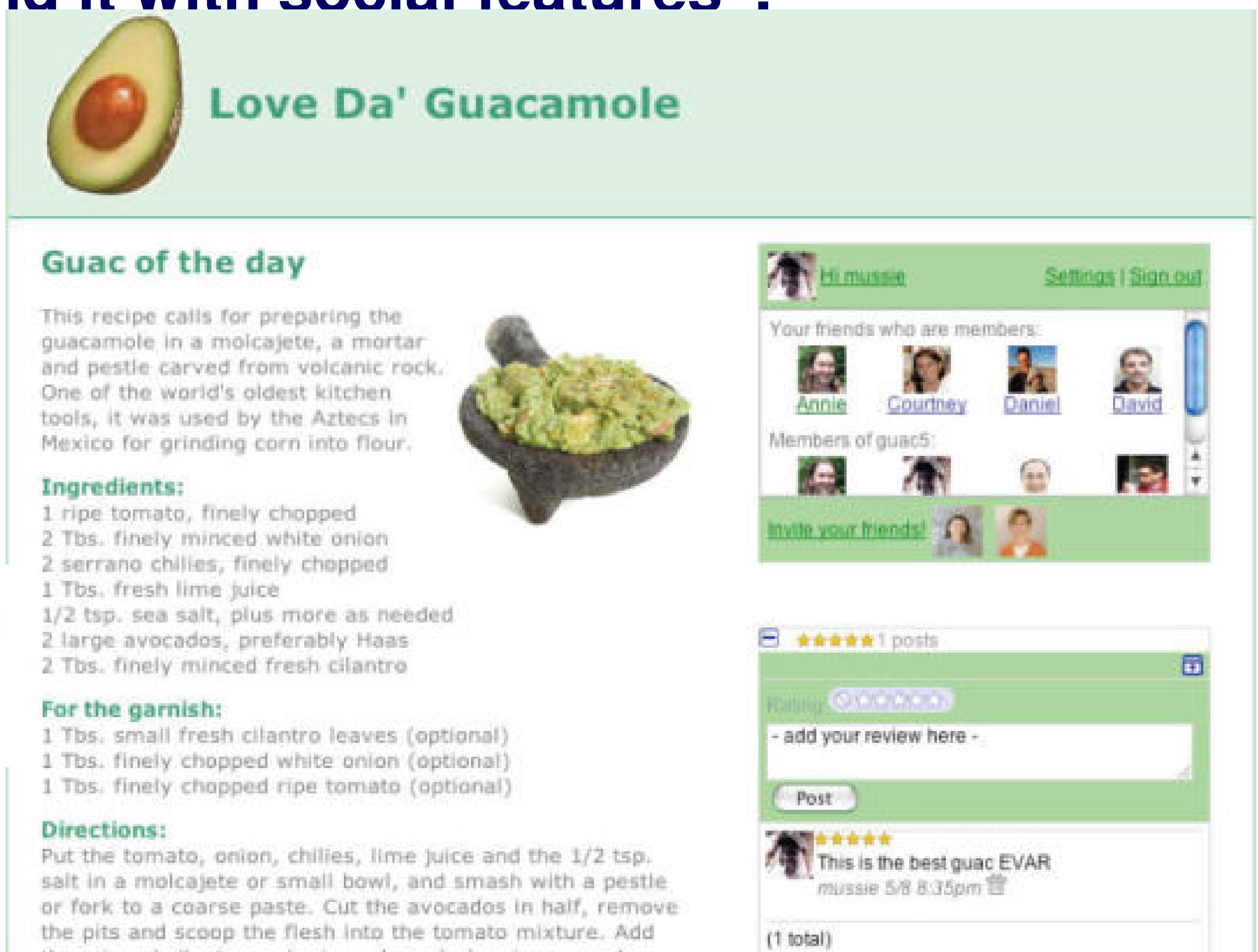
[http://blog.digg.com/]

Google Connect

- Soon websites that use Friend Connec
  a new option - add Adsense-like advertising within the Friend Connect and Open Social widgets that they've added to their websites.
- Publishers will get a percentage of the revenue generated from the advertising.
- And that's the big monetization scheme behind Open Social and Friend Connect for Google. And that's why they call it Friendsense internally.
- And occasionally let it slip to outsiders.

http://www.techcrunch.com/2008/12/26/why-googlers-call-friend-connect-friendsense/

- **“Google Friend Connect instantly awakens and strengthens the community that visits your site by enriching it with social features”.**



Google Friend Connect BETA

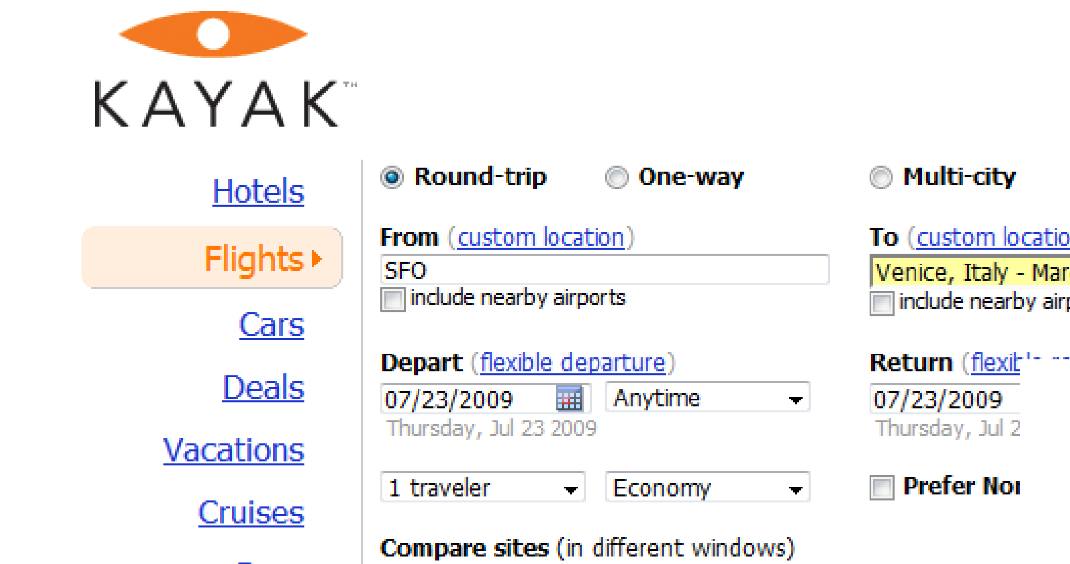
Outline

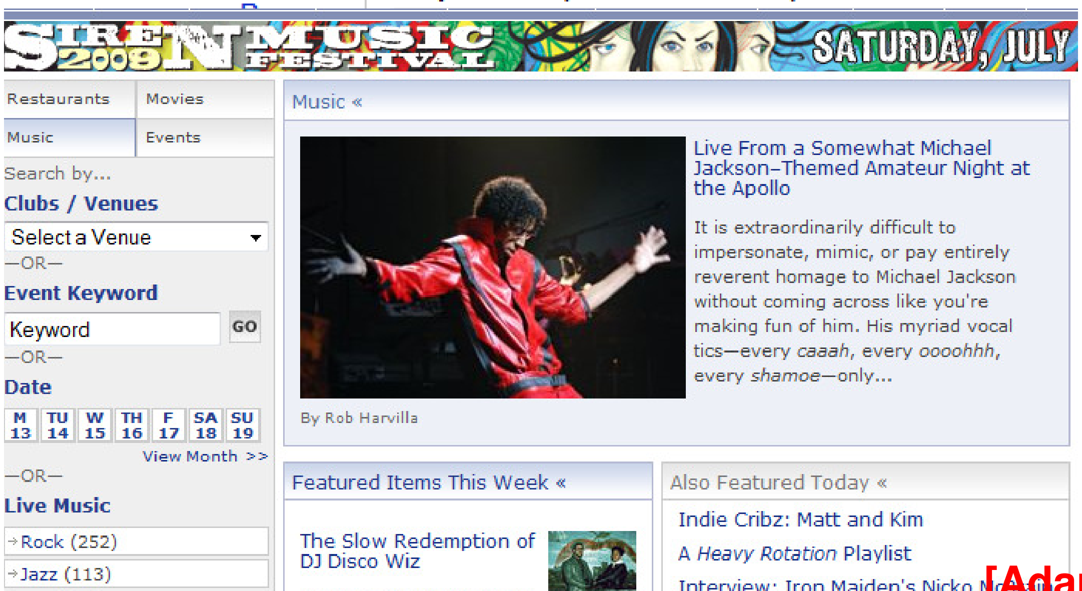
- **Introduction**
- **Online advertising background**
- **Business models**
- **Creating an online ad campaign**
- **Technology and Economics**
- **New Directions**
  - Behavioral Targeting
  - Mobile
  - Web 2.0
  - Social Advertising
  - Data Exchanges
- **Challenges in online advertising**
- **Summary**

Data Exchanges

- **Publishers get paid for their audiences and content**
- **eCommerce sites get paid for their audiences and transactions**
- **Data Exchanges**
  - Categorize users based on their behaviors
  - Auction off audiences to publishers/ad networks
  - E.g., BlueKai, eXelate and Datran (behavioral exchanges)
    - E.g., BlueKai: Intent mining from eCommerce sites
    - Lotame, SocialMedia and Media6Degrees working the social-media world.

Commercial Intent: Shop/Browse





[Adapted from Lucian Vlad Lita 2009, BlueKai]

BlueKai Commercial Intent Category Verticals

IN-MARKET AUTO:

- Autos
  - *by Make & Model*
- Class
  - *Compacts & Sub-Compacts*
  - *Convertibles*
  - *Family Sedans*
  - *Green Cars*
  - *Luxury Cars*
  - *Midsize Cars*
  - *Minivans*
  - *Pickup Trucks*
  - *Sport Utility Vehicles*
  - *Sports Cars*
  - *Station Wagons*

IN-MARKET TRAVEL:

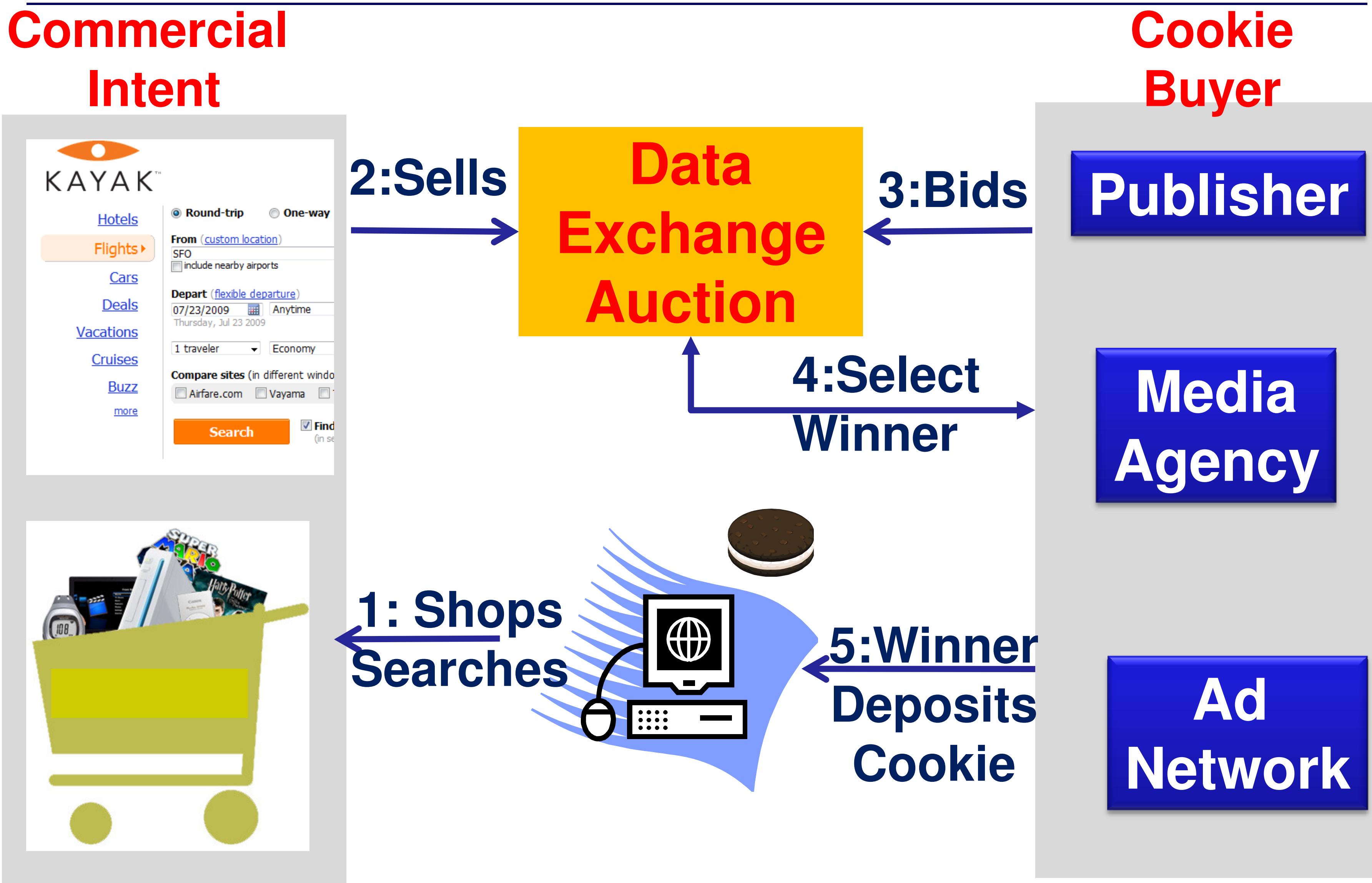
- Air Travel
  - *International Flyers*
  - *Domestic Flyers*
  - *By Destination City*
  - *By Departure City*
  - *Length of Stay*
  - *By Advanced Booking*
- Hotels & Lodging
  - *By City*
  - *By Star Rating*
  - *Length of Stay*
  - *By Advanced Booking*
- Car Rentals
  - *By City*
  - *By Rental Company*
  - *By Type of Car*
  - *By Advanced Booking*
  - *Saturday Rental*
- Cruises
  - *By Destination*
  - *By Cruise line*
  - *By Class*
  - *Duration*
  - *Departure Month*
- Packages
  - *By Domestic Destination*
  - *By International Destination*

IN-MARKET SHOPPING:

- Auto Parts & Accessories
- Bags
- Books & Magazines
- Cameras
- Cell Phones & PDA's
- Clothing & Accessories
  - *Women's > Jeans, Dresses, Shoes, Outerwear*
  - *Men's > Accessories, Shoes, Outerwear, etc*
  - *Boys*
  - *Girls*
  - *Infants & Toddlers*
  - *Jewelry & Watches, etc.*
- Computers
- Consumer Electronics
  - *Apple iPod, MP3 players*
  - *DVD and Home Theater*
  - *MP3 Accessories*
  - *Satellite, Cable TV, etc.*
- DVD & Movies
- Health
  - *Fragrance*
  - *Hair Care*
  - *Makeup*
- Home & Garden
  - *Furniture*
  - *Appliances*
  - *Kitchen, etc*
- Toys, Games, Video Games

[Lucian Vlad Lita 2009, BlueKai]

Mining Commercial Intent



Context + User Behavior


News
write to the editor | email a friend | print article | show comments (1)
News
Trash Talking with Harvey Weinstein
Juicy details plucked from the garbage of a movie mogul
By Tony Ortega
Tuesday, July 1st 2008 at 12:09pm
Getty Images
New York is so filthy with moguls and millionaires that even the trash is rich.
I mean, movie impresario Harvey Weinstein is such a big shot that he doesn't even promptly return text messages from Nicole Kidman? You can understand that Weinstein might be too busy to blow off Knicks owner Jim Dolan's request for a "follow-up"; phone call—which Harvey did for at least a few days. Who would want to talk to Dolan anyway?
But if I were Harvey, I would text the hell out of Kidman every chance I got. And I would immediately return that call from
Related Content
Related Stories ...
The Weinstein Files: The Smoking Gun Notes that 'Nicole Kidman Got Trashed'
July 3, 2008
Oscar to Sicko: The Moore The Merrier
January 25, 2008
Michael Moore Wins
Site Search «
SEARCH
Advanced Archive Search >>
Weekly
E-Mail Blasts
Weekly
Music
Promotions
Offstage Voice
Dining
NY Movie Club
Events
Subscribe
E-Mail Address
Go
EMBASSY SUITES HOTELS
STAY 3 NIGHTS AT EMBASSY SUITES AND WIN $50K FOR COLLEGE.
LEARN MORE
Viewed
Commented
Emailed
Michael Jackson: The Man in Our Mirror
Black America's eulogies for the King of Pop also let us resurrect his best self

Context

VERSUS

Context + Behavior Class

CPM = ~$1-$2

[http://www.nytimes.com/2009/03/26/business/media/26adco.html?_r=1&scp=1&sq=bluekai&st=cse]

CPM = ~$4-15

How does the data exchange works?

- **When someone does a search, for example, on Kayak.com for first-class flights to Paris in September, that information can be captured by a cookie, and Kayak.com can sell that cookie using eXelate or BlueKai.**
- **More targeted**
  - A buyer would want that cookie so the company could cut down on wasted ads. Sure, Hilton could blanket sites with its ads, but it would rather show an ad to someone who has searched for flights to Paris recently.
  - "When you see a cookie on that user, you can show them travel to Paris even though they may be on MySpace or The New York Times," said Mark Zagorski, the chief revenue officer of eXelate.
- **Who buys this data?:**
  - the Hilton itself or its media agency may buy the information.
  - A publisher may buy data so it can sell more expensive ads to Parisian hoteliers and restaurants.
  - Or the buyer may be an ad network — a company that handles sales for big groups of sites at once.

Outline

- **Introduction**
- **Online advertising background**
- **Business models**
- **Creating an online ad campaign**
- **Technology and Economics**
- **New Directions**
  - Behavioral Targeting
  - Mobile
  - Web 2.0
- **Challenges in online advertising**
- **Summary**

Challenges in Online Advertising

- **Privacy**
- **Modeling the user**
- **Ad placement**
- **Social responsibility – Long term user value**
- **Deception (e.g., phishing)**
- **Green advertising**
  - 67% of printing magazines are returned [Sierra Club, AdTech 2008]; inflated subscription rates to fulfill advertiser commitments.
- **Fraud**
  - Impression Fraud
  - Click fraud

Privacy Concerns (US view)

- "the data game right now is a little bit of the Wild West."
- All this tracking has raised privacy concerns. Some privacy advocates have asked Congress and the Federal Trade Commission to investigate the issue, seeking clear policies about sensitive data, more information on the way companies are tracking consumers and options for consumers to avoid online tracking.
- So far, the commission has recommended that the industry police itself.
- But Jon Leibowitz, one of the commissioners, warned in February that the industry needed to do a better job or face new laws and regulations.

[http://www.nytimes.com/2009/05/31/business/media/31ad.html?_r=2&em=&pagewanted=print]

Data Exchanges and Privacy

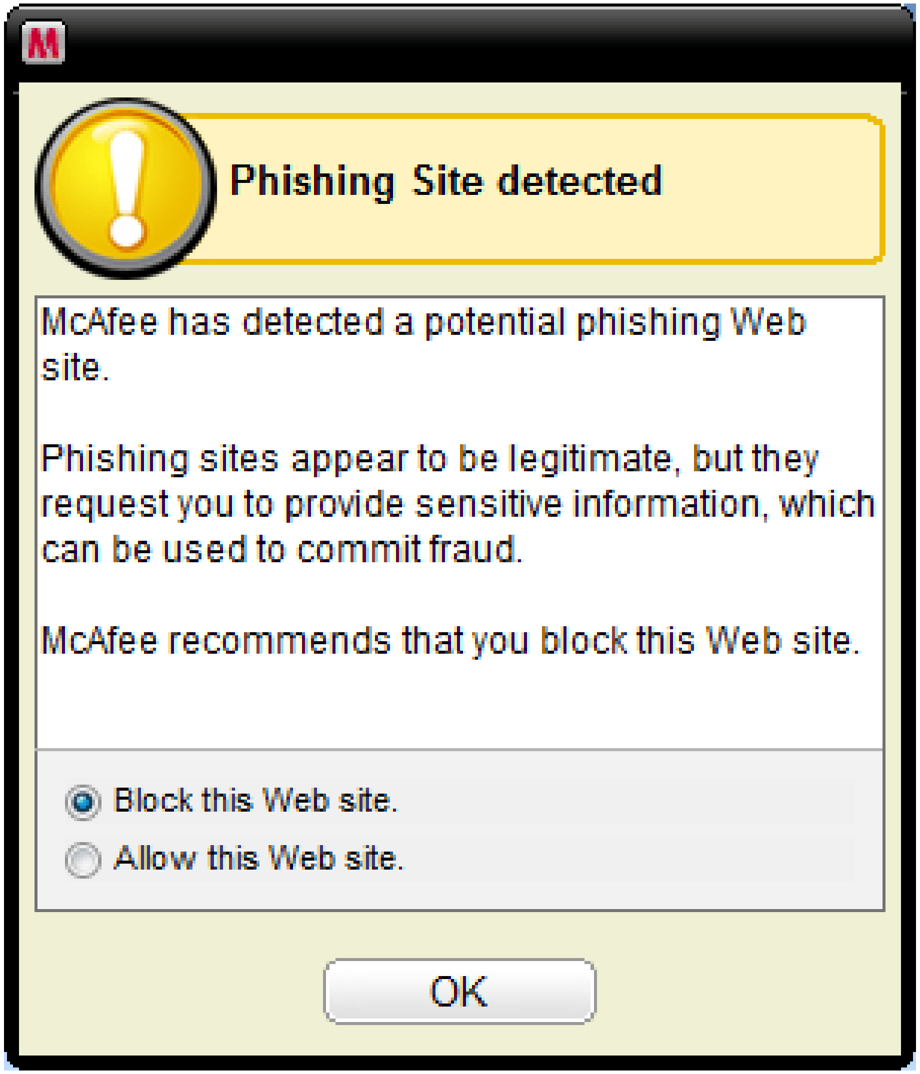
- **Both BlueKai and eXelate**
  - provide a page where consumers can refuse all targeting,
  - they are allowing consumers to see what information has been collected about them,
    - exelate.com/new/consumers-optoutpreferencemanager.html for eXelate,
    - and tags.bluekai.com/registry for BlueKai.

Fraud on Internet

- **Fraud is prevalent on the internet today**
  - Publisher side (e.g., domain-name parking, Internet pirates, **domain name bandits,** cybersquatters,typosquatters)
  - Advertiser side (both click and impression fraud)
- **Difficult to quantify**
  - Studies/organizations will claim [5, 40]% clicks are fraudulent
- **CPA Advertising model directly address click fraud**

A Bitter-Sweet Advertising Moment!

Deception


Phishing Site detected
McAfee has detected a potential phishing Web site.
Phishing sites appear to be legitimate, but they request you to provide sensitive information, which can be used to commit fraud.
McAfee recommends that you block this Web site.
Block this Web site.
Allow this Web site.
OK

Click Fraud versus SPAM

- $21.4 billion in U.S. advertising spending, according to an Oct. 16 report by research firm eMarketer. (Google has 32% of the market)
- Average click fraud rate of PPC advertisements on search engine content networks, including Google AdSense and the Yahoo Publisher Network, was 28.1% for Q3/2007.
- The California legislature found that spam cost United States organizations alone more than $13 billion in 2007, including lost productivity and the additional equipment, software, and manpower needed to combat the problem.[

Click Fraud

- **Click fraud is a type of internet crime that defrauds advertisers**
- **The act of clicking on an ad (e.g., text or banner) with the intention of either:**
  - To inflate a Publisher's revenue
  - Or manipulating a competitor's advertising behavior by artificially decreasing their ROI or depleting their budget
- **Committed by a person, or by a computer program that imitates a legitimate user of a web browser**
- **Difficult to quantify**
  - Studies/organizations will claim [10, 20]% of clicks are fraudulent
  - Dizzying collection of scams and deceptions that inflate advertising bills for thousands of organizations of all sizes

Conferences/Workshops

- **WWW 2008: WS6 - Targeting and Ranking for Online Advertising**
  - Ewa Dominowska and Vanja Josifovski
- **SIGIR 2008: Information Retrieval in Advertising**
  - Ewa Dominowska, Eugene Agichtein, James G. Shanahan, Evgeniy Gabrilovich
- **DIMACS Workshop on the Boundary between Economic Theory and Computer Science, 10/2007**
- **Click fraud workshop at Stanford (9/2007)**
- **KDD Workshop on Advertising**
  - Focused on the online and offline advertising worlds
  - Cross fertilization of both worlds (digital marketplaces, audience understanding)
- **Many workshops and conference session on online advertising**
  - See WWW, SIGIR, KDD, CIKM
  - Duplicate detection in click streams, http://www2005.org/cdrom/docs/p12.pdf

Outline

- Introduction
- Online advertising background
- Business models
- Creating an online ad campaign
- Technology and Economics
- New Directions
- Challenges in online advertising
- Summary

Business, Gold rush

Tech

Hot Areas

Bridging the gap: DIY Software+Data

- **OpenX**
  - OpenX Ad Server is a powerful open source platform that manages advertising for more than 150,000 websites in over 100 countries. More than 300 billion ads run through OpenX Ad Servers each month. And since it's available as a software download or as a hosted service, you can decide to run it yourself or let us take care of managing the infrastructure for you.
- **Solr (*Solr* is an open source enterprise search server based on the Lucene Java search library, with XML/HTTP and JSON APIs, hit highlighting, faceted search, ...)**
  - http://www.pdf-search-engine.com/solr-pdf.html
- **Data:**
  - Simulators; Mechanical Turk; Collaborations with advertising industry

R: open-source statistical software

- R is a free, open-source software package/programming language for statistical computing.
  - (commercial version is S-plus)
- **Resources on R [adapted from *http://www.stat.cmu.edu/~cshalizi/350/*]**
  - The official intro, "An Introduction to R", available online in HTML , PDF
  - John Verzani, "simpleR", in PDF
  - Quick-R. This is primarily aimed at those who already know a commercial statistics package like SAS, SPSS or Stata, but it's very clear and well-organized, and others may find it useful as well.
  - Patrick Burns, The R Inferno. "If you are using R and you think you're in hell, this is a map for you."
  - T. Lumley, "R Fundamentals and Programming Techniques" (large PDF)
- **Books: [adapted from *http://www.stat.cmu.edu/~cshalizi/350/*]**
  - Braun and Murdoch's *A First Course in Statistical Programming with R* (official site, Powell's), suitable for absolute beginners
  - Venables and Ripley's *Modern Applied Statistics with S* (official site, Powell's), useful but more advanced material
  - John M. Chambers, *Software for Data Analysis: Programming with R* (official site, Powell's) is the best book on writing programs in R.

MapReduce/Hadoop

- MapReduce is a software framework introduced by Google to support distributed computing on large data sets on clusters of computers.
- See http://en.wikipedia.org/wiki/MapReduce
- See Hadoop for open source implementation in Java

- Can be used for simple tasks such as counting, to inverted index creation to PageRank calculation for extremely large datasets (billions of rows).

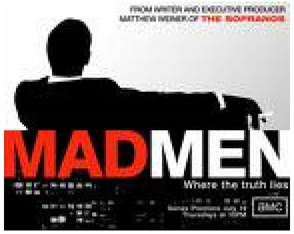
Homegrown Evaluations

- **Use publicly available datasets**
  – LETOR, Yandex LETOR, Netflix, MSAd data, TREC, CLEF, INEX
- **Simulate**
  – •M. Langheinrich, A. Nakamura, N. Abe, T. Kamba and Y. Koseki, (http://www8.org/w8-papers/2b-customizing/unintrusive/unintrusive.html) Unintrusive customization techniques for Web advertising, *Computer Networks* 31,pp.1259-1272, 1999. Targeted Internet Advertising Using Predictive Clustering and Linear Programming
- **Label data**
  – Berthier A. Ribeiro-Neto, Marco Cristo, Paulo Braz Golgher, Edleno Silva de Moura: Impedance coupling in content-targeted advertising. SIGIR 2005: 496-503
  – Use Mechanical Turk
- **Collaborate with advertising industry; intern**

Outline

- Introduction
- Online advertising background
- Business models
- Creating an online ad campaign
- Technology and Economics
- New Directions
- Challenges in online advertising
- Summary

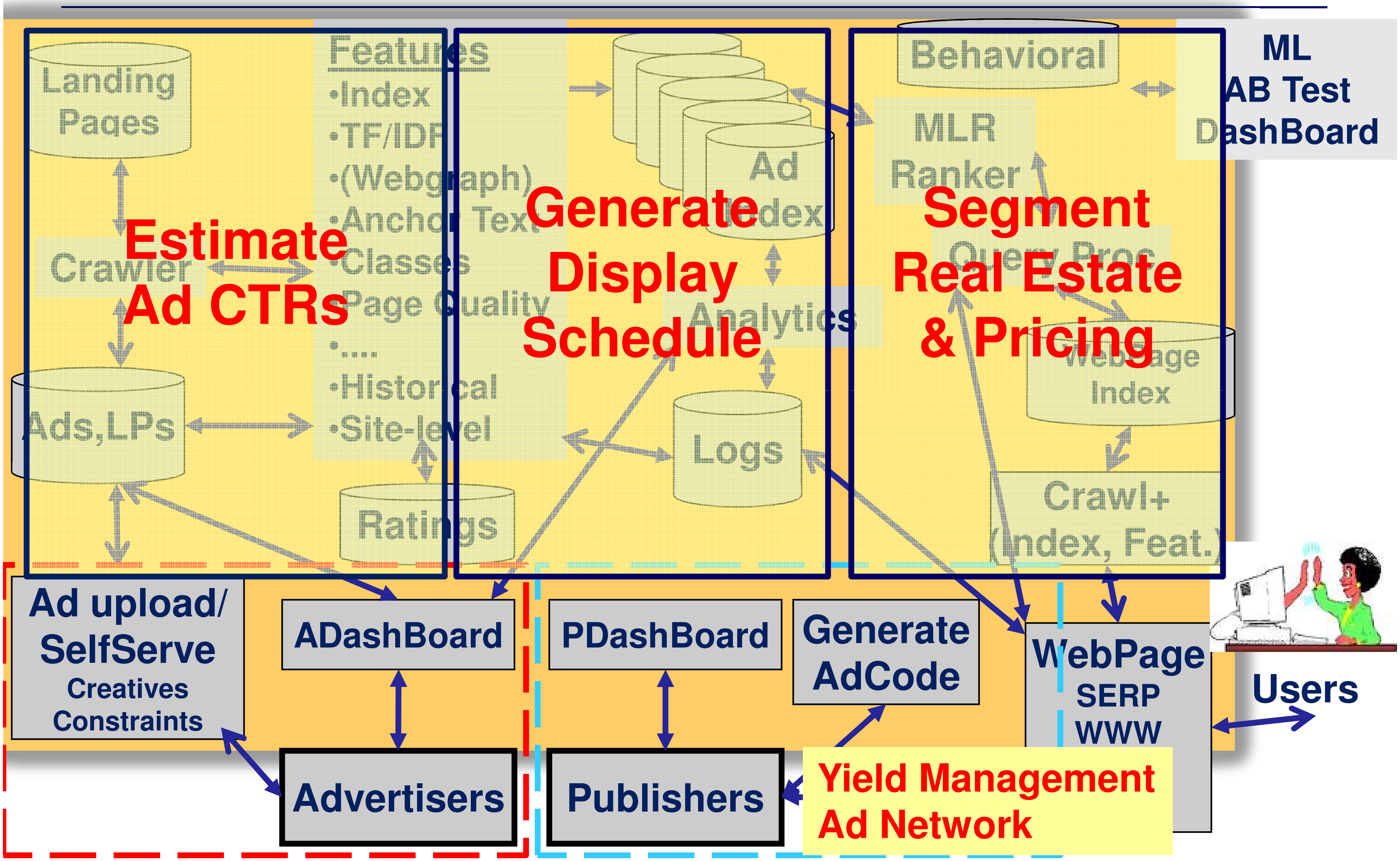
Business, Gold rush



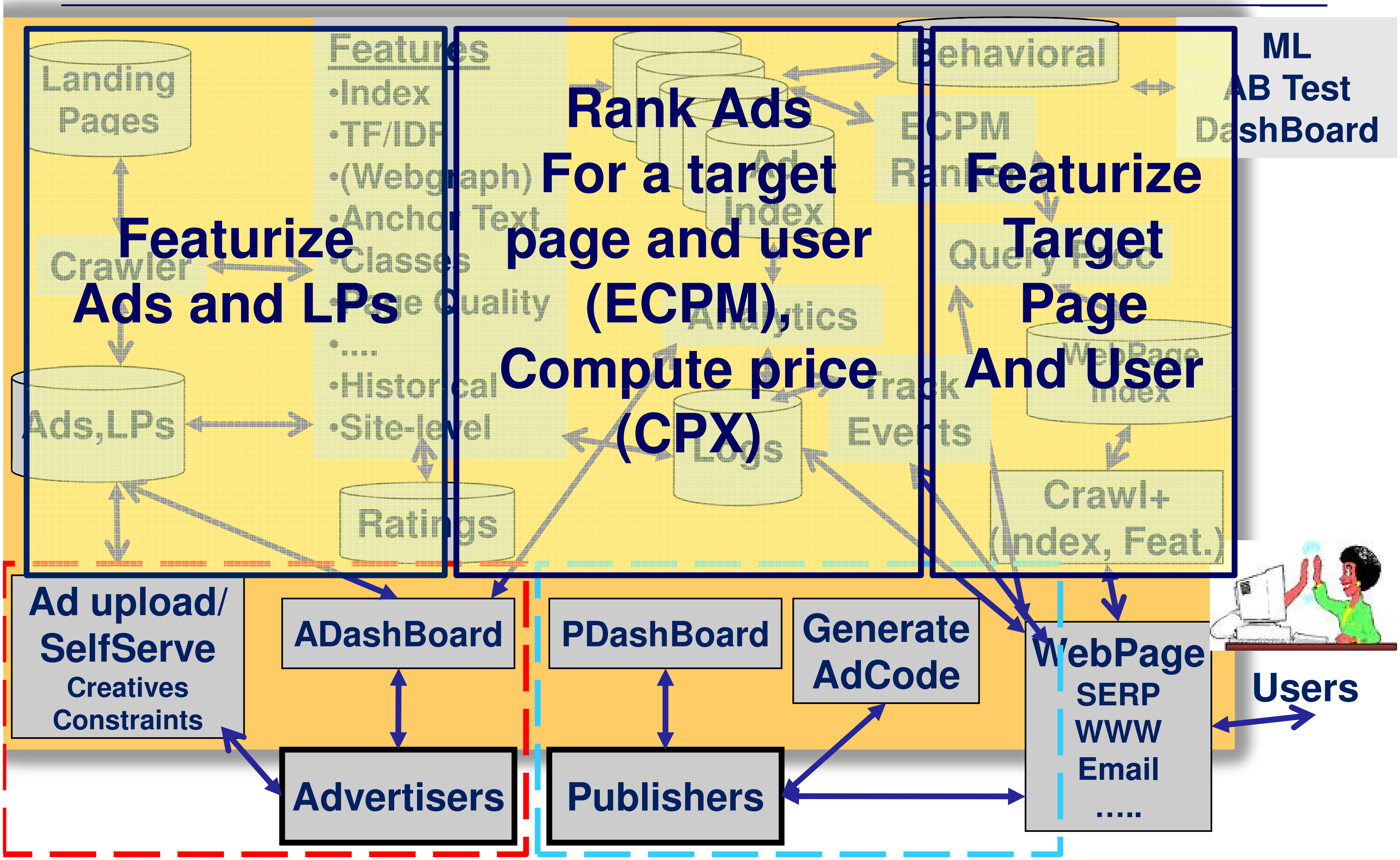
Tech

Hot Areas

Ad Network Architecture: Forward Market



Ad Network Architecture: Spot Market



Executive Summary

- **Advertisers deliver a message to consumers via publishers**
- **Online advertising revenue growing annually @ 10% (from26%)**
  - $23.4 billion in USA (2008), $65B worldwide (10% of overall ad spend)
  - Business Models: CPM, CPC, CPA (not popular yet but will be), dCPM
  - Online advertising is location, location and location ($11B M&A, 2007)
  - ….*and* data, features, objectives and optimization
- **Key directions**
  - Forward markets → **Spot Markets**
    - Advertising transformed from a low-tech, "Mad Men" process to highly optimized, mathematical, computer-centric (Wall Street-like) process
  - **Technology:** Economics, IR, ML, OR, Game Theory, Stats, Social Sciences (estimating CTR/AR; auction systems; learning algorithms; targeting display advertising; behavioral targeting)
  - **Click fraud** is the spam of online advertising ($Billion problem)
  - **New Directions**: Web 3.0; games; social advertising; data exchanges
  - **Gap:** perceived gap between academia and industry?

From Mad Men To Wall Street

- Set in New York City, *Mad Men* begins in 1960 at the fictional Sterling Cooper advertising agency on New York City's Madison Avenue.

Human Intensive
Lots of guess work
Forward Market

Increasingly

Technology
Data Driven
Forward Market
Spot Markets

Course philosophy

- **Socratic Method**
  - participation strongly encouraged (please state your name and affiliation)
- **Highly interactive and adaptable**
  - Questions welcome!!
- **Lectures emphasize intuition, not rigor or detail**
  - Build on lectures from other faculty
  - Background reading will have rigor & detail
- **Action Items**
  - Read suggested books first (and then papers), read/write Wikipedia, watch/make YouTube videos, take courses, participate in competitions, do internships, network
  - Prototype, publish, participate, simulate
  - Classic (core) versus trendy (applications)

Thanks

EMAIL:
James_DOT_Shanahan_AT_gmail_DOT_com

Bibliography

- Shanahan James G., 2010, Online Advertising: From “Mad Men” to Wall Street
- Nelson, Philip, 1974. Advertising as Information, Journal of Political Economy, University of Chicago Press, vol. 82(4), pages 729-54.
- R. Baeza-Yates, A. Broder, P. Raghavan, WWW 2007, Tutorial, Foundations and Challenges of Web Advertising
- R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley-Longman, 1st edition, 1999.
- Y. Kohda and S. Endo. Ubiquitous advertising on the www: merging advertisement on the browser. Comput. Netw. ISDN Syst., 28(7-11):1493–1499, 1996.
- The Google adwords. Google content-targeted advertising. http://adwords.google.com/select/ct_faq.html
- M. Weideman and T. Haig-Smith. An investigation into search engines as a form of targeted advert delivery. In Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology, pages 258-258. 2002.
- Large-Scale Behavioral Targeting - Ye Chen* Yahoo! Labs; Dmitry Pavlov Yahoo! Labs; John Canny Computer Science Division University of California Berkeley
- *Optimal Auction Design in a Multi-unit Environment: The Case of Sponsored Search Auctions*. With Michael Schwarz. (December 2006)
- Statistics for Experimenters: Design, Innovation, and Discovery, 2nd Edition by George Box, Stuart Hunter and William G. Hunter, 2005

Bibliography

- David Maxwell Chickering , David Heckerman, Targeted advertising on the Web with inventory management, Interfaces, v.33 n.5, p.71-77, September 2003
- Atsuyoshi Nakamura, Improvements in practical aspects of optimally scheduling web advertising, Proceedings of the eleventh international conference on World Wide Web, May 07-11, 2002, Honolulu, Hawaii, USA.
- P Berthon, LF Pitt, RT Watson , The world wide web as an advertising medium, Journal of Advertising Research, 1996.
- A Mehta, A Saberi, U Vazirani, V Vazirani. AdWords and Generalized On-line Matching. Harvard Univ, 2005.
- Ribeiro-Neto, B., Cristo, M., Golgher, P. B., and Silva de Moura, E. 2005. Impedance coupling in content-targeted advertising. In *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05, 496-503.
- Edelman, B., Ostrovsky, M., Schwarz, M. Internet advertising and the generalized second price auction: selling billions of dollars worth of keywords. NBER Working Paper No. W11765, 2005
- http://www.cis.upenn.edu/~mkearns/teaching/SponsoredSearch/
- S. Pandey, C. Olston, 2006, Handling Advertisements of Unknown Quality in Search Advertising
- Esteban Feuerstein, Pablo Heiber, Javier Martinez-Viademonte and Ricardo Baeza-Yates., New Stochastic Algorithms for Placing Ads in Sponsored Search. LA-Web, Santiago, Chile 2007.
- TNS Media Intelligence
- Global Entertainment and Media Outlook: 2006-2010

Bibliography

Machine Learning

- Logistic Regression http://en.wikipedia.org/wiki/Logistic_regression
- Logistic Regression http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_4.html
- Friedman, J. H. "Tutorial: Getting Started with MART in R." (April 2002) (software)
- Friedman, J. H. "Stochastic Gradient Boosting ." (March 1999b) (software)
- Friedman, J. H. "Greedy Function Approximation: A Gradient Boosting Machine." (Feb. 1999a) (software)
- Hosmer DW, Lemeshow S. Applied logistic regression. Wiley & Sons, New York, 1989
- Duda, Hart, & Stork (2000). Pattern Classification. http://rii.ricoh.com/~stork/DHS.html
- Russell & Norvig (2004). Artificial Intelligence. http://aima.cs.berkeley.edu/

Online Advertising Background:

- http://en.wikipedia.org/wiki/Advertising
- http://en.wikipedia.org/wiki/Behavioral_targeting

Bibliography

- **Matthew Richardson, Ewa Dominowska, Robert Ragno: Predicting clicks: estimating the click-through rate for new ads. WWW 2007 pages 521-530, http://www2007.org/papers/paper784.pdf**
- **Atsuyoshi Nakamura, Improvements in practical aspects of optimally scheduling web advertising, Proceedings of the eleventh international conference on World Wide Web, May 07-11, 2002, Honolulu, Hawaii, USA.**
- **Online Advertising. In *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation Inc. Encyclopedia on-line. Available from http://en.wikipedia.org/wiki/Online_advertising. Internet. Retrieved June 11, 2009.**
- **Behavioral Targeting. In *Wikipedia: The Free Encyclopedia*. Wikimedia Foundation Inc. Encyclopedia on-line. Available from http://en.wikipedia.org/wiki/Behavioral_targeting. Internet. Retrieved June 11, 2009.**
- **OTHER Resources**
  - See Proceedings WWW 2008, WWW 2009 for other papers on Ranking ads, economic models etc.
  - See OMMA
  - www.eMarkerter.com

THANKS!

The End