ISM 280-290: Data Mining, Analytics and Information Extraction in Intelligent Business Services: Online Ads, Healthcare, and Service Centers

James G. Shanahan[1] and Ram Akella

[1]*Independent Consultant*

EMAIL: James_DOT_Shanahan_AT_gmail_DOT_com

ISM 280

UC Berkeley & UC Santa Cruz

Wednesday January 19, 2011

ISM 280-290: Data Mining, Analytics and Information Extraction

- **Data Mining, Analytics and Information Extraction in Intelligent Business Services: Online Ads, Healthcare, and Service Centers**
- **Course Description:**
  - The purpose of this course is to provide an Online Marketing and Ads, Healthcare, and Service Center industry context for data mining, information extraction, and analytics, as an important element for student work in information systems and analytics.
- **Specifically, we hope to:**
  - Provide an overview of issues and trends which will shape the need for and structures of data mining, information extraction, and analytics in business information systems within online marketing and ads, healthcare, and service centers.
  - Identity and explore key topics, followed by the development of analytics methods, for data mining, analytics, and information extraction, in these contexts.
  - We will have industry speakers and industry projects as well, to provide real world perspective and real world engagement.

Brief Bio James G. Shanahan

- **20 years in the field AI and information management**
  - Principal and Founder, Boutique Data Consultancy
    - Clients include: Digg, SearchMe, AT&T, SkyGrid, MyOfferPal,
  - Affiliated with University of California Santa Cruz (UCSC, ISM250,251,209)
  - Chief Scientist, Turn Inc. (A CPX ad network, DSP)
  - Principal Scientist, Clairvoyance Corp (CMU spinoff; sister lab to JRC)
  - Research Scientist, Xerox Research
  - Research Engineer, Mitsubishi Group
  - PhD in machine learning (1998), University of Bristol, UK;
    B.Sc. Comp. Science (1989), Uni. of Limerick, Ireland
- **Now: Machine Learning Consultant (San Francisco)**
  - IF *(you have large **data problems** and need a consultant)*
    THEN *{email me at James.Shanahan_AT_gmail,com}*
  - Where **problems** $\in$ *{web search, online advertising, machine learning, ranking, user modeling, statistics, social networks, "*" }*

Topics 1/2

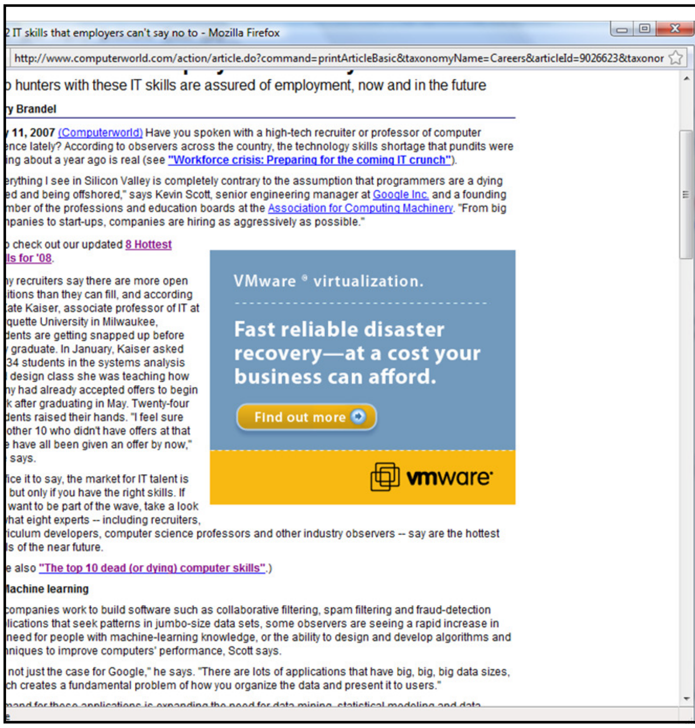
- **Part 1 of the course**
  - Machine Learning, Logistic regression, SVD, constrained optimisation, text mining, information retrieval, prediction, clustering

Topics 2/2

- **Online Advertising (1 Lecture 2/23)**
- **Information extraction (6 Lectures, March/April)**
  - NLP Basics and Named Entity Recognition
  - NER as classification
  - Hidden Markov models and Maxent Markov Models
  - Basic information retrieval
  - Conditional Random fields with applications
  - Query/sentence parsing for web search and local search
  - Sentiment Analysis
- **Lnaguages: R and Python, with Lucene and LingPipe**

ISM 250 is timely!

- **ISM280 core**
  - Data Mining, Analytics and Information Extraction in Intelligent Business Services
- **…. with applications in digital advertising**
  - Online Ads, Healthcare, and Service Centers Convex Optimization
- **Timely:**
  - Growing flood of online data, Budding industries (e.g., digital advertising)
  - Computational power is available (PC, Cloud computing, Hadoop)
  - Progress in algorithms and theory and applications


2 IT skills that employers can't say no to - Mozilla Firefox

http://www.computerworld.com/action/article.do?command=printArticleBasic&taxonomyName=Careers&articleId=9026623&taxonor

b hunters with these IT skills are assured of employment, now and in the future

y Brandel

/ 11, 2007 (Computerworld) Have you spoken with a high-tech recruiter or professor of computer nce lately? According to observers across the country, the technology skills shortage that pundits were ing about a year ago is real (see "Workforce crisis: Preparing for the coming IT crunch").

erything I see in Silicon Valley is completely contrary to the assumption that programmers are a dying ed and being offshored," says Kevin Scott, senior engineering manager at Google Inc. and a founding mber of the professions and education boards at the Association for Computing Machinery. "From big npanies to start-ups, companies are hiring as aggressively as possible."

b check out our updated 8 Hottest ls for '08.

y recruiters say there are more open itions than they can fill, and according ate Kaiser, associate professor of IT at quette University in Milwaukee, dents are getting snapped up before y graduate. In January, Kaiser asked 34 students in the systems analysis design class she was teaching how y had already accepted offers to begin k after graduating in May. Twenty-four dents raised their hands. "I feel sure other 10 who didn't have offers at that e have all been given an offer by now," says.

ice it to say, the market for IT talent is but only if you have the right skills. If want to be part of the wave, take a look hat eight experts -- including recruiters, iculum developers, computer science professors and other industry observers -- say are the hottest s of the near future.

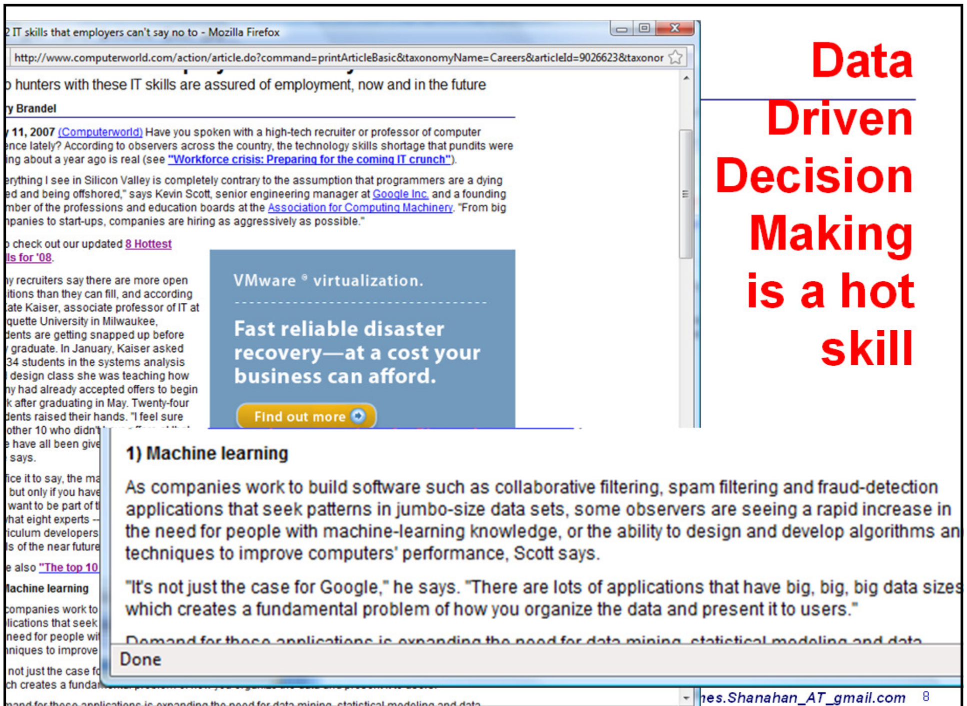
e also "The top 10 dead (or dying) computer skills".)

lachine learning

companies work to build software such as collaborative filtering, spam filtering and fraud-detection lications that seek patterns in jumbo-size data sets, some observers are seeing a rapid increase in need for people with machine-learning knowledge, or the ability to design and develop algorithms and hniques to improve computers' performance, Scott says.

not just the case for Google," he says. "There are lots of applications that have big, big, big data sizes, ch creates a fundamental problem of how you organize the data and present it to users."

Data Driven Decision Making is hot skill


2 IT skills that employers can't say no to - Mozilla Firefox

http://www.computerworld.com/action/article.do?command=printArticleBasic&taxonomyName=Careers&articleId=9026623&taxonor

b hunters with these IT skills are assured of employment, now and in the future

1) Machine learning

As companies work to build software such as collaborative filtering, spam filtering and fraud-detection applications that seek patterns in jumbo-size data sets, some observers are seeing a rapid increase in the need for people with machine-learning knowledge, or the ability to design and develop algorithms an techniques to improve computers' performance, Scott says.

"It's not just the case for Google," he says. "There are lots of applications that have big, big, big data sizes which creates a fundamental problem of how you organize the data and present it to users."

Done

Data Driven Decision Making is a hot skill

Course Modus Operandi

- **ISM250 will focus on getting students familiar with core principles in Stochastic Optimization**
- **Grounding these principles in both**
  - (1) examples taken primarily form online advertising (a $65 Billion industry)
  - And in (2) example projects and code in R.
- **Each class will be composed of theory, practice and problems, thereby informing and inspiring students on how to apply theory to practice.**

Some Practical Skills

- **Problem solving**
- **Data analysis**
- **Coding up algorithms**
- **Real-world datasets**
- **Evaluations and metrics**
- **Collaboration**
- **Presentation**
- **Teamwork**

Audience Participation

Questionnaire

- **Background**
  - Industry/Academia
  - Major
  - Programming experience
- **Expectations from taking ISM280**

Course philosophy

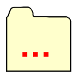
- **Socratic Method (both inspiration and information)**
  - participation strongly encouraged (please state your name and affiliation)
- **Highly interactive and adaptable**
  - Questions welcome!!
- **Lectures emphasize intuition, rigor and detail**
  - Build on lectures
  - Background reading will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

Disclaimer

- **The Authors retains all rights, including copyrights and distribution rights.**
- **No publication or further distribution in full or in part permitted without explicit written permission from the author**
- **Living vicariously!**

Course Topics: Part 2Outline

Bag of words



- **White space tokenization**
- **Good Classification Technology**
  - Thresholded SVMs

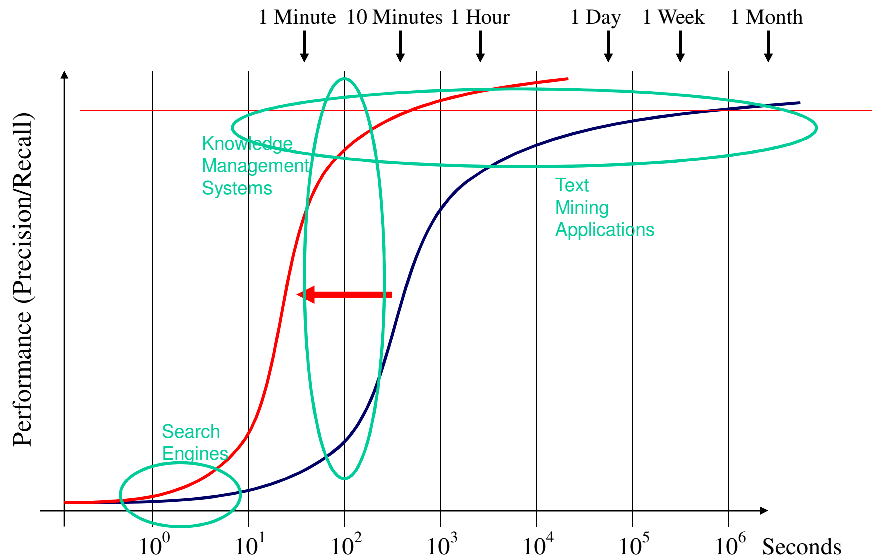
- **Extra semantic processing**
  - Affect/Opinion

- **Process Mining**
  - Bayesian Network Approach

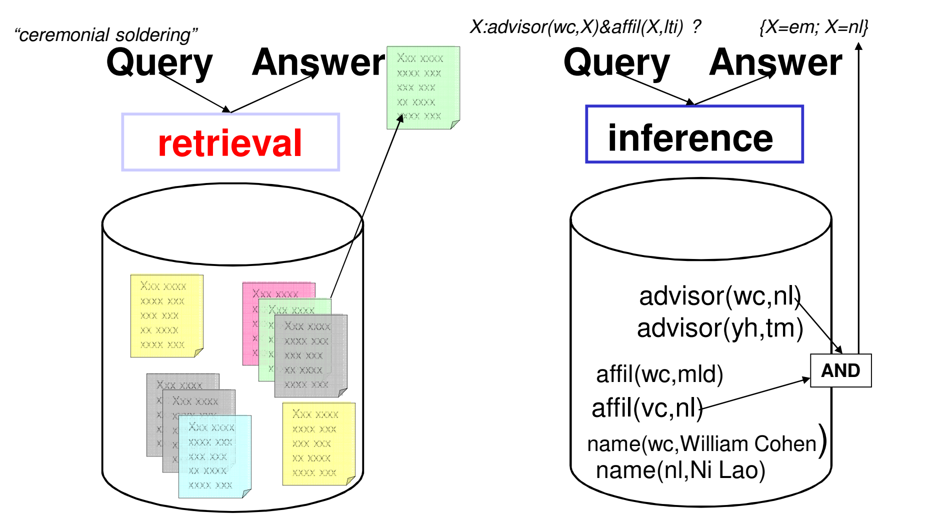
Rest of Lecture 1 Outline

- **Background:**
  - Information extraction vs information retrieval
- **Advertising 101 and Digital advertising**
  - Predicting CTR
- **Information Extraction Overview**
- **Sentiment Analysis**
- **Candidate Project**

The Information Access Curve



Two ways to manage information



Management Science Core App Areas

- **In short, management sciences help businesses to achieve their goals using the scientific methods of analytics and operational research.**
  - mathematical modeling, statistics and numerical algorithms
  - optimal or near optimal solutions to complex decision problems.
- **Airlines, manufacturing companies, service organizations, military branches, government, and *internet companies*.**
  - ***Real time decision making in data rich environments (internet information systems, digital advertising, stock trading, healthcare)***
  - Scheduling airlines, including both planes and crew
  - Place new facilities such as a warehouse, factory or fire station
  - Managing the flow of water from reservoirs
  - Identifying possible future development paths for parts of the telecommunications industry, health service

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© *2011 James G. Shanahan James.Shanahan_AT_gmail.com* 19

Management Science Core App Areas

- **In short, management sciences help businesses to achieve their goals using the scientific methods of operational research.**
  - mathematical modeling, statistics and numerical algorithms
  - optimal or near optimal solutions to complex decision pro…
- **Airlines, manufacturing companies, servic… …nizations, military branches, government, and *int… companies*.**
  - ***Real time decision making in data ric… …nments (internet information systems, digital adve… …stock trading)***
  - Scheduling airlines, including b… …es and crew
  - Place new facilities such … …ehouse, factory or fire station
  - Managing the flow … …rom reservoirs
  - Identifying pos… …ure development paths for parts of the telecomm… …s industry, health service

2000s: Realtime, huge data, lots of uncertainty

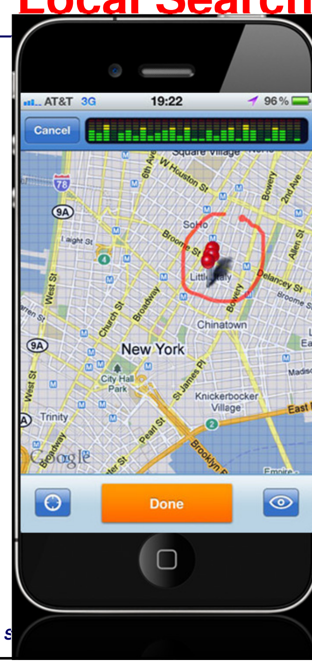
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© *2011 James G. Shanahan James.Shanahan_AT_gmail.com* 20

iPhone4 App for Local Search

- **Speach**
- **Speak4it is the original multimodal voice-driven local search app for the iPhone, iPad, and iPod touch. Just press the "Push to speak" button and say what you'd like to find. You can even point to a spot on the map and ask what's there.**

iPhone4 App for Local Search

- **Here are some things you might try saying:**
  - "Coffee shops"
  - "Pizza"
  - "Walgreens near me"
  - **Point to a spot on the map and say "Thai food around here"**
  - "Hotels near Disneyland"



Rest of Lecture 1 Outline

- **Background:**
  – Information extraction vs information retrieval
- **Advertising 101 and Digital advertising**
  – Predicting CTR
- **Information Extraction Overview**
- **Sentiment Analysis**
- **Candidate Project**

Advertising

- **Advertising is a paid, one-way communication**
  1. Deliver marketing messages and attract new customers
  2. To inform potential customers about products and services and how to obtain and use them.
  3. Branding → Direct action
     - Many advertisements are also designed to generate increased consumption of those products and services through the creation and reinforcement of brand image and brand loyalty (ads contain both factual information and persuasive messages).
  4. Use every major medium
     - To deliver these messages, including: television, radio, movies, magazines, newspapers, video games, the Internet, and billboards

Digital Advertising

- **Online advertising is a form of advertising utilizing the Internet and World Wide Web in order to deliver marketing messages and attract customers** [wikipedia.com]
- **Advertising annoys people! Advertising works!**
  - "*Half the money I spend on advertising is wasted; the trouble is, I don't know which half.*" - John Wanamaker, father of modern advertising. [Credit assignment]
  - "*I do not regard advertising as entertainment or an art form, but as a medium of information...*", "*Ogilvy on Advertising*" by David Ogilvy
- **Goals of Online advertising**

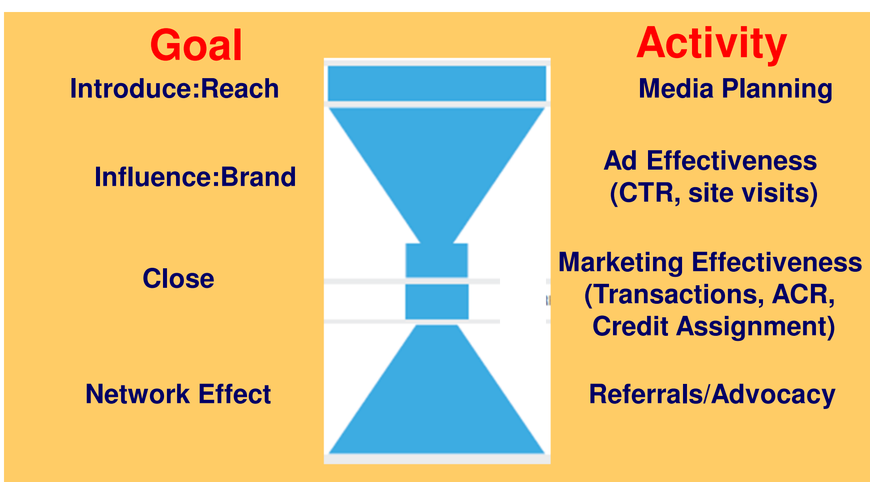
A – *Deliver/push an advertiser's message with* quantifiable measures of consumer interest

A+P – Generate ROI for the advertiser and revenue for the publisher

P+C – Enable ads as *a medium of information* (true in the case of search)!

What marketers want?

- **Deliver marketing messages and attract customers and sell products/services**



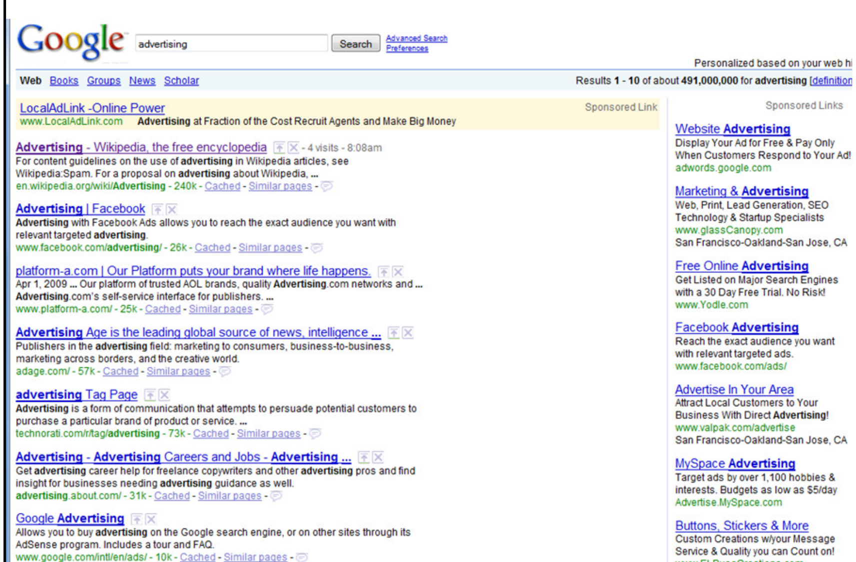
Advertising makes up ~2% of US GDP

Despite its problems (lack of credit assignment etc.)

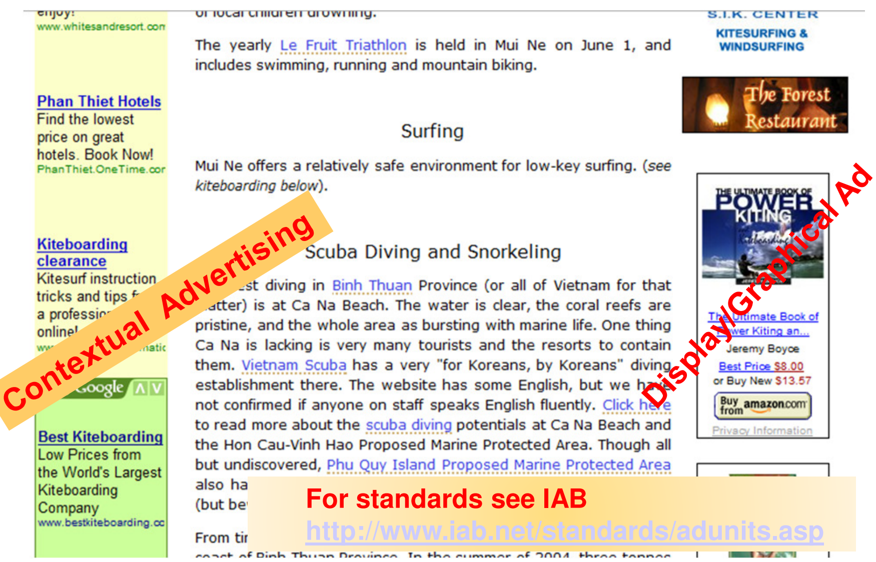
- **US GDP = $14.1 Trillion (Global $56 Trillion, 56×10^{12})**
- **US Advertising Spend**
  - ~$275 Billion across all media (2% of GDP since the early 1900s)
  - ~$23 Billion in Digital Advertising (8.4% of overall spend)
- **In 2008, Worldwide online advertising was $65B**
- **I.e., about 10% of all ad spending across all media [IDC, 2008]**

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

Online Advertising: Sponsored Search



Contextual Advertising on Webpages

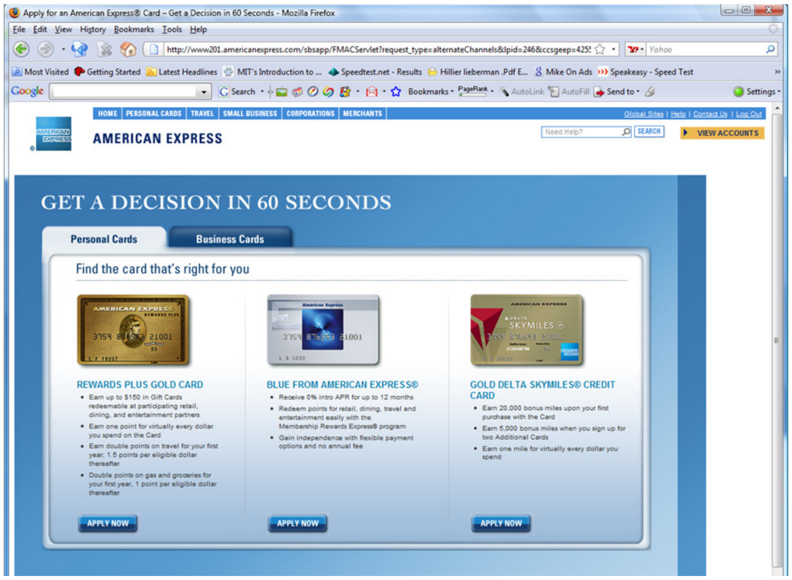


Ad Creative Formats and sizes

- **Text Ads**
- **Display Ads**
  - Graphical, Flash, Rich Media (sound, video)
- **Different sizes types:**
  - For details see http://www.iab.net/standards/adunits.asp
  - Rectangular, R and Buttons
- **See IAB for sta**

[source: www.activeresponsegroup.com]

House Ads at AMEX

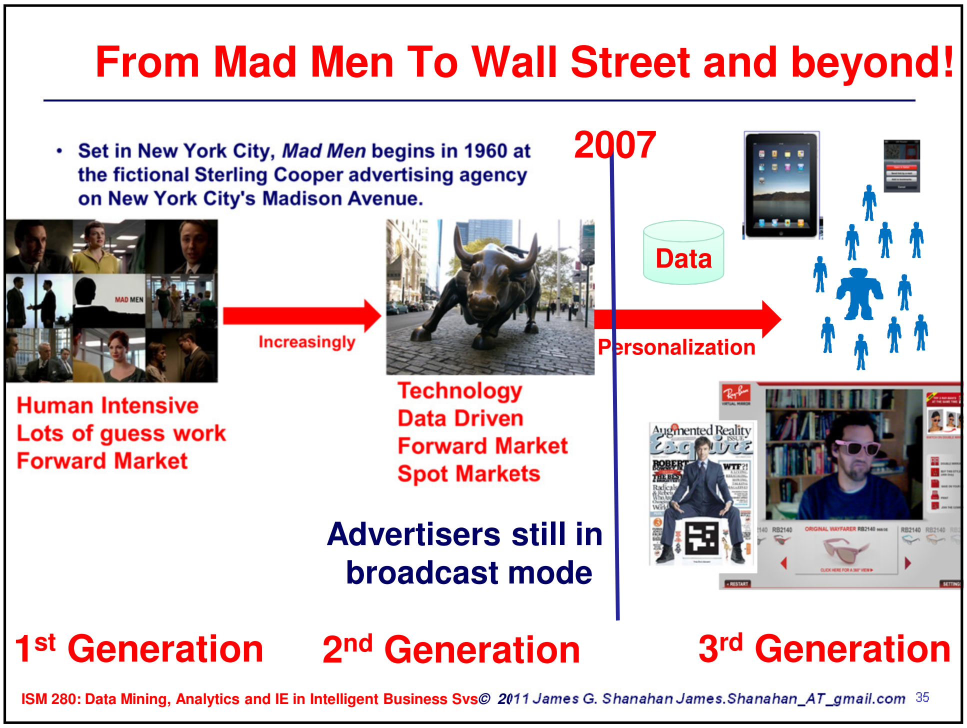


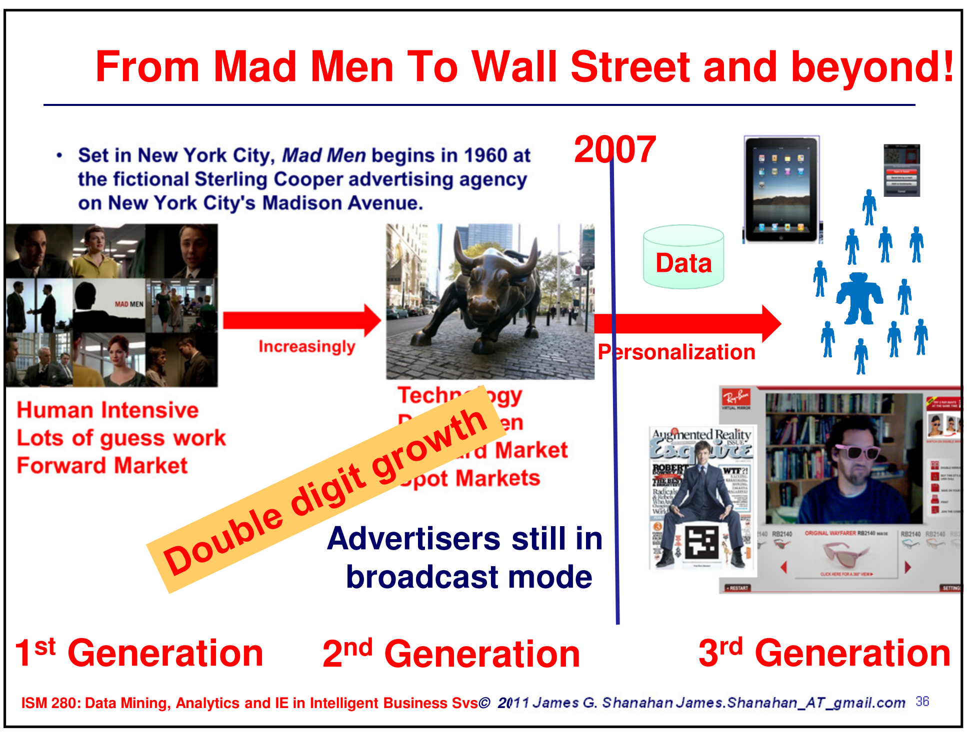
House Ads at Amazon

OA is cavalier! : business models; ad placement; e:b wants to be online

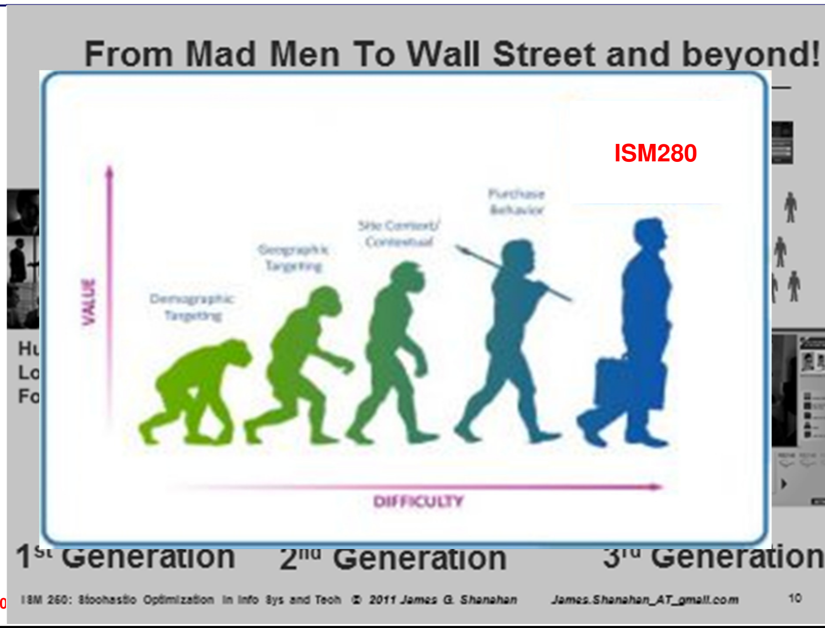
Business Models

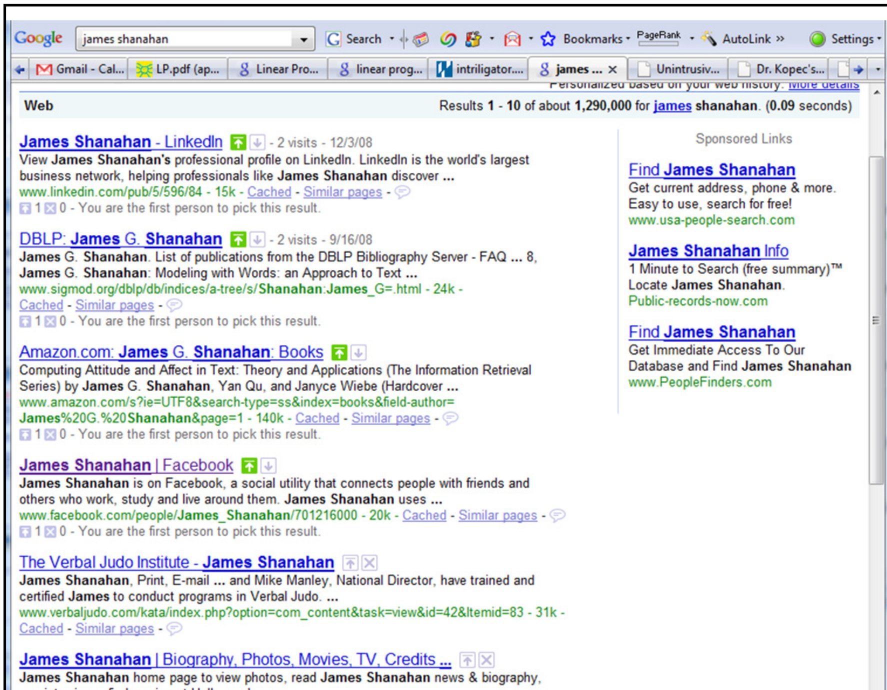
- **CPM (Cost Per Mille/Thousand)**
  - Advertisers pay for exposure of their message to a specific audience. (*M* in the acronym is the Roman numeral for one thousand)
- **CPC (Cost Per Click) aka Pay per click (PPC)**
  - Advertisers pay every time a user clicks on their listing and is redirected to their website.
- **CPA (Cost Per Action) or (Cost Per Acquisition)**
  - The publisher takes all the risk of running the ad, and the advertiser pays only for the amount of users who complete a **transaction, such as a purchase or sign-up**.


From Mad Men To Wall Street and beyond!
• Set in New York City, *Mad Men* begins in 1960 at the fictional Sterling Cooper advertising agency on New York City's Madison Avenue.
2007
Data
Increasingly
Personalization
Human Intensive
Lots of guess work
Forward Market
Technology
Data Driven
Forward Market
Spot Markets
Advertisers still in broadcast mode
1st Generation
2nd Generation
3rd Generation
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 35

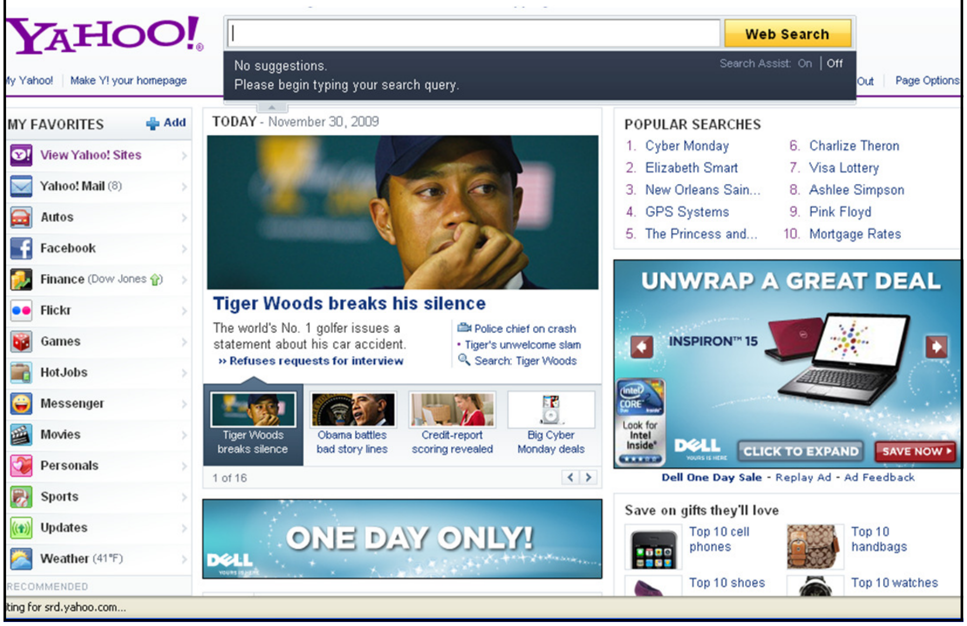

From Mad Men To Wall Street and beyond!
• Set in New York City, *Mad Men* begins in 1960 at the fictional Sterling Cooper advertising agency on New York City's Madison Avenue.
2007
Data
Increasingly
Personalization
Human Intensive
Lots of guess work
Forward Market
Double digit growth
Advertisers still in broadcast mode
1st Generation
2nd Generation
3rd Generation
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 36

Internet Age Management Science



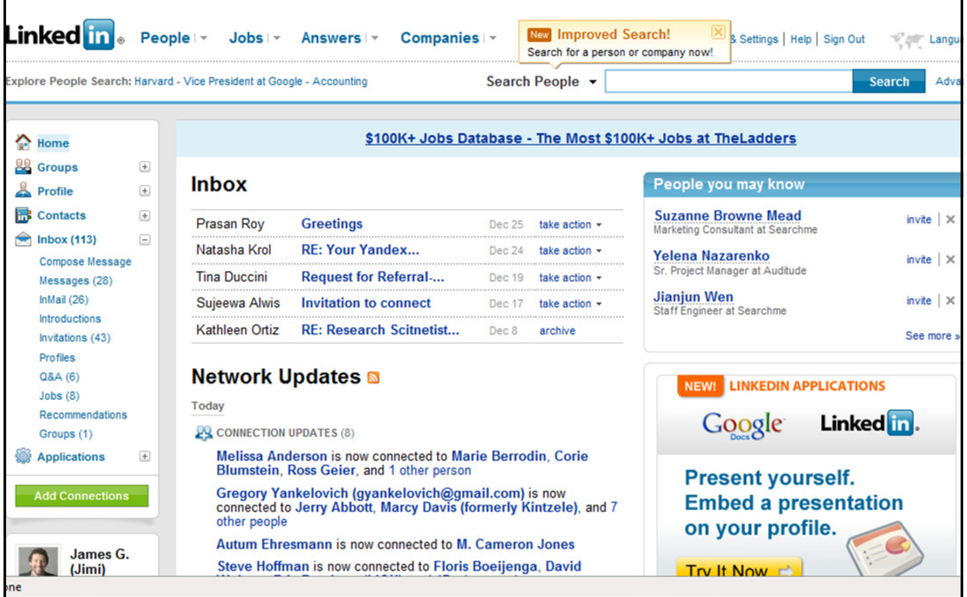


Over 300-400 applications of ML on this page



eCommerce: Online Shopping Over $20B

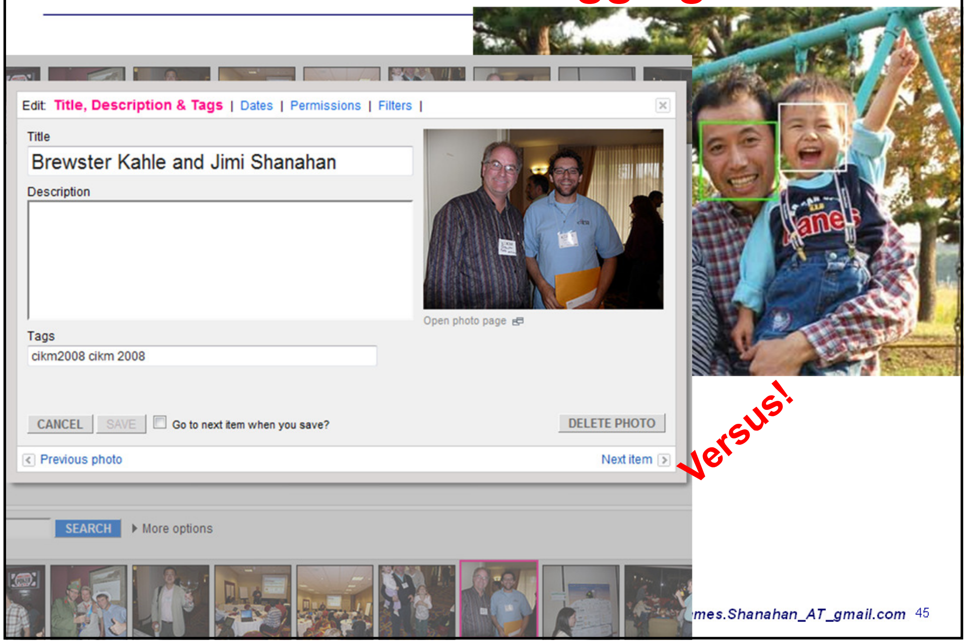
Online Advertising: $65B WorldWide



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© *2011 James G. Shanahan James.Shanahan_AT_gmail.com* 43

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© *2011 James G. Shanahan James.Shanahan_AT_gmail.com* 44

Automatic Tagging of Faces



Versus!

Branded Landing Page & Click-to-Video

Gallery/Image

Movie Trailer

Versus

CPM Banner Ad

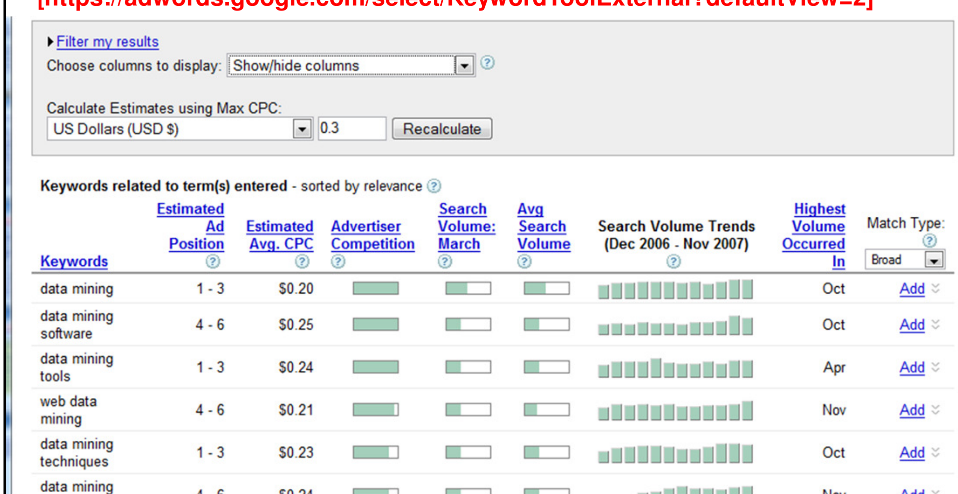
Branded Landing Page

2005-2007 AdMob, Inc.

E.g., Google AdWords

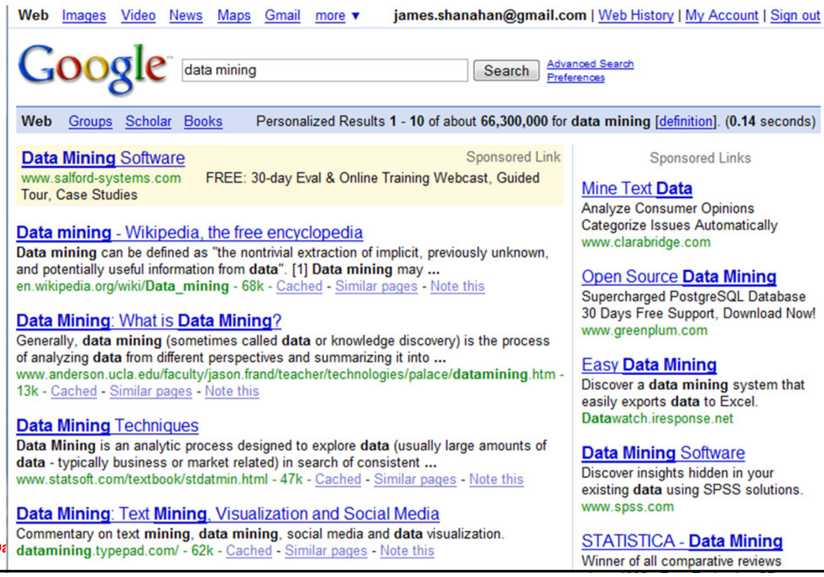
Select Portfolio of Keywords

[https://adwords.google.com/select/KeywordToolExternal?defaultView=2]

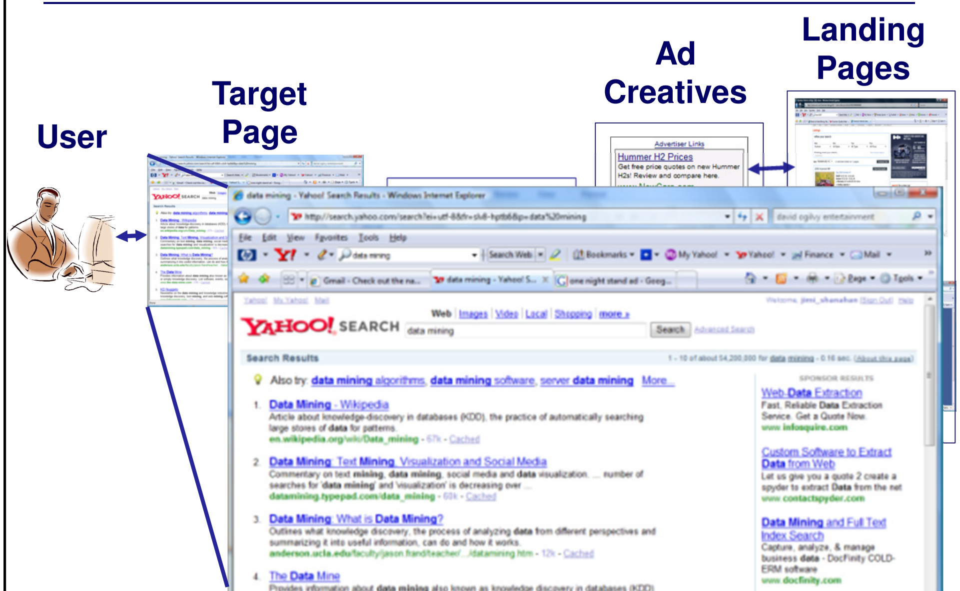


Keywords	Estimated Ad Position	Estimated Avg. CPC	Advertiser Competition	Search Volume: March	Avg Search Volume	Search Volume Trends (Dec 2006 - Nov 2007)	Highest Volume Occurred In	Match Type: Broad
data mining	1 - 3	$0.20					Oct	Add
data mining software	4 - 6	$0.25					Oct	Add
data mining tools	1 - 3	$0.24					Apr	Add
web data mining	4 - 6	$0.21					Nov	Add
data mining techniques	1 - 3	$0.23					Oct	Add
data mining	4 - 6	$0.24					Nov	Add

Which Keywords? How much for a click?

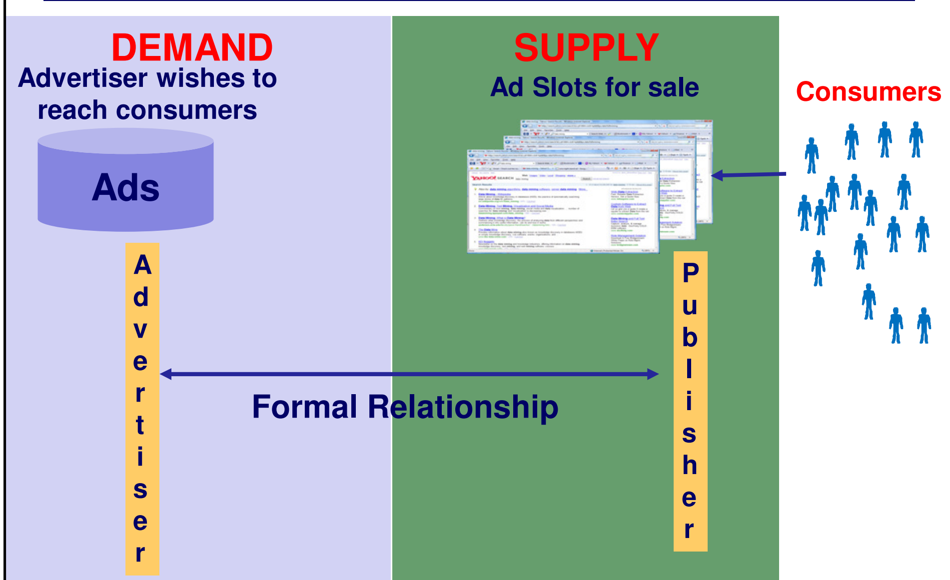


CPC Paid Search (KW Market place)

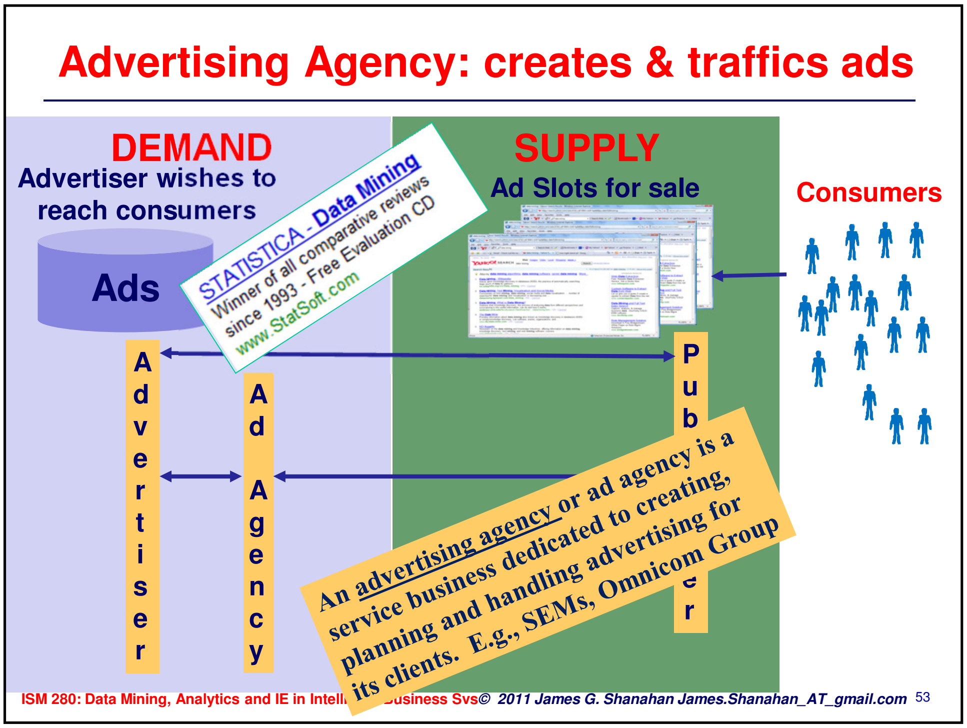


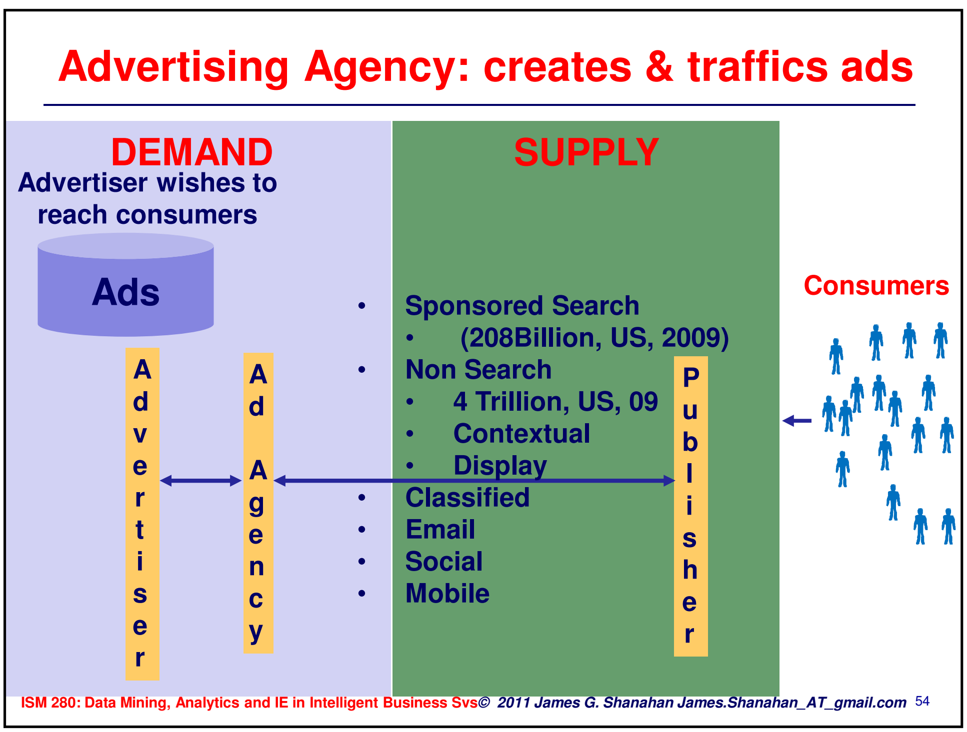
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 51

Advertising: a supply-demand marketplace



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 52

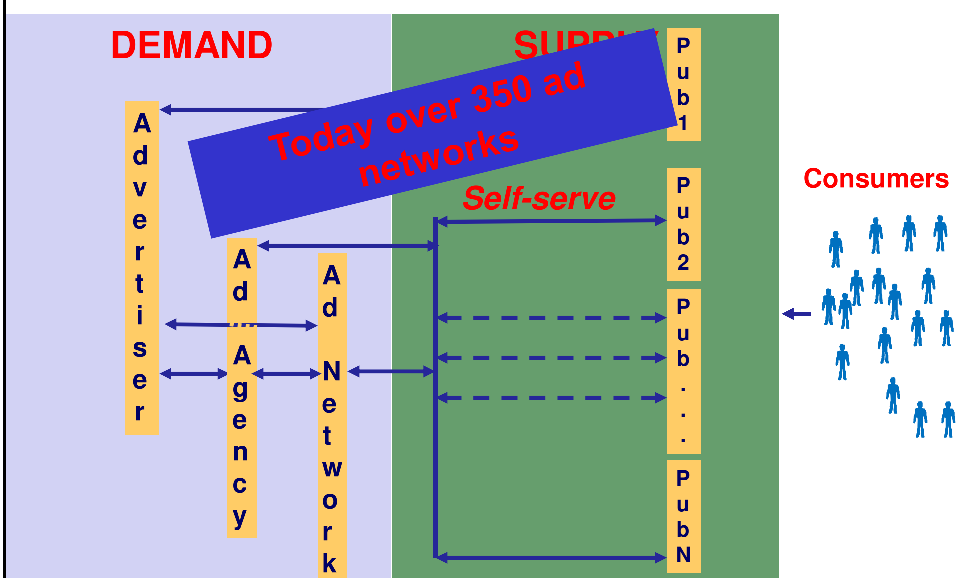

Advertising Agency: creates & traffics ads
DEMAND
Advertiser wishes to reach consumers
Ads
SUPPLY
Ad Slots for sale
Consumers
STATISTICA - Data Mining
Winner of all comparative reviews since 1993 - Free Evaluation CD
www.StatSoft.com
Advertiser
Ad Agency
Publisher
An advertising agency or ad agency is a service business dedicated to creating, planning and handling advertising for its clients. E.g., SEMs, Omnicom Group
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 53

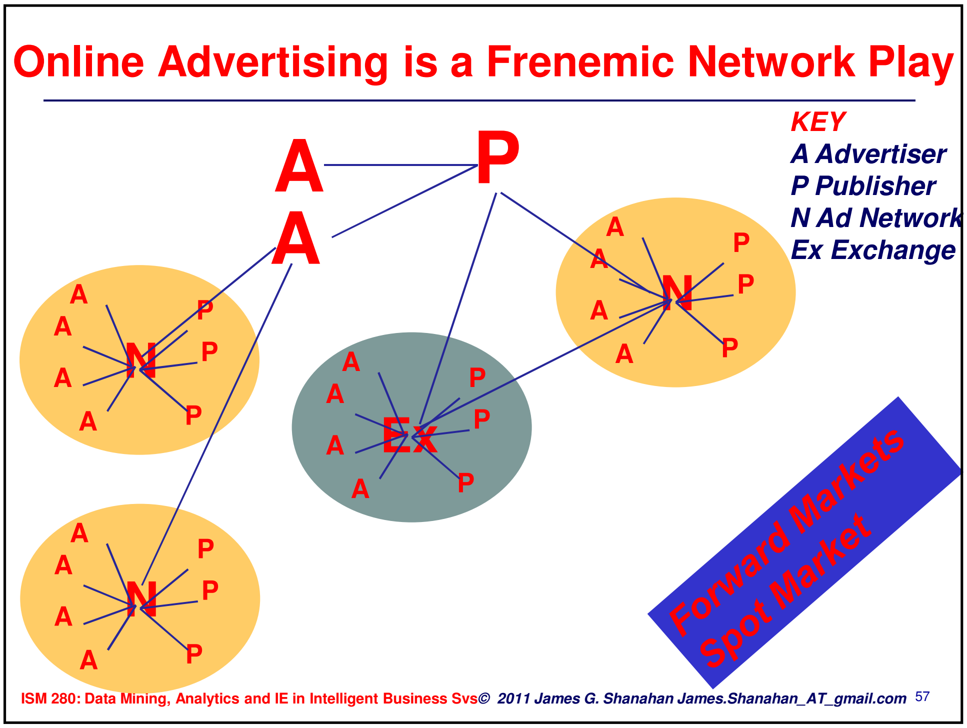

Advertising Agency: creates & traffics ads
DEMAND
Advertiser wishes to reach consumers
Ads
SUPPLY
• Sponsored Search
• (208Billion, US, 2009)
• Non Search
• 4 Trillion, US, 09
• Contextual
• Display
• Classified
• Email
• Social
• Mobile
Consumers
Advertiser
Ad Agency
Publisher
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 54

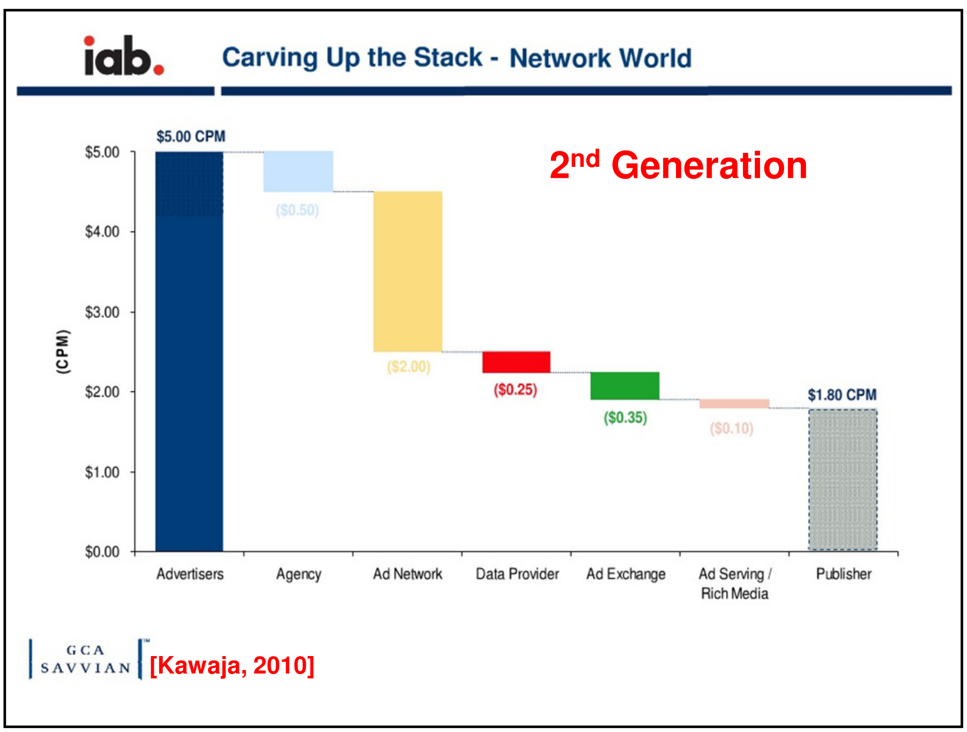
2nd Generation

- CPC, CPA
- Quant driven and quant support
- Supply can be fragmented → Ad Networks
  - Outside of search supply can be fragmented
  - Publishers maybe small and not have a sales team

Advertising Network: Aggregates Publishers



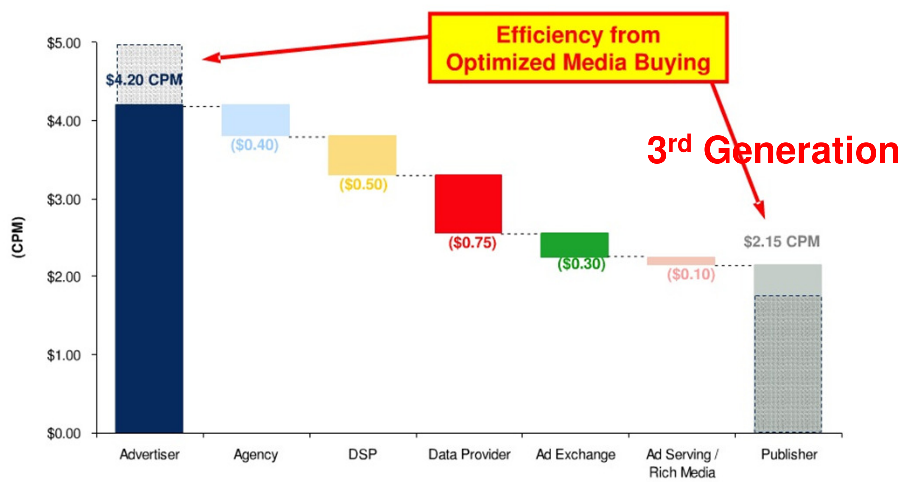

Online Advertising is a Frenemic Network Play
KEY
A Advertiser
P Publisher
N Ad Network
Ex Exchange
A
P
A
N
Ex
Forward Markets
Spot Market
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 57


iab.
Carving Up the Stack - Network World
2nd Generation
(CPM)
$5.00 CPM
$5.00
$4.00
$3.00
$2.00
$1.00
$0.00
($0.50)
($2.00)
($0.25)
($0.35)
($0.10)
$1.80 CPM
Advertisers
Agency
Ad Network
Data Provider
Ad Exchange
Ad Serving / Rich Media
Publisher
GCA SAVVIAN
[Kawaja, 2010]

3rd Generation

- **New more efficient market places**
  - Ad Exchanges
  - Data exchanges
- **Audience-based targeting**
- **Very complex pipeline**
  - Yield mgt and Demand side platforms

Carving Up the Stack - DSP / Exchange World

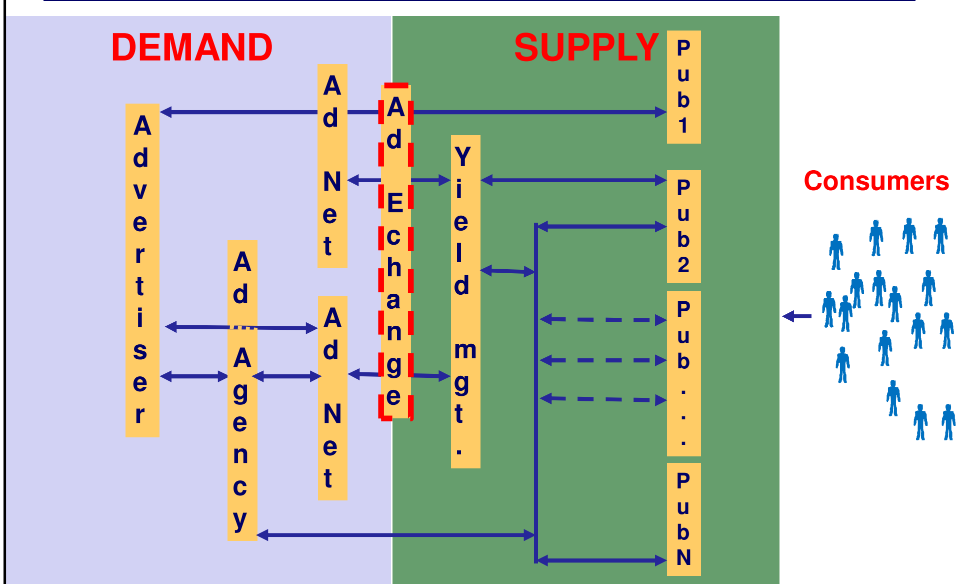


GCA SAVVIAN **[Kawaja, 2010]**

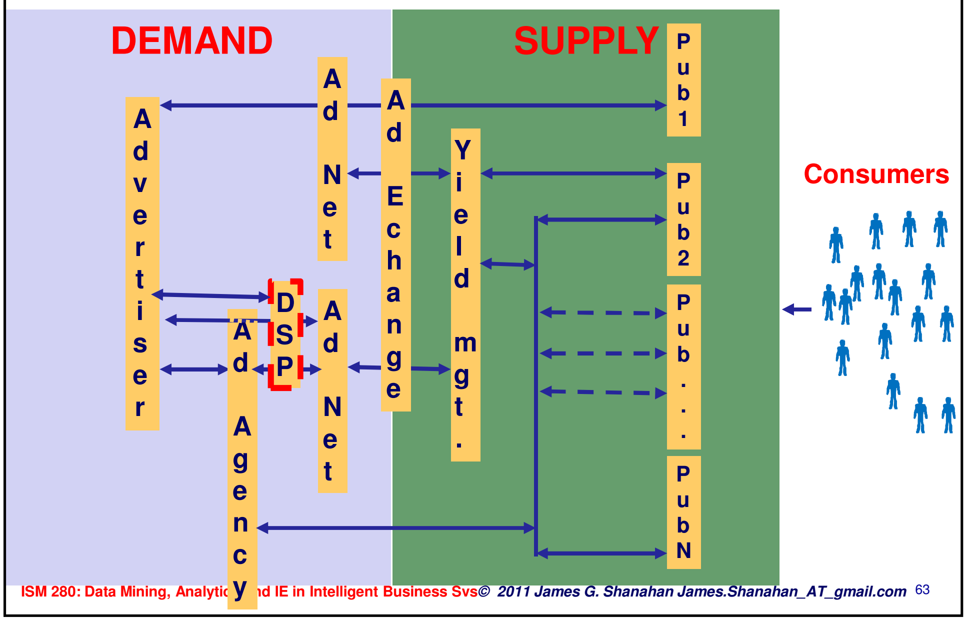
Ad Exchanges: a new SD Marketplace

- **The ad exchange is a real time marketplace**
  - with an auction-based system where the participants - advertisers and publishers – transact on a common platform to purchase and sell online graphical advertising.
- **Currently, publishers sell remnant inventory**
  - on the exchange for advertisers to purchase through bidding on a user-friendly interface.
- **Ad Exchanges do not compete with ad networks**
  - targeting technologies, or publishers, but rather serve as a more efficient way for the exchange of inventory within these groups
- **Googles acquired DoubleClick, Yahoo acq RightMedia, etc.. $11 in M&A in 2007**

Ad Exchange: auctioneer-centric marketplace



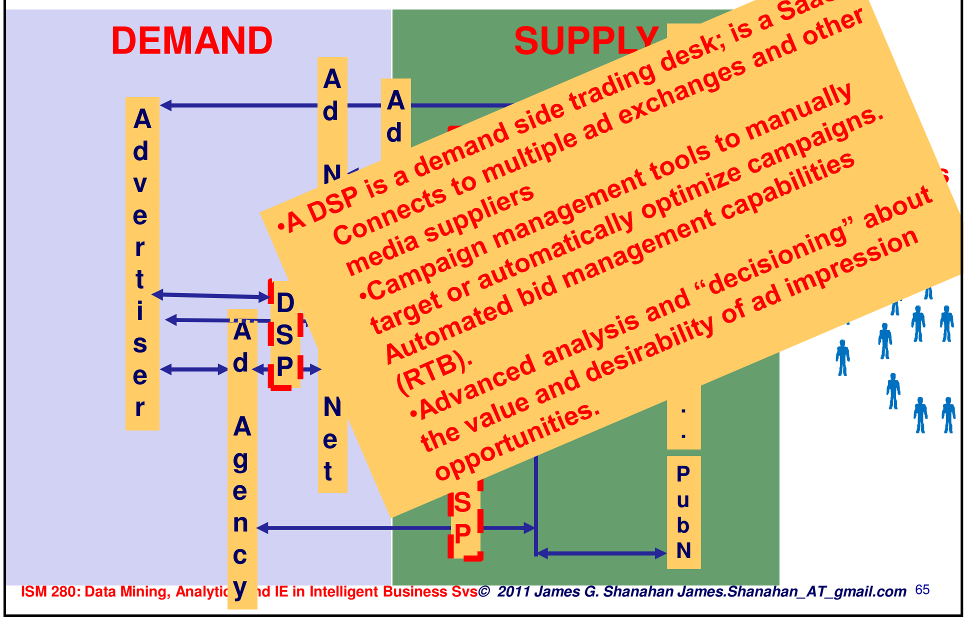
Demand-Side Platform: A trading desk for Adv.



Demand side (advertiser-side) platforms

[E.g., Turn, DataXu]

Demand Side, Supply side platforms



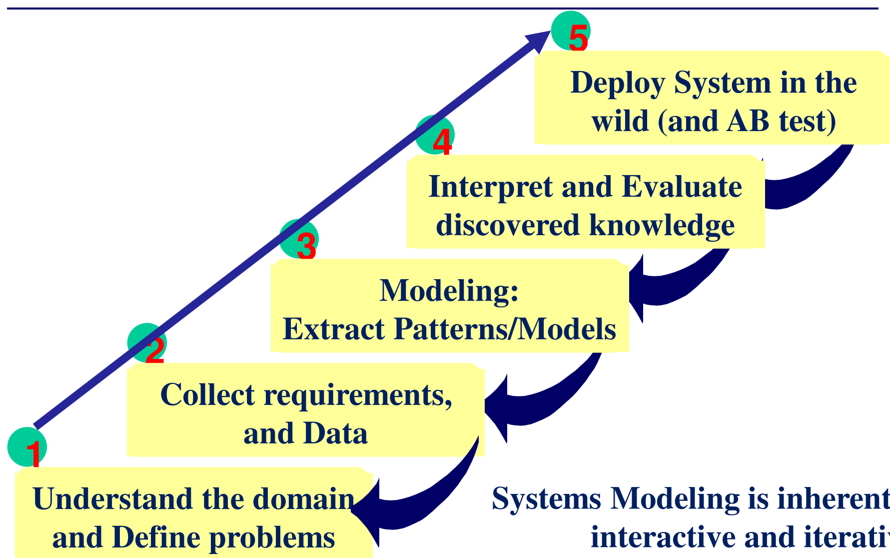
Key Features of DSP

- **Advanced and accurate audience targeting capabilities**
- **Easy-to-use inventory control**
- **Bidding dashboards**
- **Ability to set frequency caps on the ads being served**
  - reaching the "right consumer" too many times can lead to a significant decline in interest

Rest of Lecture 1 Outline

- **Background:**
  - Information extraction vs information retrieval
- **Advertising 101 and Digital advertising**
  - Predicting CTR
- **Information Extraction Overview**
- **Sentiment Analysis**
- **Candidate Project**

Machine Learning in Practice



Systems Modeling is inherently interactive and iterative

Generalized 2nd Price (GSP) Auction

1. In a GSP, multiple items are up for auction;
2. The highest bidder wins the first item at the second price (+delta)
3. The second-highest bidder wins the third-highest price, and so on

Bid = $10 PPC = $5	[Mine Text **Data**](#) Analyze Consumer Opinions Categorize Issues Automatically www.clarabridge.com
Bid = $5 PPC = $2	[Open Source **Data Mining**](#) Supercharged PostgreSQL Database 30 Days Free Support, Download Now! www.greenplum.com
Bid = $2 PPC = $1	[Easy **Data Mining**](#) Discover a **data mining** system that easily exports **data** to Excel. **Data**watch.iresponse.net
Bid = $1 PPC = $0.57	[**Data Mining** Software](#) Discover insights hidden in your existing **data** using SPSS solutions. www.spss.com

Introduced by Google in Feb 2002 (AdWords); overcomes the instability of GFP because by design the bidder is incentivized to pay the true value?!

ECPM-based rankg and payment for CPC

- **Ranks ads based on Expected-Revenue$_{Ad}$ (aka ECPM)**
  - Google, MSN and, as of 2/2007, Yahoo use ECPM-based ranking

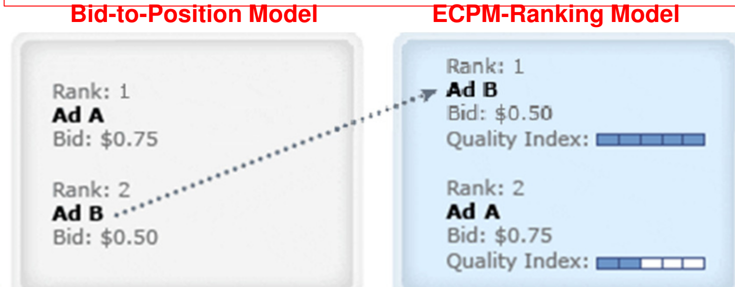
$$ECPM_{Ad} = CTR_{Ad} * Bid_{Ad}$$

$$ECPM_{Ad} = AdQualityIndex_{Ad} * Bid_{Ad}$$

PAY

$$\boxed{CPC_{Ad@i} = \frac{AdQualityIndex_{Ad@i+1}}{AdQualityIndex_{Ad@i}} * Bid_{Ad@i+1}}$$

Bid-to-Position Model **ECPM-Ranking Model**



CPC Calculation

Payoff = Value – Price
Payoff = ValuePerClick – CPC

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

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

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

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

Quality Score helps avoid Ad Spam

- **Quality Score can prohibit advertisers from simply bidding high enough to show in the top position.**
- **E.g., Below, Cameron is bidding well above all of his competitors, he will show in the fourth position due to his low Quality Score.**
- **Determining Click Cost:**
  - $ChargeToAdvertiser_i = (AdQuality_{i+1} / AdQuality_i)^* (Bid_{i+1}) + \$.01$
  - E.g., 1.6/10 + 0.1 = $0.17 Cost for the Mark (ad at ranked 1)

Rank by ECPM

Advertiser	Max CPC	Quality Score	AdRank	Position	Actual CPC
Mallory	$0.40	4	$0.4 x 4 = **1.6**	2	(1.2 / 4) + $.01 = $.31
Mark	$0.50	10	$0.50 x 10 = **5**	1	(1.6 / 10) + $.01 = $.17
Laura	$0.20	6	$0.20 x 6 = **1.2**	3	(1 / 6) + $.01 = $.17
Cameron	$2.00	0.5	$2.00 x .5 = **1**	4	(.8 / .5) + $.01 = $1.61
Alison	$0.05	16	$.05 x 16 = **.8**	5	(.2 / 2) + $.01 = $.11
Will	$0.10	2	$0.10 x 2 = **.2**	6	Minimum Bid

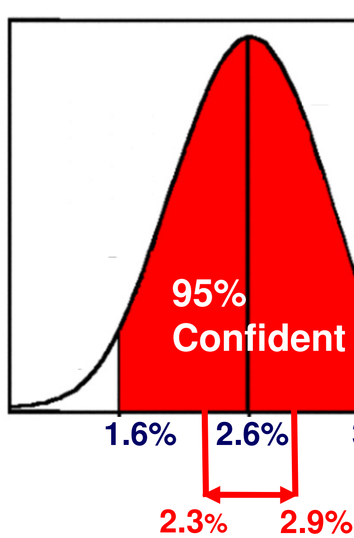
Accurate CTR Estimates are Crucial

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

- **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)



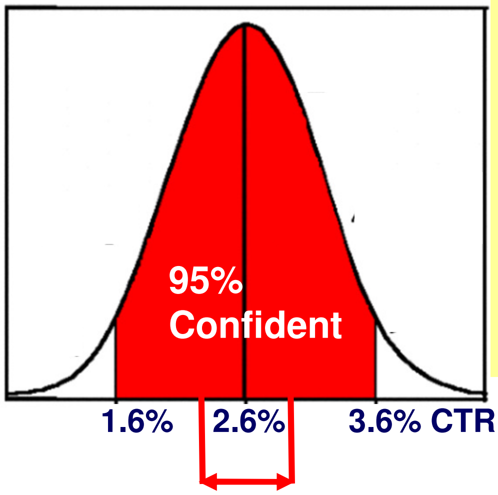
Estimate using Binomial MLE Estimates
I.e., #Clicks/#Impression

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

CTR (after 1,000 impressions)

(after 10,000 impressions)

Estimating CTR (and later AR)



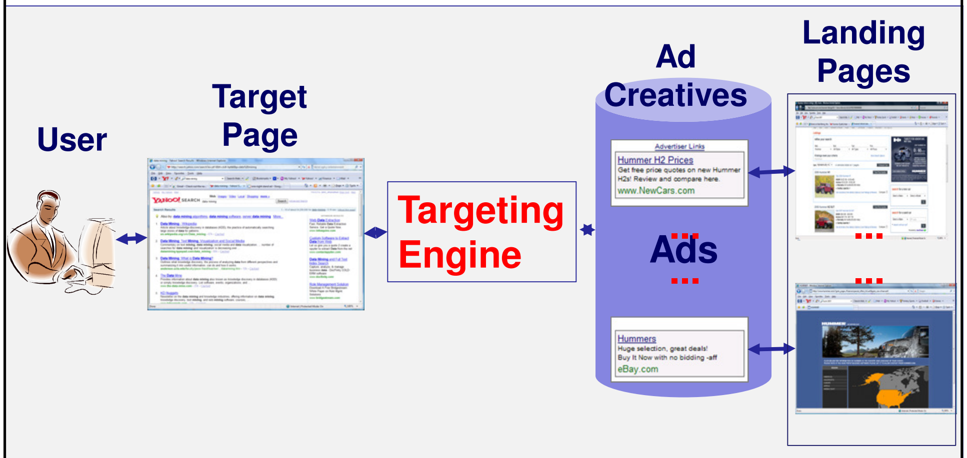
For a network of
~10^9 target pages,
~10^6 ads
~10^7 users
……
- Cannot afford this evaluation/auditioning
- Borrow strength, marginalize
- CoD (curse of $\text{dim}^{\text{ality}}$)

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!!

Ranking Ads using IR



Features

View TP
as query

IR
Engine

View Ad as document
with different sections
Keywords
Title + Description +URL
Landing Page

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 (P… decision trees using stochastic gradient de… [Friedman 2000])
  - Esteban Feuerstein, Pablo Heiber, Ja… …tinez-Viademonte and Ricardo Baeza-Y… …ew 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

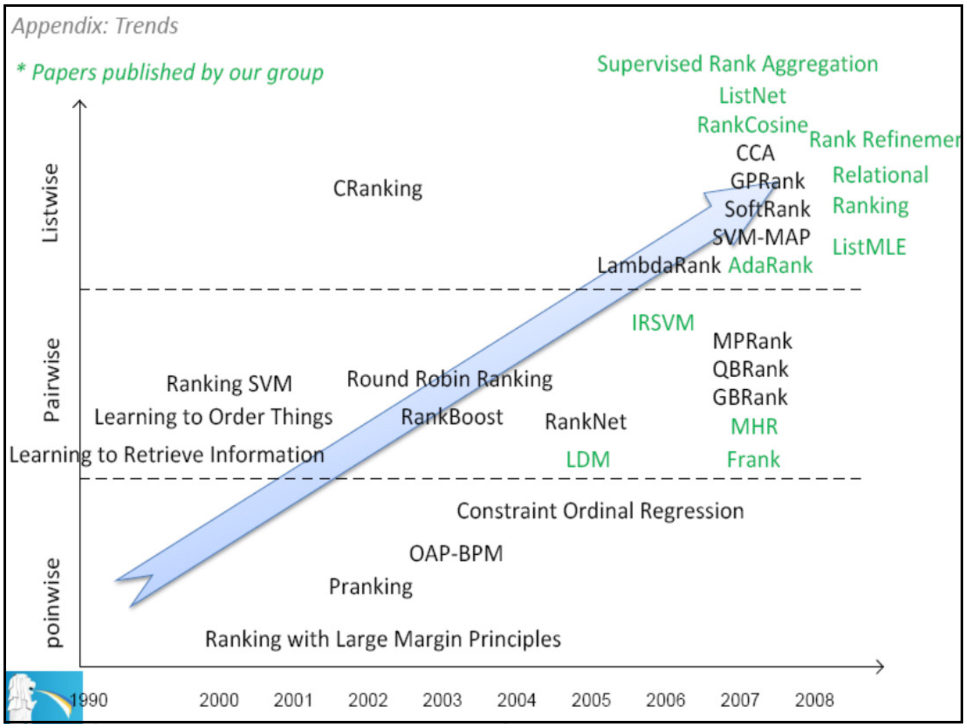
$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i})}}.$$

- **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

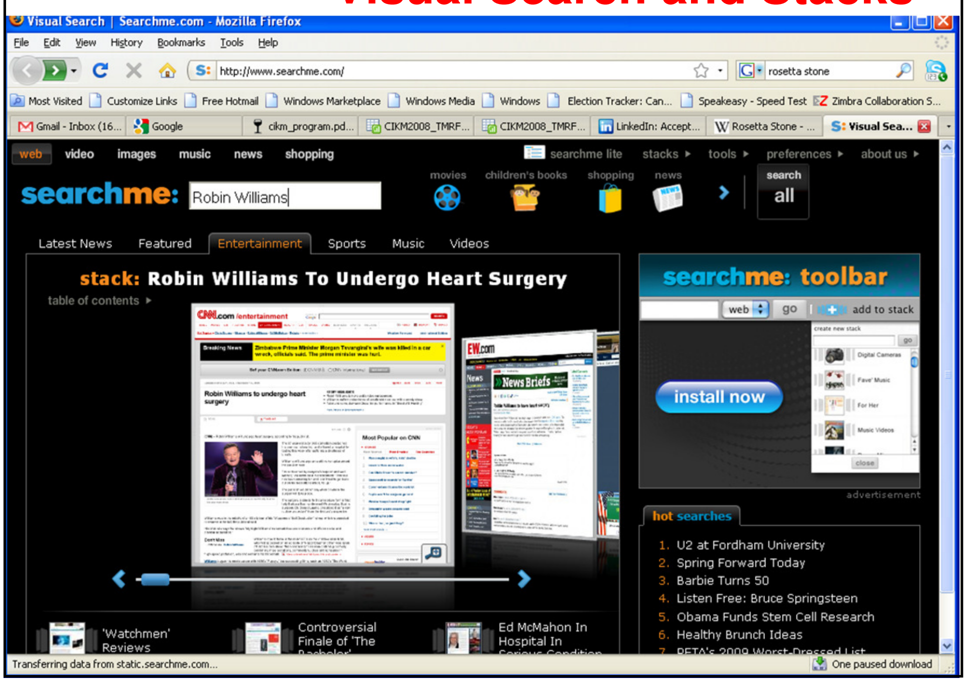
[Richardson et al., 2007]

Exercise

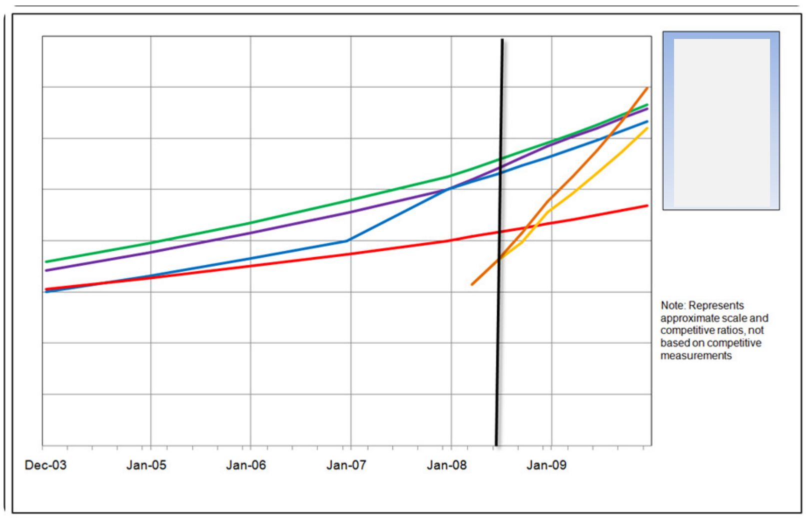
- **What are other features that could be used to modeling CTR prediction in a mobile setting?**
- **What are the three primary business models for advertising?**
- **Explain the differences between them from a publisher's perspective and from an advertiser's perspective**
- **What is the dominant business model in sponsored search?**
- **What is ECPM –based ranking? What is a key component of ECPM? How does high variance effect the publish and the advertiser?**

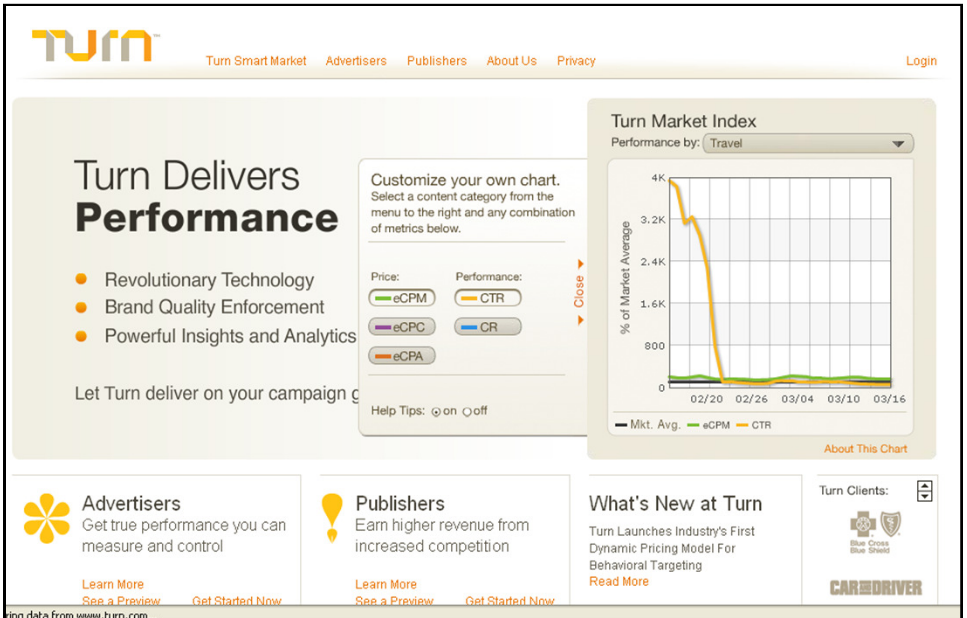


Visual Search and Stacks



Tracking Progress is key thru evaluations and metrics





Digital Advertising: Open research Areas

- **Forecasting**
- **Segmentation**
- **Prediction**
- **Ranking**
- **Allocation**
- **Targeting**
- **Mechanism design**
- **Realtime bidding**
- **Largescale distributed systems**
  - 100millisecond decisioning (Billions per day)

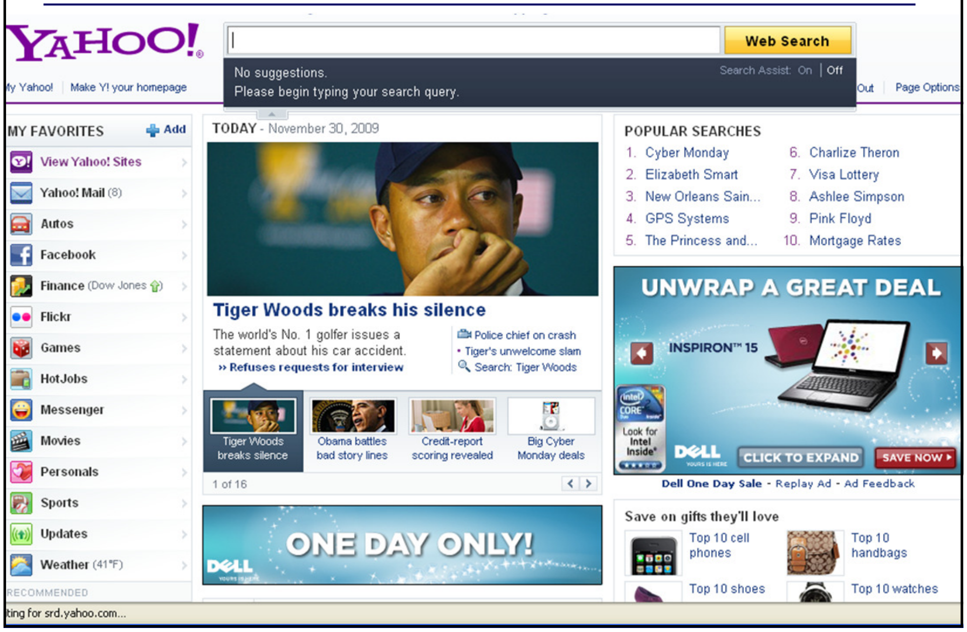
Text Classification
CTR Prediction
Clustering
Parsing/Extraction
Summarization/Mining
Ranking

Rest of Lecture 1 Outline

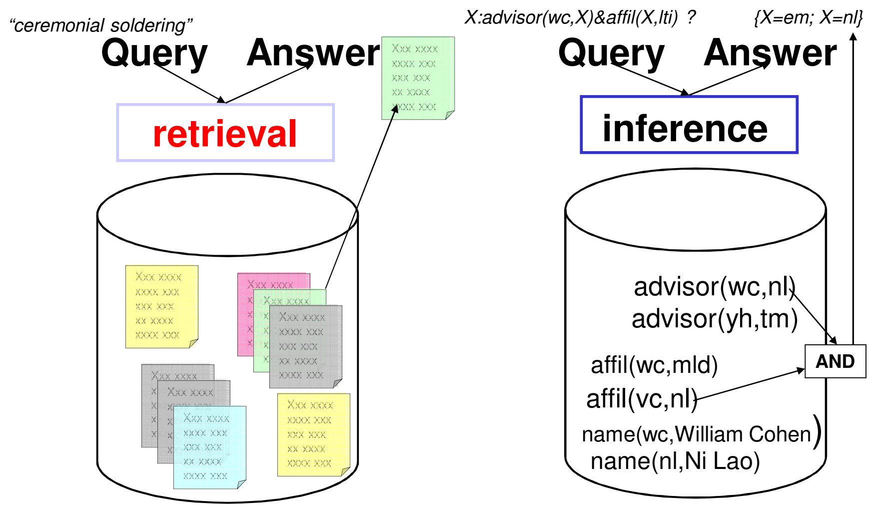
- **Background:**
  - Information extraction vs information retrieval
- **Advertising 101 and Digital advertising**
  - Predicting CTR
- **Information Extraction Overview**
- **Sentiment Analysis**
- **Candidate Project**

- Climbing the NLP foodchain
- Information Extraction
- Sentiment

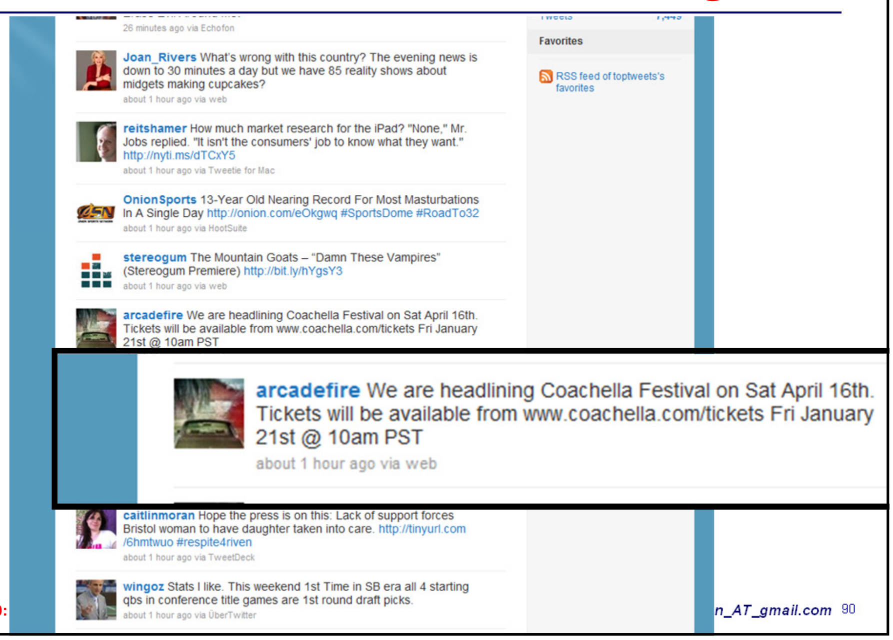
Over 300-400 applications of ML on this page



Two ways to manage information



Understanding Text



How do you extract information?

[Cohen / McCallum tutorial, NIPS 2002, KDD 2003, …]

[Some pilfering from Tom Mitchell's and William Cohen (CMU)invited talks]

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

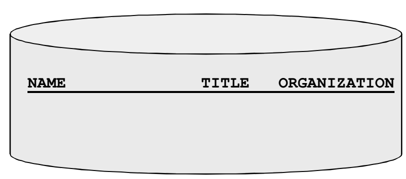
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

IE

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

IE

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

QA

End User

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

aka "named entity extraction"

What is "Information Extraction"

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

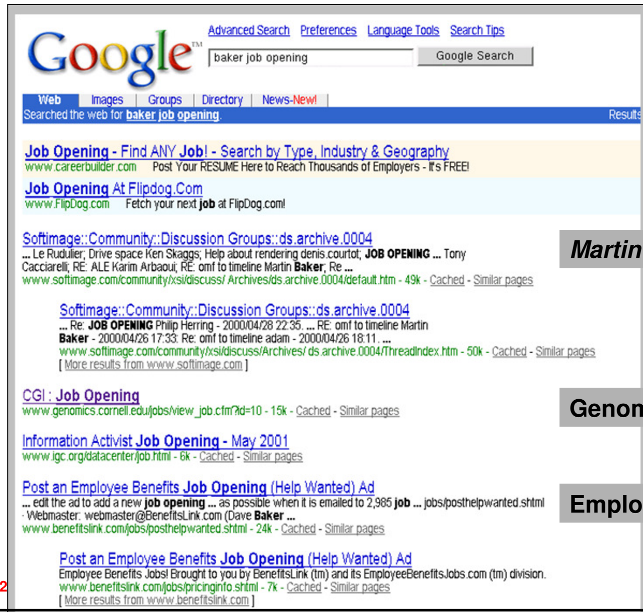
* Microsoft

Bill Veghte
* Microsoft
VP

Richard Stallman
founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Example: Finding Jobs Ads on the Web

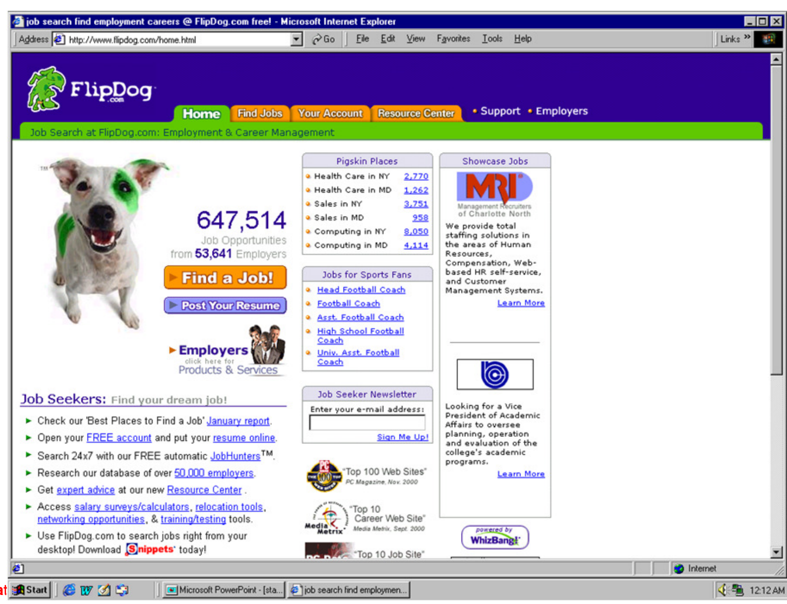


***Martin Baker*, a person**

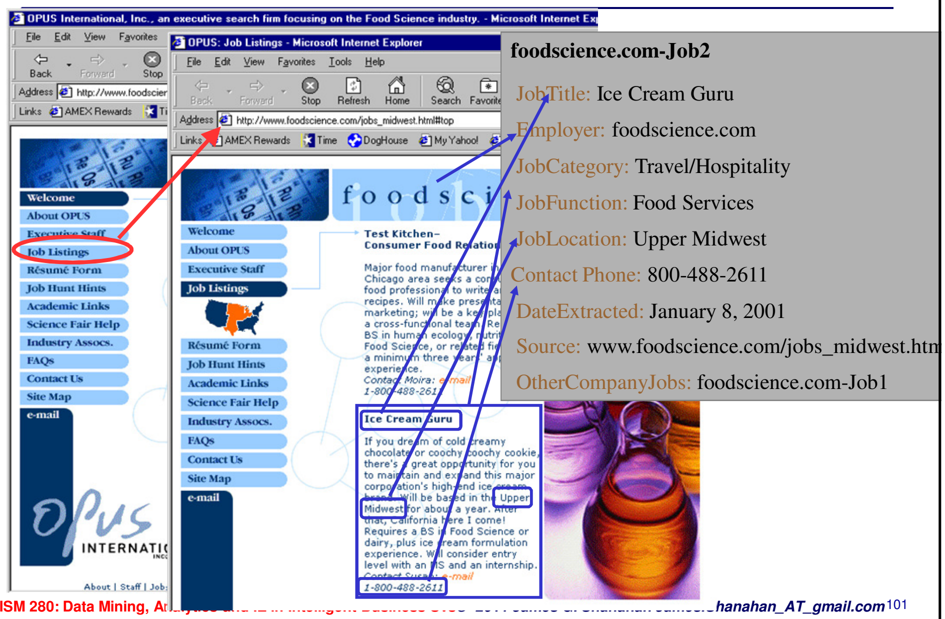
Genomics job

Employers job posting form

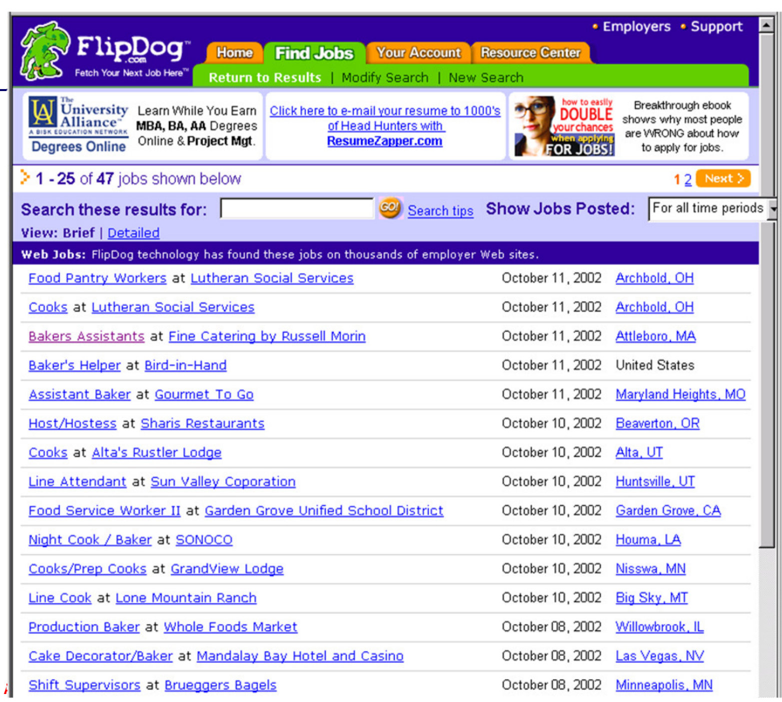
Example: A Solution



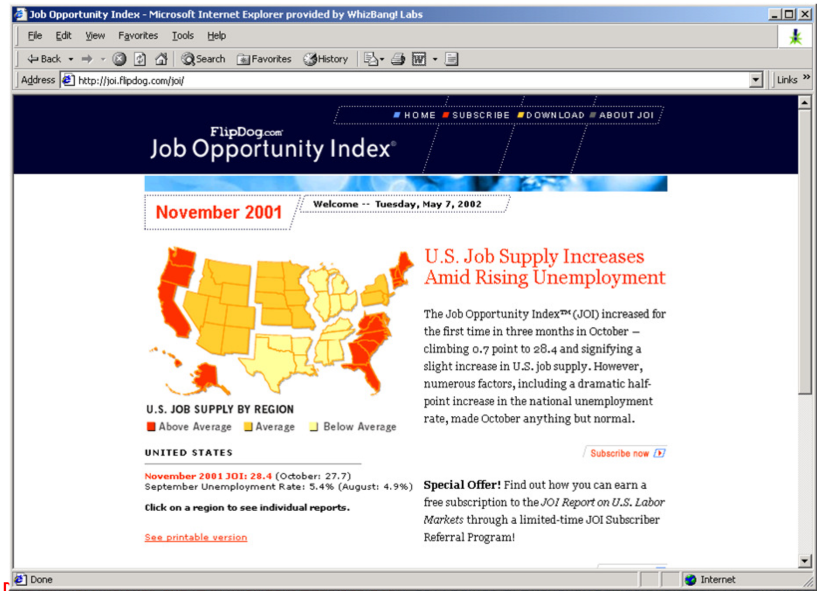
Extracting Job Openings from the Web

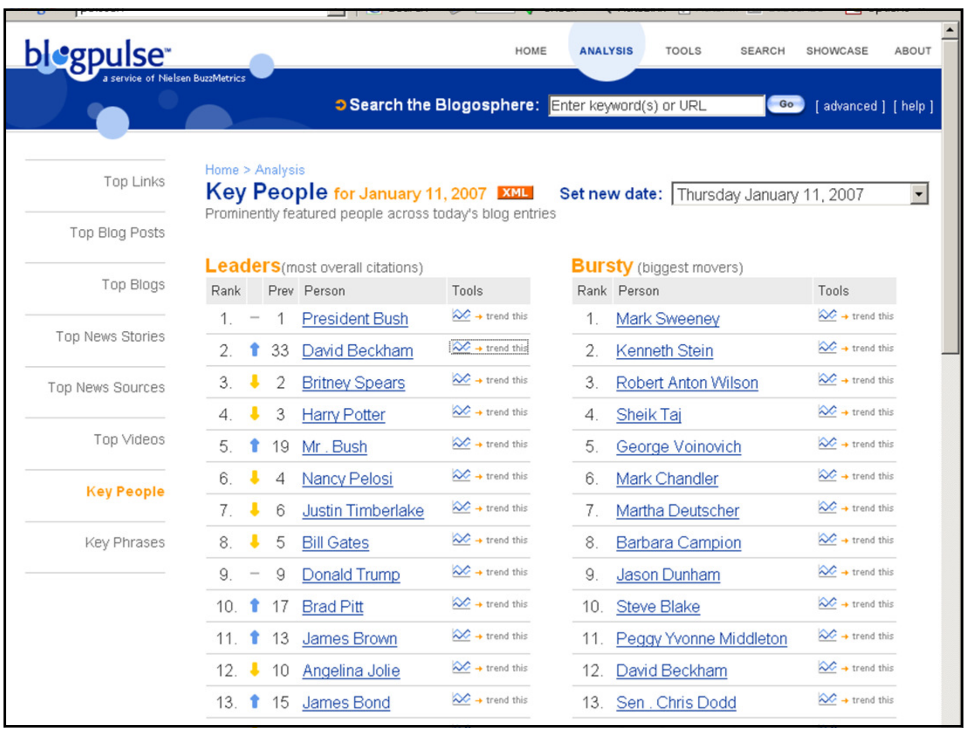


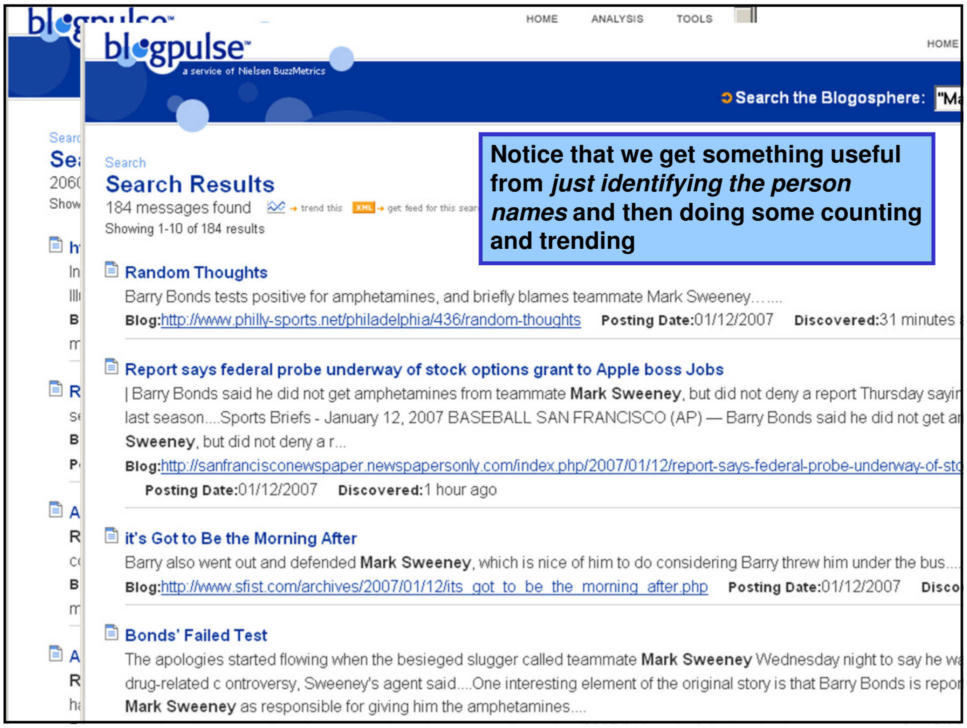
Job Openings:
Category = Food Services
Keyword = Baker
Location = Continental U.S.

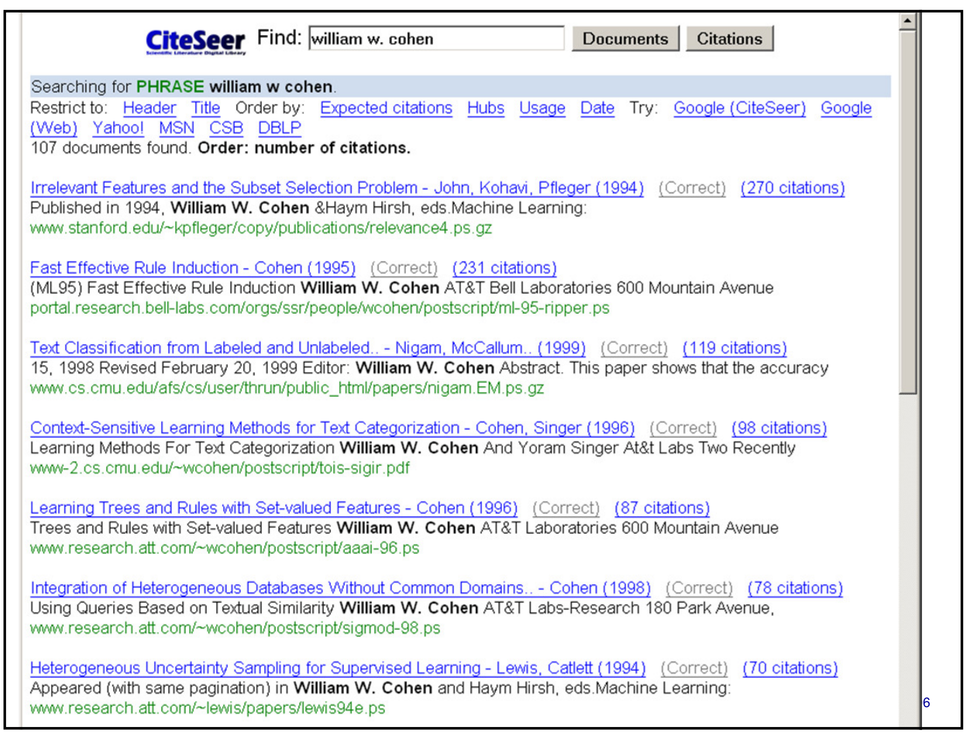


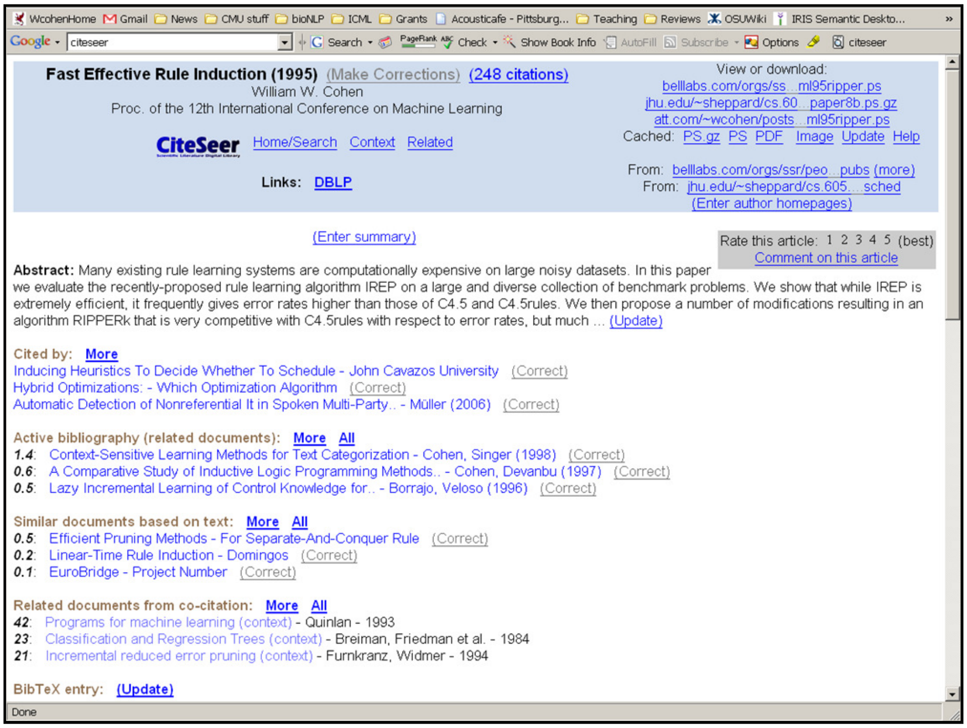
Data Mining the Extracted Job Information

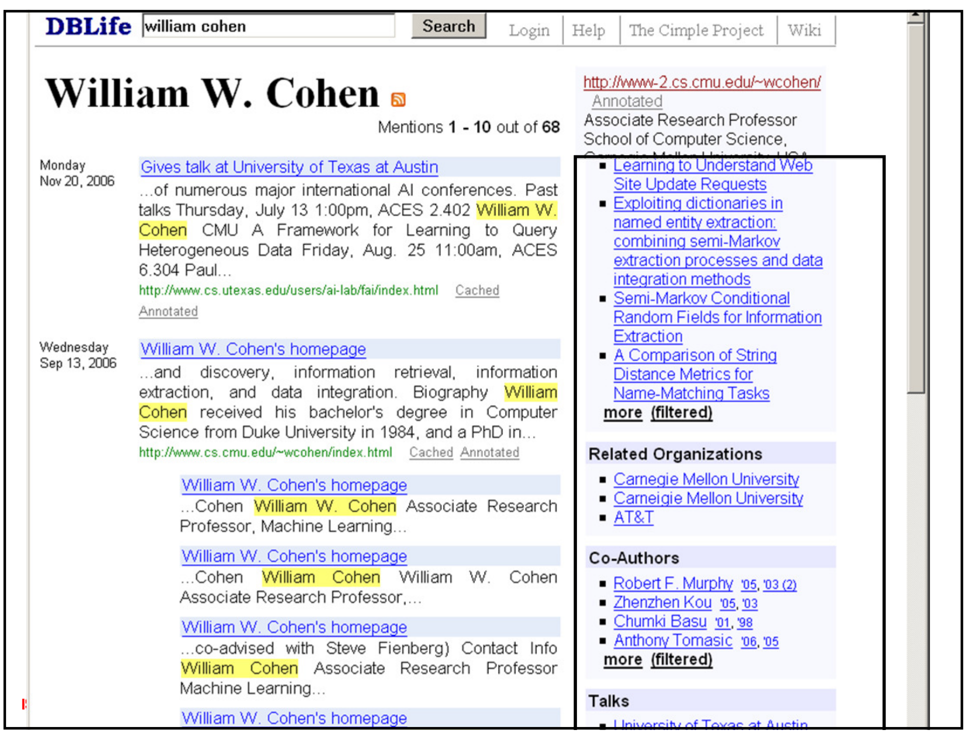





blogpulse
a service of Nielsen BuzzMetrics
HOME ANALYSIS TOOLS
Search the Blogosphere:
Search
Search Results
184 messages found
Showing 1-10 of 184 results
Notice that we get something useful from *just identifying the person names* and then doing some counting and trending
Random Thoughts
Barry Bonds tests positive for amphetamines, and briefly blames teammate Mark Sweeney.......
Blog:http://www.philly-sports.net/philadelphia/436/random-thoughts Posting Date:01/12/2007 Discovered:31 minutes
Report says federal probe underway of stock options grant to Apple boss Jobs
| Barry Bonds said he did not get amphetamines from teammate Mark Sweeney, but did not deny a report Thursday sayin
last season....Sports Briefs - January 12, 2007 BASEBALL SAN FRANCISCO (AP) — Barry Bonds said he did not get a
Sweeney, but did not deny a r...
Blog:http://sanfranciscoNewspaper.newspapersonly.com/index.php/2007/01/12/report-says-federal-probe-underway-of-sto
Posting Date:01/12/2007 Discovered:1 hour ago
it's Got to Be the Morning After
Barry also went out and defended Mark Sweeney, which is nice of him to do considering Barry threw him under the bus....
Blog:http://www.sfist.com/archives/2007/01/12/its_got_to_be_the_morning_after.php Posting Date:01/12/2007
Bonds' Failed Test
The apologies started flowing when the besieged slugger called teammate Mark Sweeney Wednesday night to say he wa
drug-related c ontroversy, Sweeney's agent said....One interesting element of the original story is that Barry Bonds is repor
Mark Sweeney as responsible for giving him the amphetamines....


CiteSeer Find: william w. cohen Documents Citations
Searching for PHRASE william w cohen.
Restrict to: Header Title Order by: Expected citations Hubs Usage Date Try: Google (CiteSeer) Google (Web) Yahoo! MSN CSB DBLP
107 documents found. Order: number of citations.
Irrelevant Features and the Subset Selection Problem - John, Kohavi, Pfleger (1994) (Correct) (270 citations)
Published in 1994, William W. Cohen &Haym Hirsh, eds.Machine Learning:
www.stanford.edu/~kpfleger/copy/publications/relevance4.ps.gz
Fast Effective Rule Induction - Cohen (1995) (Correct) (231 citations)
(ML95) Fast Effective Rule Induction William W. Cohen AT&T Bell Laboratories 600 Mountain Avenue
portal.research.bell-labs.com/orgs/ssr/people/wcohen/postscript/ml-95-ripper.ps
Text Classification from Labeled and Unlabeled.. - Nigam, McCallum.. (1999) (Correct) (119 citations)
15, 1998 Revised February 20, 1999 Editor: William W. Cohen Abstract. This paper shows that the accuracy
www.cs.cmu.edu/afs/cs/user/thrun/public_html/papers/nigam.EM.ps.gz
Context-Sensitive Learning Methods for Text Categorization - Cohen, Singer (1996) (Correct) (98 citations)
Learning Methods For Text Categorization William W. Cohen And Yoram Singer At&t Labs Two Recently
www-2.cs.cmu.edu/~wcohen/postscript/tois-sigir.pdf
Learning Trees and Rules with Set-valued Features - Cohen (1996) (Correct) (87 citations)
Trees and Rules with Set-valued Features William W. Cohen AT&T Laboratories 600 Mountain Avenue
www.research.att.com/~wcohen/postscript/aaai-96.ps
Integration of Heterogeneous Databases Without Common Domains.. - Cohen (1998) (Correct) (78 citations)
Using Queries Based on Textual Similarity William W. Cohen AT&T Labs-Research 180 Park Avenue,
www.research.att.com/~wcohen/postscript/sigmod-98.ps
Heterogeneous Uncertainty Sampling for Supervised Learning - Lewis, Catlett (1994) (Correct) (70 citations)
Appeared (with same pagination) in William W. Cohen and Haym Hirsh, eds.Machine Learning:
www.research.att.com/~lewis/papers/lewis94e.ps


WcohenHome Gmail News CMU stuff bioNLP ICML Grants Acousticafe - Pittsburg... Teaching Reviews OSUWiki IRIS Semantic Deskto...
Google citeseer Search PageRank Check Show Book Info AutoFill Subscribe Options citeseer
Fast Effective Rule Induction (1995) (Make Corrections) (248 citations)
William W. Cohen
Proc. of the 12th International Conference on Machine Learning
CiteSeer Home/Search Context Related
Links: DBLP
View or download:
belllabs.com/orgs/ss...ml95ripper.ps
jhu.edu/~sheppard/cs.60...paper8b.ps.gz
att.com/~wcohen/posts...ml95ripper.ps
Cached: PS.gz PS PDF Image Update Help
From: belllabs.com/orgs/ssr/peo...pubs (more)
From: jhu.edu/~sheppard/cs.605...sched
(Enter author homepages)
(Enter summary)
Rate this article: 1 2 3 4 5 (best)
Comment on this article
Abstract: Many existing rule learning systems are computationally expensive on large noisy datasets. In this paper we evaluate the recently-proposed rule learning algorithm IREP on a large and diverse collection of benchmark problems. We show that while IREP is extremely efficient, it frequently gives error rates higher than those of C4.5 and C4.5rules. We then propose a number of modifications resulting in an algorithm RIPPERk that is very competitive with C4.5rules with respect to error rates, but much ... (Update)
Cited by: More
Inducing Heuristics To Decide Whether To Schedule - John Cavazos University (Correct)
Hybrid Optimizations: - Which Optimization Algorithm (Correct)
Automatic Detection of Nonreferential It in Spoken Multi-Party.. - Müller (2006) (Correct)
Active bibliography (related documents): More All
1.4: Context-Sensitive Learning Methods for Text Categorization - Cohen, Singer (1998) (Correct)
0.6: A Comparative Study of Inductive Logic Programming Methods.. - Cohen, Devanbu (1997) (Correct)
0.5: Lazy Incremental Learning of Control Knowledge for.. - Borrajo, Veloso (1996) (Correct)
Similar documents based on text: More All
0.5: Efficient Pruning Methods - For Separate-And-Conquer Rule (Correct)
0.2: Linear-Time Rule Induction - Domingos (Correct)
0.1: EuroBridge - Project Number (Correct)
Related documents from co-citation: More All
42: Programs for machine learning (context) - Quinlan - 1993
23: Classification and Regression Trees (context) - Breiman, Friedman et al. - 1984
21: Incremental reduced error pruning (context) - Furnkranz, Widmer - 1994
BibTeX entry: (Update)
Done


DBLife william cohen Search Login Help The Cimple Project Wiki
William W. Cohen
Mentions 1 - 10 out of 68
Monday Nov 20, 2006
Gives talk at University of Texas at Austin
...of numerous major international AI conferences. Past talks Thursday, July 13 1:00pm, ACES 2.402 William W. Cohen CMU A Framework for Learning to Query Heterogeneous Data Friday, Aug. 25 11:00am, ACES 6.304 Paul...
http://www.cs.utexas.edu/users/ai-lab/fai/index.html Cached Annotated
Wednesday Sep 13, 2006
William W. Cohen's homepage
...and discovery, information retrieval, information extraction, and data integration. Biography William Cohen received his bachelor's degree in Computer Science from Duke University in 1984, and a PhD in...
http://www.cs.cmu.edu/~wcohen/index.html Cached Annotated
William W. Cohen's homepage
...Cohen William W. Cohen Associate Research Professor, Machine Learning...
William W. Cohen's homepage
...Cohen William Cohen William W. Cohen Associate Research Professor,...
William W. Cohen's homepage
...co-advised with Steve Fienberg) Contact Info William Cohen Associate Research Professor Machine Learning...
William W. Cohen's homepage
http://www-2.cs.cmu.edu/~wcohen/
Annotated
Associate Research Professor
School of Computer Science,
Learning to Understand Web Site Update Requests
Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods
Semi-Markov Conditional Random Fields for Information Extraction
A Comparison of String Distance Metrics for Name-Matching Tasks
more (filtered)
Related Organizations
Carnegie Mellon University
Carneigie Mellon University
AT&T
Co-Authors
Robert F. Murphy '05, '03 (2)
Zhenzhen Kou '05, '03
Chumki Basu '01, '98
Anthony Tomasic '06, '05
more (filtered)
Talks

Sunita's Breakdown of IE

- **What's the end goal (application?)**
- **What's the input (corpus)? How is it preprocessed? How is output postprocessed (to make querying easier)?**
- **What structure is extracted?**
  - Entity names? *("William Cohen, "Anthony 'Van' Jones")*
  - Relationships between entities? *("Richard Wang" studentOf "William Cohen")*
  - Features/properties/adjectives describing entities? *("iPhone 3G" → "expensive service plan", "color screen")*
- **What (learning) methods are used?**

Landscape of IE Tasks (1/4): Degree of Formatting

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Grammatical sentences and some formatting & links

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

- Press

Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links

Barto, Andrew G. (413) 545-2109 barto@cs.umass.edu CS276
Professor.
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.

Berger, Emery D. (413) 577-4211 emery@cs.umass.edu CS344
Assistant Professor.

Brock, Oliver (413) 577-0334 oli@cs.umass.edu CS246
Assistant Professor.

Clarke, Lori A. (413) 545-1328 clarke@cs.umass.edu CS304
Professor.
Software verification, testing, and analysis; software architecture and design.

Cohen, Paul R. (413) 545-3638 cohen@cs.umass.edu CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

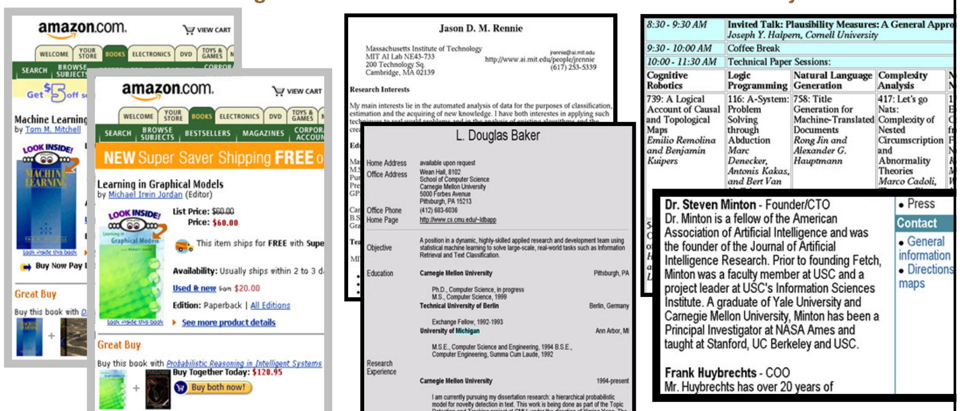
Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	**Logic Programming**	**Natural Language Generation**	**Complexity Analysis**	**Neural Networks**	**Games**
739: A Logical Account of Causal and Topological Maps *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz*	131: A Comparative Study of Logic Programs with	246: Dealing with Dependencies between Content Planning and	470: A Perspective on Knowledge Compilation	258: Violation-Guided Learning for Constrained	353: Temporal Difference Learning Applied to a

Landscape of IE Tasks (2/4): Intended Breadth of Coverage

If you were going to use formatting to do extraction, how often (with what granularity) would you have to re-train your models?

Web site specific	Genre specific	Wide, non-specific
Formatting	Layout	Language
Amazon.com Book Pages	Resumes	University Names



Landscape of IE Tasks (3/4): Complexity of extraction task

E.g. word patterns:

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Landscape of IE Tasks (4/4):
Single Field/Record

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: **Jack Welch**

Person: **Jeffrey Immelt**

Location: **Connecticut**

Binary relationship

Relation: **Person-Title**
Person: **Jack Welch**
Title: **CEO**

Relation: **Company-Location**
Company: **General Electric**
Location: **Connecticut**

N-ary record

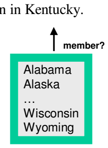
Relation: **Succession**
Company: **General Electric**
Title: **CEO**
Out: **Jack Welsh**
In: **Jeffrey Immelt**

"Named entity" extraction

Models for NER

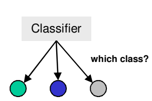
Lexicons

Abraham Lincoln was born in Kentucky.



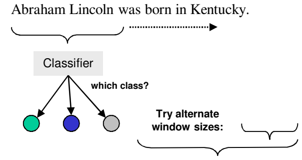
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.



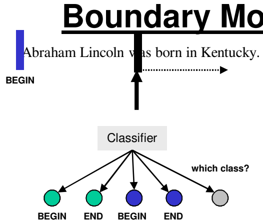
Sliding Window

Abraham Lincoln was born in Kentucky.



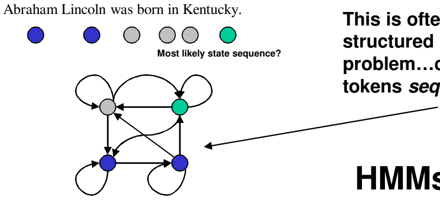
Boundary Models

Abraham Lincoln was born in Kentucky.



Token Tagging

Abraham Lincoln was born in Kentucky.



This is often treated as a structured prediction problem…classifying tokens *sequentially*

HMMs, CRFs, ….

Sliding Windows

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© *2011 James G. Shanahan James.Shanahan_AT_gmail.com* 115

Extraction by Sliding Window

E.g. Looking for seminar location

```
GRAND CHALLENGES FOR MACHINE LEARNING

         Jaime Carbonell
     School of Computer Science
    Carnegie Mellon University

             3:30 pm
           7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© *2011 James G. Shanahan James.Shanahan_AT_gmail.com* 116

Extraction by Sliding Window

E.g. Looking for seminar location

```
        GRAND CHALLENGES FOR MACHINE LEARNING

                    Jaime Carbonell
                School of Computer Science
               Carnegie Mellon University

                        3:30 pm
                     7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Extraction by Sliding Window

E.g. Looking for seminar location

```
        GRAND CHALLENGES FOR MACHINE LEARNING

                    Jaime Carbonell
                School of Computer Science
               Carnegie Mellon University

                        3:30 pm
                     7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Extraction by Sliding Window

E.g.
Looking for seminar location

```
        GRAND CHALLENGES FOR MACHINE LEARNING

               Jaime Carbonell
            School of Computer Science
           Carnegie Mellon University

                   3:30 pm
                7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

A "Naïve Bayes" Sliding Window Model

[Freitag 1997]

… 00 : pm Place : $\boxed{\text{Wean Hall Rm 5409}}$ Speaker : Sebastian Thrun …

$\underbrace{w_{t-m} \quad \ldots \quad w_{t-1}}_{\text{prefix}} \; \underbrace{w_t \quad \ldots \quad w_{t+n}}_{\text{contents}} \; \underbrace{w_{t+n+1} \quad \ldots \quad w_{t+n+m}}_{\text{suffix}}$

Estimate Pr(LOCATION|window) using Bayes rule

Try all "reasonable" windows (vary length, position)

Assume independence for length, prefix words, suffix words, content words

Estimate from data quantities like: Pr("Place" in prefix|LOCATION)
If P("Wean Hall Rm 5409" = LOCATION) **is above some threshold, extract it.**

A "Naïve Bayes" Sliding Window Model

[Freitag 1997]

… 00 : pm Place : Wean Hall Rm 5409 Speake

w_{t-m} … w_{t-1} w_t … w_{t+n} w_{t+n+1} … w_{t+n+m}

prefix — contents — suffix

1. **Create dataset of examples like these:**
   +(prefix00,…,prefixColon, contentWean,contentHall,….,suffixSpeaker,…)
   - (prefixColon,…,prefixWean,contentHall,….,ContentSpeaker,suffixColon,….)
   …
2. **Train a NaiveBayes classifier (or YFCL), treating the examples like BOWs for text classification**
3. **If Pr(class=+|prefix,contents,suffix) > threshold, predict the content window is a location.**
   - To think about: what if the extracted entities aren't consistent, eg if the location overlaps with the speaker?

"Naïve Bayes" Sliding Window Results

Domain: CMU UseNet Seminar Announcements

```
     GRAND CHALLENGES FOR MACHINE LEARNING

              Jaime Carbonell
          School of Computer Science
         Carnegie Mellon University

                  3:30 pm
               7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence during
the 1980s and 1990s.   As a result of its
success and growth, machine learning is
evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning), genetic
algorithms, connectionist learning, hybrid
systems, and so on.
```

Field	F1
Person Name:	30%
Location:	61%
Start Time:	98%

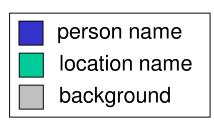
Token Tagging

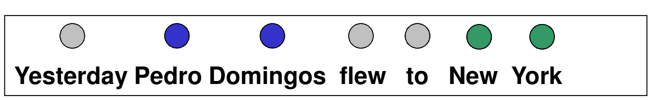
NER by tagging tokens

Given a sentence:

Yesterday Pedro Domingos flew to New York.

1) Break the sentence into *tokens*, and ***classify*** each token with a label indicating *what sort of entity* it's part of:





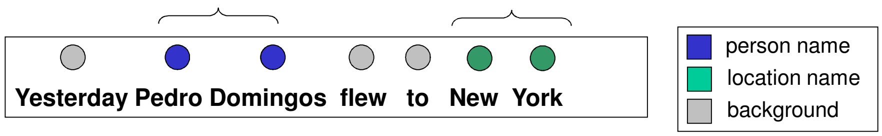
2) Identify names based on the entity labels

Person name: **Pedro Domingos**
Location name: **New York**

3) To learn an NER system, use YFCL.

NER by tagging tokens

Similar labels tend to *cluster together* in text



Another common labeling scheme is BIO (begin, inside, outside; e.g. beginPerson, insidePerson, beginLocation, insideLocation, outside)

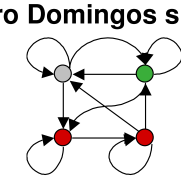
BIO also leads to *strong dependencies between nearby labels* (eg inside follows begin)

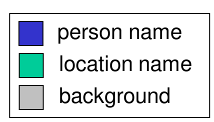
NER with Hidden Markov Models

Given a sequence of observations:

Yesterday Pedro Domingos spoke this example sentence.

and a trained HMM:





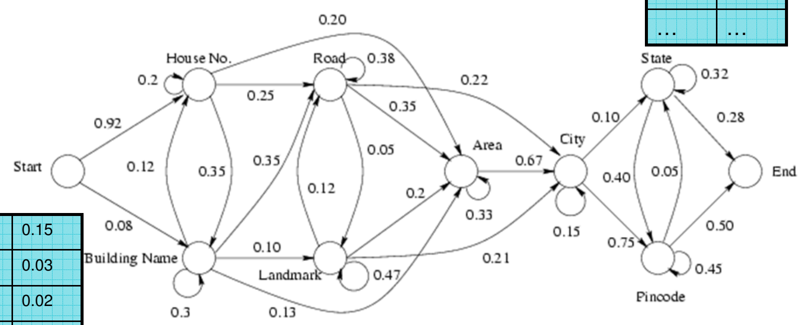
Find the most likely state sequence: (Viterbi) $\arg\max_{\vec{s}} P(\vec{s},\vec{o})$

Any words said to be generated by the designated "person name" state extract as a person name:

Person name: Pedro Domingos

HMM for Segmentation of Addresses

CA	0.15
NY	0.11
PA	0.08
…	…



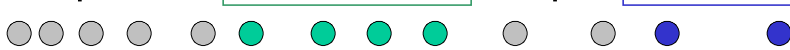
Hall	0.15
Wean	0.03
N-S	0.02
…	…

- **Simplest HMM Architecture: One state per entity type**

[Pilfered from Sunita Sarawagi, IIT/Bombay]

HMMs for Information Extraction

… 00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thr …



1. **The HMM consists of two probability tables**
   - *Pr(currentState=s|previousState=t)* for s=background, location, speaker,
   - *Pr(currentWord=w|currentState=s)* for s=background, location, …
2. **Estimate these tables with a (smoothed) CPT**
   - Prob(location|location) = #(loc->loc)/#(loc->*) transitions
3. **Given a new sentence, find the most likely sequence of hidden states using Viterbi method:**

MaxProb(curr=s|position k)=

$\text{Max}_{\text{state } t}$ MaxProb(curr=t|position=k-1) * Prob(word=w_{k-1}|t)*Prob(curr=s|prev=t)

"Naïve Bayes" Sliding Window vs HMMs

Domain: CMU UseNet Seminar Announcements

```
GRAND CHALLENGES FOR MACHINE LEARNING

Jaime Carbonell
School of Computer Science
Carnegie Mellon University

3:30 pm
7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence during
the 1980s and 1990s.   As a result of its
success and growth, machine learning is
evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning), genetic
algorithms, connectionist learning, hybrid
systems, and so on.
```

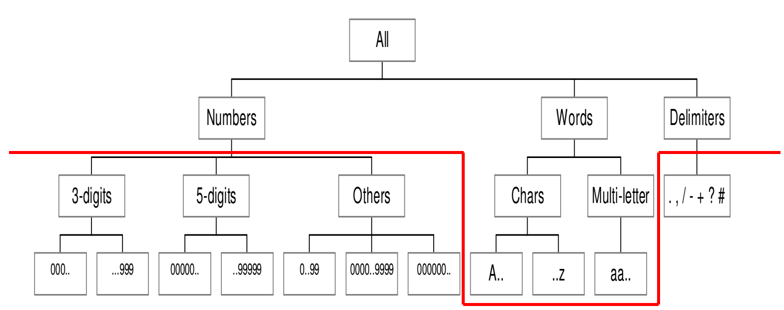
Field	F1
Speaker:	30%
Location:	61%
Start Time:	98%

Field	F1
Speaker:	77%
Location:	79%
Start Time:	98%

What is a "symbol" ???

Cohen => "Cohen", "cohen", "Xxxxx", "Xx", … ?

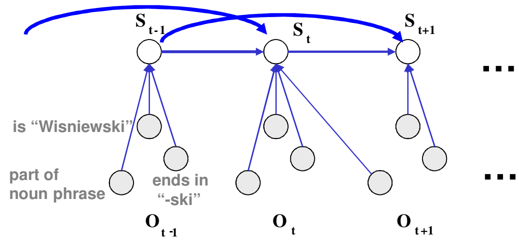
5317 => "5317", "9999", "9+", "number", … ?



Datamold: **choose best** abstraction level using **holdout** set

What is a symbol?

identity of word
ends in "-ski"
is capitalized
is part of a noun phrase
is in a list of city names
is under node X in WordNet
is in bold font
is indented
is in hyperlink anchor
…



Idea: replace **generative** model in HMM with a **maxent** model, where **state** depends on **observations** and **previous state** history

$$\Pr(s_t \mid x_t, s_{t-1}, s_{t-2}, \ldots) = \ldots$$

Ratnaparkhi's MXPOST

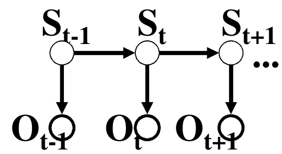
- Sequential learning problem: predict POS tags of words.
- Uses MaxEnt model described above.
- Rich feature set.
- To smooth, discard features occurring < 10 times.

Condition	Features	
w_i is not rare	$w_i = X$	$\& \; t_i = T$
w_i is rare	X is prefix of w_i, $\lvert X \rvert \leq 4$	$\& \; t_i = T$
	X is suffix of w_i, $\lvert X \rvert \leq 4$	$\& \; t_i = T$
	w_i contains number	$\& \; t_i = T$
	w_i contains uppercase character	$\& \; t_i = T$
	w_i contains hyphen	$\& \; t_i = T$
$\forall \; w_i$	$t_{i-1} = X$	$\& \; t_i = T$
	$t_{i-2} t_{i-1} = XY$	$\& \; t_i = T$
	$w_{i-1} = X$	$\& \; t_i = T$
	$w_{i-2} = X$	$\& \; t_i = T$
	$w_{i+1} = X$	$\& \; t_i = T$
	$w_{i+2} = X$	$\& \; t_i = T$

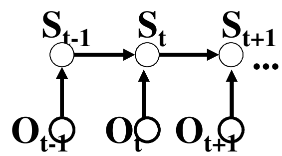
Table 1: Features on the current history h_i

Conditional Markov Models (CMMs) aka MEMMs aka Maxent Taggers *vs* HMMS

$$\Pr(s,o) = \prod_i \Pr(s_i \mid s_{i-1}) \Pr(o_i \mid s_{i-1})$$



$$\Pr(s \mid o) = \prod_i \Pr(s_i \mid s_{i-1}, o_{i-1})$$



HMMs vs MEMM vs CRF

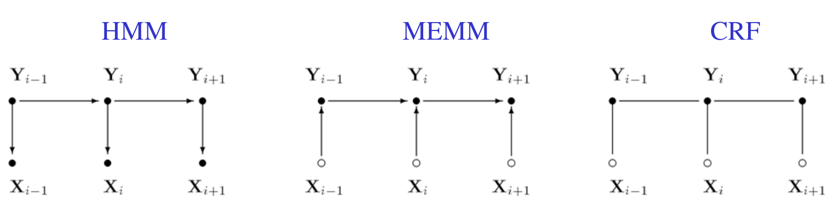


Figure 2. Graphical structures of simple HMMs (left), MEMMs (center), and the chain-structured case of CRFs (right) for sequences. An open circle indicates that the variable is not generated by the model.

Some things to think about

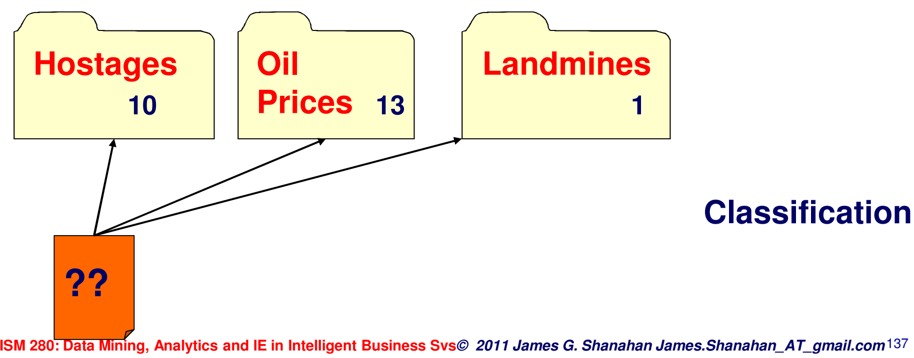
- **We've seen sliding windows, non-sequential token tagging, and sequential token tagging.**
  - Which of these are likely to work best, and when?
  - Are there other ways to formulate NER as a learning task?
  - Is there a benefit from using more complex graphical models? What potentially useful information does a linear-chain CRF not capture?
  - Can you combine sliding windows with a sequential model?
- **Next lecture will survey IE of sets of *related* entities (e.g., person and his/her affiliation).**
  - How can you formalize that as a learning task?

Rest of Lecture 1 Outline

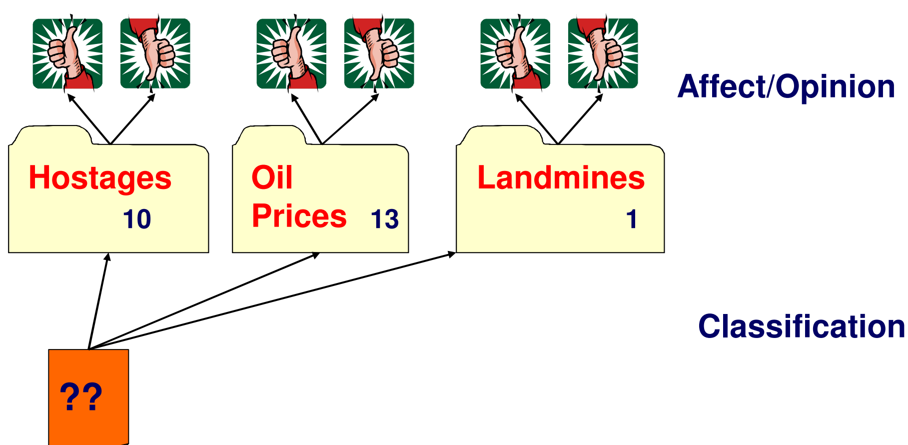
- **Background:**
  - Information extraction vs information retrieval
- **Advertising 101 and Digital advertising**
  - Predicting CTR
- **Information Extraction Overview**
- **Sentiment Analysis**
- **Candidate Project**

Theme 1: Text Classification

Thresholded SVMs



Theme 2: Affect/Opinion



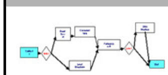
Ranking in Web Search

- **Ranking Is The Key**
- **Ideal ranking function: Relevance + Quality**
- **Relevance (query dependent)**
  - TF, IDF
  - Title, Body, Anchor, URL
  - Proximity
  - …
- **Quality**
  - PageRank
  - PageQuality, Spam
  - …
- **SVM-MAP**

Talk Outline

- **Good Classification Technology**
  - Thresholded SVMs

- **Extra semantic processing**
  - Affect/Opinion



- **Process Mining**
  - Bayesian Network Approach

Outline: Opinion Mining

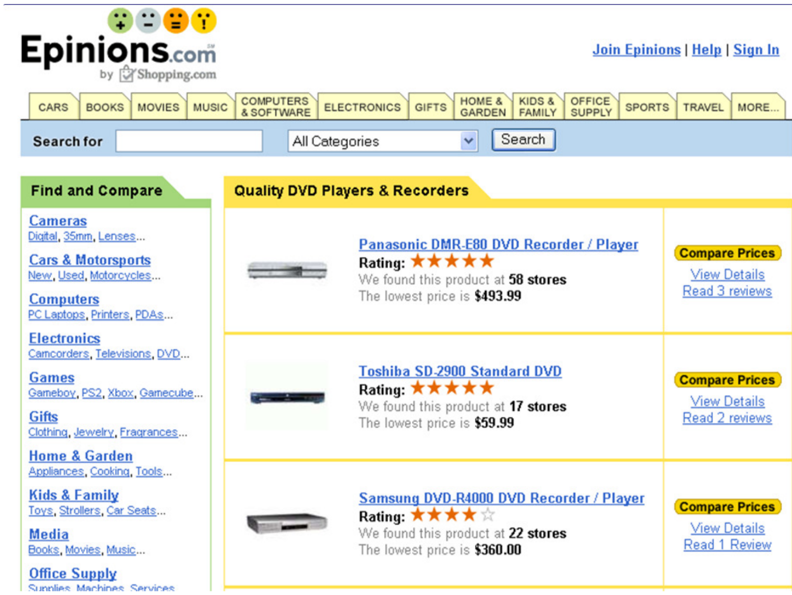
- **Motivation**
- **Review Main Approaches**
- **Evaluation and Application**
- **Conclusions**

Opinion Mining

Motivation and Background

- **Current information management systems operate at a low level with only some semantics**
- **Much of product feedback is web-based**
  - provided by customers/critiques online through websites, discussion boards, mailing lists, and blogs, CRM Portals.
- **Market research is becoming unwieldy**
  - Sources are heterogeneous and, increasingly, multilingual in nature

Examples of Opinion on WWW



Amazon.co.jp

Affect in a Reporting Point of View

"Microsoft Togetherness"
Economist, January 22–28th, 2000, Business

There is both more and less than meets the eye to the decision of Bill Gates to pass the chief executive's mantle to his best friend, Steve Ballmer. It is still business as usual at the world's biggest software company. … Nor does the move presage a change in strategy. A belligerent Mr Ballmer reaffirmed the company's hardline approach to defending the continuing antitrust action, predictably describing the break-up of the company that the government is rumoured to favour as reckless and irresponsible. Although Mr Gates spoke excitedly about Next Generation Windows Services (NGWS), a new idea that he would be working on, it is, in effect, just an ugly umbrella name for the grand Internet strategy under development at Redmond for some time. …

CRM: Support Desk Inquiries

I spoke today with an hp technican and he really upset me.
He told me that sj 4100 (usb) will be not supported.
There won't be any patches.
Can someone confirm that because I'm really pissed off.

Related Work

- **Scoring Reviews**
  - "Cold Start Recommendations" (Schein *et al.*, 2002)
  - "Thumbs Up Thumbs Down" (Turney, ACL, 2002)
  - "Mining Peanut Gallery" (Kushal *et al.*, 2003)
  - "Measuring Praise/Criticism" (Turney & Littman, 2003)
- **Affect/Opinion Detection**
  - *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications AAAI-EAAT, Stanford March 2004*
  - *SIGIR Workshop 2005*
- **Niche Browsers**
  - Citeseer (Lawrence e*t al.*, 1999)
  - PROGENIE (Duboué *et al.*, 2003)
  - HPSearch

[SIGIR 2005](#) Workshop

Stylistic Analysis Of Text For Information Access

August 19, 2005
Salvador, Bahia, Brazil

[Theme and goals of the workshop](#)

[Submission guidelines](#)

[Important dates](#)

[Organizers](#)

[Program committee](#)

Theme and goals of the workshop

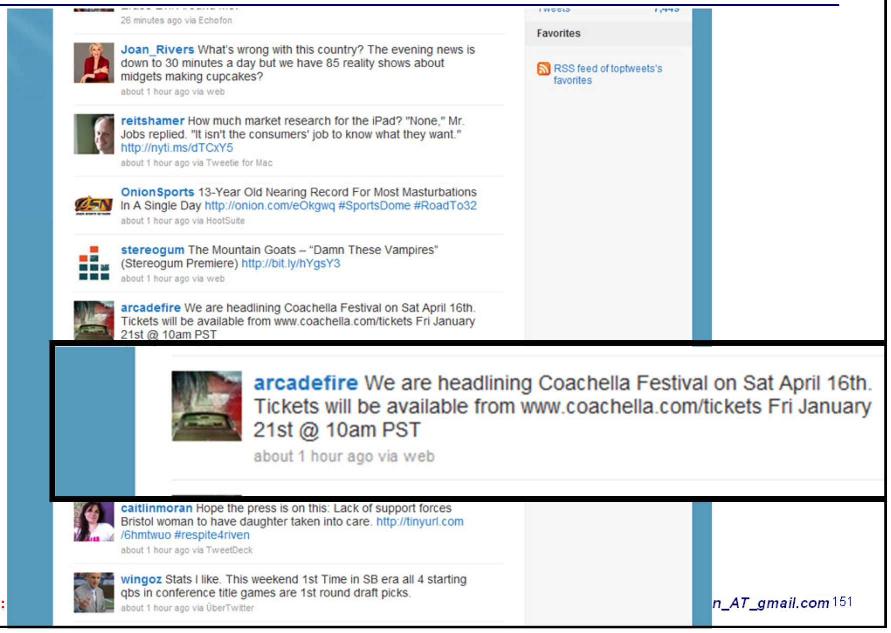
Information management systems have typically focused on the "factual" aspect of content analysis. Other aspects, including pragmatics, opinion, and style, have received much less attention. However, to achieve an adequate understanding of a text, these aspects cannot be ignored.

This workshop will be the first ever to specifically address the automatic analysis and extraction of stylistic aspects of natural language texts for purposes of improving information access. Style may be roughly defined as the 'manner' in which something is expressed, as opposed to the 'content' of a message. Stylistic variation depends on author preferences and competence, familiarity, genre, communicative context, expected characteristics of the intended audience and untold other factors, and it is expressed through subtle variation in frequencies of otherwise insignificant features of a text that, taken together, are understood as stylistic indicators by a particular reader community. Modeling, representing, and utilizing this variation is the business of stylistic analysis.

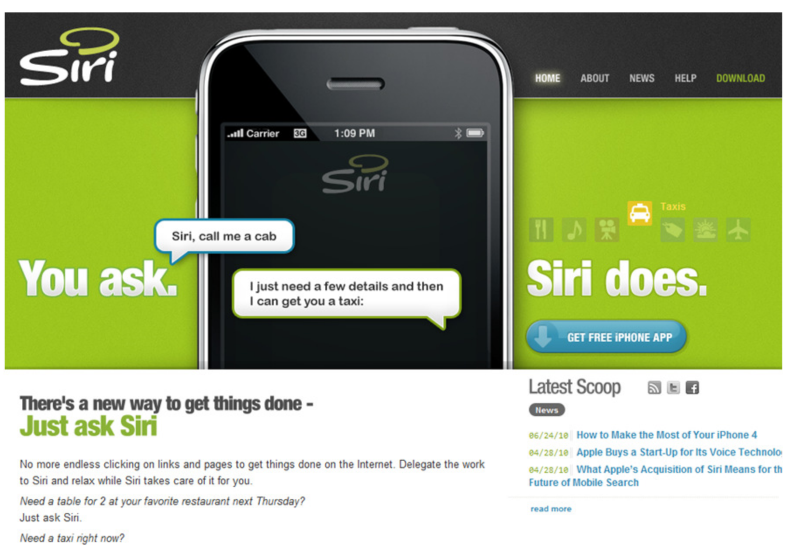
Lecture 1 Outline

- **Background:**
  - Information extraction vs information retrieval
- **Advertising 101 and Digital advertising**
  - Predicting CTR
- **Information Extraction Overview**
- **Sentiment Analysis**
- **Candidate Project**

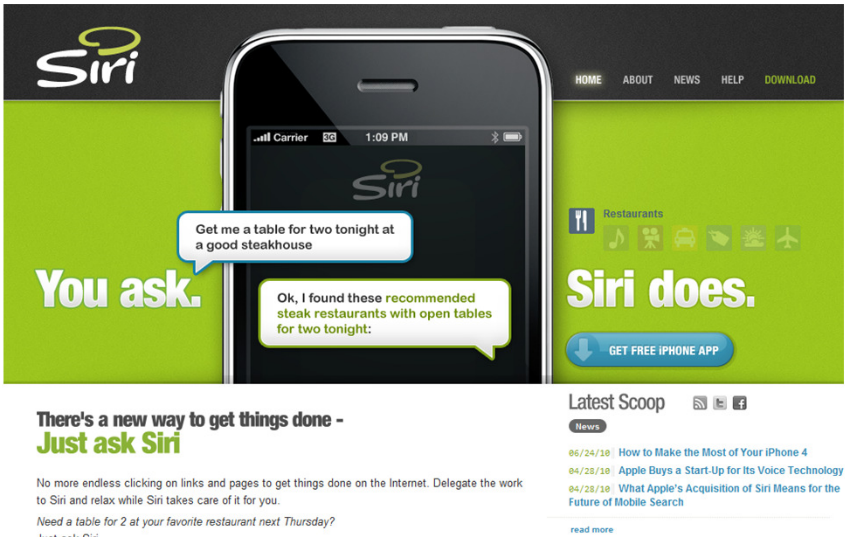
Understanding Text



SIRI: Virtual Personal Assistant



SIRI: Virtual Personal Assistant



Project

- **Build a sentiment-based search engine for people**
  - How happy are people about “Barrack Obama”?

- **Extra Slides**
  - Document Souls

Document Souls

a new paradigm for information access

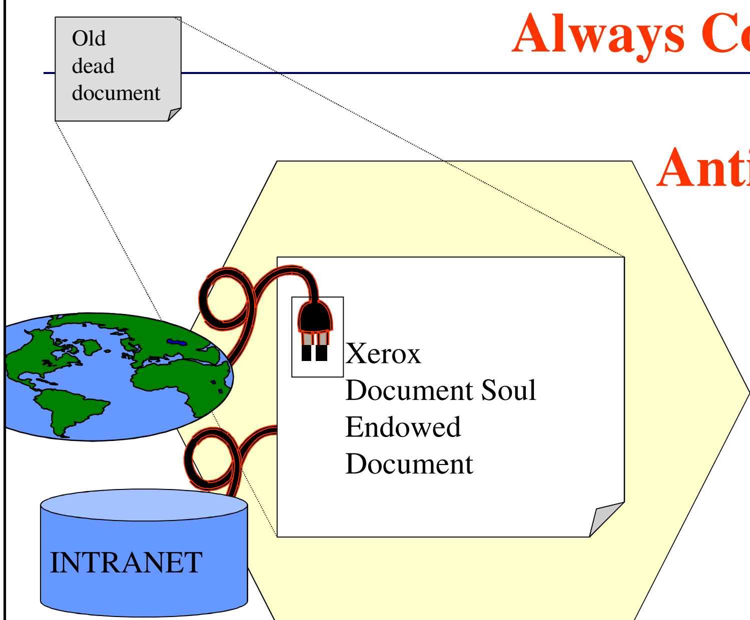
James Shanahan* and Gregory Grefenstette*

** Work performed at Xerox Research*

Interesting juncture

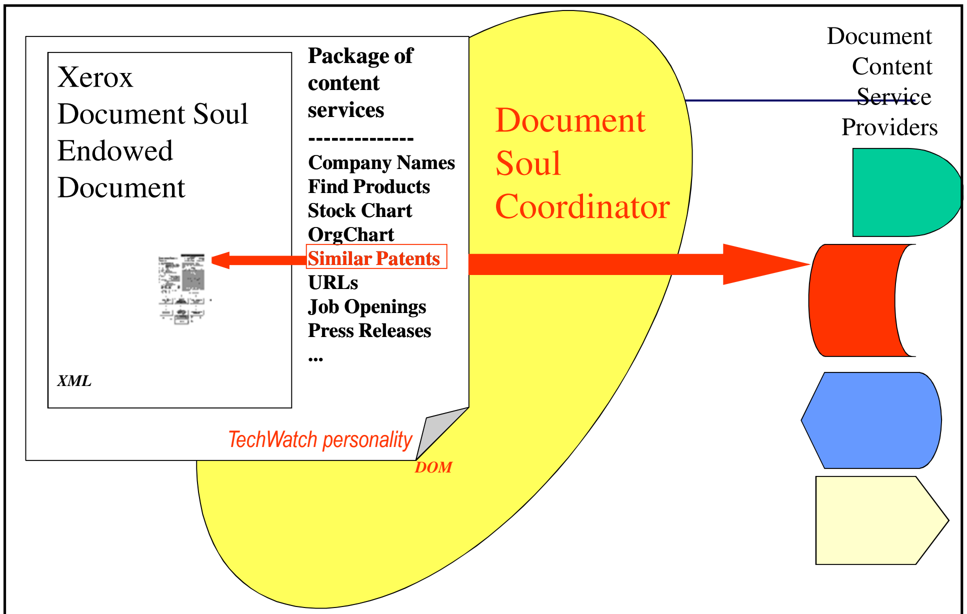
- **High Bandwidth, lots of sleeping computers, very cheap memory/disks**
- **Niche browsers, very specialised information services**
- **Recent studies have estimated the size of the hidden web to be 500 billion pages, while the size of the indexed web is three billion.(http://www.completeplanet.com/Tutorials/DeepWeb/index.asp)**
- **Search Engines have significant limitations**
  - Out of date, only index 1% of online pages, documents with authentication requirements generally are not indexed.
  - Context is ignored
  - Anticipatory services

Always Connected

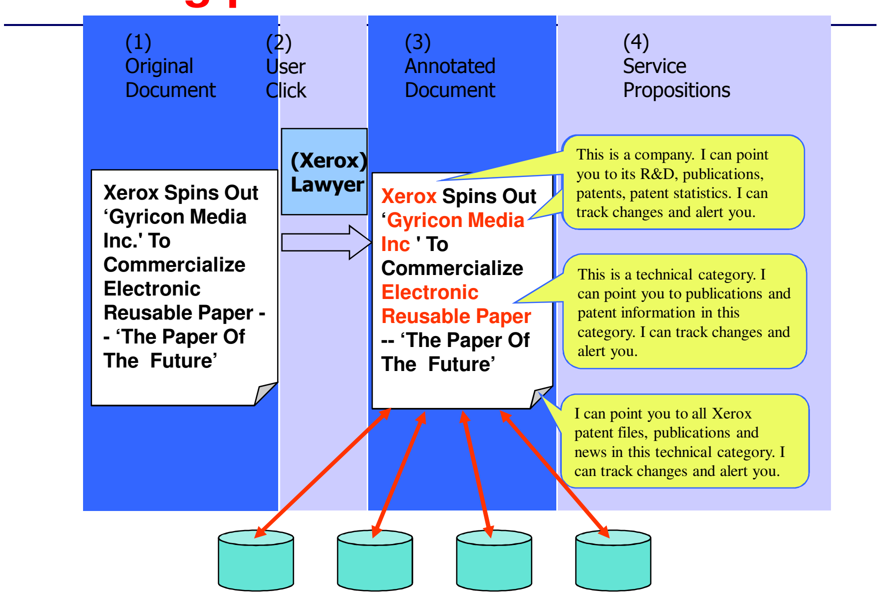


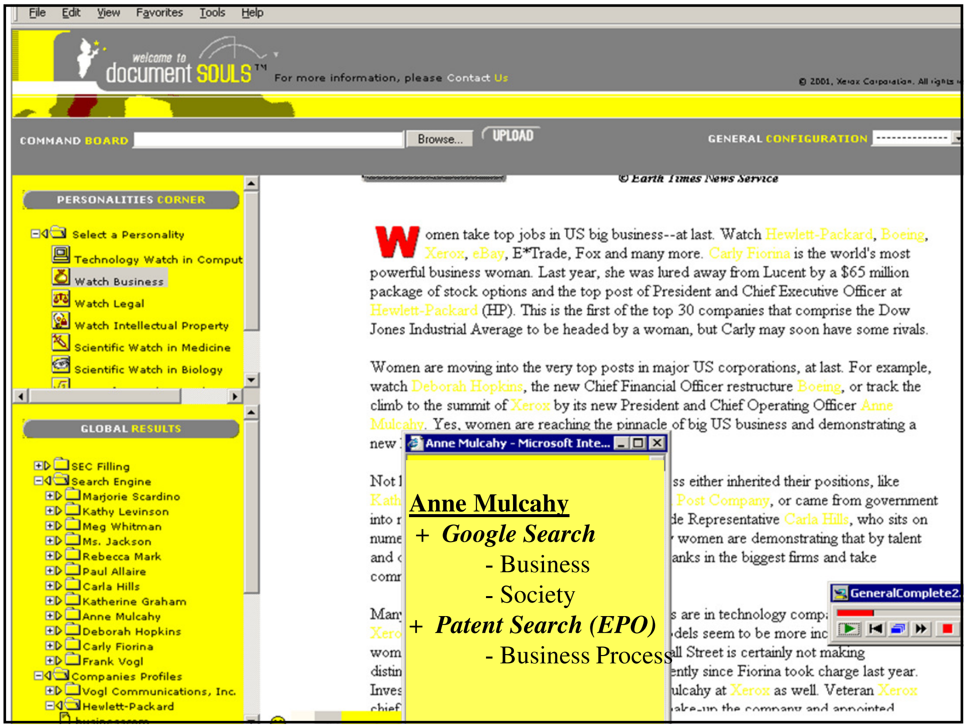
Anticipating

Adding Value



Giving personalities to documents…





Document Souls System

- **A new paradigm of information access**
- **Document gets a life**
- ***Constantly anticipating* your information needs**
- **DS System (beta product)**

- **Related Systems**
  - Contextual search (Yahoo!)
  - Kenjin (Autonomy)
  - And many others


Xerox Documents
NEVER SLEEP
you
your docume