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

James G. Shanahan¹ 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: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 1

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.
 - Identify 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.

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 2

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, “*”}*

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 3

Topics 1/2

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 4

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
- **Languages: R and Python, with Lucene and LingPipe**

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 5

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 6

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

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

by Brandel

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

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

check out our updated & Hottest lists for '08.

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

Demand for these applications is expanding the need for data mining, statistical modeling and data

VMware® virtualization.

Fast reliable disaster recovery—at a cost your business can afford.

Find out more



nes.Shanahan_AT_gmail.com 7

**Data
Driven
Decision
Making
is hot
skill**

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

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

by Brandel

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

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

check out our updated & Hottest lists for '08.

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

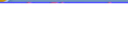
It's not just the case for Google," he says. "There are lots of applications that have big, big, big data sizes, which creates a fundamental problem of how you organize the data and present it to users."

Demand for these applications is expanding the need for data mining, statistical modeling and data

VMware® virtualization.

Fast reliable disaster recovery—at a cost your business can afford.

Find out more



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 and 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."

Demand for these applications is expanding the need for data mining, statistical modeling and data

Done

nes.Shanahan_AT_gmail.com 8

**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 from 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.**

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 9

Some Practical Skills

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 10

Audience Participation



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 11

Questionnaire

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



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 12

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)

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 13

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!**

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 14

Course Topics: Part 2 Outline

Bag of words



- **White space tokenization**
- **Good Classification Technology**
 - Thresholded SVMs
- **Extra semantic processing**
 - Affect/Opinion
- **Process Mining**
 - Bayesian Network Approach



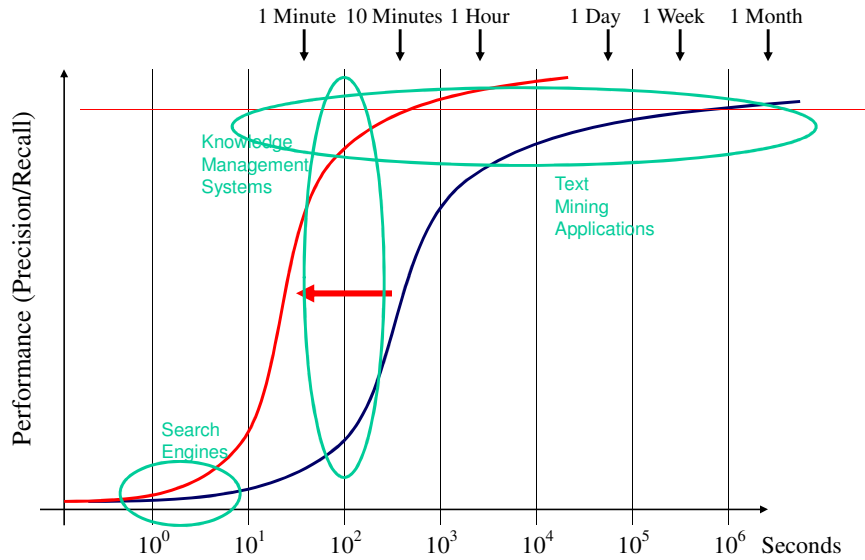
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 15

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**

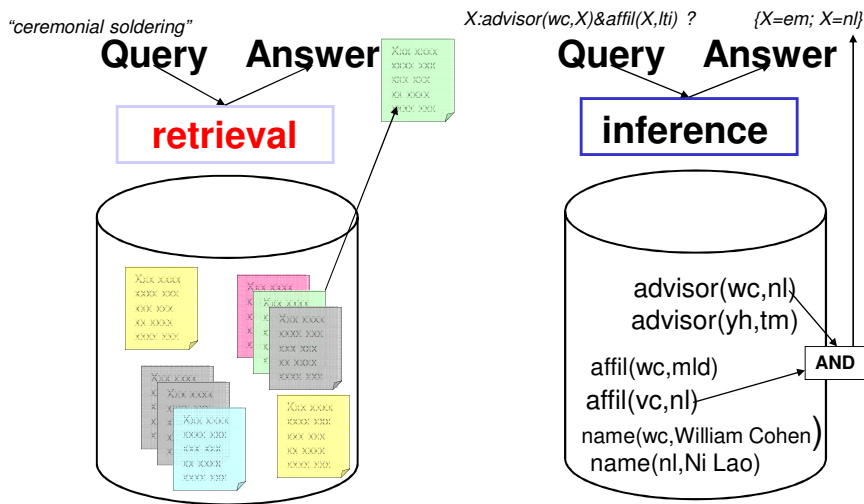
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 16

The Information Access Curve



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 17

Two ways to manage information



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 18

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 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)**
 - 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 20

2000s: Realtime, huge data, lots of uncertainty

iPhone4 App for Local Search

- **Speech**
- **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.**

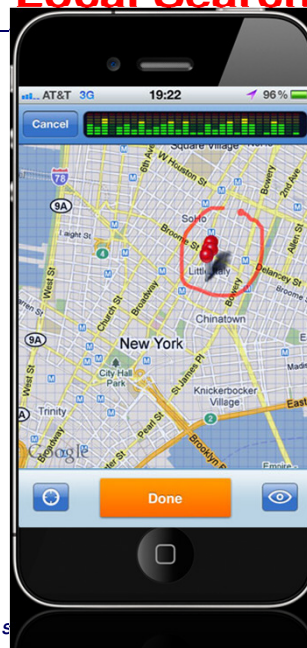


ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shar

21

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"



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. S

22

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**

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 23

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 24

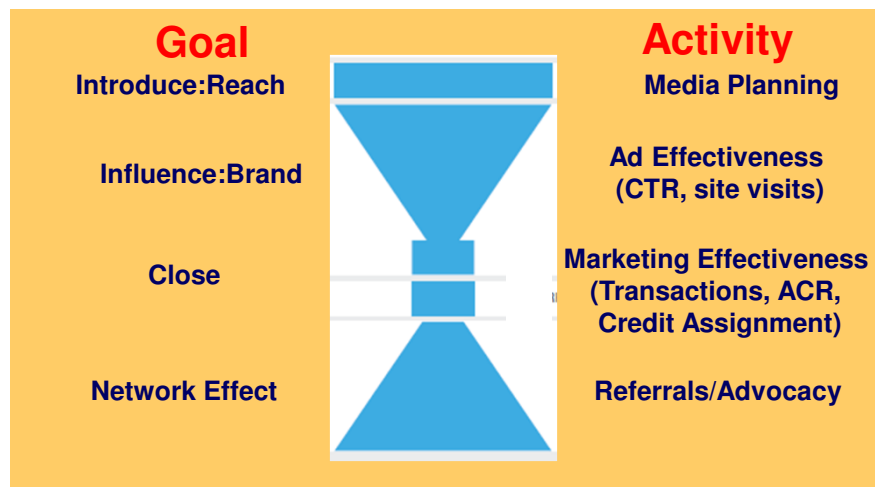
Digital Advertising

- **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)!

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 25

What marketers want?

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



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 26

Advertising makes up ~2% of US GDP

Despite its problems (lack of credit assignment etc.)

- **US GDP = \$14.1 Trillion (Global \$56 Trillion, 56x10¹²)**
- **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>

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 27

Online Advertising: Sponsored Search

The screenshot shows a Google search for 'advertising'. The search bar is at the top left, with 'advertising' entered and a search button. Below the search bar, there are navigation links for 'Web', 'Books', 'Groups', 'News', and 'Scholar'. The search results are displayed in a list format. The first result is a sponsored link from LocalAdLink.com titled 'Advertising at Fraction of the Cost Recruit Agents and Make Big Money'. Below this, there are several organic search results, including a Wikipedia entry for 'Advertising', a Facebook page for 'Advertising | Facebook', a platform-a.com article 'Our Platform puts your brand where life happens', an adage.com article 'Advertising Age is the leading global source of news, intelligence ...', a technorati.com tag 'advertising', an advertising.about.com article 'Advertising - Advertising Careers and Jobs - Advertising ...', and a Google AdSense program page 'Google Advertising'. On the right side of the page, there are several sponsored links for 'Website Advertising', 'Marketing & Advertising', 'Free Online Advertising', 'Facebook Advertising', 'Advertise In Your Area', and 'MySpace Advertising'. The page is personalized based on the user's web history.

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 28

Contextual Advertising on Webpages

enjoy:
www.whitesandresort.com

Phan Thiet Hotels
Find the lowest price on great hotels. Book Now!
PhanThiet.OneTime.com

Kiteboarding clearance
Kitesurf instruction tricks and tips from a professional online!

Google

Best Kiteboarding
Low Prices from the World's Largest Kiteboarding Company
www.bestkiteboarding.com

of local children drowning.

The yearly **Le Fruit Triathlon** is held in Mui Ne on June 1, and includes swimming, running and mountain biking.

Surfing

Mui Ne offers a relatively safe environment for low-key surfing. (see kiteboarding below).

Scuba Diving and Snorkeling

Best diving in **Binh Thuan** Province (or all of Vietnam for that matter) is at Ca Na Beach. The water is clear, the coral reefs are pristine, and the whole area is bursting with marine life. One thing Ca Na is lacking is very many tourists and the resorts to contain them. **Vietnam Scuba** has a very "for Koreans, by Koreans" diving establishment there. The website has some English, but we have not confirmed if anyone on staff speaks English fluently. [Click here](#) to read more about the **scuba diving** potentials at Ca Na Beach and the Hon Cau-Vinh Hao Proposed Marine Protected Area. Though all but undiscovered, **Phu Quy Island Proposed Marine Protected Area** also has

(but be

From tir

east of Binh Thuan Province. In the summer of 2004, three teenagers

S.I.K. CENTER
KITESURFING & WINDSURFING

The Forest Restaurant

THE ULTIMATE BOOK OF POWER KITING
The Ultimate Book of Power Kiting by Jeremy Boyce
Best Price \$9.00 or Buy New \$13.57
Buy from amazon.com
Privacy Information

For standards see IAB
<http://www.iab.net/standards/adunits.asp>

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 29

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-pan-4-11-04>
 - Rectangular, Rich Media and Buttons
- **See IAB for standards**

STATISTICA - Data Mining
Winner of all comparative reviews since 1993 - Free Evaluation CD
www.StatSoft.com

Purex Natural Elements
Gentle on you & your planet
Click here for a chance to win

728 x 90

Intuition
The effortless way to loathe, shave and add moisture in one easy step.
FREE YOUR SKIN™ www.activeresponse.com

300 x 250

CLIVE HAIR CLINICS
HAIR LOSS TREATMENT AND RESTORATION FOR MEN AND WOMEN
FREE PHONE 0800 40 42 47

468 x 60

Upgrade your Career
FREE information
CLICK HERE

120 x 600

ISM 280: Data Mining, Analytics and IE in [source: www.activeresponsegroup.com]

House Ads at AMEX

The screenshot shows the American Express website interface. At the top, there's a navigation bar with 'HOME', 'PERSONAL CARDS', 'TRAVEL', 'SMALL BUSINESS', 'CORPORATIONS', and 'MERCHANTS'. Below this is a search bar and a 'VIEW ACCOUNTS' button. The main content area features a large blue banner with the text 'GET A DECISION IN 60 SECONDS'. Underneath, there are two tabs: 'Personal Cards' and 'Business Cards'. A sub-header reads 'Find the card that's right for you'. Three credit card options are displayed in a grid:

- REWARDS PLUS GOLD CARD:** Earn up to \$100 in Gift Cards, redeemable at participating retail, dining, and entertainment partners. Earn one point for virtually every dollar you spend on the Card. Earn double points on travel for your first year. 1.5 points per eligible dollar thereafter. Double points on gas and groceries for your first year. 1 point per eligible dollar thereafter. [APPLY NOW](#)
- BLUE FROM AMERICAN EXPRESS®:** Receive 0% intro APR for up to 12 months. Receive points for retail, dining, travel and entertainment easily with the Membership Rewards Earnest program. Gain independence with flexible payment options and no annual fee. [APPLY NOW](#)
- GOLD DELTA SKYMILES® CREDIT CARD:** Earn 20,000 bonus miles upon your first purchase with the Card. Earn 5,000 bonus miles when you sign up for two Additional Cards. Earn one mile for virtually every dollar you spend. [APPLY NOW](#)

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 31

House Ads at Amazon

The screenshot shows the Amazon.com website. The top navigation bar includes 'Hello, James Shanahan. We have recommendations for you. (Not James?)', 'Earth Day Savings: Pro-Planet & Wallet-Wise', and 'Your Account | Help'. The main content area features a large advertisement for the Kindle 2, with the headline 'Kindle 2 Has Arrived'. Below this, there are several promotional banners and product recommendations:

- Kindle 2 Has Arrived:** Sleeker design. More storage. Longer battery life. Choose from over 265,000 books all available in under 60 seconds. And now Kindle can read to you. Our revolutionary wireless reading device just got better. [Learn more](#)
- Digital SLR Store:** Find Top Digital SLRs, Lenses, Buying Guides, and More. [Shop Amazon.com/dslr](#)
- Check This Out:** Tax Downloads, Earth Day Sweepstakes, Amazon BlackBerry App, New from Flip Video, Selling on Amazon.
- More to Explore:** You looked at 'The Theory of Industrial Organization' and 'Putting Auction Theory to Work'. You might also consider 'Repeated Games and Reputations...', 'The Theory of Industrial Organization' Hardcover by Jean Tirole, and 'A Mosaic of Tiles'.
- Frequently Bought Together:** When you buy more than 1 book, you can save up to 10% on your purchase.

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 32

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

The Million Dollar Homepage™ 1,000,000 pixels • \$1 per pixel • Own a piece of internet history! Sold: SOLD OUT!

Get my newsletter: your@email address Go! Homepage | Buy Pixels | FAQ | Blog | Pixel List | Press | Testimonials | Contact me

WebHosting, XBOX, make money, THE TIMES, UK Jobs, Irish Art, GIFTTLES, MANGOSTEEN, HANDBAGS, CODD WUB, I AM BETTER THAN YOU AND I AM FILTHY RICH--I AM A JERK/

ism.com/I_AM_BETTER_THAN_YOU_AND_I_AM_FILTHY_RICH--I_AM_A_JERK/
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 33

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



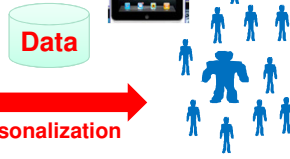
Human Intensive
Lots of guess work
Forward Market

Increasingly



Technology
Data Driven
Forward Market
Spot Markets

Personalization



Advertisers still in
broadcast mode

1st Generation

2nd Generation

3rd Generation

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 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



Human Intensive
Lots of guess work
Forward Market

Increasingly



Technology
Data Driven
Forward Market
Spot Markets

Personalization



Advertisers still in
broadcast mode

1st Generation

2nd Generation

3rd Generation

Double digit growth

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 36

Internet Age Management Science

From Mad Men To Wall Street and beyond!

1st Generation 2nd Generation 3rd Generation

ISM 280: Stochastic Optimization in Info Sys and Tech © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 10 37

Google james shanahan

Search

Bookmarks PageRank AutoLink Settings

Web Results 1 - 10 of about 1,290,000 for james shanahan. (0.09 seconds)

James Shanahan - LinkedIn - 2 visits - 12/3/08
View James Shanahan's professional profile on LinkedIn. LinkedIn is the world's largest business network, helping professionals like James Shanahan discover ...
www.linkedin.com/pub/5/596/84 - 15k - Cached - Similar pages

DBLP: James G. Shanahan - 2 visits - 9/16/08
James G. Shanahan. List of publications from the DBLP Bibliography Server - FAQ ... 8,
James G. Shanahan: Modeling with Words: an Approach to Text ...
www.sigmod.org/dblp/db/indices/a-tree/s/Shanahan_James_G.html - 24k -
Cached - Similar pages

Amazon.com: James G. Shanahan: Books
Computing Attitude and Affect in Text: Theory and Applications (The Information Retrieval Series) by James G. Shanahan, Yan Qu, and Janyce Wiebe (Hardcover ...
www.amazon.com/s?ie=UTF8&search-type=ss&index=books&field-author=James%20G%20Shanahan&page=1 - 140k - Cached - Similar pages

James Shanahan | Facebook
James Shanahan is on Facebook, a social utility that connects people with friends and others who work, study and live around them. James Shanahan uses ...
www.facebook.com/people/James_Shanahan/701216000 - 20k - Cached - Similar pages

The Verbal Judo Institute - James Shanahan
James Shanahan, Print, E-mail ... and Mike Manley, National Director, have trained and certified James to conduct programs in Verbal Judo. ...
www.verbaljudo.com/kata/index.php?option=com_content&task=view&id=42&Itemid=83 - 31k -
Cached - Similar pages

James Shanahan | Biography, Photos, Movies, TV, Credits ...
James Shanahan home page to view photos, read James Shanahan news & biography, see interviews, find movies at Hollywood.com ...
www.hollywood.com/.../james-shanahan

Sponsored Links

Find James Shanahan
Get current address, phone & more. Easy to use, search for free!
www.usa-people-search.com

James Shanahan Info
1 Minute to Search (free summary)™
Locate James Shanahan.
Public-records-now.com

Find James Shanahan
Get Immediate Access To Our Database and Find James Shanahan
www.PeopleFinders.com

38

Over 300-400 applications of ML on this page

The screenshot shows the Yahoo! homepage with a search bar at the top. Below the search bar, there are several sections: "MY FAVORITES" with links to various services like Mail, Autos, Facebook, and Finance; "TODAY - November 30, 2009" featuring a news article about Tiger Woods; "POPULAR SEARCHES" with a list of trending terms; a "DELL" advertisement for the "UNWRAP A GREAT DEAL" on an INSPIRON™ 15 laptop; and a "ONE DAY ONLY!" promotion. The page is filled with numerous small, interactive elements and advertisements, illustrating the complexity of the ML applications used.

eCommerce: Online Shopping Over \$20B

The screenshot shows the Amazon.com homepage with a navigation bar at the top. The main content area features a large advertisement for the Kindle e-reader, promoting it as the "#1 Bestselling, #1 Most Wished For, and #1 Most Gifted Product on Amazon." There are also advertisements for "Huge Savings" on holiday toys and "TERMINATOR SALVATION" on Blu-ray and DVD. The page includes a search bar, a shopping cart, and various recommendation widgets, demonstrating the extensive use of ML in e-commerce for product discovery and personalization.

Online Advertising: \$65B WorldWide

Welcome to TimesPeople
TimesPeople Lets You Share and Discover the Best of NYTimes.com

Home Page Today's Paper Video Most Popular Times Topics Most Recent

The New York Times
Tuesday, December 1, 2009 Last Update: 8:34 PM ET

Click here to watch the video

Switch to Global Edition

Obama Vows to Fight Al Qaeda 'Cancer'
But Warns of No 'Blank Check' for Afghan Government
By DAVID STOUT 8 minutes ago
President Obama said he would begin to draw U.S. forces out of Afghanistan in July 2011, even after sending some 30,000 more by mid-2010.
Text of the Speech

Obama's Speech on Afghanistan
LIVE VIDEO FROM WEST POINT

OPINION
OP-ED: HENRY MORGENTHAU III
Crashing F.D.R.'s Party
Crashing a White House party, as a Virginia couple did last week, is nothing new. Earlier intruders got a scolding from Eleanor Roosevelt.
Brooks: Clear, Hold and Duct Tape | Comments
Herbert: A Tragic Mistake party, as a Virginia couple
Cohen: A Jew in England
Editorial: Swine Flu
Schott: Plastic Kettles
Douthat Blog: Fantastic Mr. Fox

TRAVEL
Art in Hong Kong
Hong Kong's thriving art scene has new ways to enjoy

MARKETS
JAPAN CHINA
Nikkei HangSeng Shanghai
9,580.40 22,113.15 3,252.98

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 41

LinkedIn People Jobs Answers Companies

Explore People Search: Harvard - Vice President at Google - Accounting

\$100K+ Jobs Database - The Most \$100K+ Jobs at TheLadders

Inbox

Prasan Roy	Greetings	Dec 25	take action
Natasha Krol	RE: Your Yandex...	Dec 24	take action
Tina Duccini	Request for Referral...	Dec 19	take action
Sujeewa Alwis	Invitation to connect	Dec 17	take action
Kathleen Ortiz	RE: Research Scitnetist...	Dec 8	archive

Network Updates

Today

CONNECTION UPDATES (8)

Melissa Anderson is now connected to Marie Berroddin, Corie Blumstein, Ross Geier, and 1 other person

Gregory Yankelovich (gyankelovich@gmail.com) is now connected to Jerry Abbott, Marcy Davis (formerly Kintzele), and 7 other people

Autum Ehresmann is now connected to M. Cameron Jones

Steve Hoffman is now connected to Floris Boeienga, David

James G. (Jim)

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 42

Home | Global Sites | Search


ADVERTISING.COM

ADVERTISERS PUBLISHERS SEM ABOUT US CAREERS CONTACT

We reach 9 out of 10... Elvis Impersonators


How We Grab 9 out of 10 Sets of Eyeballs
(and make sure they like what they see).

- 1 We have the largest online ad network.
- 2 We have the most advanced targeting solutions and optimization in the industry.
- 3 We treat our publishers like rock-stars, so we get the best in the industry.



WHAT THIS MEANS FOR YOU.


ADVERTISERS	PUBLISHERS	SEARCH MARKETING
<ol style="list-style-type: none"> 1 Your online ads will meet your goals. 2 We are your one-stop-shop to reach online customers how you want. 	<ol style="list-style-type: none"> 1 More publishers turn to us to monetize their inventory than any other network. 2 More advertisers turn to us giving 	<ol style="list-style-type: none"> 1 We handle everything from planning to reporting to creative. 2 Superior optimization - for every kind of search campaign.



Be the first to use our new management tool for your publishers.
[Learn More >](#)
[Watch the Demo >](#)


Done

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

 INTERWOVEN

Solutions | Products | Customers | Resources | Partners | Education | Consulting | News & Events | Company | Search


PRODUCTS



Website targeting has evolved...
Interwoven Optimost Adaptive Targeting

[Learn more >](#)

Home > Products > Optimost



Hosted software and services empowers marketers to optimize their online presence and drive increased conversions, sales and customer engagement.

- Multivariable Testing
- AB Testing
- Persona Recognition & Adaptive Targeting
- Experienced managed services team

With Optimost You Can Quickly and Easily Optimize

Everywhere	Everything	Everyone
------------	------------	----------

Podcasts

- ▶ Improving Conversion Through Multivariable Optimization and Adaptive Targeting
- ▶ Customer Lifecycle Steps 1 and 2: Attract and Engage
- ▶ User Generated Content (UGC) and its Effect On Your Brand

Webcasts

- ▶ What's Clicking: Building a Powerful Online Marketing Strategy
- ▶ Finding and Targeting Your Critical Audience Segments
- ▶ Best Practices for Website Design

Demos

- ▶ Optimost Self-Running Demo

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

Edit: Title, Description & Tags | Dates | Permissions | Filters |

Title
Brewster Kahle and Jimi Shanahan

Description

Tags
cikm2008 cikm 2008

Go to next item when you save?

More options

mes.Shanahan_AT_gmail.com 45

Versus!

Branded Landing Page & Click-to-Video

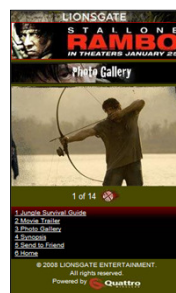


CPM Banner Ad

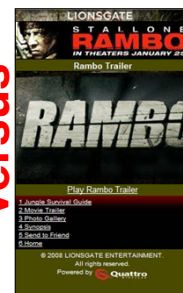


Branded Landing Page

Gallery/Image



Movie Trailer



Versus!

2005-2007 AdMob, Inc.

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 46

E.g., Google AdWords

Video News Maps Gmail more ▾ james.shanahan@gmail.com | [iGoogle](#) | [My Account](#) | [Sign Out](#)

Google™

Google Search I'm Feeling Lucky [Advanced Search](#) [Preferences](#) [Language Tools](#)

[Advertising Programs](#) **[Business Solutions](#)** [About Google](#) [Go to Google Italia](#)

[Make Google Your Homepage!](#)

©2007 Google

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 47

Select Portfolio of Keywords

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

Filter my results

Choose columns to display: ?

Calculate Estimates using Max CPC:

US Dollars (USD \$)

Keywords related to term(s) entered - sorted by relevance ?

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: ?
data mining	1 - 3	\$0.20	<div style="width: 100%;"></div>	<div style="width: 50%;"></div>	<div style="width: 50%;"></div>		Oct	Add ▾
data mining software	4 - 6	\$0.25	<div style="width: 100%;"></div>	<div style="width: 50%;"></div>	<div style="width: 50%;"></div>		Oct	Add ▾
data mining tools	1 - 3	\$0.24	<div style="width: 100%;"></div>	<div style="width: 50%;"></div>	<div style="width: 50%;"></div>		Apr	Add ▾
web data mining	4 - 6	\$0.21	<div style="width: 100%;"></div>	<div style="width: 50%;"></div>	<div style="width: 50%;"></div>		Nov	Add ▾
data mining techniques	1 - 3	\$0.23	<div style="width: 100%;"></div>	<div style="width: 50%;"></div>	<div style="width: 50%;"></div>		Oct	Add ▾
data mining	4 - 6	\$0.24	<div style="width: 100%;"></div>	<div style="width: 50%;"></div>	<div style="width: 50%;"></div>		Nov	Add ▾

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 48

Which Keywords? How much for a click?

Web Images Video News Maps Gmail more james.shanahan@gmail.com | Web History | My Account | Sign out

Google data mining Search Advanced Search Preferences

Web Groups Scholar Books Personalized Results 1 - 10 of about 66,300,000 for data mining [definition]. (0.14 seconds)

Data Mining Software Sponsored Link
www.salford-systems.com FREE: 30-day Eval & Online Training Webcast, Guided Tour, Case Studies

Data mining - Wikipedia, the free encyclopedia
Data mining can be defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data". [1] Data mining may ...
en.wikipedia.org/wiki/Data_mining - 68k - Cached - Similar pages - Note this

Data Mining: What is Data Mining?
Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into ...
www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm - 13k - Cached - Similar pages - Note this

Data Mining Techniques
Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent ...
www.statsoft.com/textbook/stdatmin.html - 47k - Cached - Similar pages - Note this

Data Mining: Text Mining, Visualization and Social Media
Commentary on text mining, data mining, social media and data visualization.
datamining.typepad.com/ - 62k - Cached - Similar pages - Note this

Sponsored Links

Mine Text Data
Analyze Consumer Opinions
Categorize Issues Automatically
www.clarabridge.com

Open Source Data Mining
Supercharged PostgreSQL Database
30 Days Free Support, Download Now!
www.greenplum.com

Easy Data Mining
Discover a data mining system that easily exports data to Excel.
Datawatch.iresponse.net

Data Mining Software
Discover insights hidden in your existing data using SPSS solutions.
www.spss.com

STATISTICA - Data Mining
Winner of all comparative reviews

ISM 280: D

web video images music shopping searchme lite stacks tools preferences about us

computer science movies history foot search all

searchme: james g shanahan

Universität Trier

James G. Shanahan

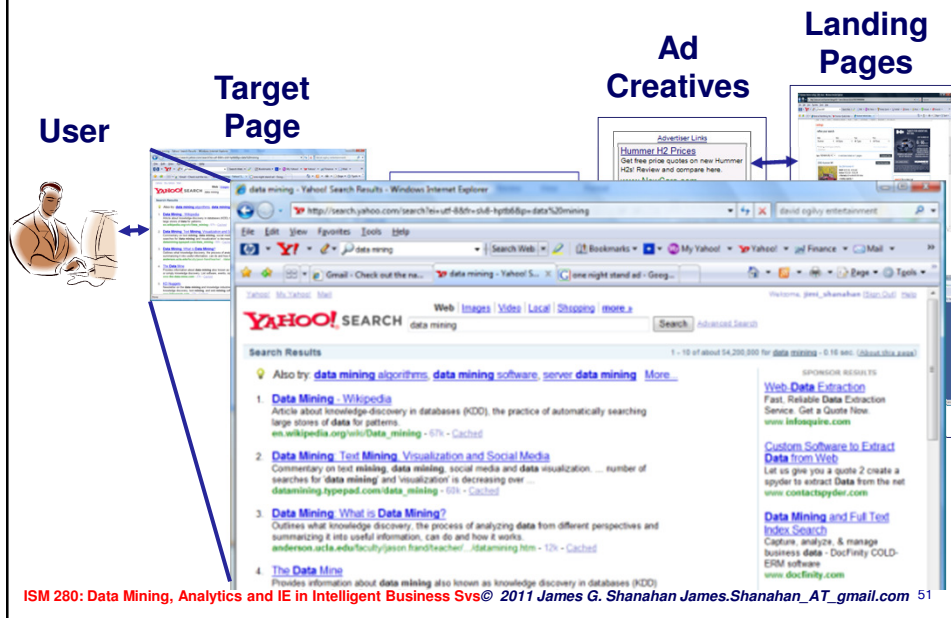
List of publications from the DLP Bibliography Server PAC

Coauthor Index - A-Z: ACM DL/DBase - CiteSeer - CSB - Google - MEN - Yahoo

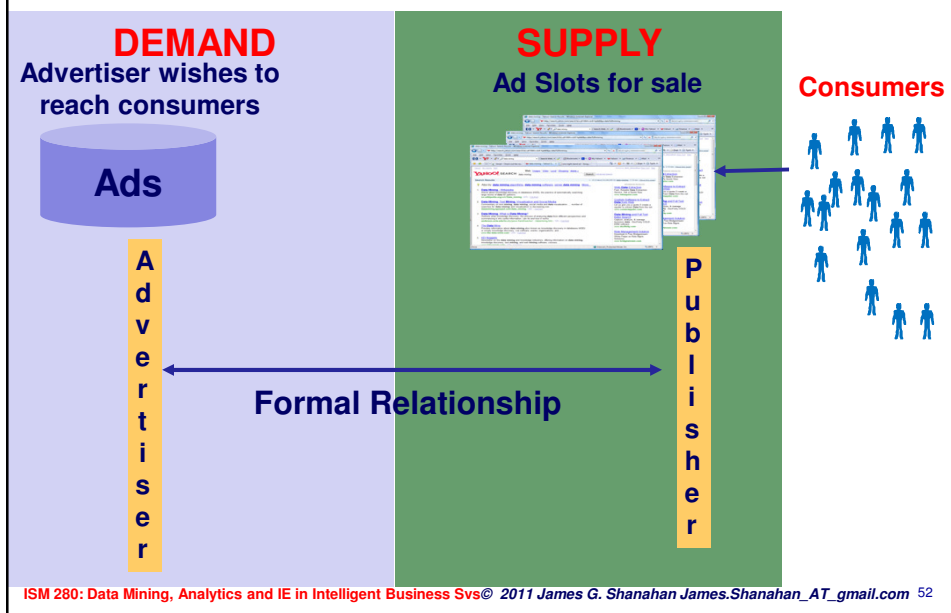
Year	Author(s)	Title
2008	James G. Shanahan, Steve Aracivchia, Irena Vindencic, Yi Zhang, David A. Evans, Ankur Arora, Samir Arora, ...	Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008
2005	James G. Shanahan, ...	Probabilistic workflow mining...
2004	James G. Shanahan, ...	Analysis of ...
2004	James G. Shanahan, ...	Topic structure modeling...
2002	David A. Evans, James G. Shanahan, Victor Shafiq, ...	Topic structure modeling...
2001	James G. Shanahan, ...	Modeling with Words: an Approach to Text Categorization...
1999	James G. Shanahan, ...	Query Logic in Artificial Intelligence...

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 50

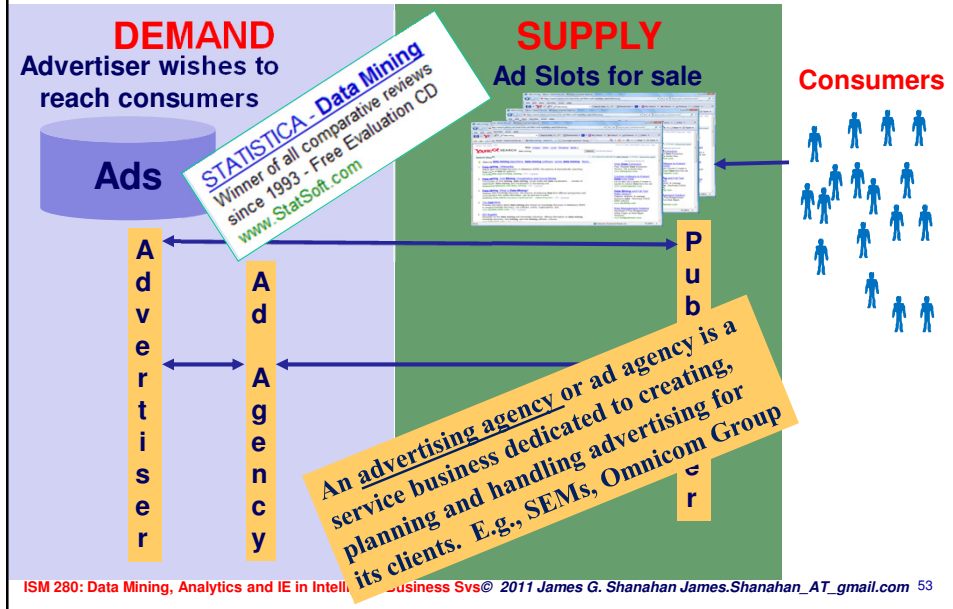
CPC Paid Search (KW Market place)



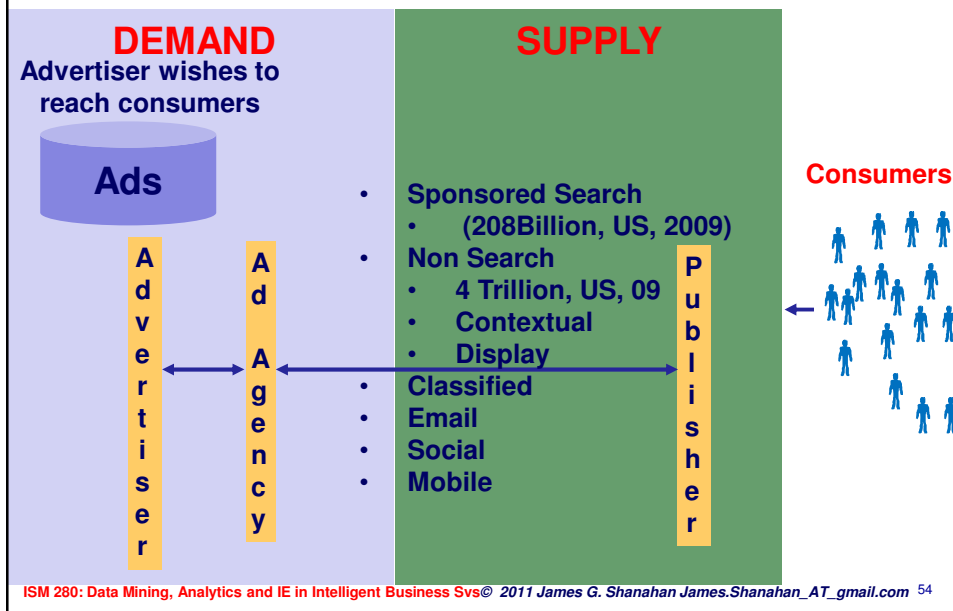
Advertising: a supply-demand marketplace



Advertising Agency: creates & traffics ads



Advertising Agency: creates & traffics ads

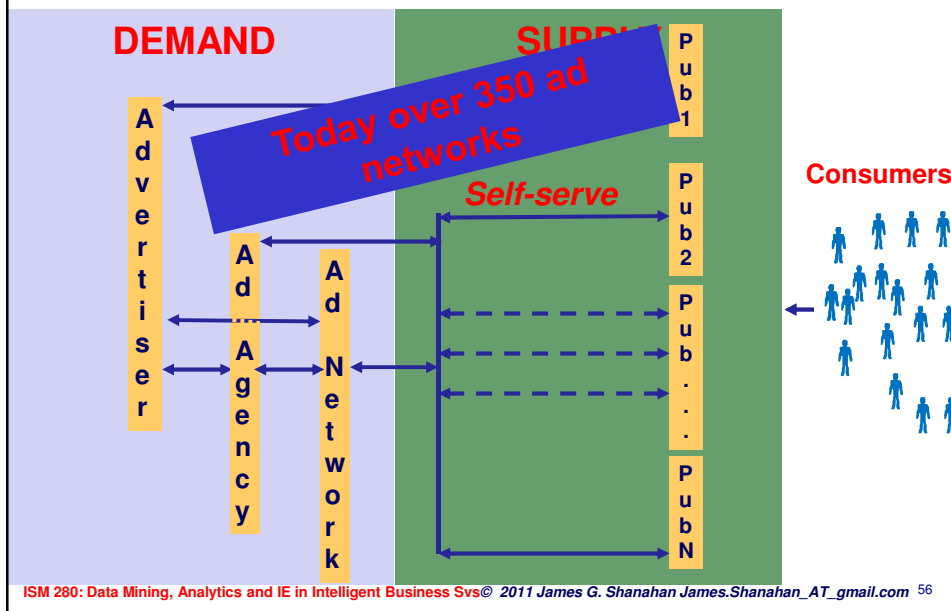


2nd Generation

- **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

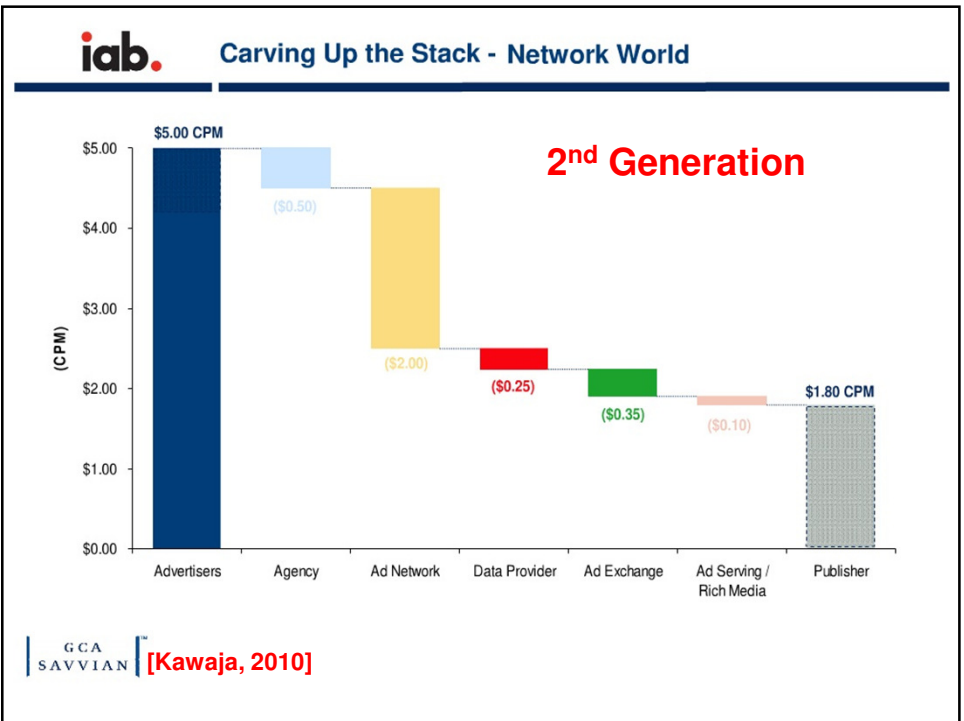
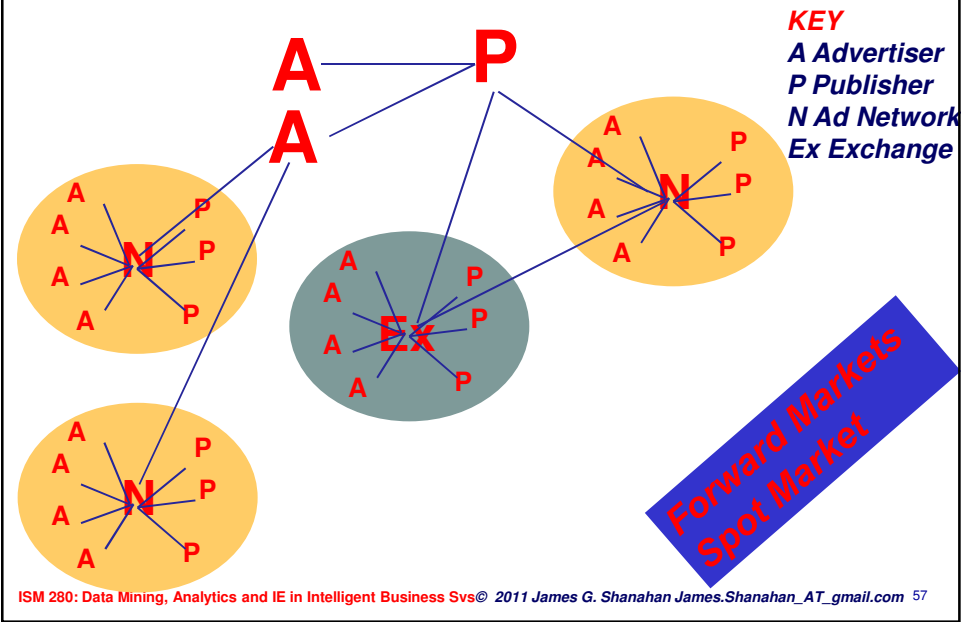
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 55

Advertising Network: Aggregates Publishers



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 56

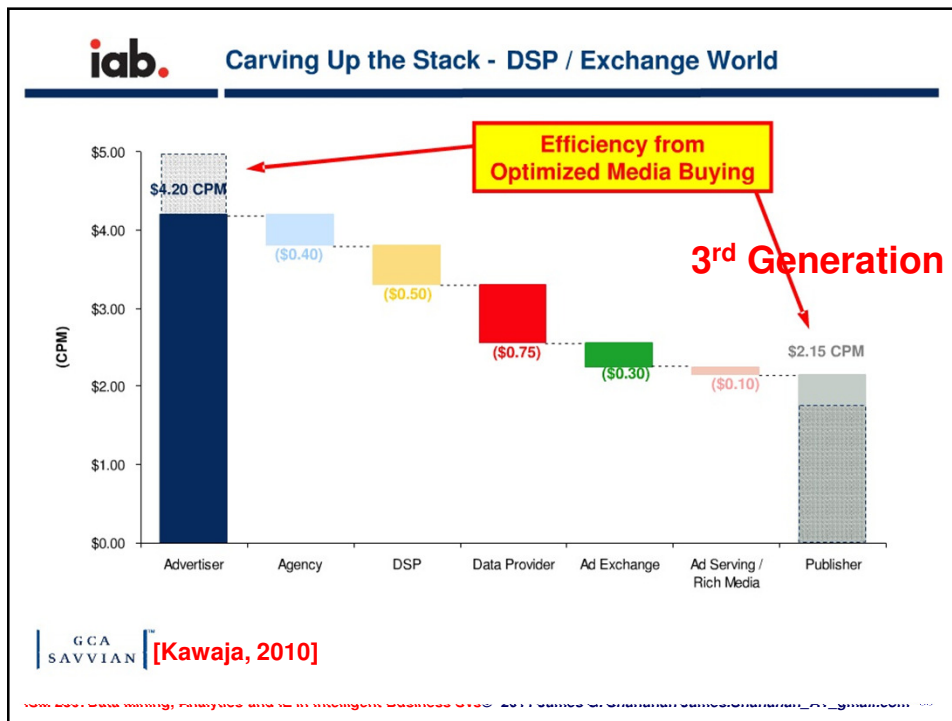
Online Advertising is a Frenemic Network Play



3rd Generation

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 59

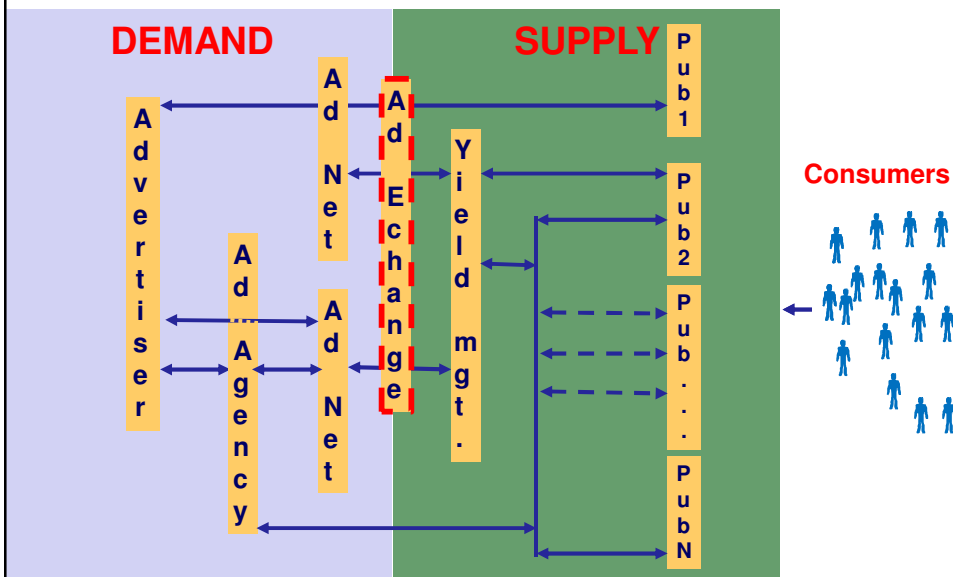


Ad Exchanges: a new SD Marketplace

- **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**

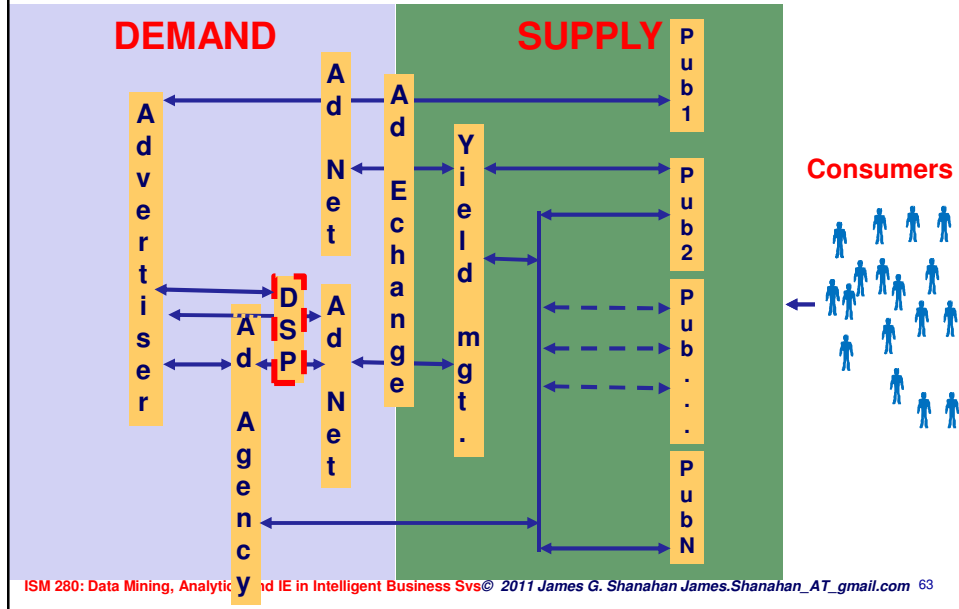
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 61

Ad Exchange: auctioneer-centric marketplace

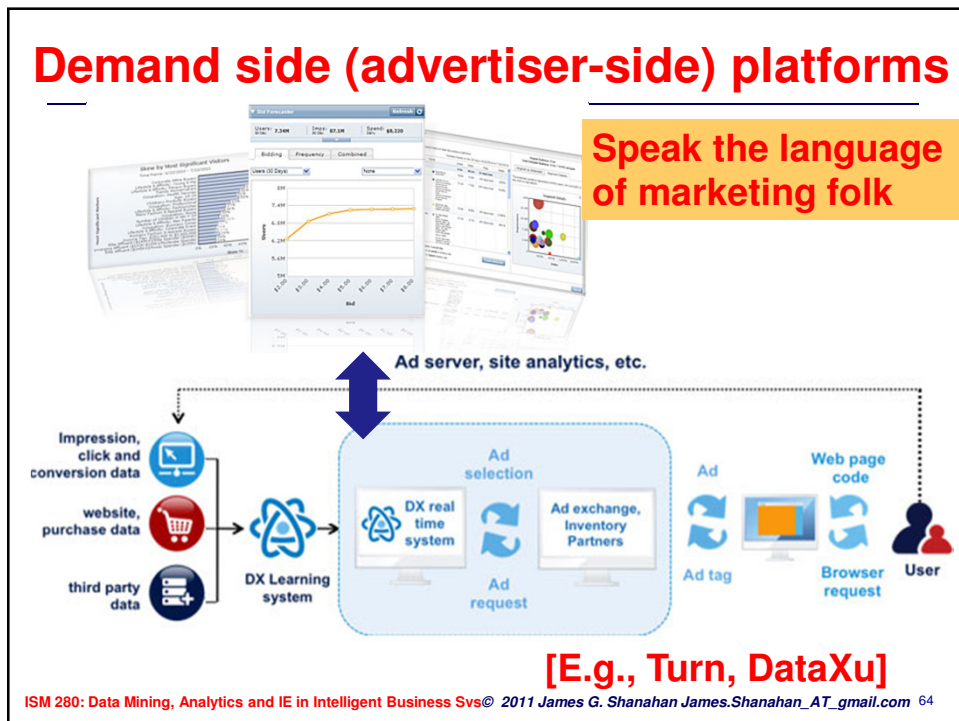


ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 62

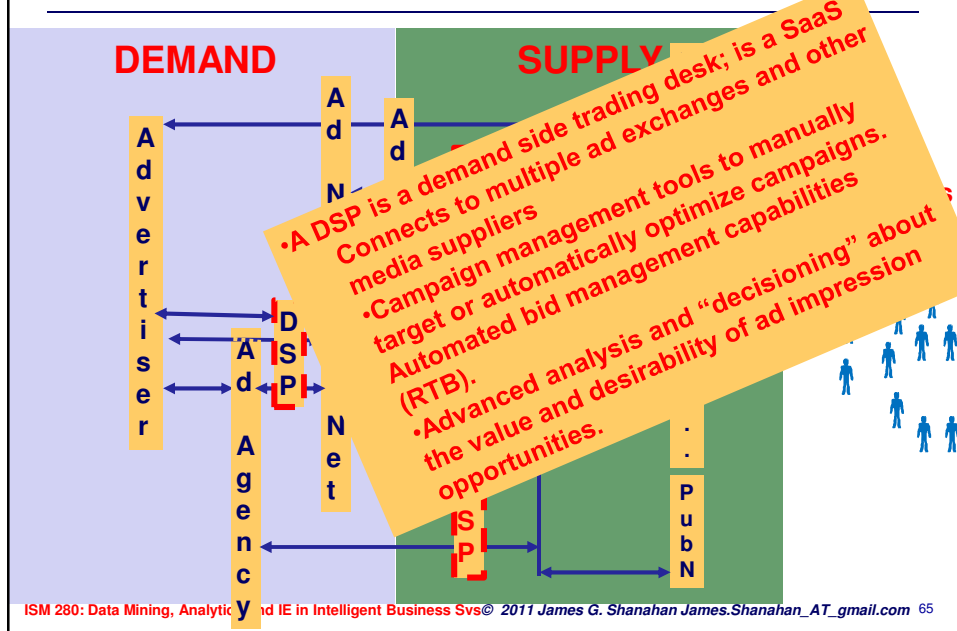
Demand-Side Platform: A trading desk for Adv.



Demand side (advertiser-side) platforms



Demand Side, Supply side platforms



Key Features of DSP

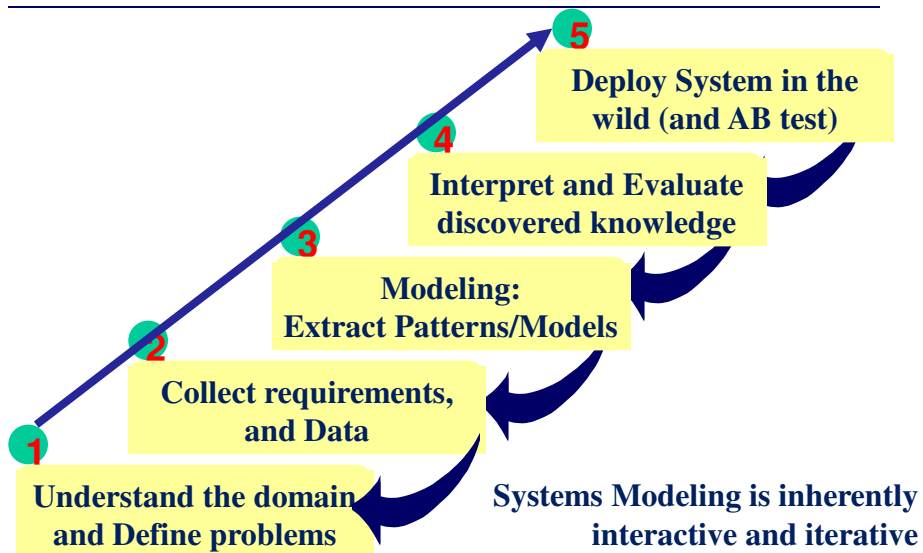
- **Advanced and accurate audience targeting capabilities**
- **Easy-to-use inventory control**
- **Bidding dashboards**
- **Ability to set frequency caps on the ads being served**
 - reaching the "right consumer" too many times can lead to a significant decline in interest

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**

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 67

Machine Learning in Practice



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 68

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 second item at the second item at the third-highest price, and so on

Bid = \$10
PPC = \$5

Bid = \$5
PPC = \$2

Bid = \$2
PPC = \$1

Bid = \$1
PPC = \$0.57

Mine Text Data

Analyze Consumer Opinions
Categorize Issues Automatically
www.clarabridge.com

Open Source Data Mining

Supercharged PostgreSQL Database
30 Days Free Support, Download Now!
www.greenplum.com

Easy Data Mining

Discover a **data mining** system that easily exports **data** to Excel.
Datawatch.iresponse.net

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?!

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 69

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

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

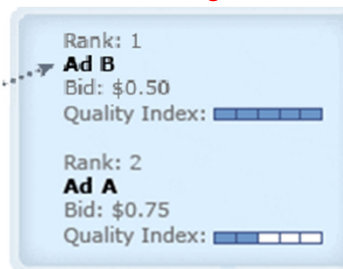
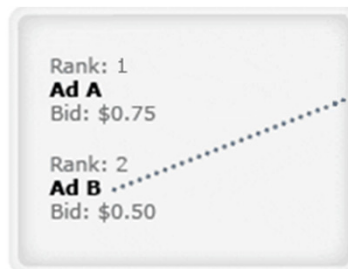
$$ECPM_{Ad} = AdQualityIndex_{Ad} * Bid_{Ad}$$

PAY

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

Bid-to-Position Model

ECPM-Ranking Model



ISM 280: Dat

ail.com 70

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:

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

	1. Receive	2. Assess	3. Calculate	4. Set CPC
Ad Id	Bid	Quality	Rank	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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 71

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}) + \0.01
 - 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	$\$0.4 \times 4 = 1.6$	2	$(1.2 / 4) + \$0.01 = \0.31
Mark	\$0.50	10	$\$0.50 \times 10 = 5$	1	$(1.6 / 10) + \$0.01 = \0.17
Laura	\$0.20	6	$\$0.20 \times 6 = 1.2$	3	$(1 / 6) + \$0.01 = \0.17
Cameron	\$2.00	0.5	$\$2.00 \times .5 = 1$	4	$(.8 / .5) + \$0.01 = \1.61
Alison	\$0.05	16	$\$.05 \times 16 = .8$	5	$(.2 / 2) + \$0.01 = \0.11
Will	\$0.10	2	$\$.10 \times 2 = .2$	6	Minimum Bid

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 72

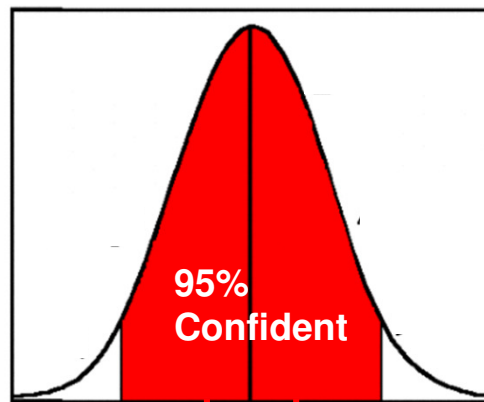
Accurate CTR Estimates are Crucial

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

- **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]

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 73

Estimating CTR (and later AR)



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

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

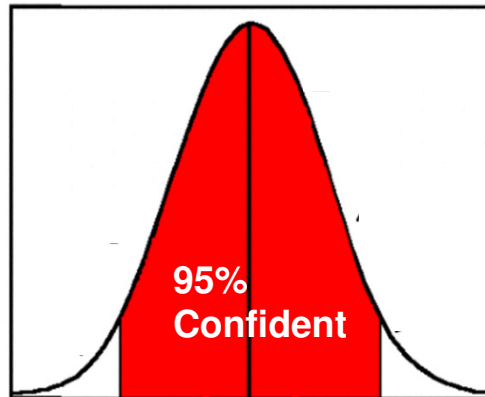
1.6% 2.6% 3.6% CTR (after 1,000 impressions)

2.3% 2.9%

(after 10,000 impressions)

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 74

Estimating CTR (and later AR)



For a network of
 $\sim 10^9$ target pages,
 $\sim 10^6$ ads
 $\sim 10^7$ users

-
- Cannot afford this evaluation/auditioning
 - Borrow strength, marginalize
 - CoD (curse of dim^{ality})

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

2.3% 2.9%

(after 10,000 impressions)

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 75

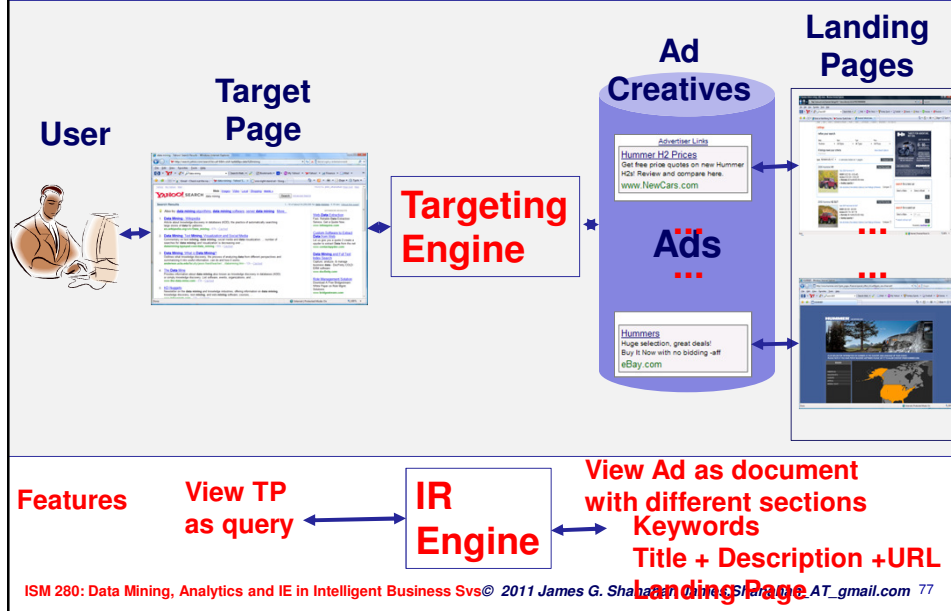
Accurate CTR Estimates are Crucial

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

- **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!!

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 76

Ranking Ads using IR



Estimating CTRs using ML

- Estimate CTR using $Pr_{Ad}(\text{Click}|\text{Keyword})$
- 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, Javier Luque, Juan Viademonte and Ricardo Baeza-Yates: New Stochastic Algorithms for Placing Ads in Sponsored Search. LA-Web, Santiago, Chile 2007

What features could be used?

ML Features 1/2

Features(KW,AD, LP)->CTR
 $X_i \rightarrow \text{CTR}_i$

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

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

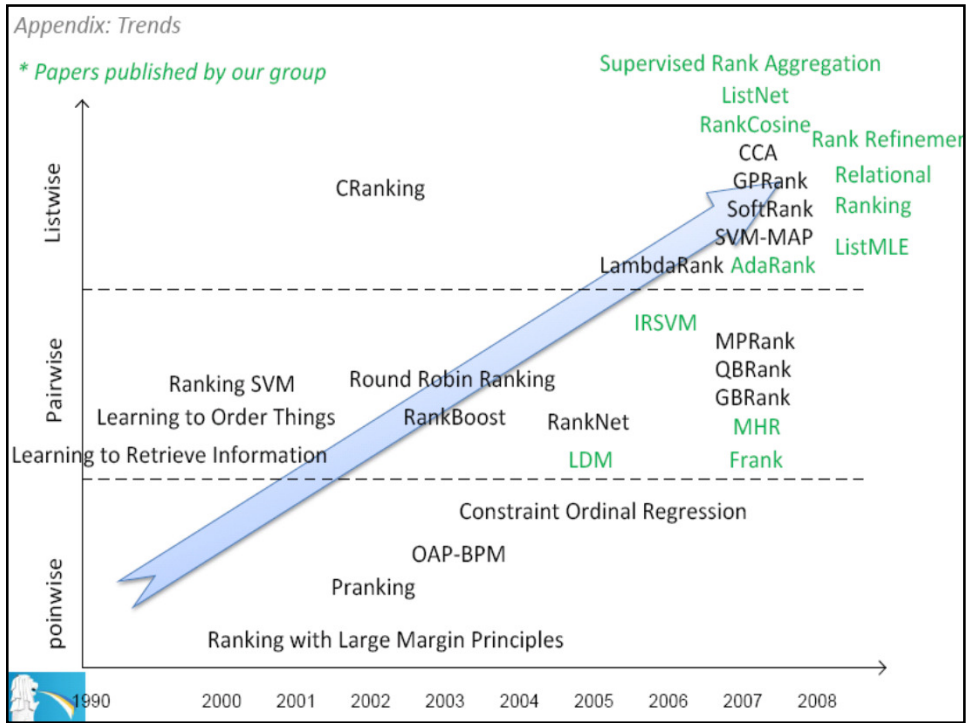
[Richardson et al., 2007]

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 79

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?

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 80



Visual Search and Stacks

Visual Search | Searchme.com - Mozilla Firefox

http://www.searchme.com/

searchme: Robin Williams

stack: Robin Williams To Undergo Heart Surgery

table of contents >

- CNN.com Entertainment: Robin Williams to undergo heart surgery
- EW.com News Briefs: Prince William Pregnant? It's Not Official Yet

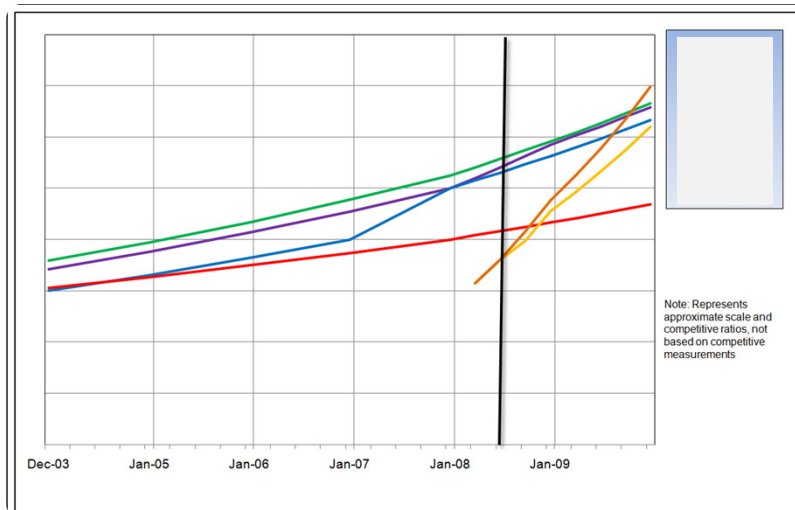
searchme: toolbar

install now


hot searches

- U2 at Fordham University
- Spring Forward Today
- Barbie Turns 50
- Listen Free: Bruce Springsteen
- Obama Funds Stem Cell Research
- Healthy Brunch Ideas
- PETA's 2009 Worst Dressed List

Tracking Progress is key thru evaluations and metrics



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 83



[Turn Smart Market](#)
[Advertisers](#)
[Publishers](#)
[About Us](#)
[Privacy](#)

[Login](#)

Turn Delivers Performance

- Revolutionary Technology
- Brand Quality Enforcement
- Powerful Insights and Analytics

Let Turn deliver on your campaign

Customize your own chart. Select a content category from the menu to the right and any combination of metrics below.

Price:

eCPM

eCPC

eCPA

Performance:

CTR

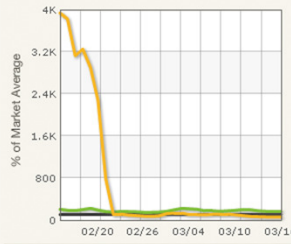
CR

Close

Help Tips: on off

Turn Market Index

Performance by: Travel



About This Chart

Advertisers

Get true performance you can measure and control

[Learn More](#)
[See a Preview](#) [Get Started Now](#)

Publishers

Earn higher revenue from increased competition



[Learn More](#)
[See a Preview](#) [Get Started Now](#)

What's New at Turn

Turn Launches Industry's First Dynamic Pricing Model For Behavioral Targeting

[Read More](#)

Turn Clients:

ing data from www.turn.com...

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 84

42

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 85

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**

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 86

- Climbing the NLP foodchain
- Information Extraction
- Sentiment

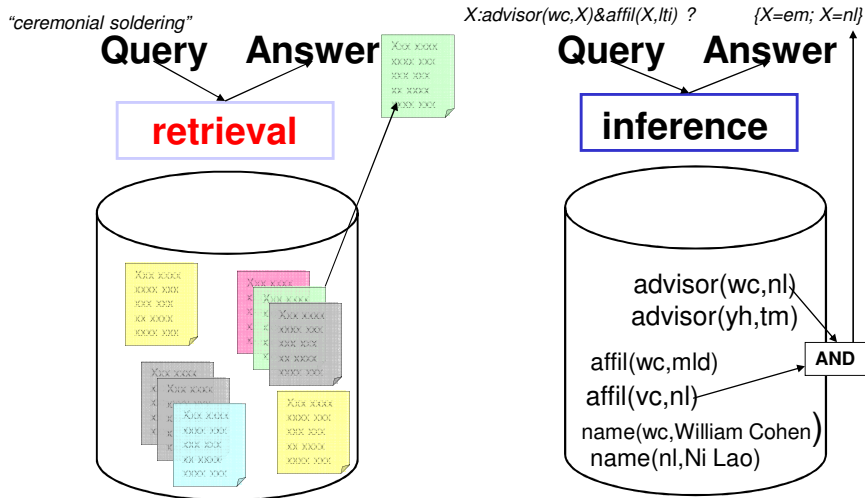
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 87

Over 300-400 applications of ML on this page

The screenshot shows the Yahoo! homepage with several key features demonstrating machine learning applications:

- Search Bar:** A search bar with a "Web Search" button and a dropdown menu showing "No suggestions. Please begin typing your search query." This likely uses a recommendation engine.
- MY FAVORITES:** A sidebar with icons for "View Yahoo! Sites", "Yahoo! Mail (8)", "Autos", "Facebook", "Finance (Dow Jones)", "Flickr", "Games", "HotJobs", "Messenger", "Movies", "Personals", "Sports", and "Updates".
- TODAY - November 30, 2009:** A main content area featuring a large image of Tiger Woods with the headline "Tiger Woods breaks his silence". Below the headline, there are several smaller news snippets: "Tiger Woods breaks silence", "Obama battles bad story lines", "Credit-report scoring revealed", and "Big Cyber Monday deals".
- POPULAR SEARCHES:** A list of ten popular search terms: 1. Cyber Monday, 2. Elizabeth Smart, 3. New Orleans Sain..., 4. GPS Systems, 5. The Princess and..., 6. Charlize Theron, 7. Visa Lottery, 8. Ashlee Simpson, 9. Pink Floyd, 10. Mortgage Rates.
- UNWRAP A GREAT DEAL:** A promotional banner for Dell's "INSPIRON™ 15" laptop, featuring a "CLICK TO EXPAND" button and a "SAVE NOW" button.
- ONE DAY ONLY!:** A large blue banner with the text "ONE DAY ONLY!" and the Dell logo.
- Save on gifts they'll love:** A section with four product categories: "Top 10 cell phones", "Top 10 handbags", "Top 10 shoes", and "Top 10 watches".

Two ways to manage information



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com 89

Understanding Text

28 minutes ago via Echofon

Joan Rivers What's wrong with this country? The evening news is down to 30 minutes a day but we have 85 reality shows about midgets making cupcakes?
about 1 hour ago via web

reitshamer How much market research for the iPad? "None," Mr. Jobs replied. "It isn't the consumers' job to know what they want."
<http://mytl.ms/dTCxY5>
about 1 hour ago via Tweetie for Mac

OnionSports 13-Year Old Nearing Record For Most Masturbations In A Single Day <http://onion.com/eOkwq #SportsDome #RoadTo32>
about 1 hour ago via HootSuite

stereogum The Mountain Goats – "Damn These Vampires" (Stereogum Premiere) <http://bit.ly/hYgsY3>
about 1 hour ago via web

arcadefire We are headlining Coachella Festival on Sat April 16th. Tickets will be available from www.coachella.com/tickets Fri January 21st @ 10am PST
about 1 hour ago via web

caitlinmoran Hope the press is on this: Lack of support forces Bristol woman to have daughter taken into care. <http://tinyurl.com/6hmitwuo #respite4nven>
about 1 hour ago via TweetDeck

wingoz Stats I like. This weekend 1st Time in SB era all 4 starting qbs in conference title games are 1st round draft picks.
about 1 hour ago via UberTwitter

ISM 280: n_AT_gmail.com 80

How do you extract information?



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

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



IS Mining, Analytics and IE in Intelligent Business Sys © 2011 James G. Shanahan James.Shanahan_AT_g

What is "Information Extraction"

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

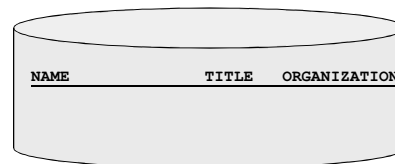
October 14, 2002, 4:00 a.m. PT

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...



IS Mining, Analytics and IE in Intelligent Business Sys © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 92

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...



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

IBM Gov. Data Mining, Analytics and e-Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 93

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...



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



End User

IBM Gov. Data Mining, Analytics and e-Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 94

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”

IBM Gov. Data Mining, Analytics and e-Intelligence Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 95

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

IBM Gov. Data Mining, Analytics and e-Intelligence Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 96

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](#)

IBM Gov. Data Mining, Analytics and e-Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 97

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."

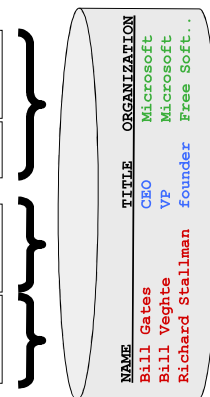
[Richard Stallman](#), founder of the [Free Software Foundation](#), countered saying...

* [Microsoft Corporation](#)
CEO
[Bill Gates](#)

* [Microsoft](#)
[Gates](#)

* [Microsoft](#)
[Bill Veghte](#)
[Microsoft](#)
VP

[Richard Stallman](#)
founder
[Free Software Foundation](#)



IBM Gov. Data Mining, Analytics and e-Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com 98

Example: Finding Jobs Ads on the Web

The screenshot shows a Google search for "baker job opening". The search results include:

- Job Opening - Find ANY Job! - Search by Type, Industry & Geography** (www.careerbuilder.com) - Callout: **Martin Baker, a person**
- Job Opening At Flipdog.Com** (www.FlipDog.com) - Callout: **Genomics job**
- Softimage:Community:Discussion Groups:ds.archive.0004** - Callout: **Employers job posting form**
- CGI: Job Opening** (www.genomics.cornell.edu/jobs/view_job.cfm?td=10) - Callout: **Genomics job**
- Information Activist Job Opening - May 2001** (www.igc.org/dacenter/job.html) - Callout: **Employers job posting form**
- Post an Employee Benefits Job Opening (Help Wanted) Ad** (www.benefitslink.com) - Callout: **Employers job posting form**
- Post an Employee Benefits Job Opening (Help Wanted) Ad** (www.benefitslink.com) - Callout: **Employers job posting form**

Additional text on the page includes "ISM 2" on the left and "nes.Shanahan_AT_gmail.com 99" on the right.

Example: A Solution

The screenshot shows the FlipDog website interface. Key features include:

- Navigation:** Home, Find Jobs, Your Account, Resource Center, Support, Employers.
- Statistics:** 647,514 Job Opportunities from 63,641 Employers.
- Find a Job! Post Your Resume** buttons.
- Employers Products & Services** section.
- Job Seekers:** Find your dream job! with links to reports, accounts, and search tools.
- Jobs for Sports Fans:** Head Football Coach, Football Coach, Asst. Football Coach, High School Football Coach, Univ. Asst. Football Coach.
- Job Seeker Newsletter:** Enter your e-mail address.
- Top 100 Web Sites:** PC Magazine Nov 2000, Career Web Site, Media Mesh, Sept. 2000, Top 10 Job Site.
- Powered by WhizBang!**

Additional text at the bottom left includes "ISM 280: Data" and "Microsoft PowerPoint - [ita...]" and the bottom right shows "12:12 AM 7/7/100".

Extracting Job Openings from the Web

foodscience.com-Job2

JobTitle: Ice Cream Guru
Employer: foodscience.com
JobCategory: Travel/Hospitality
JobFunction: Food Services
JobLocation: Upper Midwest
Contact Phone: 800-488-2611
DateExtracted: January 8, 2001
Source: www.foodscience.com/jobs_midwest.htm
OtherCompanyJobs: foodscience.com-Job1

Ice Cream Guru
 If you dream of cold, creamy chocolate or gooey, gooey cookie there's a great opportunity for you to maintain and expand this major corporation's high-end ice cream line. Will be based in the Upper Midwest for about a year. later on, California. Here I come! Requires a BS in Food Science or dairy, plus ice cream formulation experience. Will consider entry level with an BS and an internship.
 Contact: Suzanne@foodscience.com
 1-800-488-2611

ISM 280: Data Mining, A... hanahan_AT_gmail.com101

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

FlipDog
 Fetch Your Next Job Here™

Home Find Jobs Your Account Resource Center
 Return to Results | Modify Search | New Search

1 - 25 of 47 jobs shown below

Search these results for: [] Search tips Show Jobs Posted: For all time periods

View: Brief | Detailed

Web Jobs: FlipDog technology has found these jobs on thousands of employer Web sites.

Food Pantry Workers at Lutheran Social Services	October 11, 2002	Archbold, OH
Cooks at Lutheran Social Services	October 11, 2002	Archbold, OH
Bakers Assistants at Fine Catering by Russell Morin	October 11, 2002	Attleboro, MA
Baker's Helper at Bird-in-Hand	October 11, 2002	United States
Assistant Baker at Gourmet To Go	October 11, 2002	Maryland Heights, MO
Host/Hostess at Sharis Restaurants	October 10, 2002	Beaverton, OR
Cooks at Alta's Rustler Lodge	October 10, 2002	Alta, UT
Line Attendant at Sun Valley Coporation	October 10, 2002	Huntsville, UT
Food Service Worker II at Garden Grove Unified School District	October 10, 2002	Garden Grove, CA
Night Cook / Baker at SONOCO	October 10, 2002	Houma, LA
Cooks/Prep Cooks at GrandView Lodge	October 10, 2002	Nisswa, MN
Line Cook at Lone Mountain Ranch	October 10, 2002	Big Sky, MT
Production Baker at Whole Foods Market	October 08, 2002	Willowbrook, IL
Cake Decorator/Baker at Mandalay Bay Hotel and Casino	October 08, 2002	Las Vegas, NV
Shift Supervisors at Brueggers Bagels	October 08, 2002	Minneapolis, MN

ISM 280: Data Mining, A...

Data Mining the Extracted Job Information

Job Opportunity Index - Microsoft Internet Explorer provided by WhizBang! Labs

Address: <http://joi.flipdog.com/joi/>

FlipDog.com
Job Opportunity Index

November 2001 Welcome -- Tuesday, May 7, 2002

U.S. Job Supply Increases Amid Rising Unemployment

The Job Opportunity Index™ (JOI) increased for the first time in three months in October – climbing 0.7 point to 28.4 and signifying a slight increase in U.S. job supply. However, numerous factors, including a dramatic half-point increase in the national unemployment rate, made October anything but normal.

U.S. JOB SUPPLY BY REGION

- Above Average
- Average
- Below Average

UNITED STATES

November 2001 JOI: 28.4 (October: 27.7)
September Unemployment Rate: 5.4% (August: 4.9%)

Click on a region to see individual reports.
[See printable version](#)

Special Offer! Find out how you can earn a free subscription to the *JOI Report on U.S. Labor Markets* through a limited-time JOI Subscriber Referral Program!

ISM 280: C

blogpulse
a service of Nielsen BuzzMetrics

HOME ANALYSIS TOOLS SEARCH SHOWCASE ABOUT

Search the Blogosphere: Enter keyword(s) or URL Go [advanced] [help]

Home > Analysis

Key People for January 11, 2007

Prominently featured people across today's blog entries

Set new date: Thursday January 11, 2007

Leaders (most overall citations)				Bursty (biggest movers)			
Rank	Prev	Person	Tools	Rank	Person	Tools	
1.	—	President Bush	trend this	1.	Mark Sweeney	trend this	
2.	↑ 33	David Beckham	trend this	2.	Kenneth Stein	trend this	
3.	↓ 2	Britney Spears	trend this	3.	Robert Anton Wilson	trend this	
4.	↓ 3	Harry Potter	trend this	4.	Sheik Taj	trend this	
5.	↑ 19	Mr. Bush	trend this	5.	George Voinovich	trend this	
6.	↓ 4	Nancy Pelosi	trend this	6.	Mark Chandler	trend this	
7.	↓ 6	Justin Timberlake	trend this	7.	Martha Deutscher	trend this	
8.	↓ 5	Bill Gates	trend this	8.	Barbara Campion	trend this	
9.	—	Donald Trump	trend this	9.	Jason Dunham	trend this	
10.	↑ 17	Brad Pitt	trend this	10.	Steve Blake	trend this	
11.	↑ 13	James Brown	trend this	11.	Peggy Yvonne Middleton	trend this	
12.	↓ 10	Angelina Jolie	trend this	12.	David Beckham	trend this	
13.	↑ 15	James Bond	trend this	13.	Sen. Chris Dodd	trend this	

blogpulse
a service of Nielsen BuzzMetrics

HOME ANALYSIS TOOLS

Search the Blogosphere: "M

Search
2006
Showing 1-10 of 184 results

Search Results
184 messages found trend this XML get feed for this search

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/philadelphia436/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 saying last season....Sports Briefs - January 12, 2007 BASEBALL SAN FRANCISCO (AP) — Barry Bonds said he did not get at **Sweeney**, but did not deny a r...
Blog:<http://sanfrancisco.newspaper.newspaperonly.com/index.php/2007/01/12/report-says-federal-probe-underway-of-st>
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 Discovered:1 hour ago

Bonds' Failed Test
The apologies started flowing when the besieged slugger called teammate **Mark Sweeney** Wednesday night to say he was sorry about the drug-related controversy, Sweeney's agent said...One interesting element of the original story is that Barry Bonds is reported to have called **Mark Sweeney** as responsible for giving him the amphetamines....

CiteSeer Find: 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 ICMIL Grants Acousticafe - Pittsburgh... Teaching Reviews OSUWiki IRIS Semantic Deskto...

Google | citespace

Fast Effective Rule Induction (1995) (Make Corrections) (248 citations)

William W. Cohen
Proc. of the 12th International Conference on Machine Learning

View or download:
belllabs.com/orgs/ss_m195nripper.ps
jhu.edu/~sheppard/cs.60_paper8b.ps.gz
att.com/~wcohen/posts_m195nripper.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)

Links: [DBLP](#)

(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 [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,
 Carnegie Mellon University

- 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
- Carnegie Mellon University
- AT&T

Co-Authors

- Robert F. Murphy '05, '03 (2)
- Zhenzhen Kou '05, '03
- Churnki Basu '01, '98
- Anthony Tomasic '06, '05

[more \(filtered\)](#)

Talks

- University of Texas at Austin

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?

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁰⁹

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

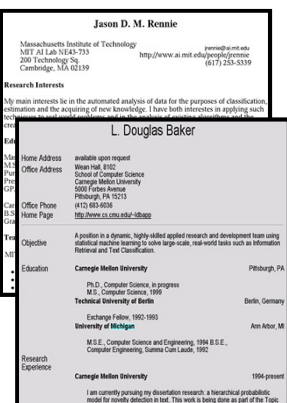
8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty Joseph Y. Halpern, Cornell University				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	Logge	Natural Language Generation	Complexity Analysis	Neural Networks	Games
739: A Logical Account of Causal and Topological Maps	116: A-System Solving through Abduction	758: Title Generation for Machine-Translated Documents	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories	179: Knowledge Extraction and Comparison from Local Function Networks	71: Iterative Widening Tristram Cazenave
Emilio Remolina and Benjamin Kuipers	Denscher, Antonis Kakas, and Bert Van Nuffelen	Rong An and Alexander G. Hauptmann	Morco Cadoli, Thomas Eiter, and Georg Gottlob	Kenneth McGeer, Stefan Wornes, and John MacIntyre	
549: Online-Execution of ecolog Plans	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 for Amplified

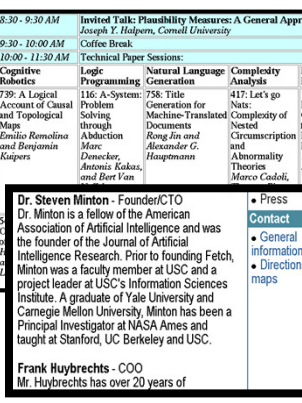
Landscape of IE Tasks (2/4): Intended Breadth of Coverage

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







ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹¹¹

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...

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

Regular set

U.S. phone numbers

Phone: (413) 545-1323

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

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.

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹¹²

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

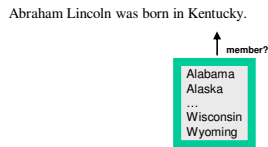
Relation: Succession
Company: General Electric
Title: CEO
Out: Jack Welch
In: Jeffrey Immelt

"Named entity" extraction

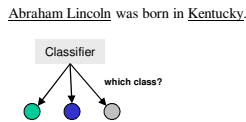
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹¹³

Models for NER

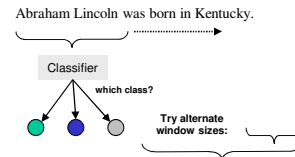
Lexicons



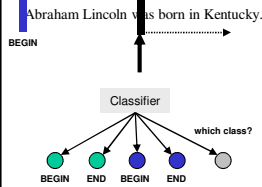
Classify Pre-segmented Candidates



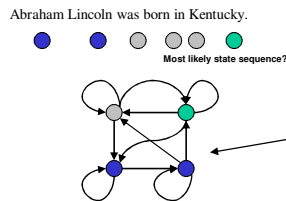
Sliding Window



Boundary Models



Token Tagging



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

HMMs, CRFs,

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹¹⁴

Sliding Windows

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹¹⁵

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.com¹¹⁶

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.com¹¹⁷

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.com¹¹⁸

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.com¹¹⁹

A “Naïve Bayes” Sliding Window Model

[Freitag 1997]

... 00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun ...

W_{t-m} W_{t-l} W_t W_{t+n} W_{t+n+l} W_{t+n+m}

prefix
contents
suffix

Estimate $\Pr(\text{LOCATION}|\text{window})$ 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: $\Pr(\text{“Place” in prefix}|\text{LOCATION})$

If $\Pr(\text{“Wean Hall Rm 5409”} = \text{LOCATION})$ is above some threshold, extract it.

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹²⁰

A “Naïve Bayes” Sliding Window Model

[Freitag 1997]

... 00 : pm Place : Wean Hall Rm 5409 Speake

W_{t-m} W_{t-l} W_t W_{t+r} W_{t+n+l} 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 $\Pr(\text{class}=+|\text{prefix},\text{contents},\text{suffix}) > \text{threshold}$, 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?

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹²¹

“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.
```

<u>Field</u>	<u>F1</u>
Person Name:	30%
Location:	61%
Start Time:	98%

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹²²

Token Tagging

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹²³

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:

■	person name
■	location name
■	background



2) Identify names based on the entity labels

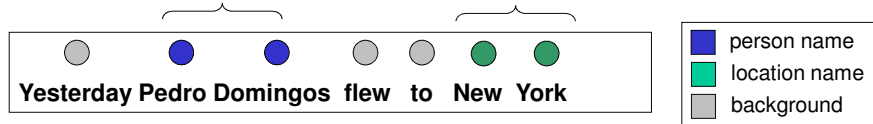
3) To learn an NER system, use YFCL.

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹²⁴

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)

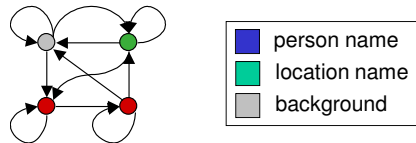
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹²⁵

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) $\arg \max_{\bar{s}} P(\bar{s}, \bar{o})$

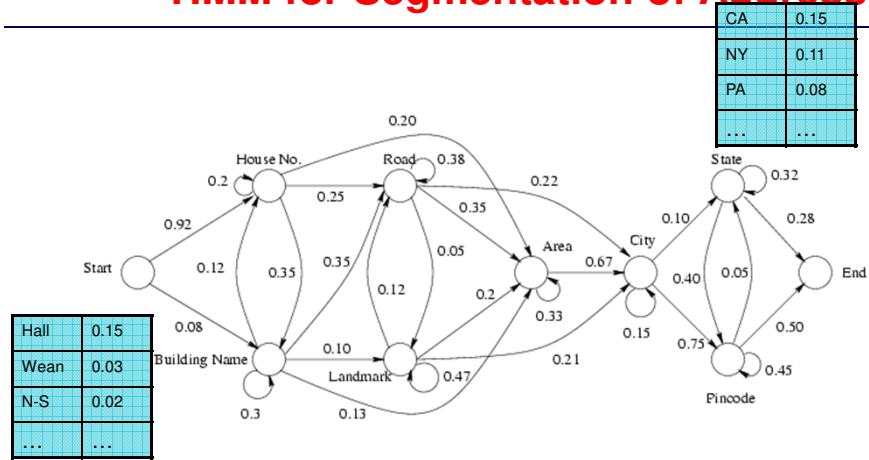


Any words said to be generated by the designated “person name”
state extract as a person name:

Person name: Pedro Domingos

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹²⁶

HMM for Segmentation of Addresses



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

[Pilfered from Sunita Sarawagi, IIT/Bombay]

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.c



HMMs for Information Extraction

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

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

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

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹²⁸

“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%

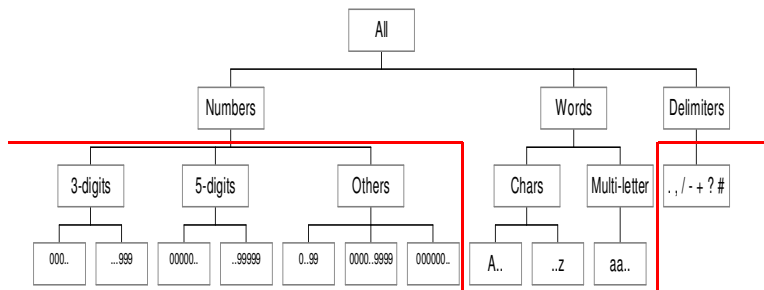
Field	F1
Speaker:	77%
Location:	79%
Start Time:	98%

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹²⁹

What is a “symbol” ???

Cohen => “Cohen”, “cohen”, “Xxxxx”, “Xx”, ... ?

5317 => “5317”, “9999”, “9+”, “number”, ... ?

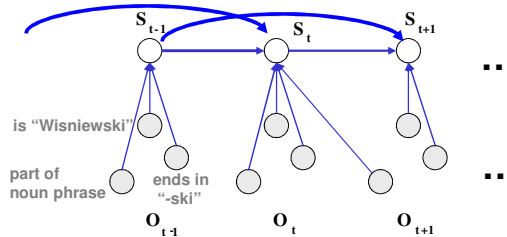


Datamold: **choose best** abstraction level using **holdout** set

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³⁰

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**

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³¹

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.**

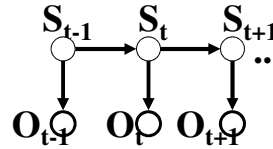
Condition	Features
w_i is not rare	$w_i = X$ & $t_i = T$
w_i is rare	X is prefix of w_i , $ X \leq 4$ & $t_i = T$
	X is suffix of w_i , $ X \leq 4$ & $t_i = T$
	w_i contains number & $t_i = T$
	w_i contains uppercase character & $t_i = T$
$\forall w_i$	w_i contains hyphen & $t_i = T$
	$t_{i-1} = X$ &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

Table 1: Features on the current history h_i

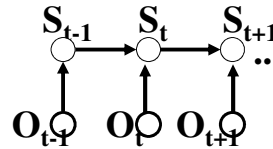
ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³²

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

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



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



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³³

HMMs vs MEMM vs CRF

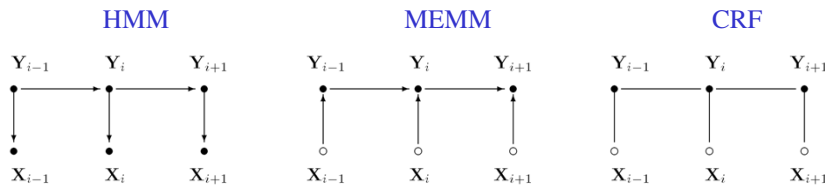


Figure 2. Graphical structures of simple HMMs (left), MEMMs (center), and the chain-structured case of CRFs (right) for sequences. An open circle indicates that the variable is not generated by the model.

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³⁴

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?

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³⁵

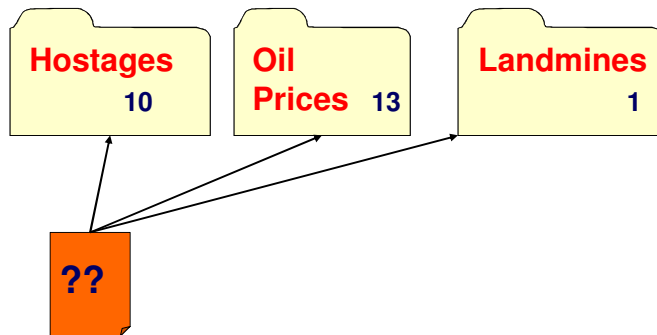
Rest of Lecture 1 Outline

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³⁶

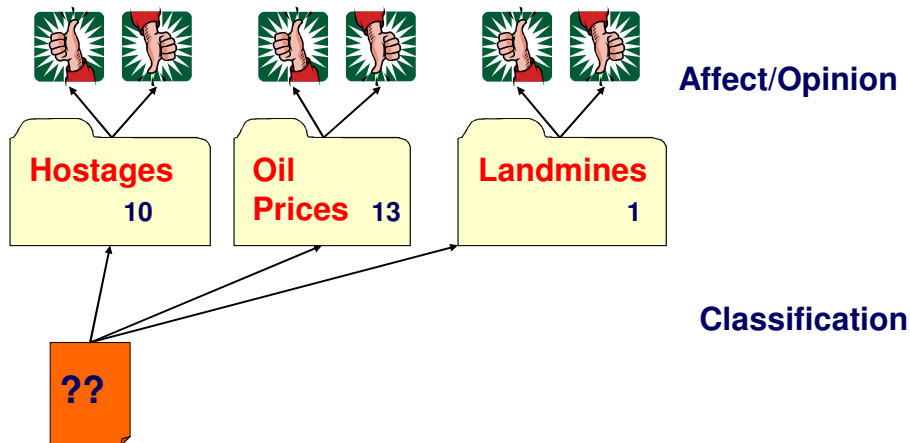
Theme 1: Text Classification

Thresholded SVMs



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³⁷

Theme 2: Affect/Opinion



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³⁸

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**

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹³⁹

Talk Outline



- **Good Classification Technology**
 - Thresholded SVMs



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



- **Process Mining**
 - Bayesian Network Approach

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁴⁰

Outline: Opinion Mining

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁴¹


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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁴²

Examples of Opinion on WWW



[Join Epinions](#) | [Help](#) | [Sign In](#)

CARS BOOKS MOVIES MUSIC COMPUTERS & SOFTWARE ELECTRONICS GIFTS HOME & GARDEN KIDS & FAMILY OFFICE SUPPLY SPORTS TRAVEL MORE...

Search for All Categories

Find and Compare

Cameras
Digital, 35mm, Lenses...

Cars & Motorsports
New, Used, Motorcycles...

Computers
PC Laptops, Printers, PDAs...

Electronics
Camcorders, Televisions, DVD...

Games
Gameboy, PS2, Xbox, Gamecube...

Gifts
Clothing, Jewelry, Fragrances...




Home & Garden
Appliances, Cooking, Tools...

Kids & Family
Toys, Strollers, Car Seats...

Media
Books, Movies, Music...

Office Supply
Supplies, Machines, Services

Quality DVD Players & Recorders

	<p>Panasonic DMR-E80 DVD Recorder / Player Rating: ★★★★★ We found this product at 58 stores The lowest price is \$493.99</p>	<p>Compare Prices View Details Read 3 reviews</p>
	<p>Toshiba SD-2900 Standard DVD Rating: ★★★★★ We found this product at 17 stores The lowest price is \$59.99</p>	<p>Compare Prices View Details Read 2 reviews</p>
	<p>Samsung DVD-R4000 DVD Recorder / Player Rating: ★★★★★ We found this product at 22 stores The lowest price is \$360.00</p>	<p>Compare Prices View Details Read 1 Review</p>

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁴³

Amazon.co.jp



買うなら今!

定価: ¥59,800

価格: ¥32,800

OFF: ¥17,200 (34%)

発送可能時期: 通常24時間以内に発送します。

製品概要・仕様:

- メーカー型番: PW-A8700
 - サイズ: 141(W)mm×106.8(D)×10.8(H)mm(閉時)
 - 重量: 約220g(電池含む)
 - 収録辞書: リーダーズ英和辞典[第2版](研究社)、リーダーズ・プラス(研究社)...
- ▶ [詳しい情報を見る](#)

ASIN: B0000DKMK5
 国外配送の制限: この商品は日本国外へお届けできません。

Amazon.co.jp 売上ランキング: 47
 カスタマーのおすすめ度: ★★★★★

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁴⁴

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. ...

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁴⁵

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.


ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁴⁶



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 *et al.*, 1999)
 - PROGENIE (Duboué *et al.*, 2003)
 - HPSearch

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs © 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁴⁸



SIGIR 2005

[Theme and goals of the workshop](#)

[Submission guidelines](#)

[Important dates](#)

[Organizers](#)

[Program committee](#)

[SIGIR 2005](#) Workshop

Stylistic Analysis Of Text For Information Access

August 19, 2005
Salvador, Bahia, Brazil

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.

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁴⁹

Lecture 1 Outline

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁵⁰

Understanding Text

25 minutes ago via Echofon

Joan Rivers What's wrong with this country? The evening news is down to 30 minutes a day but we have 65 reality shows about midgets making cupcakes?
about 1 hour ago via web

reitshamer How much market research for the iPad? "None." Mr. Jobs replied. "It isn't the consumers' job to know what they want."
<http://nyti.ms/dTCxY5>
about 1 hour ago via Tweetie for Mac

OnionSports 13-Year Old Nearing Record For Most Masturbations In A Single Day <http://onion.com/eOkwq> #SportsDome #RoadTo32
about 1 hour ago via HootSuite

stereogum The Mountain Goats – "Damn These Vampires" (Stereogum Premiere) <http://bit.ly/hYgsY3>
about 1 hour ago via web

arcadefire We are headlining Coachella Festival on Sat April 16th. Tickets will be available from www.coachella.com/tickets Fri January 21st @ 10am PST
about 1 hour ago via web

caitlinmoran Hope the press is on this: Lack of support forces Bristol woman to have daughter taken into care. <http://tinyurl.com/6hmtwuo> #respite4riven
about 1 hour ago via TweetDeck

wingoz Stats I like. This weekend 1st Time in SB era all 4 starting qbs in conference title games are 1st round draft picks.
about 1 hour ago via UberTwitter

ISM 280: n_AT_gmail.com 151

SIRI: Virtual Personal Assistant

Siri

HOME ABOUT NEWS HELP DOWNLOAD

You ask.

Siri, call me a cab

Siri does.

I just need a few details and then I can get you a taxi:

GET FREE IPHONE APP

There's a new way to get things done - Just ask Siri

No more endless clicking on links and pages to get things done on the Internet. Delegate the work to Siri and relax while Siri takes care of it for you.

Need a table for 2 at your favorite restaurant next Thursday?
Just ask Siri.

Need a taxi right now?

Latest Scoop

News

- 06/24/10 [How to Make the Most of Your iPhone 4](#)
- 04/28/10 [Apple Buys a Start-Up for its Voice Technology](#)
- 04/28/10 [What Apple's Acquisition of Siri Means for the Future of Mobile Search](#)

[read more](#)

SIRI: Virtual Personal Assistant



There's a new way to get things done - Just ask Siri

No more endless clicking on links and pages to get things done on the Internet. Delegate the work to Siri and relax while Siri takes care of it for you.

Need a table for 2 at your favorite restaurant next Thursday?

Latest Scoop

- 06/24/10 How to Make the Most of Your iPhone 4
- 04/28/10 Apple Buys a Start-Up for Its Voice Technology
- 04/28/10 What Apple's Acquisition of Siri Means for the Future of Mobile Search

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁵³

Project

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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁵⁴

-
- **Extra Slides**
- **Document Souls**

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁵⁵

Document Souls

a new paradigm for information access

James Shanahan* and Gregory Grefenstette*



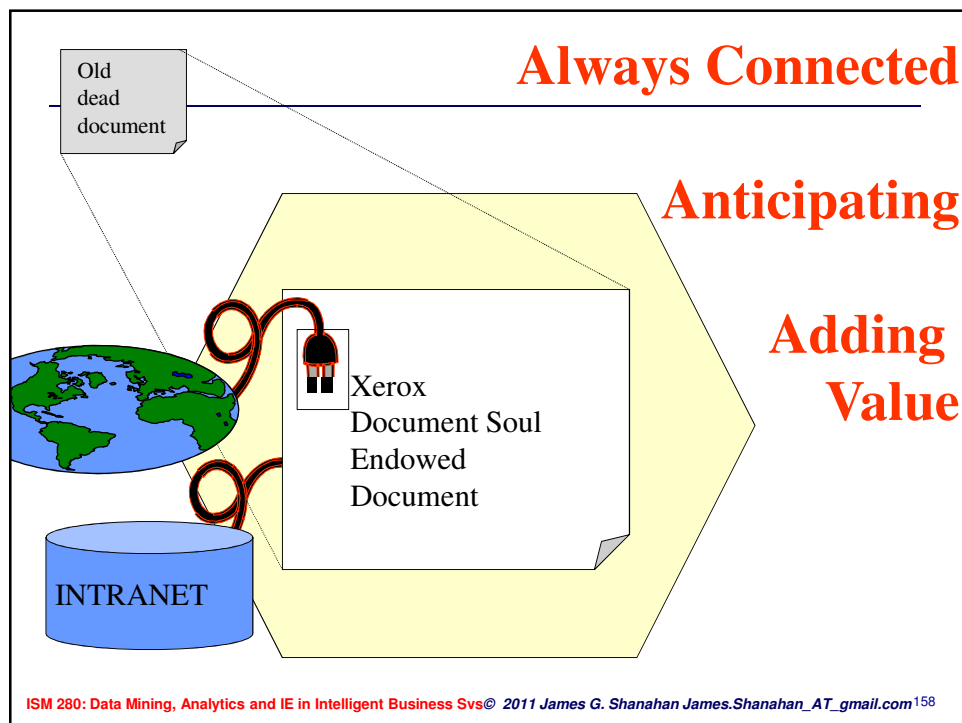
*** *Work performed at Xerox Research***

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svcs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁵⁶

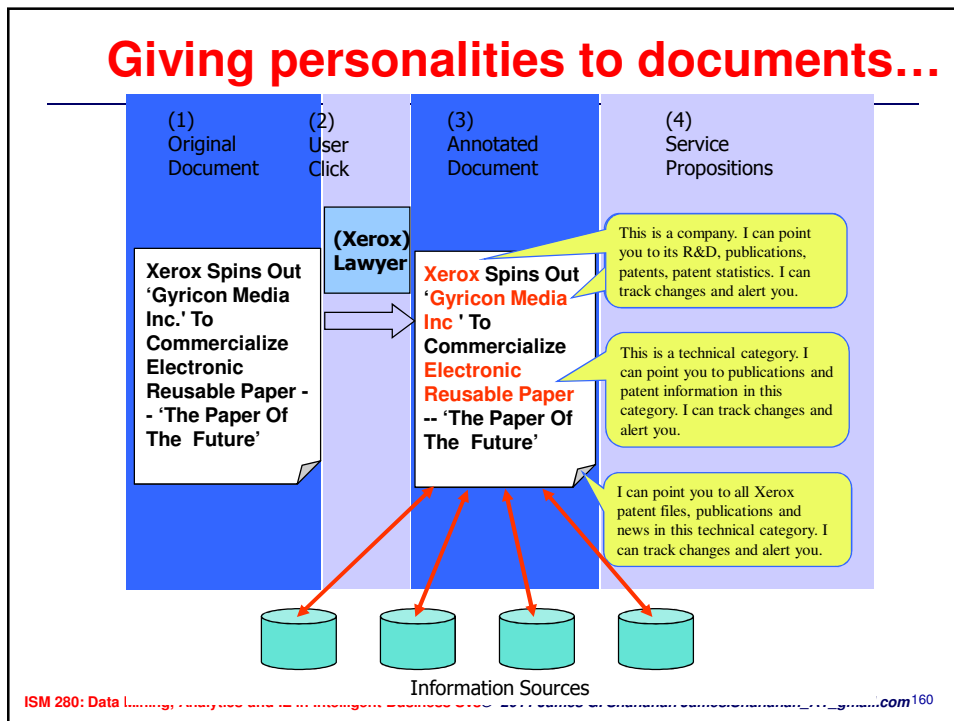
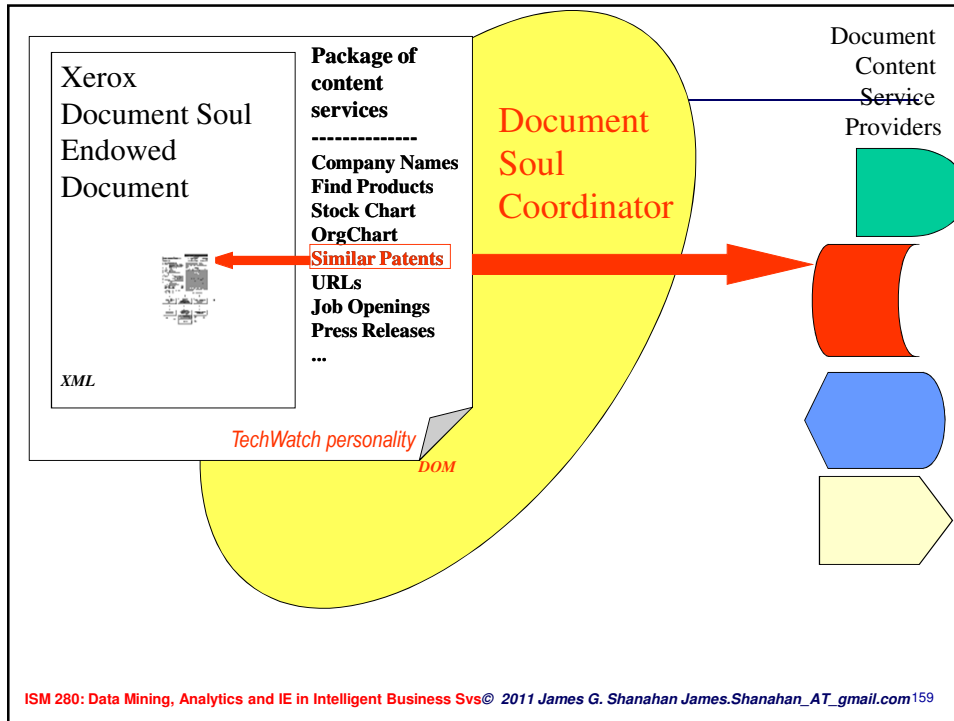
Interesting juncture

- 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

ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁵⁷



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁵⁸





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



ISM 280: Data Mining, Analytics and IE in Intelligent Business Svs© 2011 James G. Shanahan James.Shanahan_AT_gmail.com¹⁶³