Text Mining Lessons Learned from Real Applications

ASIAS Technology and Tools Symposium

July 27, 2009

Elder Research, Inc.
300 W Main St., Suite 301
Charlottesville, Virginia 22903
434-973-7673
www.datamininglab.com
Elder@datamininglab.com

Elder Research, Inc.
Data Mining & Pattern Discovery

Outline

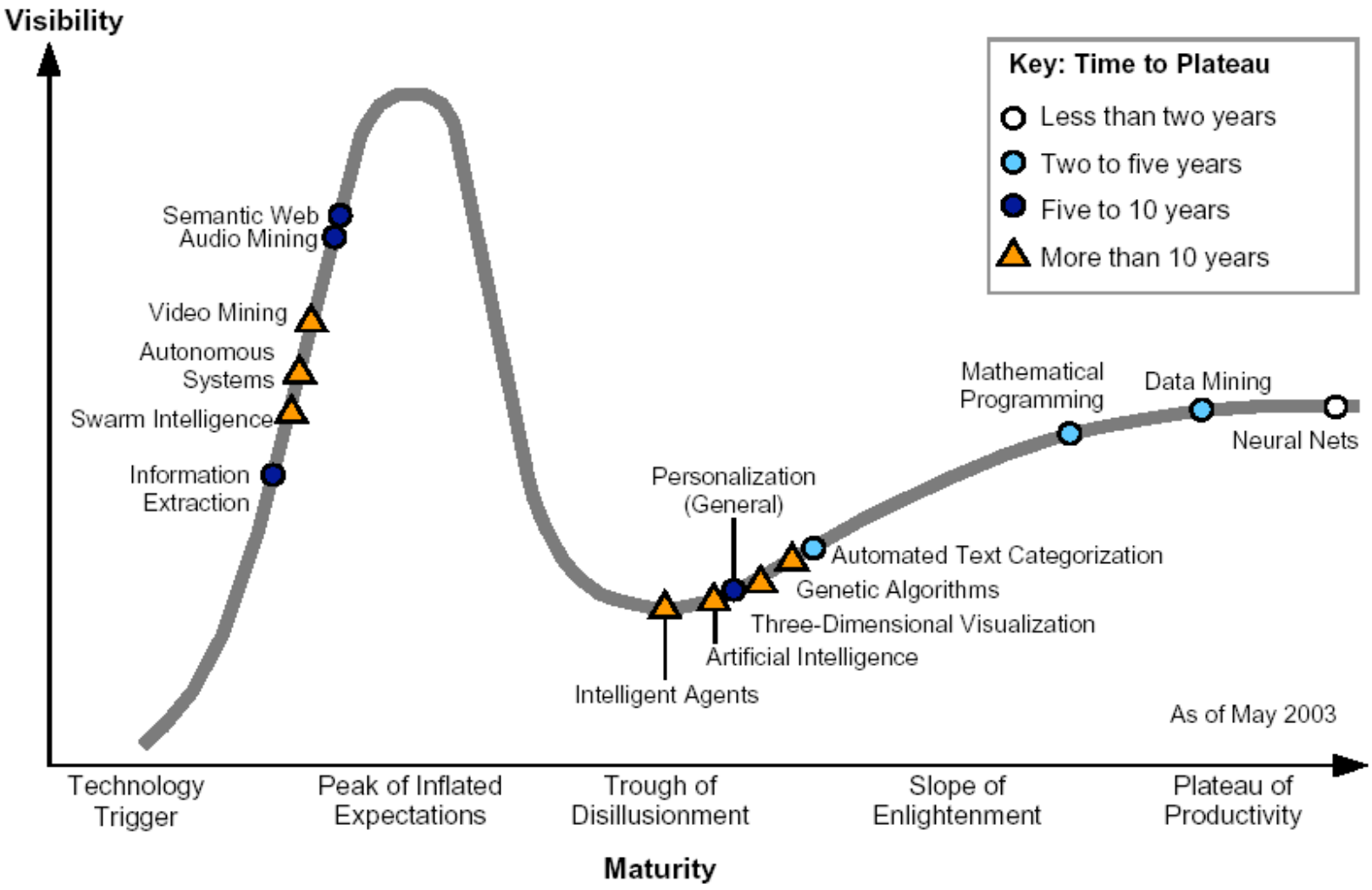
- Humans v. Machines?
- Mining tables and text for valuable information
  - Application areas at Elder Research
  - Ex: IRS Fraud detection
- Example Government text mining projects
  - Risk Profiling for NSA
  - Prioritizing CBP searches
  - Quick decisions for SSA disability
  - Document discovery for NGIC
  - Disease discovery for NCMI
- Factors for data/text mining project success

"Of course machines can think. After all, humans are just machines made of meat."
- MIT CS professor

Our view:

Human and computer strengths are more complementary than alike.

Data Mining and the Hype Cycle



Source: Gartner Group

Selected ERI Clients

Text Mining
- Social Security Administration
- National Ctr. for Medical Intelligence
- US Customs and Border Protection
- National Ground Intelligence Center
- Digital Reasoning Systems
- US Dept. Education, NIDRR
- Northrup Grumman
- Ctr. for Navy SEAL Operations
- US Army Tank Automotive RDEC

Investment
- Westwind Foundation
- Oppenheimer Funds
- CPTR
- Vantage Consulting Group
- Two Rivers Capital Mgmt.
- N. de Rothschild Holdings
- SAC Capital Advisors
- R-Squared Trading, LLC
- Energy Service Providers

Biometric & Pharmaceutical
- AstraZeneca
- Pharmacia & UpJohn
- SmithKline Beecham Pharm.
- Rio Grande Medical Technology
- Epsilon Group
- Lumidigm
- Dekalb Genetics
- VeriLight

CRM & Cross-Selling
- Hewlett Packard
- AAA Michigan
- Woodworker's Supply
- Subscription Partners
- HSBC

Behavioral & Web Analytics
- PDI Solutions
- 3eDC
- Richmond Police Dept.

Collaborative Filtering
- FindMoreFives.com (Netflix)

Tool Evaluations
- Defense Finance & Accounting Service
- Fair Isaac
- Electronic Warfare Assoc.

Reliability
- Lockheed Martin

Entity Extraction & Link Analysis
- Department of Defense
- Army Labs

Data Mining Technology
- Lucent

Data Warehousing & Business Intelligence
- Darden Solutions
- Albemarle County
- Georgetown University

Image Recognition
- Anheuser-Busch

Credit Scoring
- Capital One
- Dealer Services
- Grupo Ficohsa (Honduras)
- Dollar Financial Group

Training & Courses
- Dozens of Corporations, Government Agencies, Non-Profits, Universities, and Professional Associations

Optimization & Simulation
- Peregrine Systems
- Westwind Foundation
- NuTech Solutions
- Finch Asset Mgmt. (Bermuda)
- Commonwealth Comp. Research

Fraud Detection
- Internal Revenue Service
- Defense Finance & Accounting Service
- Federal Data Corporation
- Mantas
- Hewlett Packard

Example: Government, Financial Enforcement, Large-Scale Production System, Strong Success

ERI Support to IRS Office of Refund Crimes

- ERI served as technical lead on a multi-year program to improve the detection of refund fraud in individual tax returns
  - Objective was to prioritize returns based on likelihood of fraud in order to optimize scarce resources (i.e., investigative analysts)
- Data mining system far out-performed existing methods
  - Found orders of magnitude more fraud than previous approach
  - Reduced workload with increased reliability, allowing analysts to focus on most promising and challenging cases
  - Improved refund time for honest filers because fewer "good" returns were held for initial review
- Resulting system was rolled-out nationwide ahead of schedule and with no adverse impact on existing workflow processes

ERI Support to US Customs & Border Protection Targeting & Analysis Systems Program Office

- Detecting unusual patterns of activity in land border crossings to identify criminal activity
- Mining free-form text to accurately identify commodities in sea-going transport
  - Mapping free-form text descriptions to established CBP codes which are used to identify unusual patterns of behavior

Social Security Administration

- President inquired why the SSA can't more quickly approve applicants with severe diseases
- Process can take up to 2 years
- Applicants need to be poor and unable to work (anywhere)
- Half of appealed applications are overturned
  - 5 layers of appeal
  - applicants actively worsening
- Commissioner Barnhart promised that SSA would use 21st century technology for Quick Approval of a sure subset

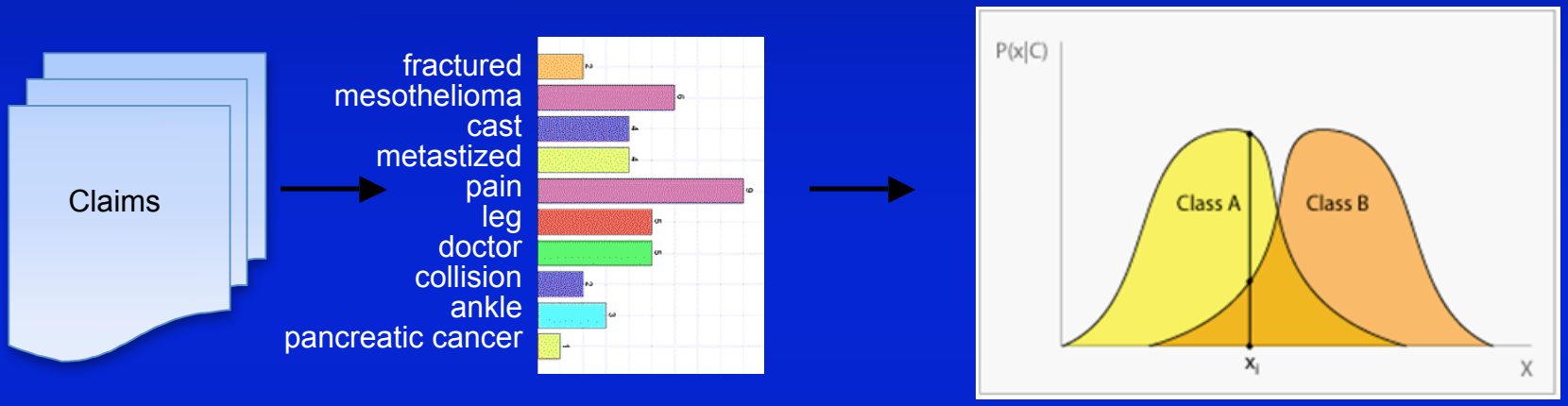
Text Mining Challenges

- Multi-Word Phrases (Concepts or Lemmas)
- Stemming (ex: Learn = Learning, Learned, Learns…)
- Synonyms (ex: ALS = Lou Gherig's Disease)
- Misspellings

Ex: 51 phrases found (by SPSS Clementine) for "Learning Disability"

learning disablitiy, learning deisability, learning disablity, learining disabilities, learning disabiality, learningdisabilty, learning disabiilty, learning disabilitty, learning disabilty, learning disablety, learning disabilitiy, learnoing disability, learning disabilties, learning disabiltiy, learning dsblty, learning disibility, learnings disabilty, learningdisability, larning disabilities, learning disabilitiies, learning disabilitties, learning disibilities, learning diasability, learning dasability, learnning disability, learning disabilities, lerning disability, learning disabilites, learneing disability, learninig disability, learning disaiblities, leraning disability, learning disaiblity, learnings disability, learning disabilitys, learning disabillity, learnings disabilities, learning diasbility, learning disabiliites, learning dsiability, learning disabliity, learning disibilty, learning disbilities, learning disbality, learning disbility, learning disabilit, learningdisabilities, learningi disability, lerniung disabilities, learning disabliities, learning disaability, learning disablities

Technology Approach at SSA: Bag of Words / Density-based Classification

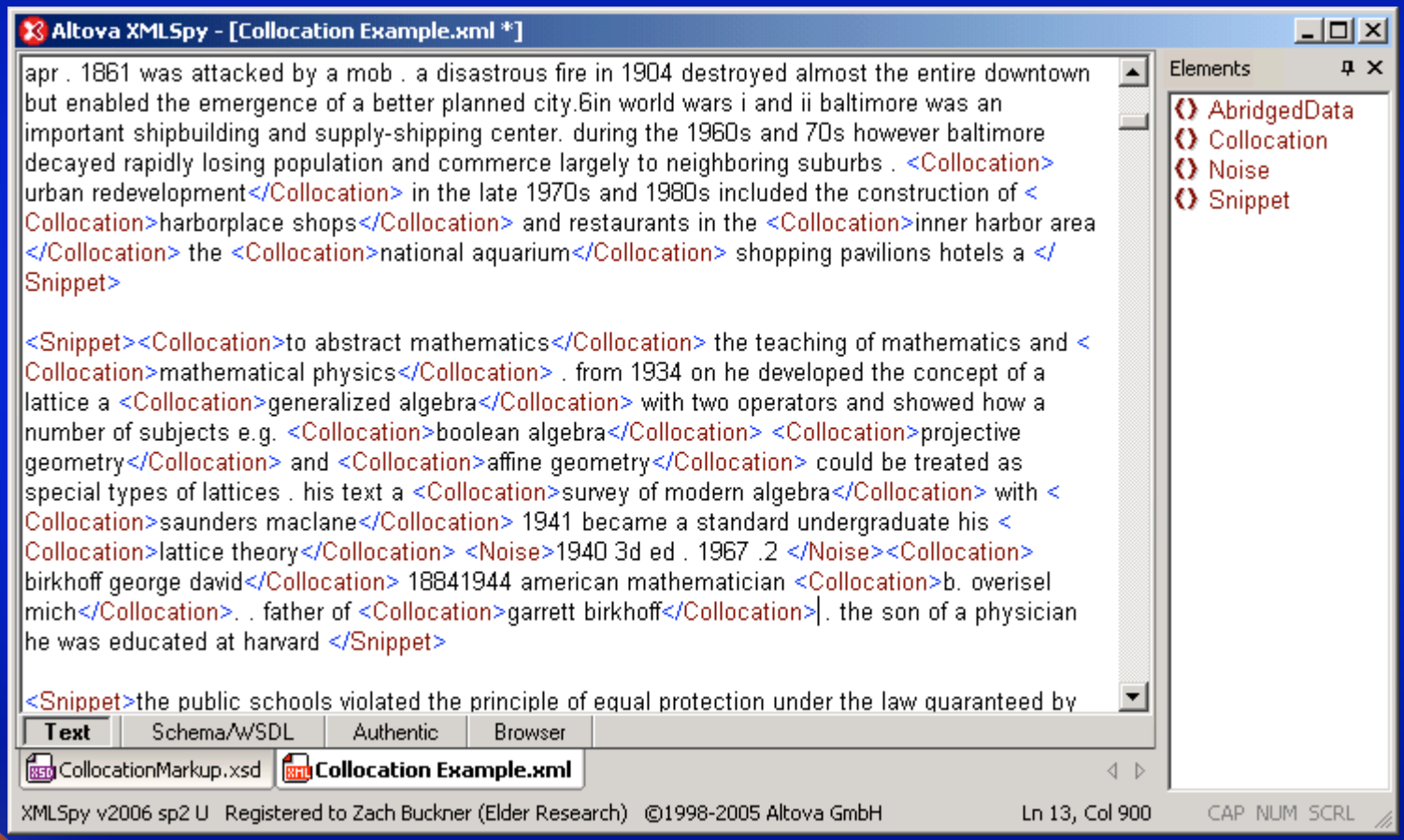


P(C|X) is the probability of case *X* being in class *C*. Assign the class -- considering *P* and the *costs of misclassifications* -- that leads to the *minimum cost* decision.

30% baseline -> 90% model accuracy

Elder Research, Inc.
Data Mining & Pattern Discovery

Inductive Modeling Approach: Label cases for training (using analysts)



Build Statistical Model from Labeled Data

Microsoft Excel - bigram.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF — Type a question for help

E528 | at

	C	D	E	F	G	H	I	J	K	L	M	N	
1	V1	V1 Co	V2	V2 Co	L	C(V1)	C(V2)	C(V1V	Mean O	STD Ou	Mean In	STD In	
517	at	TRUE	shiat	FALSE		4628	22	1	8.66E+00	6.65E+01	1.69E+00	1.32E+00	sunni people kept firing AT SH
518	leave	FALSE	except	FALSE		515	1240	1	5.08E+00	9.00E+00	2.81E+00	1.16E+01	shall neither take nor LEAVE E
519	activity	FALSE	that	TRUE		109	25965	1	4.04E+00	8.17E+00	8.79E+00	5.69E+01	security plan for each ACTIVIT
520	al-masjid-al-haram.	FALSE	.	TRUE		1	279751	1	1	0	8.73E+00	1.71E+02	any other mosque excepting A
521	_o_8_	FALSE	.	TRUE	2	4	279751	4	4	0	8.73E+00	1.71E+02	l_c_o_8_l q_O_8_._ _ . l_l
522	30-31	FALSE	.	TRUE	2	4	279751	2	1.33E+00	5.77E-01	8.73E+00	1.71E+02	. koran 13 . 30-31 . they also say
523	scatter	FALSE	you	TRUE		9	15384	1	1	0	1.51E+01	8.08E+01	other ways which will SCATTE
524	the	TRUE	corpus	FALSE		130279	3	1	1.25E+01	1.06E+02	1	0	. you may keep THE CORPUS
525	.	TRUE	3.36	FALSE		279751	1	1	2.07E+01	3.66E+02	1	0	from the outcast satan . 3.36 -
526	shawahid	FALSE	al-tanzil	FALSE	1	4	2	2	1.33E+00	5.77E-01	2	0	v3 . p371 6 SHAWAHID AL-TA
527	smelled	FALSE	him	TRUE		2	22843	1	1	0	1.81E+01	2.70E+02	and kissed him and SMELLED
528	matters	FALSE	at	TRUE		143	4628	1	2.88E+00	4.99E+00	3.36E+00	1.34E+01	and i discussed such MATTEI
529	those	FALSE	surrendered	FALSE		4254	47	1	9.27E+00	1.12E+02	2.24E+00	4.44E+00	but one house of THOSE SUF
530	other	TRUE	soul	FALSE		2762	331	1	3.47E+00	1.53E+01	5.61E+00	1.20E+01	at all for any OTHER SOUL . th
531	real	FALSE	prophecy	FALSE		147	28	1	1.59E+00	1.26E+00	3.11E+00	3.95E+00	real interpretation of this REA
532	not	FALSE	masterminded	FALSE		14173	1	1	7.14E+00	3.06E+01	1	0	the attacks applauded if NOT I
533	of	TRUE	architecture	FALSE		76432	38	8	1.11E+01	2.29E+02	1.81E+00	1.81E+00	issue of the use OF ARCHITE
534	imam	FALSE	in	TRUE		1328	31692	15	7.78E+00	3.13E+01	5.44E+00	4.54E+01	the qur'an by the IMAM IN the
535	.	TRUE	thenceforth	FALSE		279751	2	1	2.07E+01	3.66E+02	1	0	bright gleam of dawn . THENC
536	al-isra'	FALSE	.	TRUE		1	279751	1	1	0	8.73E+00	1.71E+02	an-nisa' . 163 & Al-ISRA' . 55 pe
537	weak	FALSE	commodity	FALSE		171	25	1	2.80E+00	6.83E+00	1.67E+00	1.50E+00	rains in 2001 . WEAK COMMO
538	was	TRUE	listening	FALSE		13895	58	6	6.82E+00	3.40E+01	2	1.69E+00	a man and imran WAS LISTEN
539	my	TRUE	auliya'	FALSE		4914	15	2	5.55E+00	2.07E+01	1.50E+00	7.07E-01	helpers and protectors . MY A
540	us	FALSE	operations	FALSE		3579	30	1	7.41E+00	4.80E+01	1.30E+00	6.57E-01	countries to condemn the US
541	benhamou	FALSE	.	TRUE		1	279751	1	1	0	8.73E+00	1.71E+02	initiatives for development mc
542	when	TRUE	protecting	FALSE		7404	76	1	1.33E+01	9.25E+01	2.42E+00	2.73E+00	person during self-defense or
543	labors	FALSE	among	FALSE		4	1872	1	1.33E+00	5.77E-01	2.96E+00	9.86E+00	labor of any that LABORS AM
544	muslims	FALSE	alone	FALSE		1285	429	1	4.84E+00	2.20E+01	3.30E+00	8.41E+00	not belong to the MUSLIMS A
545	considered	FALSE	praiseworthy	FALSE		222	15	1	2.31E+00	4.60E+00	1.88E+00	2.10E+00	possessors of which are CON
546	fis's	FALSE	armed	FALSE		1	65	1	1	0	1.91E+00	2.42E+00	by the late-1990s and FIS'S AF

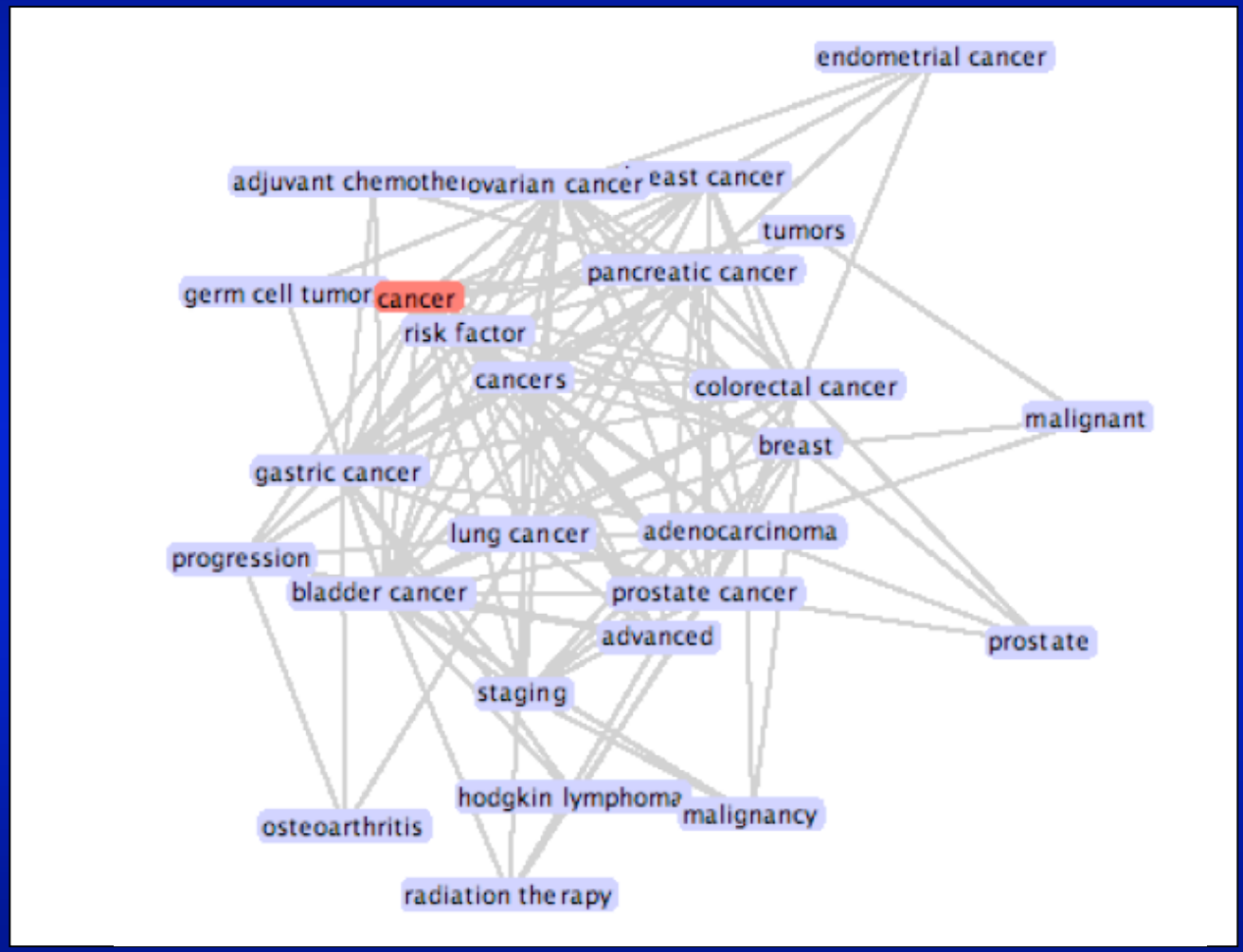
Bigrams | Common | Training

Ready

Apply Model to New Data

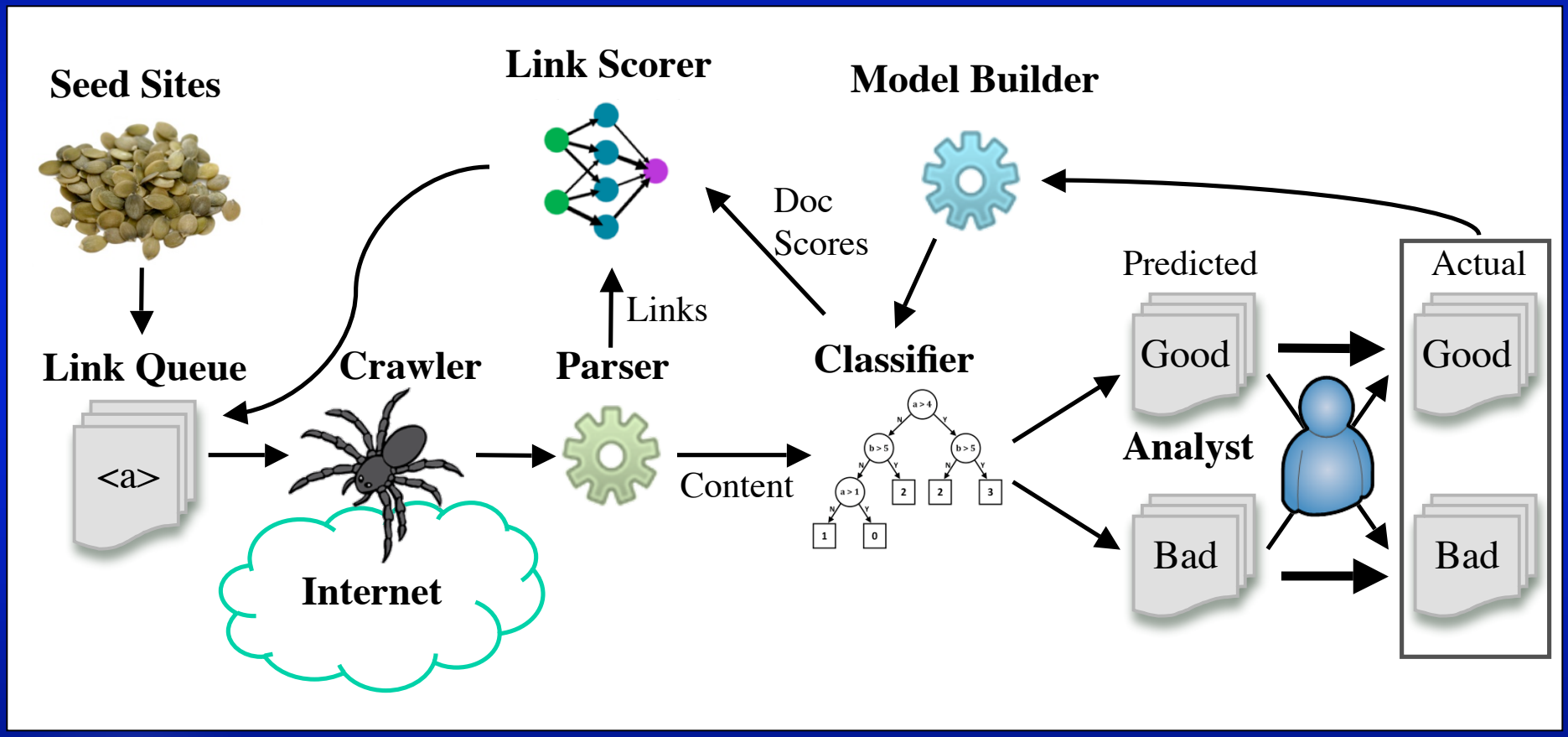
((Ahmed Omar) (Saeed Sheikh)) Known as "Sheikh Omar", ((Ahmed Omar) (Saeed Sheikh)) is a (British citizen) of Pakistani descent, with links to various Islamic-based (terrorist organisations), including Al-Qaeda and Harkat-ul-Mujahideen. He has been mentioned in many conspiracy theories linking him with the CIA and the ISI, Pakistan's intelligence agency, for whom he was allegedly an informer. Many of the sensationalist and inaccurate reports - the basis for the conspiracy theories involving him - arose in the confusion of the early weeks after the (9/11 attacks), when investigators did not have a (clear picture) of the plot, and followed a great many leads that later turned out to be (dead ends). As of 2005 the suspicion that Sheikh (played a part) in the funding of the (9/11 attacks) has been fading. However, as the factual support for his involvement has crumbled, conspiracy theories about him have thrived. Much of what follows is perhaps best seen in that context. In his youth he attended Forest School Snaresbrook, a (public school) in (North-East London), whose alumni include England cricket captain Nasser Hussain. He also attended the prestigious London School of Economics. (The Times) describes Saeed Sheikh as "no ordinary terrorist but a man who has connections that reach high into Pakistan's (military and (intelligence) elite) and into the innermost circles of Osama Bin Laden and the (al-Qaeda organization)." According to ABC, Sheikh began working for the ISI in 1993. By 1994 he was operating (terrorist training camps) in Afghanistan and had earned the title of (bin Laden's) "special son." At the time, the Taliban were beginning to dominate Afghanistan, much due to support received from the ISI. In (May 2002), the Washington Post quotes an unnamed Pakistani as saying that the ISI paid Sheikh's legal fees during his 1994 trial in India on charges of kidnap. However, this claim has not been confirmed by any other source. An unnamed senior-level U.S. government (source told) CNN in (October of 2001) that (U.S. investigators had dicovereed) that someone using the alias (Mustafa (Muhammad Ahmad)) possibly ((Ahmed Omar) (Saeed Sheikh)), allegedly a long-time ISI informer, had sent about $100,000 from the United Arab Emirates to Mohammed Atta, the suspected hijack ringleader of the (September 11, 2001 attacks). "Investigators said Atta then distributed the funds to conspirators in Florida in the weeks before the deadliest (acts of terrorism) on U.S. soil that destroyed the World Trade Center, (heavily damaged) the Pentagon and (left thousands) dead. In

Collocation & Association Networks (National Ground Intelligence Center)

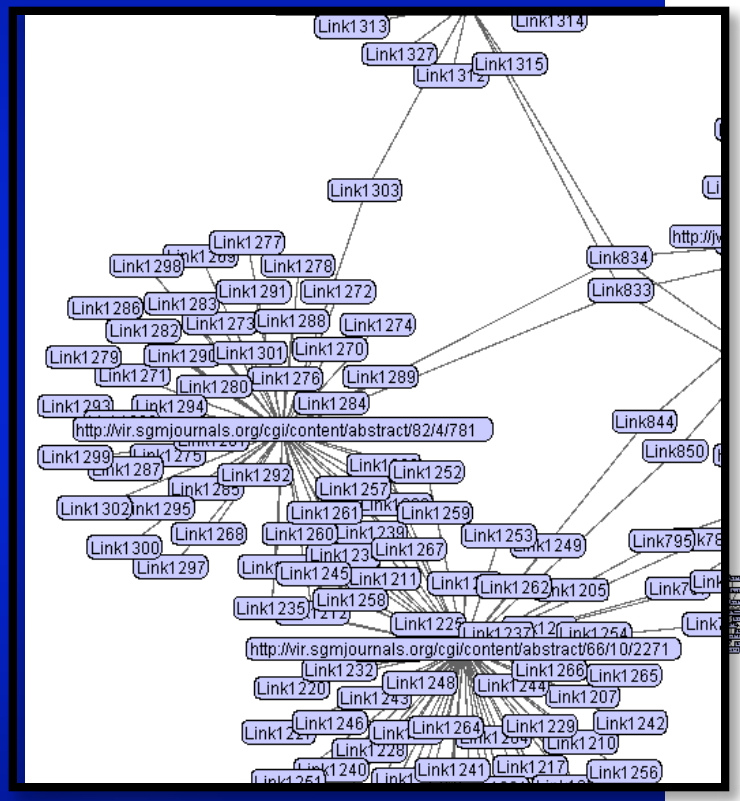


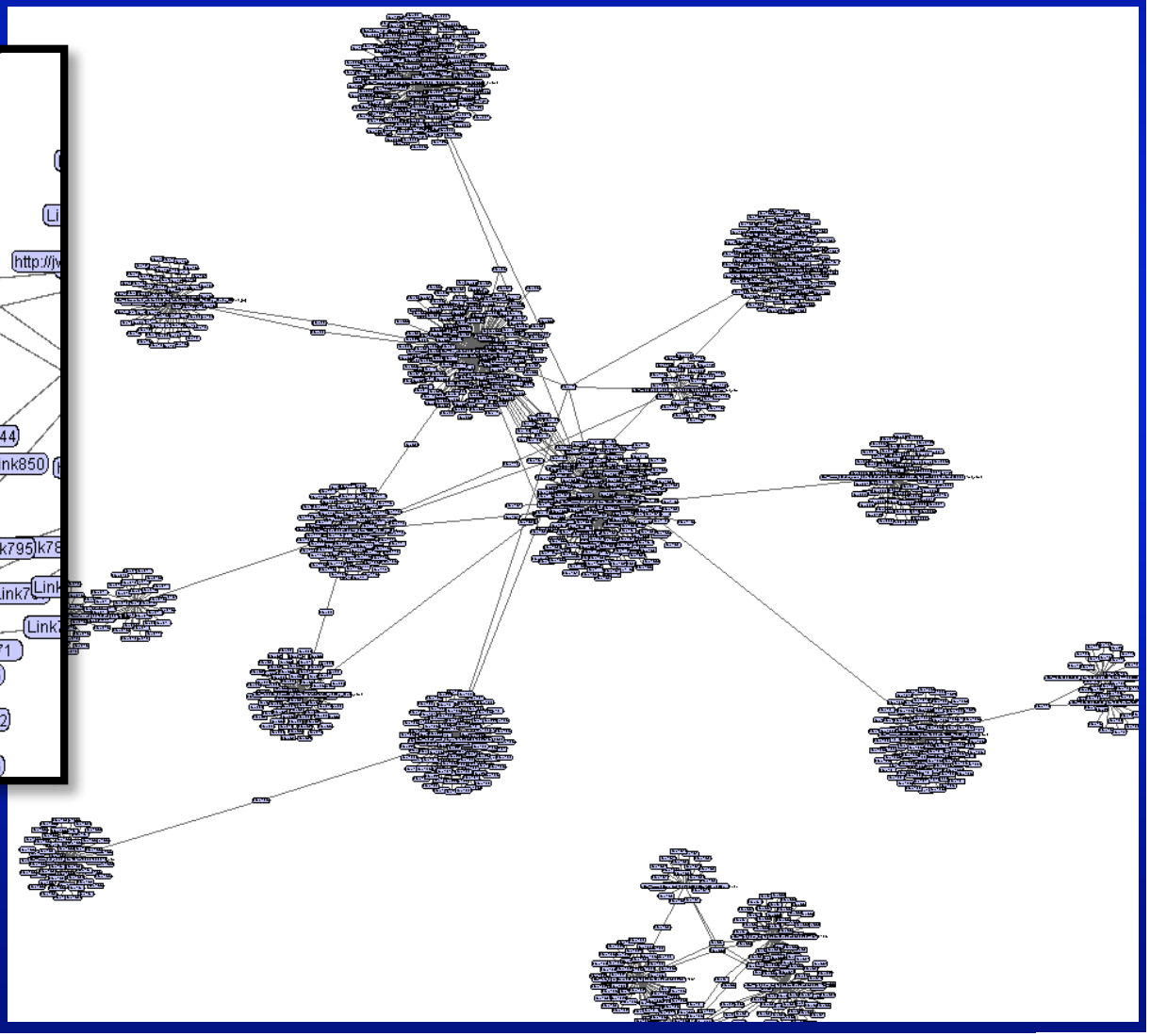
Improving Search with Collocation and Association

Monitoring the Web for Infectious Animal Diseases (National Center for Medical Intelligence)



Exploring Web of References





Elder Research, Inc.
Data Mining & Pattern Discovery

Visualize Document Similarity (ERI Tool)

Elder Research, Inc.
Data Mining & Pattern Discovery

Documents are linked if *content* is similar (even if no hyperlink)

Prioritize Documents for Analyst Attention through Discovered Clusters

Elder Research, Inc.
Data Mining & Pattern Discovery

Those close to document of interest are likely worth exploring

Necessary Ingredients for Data/Text Mining Project Success

- Gain Expected: either:
  - Leverageable - an incremental improvement will matter, or
  - "Low-hanging fruit" - nobody's yet dared attack the problem
- Interdisciplinary Team: experts needed in business area, statistics, algorithms, and databases
- Data Vigilance: capture and maintain the accumulating information stream
- Time: learning occurs over multiple cycles
- Business Champion is essential

How to Manage Data/Text Mining Projects

- Assess data assets (what treasure could be hidden in our sludge?)
  - Data caretaker must be on board
- Identify pain points in current production process
  - What improvements would have the biggest impact?
- Brainstorm ideal process
  - External expertise extremely efficient here
- Conduct a pilot project. Simultaneously:
  - "Hit a single": automate key task, create dashboard, or graphic
  - "Swing for fences": attack core weakness
- Have key staff work closely with analytic experts
  - Transfer technology
  - Internalize essential steps
- Prove ROI. Make allies and decision-makers look good

John F. Elder IV
Chief Scientist, Elder Research, Inc.

Dr. John Elder heads a data mining consulting team with offices in Charlottesville, Virginia and Washington DC (*www.datamininglab.com*). Founded in 1995, Elder Research, Inc. focuses on investment, commercial and security applications of advanced analytics, including text mining, stock selection, image recognition, process optimization, cross-selling, biometrics, drug efficacy, credit scoring, market timing, and fraud detection.

John obtained a BS and MEE in Electrical Engineering from Rice University, and a PhD in Systems Engineering from the University of Virginia, where he's an adjunct professor teaching Optimization or Data Mining. Prior to 14 years at ERI, he spent 5 years in aerospace defense consulting, 4 heading research at an investment management firm, and 2 in Rice's *Computational & Applied Mathematics* department.

Dr. Elder has authored innovative data mining tools, is a frequent keynote speaker, and was co-chair of the 2009 *Knowledge Discovery and Data Mining* conference, in Paris. John's courses on analysis techniques – taught at dozens of universities, companies, and government labs – are noted for their clarity and effectiveness. Dr. Elder was honored to serve for 5 years on a panel appointed by the President to guide technology for National Security. His book on Practical Data Mining, with Bob Nisbet and Gary Miner, appeared in May 2009.

Elder Research, Inc.
Data Mining & Pattern Discovery

John is a follower of Christ and the proud father of 5.