Improving Analysis with Information Extraction Technology

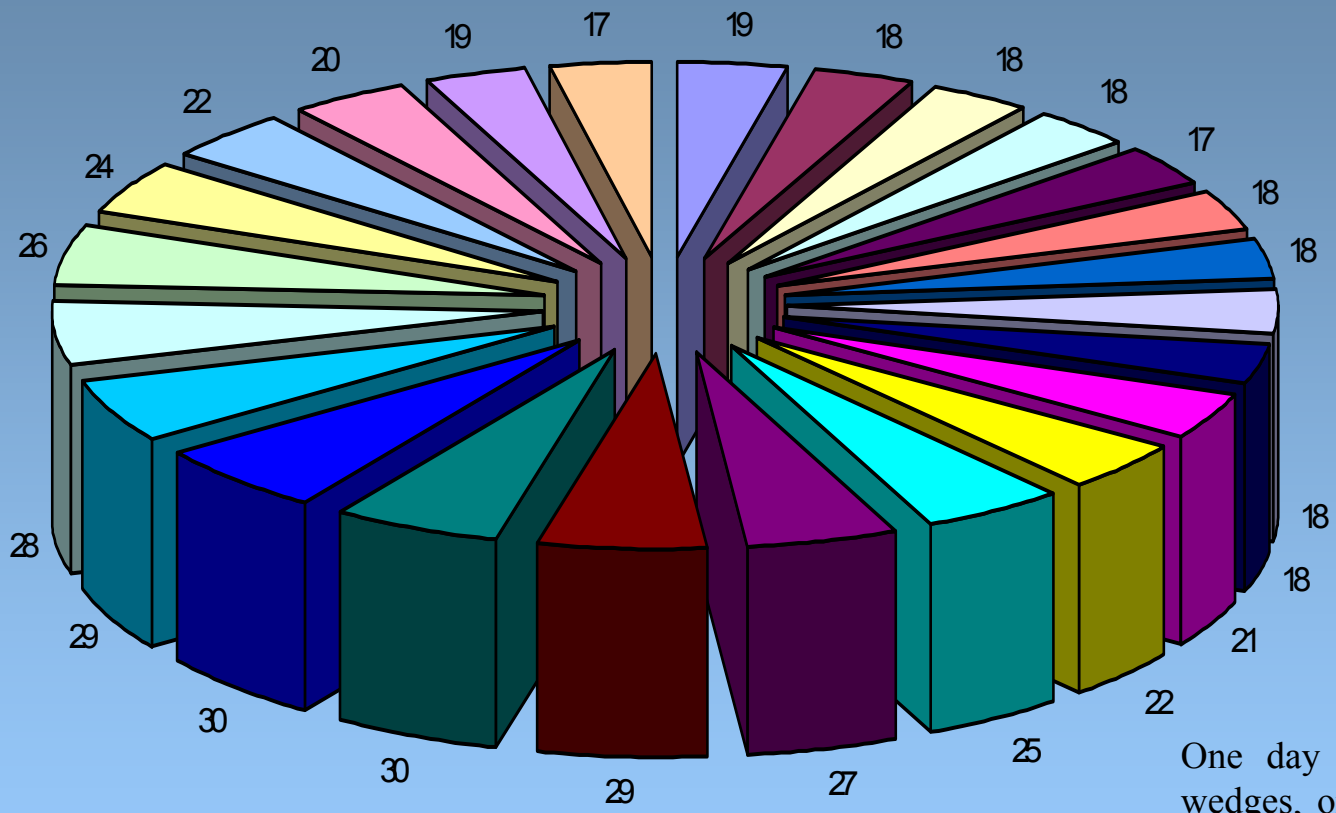
June 2003

Sarah M. Taylor, PhD
Principal Systems Engineer
Global Information Management Systems
Lockheed Martin M&DS
Sarah.M.Taylor@lmco.com

Outline

- Presenting data
- A problem and an opportunity
- Information extraction state of the art
- Benefits of IE
- Examples of IE plus visualization supporting better analysis
- Conclusion

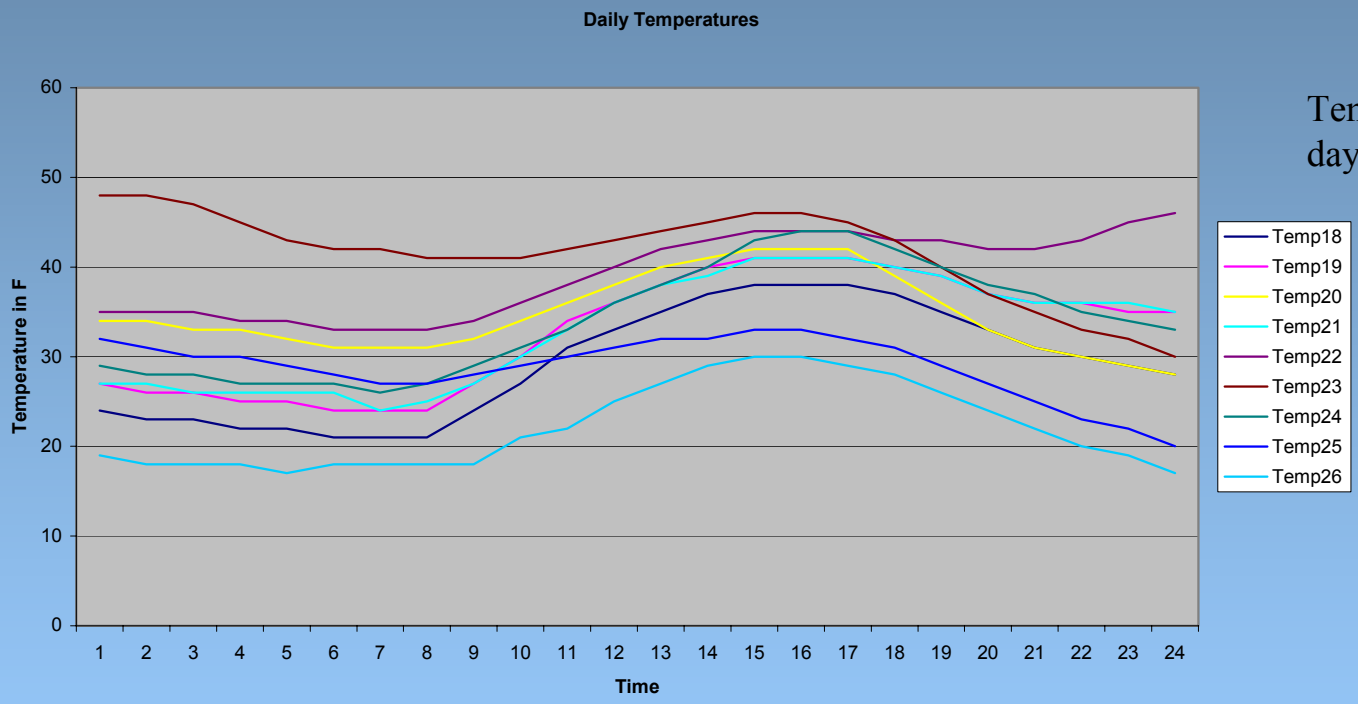
Presenting Data – an example



This is eye-catching, but no pattern is obvious….

One day in February (There are 24 wedges, one for each hour of the day; the numbers are the temperatures for each hour. The size of the wedge represents the share of that temperature in a total of all temperature readings for the day.)

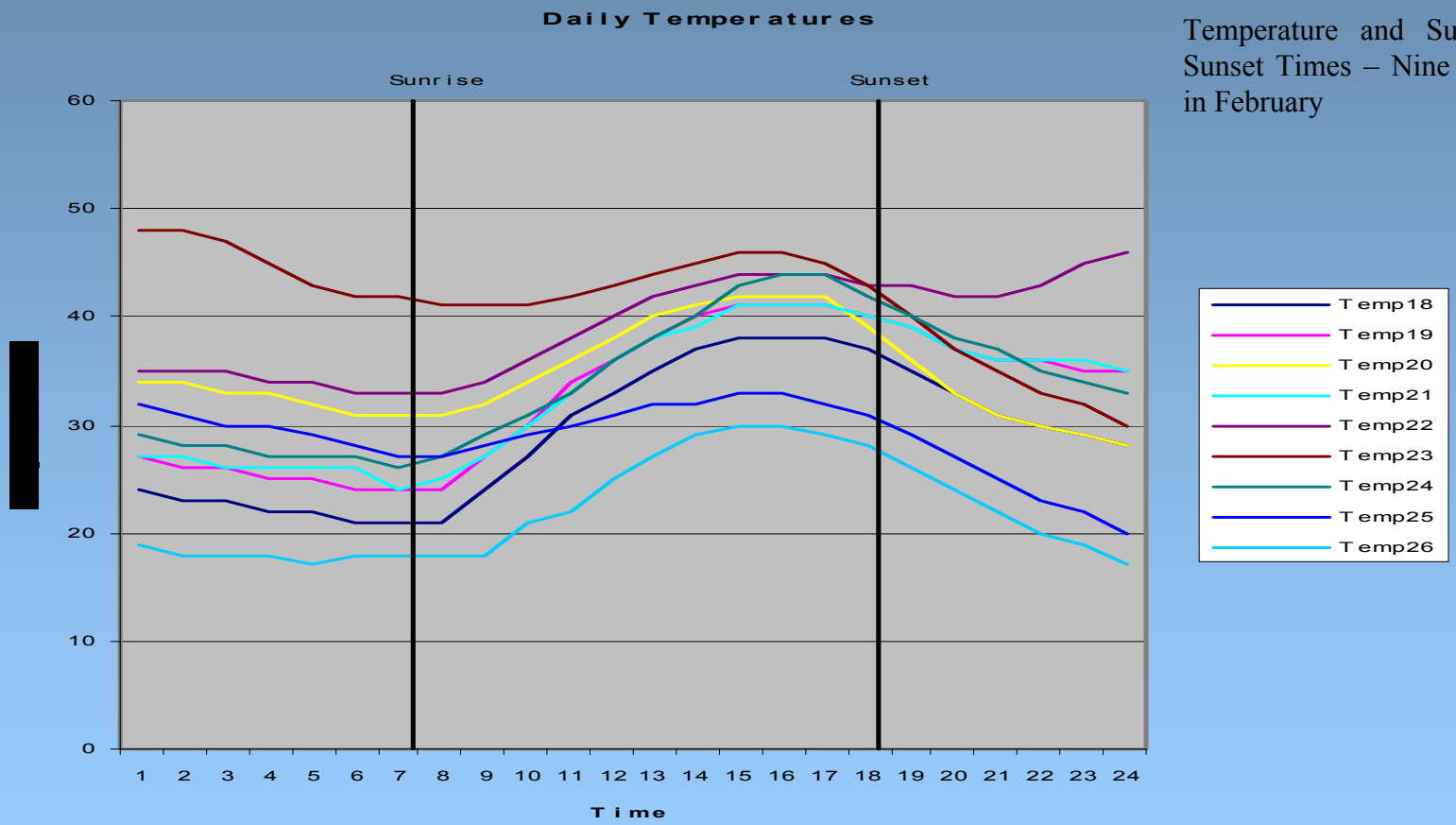
Same data – different picture



Temperature – Nine days in February

Here the pattern emerges, easily, to the eye……

Same picture – add some data



Temperature and Sunrise, Sunset Times – Nine Days in February

Cause and effect become visible.

A Problem and an Opportunity

- Problem: Information Overload
- Opportunity: Use more information to support more thorough and complete analysis
- Requirement: Good visualization and analysis tools
- Requirement: Good information extraction technology
- Results: Ability to aggregate information from text, discover broad trends and patterns, as well as find the “nuggets”

What is Information Extraction?

- Find pre-specified types of information in text and format it so it is useful –
  - Entity extraction: find all the person names, organization names, dates, locations, facilities, identification numbers, equipment names
  - Relationship extraction: find person to person, person to organization, or more specific links such as person to family member, employee to organization, links stated in the text
  - Event extraction: find sales, acquisitions, travel, attacks, drug seizures and their associated actors, locations, means, results
- Typically, this extracted information is placed in a database, or tagged with xml or related tagging schema
- Information can be used for everything from text summaries, to improving full text indexing, to improving text categorization and clustering, to link analysis and displays, gis overlays and linking, timelines, transaction analysis and other types of data mining.

State of the Art in Information Extraction

- Several companies working on finding commercial applications – business intelligence, pharmaceutical research, financial services compliance
- US Government was initial funder of this technology and remains the major interested customer
- Most systems use a human built rule-base, although there is increasing interest in augmenting with automated rule development of some kind
- Out of the box accuracy (e.g. with NO tuning) –
  - Entity extraction: Excellent to Adequate depending on system and task
  - Relationship extraction: Adequate to poor, depending on system and task
  - Event extraction: No real out of the box capability, must be tuned

Benefits of Information Extraction

- Moves us away from document by document review as the sole means to exploit text
- Dramatically (4x – 10x) increases the speed with which an analyst can review material
  - Finding and focusing on relevant material more specifically than search and retrieval alone
- Speeds link and transaction analysis
  - Analyst can get the basic idea, often, without first entering data or correcting input
- Supports data mining over large volumes of text, which is otherwise impossible
  - Analyst can repeatedly query data from different views
  - Analyst can discover trends and patterns otherwise not visible

Data Used for the Examples

- With a Google query, and rapid filter for duplicates and obvious polemic, found 250 web documents on Tariq Aziz
- Extracted entities automatically of five types: person names, facility names, organization names, times (dates and hours), and place names - with their text off-sets
- Storage and analysis of extracted data in MS Access
- For graphics used MS Excel
- All examples mimic our own operational experience
- Warning: I used uncorrected extractions for these examples
  - No manual review of documents
  - Used proximity as a way to mimic relationships
  - Shows what uncorrected data looks like

Example 1: Reviewing Hypotheses

- Description: Extraction completed, the analyst reviews extracted material
  - To determine interest
  - To determine accuracy
  - To augment or correct
- Application must be set up to make this review easy
- Analyst is presented with a "hypothesis" from the system - that this person, organization or relationship is valid and is of interest
- Analyst review task is far less stressful and quicker than the unassisted task of locating and mentally testing for possible relevance every entity in a pile of documents

Value	ShortSource
Lester Holt	DocOil
Madeleine Albright	Washpost
Margaret Warner	NewsHour Aziz97
Massoud Barzani	Turkish Daily News2
Massoud Rajavi	DocIran95
Mbeki	DocConsider
MEETING	France and Iraq
Mehmet Ali Sahin	Zaman Daily
Menachem Gantz	Ottawa Sun
Michael Vincent	DocInspectors02
Michael Voss	BBC US deal
Michael Yuhanna	Aziz role
Michel Collon	Aziz Analysis

Countdown: Iraq's Lester Holt interviewed Peter Arnett last Wednesday for more on his conversation with Aziz. MSNBC News, Dec 5, 2002

Tariq Aziz gave Ankara an important message along the lines of, "We can accept Massoud Barzani controlling northern Iraq with our approval until we return to the region. Do not react against that."
15 November, 1996, Copyright © Turkish Daily News

Reuters, Baghdad, May 30 - Iraqi Deputy Prime Minister Tariq Aziz met Massoud Rajavi, leader of the exiled Iranian opposition Mujahideen Khalq, a Mujahideen statement said on Tuesday. (1995)

Istanbul, TURKEY, February 17, 2003 - At the Ataturk International Airport, AK Party Leader Recep Tayyip Erdogan and State Minister and Vice Prime Minister Mehmet Ali Sahin met with Iraqi Vice Prime Minister Tariq Aziz regarding the last point reached in the Iraqi Issue. Zaman Daily Newspaper

as Michael Vincent reports, former weapons inspectors are regarding Iraq's change of heart with some scepticism.Tuesday, September 17, 2002 18:05, Australian Broadcasting Corporation

On the left Michel Collon, who took the initiative for the mission.
A reference to a photo showing Aziz and Collon from an International Peace Mission which met with Aziz (website unidentified and undated)

People who might have met Aziz

List of Names and Short Source names with quotations from text from which person names extracted, as analyst might view material for verification.

Results

Productivity gains (from verification task) -

Metric	Analyst 1 (used IE)	Analyst 2 (without IE)
Hours spent	48	200
Entities verified	350	200
Pages used	2601	Unrecorded estimate - 2600

Estimated recall on Aziz data, based on manual review of 25 documents (10% of total) -

Possible meetings w/Aziz	Manual count	Automated count	Percent
Document based – in 25 docs	19	13	68%
Person pair based – in whole set	19	16	84%

Ex. 2: Trends in large data sets

- Description: Extraction completed and loaded in a database
  - Extracted data may be corrected for errors which can be automatically corrected
  - Remaining errors 'accepted'
  - Analyst 'slices and dices' data for different trends
- Task impossible without Information Extraction
  - Requires entities and relationships to be in a database
  - Manual extraction at the rate of 24 entities per hour would have required 1,000 hours of extraction just for these minimal examples used in this paper

Places associated closely with Aziz **- by year, from 1998 through 2003 – one instance of each place name and spurious locations removed**

98	99	00	01	02	03
Afghanistan	Amman	Amman	Afghanistan	Ankara	Afghanistan
Amman	Beirut	Assisi	Britain	Assisi	Ankara
Assisi	China	Beirut	China	Bahrain	Assisi
Bahrain	Egypt	China	Israel	Britain	Bahrain
Britain	Iran	Damascus	Kuwait	China	Britain
China	Italy	Israel	Moscow	Damascus	Cairo
France	Jordan	Jordan	New York	Egypt	China
Geneva	Lebanon	Kuwait	Pakistan	France	Damascus
Iran	South Africa	Lebanon	Russia	Germany	Egypt
Israel	Syria	Moscow	South Africa	Hollywood	France
Italy	U.S.	Prague	Syria	Iran	Germany
Johannesburg	U.A.E.	Rome	U.S.	Israel	Hollywood
Jordan	Washington	Russia		Italy	Iran
Kuwait		Syria		Johannesburg	Israel
Lebanon		U.S.		Jordan	Italy
Moscow		Washington		Kuwait	Johannesburg
New York				Lebanon	Jordan
Pakistan				Marrakesh	Korea
Paris				Morocco	Kuwait
Russia				Moscow	Lebanon
Saudi Arabia				New York	Marrakesh
Switzerland				Northampton	Milan
Syria				Paris	Morocco
Turkey				Rome	Moscow
U.S.				Russia	New York
Vatican				Saudi Arabia	Northampton
Vienna				South Africa	Ottawa
Washington				Syria	Pakistan
				Turkey	Paris
				U.S.	Rome
				U.A.E.	Russia
				Vatican	Saudi Arabia
				Vienna	South Africa
				W.Va.	Syria
				Washington	Turkey
					U.S.
					Vatican
					Washington

Places closely associated with Aziz:

- **By year**
- **Normalized to country name & aligned by country**
- **Organized by region**

98	99	00	01	02	03
Africa					
			Morocco	Morocco	
South Africa	South Africa		South Africa	South Africa	South Africa
Americas					
					Canada
U.S.	U.S.	U.S.	U.S.	U.S.	U.S.
Central and South Asia					
Afghanistan			Afghanistan		Afghanistan
Iran	Iran			Iran	
Pakistan			Pakistan		Pakistan
East Asia					
China	China	China	China	China	China
				Korea	
Europe					
Austria				Austria	
Britain			Britain	Britain	Britain
		Czechoslavakia			
France			France	France	France
				Germany	Germany
Italy	Italy	Italy	Italy	Italy	Italy
Russia		Russia	Russia	Russia	Russia
Switzerland					
Vatican				Vatican	Vatican
Middle East					
Bahrain				Bahrain	Bahrain
	Egypt			Egypt	Egypt
Israel		Israel	Israel	Israel	
Jordan	Jordan	Jordan	Jordan	Jordan	
Kuwait		Kuwait	Kuwait	Kuwait	
Lebanon	Lebanon	Lebanon	Lebanon	Lebanon	
Saudi Arabia				Saudi Arabia	Saudi Arabia
Syria	Syria	Syria	Syria	Syria	Syria
Turkey				Turkey	Turkey
	U.A.E.			U.A.E.	

Countries closely associated with Aziz:

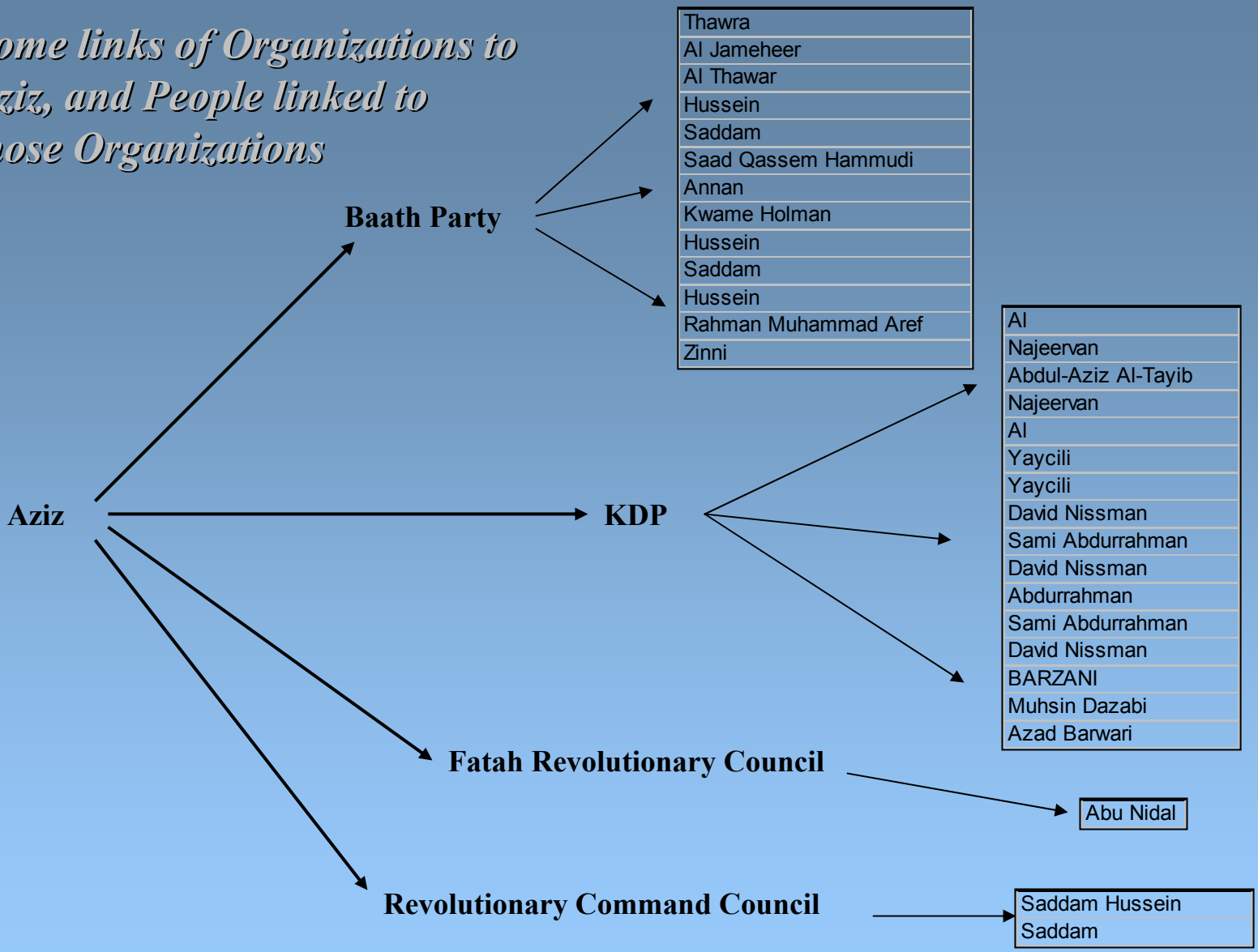
- **By year**
- **Categorized by Muslim association**
- **Ordered by number of years of close association**

98	99	00	01	02	03
Predominately non-Muslim					
China	China	China	China	China	China
Italy	Italy	Italy	Italy	Italy	Italy
U.S.	U.S.	U.S.	U.S.	U.S.	U.S.
Russia		Russia	Russia	Russia	Russia
South Africa	South Africa		South Africa	South Africa	South Africa
Britain			Britain	Britain	Britain
France			France	France	France
Israel		Israel	Israel	Israel	
Vatican				Vatican	Vatican
Austria				Austria	
				Germany	Germany
					Canada
		Czechoslavakia			
				Korea	
Switzerland					
Predominately Muslim					
Syria	Syria	Syria	Syria	Syria	Syria
Jordan	Jordan	Jordan	Jordan	Jordan	
Lebanon	Lebanon	Lebanon	Lebanon	Lebanon	
Kuwait		Kuwait	Kuwait	Kuwait	
Afghanistan			Afghanistan		Afghanistan
Bahrain				Bahrain	Bahrain
	Egypt			Egypt	Egypt
Iran	Iran			Iran	
Pakistan			Pakistan		Pakistan
Saudi Arabia				Saudi Arabia	Saudi Arabia
Turkey				Turkey	Turkey
			Morocco	Morocco	
	U.A.E.			U.A.E.	

Ex. 3: Finding Links

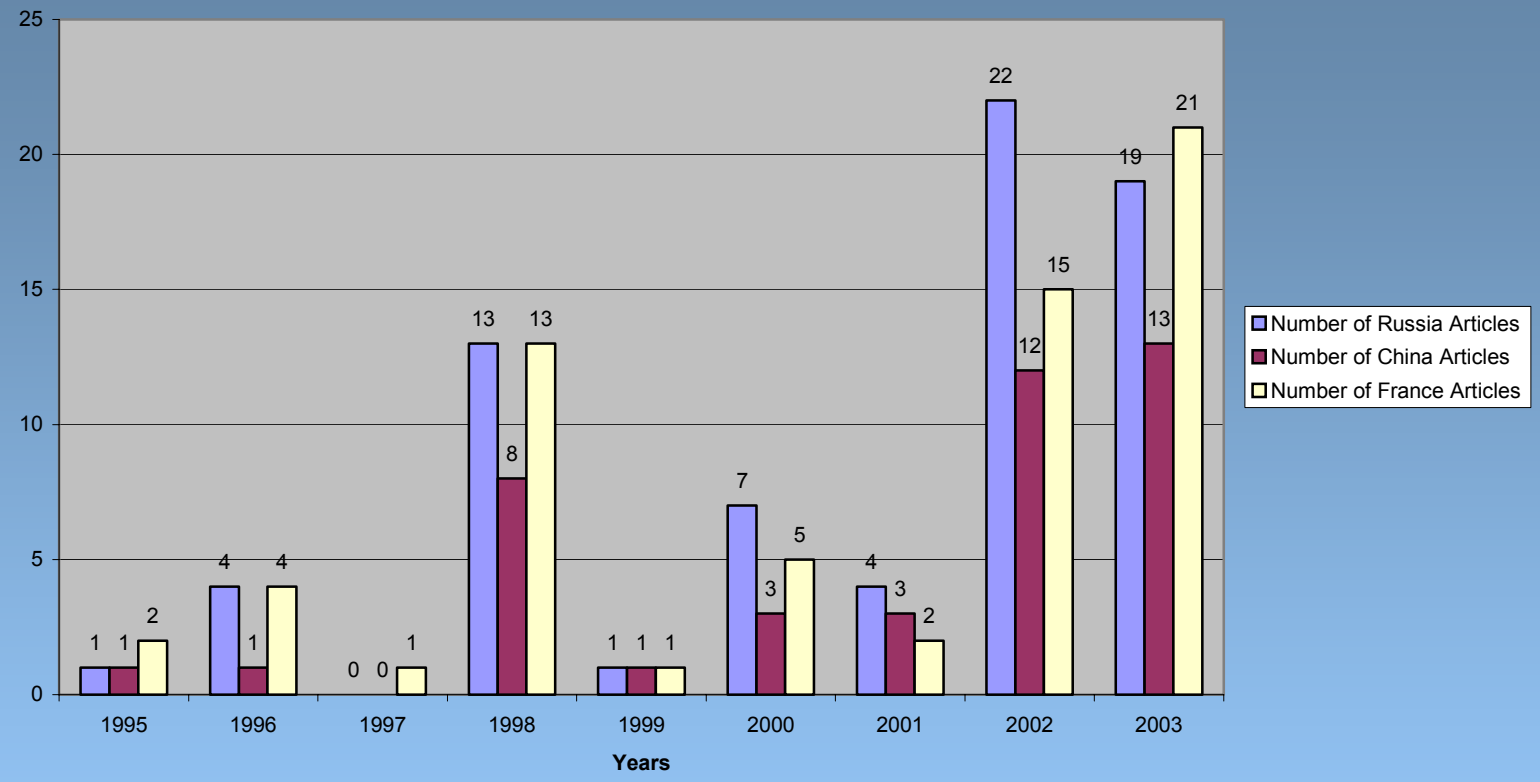
- Description: Extraction completed, entities and relationships are loaded into a database -
  - Automated cleanup can occur
  - Remaining errors accepted “for now”
  - Mapping to link analysis tool occurs
  - Analyst “slices and dices” data using link analysis tool to view the database
  - Analyst checks the validity of interesting links
- Extraction vastly increases the volume of data which can be checked for interesting material
- Only those links of interest are validated – saving resources

Some links of Organizations to Aziz, and People linked to those Organizations

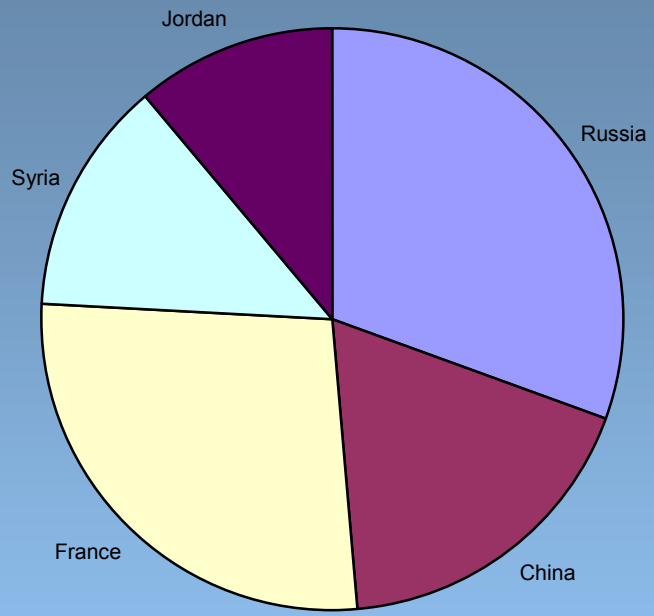


Ex. 4: Finding the Unexpected

- Description: Extraction completed, material is loaded into a database – ideally connected to several different visualization and analysis capabilities
  - Analyst in open-ended quest, "she'll know it when she sees it"
  - Analyst queries and views data in different ways, often with only marginal differences
  - Something may or may not appear…..
  - Many results can be validated against sources; some quantitative results can only be taken as indicative
- Requires extraction, one or more visualization capabilities tied to the same database, a flexible querying mechanism, extracted data tied back to its source, with automated data clean-up capabilities helpful
- Simulated this environment, using the Aziz data and MS Excel for multiple visualizations

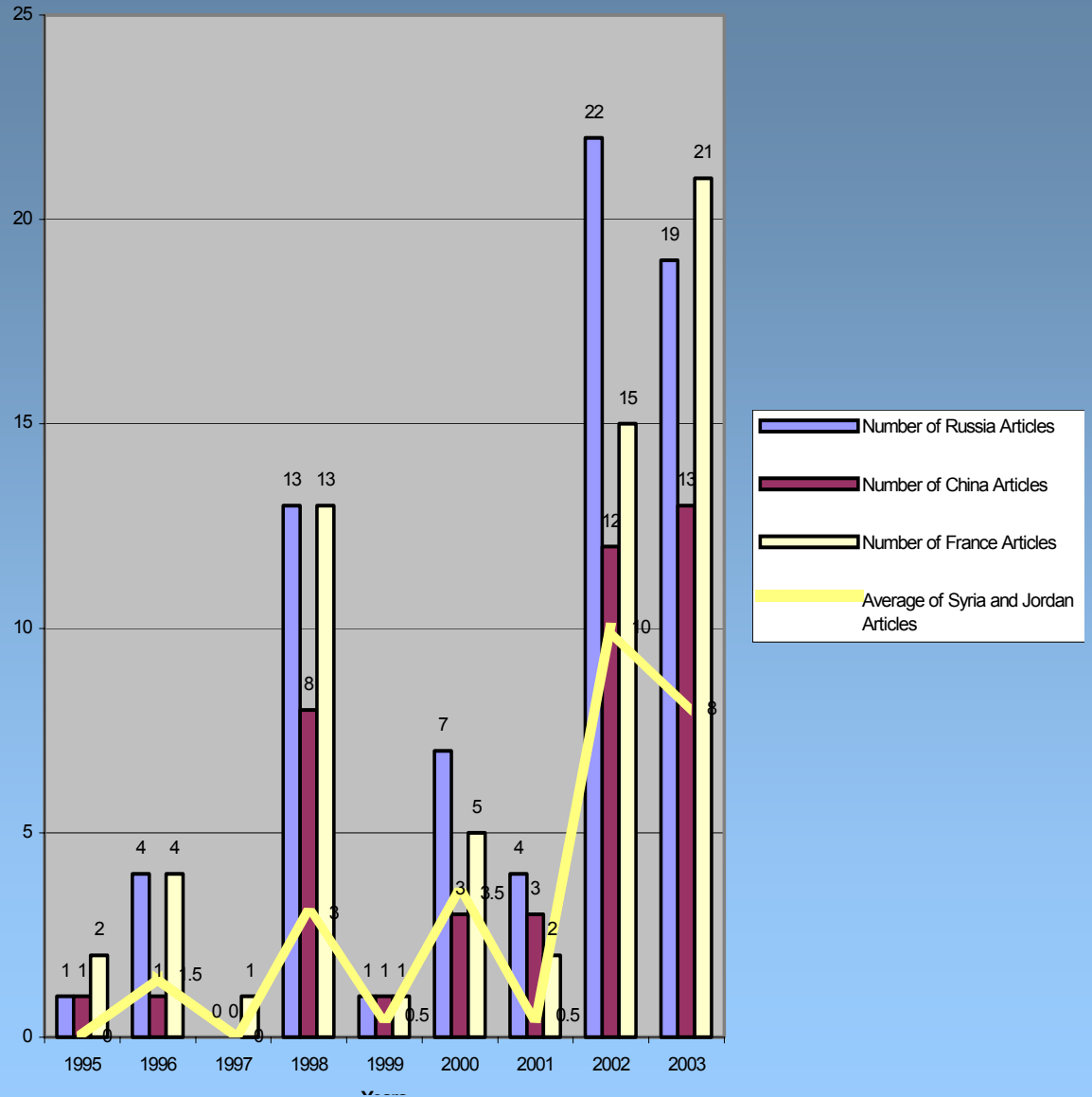


Numbers of articles, per year **- where Aziz associated with China, France, or Russia, from 1995 to 2003**



Share of articles, from 1995-2003 **- closely associating Aziz with China, France, Russia, comparing with Jordan and Syria for the same period**

Number of articles, per year - **associating Aziz with China, France, and Russia, compared to a baseline of the average number of articles per year associating Aziz with Jordan and Syria**



Conclusion

- Information Extraction enables:
  - Use of multiple tools and views of the same data
  - Analysts to use a vastly increased volume of data
- In our operational applications of Information Extraction, using our own tool, AeroText, we have observed:
  - Increased analyst productivity
  - Improved depth and breadth of analytic product