Alberta Parks’ use of Text Mining

Alberta Tourism, Parks and Recreation,

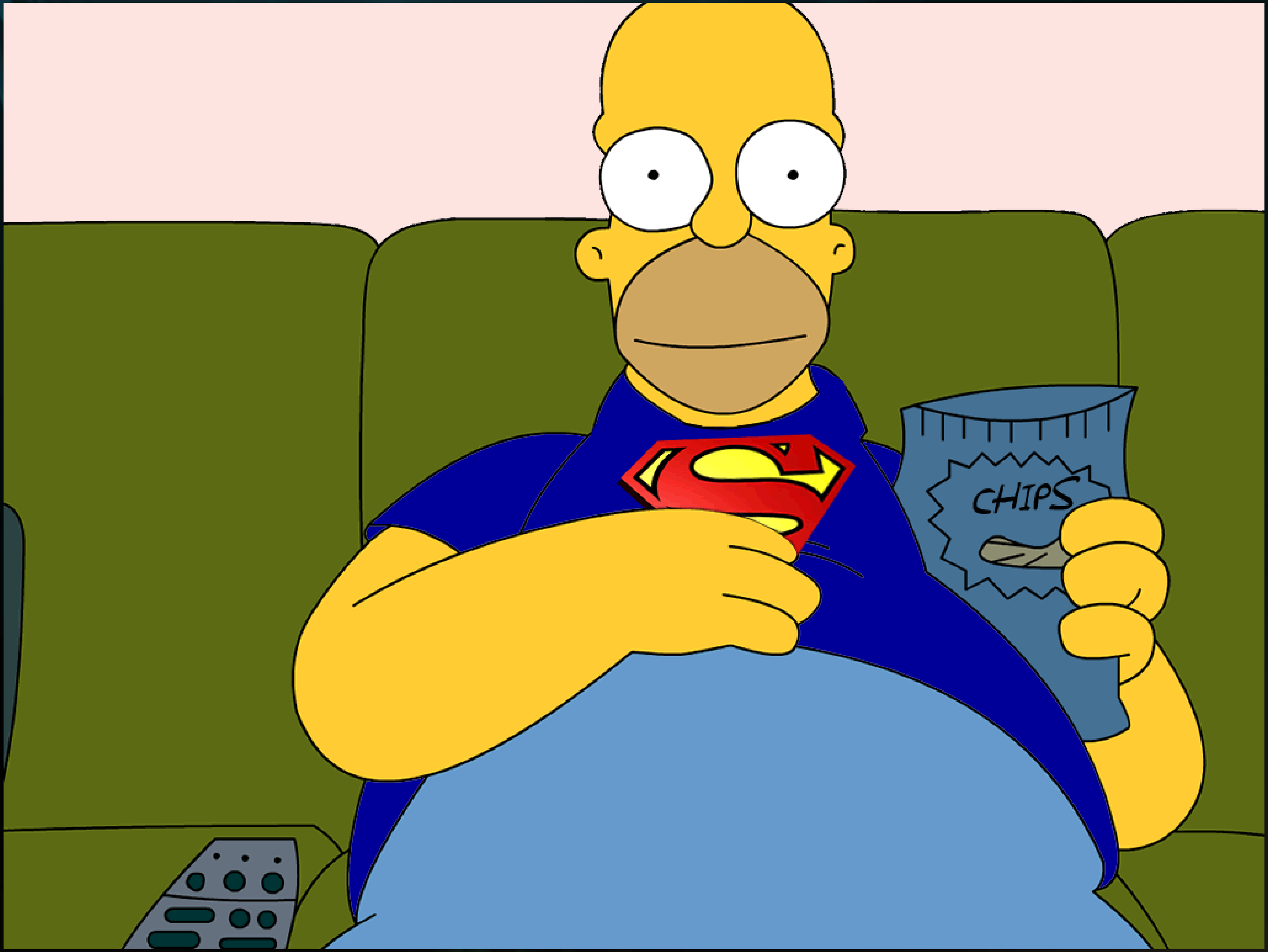
Parks Division,

Business Integration and Analysis

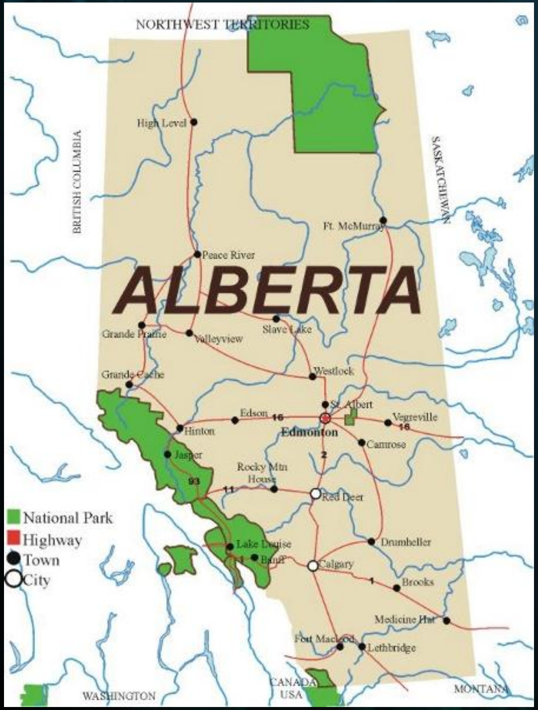
Alberta Parks

Statistics Man!

realistically...



Alberta Parks



- 209 Provincial Recreation Areas
- 75 Provincial Parks
- And more…

Total 478

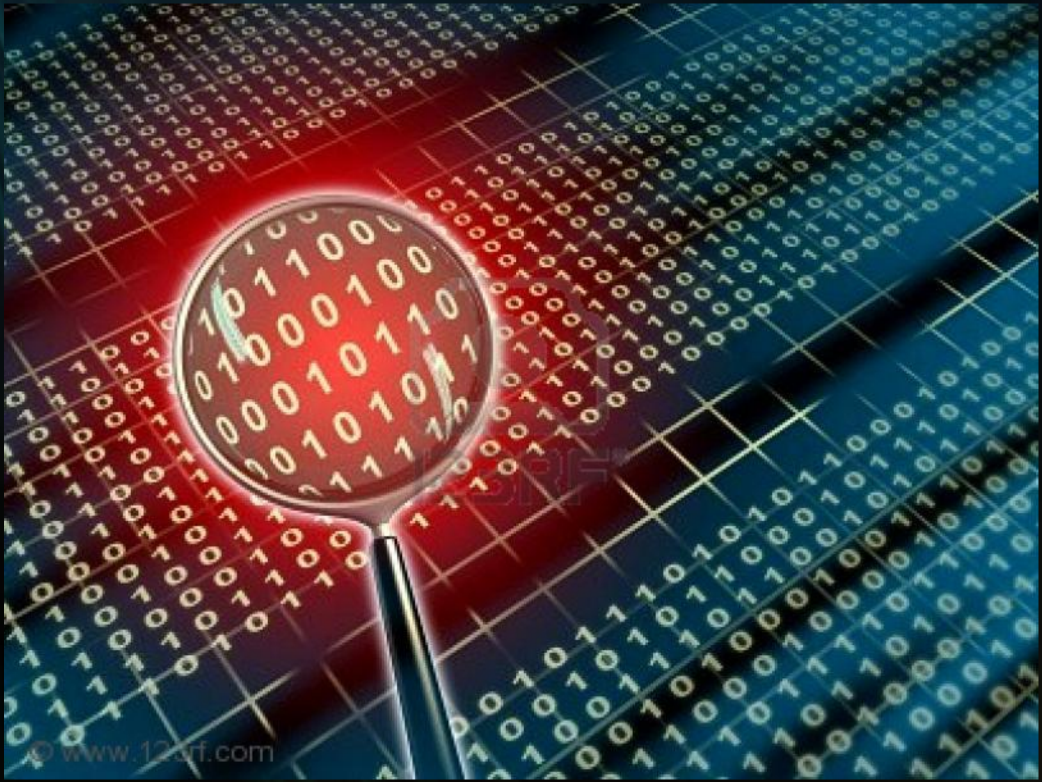
More info: http://www.albertaparks.ca/albertaparksca/library/land-reference-manual.aspx

2

Alberta Park history: http://www.tpr.alberta.ca/parks/managing/history.asp

Transforming Data

Data is shapeless, providing limited insight without the proper tools to drive fact-based analysis and decision making.



© www.123rf.com

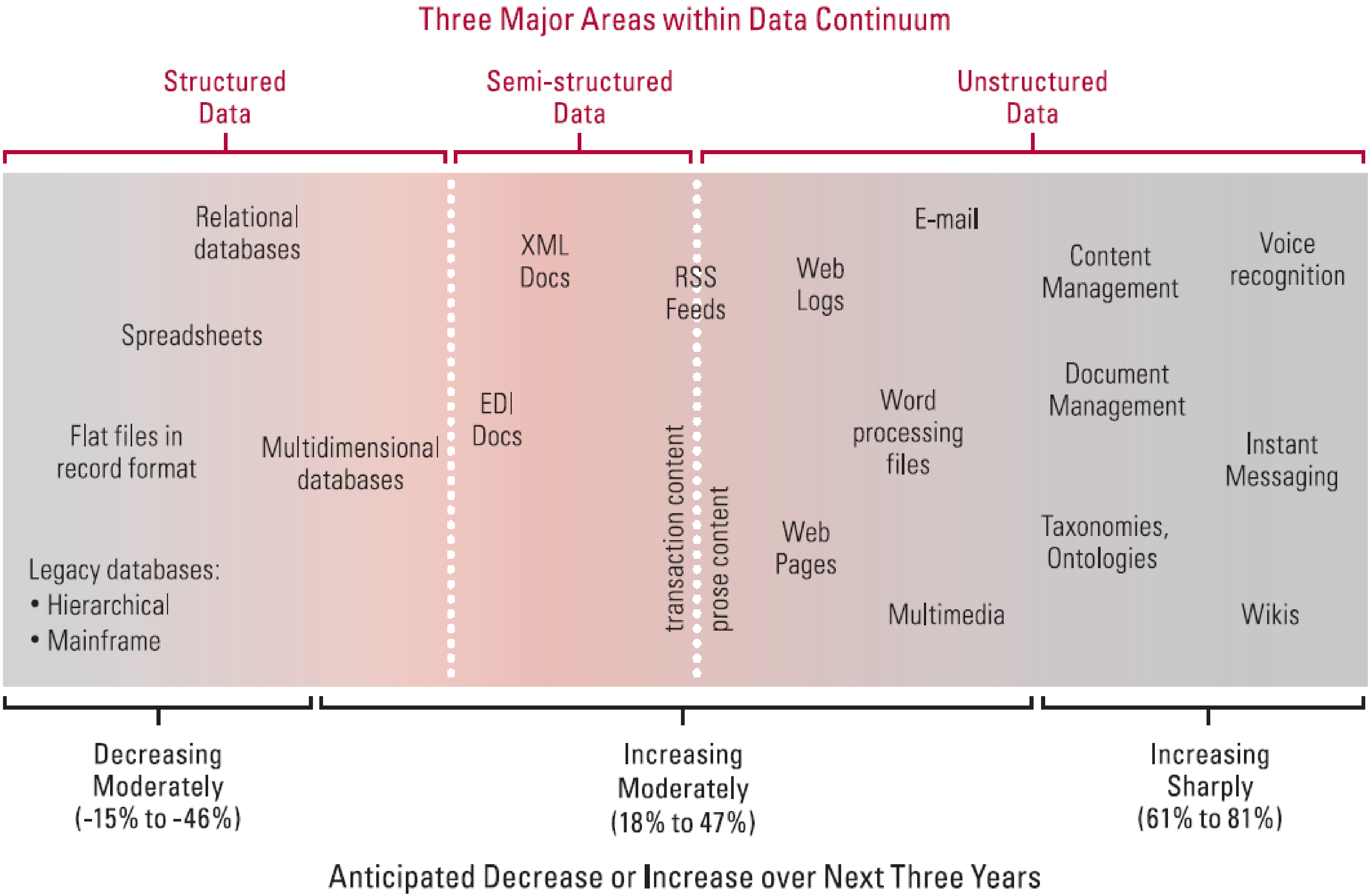
Analytics Magic?

“Most organizations are drowning in data but starving for information.”

Anonymous

Data: Structured v.s. Unstructured

Data and source types plotted on the data continuum



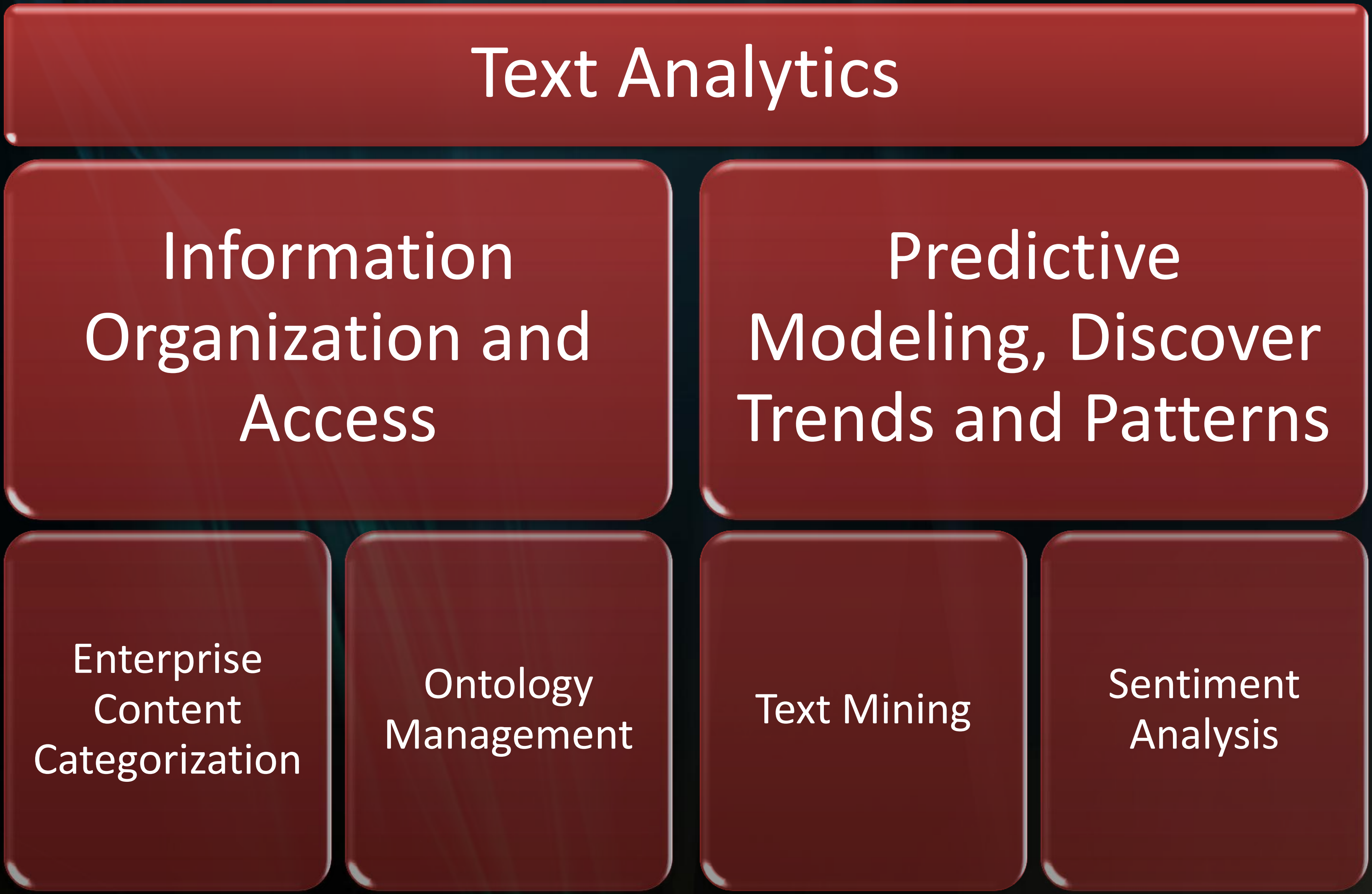
Unstructured Data Sources

- Web page
- Email
- Content management system records
- Word Document, PDF
- Telephone call
- Instant message
- SMS (text message)
- Letters from the public
- Tweet from Twitter
- Blog post
- etc…

Text Analytics

Using statistical methods to analyze and interpret the meaning of **textual data** (unstructured data).

- Visionary paper written by Hans Peter Luhn titled *"The Automatic Creation of Literature Abstracts"* for the **1958** IBM Journal marks birth of computational text analytics
- Automated **solutions** go mainstream in **early 2000's** by visionary companies such as SAS and Teragram
- **Web 2.0** has kicked off an arms race to **capture** the broad and vast **content** now being **exposed** by the web (I.e. Online social networks)



Text Analytics

Information Organization and Access

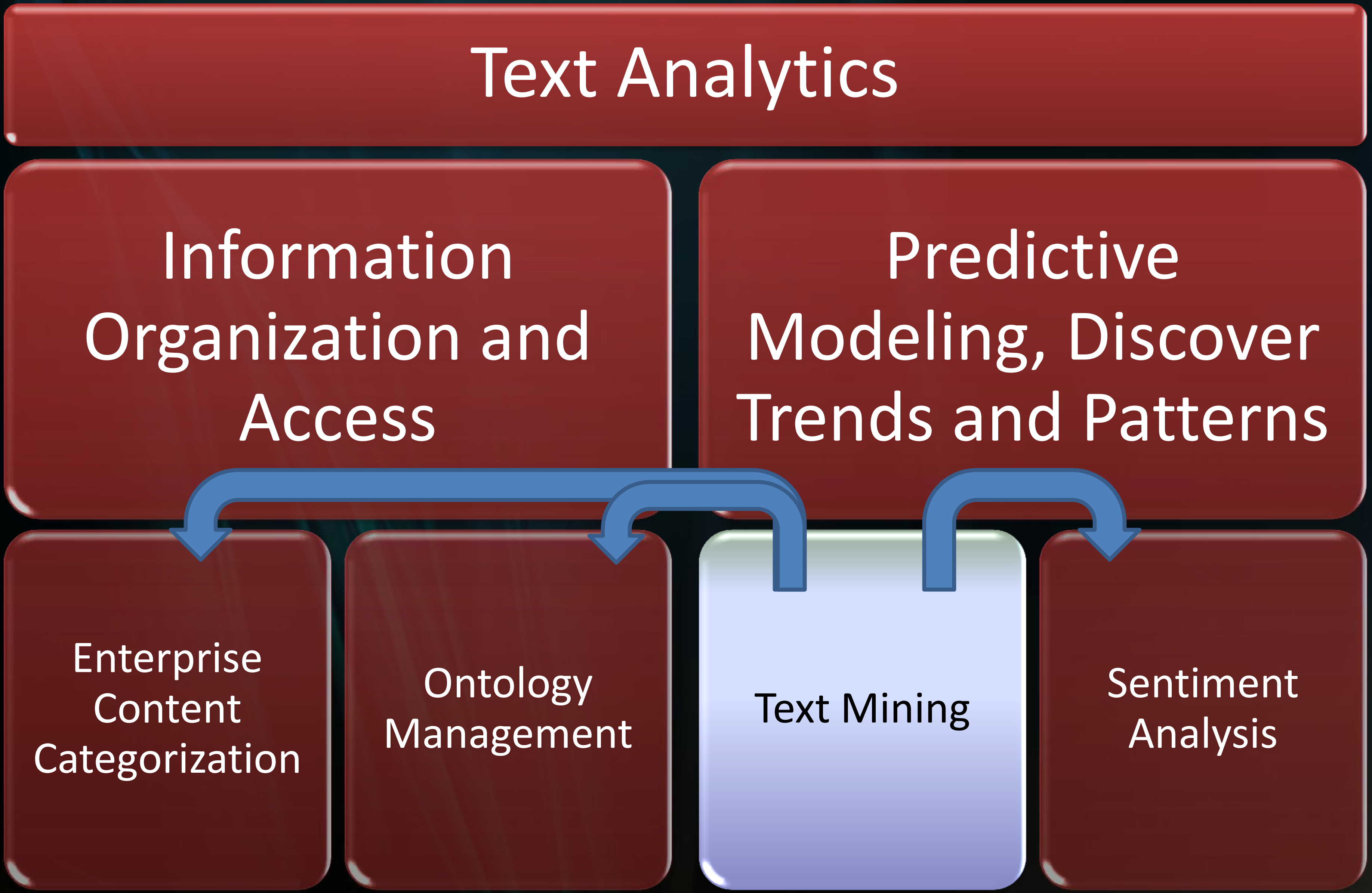
Predictive Modeling, Discover Trends and Patterns

Enterprise Content Categorization

Ontology Management

Text Mining

Sentiment Analysis



Text Analytics

Information Organization and Access

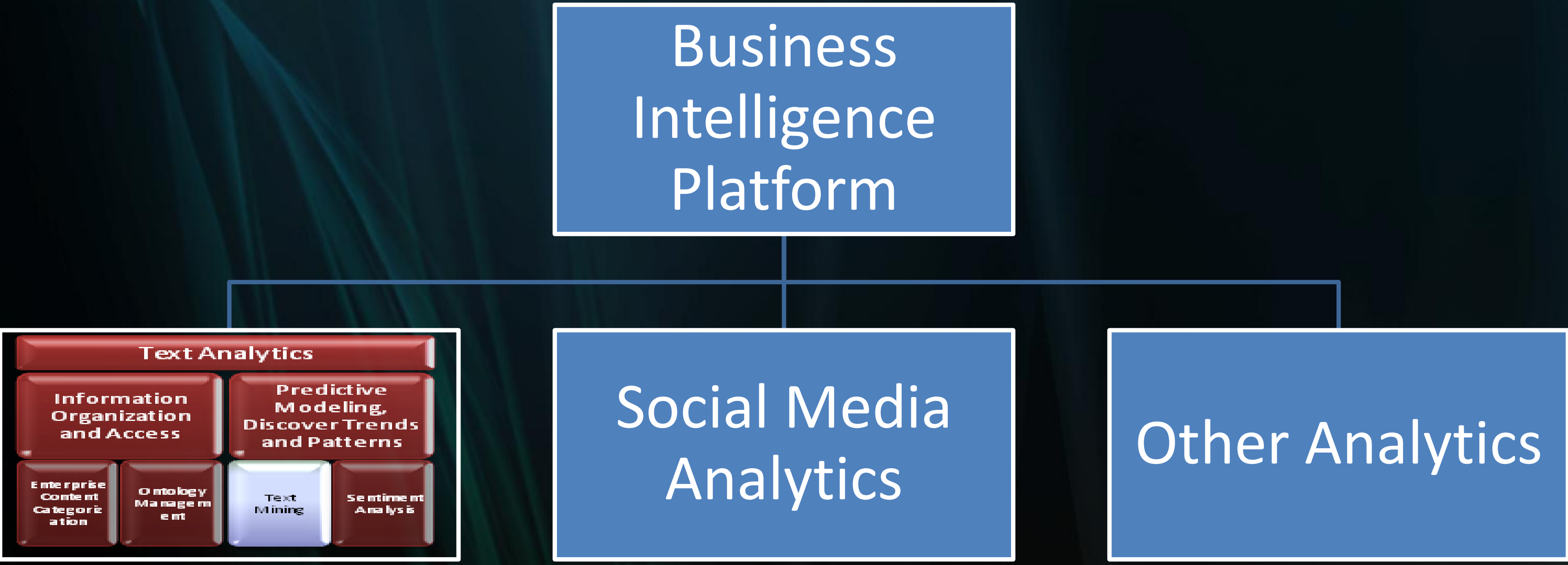
Predictive Modeling, Discover Trends and Patterns

Enterprise Content Categorization

Ontology Management

Text Mining

Sentiment Analysis



Business Intelligence Platform

Text Analytics

Information Organization and Access

Predictive Modeling, Discover Trends and Patterns

Enterprise Content Categorization

Ontology Management

Text Mining

Sentiment Analysis

Social Media Analytics

Other Analytics

Text Mining

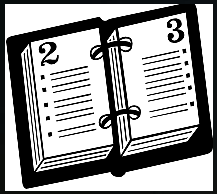
Natural Language Processing

- Stem…Stems…Stemming (park, parks, parking)
- Parts of Speech (verb, noun, adjective…)
- Dictionaries
- Entity Extraction

person

place

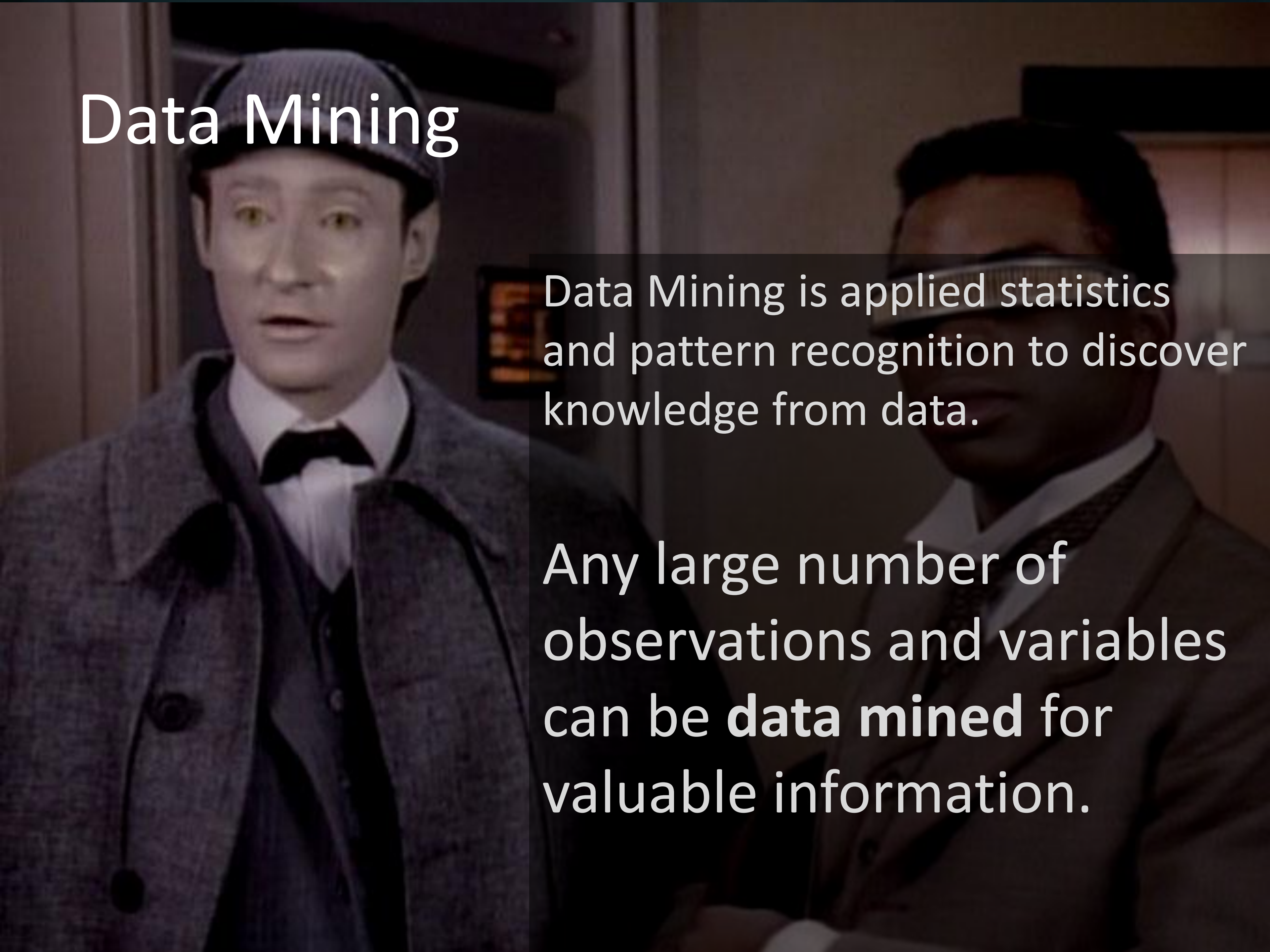
dates



Natural Language Processing

- UPPERCASE
- Miss-spelings
- A.C.R.O.N.Y.M.S
- Shrt-hnd
- Pr☺f@nity
- *Punctuation*

Data Mining

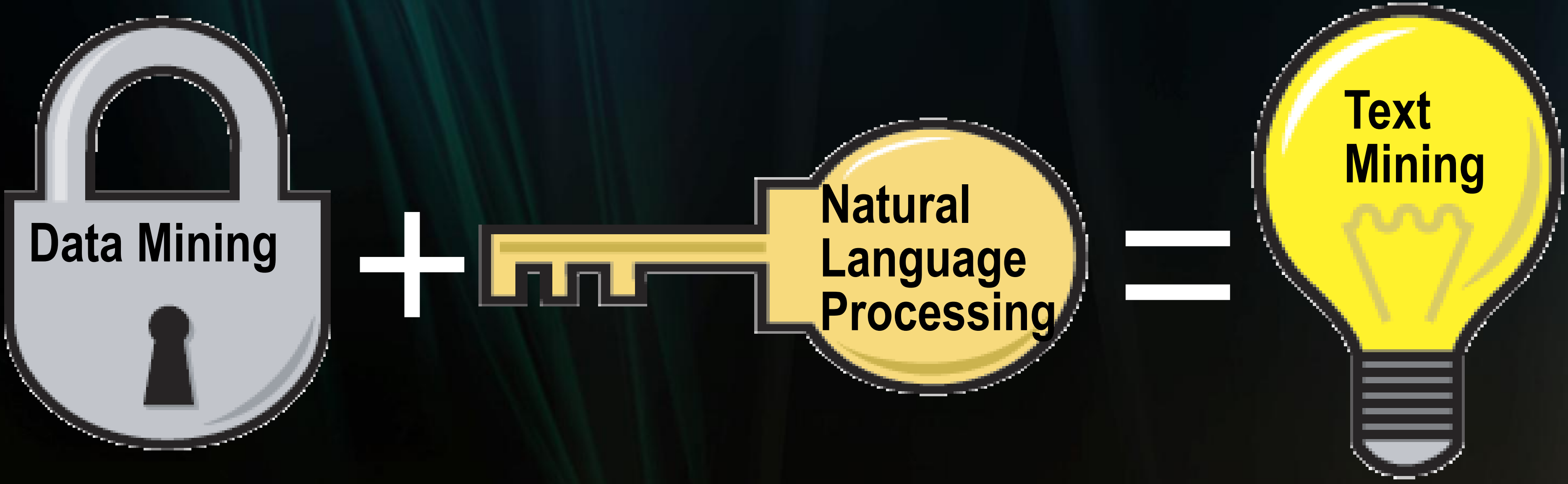


Data Mining is applied statistics and pattern recognition to discover knowledge from data.

Any large number of observations and variables can be **data mined** for valuable information.

Text Mining

“The process of discovering and extracting meaningful patterns and relationships from text collections.”



TM = Discovery ≠ Search

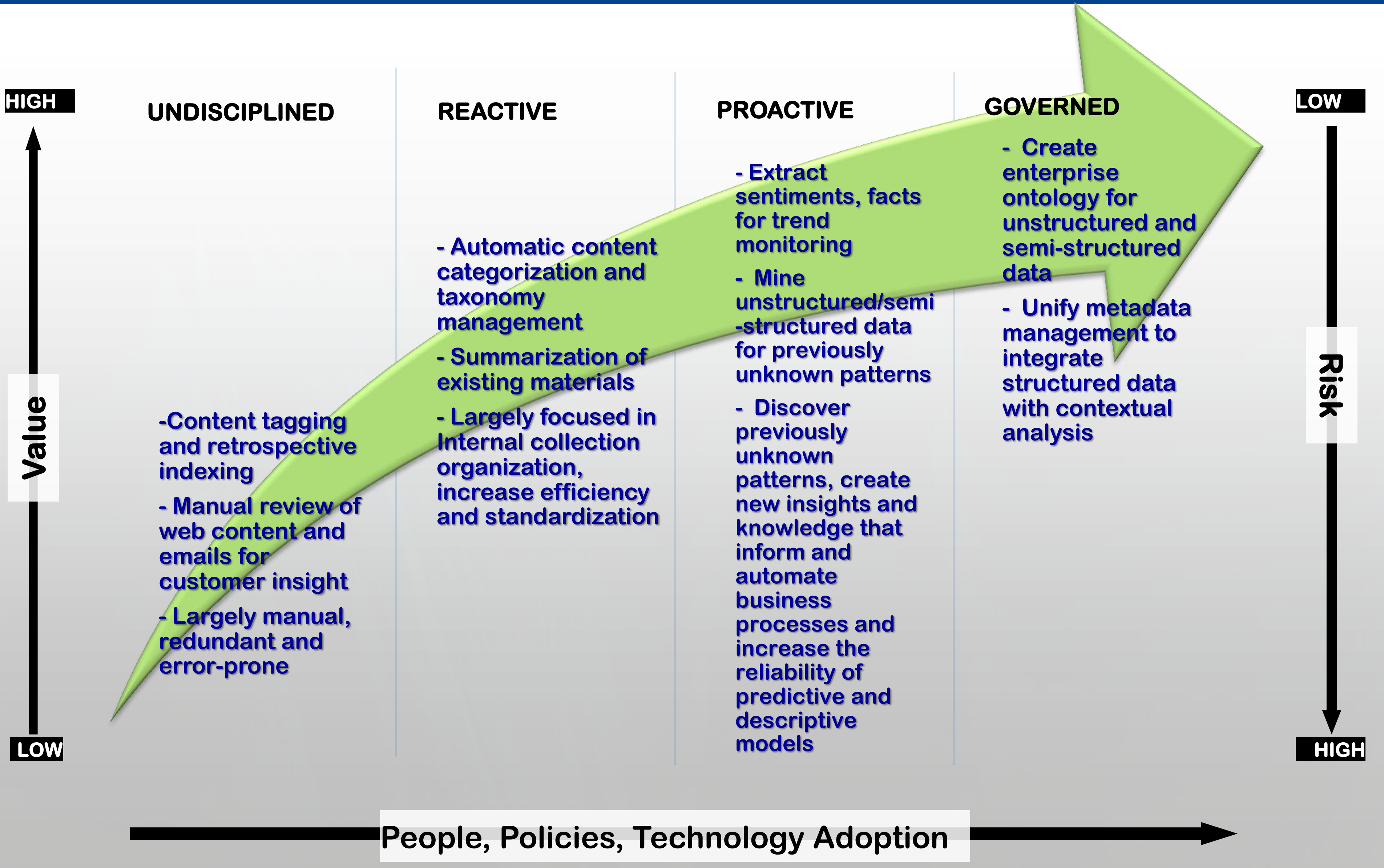
- Text Mining is more than frequency counts. Frequency excludes context and relations.
- A Microsoft Word ‘word count’ or a word cloud does not capture meaning and could even be misleading.
- Text Mining helps discover key concepts, term associations and relationships.

Search v.s. Discovery

You can search the island, but you might still be lost

Is text really unstructured data?

Text Analytics Adoption Curve



HIGH

Value

LOW

LOW

Risk

HIGH

UNDISCIPLINED

-Content tagging and retrospective indexing

- Manual review of web content and emails for customer insight

- Largely manual, redundant and error-prone

REACTIVE

- Automatic content categorization and taxonomy management

- Summarization of existing materials

- Largely focused in Internal collection organization, increase efficiency and standardization

PROACTIVE

- Extract sentiments, facts for trend monitoring

- Mine unstructured/semi-structured data for previously unknown patterns

- Discover previously unknown patterns, create new insights and knowledge that inform and automate business processes and increase the reliability of predictive and descriptive models

GOVERNED

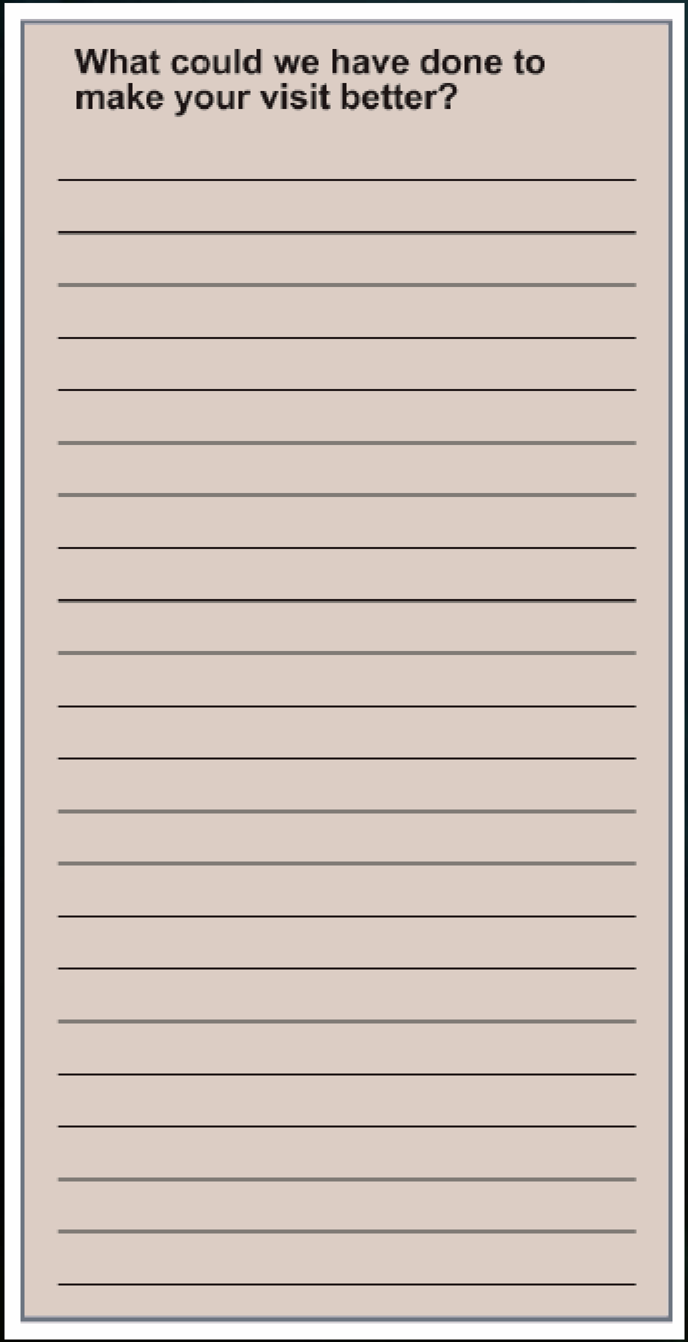
- Create enterprise ontology for unstructured and semi-structured data

- Unify metadata management to integrate structured data with contextual analysis

People, Policies, Technology Adoption

Some things in parks can’t be analysed…

Government of Alberta
How Are We Doing?
Dear Visitor,



What could we have done to make your visit better?

“What could we have done to make your visit better?”

- Annual camper satisfaction survey

“The nice looking lady that woke me up was a very good start”

“Great skinny dipping lake, nothing wrong there.”

“A great big Budweiser motorhome pulls up with 12 girls that want to party!”

“Fireworks at 4:00 in the morning”

“Daughter says more hot boys are needed”

“Park Rangers were informative – maybe it was just my pretty girlfriend”

“Conservation officer was a hottie. He should have visited more often”

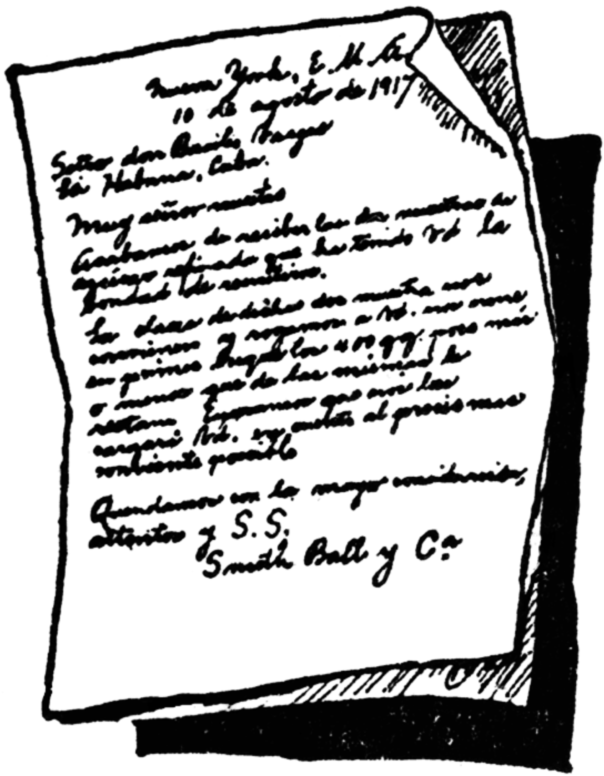
"We are two spry young gents and enjoy camping here."

“One ply toilet paper is insufficient for the task”

“Rangers need skimpier uniforms”

“Well, you could have stopped by for a little cuddle time… ”

- Hand written responses are not ideal. It comes with the nature (no pun intended) of our business.



- Comments are transferred to electronic format through typing or Speech to Text (dictation) software.

Old method – Assigning codes

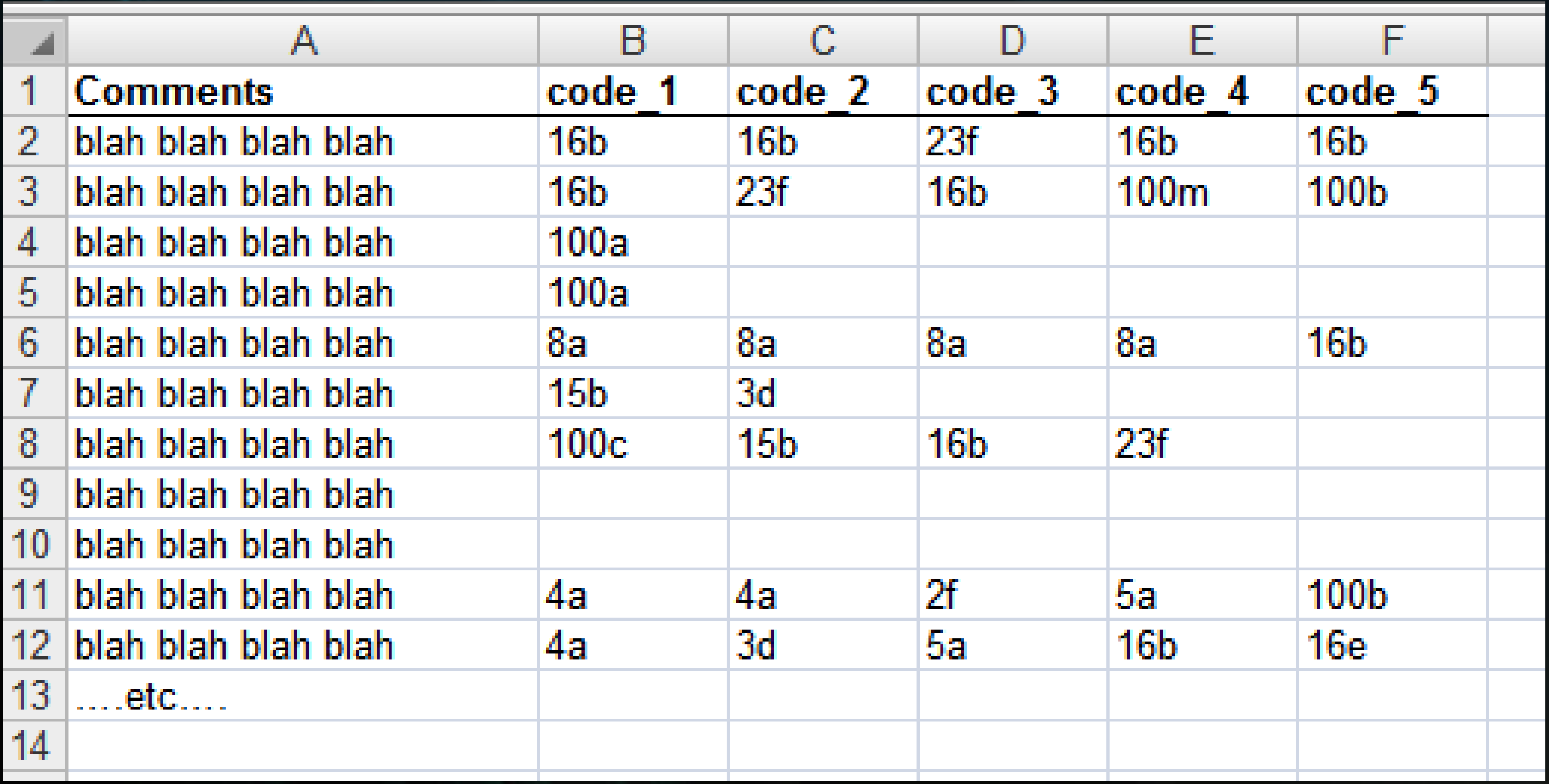
187 sub-categories across 28 General Categories.
Examples:

Washrooms
Firewood
Roads
Showers
Security
Grounds Maintenance
Operations

Information Services
Pest Control
Playgrounds
Reservation System
Fishing
Facilities
Beach / Lake

Policy
Trails
Camping Preferences
Value
Noise

Once comments are assigned codes, simple frequency counts show magnitudes of customer feedback…



	A	B	C	D	E	F
1	**Comments**	**code_1**	**code_2**	**code_3**	**code_4**	**code_5**
2	blah blah blah blah	16b	16b	23f	16b	16b
3	blah blah blah blah	16b	23f	16b	100m	100b
4	blah blah blah blah	100a				
5	blah blah blah blah	100a				
6	blah blah blah blah	8a	8a	8a	8a	16b
7	blah blah blah blah	15b	3d			
8	blah blah blah blah	100c	15b	16b	23f	
9	blah blah blah blah					
10	blah blah blah blah					
11	blah blah blah blah	4a	4a	2f	5a	100b
12	blah blah blah blah	4a	3d	5a	16b	16e
13	….etc….					
14						

…see example of frequency counts on the next slide…

2008 Camper Satisfaction Survey
General and Sub-Category Comments - Provincial Negative Comments
(Total Surveys Represented – 1,118)

General Category	Sub-Category	# of Comments	% of Category	% of All Comments	% of ALL Surveys Represented
Firewood	Too expensive	96	30.1	3.6	8.6
	Firewood Quantity (not enough/no wood)	61	19.1	2.3	5.5
	Poor Quality (too long, wet)	48	15.0	1.8	4.3
	Poor Access (location, timing)	47	14.7	1.8	4.2
	Should be free	40	12.5	1.5	3.6
	Firewood Delivery Needed and other	13	4.1	0.5	1.2
	Firewood Should be Included in Fees	12	3.8	0.4	1.1
	Firewood Shelter Needed/Upgraded	2	0.6	0.1	0.2
	Subtotal	319	100.0	11.9	28.5
Hook-ups/Dump stations/Water	Additional power campsites	86	34.8	3.2	7.7
	Full Power-Water-Sewer Hook-ups Needed	31	12.6	1.2	2.8
	Other (specific amperage, water filling station needed)	26	10.5	1.0	2.3
	More Taps / Water Locations	24	9.7	0.9	2.1
	Poor Drinking Water Quality / Need Potable Water	21	8.5	0.8	1.9
	Install power campsites	20	8.1	0.7	1.8
	Sewage Dump-stations Needed / Dirty / Full	18	7.3	0.7	1.6
	Water Hook-ups Needed	11	4.5	0.4	1.0
	Running Water Needed (not washroom related)	10	4.0	0.4	0.9
	Subtotal	247	100.0	9.2	22.1

Analysing Text

The Old Way

1. Typing comments (~3 weeks/year)
2. Every comment manually read and manually assigned special codes (~ 3 weeks/year)

The New Way

1. Dictation software types comments (~1 week/year)
2. SAS Text Miner analyses data (~ 1 minute/year)*

*First year requires a few days to create the 'black box' but becomes a production run thereafter.

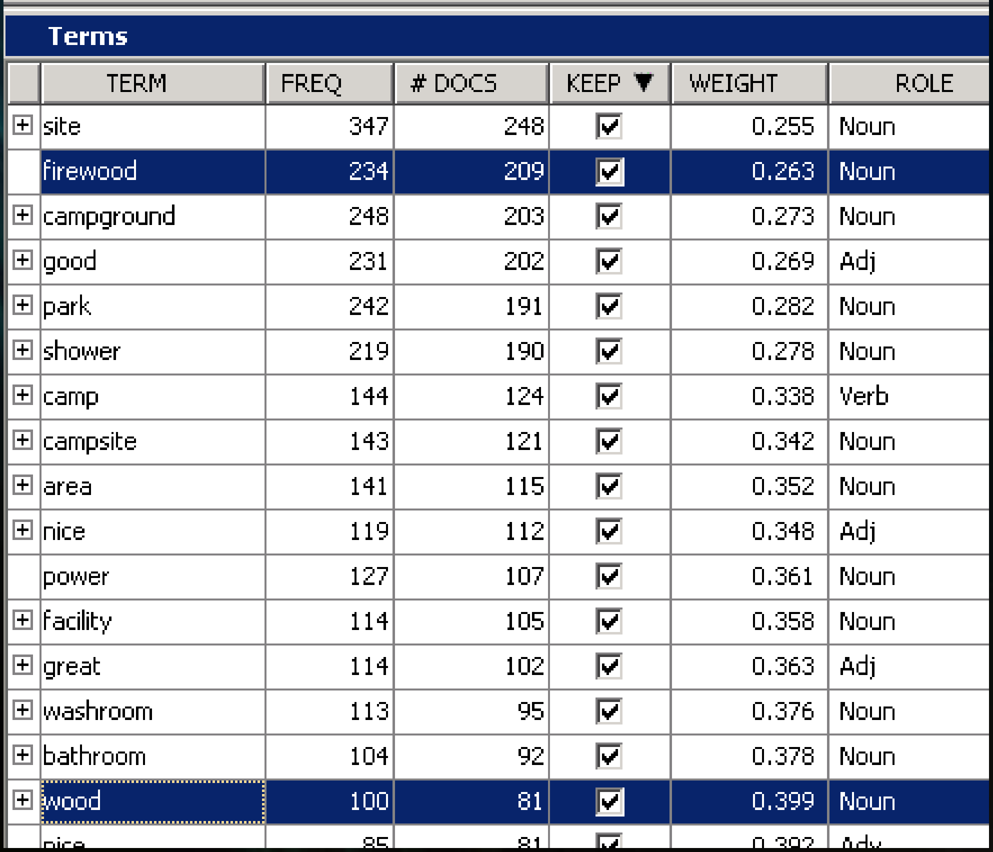
2008 Camper Satisfaction Survey
General and Sub-Category Comments - Provincial Negative Comments
(Total Surveys Represented – 1,118)

General Category	Sub-Category	# of Comments	% of Category	% of All Comments	% of ALL Surveys Represented
Firewood	Too expensive	96	30.1	3.6	8.6
	Firewood Quantity (not enough/no wood)	61	19.1	2.3	5.5
	Poor Quality (too long, wet)	48	15.0	1.8	4.3
	Poor Access (location, timing)	47	14.7	1.8	4.2
	Should be free	40	12.5	1.5	3.6
	Firewood Delivery Needed and other	13	4.1	0.5	1.2
	Firewood Should be Included in Fees	12	3.8	0.4	1.1
	Firewood Shelter Needed/Upgraded	2	0.6	0.1	0.2
	Subtotal	319	100.0	11.9	28.5
Hook-ups/Dump stations/Water	Additional power campsites	86	34.8	3.2	7.7
	Full Power-Water-Sewer Hook-ups Needed	31	12.6	1.2	2.8
	Other (specific amperage, water filling station needed)	26	10.5	1.0	2.3
	More Taps / Water Locations	24	9.7	0.9	2.1
	Poor Drinking Water Quality / Need Potable Water	21	8.5	0.8	1.9
	Install power campsites	20	8.1	0.7	1.8
	Sewage Dump-stations Needed / Dirty / Full	18	7.3	0.7	1.6
	Water Hook-ups Needed	11	4.5	0.4	1.0
	Running Water Needed (not washroom related)	10	4.0	0.4	0.9
	Subtotal	247	100.0	9.2	22.1

Using SAS Text Miner…

firewood (234) + wood (100) = 334 (v.s. 319)

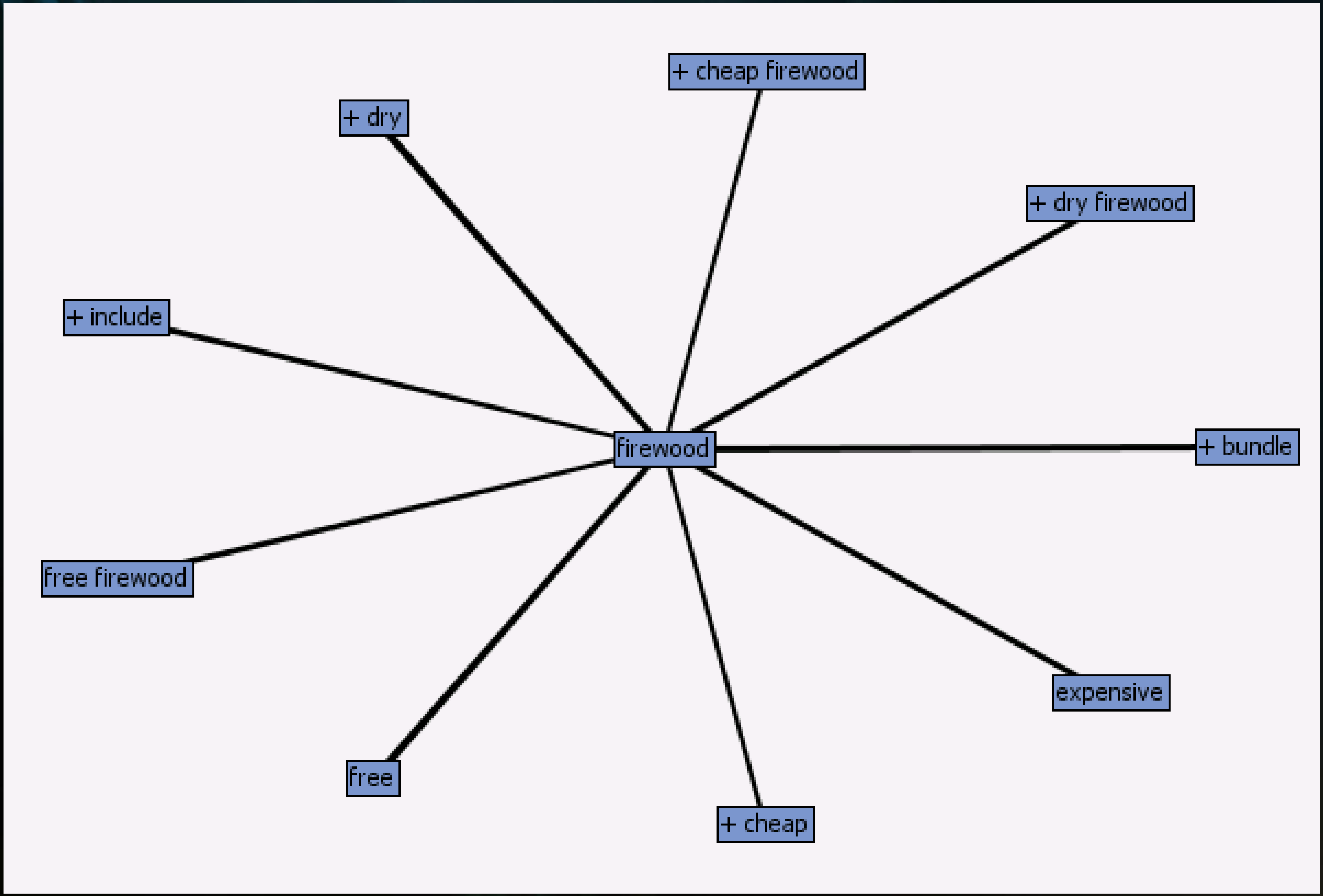
% Represented = 25% (v.s. 28.5%)



Terms

	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE
⊞	site	347	248	☑	0.255	Noun
	firewood	234	209	☑	0.263	Noun
⊞	campground	248	203	☑	0.273	Noun
⊞	good	231	202	☑	0.269	Adj
⊞	park	242	191	☑	0.282	Noun
⊞	shower	219	190	☑	0.278	Noun
⊞	camp	144	124	☑	0.338	Verb
⊞	campsite	143	121	☑	0.342	Noun
⊞	area	141	115	☑	0.352	Noun
⊞	nice	119	112	☑	0.348	Adj
	power	127	107	☑	0.361	Noun
⊞	facility	114	105	☑	0.358	Noun
⊞	great	114	102	☑	0.363	Adj
⊞	washroom	113	95	☑	0.376	Noun
⊞	bathroom	104	92	☑	0.378	Noun
⊞	wood	100	81	☑	0.399	Noun
	nice	85	81	☑	0.392	Adv

Firewood’s related terms (“sub-categories”)



Refining the model…

Synonyms:

firewood = wood

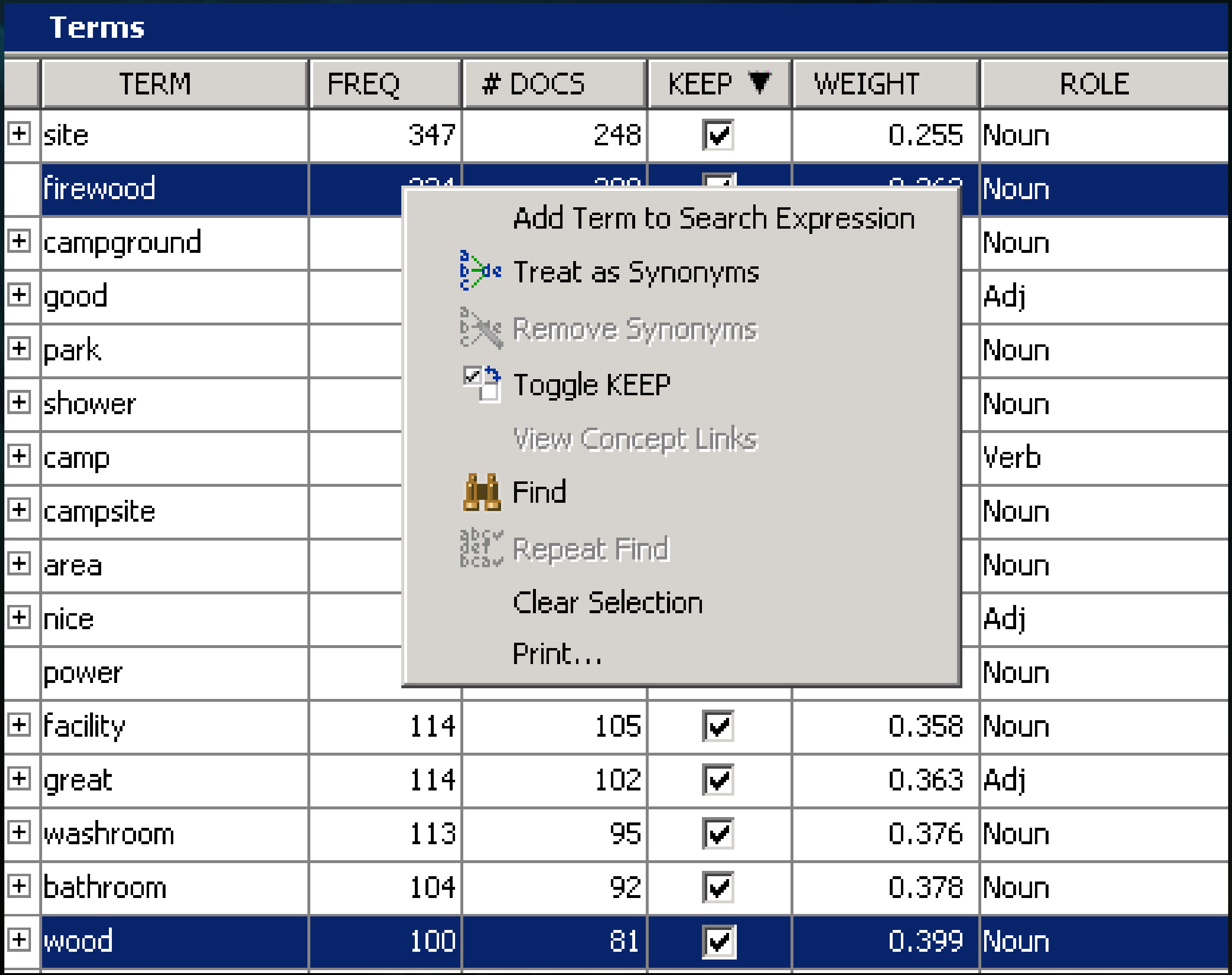
include = bundle

and more…

Text Miner’s ability to set synonyms and handle other lexical relations outweighs and outperforms days spent re-categorizing.

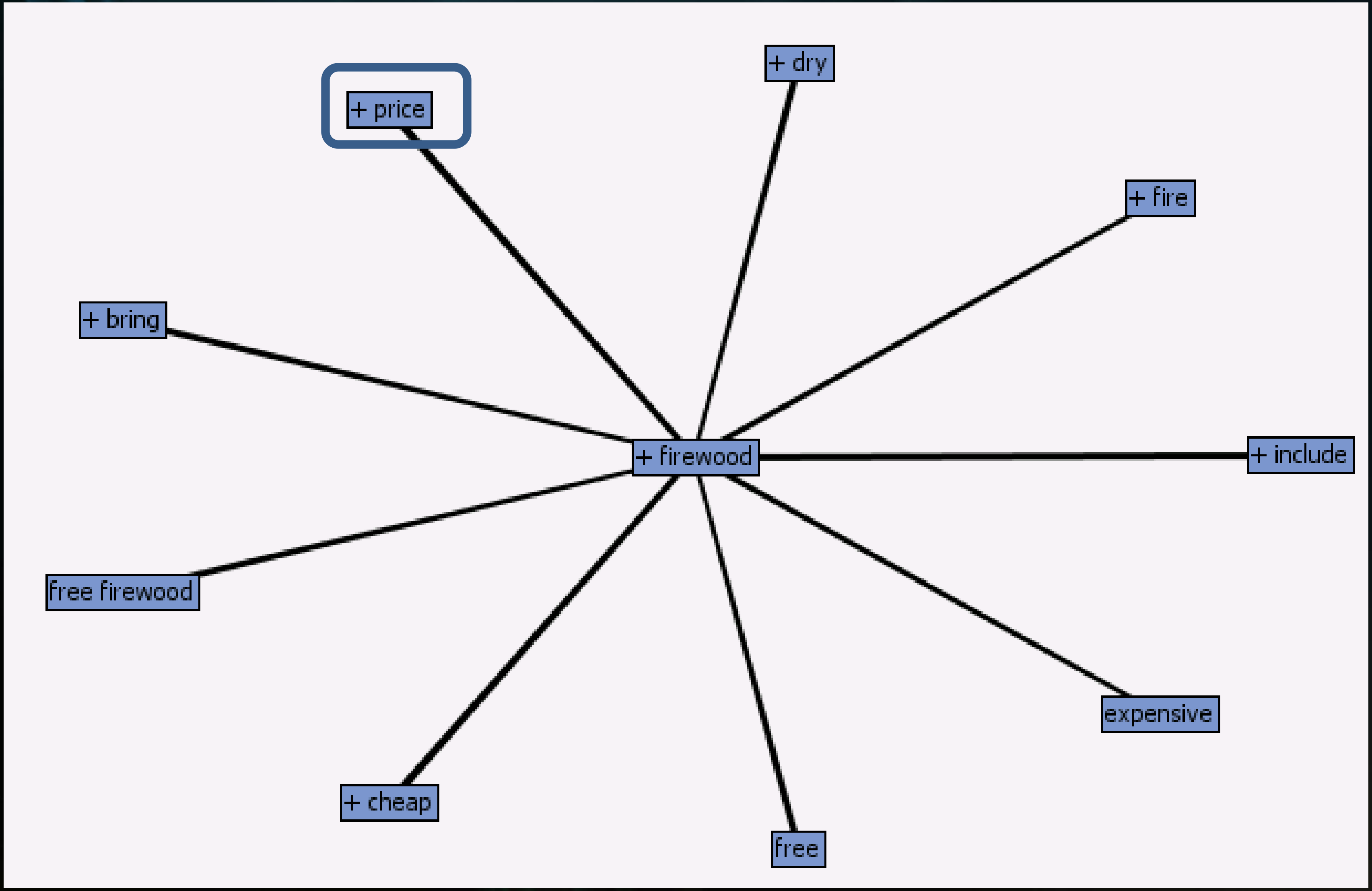
Refining the model…

Ability to handle synonyms



	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE
⊞	site	347	248	☑	0.255	Noun
	firewood	[illegible]	[illegible]	[illegible]	[illegible]	Noun
⊞	campground					Noun
⊞	good					Adj
⊞	park					Noun
⊞	shower					Noun
⊞	camp					Verb
⊞	campsite					Noun
⊞	area					Noun
⊞	nice					Adj
	power					Noun
⊞	facility	114	105	☑	0.358	Noun
⊞	great	114	102	☑	0.363	Adj
⊞	washroom	113	95	☑	0.376	Noun
⊞	bathroom	104	92	☑	0.378	Noun
⊞	wood	100	81	☑	0.399	Noun

Continuous refining

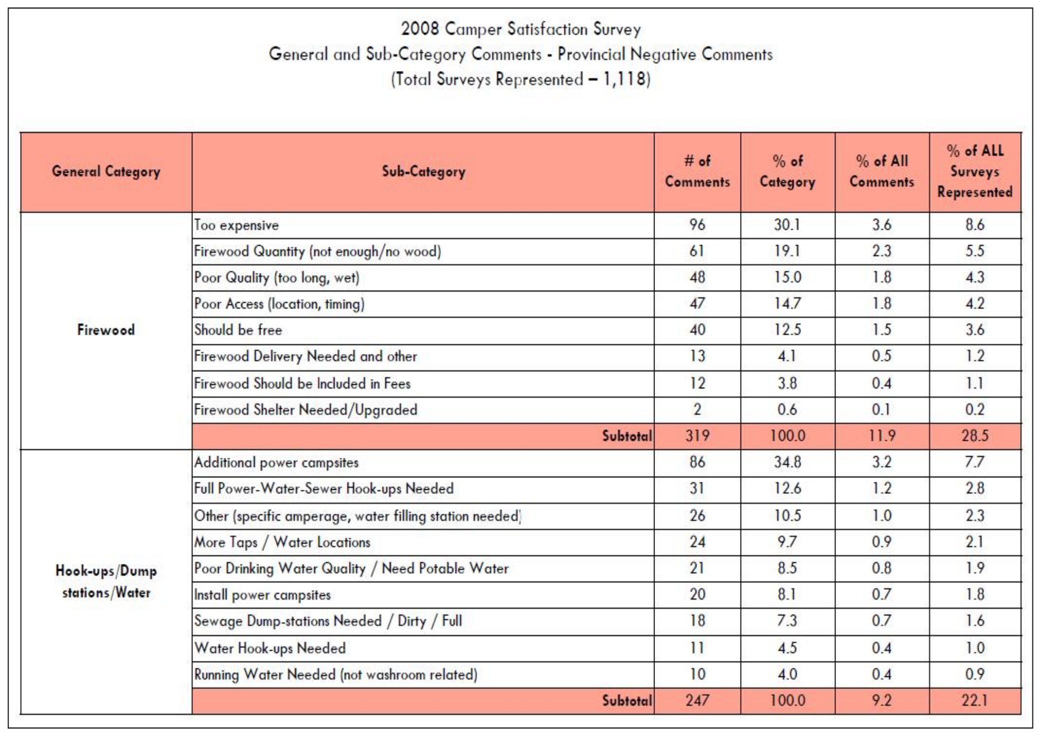


Actionable Intelligence

There is no way to determine <u>outcomes</u> like this:

“Percieved camper safety can be impacted by the level of noise, bathroom or site cleanliness, and the amount of officer patrols. Failing in any of these may contribute to campers feeling unsafe”.

From <u>output</u> like this:



2008 Camper Satisfaction Survey
General and Sub-Category Comments - Provincial Negative Comments
(Total Surveys Represented – 1,118)

General Category	Sub-Category	# of Comments	% of Category	% of All Comments	% of ALL Surveys Represented
Firewood	Too expensive	96	30.1	3.6	8.6
	Firewood Quantity (not enough/no wood)	61	19.1	2.3	5.5
	Poor Quality (too long, wet)	48	15.0	1.8	4.3
	Poor Access (location, timing)	47	14.7	1.8	4.2
	Should be free	40	12.5	1.5	3.6
	Firewood Delivery Needed and other	13	4.1	0.5	1.2
	Firewood Should be Included in Fees	12	3.8	0.4	1.1
	Firewood Shelter Needed/Upgraded	2	0.6	0.1	0.2
	Subtotal	319	100.0	11.9	28.5
Hook-ups/Dump stations/Water	Additional power campsites	86	34.8	3.2	7.7
	Full Power-Water-Sewer Hook-ups Needed	31	12.6	1.2	2.8
	Other (specific amperage, water filling station needed)	26	10.5	1.0	2.3
	More Taps / Water Locations	24	9.7	0.9	2.1
	Poor Drinking Water Quality / Need Potable Water	21	8.5	0.8	1.9
	Install power campsites	20	8.1	0.7	1.8
	Sewage Dump-stations Needed / Dirty / Full	18	7.3	0.7	1.6
	Water Hook-ups Needed	11	4.5	0.4	1.0
	Running Water Needed (not washroom related)	10	4.0	0.4	0.9
	Subtotal	247	100.0	9.2	22.1

Noise in parks?

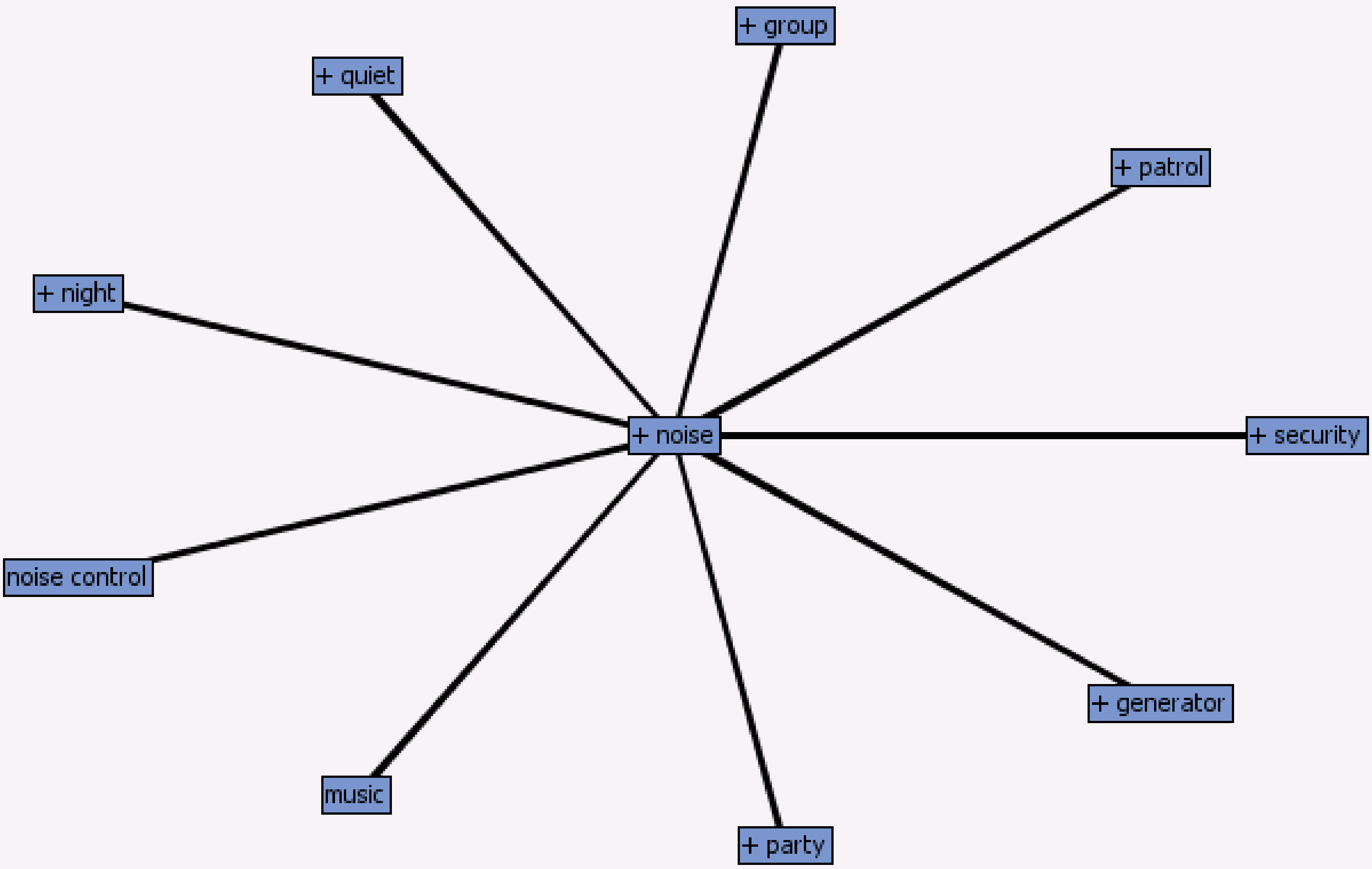
Noise in parks

Leveraging Existing Data

- Survey comments 2002 – 2011
- 18,510 comments

Analysis

- Text Mining
- No specific park stood out as a problem area
  - →Not to say there are none
  - →Only those ~100 top visited parks


+ group
+ quiet
+ patrol
+ night
+ noise
+ security
noise control
+ generator
music
+ party

Noise in Parks? - Results

Noise is 6% of all comments

- Generators = 1.8%
- Parties = 1.3%
- Music = 1.2%
- Barking = 0.7%

** some double counting **

The magnitude of the problem is no bigger than other problems (e.g. Boat launch, road issues), but the sentiment is strong, making this an important issue.

Noise in Parks? - Sentiment

Sentiment:

- 10% specifically mentioned banning generators.
- The remainder demand quiet time respect.
- More patrols to better control noise.
- A few suggested identifying sites for generator users.

- Educating (improve information services) and improved enforcement are suggested.

Letters from the Public (a.k.a. Action Requests)

Action Requests – Text Mining

Difficult because AR process is not built with the mindset that public feedback is data.

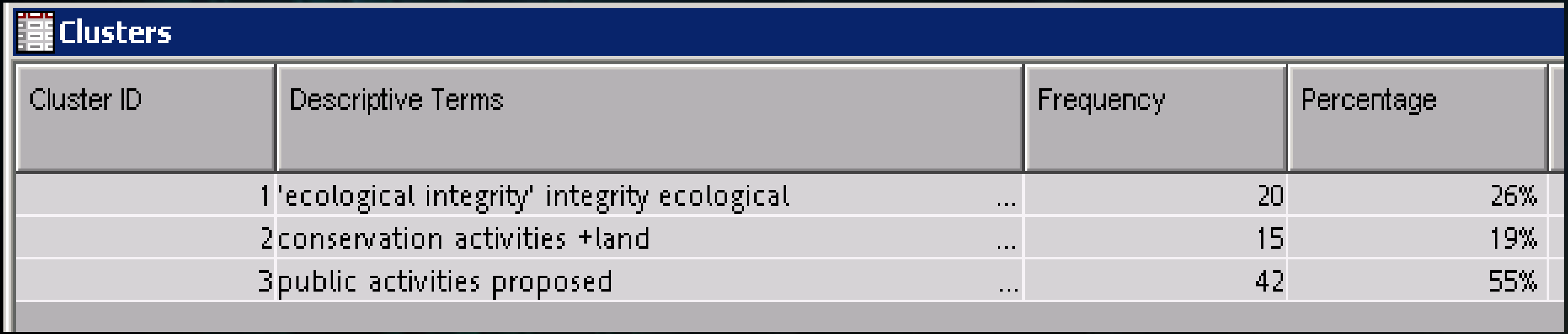
Dataset used in this example is a folder of PDF documents painstakingly downloaded from ARTS, one click at a time.

(Courtesy of Peter Weclaw ☺)

Action Requests – Bill 29 *dialogue*

Each PDF contains public letters AND our response

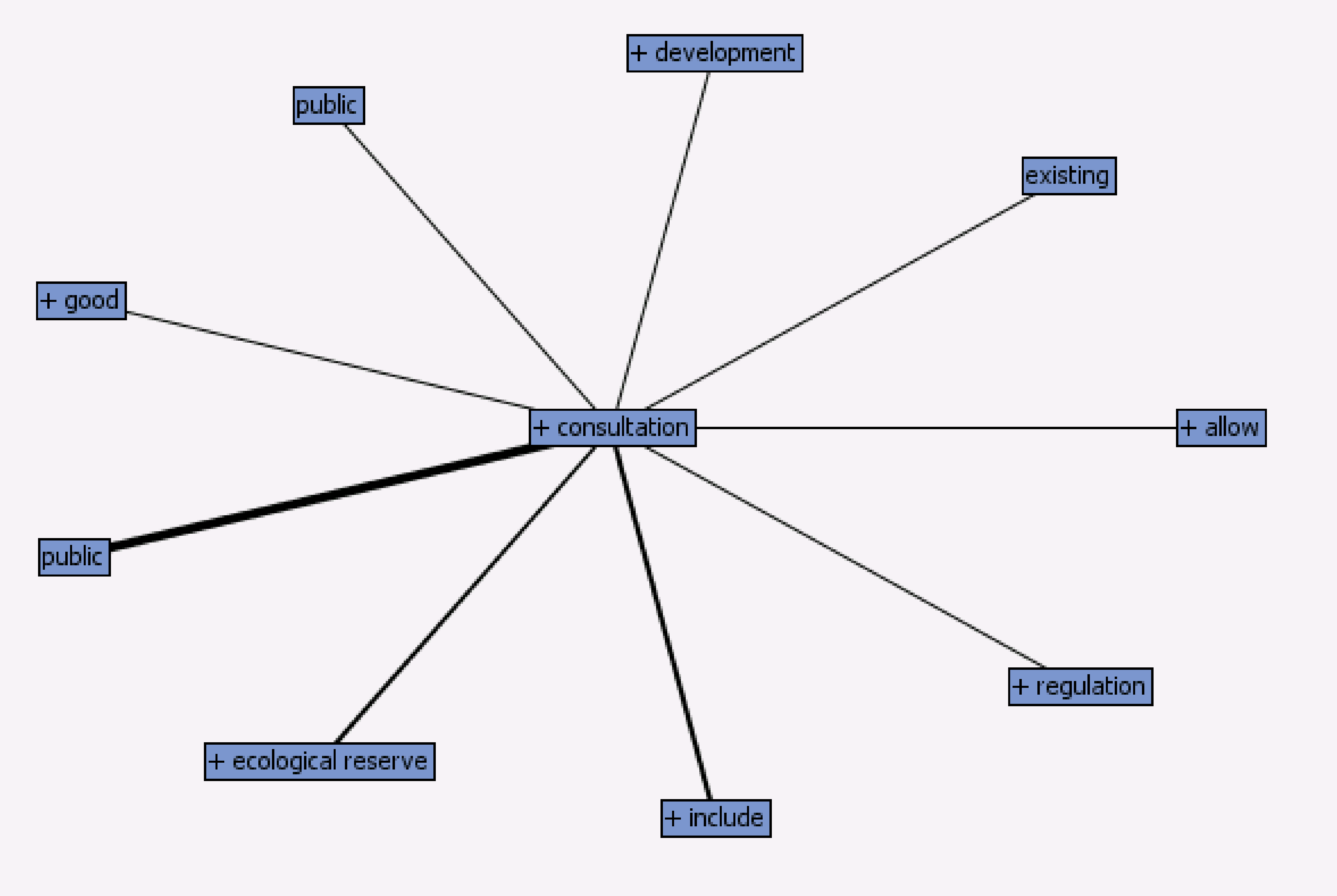
Example of TM's accuracy and Clustering feature:



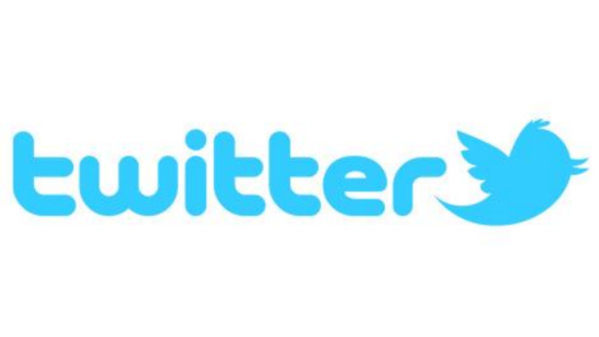
Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage
1	'ecological integrity' integrity ecological ...	20	26%
2	conservation activities +land ...	15	19%
3	public activities proposed ...	42	55%

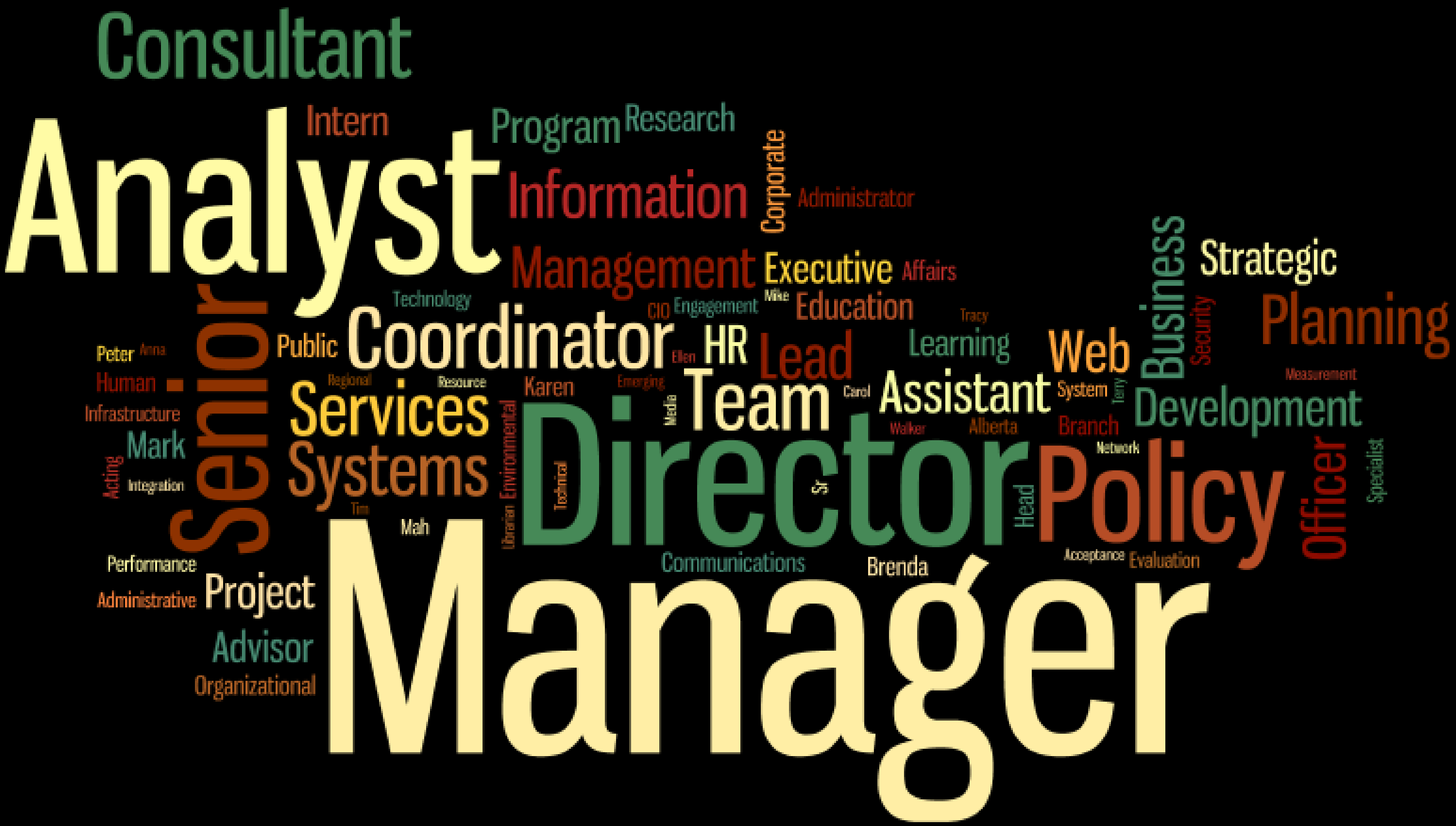
Action Requests – Bill 29



Social Media Monitoring



Who's on yammer ?



Information is the new Currency $$$

amazon
YouTube
facebook
Y!
YAHOO!
LinkedIn
yammer
Google
twitter
Meetup

New Website ROI:

www.AlbertaParks.ca

Titanus
1061
GODZILLA CONTRO I ROBOT
CON MASAAKI DAIMON KAZUYA AOYAMA REIKO TAJIMA BARBARA LYNN
AKIHIKO HIRATA HIROSHI KOIZUMI
PRODUTTORE ESECUTIVO TOMOYUKI TANAKA
SCRITTO E DIRETTO DA JUN FUKUDA
UNA PRODUZIONE TOHO COMPANY Ltd.
COLORE L.V. DI LUCIANO VITTORI
ANNO DI EDIZIONE 1975

Old v.s. New *AlbertaParks.ca* website

Comparing consistency and findability of information of the new website v.s. the old website.

Email as a data source

Spam-a-thon

- One mild mannered Wednesday, GOA was hit with massive amounts of Spam email.
- Recipients of the spam kept responding to the spam and everyone on the list would get it. (i.e. We were spamming ourselves).
- 138 employees responded to the spam → nearly 1 email every 2 minutes.

Spam-a-thon

- Demonstration of analytics and visualization of the SPAM event

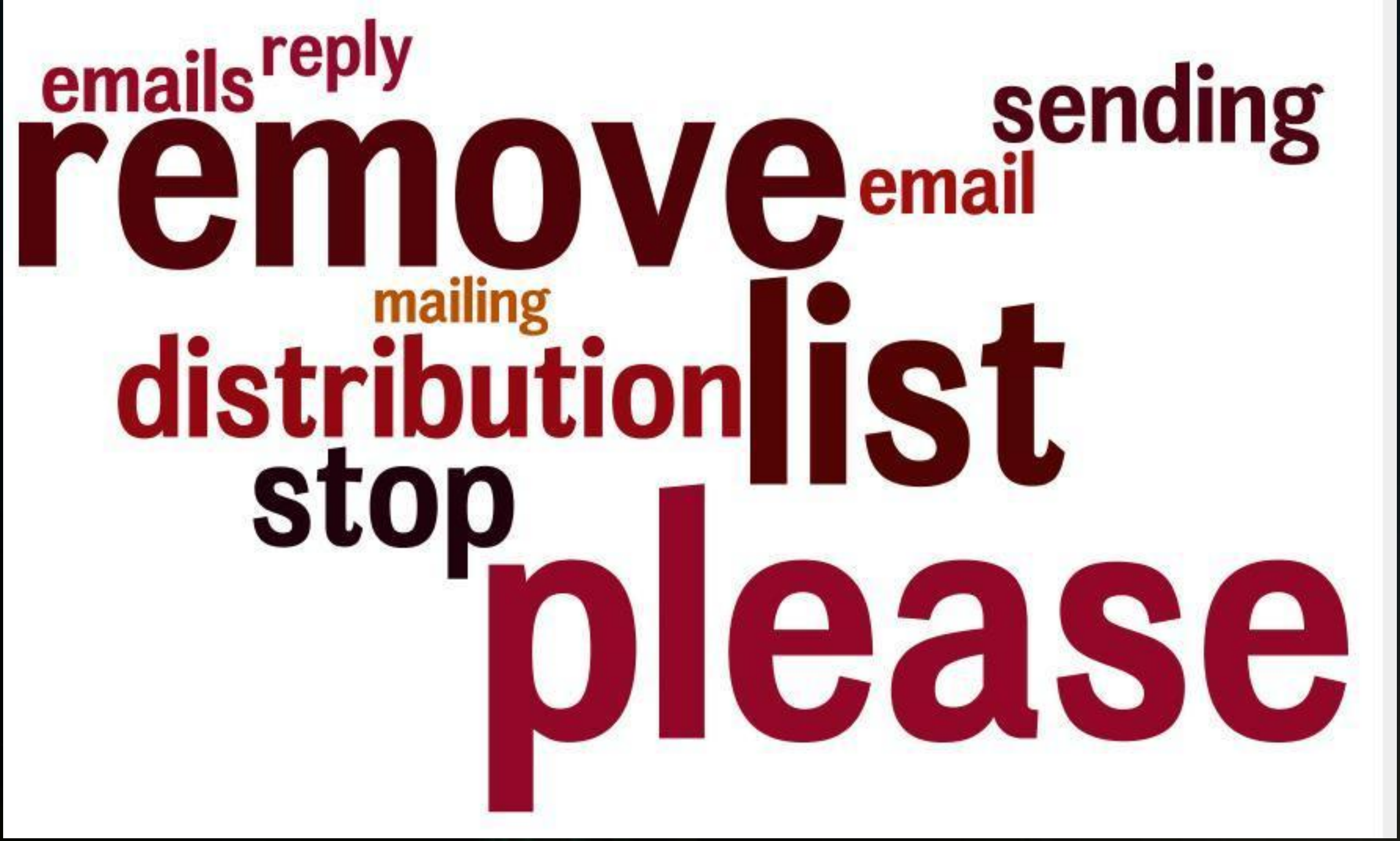
- The insight provides a new perspective on the problem and educates people on a better understanding of spam issues.

- EMAIL as a data source

Spam-a-thon

- 7% of people marked the spam with high importance.
- Half (50%) asked to be removed from the distribution list
- 27% asked the spammer to ‘just stop’.
- Our Canadian side shines brightly since 54% of all responses contained Please and/or Thank you…

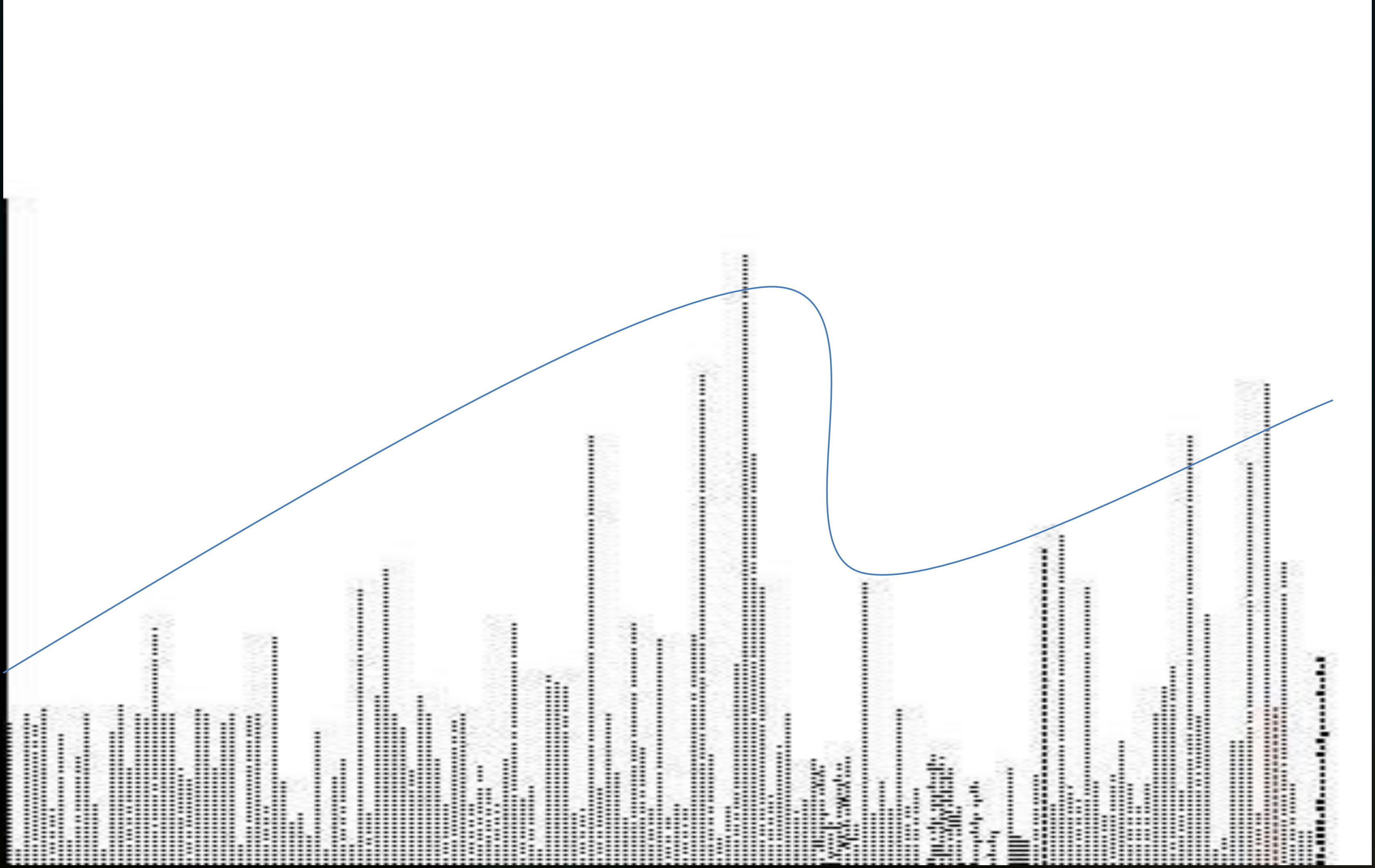
Spam-a-thon: Canadian Politeness



Spam-a-thon

- 14% of people replied with "Ditto“ (E.g. they write "Me too“)
- 70% of the responses came from women (I’m not going to read into this, but feel free to make your own interpretations with your work pals)
- 20% of people responded to tell people that by responding, they are contributing to the spam. This perpetuated the problem. Oh, the irony.

Graphing length of response text



“Nothing is more terrible than activity without insight.”

Thomas Carlyle

Contact Information

Jared Prins B.Sc., Program Analyst

Business Integration and Analysis Section
Alberta Tourism, Parks & Recreation
2nd Flr. Oxbridge Place, 9820-106 Street
Edmonton, AB
T5K 2J6

Phone: 780.427.6313
Fax: 780.427.5980
www.AlbertaParks.ca

Alberta Parks