

# DISCOVERY SUMMIT

EXPLORING DATA • INSPIRING INNOVATION

SAN DIEGO | SEPTEMBER 14-17 2015



## **Harness the Power of JMP: Big Data and Social Media for Competitor Analytics**

**Jim Wisnowski, Adsurgo**

**Flor Castillo, SABIC**

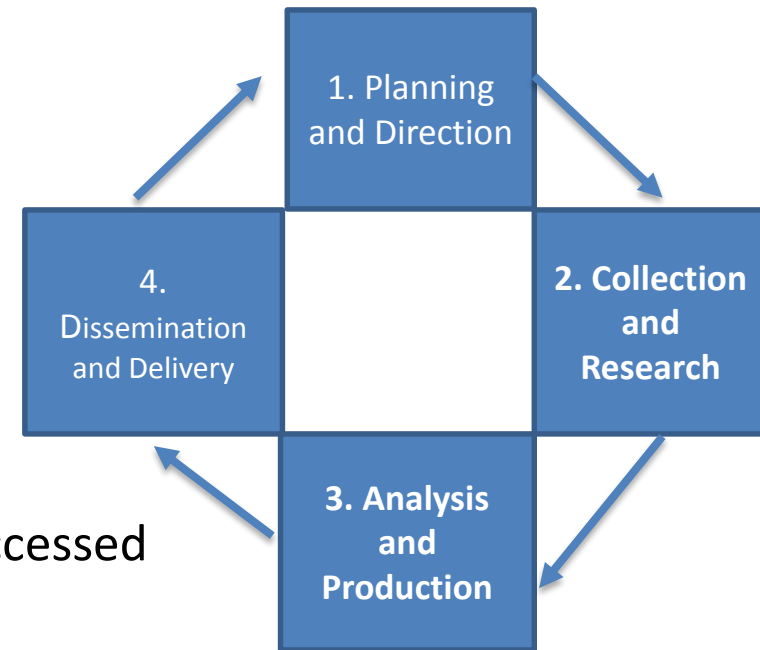
**Andrew Karl and Heath Rushing, Adsurgo**

# Objectives

- Describe competitive intelligence and data requirements
- Demonstrate analytics from web-based tools
- Demonstrate web scraping of competitors
- Show conversion of text documents to JMP data tables
- Demonstrate text analytics in JMP
  - Scholarly journal article collection
  - Patent searches
  - Topic analysis, clustering documents and clustering words

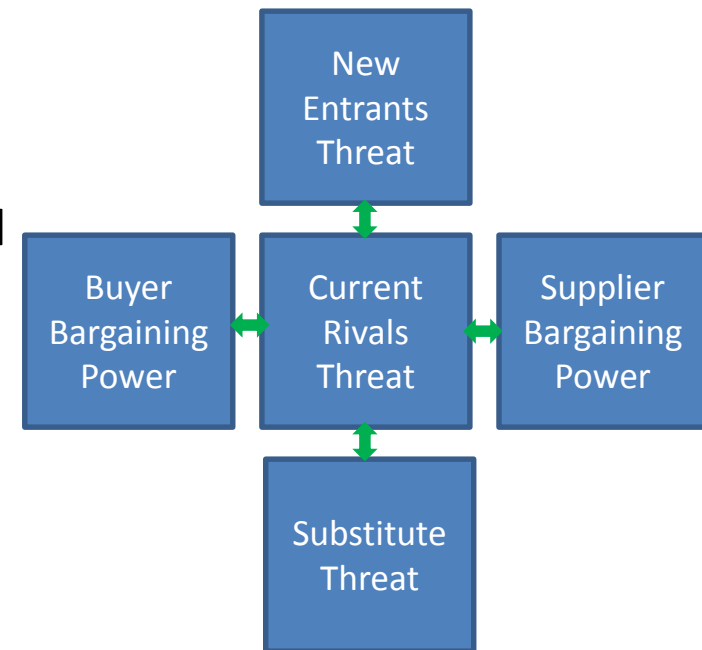
# Competitor Analysis

- Competitive Intelligence (CI) Analysis
  - Focuses on external forces to organization: products, competitors, customers
  - Decision support=>strategic and tactical, protect your own=>counter
  - Not industrial espionage
  - Open data sources
  - Ethical practices
- 4 common phases of the CI Cycle
- Our focus..
- Phase 2. Data collection and research
  - Most often unstructured, electronically-accessed
- Phase 3. Analysis and Production
  - Transform raw data to actionable intelligence; eliminate blindspots
  - Most difficult, wide variance of capabilities and interpretation
  - May take new methods and should be **persistent** surveillance



# Classical Competitor Analysis

- SWOT Analysis-> External OPPORTUNITIES and THREATS
  - PEST(LE): political, economic, social, technological, legal, environment
- Porter's 5 Forces and Porter's 4 Corners (predict competitor future moves)
- Competitor benchmarking, arrays, matrices (BCG ...)
  - KPIs: distribution channels, technological edge, pricing, market share, customer focus, financial stability, workforce, facilities, partnerships...
  - Weight each KPI and evaluate current and future competition
- Value chain analysis, Monte Carlo simulation, and many other frameworks
- ALL need reliable data for fuel



# Competitive Intelligence Data

- In the past, only CI specialists could get data, now their role is morphing into analyzing that data as well
- Value added content—new “coin of the realm” repackaging data understandable to marketing and strategy
- You won’t have the nice structured data like your enterprise data for transactions, call center transcripts, customer profiles etc.
- Many open source opportunities and many great proprietary (unfortunately) databases and tools
- Vast number of sources to paint the landscape
  - Articles, speeches, annual reports, web, trade shows, patents, ...
  - Proprietary competitor databases such as D&B Hoovers and niche-specific
  - Web presence and social media
  - Most will require retrieval and preprocessing

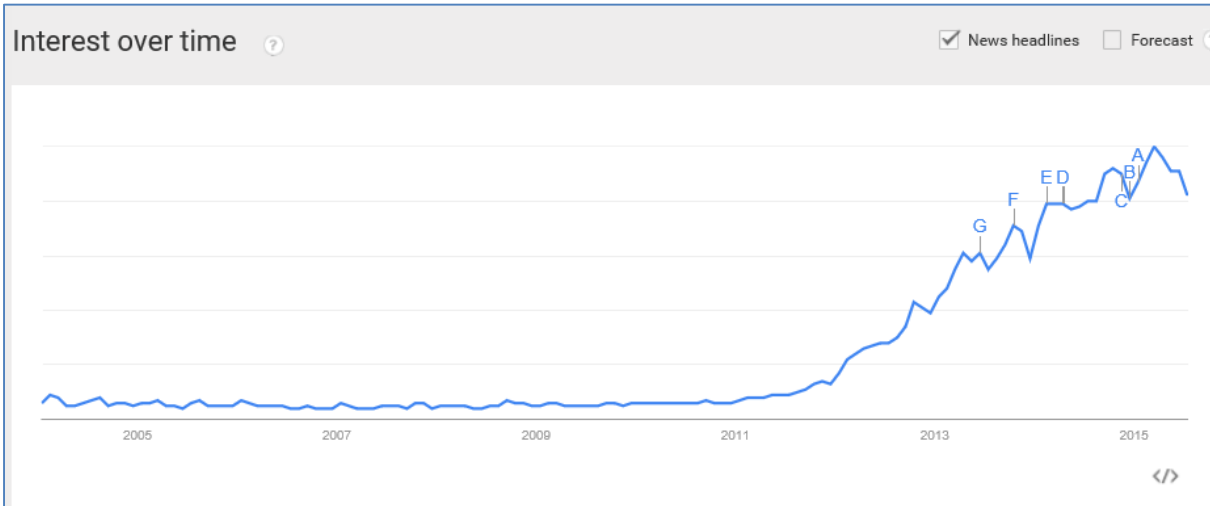
# Text Data is not Clean

- Documents—OCR errors, misspellings, code text from figures and headers, synonyms, and user-specific lingo
- Social networks—many (most!) words not standard with mix of languages, non-standard abbreviations, unusual parts of speech, and grammatically incorrect
- Voice-to-text—recognition errors (10-40%), ums & ahs, slang, same phrases repeated...”hello this is JW from ABC Corp how can I help you today.”; “Thank you and have a great day.”
- Word Error Rates (WER) are both lexical and semantic
  - Lexical=> tonight, 2nt, 2night, nite, tonite
  - Semantic => Shes a gr8 sk8r, she is a grate skatr
- Remedies require time and variety of applications
  - JMP recode very helpful
  - JSL character formula scripts
  - Text parsing utilities

# Web-Based CI Collection Tools

- Site-centric for direct competitors or known sites of interest
  - Google Analytics, Compete, and SimilarWeb for competitor online consumer behavior, demographics, referring domains
  - Marketing Grader, Majestic for SEO, keyword, landing pages, mobile, click analysis
  - AdWords Keyword Planner & Adbeat to analyze on-line advertising presence
  - Most have little free functionality apart from your own site
- Ecosystem-centric for industry, technology, broader markets
  - Google Trends
  - Raven Tools

# Google Trends: Big Data



Chinchwad	100	<div><div></div></div>
Bangalore	55	<div><div></div></div>
New Okhla Industrial Development A...	54	<div><div></div></div>
Hyderabad	44	<div><div></div></div>
Chennai	41	<div><div></div></div>
San Jose	22	<div><div></div></div>
Seoul	18	<div><div></div></div>

## Related searches ?

### Topics

	Top	Rising
Big data - Industry	100	<div><div></div></div>
Analytics - Industry	10	<div><div></div></div>
Apache Hadoop - Software	5	<div><div></div></div>
Big Data - Musical Group	5	<div><div></div></div>
Data analysis - Industry	0	<div><div></div></div>

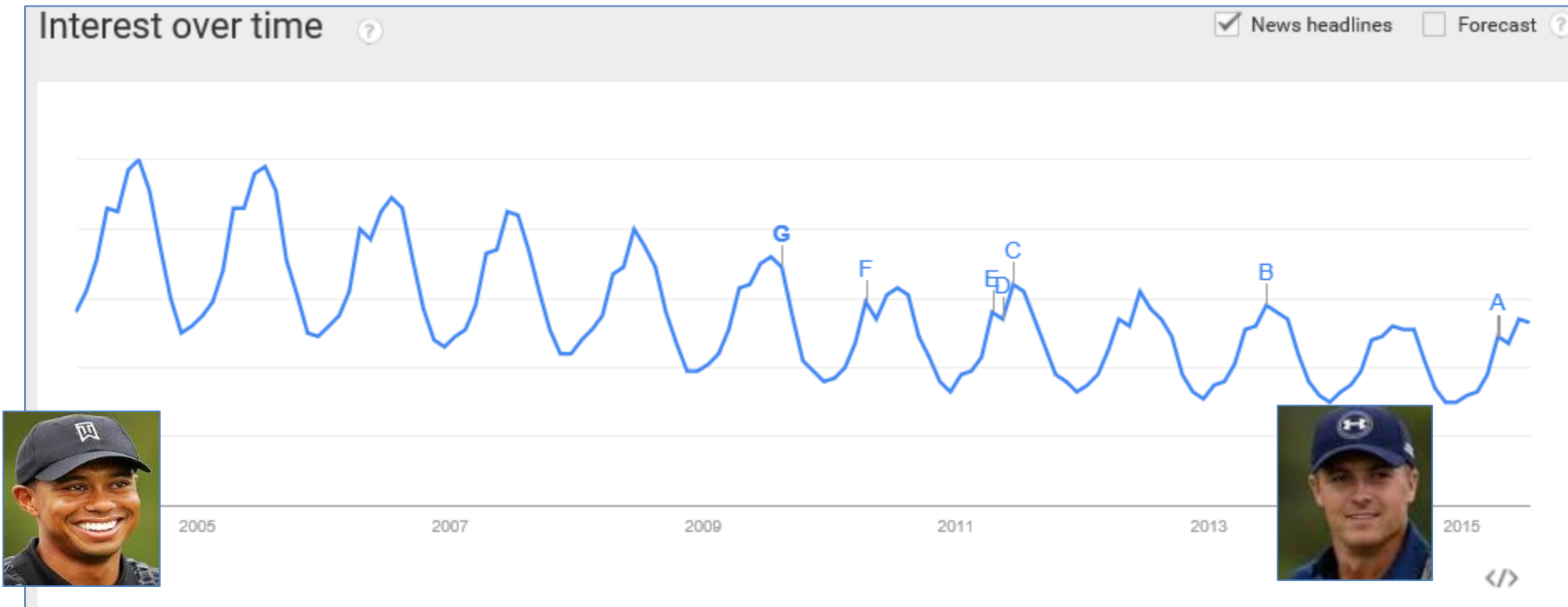
### Queries

	Top	Rising
the big data	100	<div><div></div></div>
big data analytics	95	<div><div></div></div>
data analytics	95	<div><div></div></div>
hadoop	75	<div><div></div></div>
big data hadoop	75	<div><div></div></div>
google big data	40	<div><div></div></div>
big data dangerous	30	<div><div></div></div>



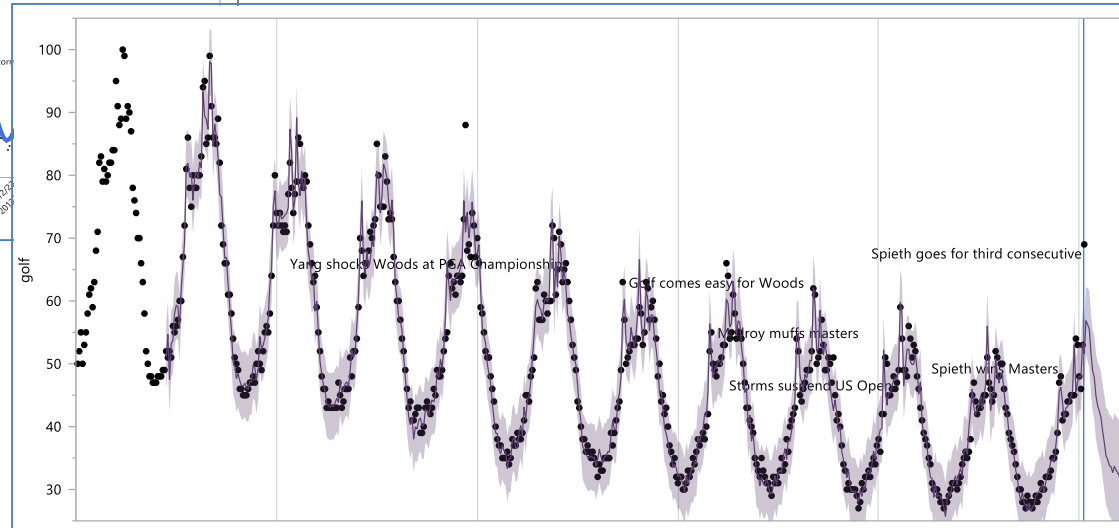
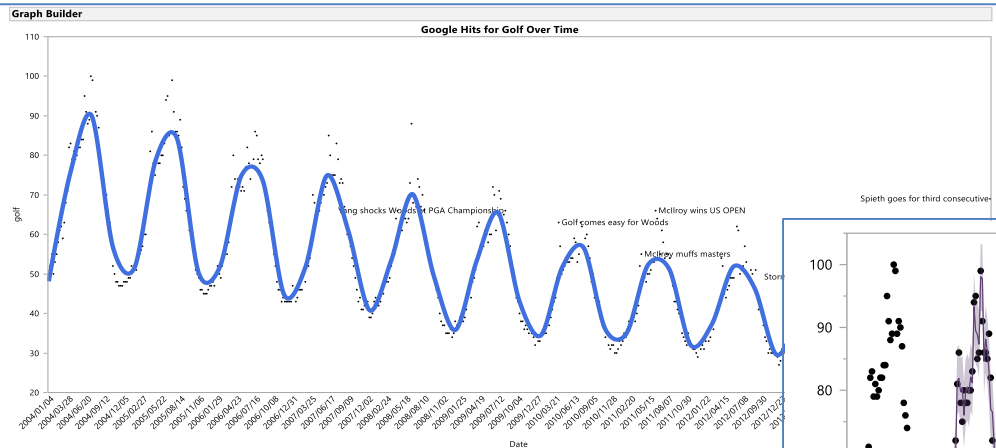
# Google Trends








- Is interest in golf waning? What does this mean for Under Armour?



- JMP Demonstration
  - Google Trends data extract
  - JMP graph builder and Seasonal ARIMA forecast

# JMP Output Google Trends



Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2 .4 .6 .8	MAPE
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Seasonal ARIMA(1, 1, 1)(0, 1, 1)52	546	7.3818947	2674.0555	2691.2952	0.967	2666.0555	0.490535		3.699284
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(1, 0, 1)(0, 1, 1)52	547	7.3420367	2675.5263	2692.7732	0.967	2667.5263	0.235124		3.640049
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(1, 1, 1)(1, 1, 1)52	545	7.3952772	2676.0510	2697.6006	0.967	2666.051	0.180863		3.698551
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(1, 0, 1)(1, 1, 1)52	546	7.3554575	2677.5219	2699.0806	0.967	2667.5219	0.086688		3.640274
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(0, 1, 1)(0, 1, 1)52	547	7.5216347	2683.4990	2696.4288	0.966	2677.499	0.004366		3.752433
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(0, 1, 1)(1, 1, 1)52	546	7.5354575	2685.4959	2702.7356	0.966	2677.4959	0.001608		3.752416
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Seasonal ARIMA(1, 0, 1)(1, 1, 1)52	547	7.5559058	2688.2147	2705.4617	0.966	2680.2147	0.000413		3.699227

## Parameter Estimates

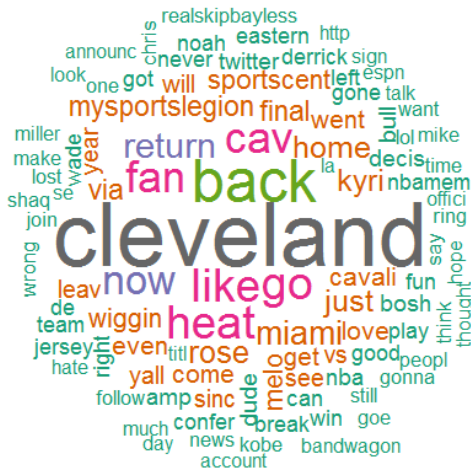
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob> t	Constant
AR1,1	1	1	0.24468215	0.0729851	3.35	0.0009 *	Estimate
MA1,1	1	1	0.78670399	0.0518144	15.18	<.0001 *	0.00726574
MA2,52	2	52	0.40213305	0.0429196	9.37	<.0001 *	
Intercept	1	0	0.00961945	0.0208197	0.46	0.6442	

# Social Media Presence

- Blogs ([google.com/googleblogsearch](http://google.com/googleblogsearch)) and other niche bulletin boards are very good hunting grounds
- LinkedIn (follow company, previous employees, new hires, jobs)
- Facebook
- Twitter
  - Follow #competitors products, # name, employees
  - Check out their lists of followers and how classify
  - Monitor text from Tweets
  - JMP Demonstration
- We don't have nice .csv flat files given to us—text analytics can help

# Twitter in JMP

- JSL script that calls R packages streamR and Twitter815
  - Under Armour's pursuit of LeBron James after he announces he is going back to Cleveland
    - Tweets for 5 mins the day LeBron made his statement
  - Sentiment analysis/opinion with text mining tabulates the number of positive terms and number of negative terms (Harvard IV dictionary)
- |      | Negative | Positive |
|------|----------|----------|
| 1122 | likable  | pleasant |



	Negative	Positive
1132	liable	pleasantry
1133	liar	please
1134	lie	pleased
1135	lifeless	pleasurable
1136	limit	pleasure
1137	limitation	pledge
1138	limp	plentiful
1139	liquidate	plenty
1140	liquidation	poetic
1141	litter	poignant
1142	load	poise
1143	lone	polish
1144	loneliness	polite
1145	lonely	politeness
1146	loner	pomp
1147	lonesome	popular
1148	loom	popularity
1149	lose	populous
1150	loser	portable

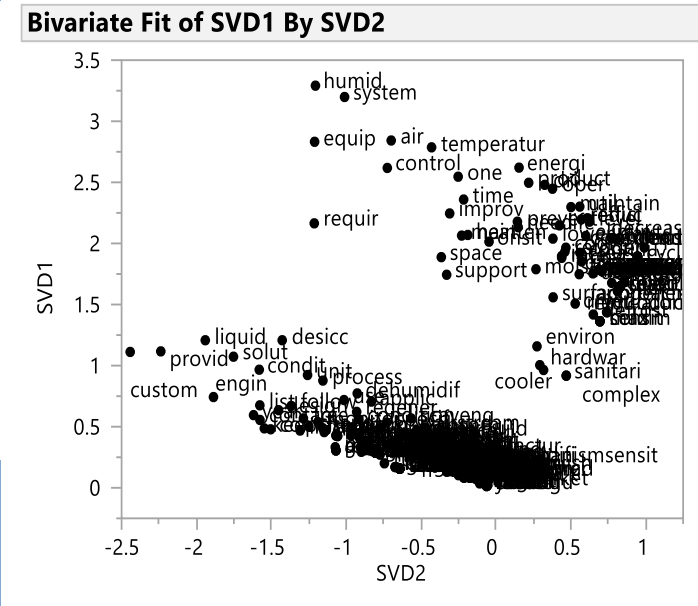
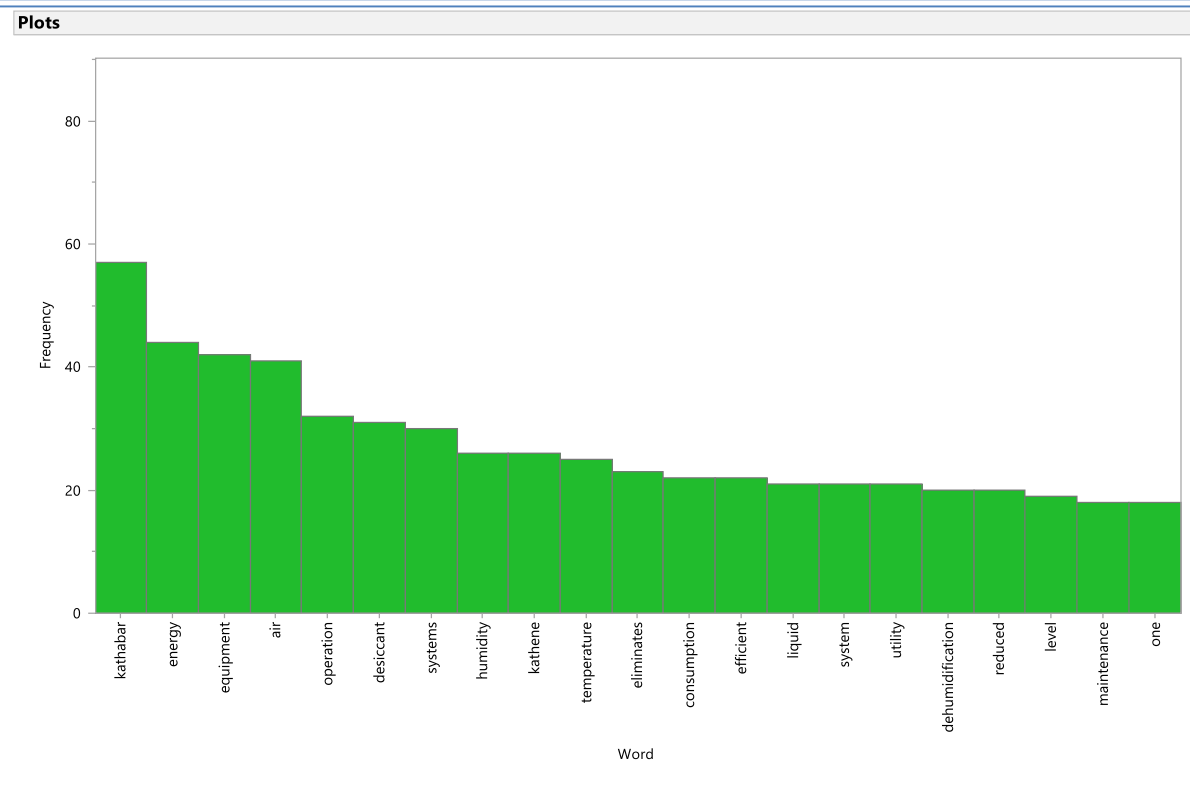
Positive	Sum	3533
Negative	Sum	1626

- Job advertisements (Indeed.com)
- Conferences and media
- Technology
- Keywords in SEO
- Website architecture really should describe whole business
- Use their best practices
- How do they “hook” visitors?

# Web Scraping Your Competitors

- One green energy technology is liquid desiccant air conditioning; we want to find out about one of the major players in this space
- Scrape [www.kathabar.com](http://www.kathabar.com) and analyze with text mining
- JSL script that calls R packages Rcurl and Boilerpipe
- Use JMP to find word counts for general impressions and text analytics for exploration and discovery
  - Consumer Research>Categorical>Response Role=Multiple>Free Text
  - Use cluster analysis of document term matrix (SVDs) to find themes and information about liquid desiccant AC
- What if have many files? Put them in a folder and read into JMP data table with JSL script

# Web Scrapping Competitors



- Frequencies from Pareto are helpful but need context from eigenanalysis and clustering

	Name	desicc
1	desicc	0
2	liquid	6.7605806587
3	system	8.5003073684
4	process	8.6240518105
5	dri	8.8911772
6	product	8.9001053874
7	applic	9.1374522553
8	humid	9.1771471813
9	manufactur	9.3465650954
10	reliabl	9.3792174176
11	solut	9.4465338254
12	condit	9.4539746553
13	learn	9.5609119003
14	moistur	9.6377016924
15	dehumidif	9.7417120967
16	effici	9.8127825986
17		

# Patents

- Patent profiles essential for many industries for CI
- Fortunately, rich and open databases exist
- World IP Organization PATENTSCOPE search abstracts
- JMP Free Text can form indicator variables for tagging your patent data for quick search and analytics

Categorical			
Free Text Word Counts for Summary			
Free Text Word Counts for Summary			
Number of Words	Number of Cases	Number of Non-empty Cases	Portion Non-empty per Case
1193	126	125	0.9921
Alphabetical Word Counts			
Frequency-Ordered Word Counts			
"air" 337 "desiccant" 295 "liquid" 157 "heat" 150 "such" 136 "systems" 128 "system" 113 "cooling" 96 "energy" 87 "generally" 85 "stream" 82 "water" 74 "space" 70 "typically" 70 "used" 64 "conditioning" 59 "however" 59 "solar" 56 "not" 55 "humidity" 54 "thermal" 54 "been" 51 "large" 51 "temperatures" 50 "use" 49 "temperature" 47 "also" 45 "modules" 45 "more" 45 "building" 42 "higher" 41 "heating" 40 "membrane" 39 "buildings" 38 "equipment" 38 "transfer" 38 "flow" 37 "fluid" 36 "oftentimes" 36 "other" 36 "required" 36 "dehumidification" 35 "moisture" 35 "between" 34 "vapor" 34 "cost" 32 "dehumidify" 32 "high" 32 "require" 32 "risk" 32 "years" 32 "application" 31 "desiccants" 31 "outside" 31 "efficient" 30 "since" 30 "carry-over" 29 "conventional" 29 "electricity" 29 "properly" 29 "absorption" 28 "coil" 28 "module" 28 "need" 28 "relates" 27 "thus" 27 "using" 27 "control" 26 "cool" 26 "furthermore" 26 "needs" 26 "sensible" 26 "conditions" 25 "power" 25 "results" 25 "filter" 24 "low" 24 "lower" 24 "media" 24 "spaces" 24 "units" 24 "where" 24 "two" 23 "chillers" 22 "costs" 22 "even" 22 "increase" 22 "membranes" 22 "compression" 21 "many" 21 "requires" 21 "side" 21 "solutions" 21 "source" 21 "well" 21 "larger"			



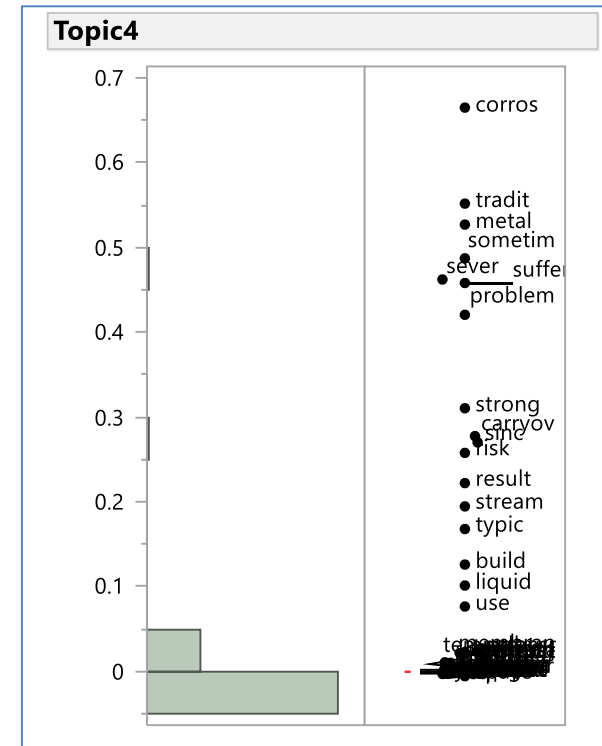
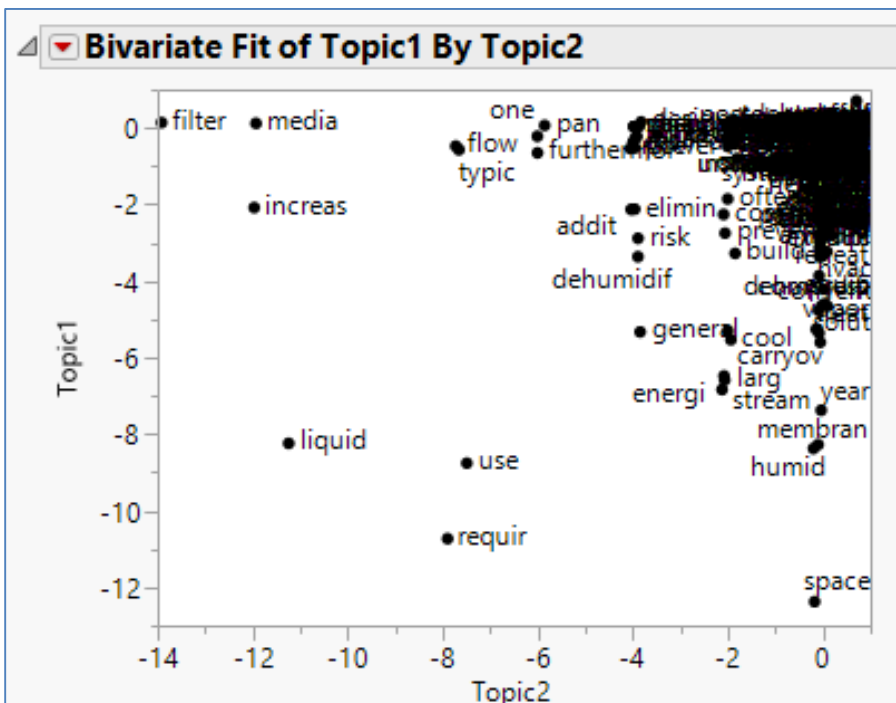
# Investigate Word Correlations

- From the indicator matrix, run multivariate platform to see significant pairwise correlation
- Negative correlations also of interest (solar vs thermal = -0.8)

Pairwise Correlations		
Variable	by Variable	Correlation
chillers	absorption	1.0000
pvt	pv	1.0000
packed	bed	1.0000
flat	collectors	1.0000
media	filter	1.0000
carryover	allow	1.0000
unglazed	glazed	1.0000
sides	side	1.0000
relates	application	0.9639
present	relates	0.9639
strongly	carry-over	0.9618
strongly	corrosive	0.9618
micro-porous	membranes	0.9618
plate	collectors	0.9388
flat	plate	0.9388
present	application	0.9284
corrosive	carry-over	0.9243
climates	itself	0.9118
adiabatic	requires	0.9118
treated	quantities	0.9050
quantities	prevent	0.9050
efficiencies	modules	0.8893
amounts	compression	0.8660
waste	higher	0.8575
climates	occupant	0.8575
climates	inside	0.8575
climates	parallel	0.8575
climates	reduce	0.8575
occupant	comfort	0.8507
quantities	solutions	0.8453
exposed	allow	0.8413

# Patent Data Analysis

- We can find themes and topics in patents
- Quickly locate the associated records with the themes by sorting on the topic
- Subject matter expertise goes a long way:  
pv=photo-voltaic; pvt=photo voltaic-thermal

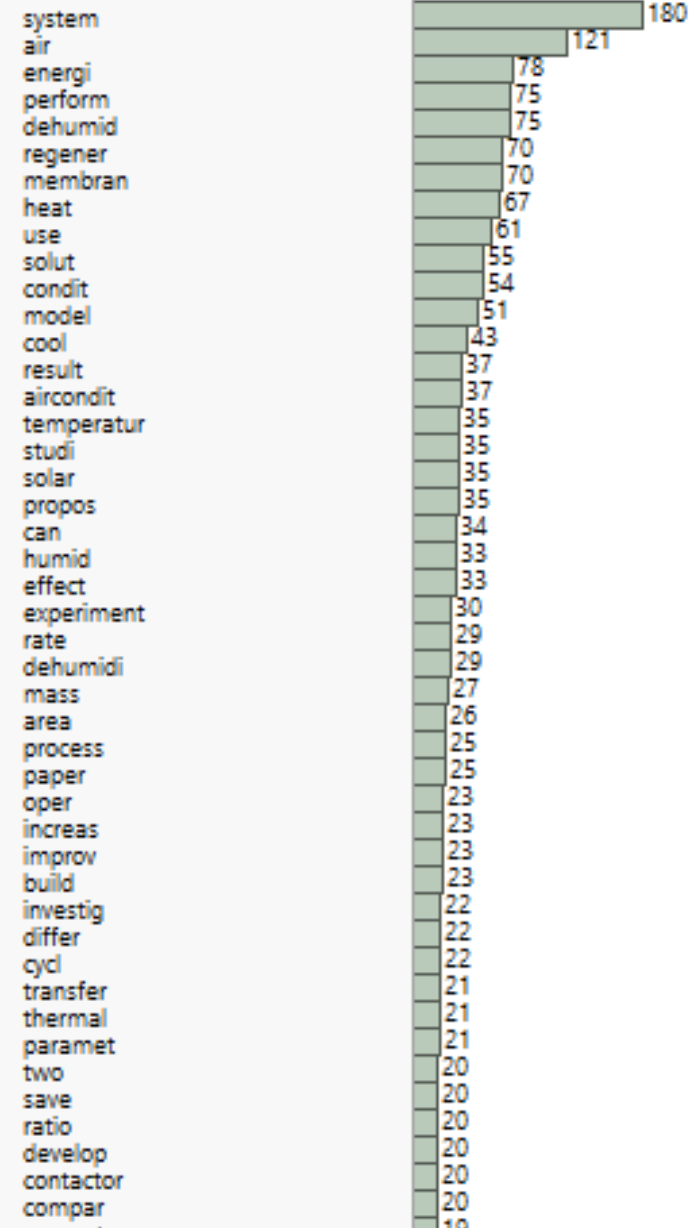


	Summary	stop	SVD1
1	Liquid desiccant systems however have traditionally suffered from the risk of desiccant carry-over into the air stream resulting in sometimes severe corrosion problems in the building since the desiccants that are used are typically strongly corrosive to metals.		
2			
3			
4			
5	[0004] Membrane modules often suffer from proble...		0.0076865286
6	[0006] Heat exchangers (mostly for 2 fluids) are very ...		-0.012228009
7	strongly corrosive, even in small quantities, so numer...		-0.002633753
8	...		...

# Liquid Desiccant Journal Articles

- Collected 45 refereed journal articles on liquid desiccant membrane
- Most from 2013-2015 though a few date to 2010
- Translating pdf to text for JMP was difficult and had varying success rates based on numerous methods
  - Equations and non-standard characters problematic
  - Text from figures fragmented
- Several improvements added to existing tools to ensure success for future conversion
- Text in References section obscured analysis so it was removed

# Liquid Desiccant Journal Articles

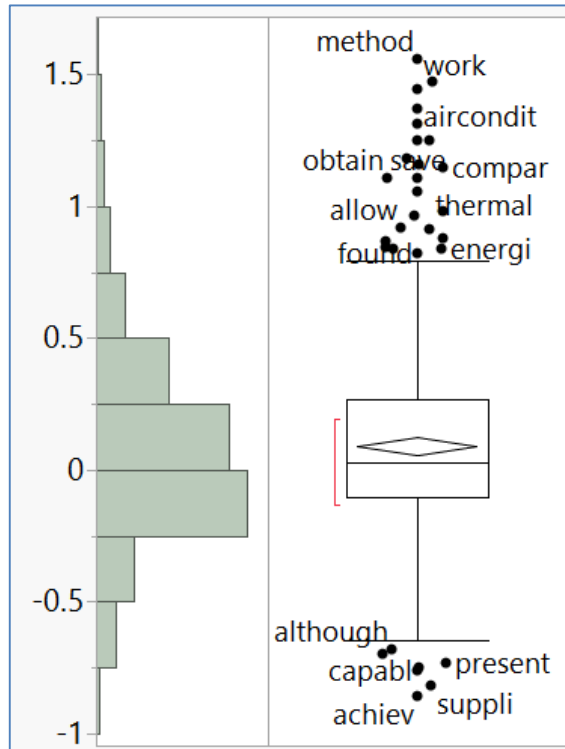


# Cluster on Journal Documents

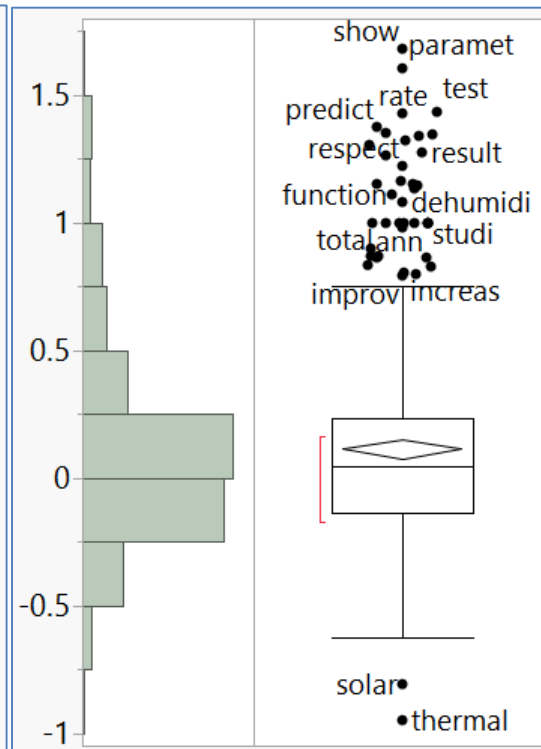
Cluster	file.name
2	2010_Niu_Performance-analysis-of-liquid-desiccant-based-air-conditioning-system-under-variable-fresh-...
2	2011_Ge_Control-strategies-for-a-liquid-desiccant-air-conditioning-system.txt
2	2011_Ge_Model-based-optimal-control-of-a-dedicated-outdoor-air-chilled-ceiling-system-using-liquid-de...
2	2011_Xiao_Control-performance-of-a-dedicated-outdoor-air-system-adopting-liquid-desiccant-dehumidifi...
2	2013_Qi_Investigation-on-wetted-area-and-film-thickness-for-falling-film-liquid-desiccant-regeneration-s...
2	2015_Angrisani_Experimental-assessment-of-the-energy-performance-of-a-hybrid-desiccant-cooling-syste...
2	2015_Das_Simulation-of-potential-standalone-liquid-desiccant-cooling-cycles.txt
2	2015_Wang_Model-based-optimization-strategy-of-chiller-driven-liquid-desiccant-dehumidifier-with-gen...
3	2012_Qi_Investigation-on-air-conditioning-load-profile-and-energy-consumption-of-desiccant-cooling-sy...
3	2014_Qi_Energy-consumption-and-optimization-of-internally-cooled-heated-liquid-desiccant-air-conditio...
3	2014_Qi_Energy-performance-of-solar-assisted-liquid-desiccant-air-conditioning-system-for-commercial-...
4	2013_Mohammad_Historical-review-of-liquid-desiccant-evaporation-cooling-technology.txt
4	2013_Mohammad_Survey-of-hybrid-liquid-desiccant-air-conditioning-systems.txt
5	2013_Woods_A-desiccant-enhanced-evaporative-air-conditioner-Numerical-model-and-experiments.txt
5	2014_Woods_Membrane-processes-for-heating-ventilation-and-air-conditioning.txt
6	2010_Bergero_Performance-analysis-of-a-liquid-desiccant-and-membrane-contactor-hybrid-air-conditioni...
6	2011_Bergero_On-the-performances-of-a-hybrid-air-conditioning-system-in-different-climatic-conditions...
7	2013_Mohammad_Artificial-neural-network-analysis-of-liquid-desiccant-regenerator-performance-in-a-so...
7	2013_Mohammad_Implementation-and-validation-of-an-artificial-neural-network-for-predicting-the-perf

- Clustering on documents shows very clean results
  - Same authors wrote multiple articles and their work grouped together
  - General research areas also clustered

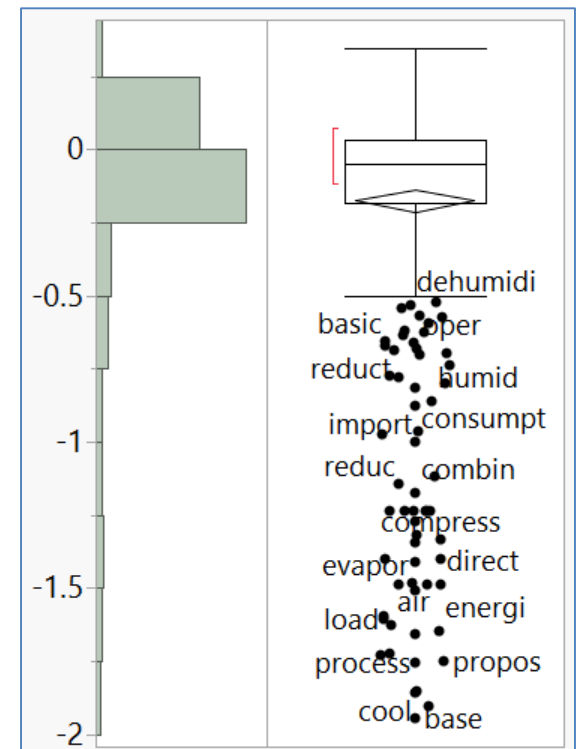
# Abstracts from 45 Journal Articles



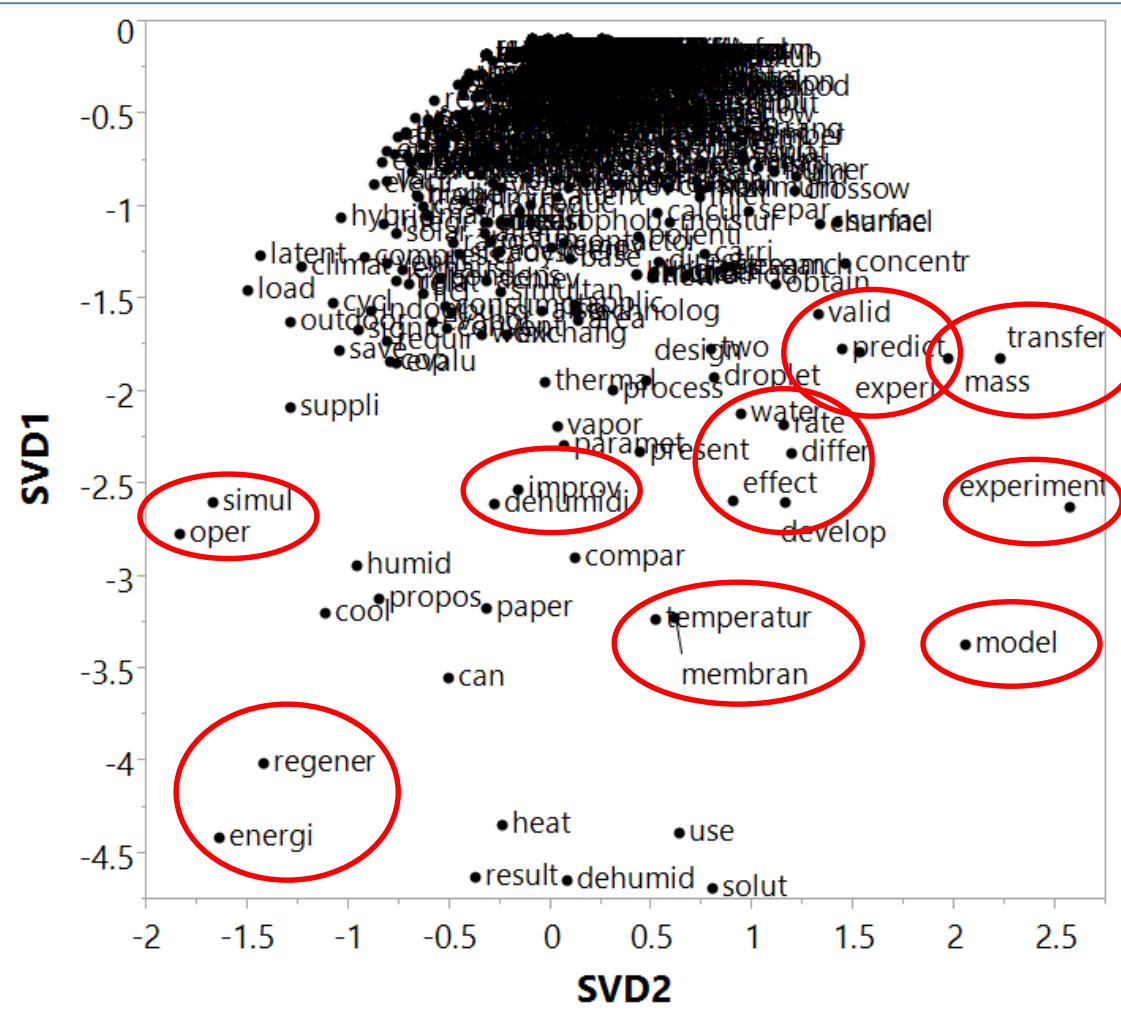
Comparative experiments validating liquid desiccant as A/C solution and increase in efficiency from regeneration method that saves energy



Experiment to predict rates/ratios; different inlet parameter values



Alternative method to remove vapor using hybrid electric compressor and liquid desiccant



- Major themes
  - Energy regeneration, improve dehumidification, simulation, mass transfer, experiment prediction, model, temperature and membrane, thermal process with water vapor



# Abstracts Word Associations

	Name		Name		Name		Name
1	cost	1	reliabl	1	lithium	1	droplet
2	payback	2	produc	2	chlorid	2	parametr
3	period	3	doascc	3	although	3	carryov
4	environment	4	simpli	4	aqueous	4	carri
5	main	5	construct	5	concern	5	one
6	instal	6	multizon	6	contact	6	inuenc
7	limit	7	variabl	7	identi	7	elimin
8	serious	8	rect	8	ambient	8	econom
9	follow	9	search	9	provid	9	directcontact
10	boiler	10	incorpor	10	input	10	trnsys
11	smldac	11	ceil	11	interact	11	life
12	storag	12	proper	12	micropor	12	annual
		13	serv	13	five	13	analysi
		14	aim	14	outlet	14	type
		15	airchil	15	major	15	great
		16	subsystem	16	characterist	16	sensit
				17	porous		

- Top word is word of interest (you can choose any of the thousands in the documents)
- Next ones are in order the “closest” based on all documents
  - Cost—concern is payback period, main installation, boiler, and storage big drivers
  - Reliability—producing multizone and ceiling units with airchilling subsystem
  - Lithium-dessicant is lithium chloride as aqueous solution; major concern is contact with ambient environment (toxic), microporous membrane is solution
  - Droplets—coming in direct contact are harmful, need to eliminate to make economically feasible



# Summary

- Competitor intelligence is essential across the organization and fueled by unstructured data
- Like military intelligence, there is an abundance of relevant open source information (e.g. journal articles, competitor websites, Twitter) but when you can put it together in meaningful ways it transitions to “classified information”
- JMP coupled with text analytics drives discovery of actionable intelligence to influence strategic decisions