



## Not Your Father's Forensics: Concept Searching for Data Forensic Investigations: Uncover what keywords miss

By

Warren G. Kruse II and Robert "Bobby" Kruse

*From the proceedings of*


The Digital Forensic Research Conference

**DFRWS 2019 USA**

Portland, OR (July 15th - 19th)

DFRWS is dedicated to the sharing of knowledge and ideas about digital forensics research. Ever since it organized the first open workshop devoted to digital forensics in 2001, DFRWS continues to bring academics and practitioners together in an informal environment. As a non-profit, volunteer organization, DFRWS sponsors technical working groups, annual conferences and challenges to help drive the direction of research and development.

<https://dfrws.org>



# Not Your Father's Forensics: Concept Searching for Data Forensic Investigations: Uncover what keywords miss

Textual Analytics  
Predictive Coding for Forensics Overview



Warren G. Kruse II

warren.kruse@Consilio.com

**Cyber Investigator/eDiscovery  
Consilio LLC**

Robert “Bobby” Kruse

**InfoSec Investigator/eDiscovery Analyst  
Internally for a Corporation**

# Analytics: Meaning Within Textual Context

---



“COCKTAILS”

## Goals of This Session

---

- Explore the foundations of analytics for investigations
- Explore the foundations of analytics
- Understand search methodologies and technologies
- Understand workflows - leveraging concept search
- Discuss integrating conceptual analytics into investigations

# Keyword vs. Concept

Extra line breaks in this message were removed.

From: Perko, Alicia  
To: Companywide  
Cc:  
Subject: Fwd: New Announcement for 11720 Sunrise Valley Drive

A new announcement has been posted for 11720 Sunrise Valley Drive:  
> -----  
> Thanksgiving Holiday  
>  
> Please be advised that the building and JBG's Reston Management Office  
> will be closed Thursday, November 26, 2009, in observance of the  
> Federal Holiday. The perimeter doors will be secured and no services  
> will be provided. Please remind your employees they will need an  
> activated Datawatch card to gain entry into the building.  
>  
> If you should require Overtime HVAC on Thursday, November 26, 2009,  
> please submit a Workspeed request no later than 12:00pm on Wednesday,  
> November 25, 2009.  
>  
> Additionally, the Management Office will close at 2:00pm on Wednesday,  
> November 25, 2009 and will be closed on Friday, November 27, 2009.  
> However, the building will remain open and services will be provided  
> on Friday, November 27th. Please continue to submit all building  
> requests via Workspeed.  
>  
> Should you have a building emergency that requires immediate attention  
> during the holiday, please call our after hours service line at (703)  
> 385-4140.  
>  
> Thank you and have a safe holiday!  
>  
>  
> Nov 23, 2009 4:59p  
> -----  
>  
> Please do not reply to this message. If you have questions regarding  
> this announcement, please contact your Property Management staff.  
>

Management

Clear

Search

6 A. These other kind of heart diseases like  
7 valvular, congenital, and cardiomyopathy can cause  
8 different kinds of demises and may not even be fatal  
9 if you die of something first; but they are far less  
10 common than atherosclerotic heart disease in our  
11 society.  
12 But to my knowledge, those have not been  
13 linked with smoking, which is the sense of your  
14 question.  
15 MS. ROSENBLATT: Page 72, line 24.  
16 Q. Does everyone who is exposed to  
17 environmental tobacco smoke develop heart disease?  
18 A. Develop clinical manifestations of heart  
19 disease?  
20 Q. Well, let's say an atheroma-induced event.  
21 A. Oh, no.  
22 MS. ROSENBLATT: Line 17.  
23 Q. If someone has atheroma bad enough to hurt  
24 you, was it necessarily caused by ETS?  
25 A. No.

Secondhand smoking is  
detrimental to your health and  
can cause cardiovascular  
disease

Clear

Search

# Value of Concept Search

---

- Avoids term mismatch issues
  - Polysemy, synonymy, group-specific jargon
- Avoids intentional obfuscation of language
  - Intentional disuse of a term
- Focuses on relevancy
  - Accounts for context of terms within documents
- Enables more natural querying
  - Allows longer queries that naturally express a concept
- Shows “truer” correlations between terms
  - Words can mean something entirely different than their denotation in a certain context

**Context is  
everything!**

# How eDiscovery Can Help Investigations

## Predictive Coding Overview

---

- Similar to Netflix and Pandora, where a machine learning application “learns” what movies or songs you like based on your selections





## Conversations (SHORT) Not Messages

---

- Start Time: 11/11/2015 4:59:41 PM  
End Time: 11/17/2015 5:40:22 PM  
Source: SMS/MMS  
Participants: Mahan Hank <9155045748>, (915) 487-0649 <9154870649>  
X-Riskcovery: EOH

(11/11/2015 4:59:41 PM) Mahan Hank <9155045748>: Did you get Altep squared away? I was playing poker with their VP of Legal and mentioned something about revising their dashboard requirements.

(11/11/2015 9:58:37 PM) Mahan Hank <9155045748>: Did this get fixed?

(11/12/2015 3:52:43 PM) (915) 487-0649 <9154870649>: Yes, sir. I made sure our team took care of it.

(11/17/2015 5:40:22 PM) Mahan Hank <9155045748>: Security upgrade

# TAR “Nuts & Bolts”: Latent Semantic Indexing (LSI)

- Dimensionality
- Creates a conceptual index of by grouping documents that are conceptually similar
- Example of a tool that uses this is Relativity Analytics Review (RAR)

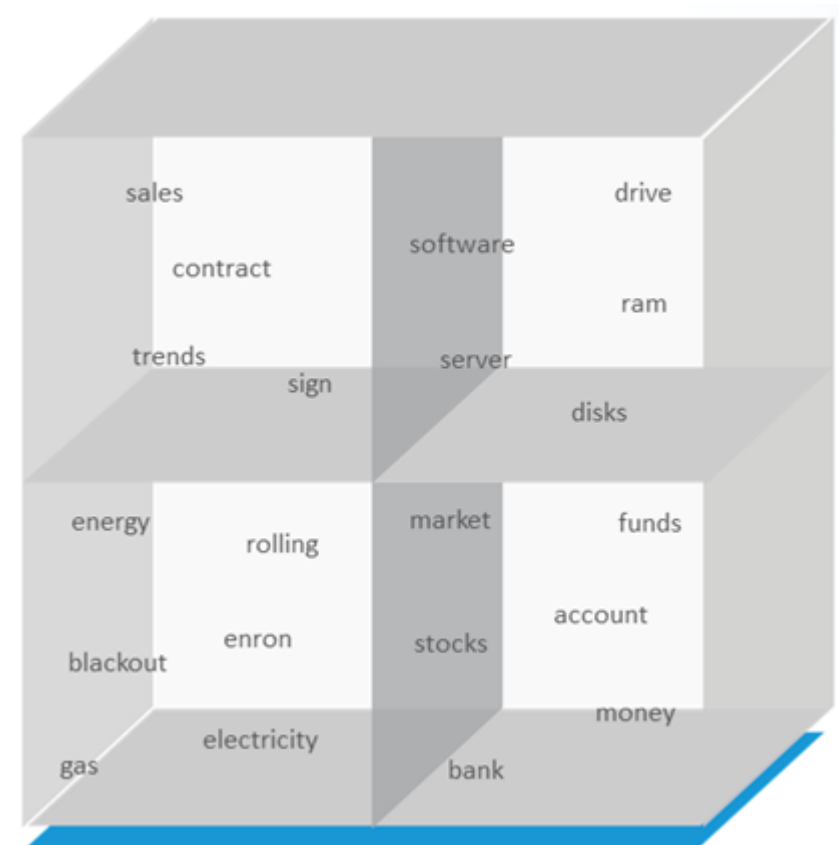


Image courtesy of kCura Corp.

## Dimensionality: An Example

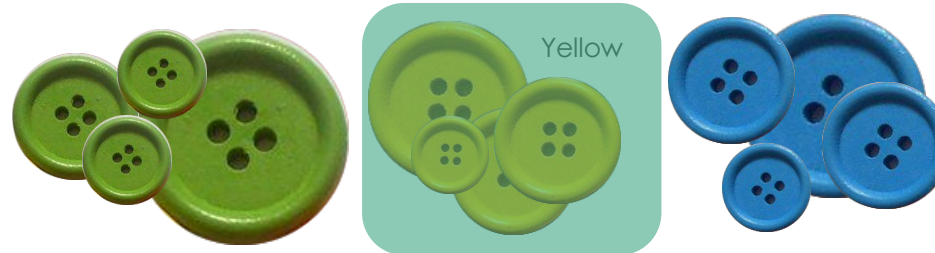
---

- Suppose you are sorting a pile of buttons.
- They are of various colors and sizes



# Dimensionality: An Example

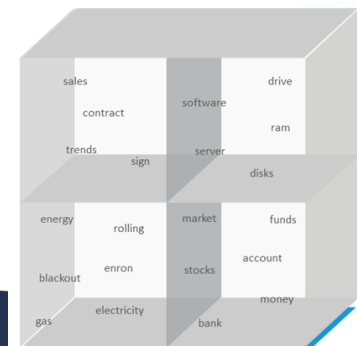
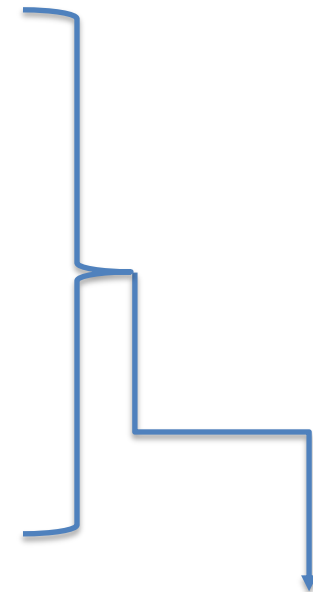
- Color



- Size



- Color and Size



# Understanding Concept Search

---

- Term Mismatch
- Term vs. Concept
- Concept Search vs. Keyword Search

*“I really fell for her”*

*“They fell head over heels”*

*“She only has eyes for him”*

*“He fell for her like his heart was a mob informant, and she was the East River”*

## Misc. Notes

---

- Bias Issue – Because training stems from an initial set of responsive documents, how do we know that we are covering all potentially responsive concepts/issues?
  - allowing Brainspace to select training documents may help

# Know What You Don't Know (clusters)

- image, aapl, width, rss, height, picture, microsoft (2,413)
    - cup, sugar, minutes, salt, flour, water, mix (354)
    - microsoft, rss, wt.rss\_a, wt.rss\_ev, xbox, pro, save (324)
    - aapl, apple, google, irrelevant, flag, news, alert (321)
    - image, picture, pillar, width, height, apod, nasa (287)
    - wine, â, page, email, mailto, policy, privacy (219)
    - construction, plans, project, building, discount, excluded, prior (170)
        - construction, building, project, cost, design, materials, contractor (80)
        - plans, discount, excluded, select, gift, prior, zalto (72)
        - network, local, smes, businesses, support, employment, micro (18)
      - email, wine, bottle, promotions, gift, click, address (168)
      - share, gmt, border, image, shark, grammy, br (117)
      - kayla, mueller, gmt, dominique, clear, height, strauss-kahn (73)
      - website, explorer, tech, internet, click, pc, sling (60)
      - tasting, table, store, altep.com, tlatulippe, free, message (60)
      - image, texas, hasbro, u.s, getty, police, ap (45)
      - u.s, appointment, citizen, consulate, general, emergency, consular (25)
      - options, irrelevant, flag, google, aapl, facebook, id (22)
      - width, http, td, height, tr, www.kunde.com, align (21)
      - electrical, log, cybercoders, plans, true, manager, full-time (12)
      - UNCLUSTERED (135)

# Tool Examples

---



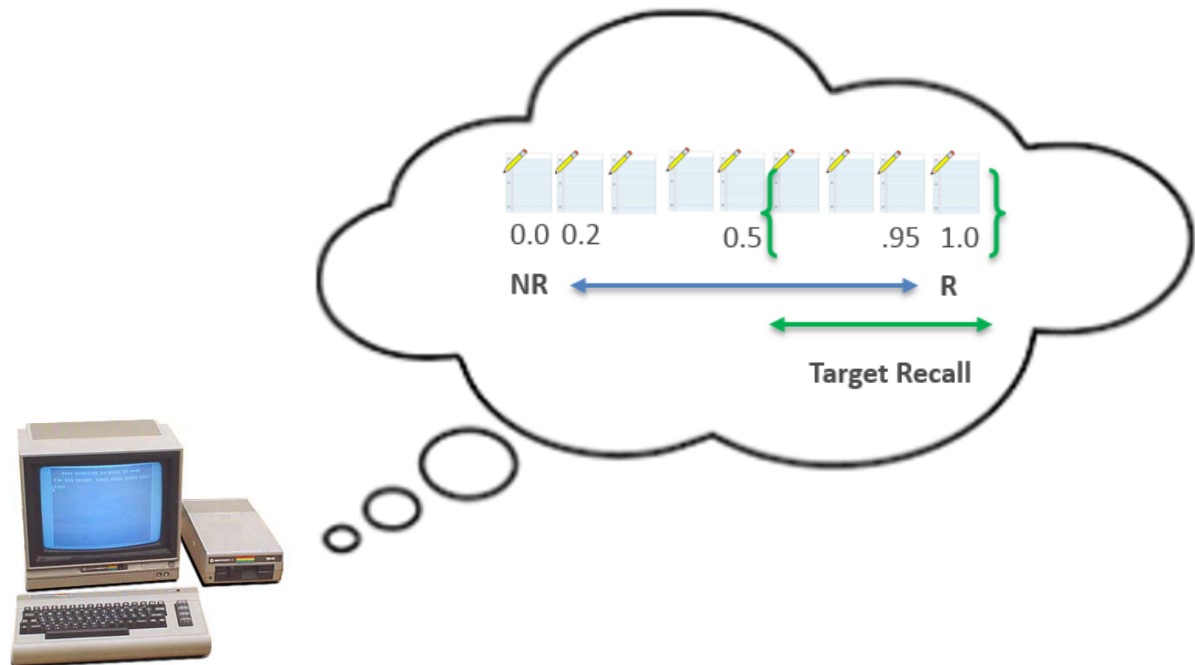
# TAR “Nuts & Bolts”: Brainspace Logistic Regression

---

- Method of taking a large number of factors or variables and generating a single “yes/no” answer
- Used in medical research to determine the probability of disease  
 +**smoker** | +parent had heart disease | -exercise daily | -**good diet**  
 = 50% probability of heart disease
- Brainspace analyzes a document’s features, including word occurrences frequencies, relationships, etc... and develops an “equation” (S-curve) that calculates the probability that a document is responsive  
 +**competition** | +market analysis | -international | -**france**  
 = 50% probability that docs is responsive
- Once the equation is fully developed, Brainspace analyzes all documents in the PC population by applying the equation to each document and assigning a probability score.

# TAR “Nuts & Bolts”: Brainspace Responsive Probability Scoring

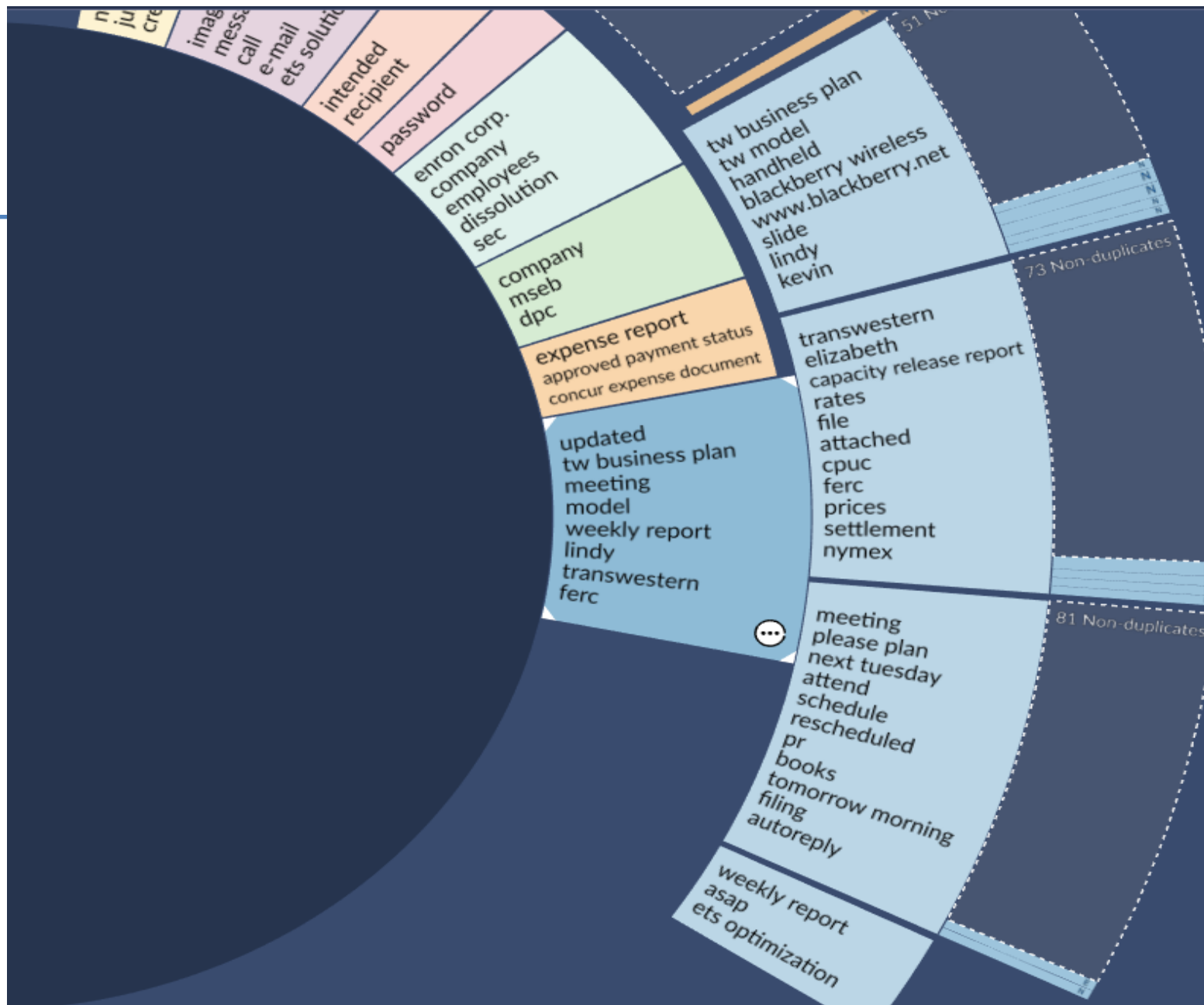
- Brainspace assigns scores to documents, which represent the probability that a document is responsive – 1.0 is highly responsive and 0.0 highly non-responsive
- Scores can be used to prioritize review or for defining a review cut-off point.

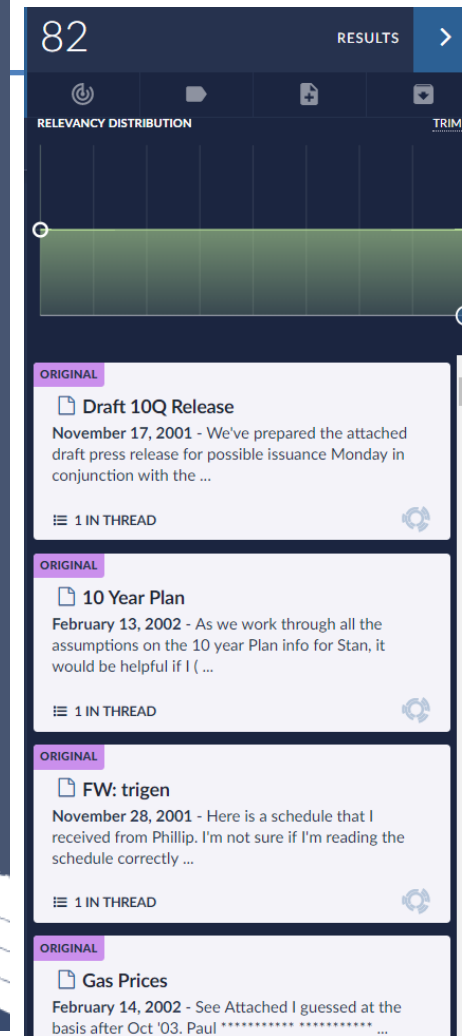


- “Enron Corp., Houston, Tuesday said its interstate natural gas pipelines will begin offering capacity on EnronOnline. Transwestern Pipeline Co. will be the first of Enron's pipelines to offer capacity for bidding through an open season conducted on EnronOnline, Enron's principal-based electronic transaction platform”

The diagram is a circular sunburst chart with a dark blue center. The first ring consists of 12 colored segments. The second ring contains text labels of varying widths, some of which are grouped within a single segment. A red circle highlights a specific segment in the first ring and its corresponding segment in the second ring.

Segment Color	Segment 1 (Inner Ring)	Segment 2 (Outer Ring)
Dark Blue	server	et solution center
Purple	updated	game
Pink	image	weekly customer
Light Purple	message points	book request
Yellow	image	conference call
Orange	intended	message points
Light Orange	password	93 News duplicate
Light Green	enron corp.	vacation
Light Green	company employees	employees
Light Green	employees houston	partnerships
Light Green	company mseb	ljm
Light Green	expense report	german
Light Green	company mseb	accounting
Light Green	power merc	enron europe
Light Green	health schlesser	operations
Light Green	enron north america	travel department
Light Green	company mseb dpc	energy
Light Green	tw business plan	corporate services
Light Green	handheld tw model	enron corp.
Light Green	tw business plan	blackberry website
Light Green	file	transwest
Light Green	elizabeth	enron corp.
Light Green	schedule attend	meeting
Light Green	next tuesday	please plan
Light Green	weekly report	meeting
Light Green	newspaper clippings	93 News duplicate
Light Green	june	weekly report
Light Green	lindy	model
Light Green	meeting	updated
Light Green	tw business plan	expense report
Light Green	expense report	enron corp.
Light Green	company mseb	enron north america
Light Green	power merc	company mseb dpc
Light Green	company mseb	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green	enron corp.	enron europe
Light Green	enron europe	operations
Light Green	operations	travel department
Light Green	travel department	energy
Light Green	energy	corporate services
Light Green	corporate services	enron corp.
Light Green		





ORIGINAL



## 10 Year Plan

**Subject:** 10 Year Plan

**From:** Harris Steven <steven.harris@enron.com>

Date: Wed, 13 Feb 2002 22:32:26 +0000

**To:** Hayslett Rod <rod.hayslett@enron.com>,  
Howard Kevin A. <kevin.howard@enron.com>

As we work through all the assumptions on the 10 year Plan info for Stan, it would be helpful if I (or a member of my staff) could be involved in the specific discussions as to what the assumptions are that are being considered. Although we have been involved in some of those meetings it appears that significant changes have been made at the last minute without ever being discussed with the Commercial Group. If I am to be responsible for answering questions as to the details then I would like to provide input prior to any future meetings with Stan on these issues.

There are several areas that I would like to sit down with you to discuss:

-First, I would like to see the forward curves used to price capacity from '06 going forward;

-Once we have the forward curves from now to '06, I would like to see those also (if we thought those prices didn't look good then how do we justify assuming resubscription at Max rates going forward?) Is Max for both East and West capacity at the same time a good assumption? The market has never allowed that kind of deal structure;

-What expansion scenarios are we assuming? I understand that they were changed sometime today yet I was not told what is in the current plan;

I would like to be able to put together the most beneficial and realistic plan for TW going forward as we all have the same interests in mind. Thanks.

Steve

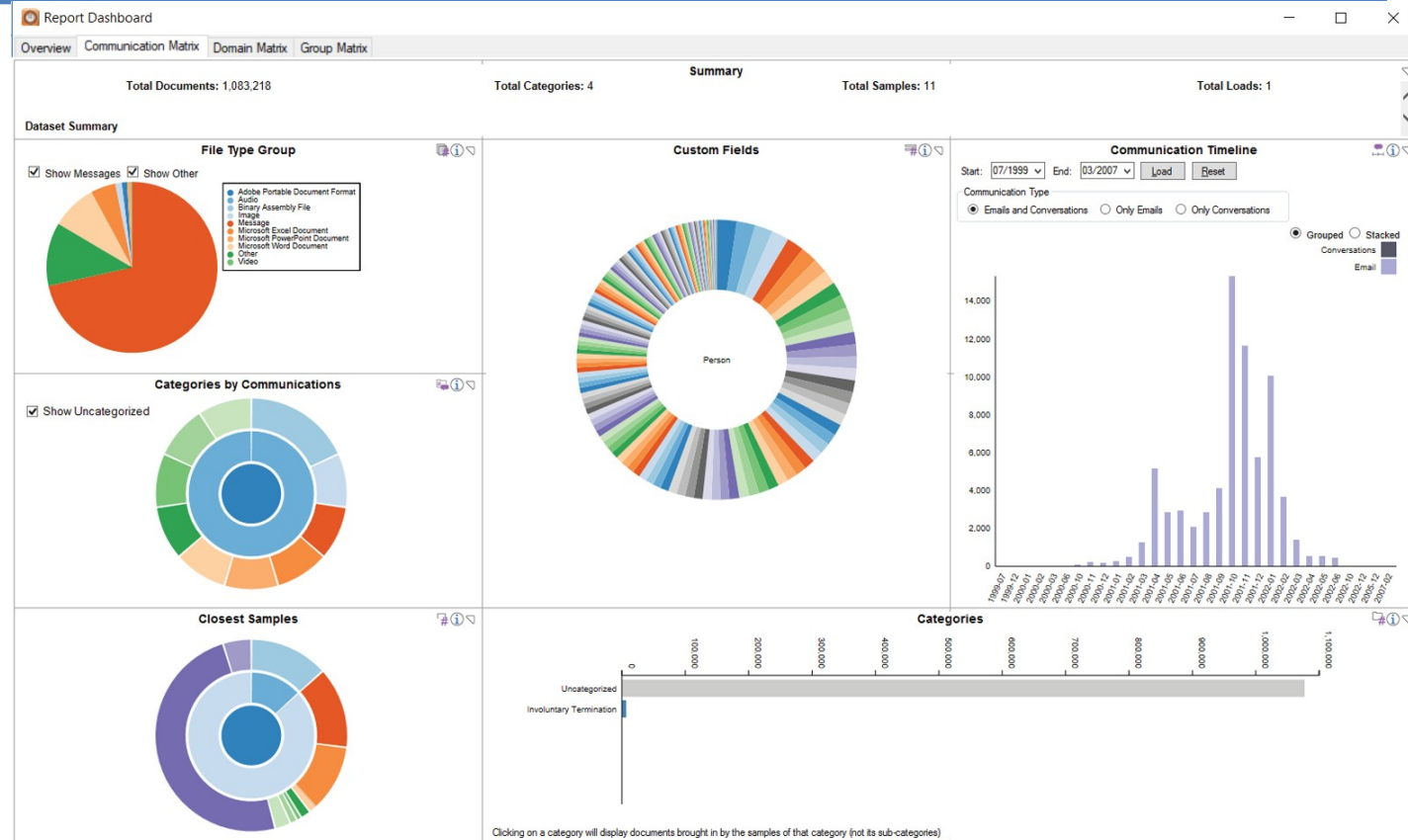
Add to Notebook

More Like This

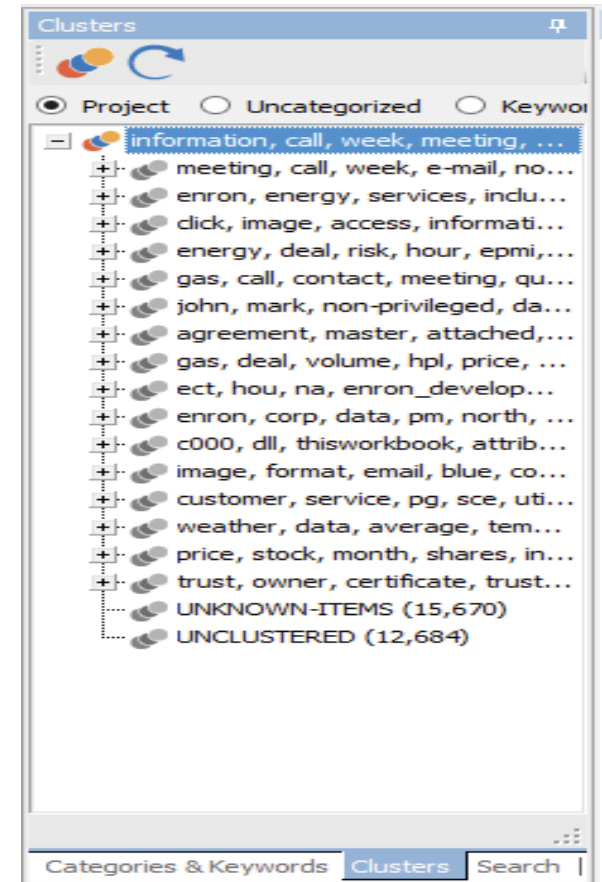
View Source Document

# Riskcovery

- Portable

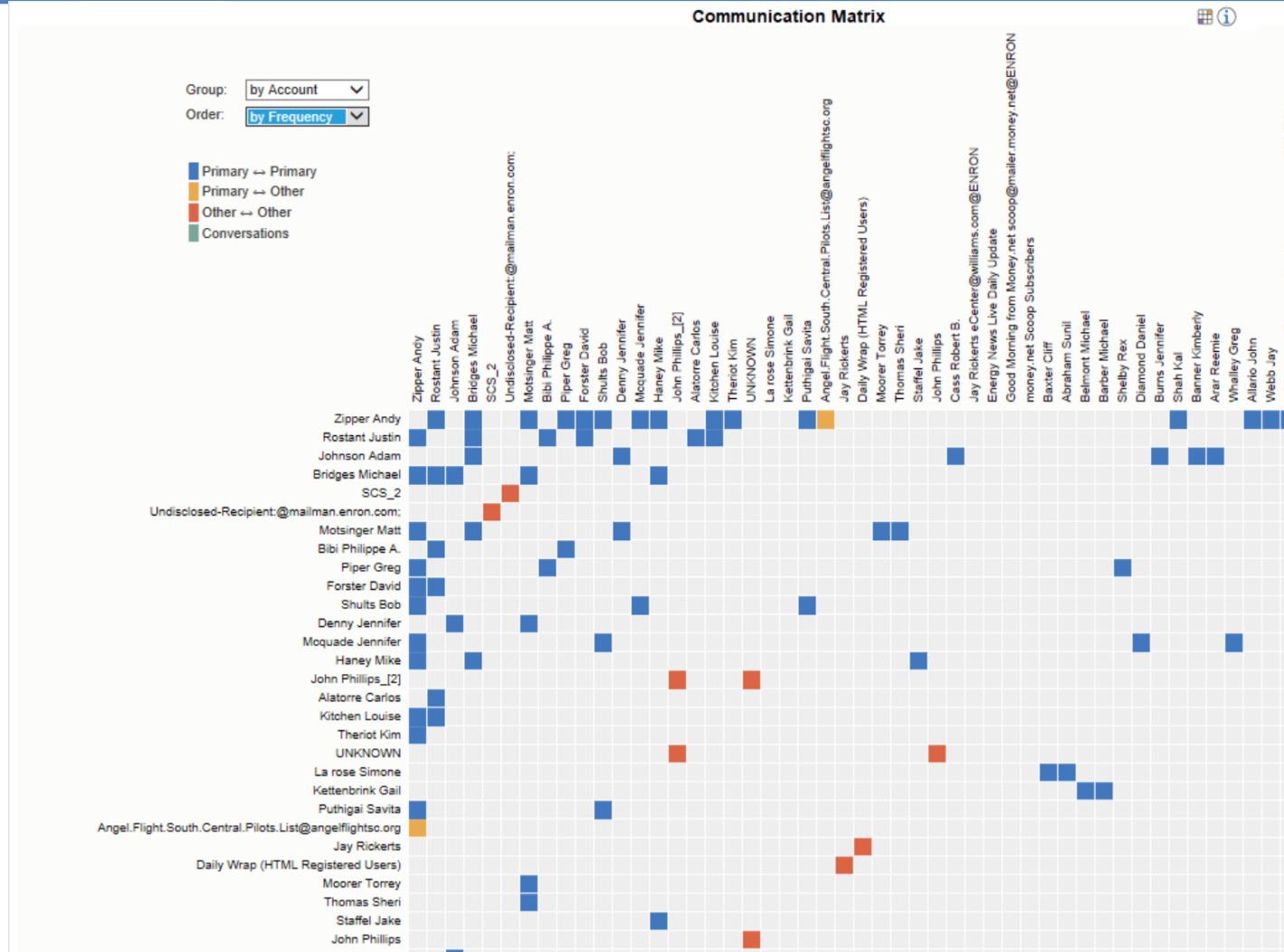


- Riskcovery automatically ascertains patterns of meaning in the ingested data and groups “like” items into clusters. You can browse these clusters to get a quick sense of the nature and content of the data population. From the Clusters pane





# Communication Heat Map



**Riskcovery**

Home Taxonomies View Reports Utilities

New Project Open Project Close Project Load Documents Export Documents Rebuild Conversations Import Back Up Restore Restore Points Logs Edit Project Settings Account Manager Delete Marked Documents Delete Project View Documents Marked for Deletion Configuration License About Release Notes Exit

Open/New Projects Documents Backup... Restore Point... Project Settings Acc. Mana... Delete Documents Project Configura... Help Exit

**1** Categories & Keywords

Uncategorized (1,083,244)  
Keywords (0)  
Categorized Documents (0)  
Uncategorized Documents (0)

**6** Document - 2097220.msg

folder: \ExMerge - Meyers, Albert\Sent Items  
date: Wed, 30 Jan 2002 03:48:57 -0800 (PST)  
X-Riskcovery: EOH

<pre>Bill:

Please note the following due to the past two days schedules have been wrong in the EPE Schedules in Excel:

Tag number 6181 has been cancelled (50mw to the CISO).

Lending is wrong in the EPE schedules (it is 75mw from PSCO instead of 50mw).

SPS is wrong for HE 08 (it is 130mw instead of 100mw).

I thought you might like to since this is the only income we have currently for real-time and a major screw-up could hurt our relationship.

**2** Regards.

**3** Document Categories

☒ Category Score Closest Sample  
☒ Uncategorized 100

**4** Document Keywords

☐ Identifier Keyword Expression

**5** Family

2097220.msg

**7** Category In (Uncategorized)

Drag a column header here to group by that column

| Result | Document ...  | Categories    | Bookmark | Forced | Score | Original Path  | Cached | Logical Date         | FamilyID |
|--------|---------------|---------------|----------|--------|-------|--|--------|----------------------|----------|
|        | andrew_je...  | Uncategorized |          |        | 100   | E:\_ENRON_V1_AH\andrew_jewis_000.pst   |        | 5/10/2009 6:05:22 PM | 3        |
|        | andy_zippe... | Uncategorized |          |        | 100   | E:\_ENRON_V1_AH\andy_zipper_000.pst  |        | 5/12/2009 5:15:38 AM | 4        |
|        | andy_zippe... | Uncategorized |          |        | 100   | E:\_ENRON_V1_AH\andy_zipper_001.pst  |        | 5/12/2009 5:34:32 AM | 5        |
|        | barry_tych... | Uncategorized |          |        | 100   | E:\_ENRON_V1_AH\barry_tycholz_000.pst  |        | 5/10/2009 6:39:58 PM | 6        |
|        | 2097188.msg   | Uncategorized |          |        | 100   | E:\_ENRON_V1_AH\albert_meyers_000.pst\Top of Personal Folders\meyers-a\ExMerge - Meyers, Albert\Sent Items\2097188.msg |        | 1/1/1601 12:00:00 AM | 7        |
|        | 2097220.msg   | Uncategorized |          |        | 100   | E:\_ENRON_V1_AH\albert_meyers_000.pst\Top of Personal Folders\meyers-a\ExMerge - Meyers, Albert\Sent Items\2097220.msg |        | 1/1/1601 12:00:00 AM | 8        |
|        | 2097252.msg   | Uncategorized |          |        | 100   | E:\_ENRON_V1_AH\albert_meyers_000.pst\Top of Personal Folders\meyers-a\ExMerge - Meyers, Albert\Sent Items\2097252.msg |        | 1/1/1601 12:00:00 AM | 9        |

1,083,244 document(s), 1,083,244 shown

**8** Category In (Uncategorized) Activity

Analytics Index Online

Licensed to: Alteq, Inc. License Expires on: Wednesday, Oct 4, 2017 (Required check in by 8/16/2017) Project FullEnronv1

# Resources

---

- [The Rand Corporation Safety and Justice Program](http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf)  
([http://www.rand.org/content/dam/rand/pubs/research\\_reports/RR200/RR233/RAND\\_RR233.pdf](http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf))
- Law [Enforcement](http://www.informationbuilders.com/solutions/gov-lea) Analytics: Intelligence-Led and Predictive Coding  
(<http://www.informationbuilders.com/solutions/gov-lea>)
- Visual Analytics Law Enforcement Toolkit: Helping Law Enforcement Stay Ahead of Crime  
(<http://www.dhs.gov/sites/default/files/publications/Visual%20Analytics%20Law%20Enforcement%20Toolkit-Helping%20Law%20Enforcement%20Stay%20Ahead%20of%20Crime-CVADA-VALET-Nov2013.pdf>)
- Connecting the Dots: Data Mining and Predictive Analytics in Law Enforcement and Intelligence Analysis:  
([http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article\\_id=121&issue\\_id=102003](http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=121&issue_id=102003))

Questions?

