Advanced Analytics: Technology Is Your Friend

Allegheny County Bar Association

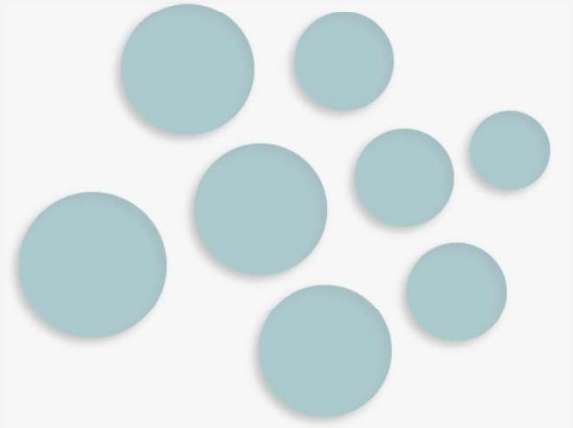
Bench-Bar Conference (June 17, 2022)

bit-x-bit SM

THE EVIDENCE. THE BACKSTORY.
THE INTELLIGENCE. **WE GET IT.** SM

Agenda

- Introduction – What Is Analytics?
- Variations: Conceptual & Structural Analytics
- Finding Key Information: Review Platforms
- Use Cases: How Does It Work?
- What Do Judges Think About This Technology?



What Is Analytics?

What Is Analytics?

- The process of identifying patterns in data.
- Useful in discovery to reduce data to relevant information.
- Technology assisted review saves time and money.
- Accuracy is greater than that of standard linear human review.

Conceptual vs. Structured Analytics

Conceptual

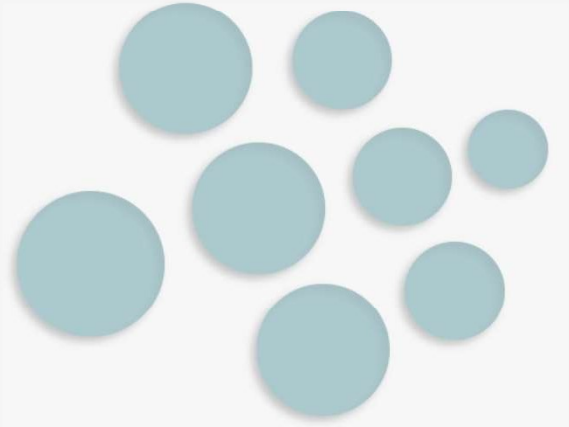
- Clustering
- Concept Searching
- Visualization

Structured

- Email threading
- Name normalization
- Textual near duplicate
- Language detection
- Automated redactions
- Active learning review

Conceptual vs. Structured – Understanding Use

- Not a case of one or the other, both have uses at various stages in the discovery process.
- Conceptual helps bring together data based on content and context.
- Structured is a culling and streamlining of the data.

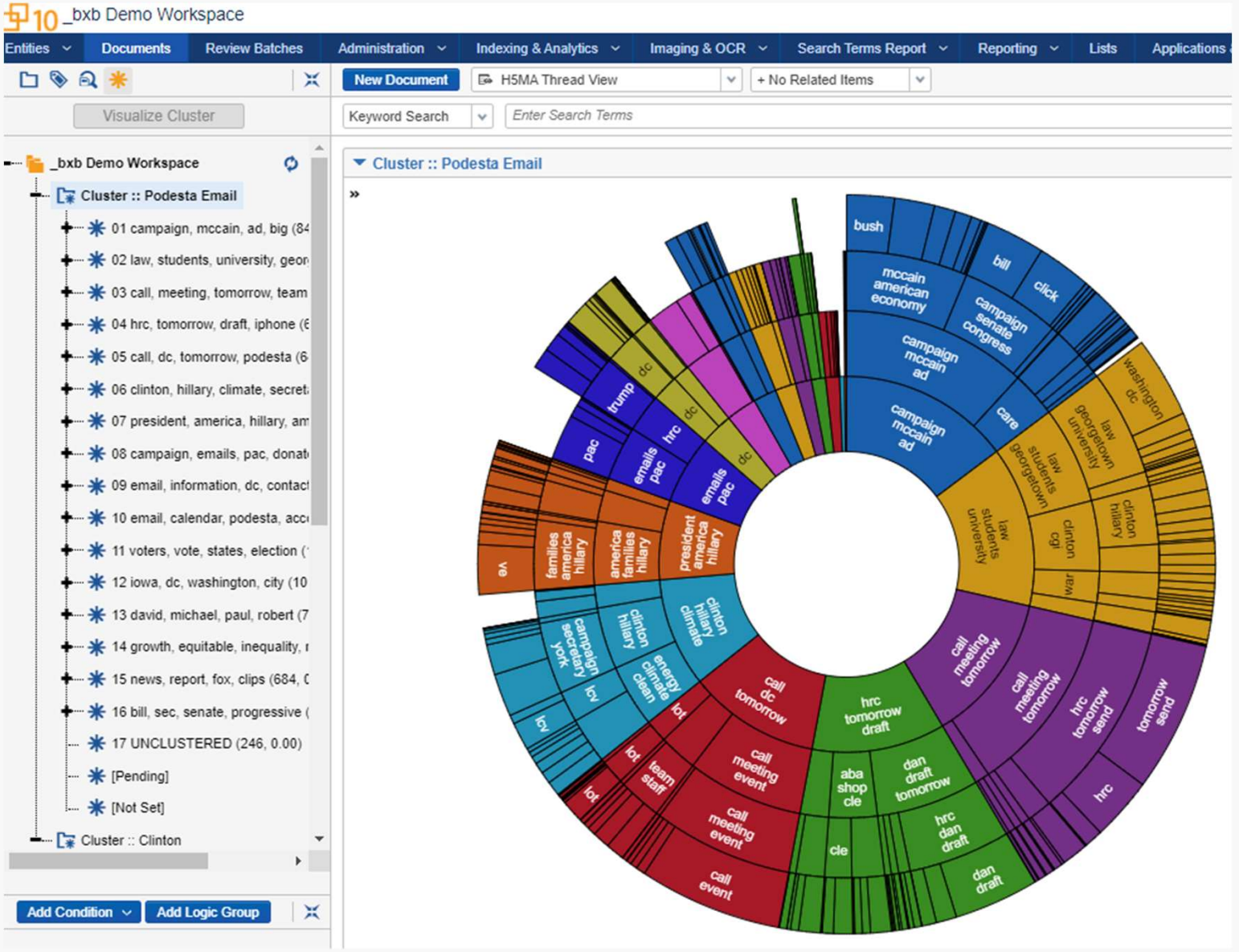


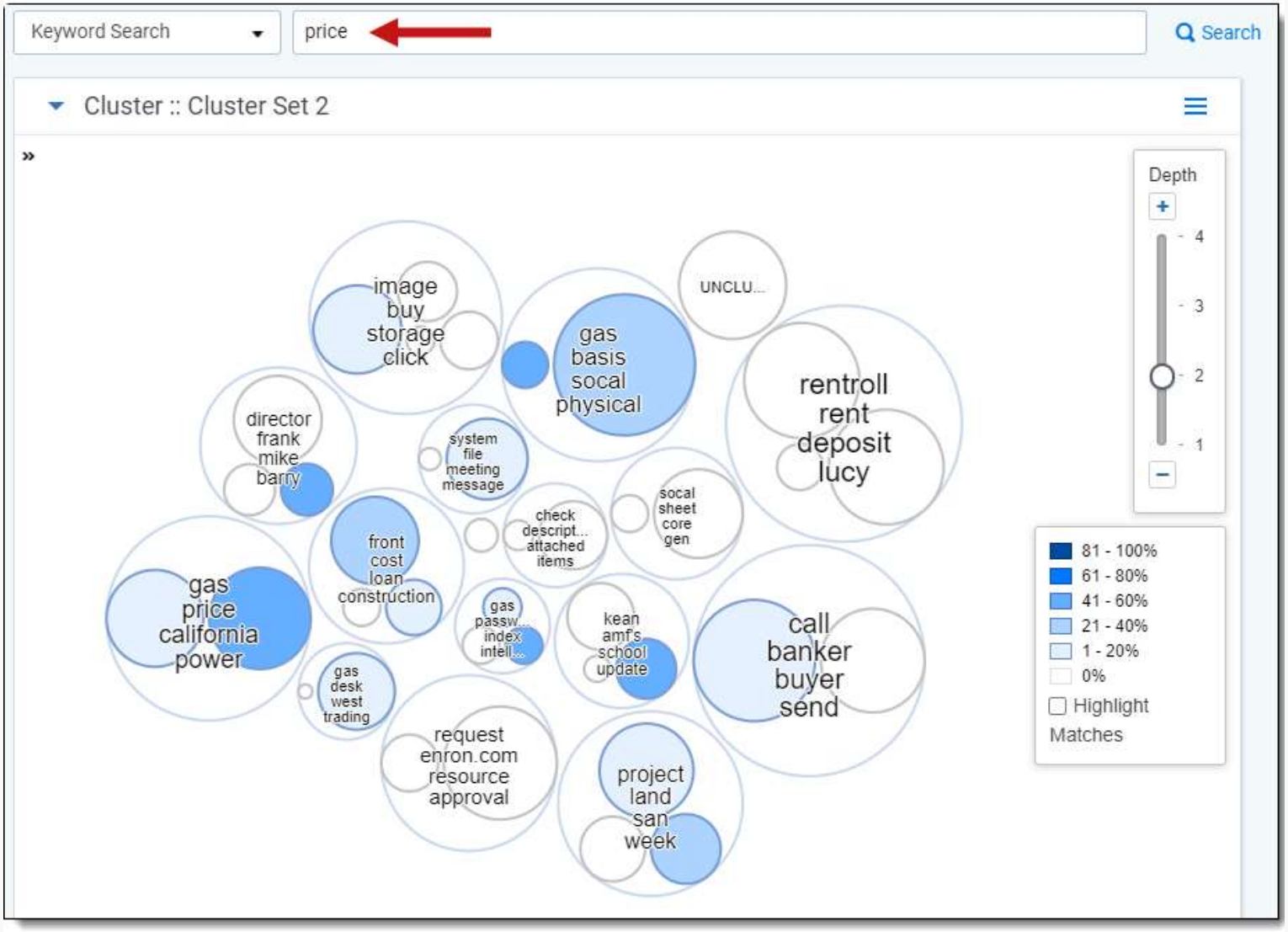
Conceptual Analytics

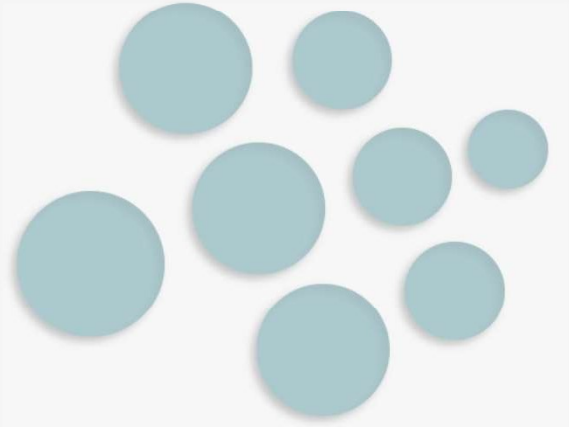
Conceptual Search

Standard Method	Analytics Method
Finds the presence (or absence) of a query (term or block of text)	Derives the potential meaning of a query
Simply looks for matches of query and indexed docs	Attempts to understand semantic meaning and context of terms
Incorporates Boolean logic	Incorporates mathematics

- Concept searching forces the focus on conceptual relevancy rather than on any single or group of specific terms.
- Encourages the user to express a concept in the way that people are used to describing ideas and concepts. Concept searching can handle very long queries.


10 _bxb Demo Workspace
Entities
Documents
Review Batches
Administration
Indexing & Analytics
Imaging & OCR
Search Terms Report
Reporting
Lists
Applications
New Document
H5MA Thread View
+ No Related Items
Visualize Cluster
Keyword Search
Enter Search Terms
_bxb Demo Workspace
Cluster :: Podesta Email
01 campaign, mccain, ad, big (84
02 law, students, university, geor
03 call, meeting, tomorrow, team
04 hrc, tomorrow, draft, iphone (6
05 call, dc, tomorrow, podesta (6
06 clinton, hillary, climate, secret
07 president, america, hillary, am
08 campaign, emails, pac, donat
09 email, information, dc, contact
10 email, calendar, podesta, acc
11 voters, vote, states, election (
12 iowa, dc, washington, city (10
13 david, michael, paul, robert (7
14 growth, equitable, inequality,
15 news, report, fox, clips (684,
16 bill, sec, senate, progressive (
17 UNCLUSTERED (246, 0.00)
[Pending]
[Not Set]
Cluster :: Clinton
Add Condition
Add Logic Group
Cluster :: Podesta Email
bush
bill
click
mccain american economy
campaign senate congress
campaign mccain ad
care
washington dc
law georgetown university
law students georgetown
clinton hillary
law students university
clinton cgi
war
call meeting tomorrow
call meeting tomorrow
hrc tomorrow send
tomorrow send
hrc
hrc tomorrow draft
dan draft tomorrow
hrc dan draft
dan draft
aba shop cle
cle
call dc tomorrow
call meeting event
call meeting event
call event
lot
lot
team staff
clinton hillary climate
energy climate clean
clinton hillary
campaign secretary york
lcv
lcv
president america hillary
america families hillary
families america hillary
ve
emails pac
emails pac
hrc
trump
pac
dc
dc
dc

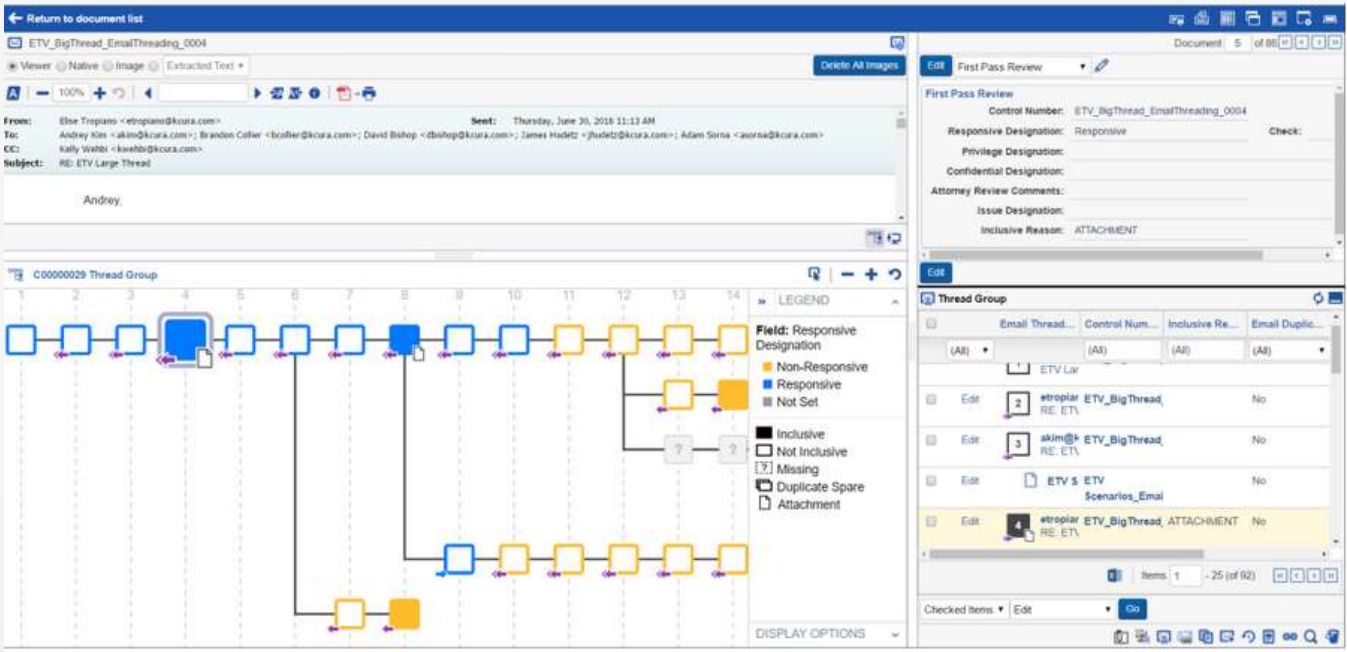

Keyword Search
price
Search
Cluster :: Cluster Set 2
Depth
4
3
2
1
81 - 100%
61 - 80%
41 - 60%
21 - 40%
1 - 20%
0%
Highlight
Matches
image buy storage click
UNCLU...
gas basis socal physical
rentroll rent deposit lucy
director frank mike barry
system file meeting message
socal sheet core gen
check descript... attached items
front cost loan construction
gas price california power
gas passw... index intell...
kean amf's school update
call banker buyer send
gas desk west trading
request enron.com resource approval
project land san week



Structured Analytics

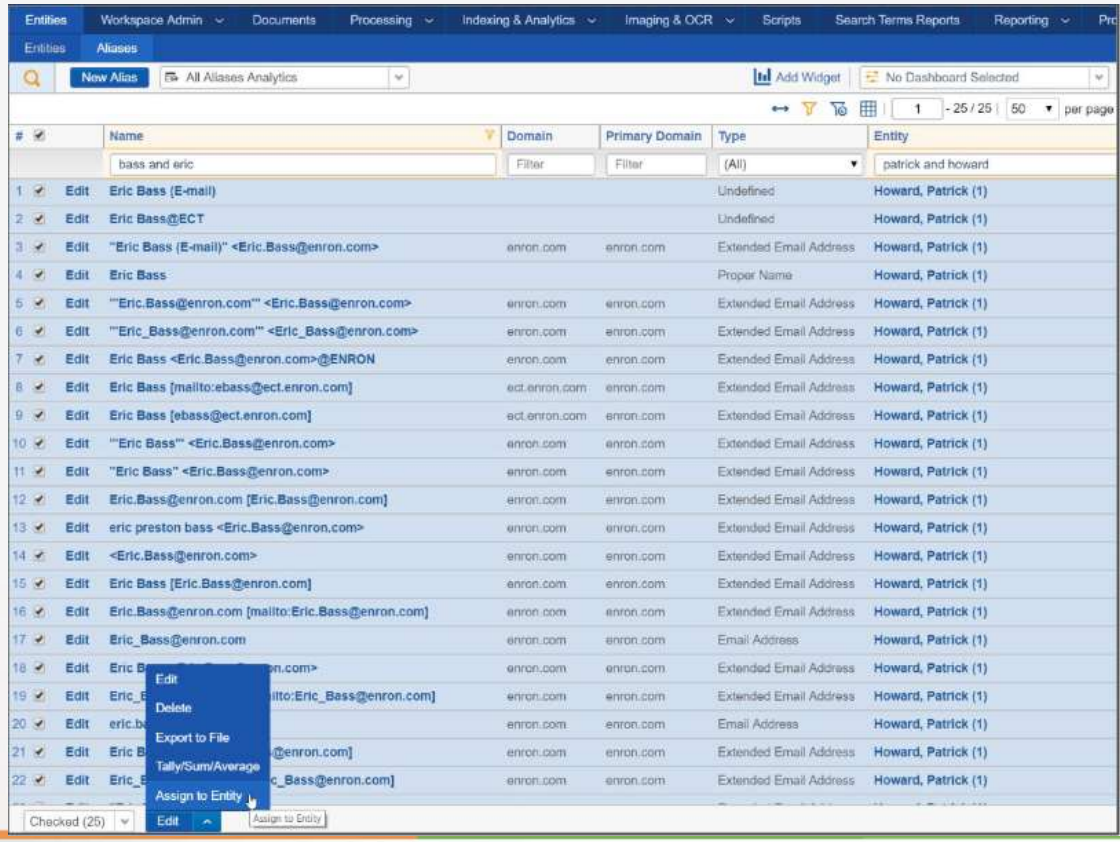
Email Threading

- Assists in reviewing the most inclusive email and ignoring the others (speed)
- Can also be used to visualize the conversation and find where branch conversations occurred
- Can assist in identifying missing emails in the thread



Name Normalization

- Analyzes email headers (both the top level and embedded) of the parent emails.
- Creates aliases that are then linked to entities (Doe, John)
- Review and recategorize the aliases, email addresses, etc., to group an entities' documents



Textual Near Duplicates

- Identifies documents that are near duplicates based on the text of the document
- Provides a percentage similarity that can be used to group documents for review or find (perhaps key) similar documents, i.e. theft/reincorporation of text in documents, contracts that are similar, etc.

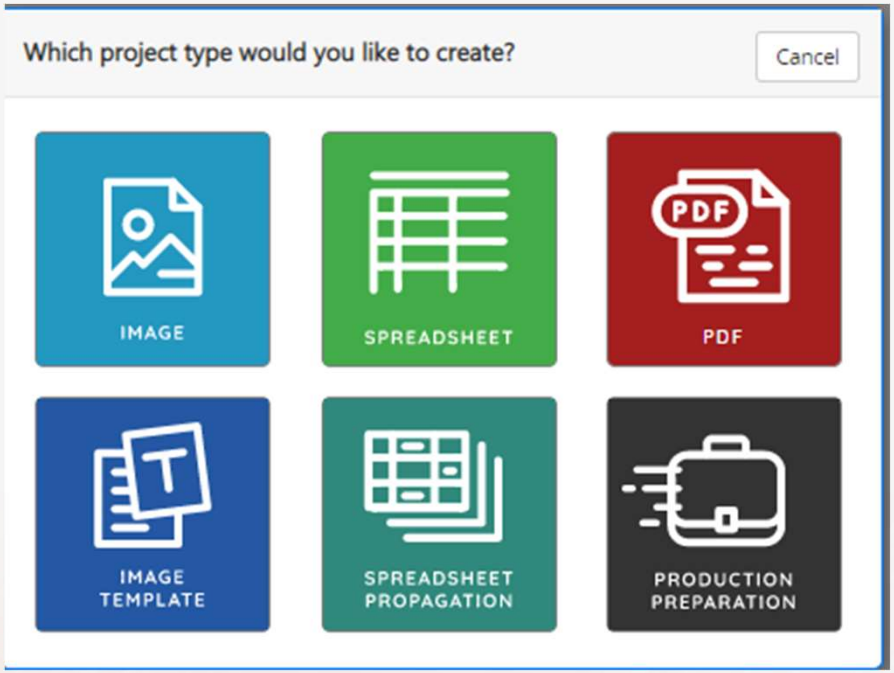
- The quick brown fox jumped over the lazy **d**og
- The quick brown fox jumped over the lazy **l**og

Language Identification

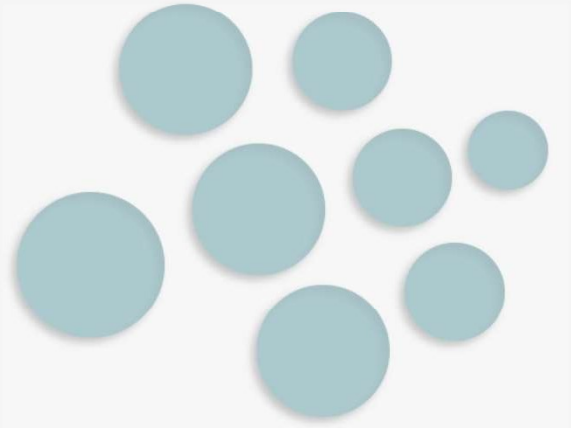
- Examines the extracted text from each document to determine the primary language and up to two secondary languages
- Works best on sets of at least 200 characters (about two sentences) per document. Covers 173 languages.

Automated Redactions

Blackout
Automated Redaction Tool



- Cut time and costs with automated redactions
- Redact images, PDFs, even native spreadsheets and other large files
- Rule-based redaction capability, i.e., PII
- Redact Information not visible in the file, i.e., metadata, file attachments, document notes/comments



Review Platforms

eDiscovery Technology

Relativity®

Blackout
Automated Redaction Tool

H5 | MA
MATTER ANALYTICS

Brainspace™

opentext™
Insight

opentext™ | Insight Predict

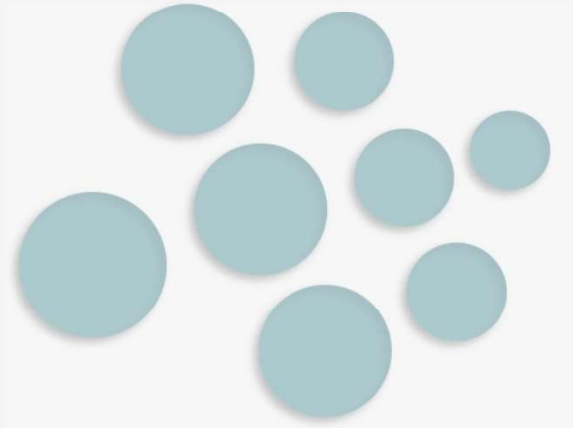
How Do I Find What I'm Looking For?

Review Platforms	
Full service/end-to-end support	**Self-service models**
• External experts help with ESI preservation, culling and upload • Attorney review overseen (technically) by third party: coding fields, structure of review, use of analytics • Your data set can be reduced before ESI is promoted to review platform	• No or limited external support • Attorney "drags and drops" a dataset into the processing/review tool • Attorney is responsible for review, quality control, and production

Review Methods

Review Methods: Linear vs. VAR vs. CAL

	Training Documents	Training Model	Document Categorization/ Prediction
Linear	N/A	None	**Manual review** and categorization of all documents (after filtering)
VAR	Coded by **SMEs**	**Static** - Based on SME's decisions on seed documents and must be **manually perfected** through several rounds of training	**Multiple rounds of training and prediction** are done. When prediction is acceptable, **then manual review** and categorization is done for the predicted documents.
CAL	Coded by **Reviewers**	**Dynamic** - Continuously updated and perfected based on Reviewer's coding decisions and content presented	**Automated training** is integrated and **occurs simultaneously** with manual review and categorization of documents.



Use Cases:
Technology Is Your Friend

Value of Technology Assisted Review

Matter: DOJ Subpoena Issued to Energy-Related Tech Company (involving oversees transactions)

Problem:

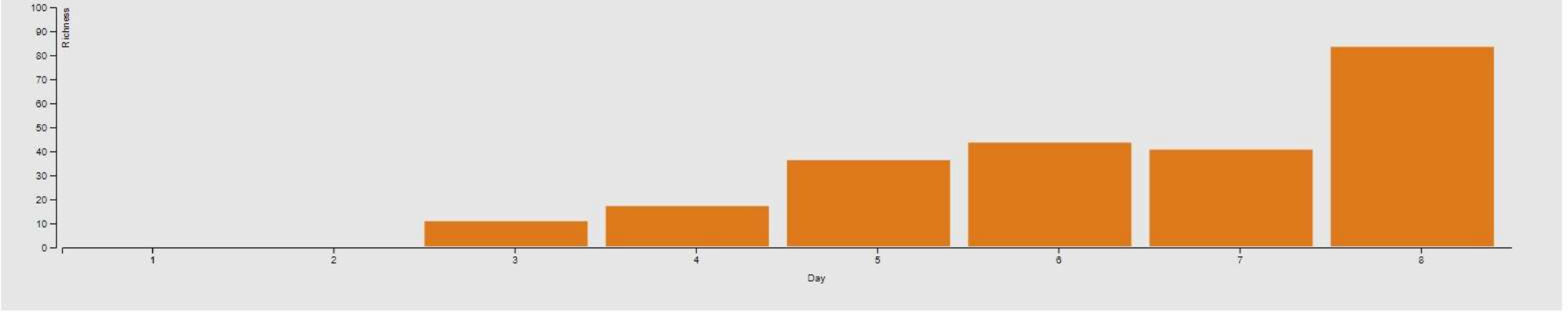
- ESI document collection of approximately 34,000 documents, resulting from key word searching.
- Only 10% of the documents were determined to be relevant and responsive (benchmarked by a random sample). Absent the use of analytics, the reviewers thus would have had to review ten documents to find one that was relevant.

Solution:

- "Continuous Active Learning" analytics noted the attorney reviewers' relevance decisions and then continuously identified more and more relevant documents.
- After reviewing only 550 documents, the percentage of relevant documents presented to the reviewers increased from 10.9% to 43.6%, and then to 83.5% after reviewing only 2,144 documents. (That's less than 1 banker's box, in old-school terms.)

Value of Technology Assisted Review

Day	1	2	3	4	5	6	7	8
Reviewed	0	0	64	209	295	808	307	461
Relevant	0	0	7	36	107	352	125	385
Richness (%)	0%	0%	10.9%	17.2%	36.3%	43.6%	40.7%	83.5%



Podesta emails – 2016 Hillary Clinton Presidential Campaign

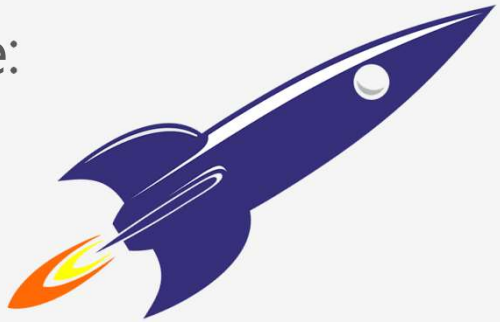
Corpus: 58,468 emails and documents

The Challenge: Locate Documents Relevant to 3 Issues:

1. Clinton's use of email server and email deletions
2. 2012 Benghazi Attack
3. Clinton's Speeches to Wall Street

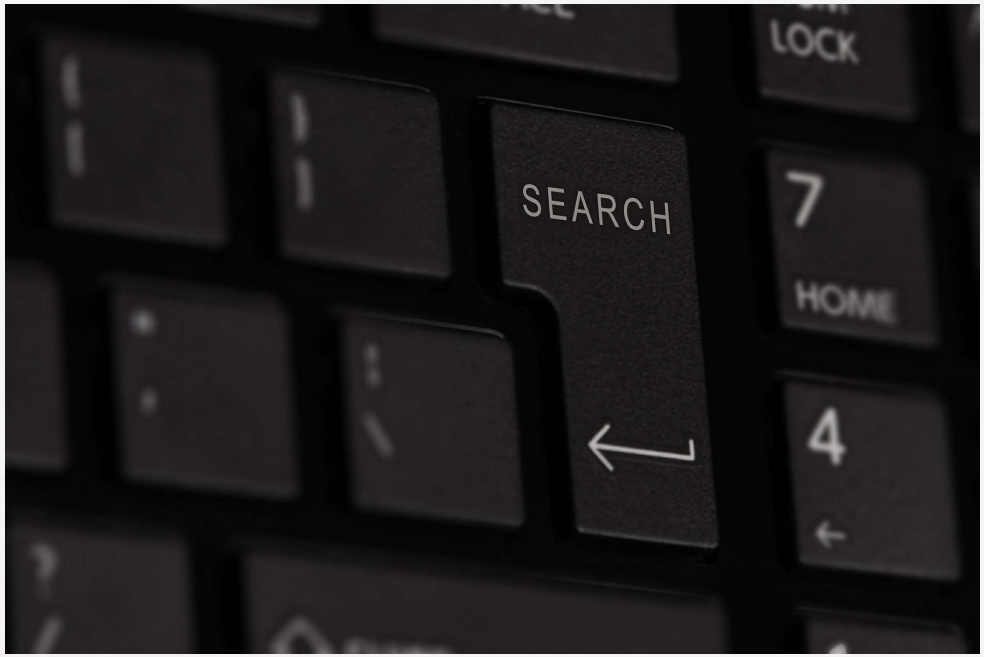
Traditional Approach: "Linear Review"

- Attorneys review 58,468 emails one by one:
  - Accuracy
  - Speed
  - Inconsistent Coding Decisions
- Cost to the Client
  - Average review speed = 70 – 100 DOCS PER HOUR
  - 600 - 800 HOURS = $24,000-$32,000 (using contract attorneys at $40/hour)
- How many docs among the 58,468 are junk/non-responsive?



Leveraging Analytics for a Successful Review – Step 1: Key Words

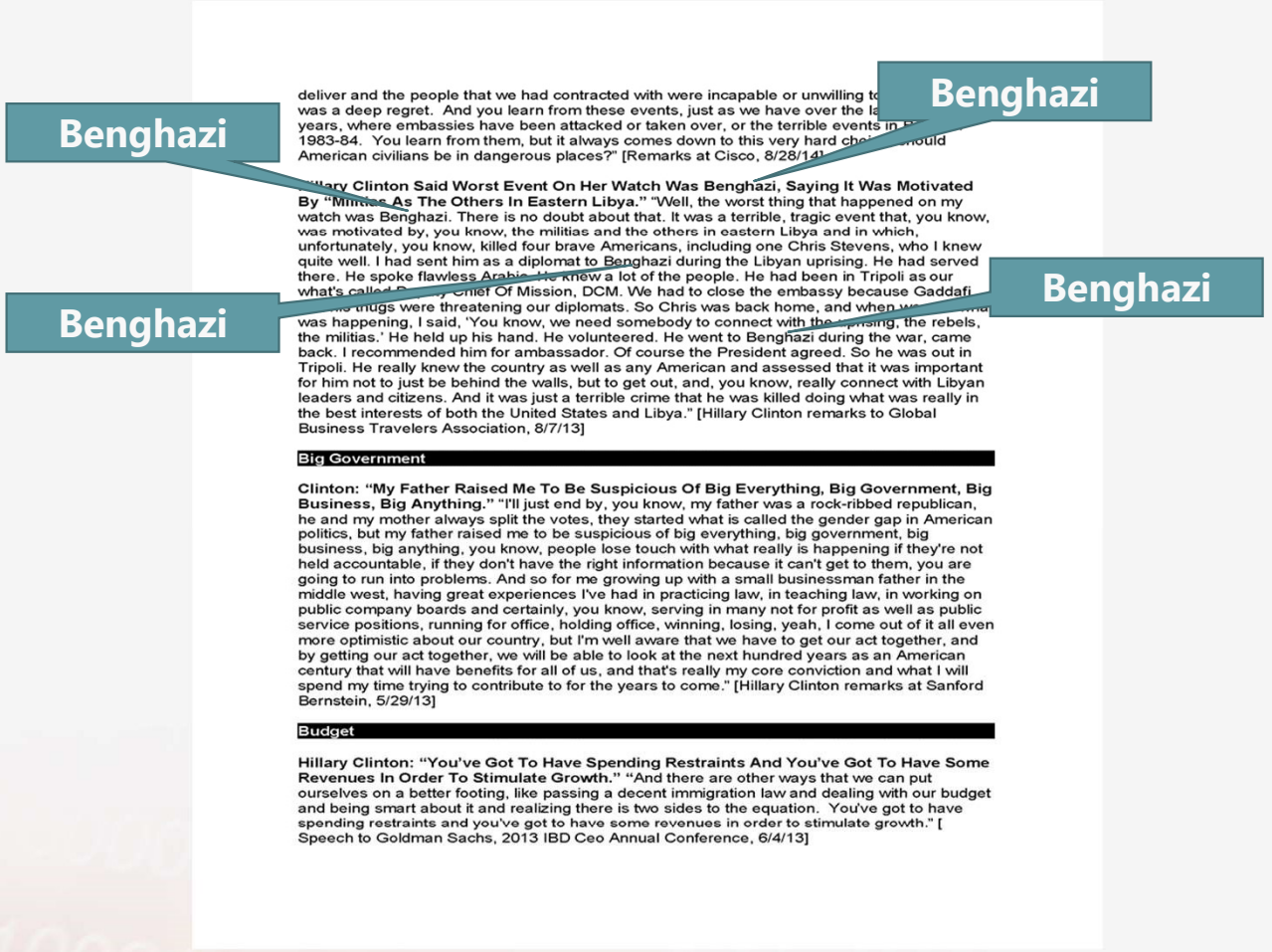
- Search the ESI with key words to unearth potentially responsive data set
- "Goldman" and "Speech" surfaced the "Speech Flags" document, a summary of important campaign topics



"HWA SPEECH FLAGS" DOC PAGE 1 (index)

HWA Speech Flags

Speech Flags



deliver and the people that we had contracted with were incapable or unwilling to
was a deep regret. And you learn from these events, just as we have over the la
years, where embassies have been attacked or taken over, or the terrible events in
1983-84. You learn from them, but it always comes down to this very hard choi
American civilians be in dangerous places?" [Remarks at Cisco, 8/28/14]

Hillary Clinton Said Worst Event On Her Watch Was Benghazi, Saying It Was Motivated By "Militias As The Others In Eastern Libya." "Well, the worst thing that happened on my watch was Benghazi. There is no doubt about that. It was a terrible, tragic event that, you know, was motivated by, you know, the militias and the others in eastern Libya and in which, unfortunately, you know, killed four brave Americans, including one Chris Stevens, who I knew quite well. I had sent him as a diplomat to Benghazi during the Libyan uprising. He had served there. He spoke flawless Arabic. He knew a lot of the people. He had been in Tripoli as our what's called Deputy Chief Of Mission, DCM. We had to close the embassy because Gaddafi thugs were threatening our diplomats. So Chris was back home, and when
was happening, I said, 'You know, we need somebody to connect with the uprising, the rebels, the militias.' He held up his hand. He volunteered. He went to Benghazi during the war, came back. I recommended him for ambassador. Of course the President agreed. So he was out in Tripoli. He really knew the country as well as any American and assessed that it was important for him not to just be behind the walls, but to get out, and, you know, really connect with Libyan leaders and citizens. And it was just a terrible crime that he was killed doing what was really in the best interests of both the United States and Libya." [Hillary Clinton remarks to Global Business Travelers Association, 8/7/13]

Big Government

Clinton: "My Father Raised Me To Be Suspicious Of Big Everything, Big Government, Big Business, Big Anything." "I'll just end by, you know, my father was a rock-ribbed republican, he and my mother always split the votes, they started what is called the gender gap in American politics, but my father raised me to be suspicious of big everything, big government, big business, big anything, you know, people lose touch with what really is happening if they're not held accountable, if they don't have the right information because it can't get to them, you are going to run into problems. And so for me growing up with a small businessman father in the middle west, having great experiences I've had in practicing law, in teaching law, in working on public company boards and certainly, you know, serving in many not for profit as well as public service positions, running for office, holding office, winning, losing, yeah, I come out of it all even more optimistic about our country, but I'm well aware that we have to get our act together, and by getting our act together, we will be able to look at the next hundred years as an American century that will have benefits for all of us, and that's really my core conviction and what I will spend my time trying to contribute to for the years to come." [Hillary Clinton remarks at Sanford Bernstein, 5/29/13]

Budget

Hillary Clinton: "You've Got To Have Spending Restraints And You've Got To Have Some Revenues In Order To Stimulate Growth." "And there are other ways that we can put ourselves on a better footing, like passing a decent immigration law and dealing with our budget and being smart about it and realizing there is two sides to the equation. You've got to have spending restraints and you've got to have some revenues in order to stimulate growth." [Speech to Goldman Sachs, 2013 IBD Ceo Annual Conference, 6/4/13]

Analytics for Success - Step 2: Random Sample Review

- Estimate the percentage of the 58,465 documents that are relevant to the core issues ("richness")
  - Method: Pull statistically valid random sample for review
  - Here: 382 documents (process will give you 95% confidence level, +- 5%)
  - Methods have been approved by courts
- Time to Review Sample: 4-5 hours
- Result: Only 14 of the 382 were marked "responsive" (3.66%, very low richness)

Analytics for Success - Step 3: Train the Analytics to Find Relevant Documents First

3.66% richness = 2,143 responsive documents

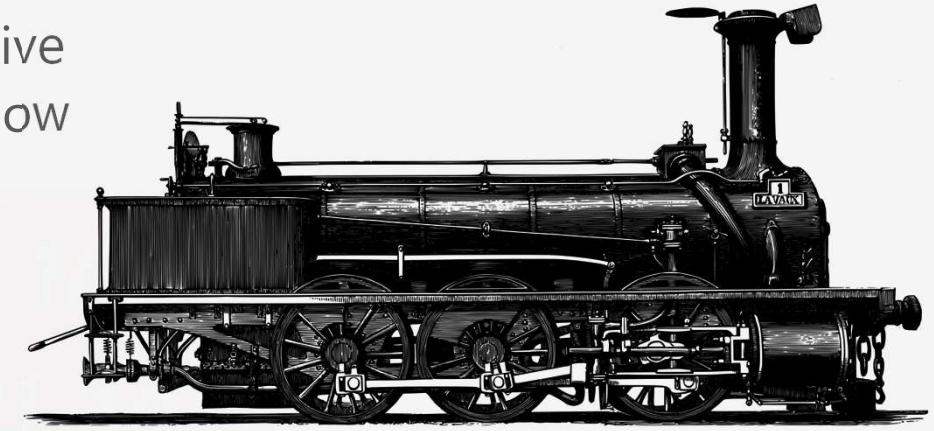
- Waste of time and money to review full set when less than 4 of 100 documents are likely to be relevant

Solution: Review the relevant emails, first.

- How? Use analytics to surface the most relevant documents earlier in the review.

Training the Data Set: 100% Relevance in the Next 50 Documents

- Use the coding decisions on the 382-document sample set to "train" the analytics software
- The software identified a set of 50 documents as potentially responsive
  - RESULT: All 50 were deemed responsive
  - Examples of a few hot documents follow



LAW OFFICES
WILLIAMS & CONNOLLY LLP
725 TWELFTH STREET, N.W.
WASHINGTON, D. C. 20005-5901
(202) 434-5000
FAX (202) 434-5029

DAVID E. KENDALL
(202) 434-5145
dkendall@wc.com

EDWARD BENNETT WILLIAMS (1920-1988)
PAUL R. CONNOLLY (1922-1978)

March 27, 2015

BY FIRST-CLASS SURFACE AND ELECTRONIC MAIL

The Honorable Trey Gowdy
United States House of Representatives
Select Committee on Benghazi
Washington, DC 20515

Dear Mr. Chairman:

This letter will respond to (1) the subpoena duces tecum issued by the Benghazi Select Committee to the Hon. Hillary R. Clinton and served by agreement on March 4, 2015; and (2) your March 19, 2015 letter requesting that former Secretary of State Clinton make her e-mail server available for third-party inspection and review.

Response to the Subpoena

As you know, the subpoena calls for the following documents, for the period January 1, 2011 through December 31, 2012, referring or relating to:

(a) Libya (including but not limited to Benghazi and Tripoli);

(b) weapons located or found in, imported or brought into, and/or exported or removed from Libya;

(c) the attacks on U.S. facilities in Benghazi, Libya on September 11, 2012 and September 12, 2012; or

(d) statements pertaining to the attacks on U.S. facilities in Benghazi, Libya on September 11, 2012 and September 12, 2012.

The subpoena requests production of any documents sent from or received by the e-mail addresses "hdr22@clintonemail.com" or "hrod17@clintonemail.com." As explained in my March 4, 2015 e-mail to your Staff Director and certain others, "hrod17@clintonemail.com" is not an address that existed during Secretary Clinton's

flag

email

email

Re: CLIPS ON EMAIL RELEASE

From:
Tyson Brody <tbrody@hillaryclinton.com>
To:
Ian Sams <isams@hillaryclinton.com>
Cc:
Josh Schwerin <jschwerin@hillaryclinton.com>, Brian Fall .com>, Christina Reynolds <creynolds@hillaryclinton.com>, HRCRR <hrcrr@hillarycl mieri <jpalmieri@hillaryclinton.com>, Kristina Schake <kschake@hillaryclinton.com>, John Podesta <john. ta@gmail.com>, Robby Mook <re47@hillaryclinton.com>, Huma Abedin <ha16@hillaryclinton.com>, Cheryl Mills <che mills@gmail.com>, Jake Sullivan <jsullivan@hillaryclinton.com>
Date:
22 May 2015 13:16:55 -0400

Want to flag that if you look at the email in this link, HRC's email is redacted. But in yesrterday's new york times dump, the same email is unredacted. Could it be because Committee leaked?

https://twitter.com/gabrielmalor/status/601796614060167169

See it unredacted from the times below:

From: Sullivan, Jacob J <SullivanJJ@state.gov>
Sent: Monday, October 1, 2012 3:37 PM
To: H
Subject: RE: H: Romney's last gambit. Got done and published. Sid

Will do.

From: H [mailto:HDR22@clintonemail.com]
Sent: Monday, October 01, 2012 3:34 PM
To: Sullivan, Jacob J
Subject: Fw: H: Romney's last gambit. Got done and published. Sid

Be sure Ben knows they need to be ready for this line of attack.

From: Sidney Blumenthal [mailto:sbwhoeop@[redacted]
Sent: Monday, October 01, 2012 10:13 AM
To: H
Subject: H: Romney's last gambit. Got done and published. Sid

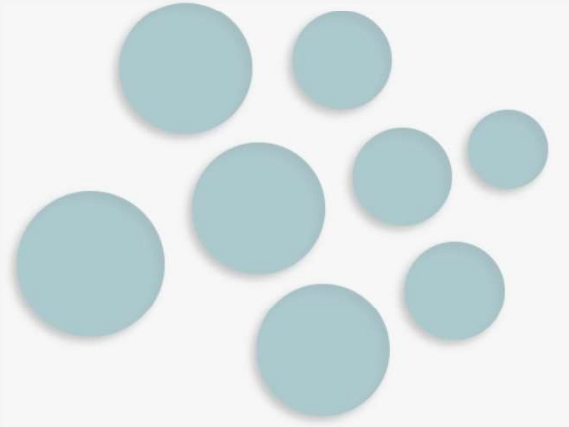
http://www.salon.com/2012/10/01/gops_october_surprise/

Monday, Oct 1, 2012 09:30 AM EDT

GOP's October surprise?

Better Results/Less Client Spend

- Unnecessary to review all 58,000 emails to achieve a reasonable "recall" rate (80%)
  - Recall = the percentage of the truly responsive documents were found by the review process
  - Manual reviewers, on average, miss 20%-75% of all relevant documents (recall = 25-80%)
- Review Hour Savings: Review for 20-30 hours rather than 600-800 (manual view)



What Do Judges Think About Analytics Technology?

Judicial Acceptance: Advanced Analytics

- First blessed by a court in 2012
  - *Da Silva Moore v. Publicis Group*, 287 F.R.D. 182 (S.D.N.Y. 2012)
    - Sex discrimination case against large advertising firm
    - Defendant sought to use AI to reduce massive ESI volumes for review, parties had dispute over methods
    - "Statistics clearly show that computerized searches are at least as accurate, if not more so, than manual review." (*Id.* at 190)
    - "While . . . computer-assisted review is not perfect, the [FRCP] do not require perfection." (*Id.* at 191)

Judicial Acceptance: Advanced Analytics

- *Da Silva Moore's* Progeny
  - *Global Aerospace Inc. v. Landow Aviation, L.P*, 2012 WL 1431215, No. CL 61040 (Va. Cir. Ct. Apr. 23, 2012)
    - 250 GB of ESI to be reviewed in commercial litigation
    - Defendants permitted to use predictive coding over plaintiffs' objections
    - Observed that analytics "is capable of locating upwards of seventy-five percent of the potentially relevant documents . . . at a fraction of the cost and in a fraction of the time of linear review." (*Id.* at *1)

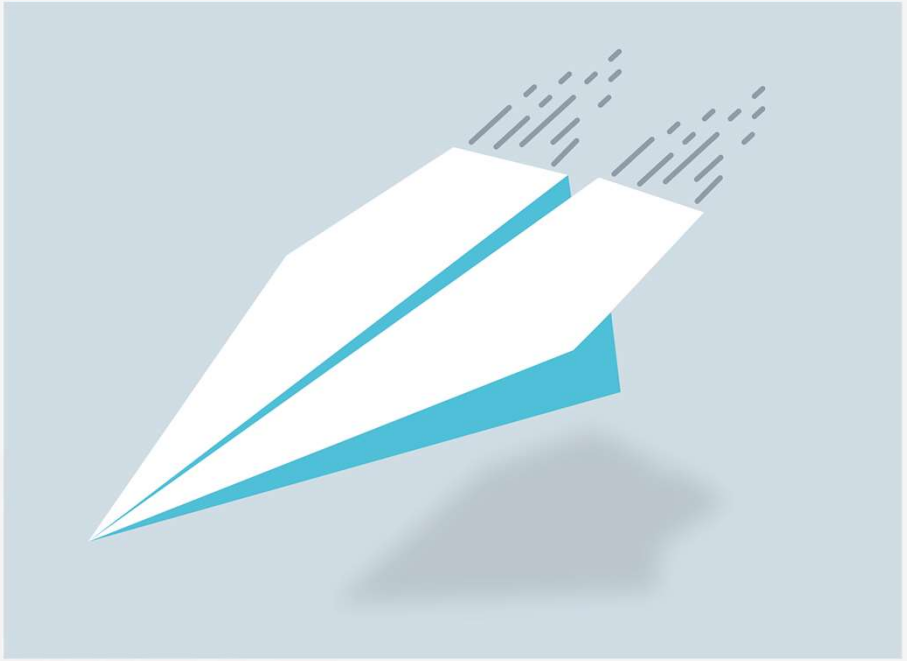
Judicial Acceptance: Advanced Analytics

- *Da Silva Moore's* Progeny
  - *Federal Housing Finance Agency v. HSBC North America Holdings, Inc.*, 2014 WL 584300 (S.D.N.Y. Feb. 14, 2014)
    - Court denied a request for reconsider of a discovery order permitting the use of analytics
    - Defendants permitted to use predictive coding over plaintiffs' objections
    - "The literature that the Court reviewed . . . Indicated that predictive coding had a better track record in the production of responsive documents than human review[.]" (*Id.* at *3)
    - "[N]o one could or should expect perfection from the discovery process. All that can be legitimately expected is a good faith . . . commitment to produce . . . responsive documents." (*Id.* at *2)

New Huntsman v. SW Airlines Co.,

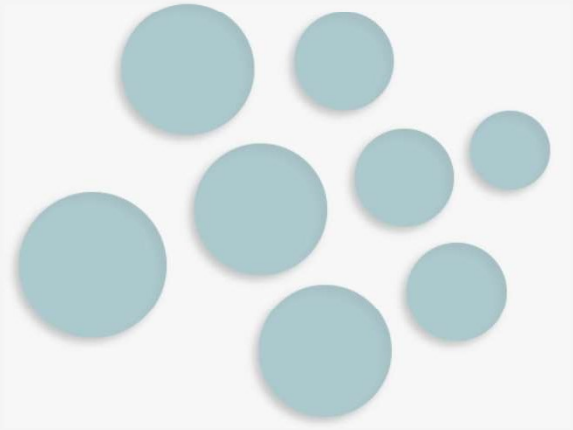
No. 19-cv-00083-PJH, 2021 BL 301052, 2021 US DistLexis 150170 (N.D. Cal. Aug. 10, 2021)

- "Southwest's approach to using keyword searches and technology-assisted review in tandem does not offend the court's expectation that the parties conduct a reasonable inquiry as required by the rules."



In re Diisocyanates Antitrust Litig.,
2021 WL 4295729 (W.D. Pa. Aug. 23, 2021)

- Plaintiffs filed a motion to require Defendants to use certain search terms and TAR methodologies to identify responsive documents. Defendants cross-moved for a 26(c) protective order to allow them to use their own search terms and TAR methodology
- Special Master: parties should continue to meet and confer on the areas of dispute using the provided guidance as a roadmap
  - "Transparency transcends cooperation. It does not mean merely that parties must discuss issues concerning the discovery of ESI; it requires that they disclose information sufficient to make those discussions, as well as any court review, meaningful."



Your Questions and Feedback

Contact Us:

John Unice, Esq.
bit-x-bit, LLC
Chief Executive Officer
john.unice@bit-x-bit.com

Brett Creasy, CISSP, GCFA, CCE
bit-x-bit, LLC
President & Director of Digital Forensics
brett.creasy@bit-x-bit.com

bit-x-bit SM

THE EVIDENCE. THE BACKSTORY.
THE INTELLIGENCE. **WE GET IT.** SM

bit-x-bit
437 Grant Street
Suite 1250
Pittsburgh, PA 15219

412.325.4033

www.bit-x-bit.com