

eDiscovery. There is a better way



Predictive Coding

Gain Earlier Insight and Reduce Document Review Costs

Tom Groom
Vice President, Discovery Engineering
tgroom@d4discovery.com
303.840.3601



D4 LLC

- Litigation support service provider since 1997
- National footprint with 100 employees
- Seasoned Professionals with Significant Industry Knowledge
- Known as “Thought Leaders” in the industry
- www.d4discovery.com



Agenda

- eDiscovery Mega Trends – The Sedona Conference
- What is Predictive Coding?
- Effective Uses and Workflows
 - Early Assessment
 - Data Reduction
 - Expedited Review
 - Post Review QA
- Example: Equivio->Relevance
- Q&A



eDiscovery Mega Trends – The Sedona Conference

- New sources of ESI becoming available
- Plaintiff bar becoming more aggressive in requesting ESI
- Corporate clients taking more control
 - To reduce overall cost
 - To use ESI as an offensive strategy
- Increasing ESI Volume
 - Available storage doubles every 18 months (Moore's Law)
 - Data expands to fill the space available for storage (Parkinson's Law)
- More acceptance in using data analytics & sampling tools
 - Early Data Assessment
 - Statistical Sampling / Predictive Coding
 - Conceptual Analytics in Review and QC

What is Predictive Coding?

- A computerized sampling system combined with the intelligence provided by a human “expert.”
- Built upon a well established framework called Predictive Analytics
- Predictive Analytics encompasses a variety of techniques from statistics, data mining, conceptual search and game theory which analyze current and historical facts to make predictions about future events
- Currently used in actuarial science, financial services, insurance, telecommunications, retail, travel, healthcare, and pharmaceuticals
- A well-known example is your FICO Score where financial scoring models process your credit history, loan application, customer data, etc., in order to rank-order your likelihood of making future credit payments on time.



Predictive Coding for Litigation Support

- Recommind
- Equivio->Relevance
- EQOD
- OrcaTec
- Relativity (via the “Rapid Review” workflow)
- More to be released by the end of 2011
 - AccessData (Next Generation “Summation”)
 - LexisNexis (Next Generation “Concordance”)

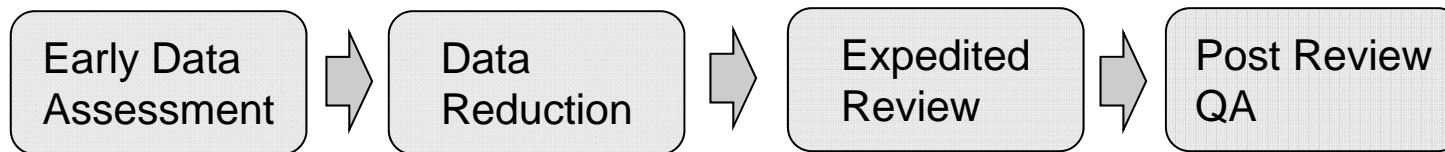


How Does it Work for eDiscovery?

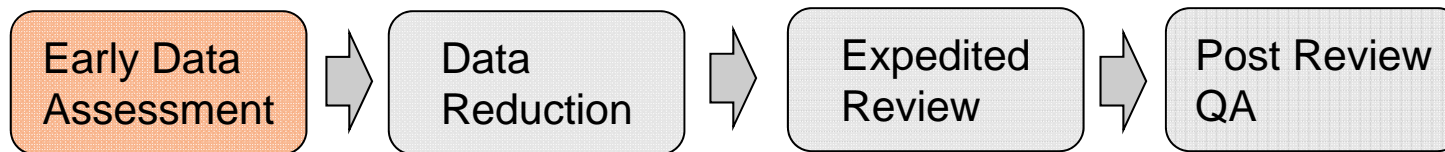
- Expert makes “yes/no” decisions on sample documents randomly selected and presented by the system
- Questions can be “Is this document responsive?” or “Does it pertain to this specific issue?” or “Is this document privileged?”, etc.
- The system builds a list of terms in the background as it learns from the expert
- At some point the system becomes “statistically stable” and can “predict” what the expert will choose
- The system can then classify or rank the rest of the collection based on the knowledge it now contains from the expert
- Many workflows can then leverage that classification and/or ranking



Where Can Predictive Coding be Applied?



Where Can Predictive Coding be Applied?



Zoom in on most relevant documents early in the case
Enables informed decisions on case strategy –

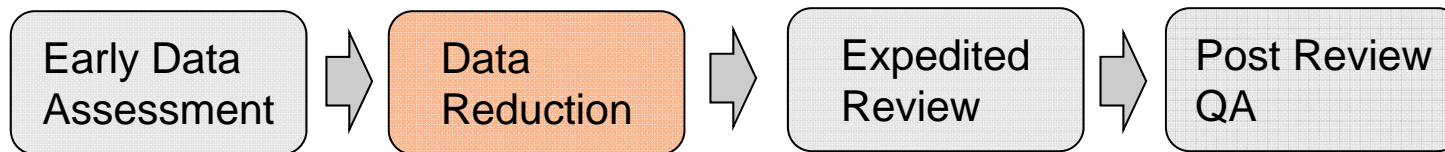
- Winnability
- Risks and costs
- Settle or defend

Richness estimates for review budgeting

Initial Search term development



Where Can Predictive Coding be Applied?



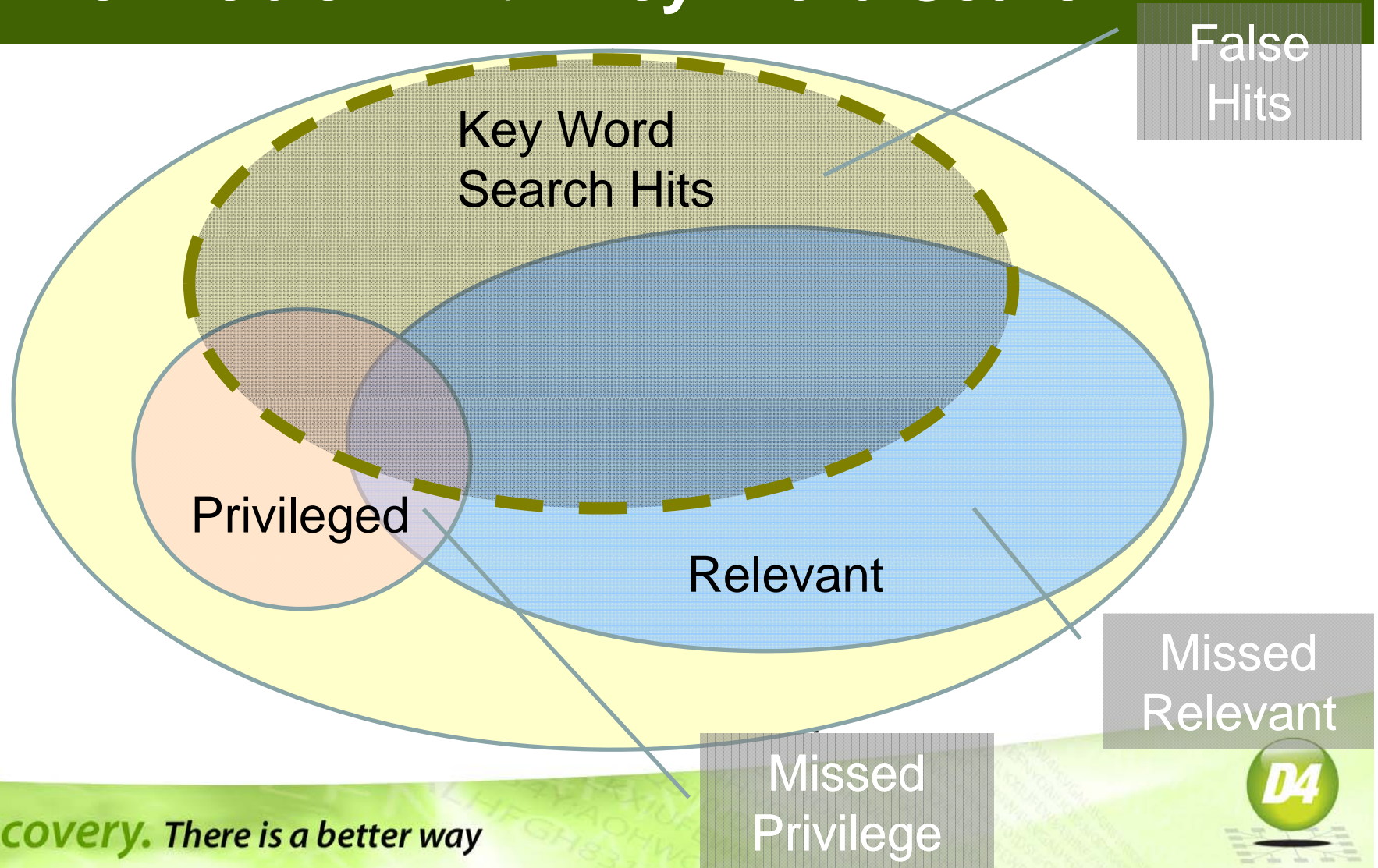
Achieves recall and precision of 70-80% (vs. 20-30% with key words alone)
Statistical model to manage over- and under-inclusive tradeoff
Read fewer documents, find more relevant documents
Finalizes key words that can be used in negotiations
Enables replacement of first pass review



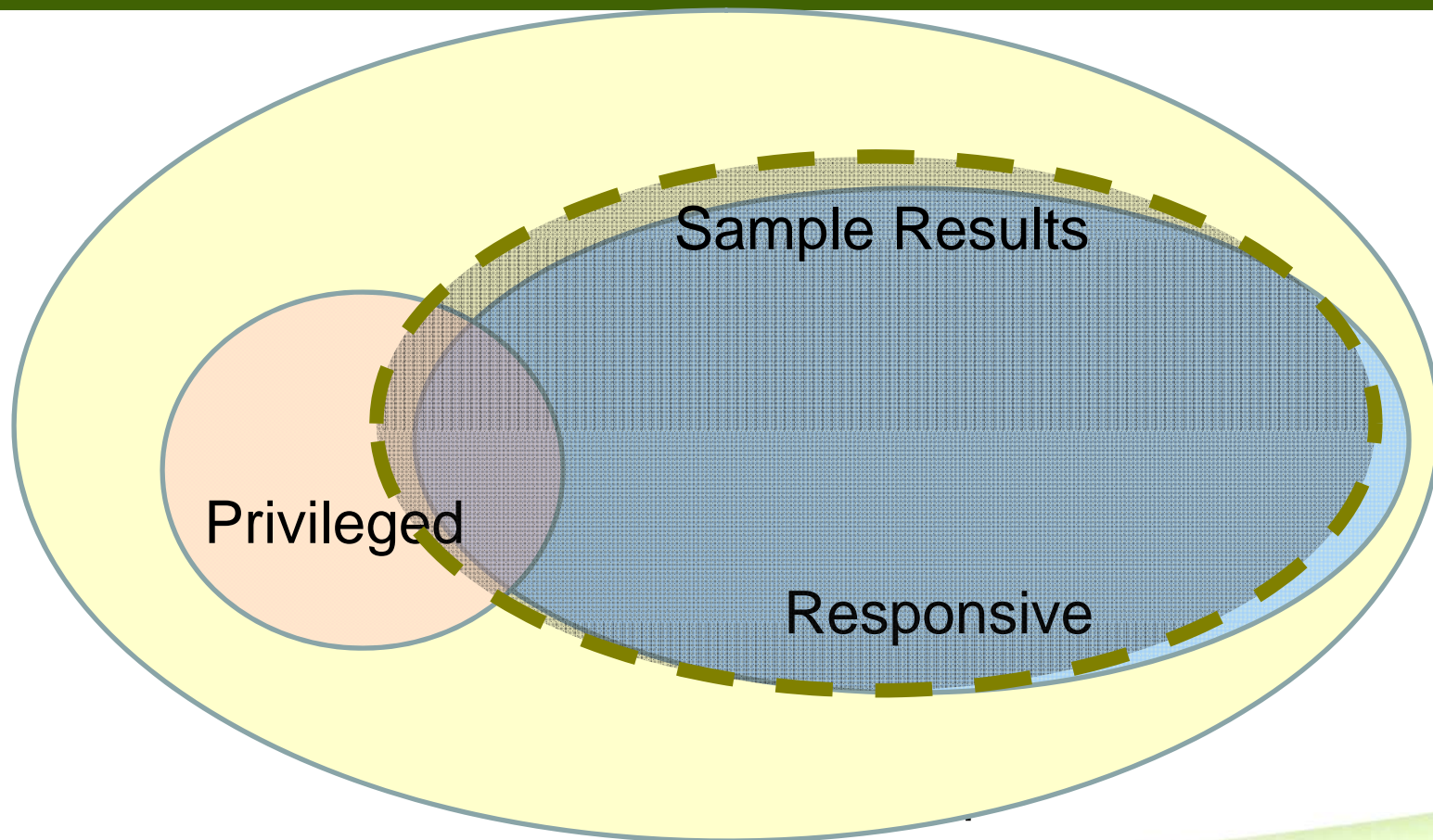
“Recall” and “Precision”

- Recall is defined as “the measure of the ability of a system to present all relevant items”. It determines “how wide to you cast the net”
 - Recall % = Number of relevant items retrieved/number of relevant items in the collection
- Precision is defined as “the measure of the ability of a system to present only relevant items” It determines “how accurate was the review process”.
 - Precision % = Number of relevant items identified/total number of items retrieved
- The “system” above is the combination of people, data, tools and workflow. The goal is to design and leverage the optimal “system” to yield sufficient recall and precision within cost limitations.

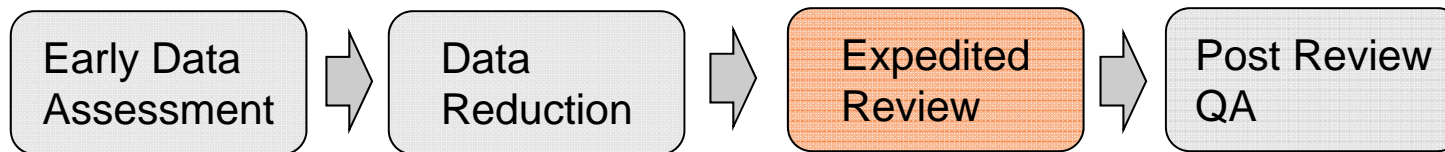
The Problem with Key Word Search



Predictive Coding Yield



Where Can Predictive Coding be Applied?



Prioritized review

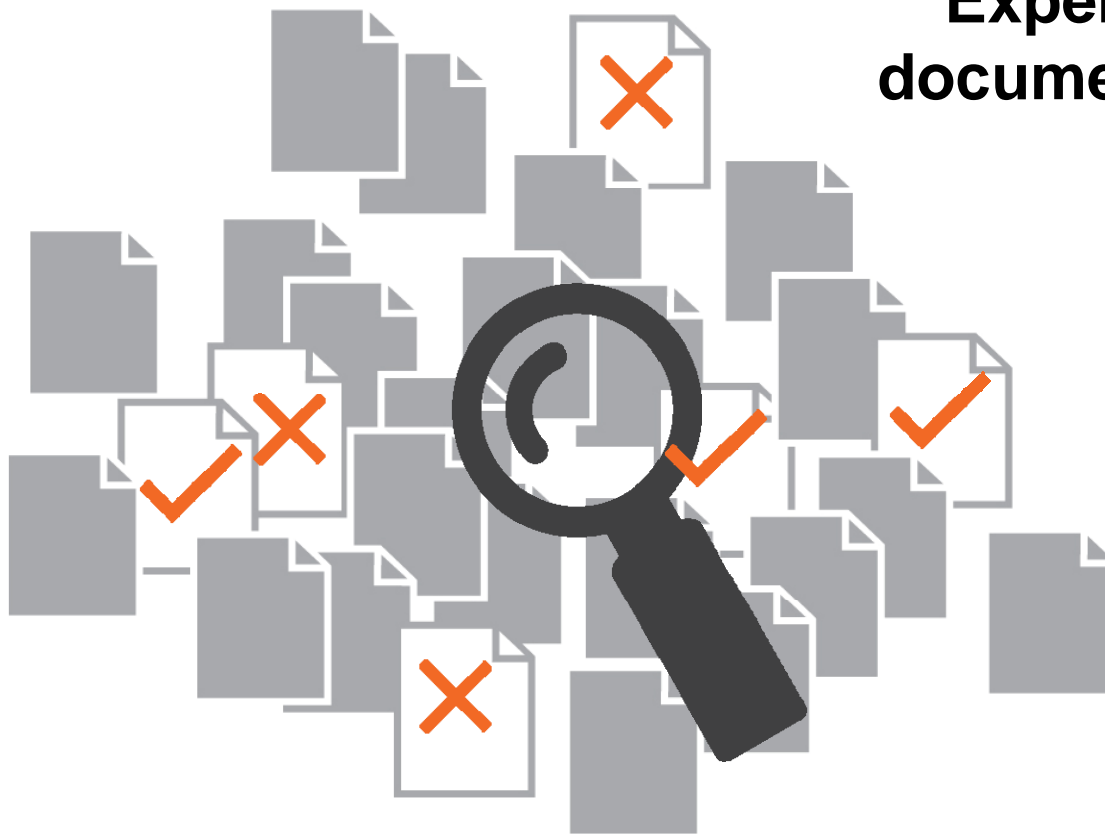
Get to the issues faster

Reduce costs by stratifying review effort

Sample “Non Responsive” documents to validate accuracy

Example – Equivio-→Relevance

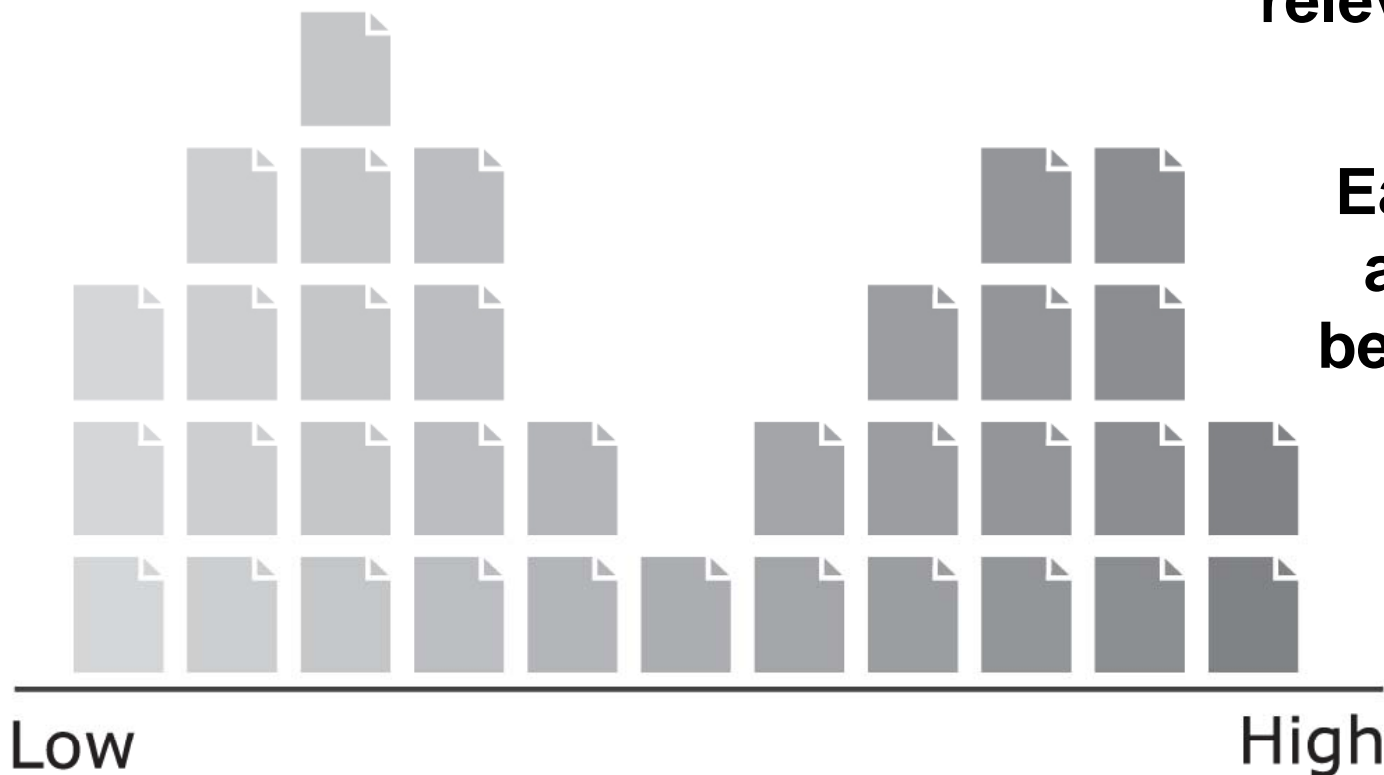
Expert reviews sample documents for relevance



Relevance End Result

Software calculates relevance scores for documents.

Each document is assigned a score between 0 and 100



Equivio > Relevance

View Actions

Navigation

equivio

3799559-pud85f00.t
5059049-vkq53d00.
837519-dul93f00.bt
4536689-tmr74f00.bx
1631079-hjt05d00.bx
3259239-nul84c00.b
1436419-gmx51c00.
5570799-xhp29c00.t
2488539-kys56d00.i
783389-doc65e00.b
4202879-rgt13a00.b
608619-cto91d00.bt
2198359-jwy64d00.t
5777229-ybi84a00.t
4868349-usk20c00.t
5303549-wia75f00.bx
32499-adv37c00.bt
3805939-put67e00.t
2746039-lxi90a00.bt
3158199-nku55d00.i
2958439-mrr57c00.t
662109-czw45a00.bx
352129-bpk97d00.b
5090689-vnr19d00.t
1050949-etp25a00.t
5218629-vzx48e00.t
1027839-ewq49c00.
2471679-kxc64d00.t
3423949-okf72c00.b
487609-cf22d00.bt

Interactive Ranking	Sample Results	Batch Ranking	Final Results	Utilities	Setup
Row Id	Document Identifier	Topic 103	Date	Size	
31	1050949-etp25a00.bt	Not Relevant	3/30/2009	6	
32	5218629-vzx48e00.bt	Not Relevant	3/30/2009	3	
33	1027839-ewq49c00.bt	Not Relevant	3/30/2009	1	
34	2471679-kxc64d00.bt	Not Relevant	3/30/2009	4	
35	3423949-okf72c00.bt	None	3/30/2009	0	
36	487609-cf22d00.bt	None	3/30/2009	7	
37	3943589-qhy15d00.bt	Relevant	3/30/2009	12	
38	4870299-usp03a00.bt	Relevant	3/30/2009	11	
39	3508249-osh39d00.bt	Not Relevant	3/30/2009	2	
40	wcv76e00.TXT	Not Relevant	3/30/2009	0	

100- Menthol Market Review
To Understand Long Term Dynamics and Patterns of the menthol "world"
To Evaluate PM Discount opportunities in the menthol market.
Discount Portfolio Optimization
To Evaluate New Menthol Brand opportunity among YAS
Brand X menthol
9699~0Ls0g
PF/MaMhol Ra.aw.2j94

pgNbr=1
Size
Menthol accounts for 25.7% (12MM March '94) of the industry.
The menthol market is declining since the beginning of the 80's
from 28.7% in 1982 to 25.7% to date (Shipments)
On a short term basis, the menthol market strongly decreased in 2nd quarter of '93 and came back to trend after PRP.
t69990L~0t
PF/M~hol Rewaw-Z94

Get More Documents

Calculate Sample Results

State

Setup
Interactive Ranking
Sampling Results
Batch Ranking
Final Results

Sample Status

All

Total: 1398/1400

Not Relevant: 953

Relevant: 445

Skip: 0

38/40

27

Relevant: 11

Skip: 0

Ranking Palette

Topic 103 Not Relevant Relevant Skip

Connected to Case: ER_30



Equivio > Relevance

View Actions

Navigation

Interactive Ranking | Sample Results | Batch Ranking | Final Results | Utilities | Setup

State

equivio

- Dashboard
 - Topic 103
 - Keywords
 - Ranking Statistics

Topic 103

Progress Indication

0 5 10 15 20 25 30 35 40 45

Not Stable
Nearly Stable
Stable

If you are satisfied with the results, click "Perform Batch Ranking" to move to the next step. Otherwise click "Continue Interactive Ranking" to continue sampling.

Continue Interactive Ranking

Perform Batch Ranking

Setup

Interactive Ranking

Sampling Results

Batch Ranking

Final Results

Sample Status

All

Total:	1400/1400
Not Relevant:	955
Relevant:	445
Skip:	0

Current

Total:	40/40
Not Relevant:	29
Relevant:	11
Skip:	0

Connected to Case: ER_30

Equivio > Relevance

View Actions

Navigation ☰ ↑ ×

Interactive Ranking | **Sample Results** | **Batch Ranking** | **Final Results** | **Utilities** | **Setup**

State ☰ ↑ ×

- Dashboard
- Topic 103
 - Keywords
 - Ranking Statistics**

Sampling Progress Graph

Estimated F-measure

Iterations

- Not Stable
- Nearly Stable
- Stable

The Sampling Progress graph shows the behavior of the F-measure from sample to sample. The graph illustrates the progressive convergence of the F-measure estimate through the sampling iterations. The orange line shows the estimated maximum F-measure. The band around the orange line delineates the confidence interval for the F-measure estimate. The shade of the band denotes the stability of the F-measure estimate, where dark blue indicates that the F-measure has stabilized, having reached the optimal level for the current case.

Sample Status

All

Total:	1400/1400
Not Relevant:	955
Relevant:	445
Skip:	0

Current

Total:	40/40
Not Relevant:	29
Relevant:	11
Skip:	0

Connected to Case: ER_30

Equivio > Relevance

View Actions

Navigation

Interactive Ranking | Sample Results | Batch Ranking | Final Results | Utilities | Setup

State

Setup

Interactive Ranking

Sampling Results

Batch Ranking

Final Results

Sample Status

All

Total:	1400/1400
Not Relevant:	955
Relevant:	445
Skip:	0

Current

Total:	40/40
Not Relevant:	29
Relevant:	11
Skip:	0

equivio

Topic 103

- Keywords
- Discrepancy Set
- Discrepancy Analy

Relevance Distribution Graph

Relevant Documents Estimators

Review documents: 0% 100%

Review set parameters (*)

Documents to review:	552,815	8.0% of total collection
Relevant documents retrieved:	498,200	67.4% of relevant documents

(*) Assumes review of documents with score above 69

Population parameters

Number of documents in collection:	6,910,192	
Estimated number of relevant documents in collection:	739,390 (+/-158,934)	10.7% (+/-2.3%)

Connected to Case: ER_30



Equivio > Relevance

View Actions

Navigation

Interactive Ranking | Sample Results | Batch Ranking | Final Results | Utilities | Setup

State

Setup

Interactive Ranking

Sampling Results

Batch Ranking

Final Results

Sample Status

All

Total: 1400/1400

Not Relevant: 955

Relevant: 445

Skip: 0

Current

Total: 40/40

Not Relevant: 29

Relevant: 11

Skip: 0

Connected to Case: ER_30

equivio

Topic 103

- Keywords
- Discrepancy Set
- Discrepancy Analy

Relevance Distribution Graph

Relevant Documents Estimators

Review documents: 0% 100%

Review set parameters (*)

Documents to review:	1,382,038	20.0% of total collection
Relevant documents retrieved:	678,464	91.8% of relevant documents

(*) Assumes review of documents with score above 17

Population parameters

Number of documents in collection:	6,910,192
Estimated number of relevant documents in collection:	739,390 (+/-158,934) 10.7% (+/-2.3%)



Equivio > Relevance

View Actions

Navigation

Interactive Ranking | Sample Results | Batch Ranking | Final Results | Utilities | Setup

State

Setup

Interactive Ranking

Sampling Results

Batch Ranking

Final Results

Sample Status

All

Total:	1400/1400
Not Relevant:	955
Relevant:	445
Skip:	0

Current

Total:	40/40
Not Relevant:	29
Relevant:	11
Skip:	0

equivio

Topic 103

- Keywords
- Discrepancy Set
- Discrepancy Analy

Relevance Distribution Graph

Relevant Documents Estimators

Review documents: 0% 100%

Review set parameters (*)

Documents to review:	276,407	4.0% of total collection
Relevant documents retrieved:	228,175	30.9% of relevant documents

(*) Assumes review of documents with score above 92

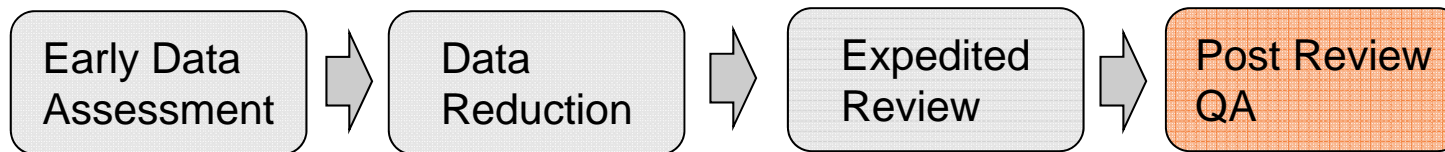
Population parameters

Number of documents in collection:	6,910,192	
Estimated number of relevant documents in collection:	739,390 (+/-158,934)	10.7% (+/-2.3%)

Connected to Case: ER_30



Where Can Predictive Coding be Applied?



Systemize QA

Focus QA on docs with high probability of error

Track accuracy of individual reviewers

Additional Resources and Links

- eDiscovery Institute Survey on Predictive Coding
 - <http://www.ediscoveryinstitute.org/pubs/PredictiveCodingSurvey.pdf>
- Predictive Coding Demystified
 - <http://www.wortzmannickle.com/ediscovery-blog/2011/04/28/predictive-coding-demystified/>
- E-Discovery And The Rise of Predictive Coding
 - <http://eddblogonline.blogspot.com/2011/03/e-discovery-and-rise-of-predictive.html>
- eDiscovery Trends: Forbes on the Rise of Predictive Coding
 - <http://blogs.forbes.com/benkerschberg/2011/03/23/e-discovery-and-the-rise-of-predictive-coding/>
- Linked in Group: Text Analysis and Predictive Coding in E-Discovery
 - http://www.linkedin.com/groups/Text-Analysis-Predictive-Coding-in-3319494?trk=myg_ugrp_ovr



Summary

- eDiscovery volume is driving the demand for tools that can prioritize review
- Predictive Coding is a class of tools that combine the consistency and efficiency of a computerized system with the human “expert” intelligence to rank/classify document collection prior to review
- Effective Uses and Workflows
 - Early Assessment
 - Data Reduction
 - Expedited Review
 - Post Review QA
- Equivio->Relevance Example
- Questions?



eDiscovery. There is a better way



Predictive Coding

Gain Earlier Insight and Reduce
Document Review Costs

Tom Groom
Vice President, Discovery Engineering
tgroom@d4discovery.com
303.840.3601

