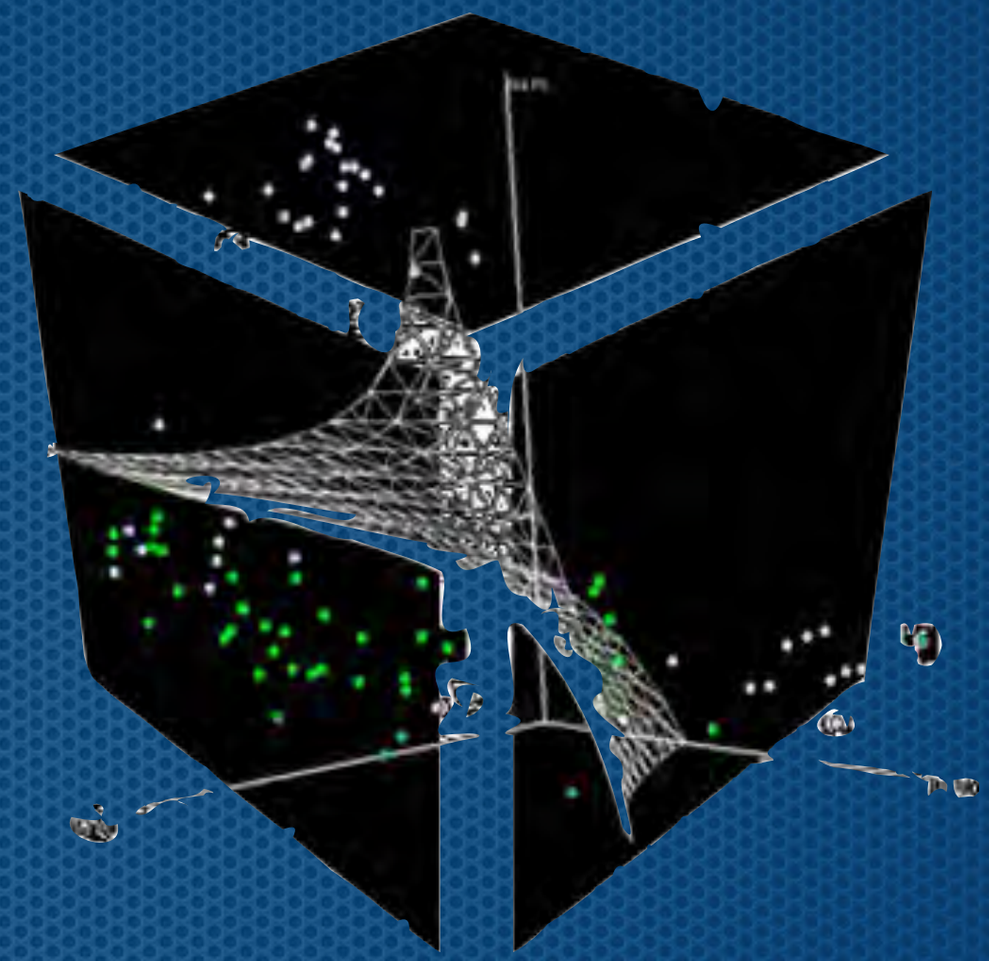
Predictive Coding:
An Introduction and Real World Example



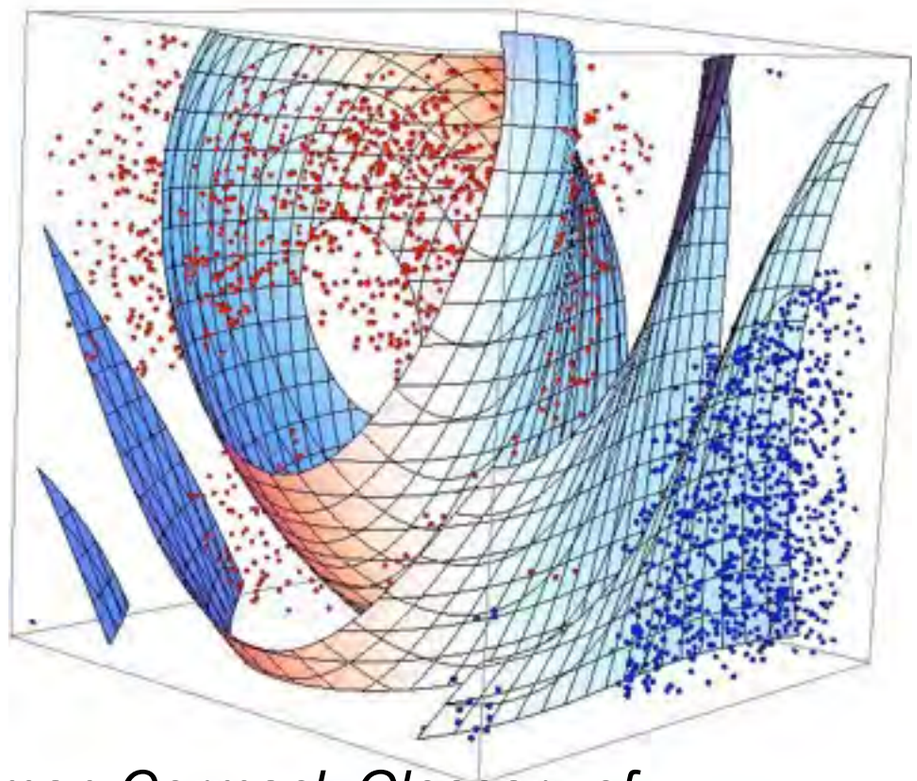
Presenters name and other information omitted to try to limit temporal anomalies

Overview of the Basic Math Underlying Predictive Coding

$$
\begin{aligned}
&P(Z_{(m,n)} = k | \boldsymbol{Z}_{-(\boldsymbol{m},\boldsymbol{n})}, \boldsymbol{W}; \alpha, \beta) \\
\propto &P(Z_{(m,n)} = k, \boldsymbol{Z}_{-(\boldsymbol{m},\boldsymbol{n})}, \boldsymbol{W}; \alpha, \beta) \\
= &\left(\frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \right)^{M} \prod_{j \neq m} \frac{\prod_{i=1}^{K} \Gamma(n_{j,(\cdot)}^{i} + \alpha_i)}{\Gamma\left(\sum_{i=1}^{K} n_{j,(\cdot)}^{i} + \alpha_i\right)} \\
&\times \left(\frac{\Gamma\left(\sum_{r=1}^{V} \beta_r\right)}{\prod_{r=1}^{V} \Gamma(\beta_r)} \right)^{K} \prod_{i=1}^{K} \prod_{r \neq v} \Gamma(n_{(\cdot),r}^{i} + \beta_r) \\
&\times \frac{\prod_{i=1}^{K} \Gamma(n_{m,(\cdot)}^{i} + \alpha_i)}{\Gamma\left(\sum_{i=1}^{K} n_{m,(\cdot)}^{i} + \alpha_i\right)} \prod_{i=1}^{K} \frac{\Gamma(n_{(\cdot),v}^{i} + \beta_v)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)} \\
\propto &\frac{\prod_{i=1}^{K} \Gamma(n_{m,(\cdot)}^{i} + \alpha_i)}{\Gamma\left(\sum_{i=1}^{K} n_{m,(\cdot)}^{i} + \alpha_i\right)} \prod_{i=1}^{K} \frac{\Gamma(n_{(\cdot),v}^{i} + \beta_v)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)} \\
\propto &\prod_{i=1}^{K} \Gamma(n_{m,(\cdot)}^{i} + \alpha_i) \prod_{i=1}^{K} \frac{\Gamma(n_{(\cdot),v}^{i} + \beta_v)}{\Gamma\left(\sum_{r=1}^{V} n_{(\cdot),r}^{i} + \beta_r\right)}..
\end{aligned}
$$

What is Predictive Coding?

Active Machine Learning, a form of **AI,** where a computer learns to rank the probable relevance of an entire collection of documents based on a Subject Matter Expert(s) ("SME") classification of a small Training Set of documents.



See: Grossman Cormack Glossary

Maura R. Grossman and Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 2013 Fed. Cts. L. Rev. 7 (January 2013)
http://cormack.uwaterloo.ca/targlossary/

Also see: LegalSearchScience.com

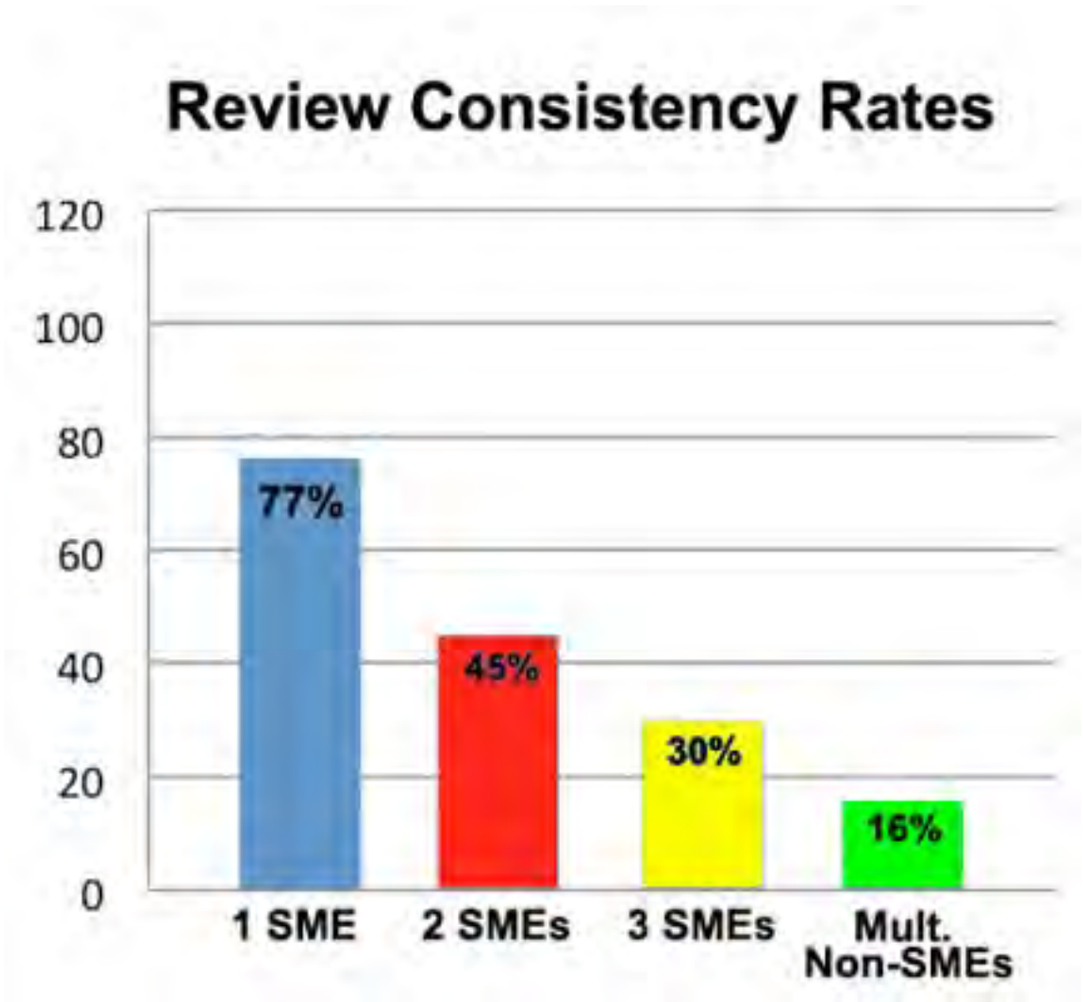
This kind of Machine Learning has Very Common in Business and Advertising for decades

Reduce Costs with Predictive Coding

- **Where the Money Goes:** *Understanding Litigant Expenditures for Producing Electronic Discovery*, 2012
- Study recommends predictive coding:
  - » "The increasing volume of digital records makes predictive coding and other computer-categorized review techniques not only a cost-effective option to help conduct review but the *only* reasonable way to handle large-scale production."
- Notes the cost advantages:
  - » "We believe that one way to achieve substantial savings in producing massive amounts of electronic information would be to let computers do the heavy lifting for review."

Predictive Coding Replaces Large Scale Document Reviews, which is a Good Thing Because Humans Are Inconsistent in Determinations

- Ellen Voorhees (1998) study using three well trained experts, retired intelligence officers.
  - Two experts agreed on relevance of documents to queries only 45% of time. (Disagreed 55%)
  - With three experts it was only 30%. (Disagreed 70%)
- Herb Roitblat, et al (2010) study found only 16% agreement among team of professional legal reviewers using quality controls between 1st and 2nd reviews. (disagreed 84%). Cormack.
- Losey - 699,082 Enron Email Study by one SME of Multimodal v Monomodal predictive coding methods found 77% agreement in two reviews six months apart (*Jaccard Index*). (23% disagreement)
  - include irrelevant docs = 98% agreement



The solution is supervised machine learning, but the role of skilled humans is still central.

- **Computers Are 100% Consistent (and Fast)**
- Review should be limited to a small number of SMEs (ideally one) who train the machines.
- Beware of sample sizes in quality assurance tests that are too large to avoid review teams.
- Beware of over-delegation to machines - the monomodal Borg approach.
  - Lucky Borg
  - Introverted Borg
  - Enlightened Borg

- Three-Cylinder Multimodal Approach To Predictive Coding
- Humans have important skills; Human–computer information retrieval (*HCIR*)

Hybrid Multimodal Approach

- CAR with Three-Cylinder Predictive Coding Search Engine:
  - Random, Machine, and Judgmental
- Judgmental uses multiple modes of search:
  - Expert Manual Review - *HCIR*
  - Parametric Boolean Keyword
  - Near Deduplication, Clusters (unsupervised machine learning)
  - Concept Searches
  - Predictive Coding

See Losey's @100 articles at LegalSearchScience.com

Consider the Advantages of the Borg's Hive Mind and Take Steps to Minimize Inconsistent Reviews:

- Limit Number of SMEs and Reviewers
- *More than one reviewers* must communicate often and try hard to keep the same mind on relevance.
- Talk early and often to the requesting party on the scope of relevancy.
  - Know what you are looking for.
  - GIGO.
- Include the judge early on relevance disagreements.

Overview of Predictive Coding Review Project

Two Pass Review:

1. ID Likely Relevant by Predictive Coding
2. Confirm Relevant, ID Confidentiality, Redact, Log

Assemble Team

- Subject Matter Experts ("SME")
- Experienced Searcher - Predictive Coding Manager
- Power User - Software Expert
- Project Managers
- Contract Reviewers
- Vendors (Software)

Planning the Predictive Coding Project

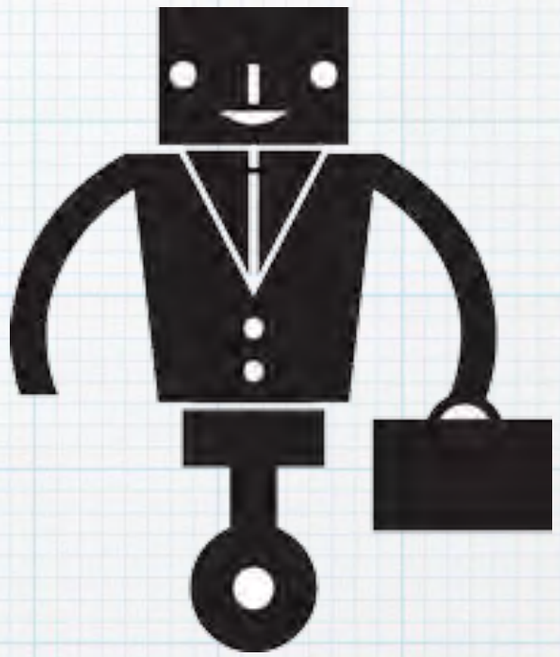
- Is Predictive Coding appropriate for the case?
- Customize plan to fit:
  - Case type, Issues
  - Size $$ and Data.
  - Deadlines.
  - Opposing Counsel.
  - Client.
  - Court - know your judges!
  - Type of Data to be reviewed.
- Articulate the goals of the review.
  - What are you looking for?
  - General recall range?
- One or Two Pass Review?

Planning the Predictive Coding Project

- Determining acceptable budget range using realistic ***Proportionality analysis*** and judge awareness.
- Cooperation Strategy and Disclosure:
  - Transparent, Translucent, Brick Wall
  - Sedona Principle 6
  - Irrelevant documents
- Pre-Review Culling: Non-Text, Files Types, Date Range, Custodians, Keywords, File Size, other.

Carrying Out a Predictive Coding Project

- Strategic selection of best documents for training the AI.
- Quality Control steps:
  - experienced personnel
  - known software with proven AI
  - small review team
  - limited coding issues
  - good horizontal and vertical communications
  - inconsistency detection and cure
  - mistake detection and cure
  - concept drift corrections

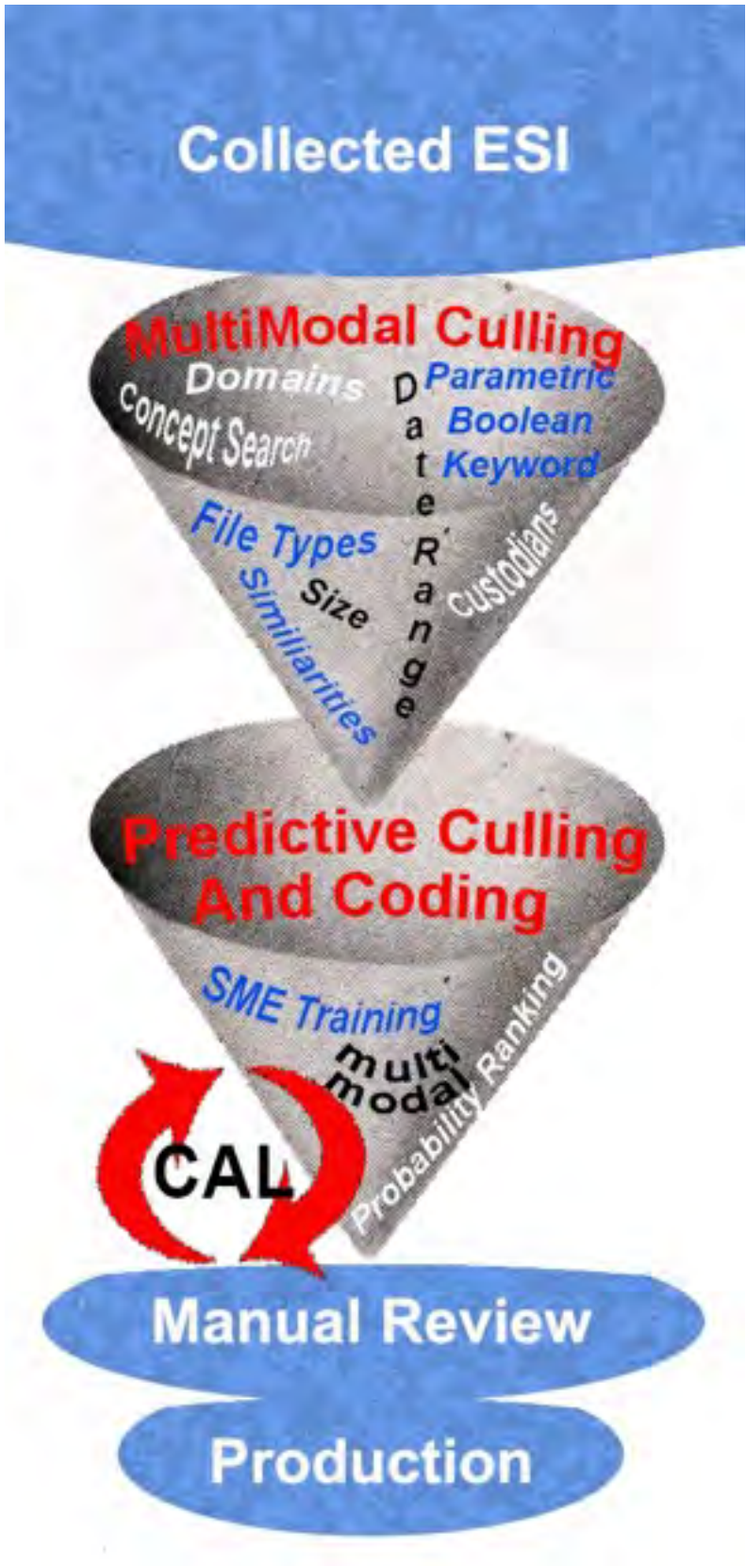


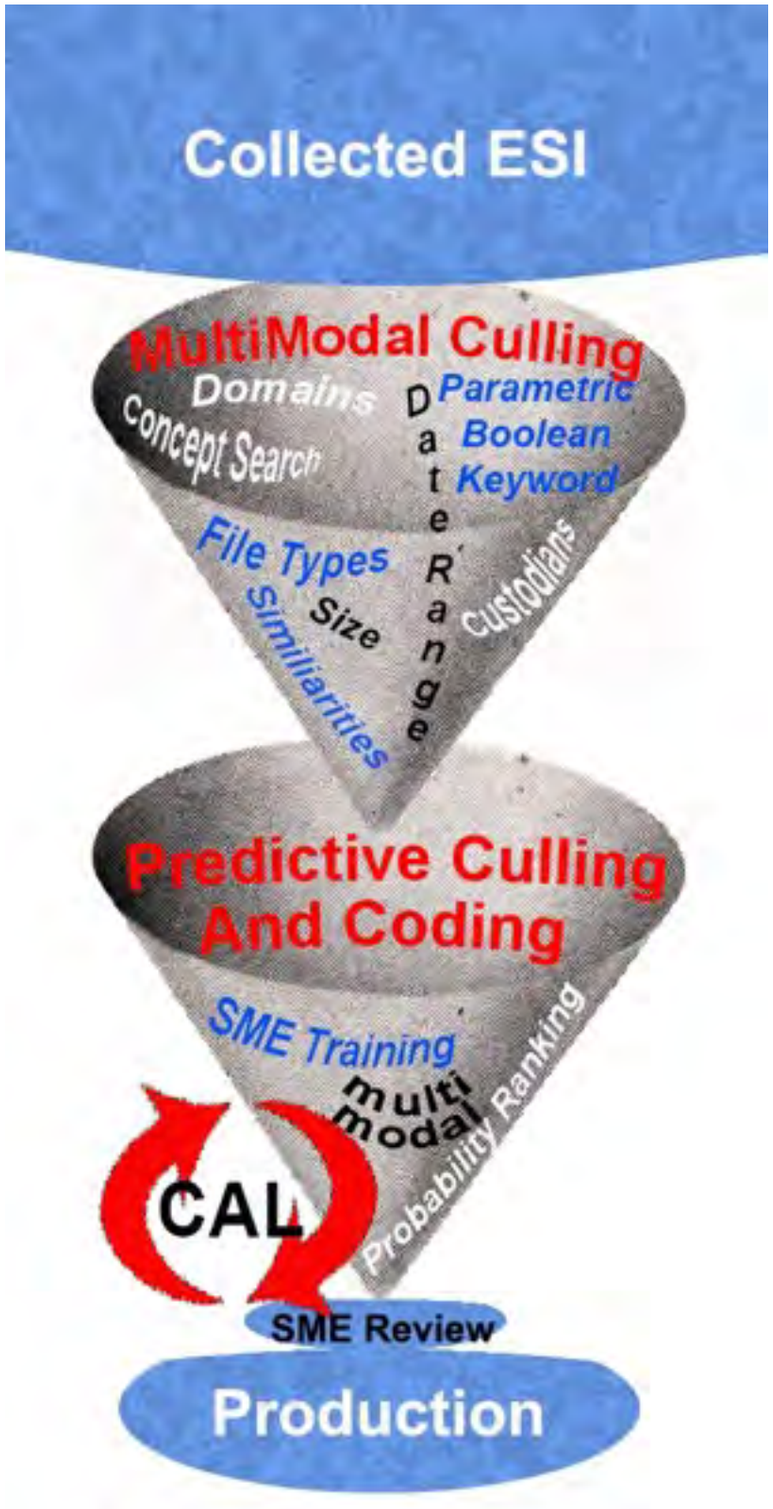
- Role of initial random sampling and ongoing recall metrics? (Prevalence based recall ranges.)
- Measuring the number and types of relevant documents found after each new round of training. (Mere Relevant, Strong Relevant, Highly relevant)
- Continuous Active Learning (CAL)
- Key metric of document ranking to track progress of predictive coding. (Ideal upside down champagne glass image showing goal of polar separation.)
- **Decision to Stop:** polarity has been attained, few new relevant documents are being found, approximate recall estimates, budgetary constraints.

Validating a Predictive Coding Project

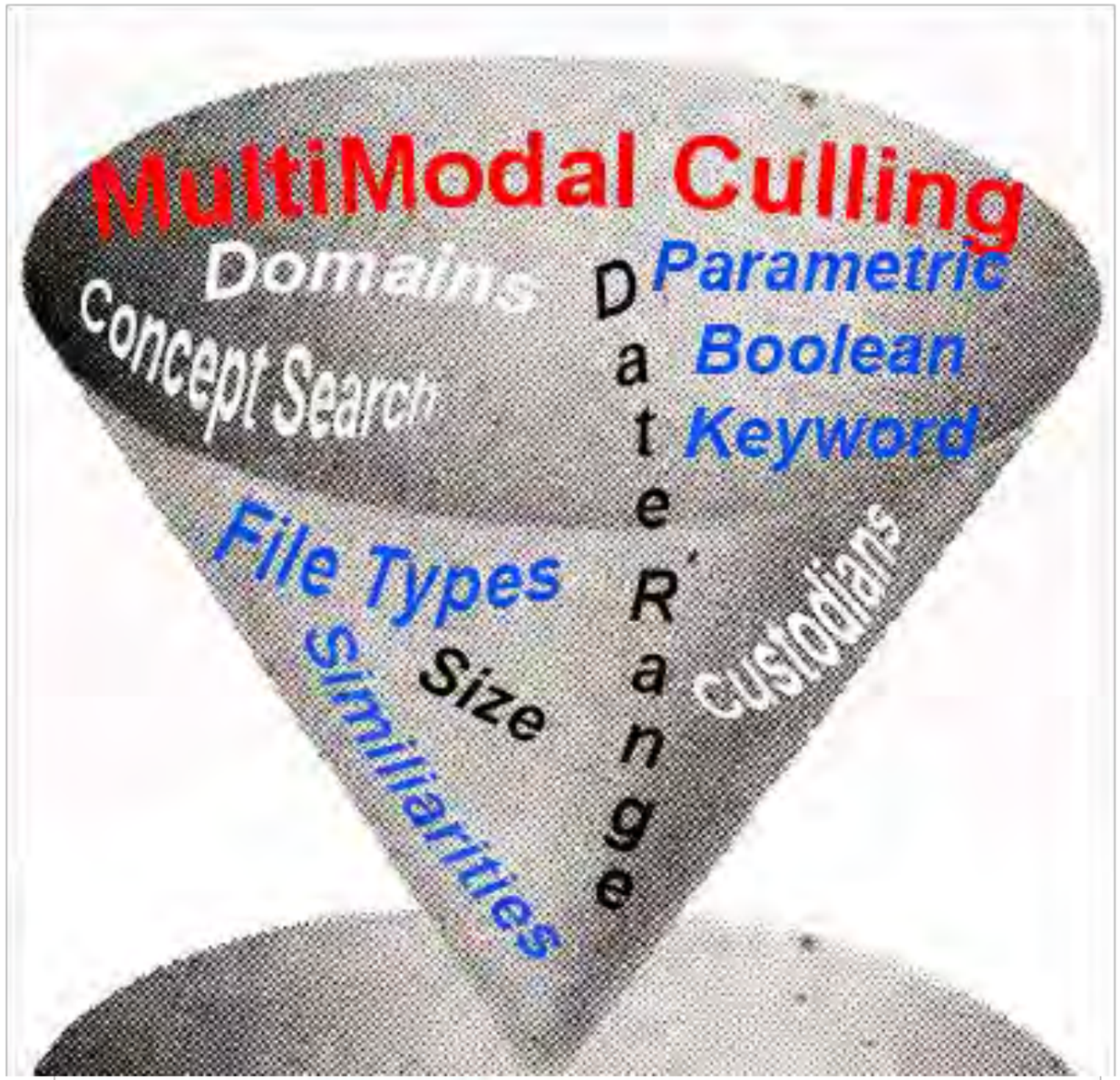
- Validating your decision to stop review by preplanned Quality Assurance tests
  - *ei-Recall*
  - accept of zero error
  - judgmental sampling
- Reporting metrics and end-project disclosures
- Keep good notes and be prepared to disclose with formal or informal reports.

Two Filter Culling

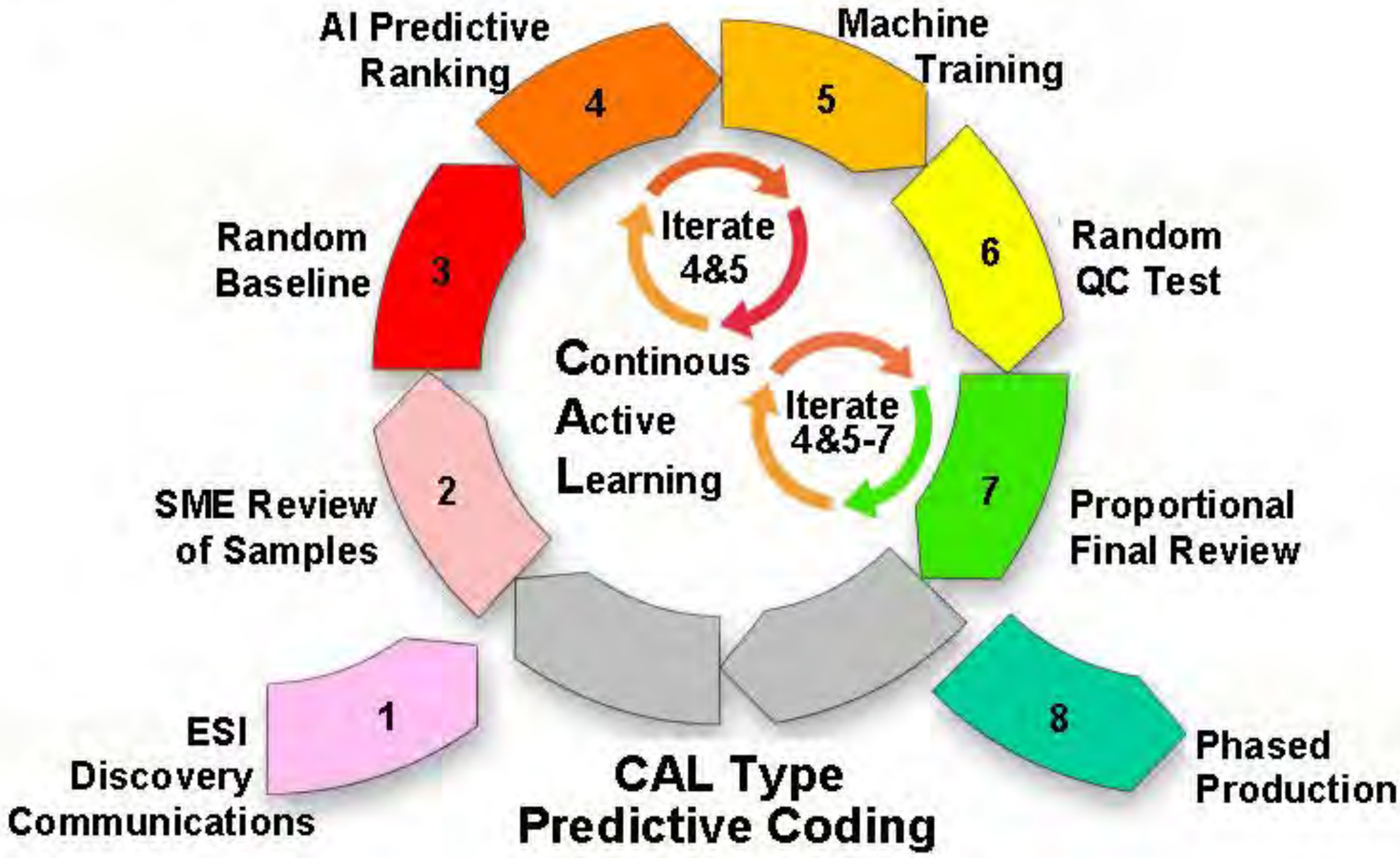




Two Filter Culling

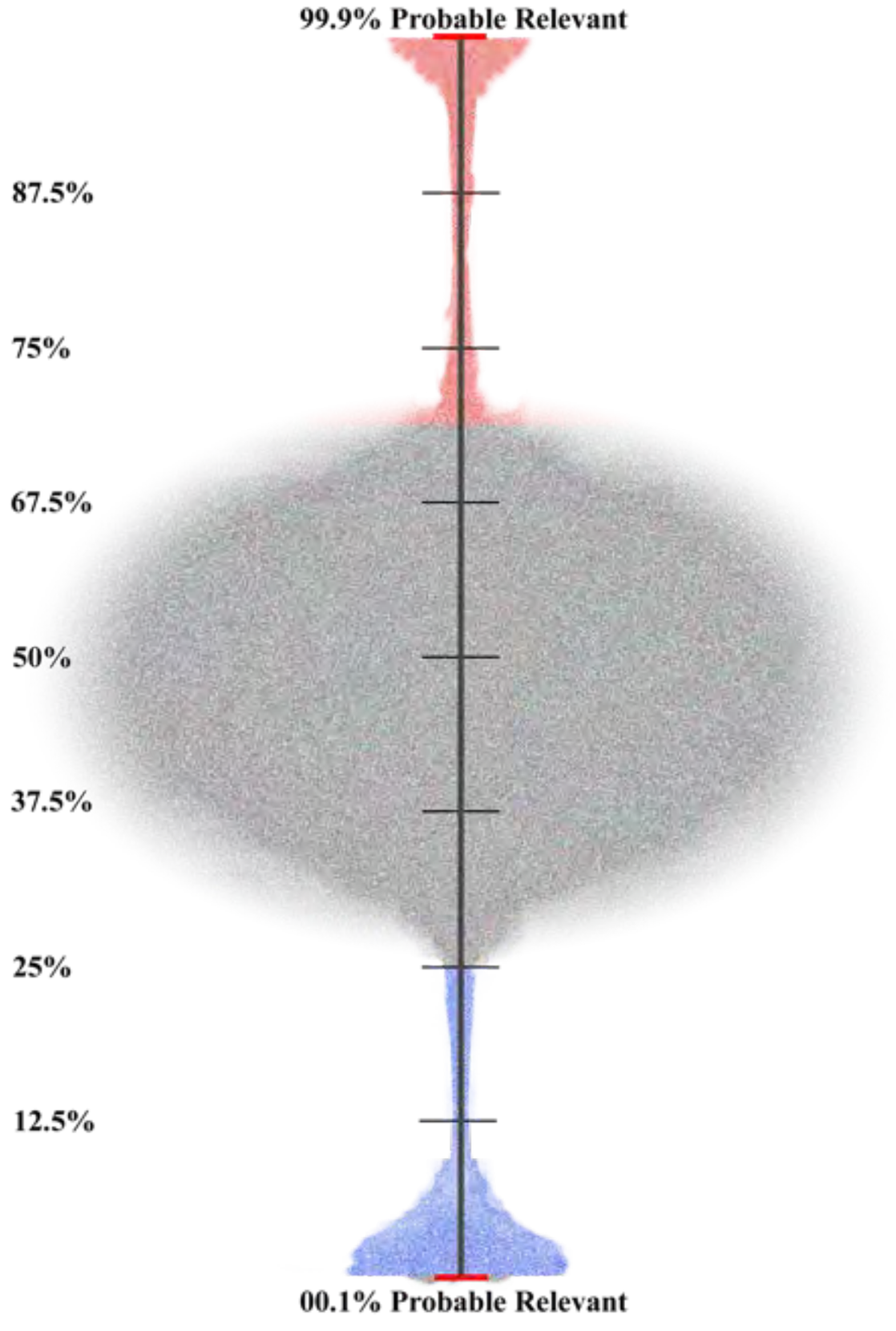


EDBP.com



How do you know when to stop training?

- Few new types of relevant files.
- No new strong relevant or highly relevant.
- Progression of probable relevance distribution:



ei-Recall

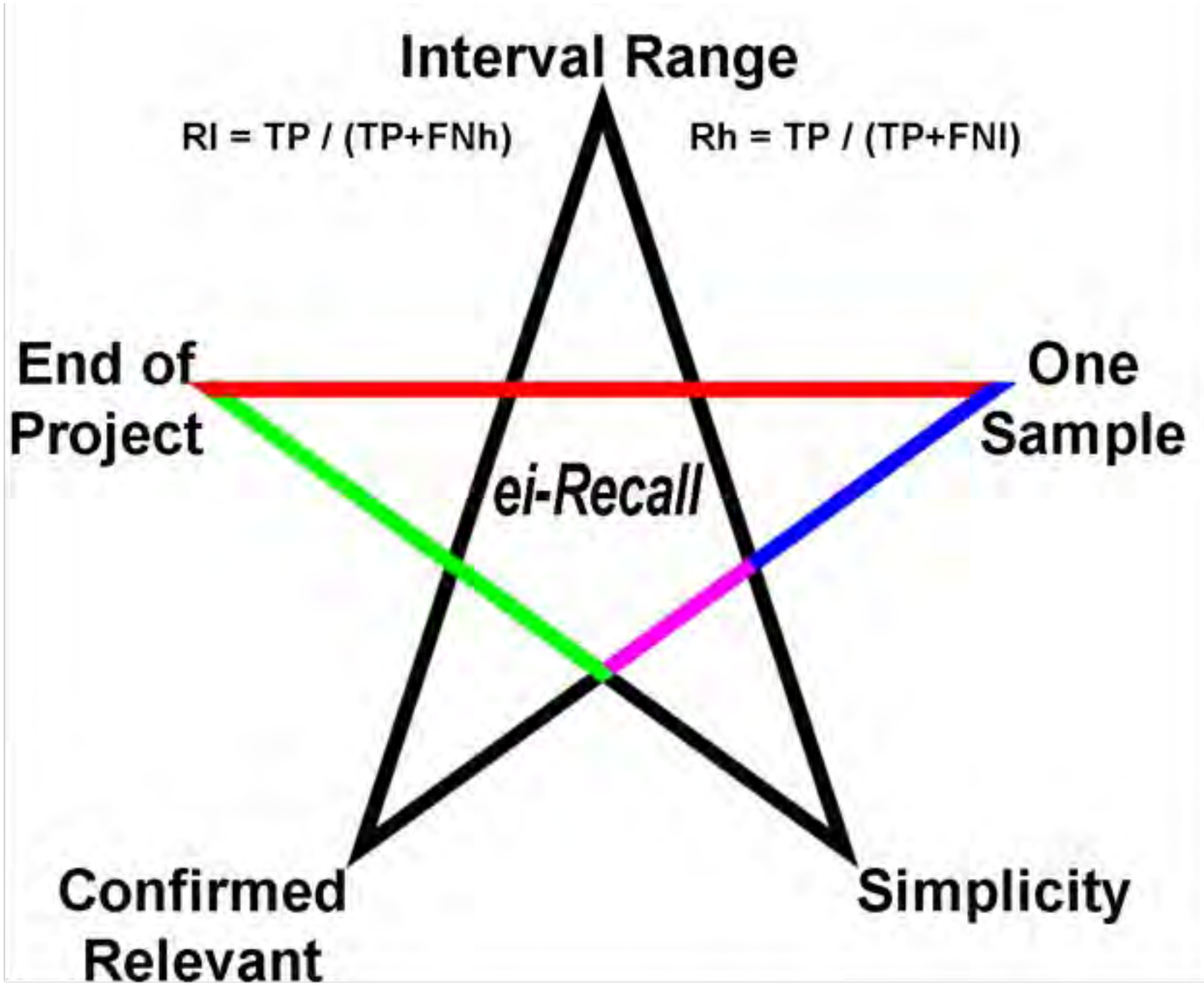
Based on Random Sample of Negatives.
Typically sample 1,500 (95% +/- 2.5%)

Formula for the low end of the recall range: **Rl = TP / (TP+FNh)**.
Formula for the high end of the recall range: **Rh = TP / (TP+FNl)**.

TP is the verified total number of relevant documents found in the course of the review project.
FNl is the low end of the False Negatives projection range based on the low end of the binomial confidence interval show by the random sample
FNh is the high end of the False Negatives projection range based on the high end of the binomial confidence interval shown by the same sample.

This formula essentially adds the extreme probability ranges to the standard formula for recall, which is: **R = TP / (TP+FN)**.

Advantages of *ei-Recall*

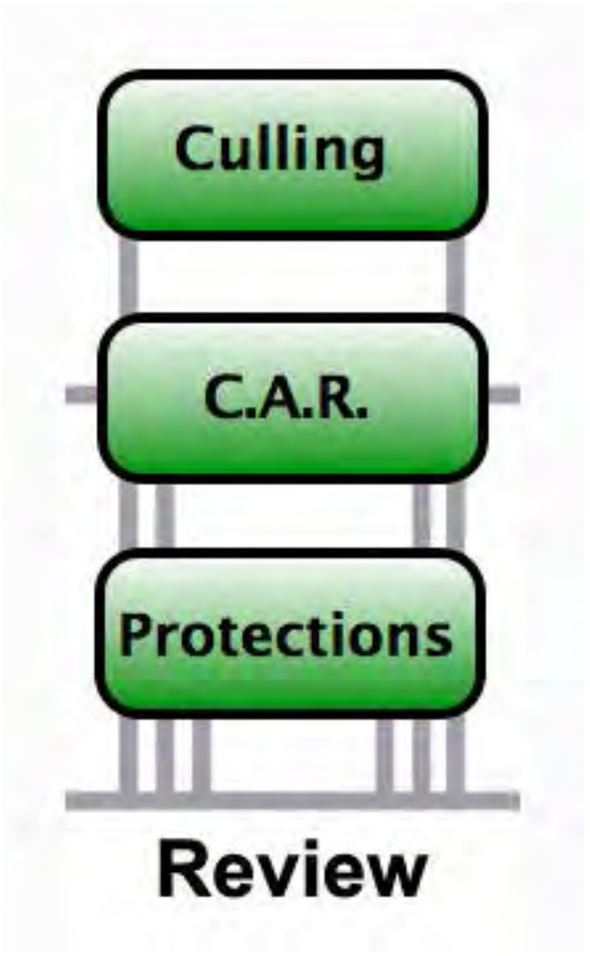


CASE STUDY - Technology Company

- **Over 1.5 Million document** review project
  - Poorly Defined Relevance Issues
  - Fourteen Days to Complete the Assignment
  - Experts on other side would challenge and test the search
  - Report and full disclosure required
  - 80% Recall Demanded
- **The Challenge: Review 1.5 Million Documents by Myself in 14 days**
- I happened to be an SME on the legal issues involved
- Consistency advantage of fewer reviewers:
  - 22% - best on record
  - 70%+ typical in multiple reviewer projects
- Used Continuous Active Learning and Multimodal
  - Heavy Reliance on High Ranking Documents for Training

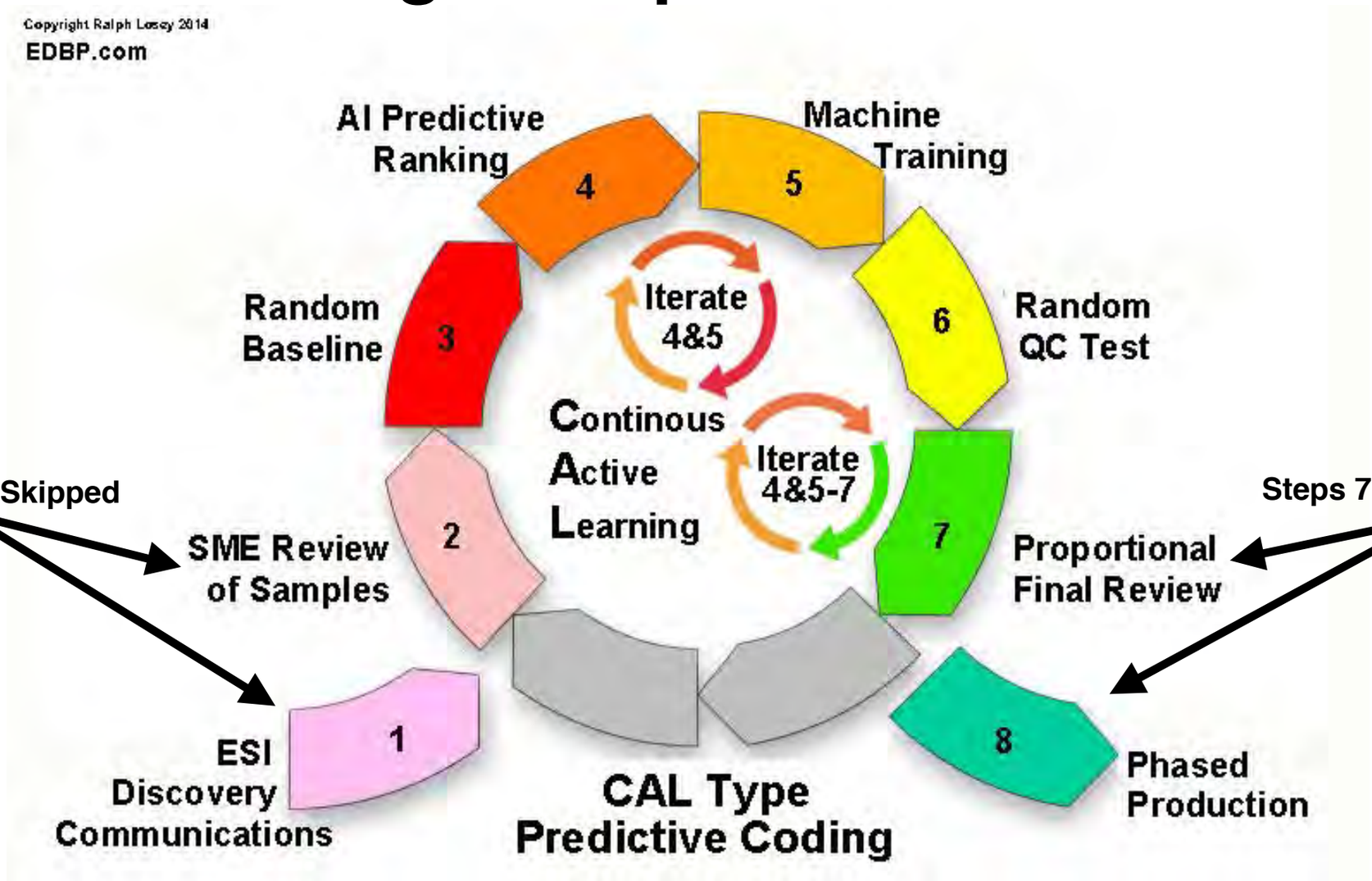
Project Overview

- **Predictive Coding Used for both Relevance and Privilege**
- **Focus on Relevance and Recall**
- **Heavy Reliance on Clawback Orders for Privilege**
  - Typical reasons a client might want to do that:
    - Money
    - Non-Litigation Production
    - Old data from another company
    - Other indications that privileged or confidential data was unlikely



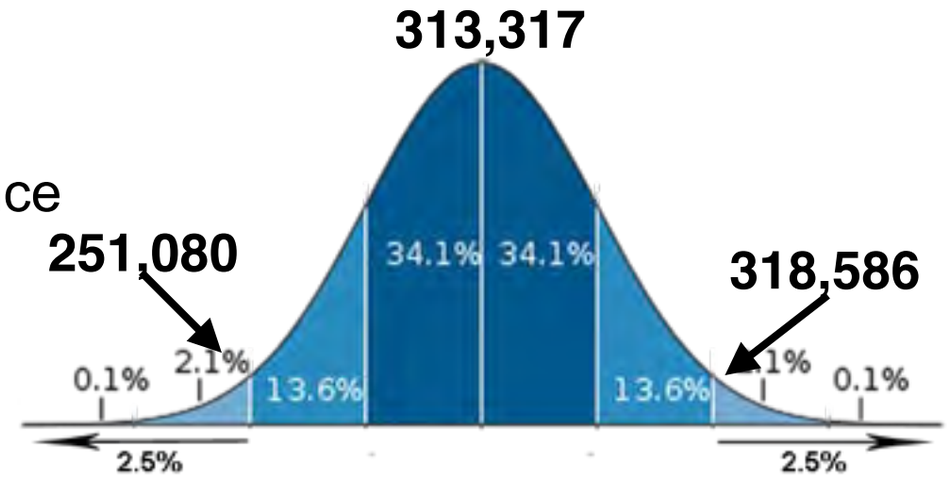
EDBP.com

Eight-Step Workflow



First Random Sample for Prevalence Calculation

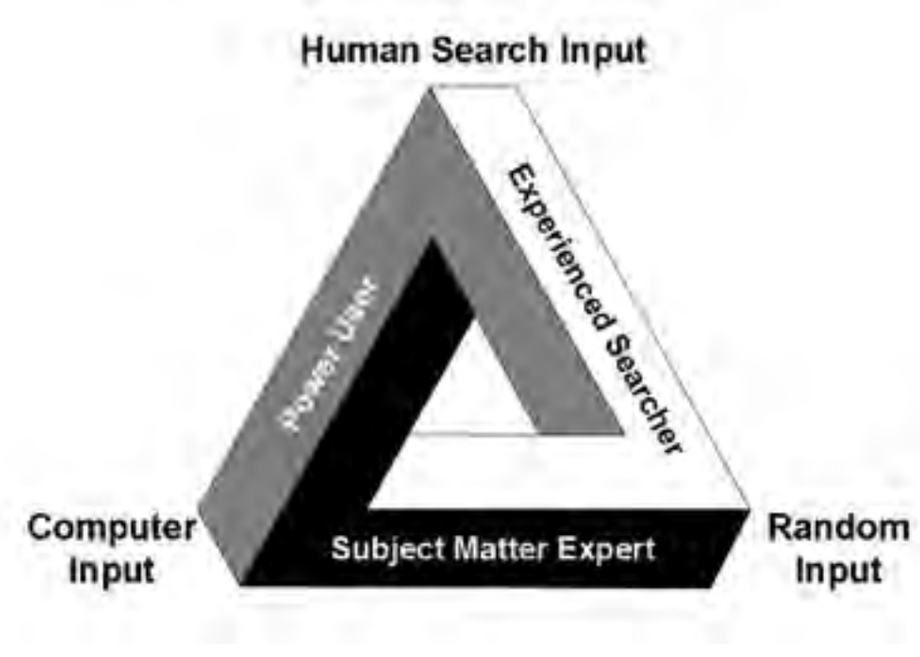
- First removed 75,832 documents from the initial machine learning dataset on various grounds, including size and non-text.
- Then took first simple Random Sample of remaining documents:
- Sample size - 1,705 (usually 1,535 for 2.5%)
  - 95% Confidence Level with 2.37% Confidence Interval. (**95% +/- 2.37%**)

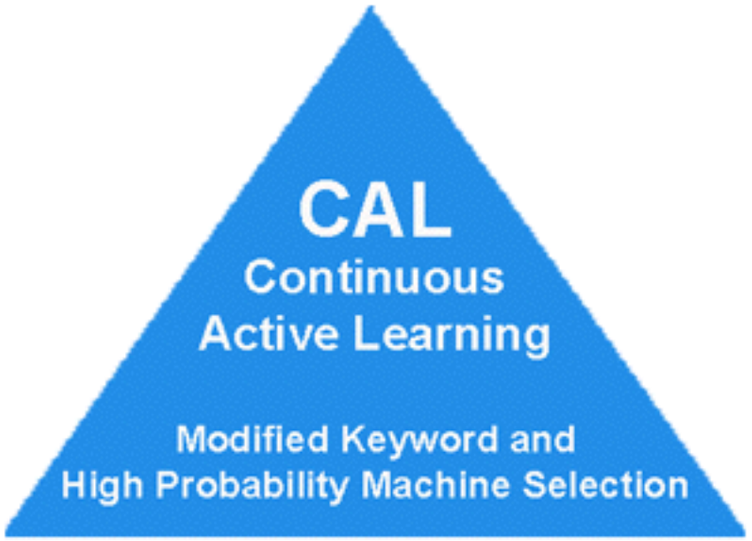


- Result of Review of 1,705 Sample
  - 375 Relevant.
  - 1,330 Irrelevant
  - 22% Prevalence with range of from 17.63% to 22.37%.
  - Spot Projection of 313,317 Relevant Documents in 1,424,168 document corpus.
  - High Low Range of 4.74%, so could be from **251,080** (17.63%) to **318,586** (24.37%) Relevant Documents.
  - The Prevalence metric is not a reliable basis to calculate Recall, but serves as adequate *general guide* to monitor progress

Metrics and Methods

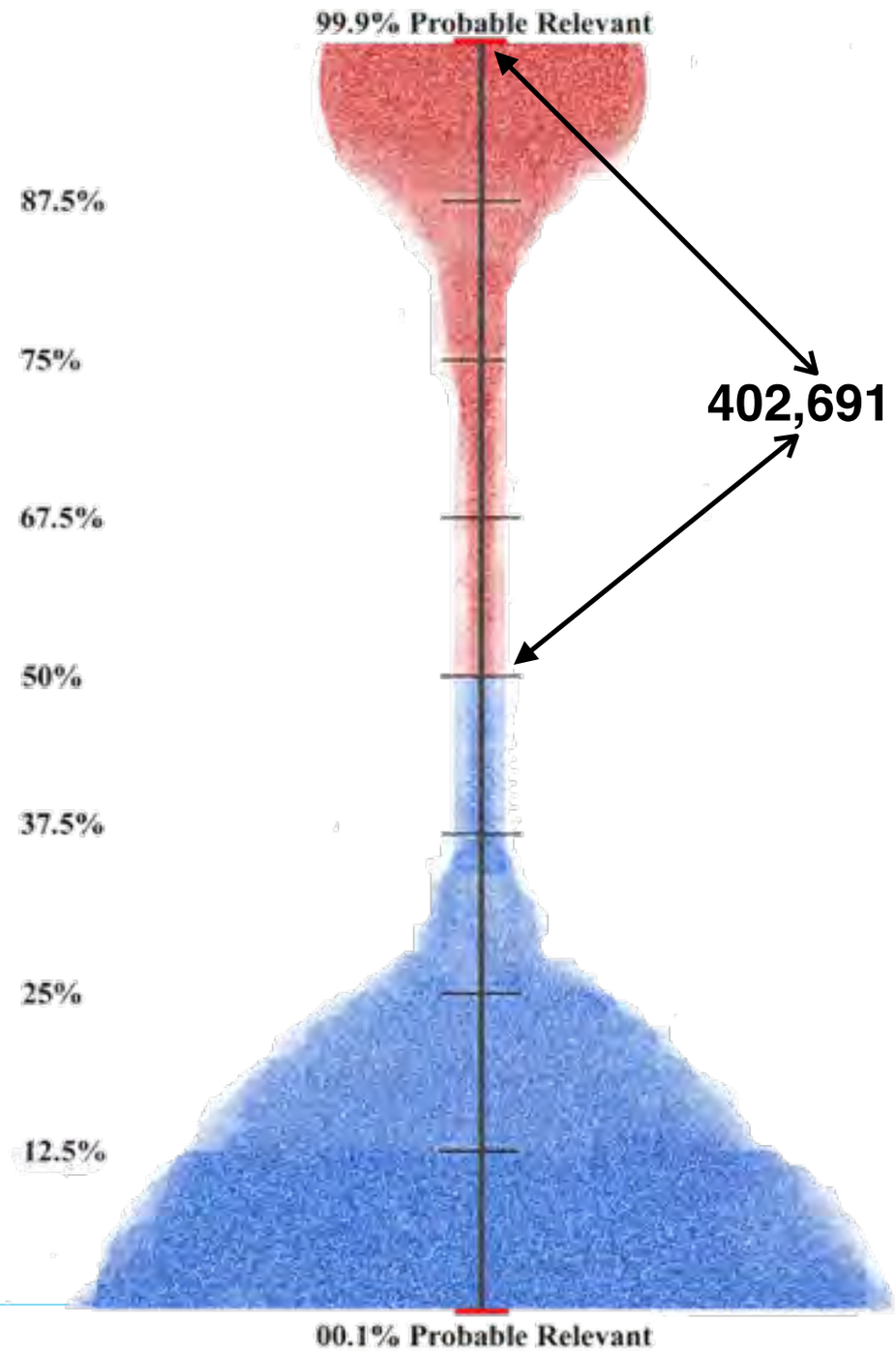
- After Ten Rounds of Multimodal Training with Focus on Relevance:
  - 318,552 documents coded relevant
  - Training used both Relevant and Irrelevant
- Used 3-Cylinder Search Engine Methodology
  - Random (least important),
  - Machine Selected (uncertain range)
  - Human Judgmental (most important)
- Heavy reliance in this particular CAL search on Keyword and Probability Machine Ranking.
  - Both are human judgmental type of searches.
  - Also used Hybrid modified *Presumed-Relevance* feedback approach.
- Completed Four More Rounds of Training (up to round 14).
  - Included QC type searches that focused on inconsistencies.





Metrics and Methods

- At Round 15 changed focus of search to Privilege; ran keywords on relevant docs only.
- Stopped at end of next Round 16.
  - **402,691 Predicted Relevant.**
    - Compared well with high end of prevalence range of **318,586**
    - **Probability distribution** looked good as per figure right
  - 2,528 of those were Predicted Privileged
  - 1,021,477 Predicted Irrelevant
- *I had personally reviewed only **5,124** documents - 3% of total.*
- Less than **3,379** of these documents were used to train the computer to categorize all 1.5 Million - **2% of total**.

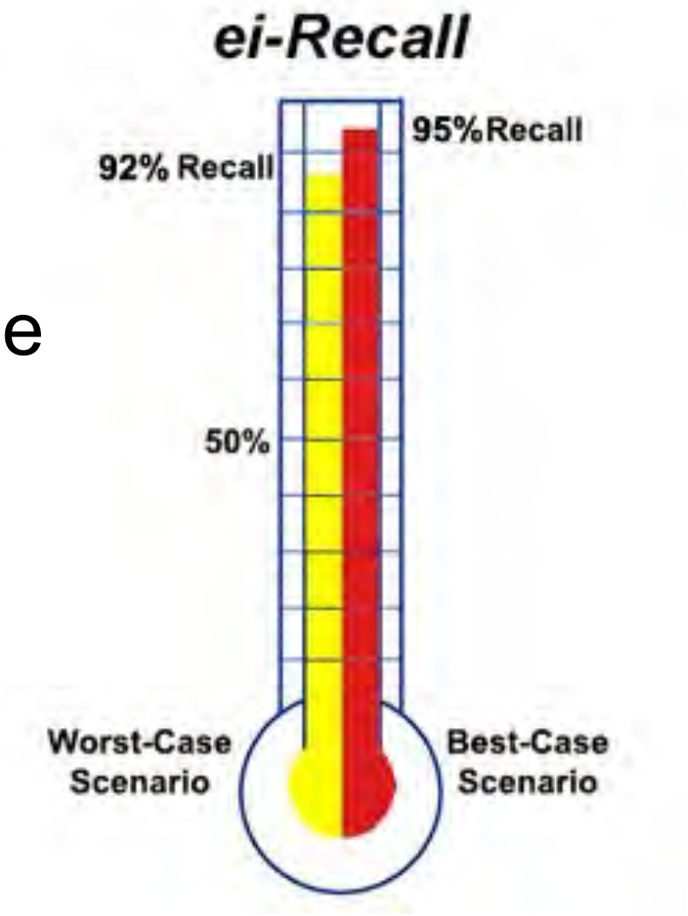


Metrics and Methods

- Sampled True Positives to Calculate **Precision**
  - 77% estimate
  - 92,619 False Positives (23%*402,691)
  - Only measured since this was one-pass review
  - Focus was always on Recall, not Precision nor F1
- Next Step - Final Random Sample Quality Assurance Test - ***ei-Recall***
  - *Random* Sample 1,535 from 1,021,477 Negatives
    - 95% +/- 2.5%
  - 41 False Negatives (FN) Found in Sample
    - Mistakes in auto-classification of Irrelevant
  - 2.67% error rate *point projection* (41/1535)
  - 1.92% - 3.61% *Binomial Interval Range*
  - Range of Total False Negatives
    - 36,875 - FN High (3.61%*1,021,477)
    - 19,612 - FN Low (1.92%*1,021,477)

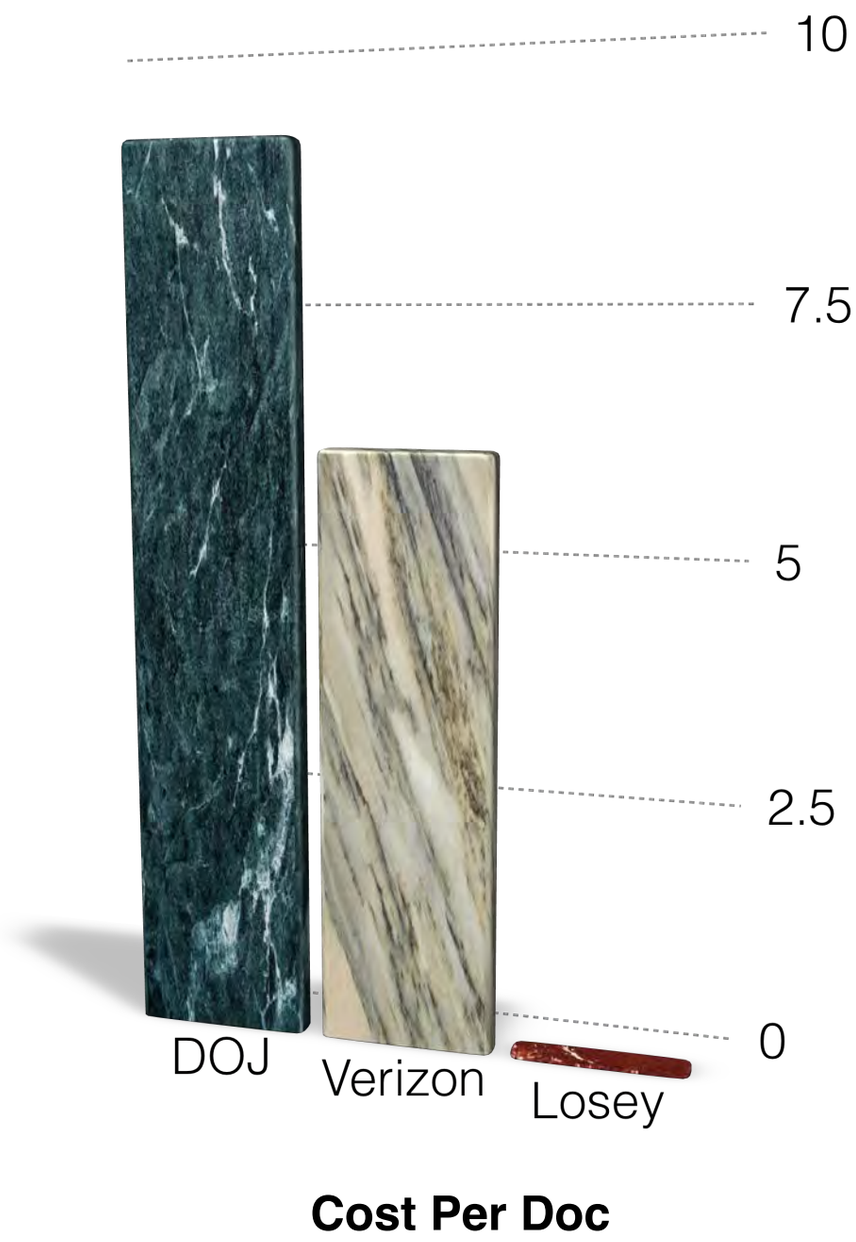
Metrics and Methods

- Recall Range
  - **91.6%** - Rl - Low End of Recall Range (402,691 / (402,691 + 36,875)
  - **95.4%** - Rh - High End of Recall Range (402,691/(402,691+19,612)
- No Hot Docs missed - *accept on zero error*
- Quality Assurance Tests passed
- Documents were Produced
- Client Happy
- No Complaints
- Facts modified somewhat to protect confidentiality
- This is one of my *best ever* results, with no *transaction costs.*



Final Analysis

- **Total Time: 64.5 Hours for One SME to review and classify 1.5 Million Documents in 14 Days.**
  - 45.75 hours document review.
  - 18.75 hours for analysis & report preparation.
- Assume $750.00 per hour rate; then attorney fee would be **$48,375**.
- Only **$0.03 per document** attorney review cost
  - Compare with DOJ using contact lawyers: **$9.09**. *Fannie Mae Securities Litig*., 552 F.3d 814, 817 (D.C. Cir. 2009) ($6,000,000/660,000 emails);
  - Verizon *2nd Review* Team of contract lawyers: **$6.09**. Roitblat, et al, *Document categorization*. J. Amer. Soc. Info. Sci. & Tech, 61(1):70–80, 2010 ($14,000,000 to review 2.3 million documents in four months)
  - Does not include e-Discovery vendor costs (Processing, Software Usage/Hosting), or privilege logging



Final Analysis of Exceptional One-Pass Project

- **Average Review Speed of Human Reviewer without AI assist: 75 files per hour**
- **Average Review Speed of AI Assisted Human Reviewer in this project**
  - **32,787 files per hour**
    - **45.75 hours of review time over two weeks divided by 1.5 Million files**
  - **AI-Enhanced Reviewer Speed is 437 Times Faster**
  - **AI and Robots are the Future!**

"The Future is already here.
It is just not evenly distributed yet."
William Gibson

QUESTIONS?

Internet Resources for Further Learning:

- LegalSearchScience.com
- e-DiscoveryTeam.com
- EDBP.com
- TREC Legal Track - trec-legal.umiacs.umd.edu/
- www.fclr.org/fclr/articles/html/2010/grossman.pdf
- PreSuit.com