Recent Developments in the Law & Technology Relating to Predictive Coding

Presented by

Paul Neale

CEO

Presented by

Gene Klimov

VP & Managing Director

Presented by

Gerard Britton

Managing Director

DOAR
LITIGATION CONSULTING

Discovery Obligations

- Proper Timing & Implementation of Litigation Hold
  - Identification of sources of ESI
  - Identification of custodians
  - Implementation of preservation protocols
  - Designation of point person
  - Signed acknowledgements by custodians
- Rule 26 Conference
  - Scheduling and Planning
  - Initial Disclosures
  - Meet-and-Confer
  - Privilege and Work-Product Protection
  - Form of Production

Meet and Confer

- Disclosure of sources of ESI
- Indication of ‘Not reasonably accessible sources’
  - Disaster Recovery Systems (Backup tapes/Archives)
  - Legacy Systems
  - Transient or Ephemeral data
- Proprietary systems and/or atypical ESI
- Culling methodologies
  - Keyword search terms, Advanced Analytics and/or Predictive Coding
- Protective order, clawback agreement & 502 protection
- Form of production

What is Predictive Coding?

Predictive Coding (aka TAR, CAR, CBAA) is the use of technology to assist in the categorization of documents to reduce the time and cost associated with document review and production.

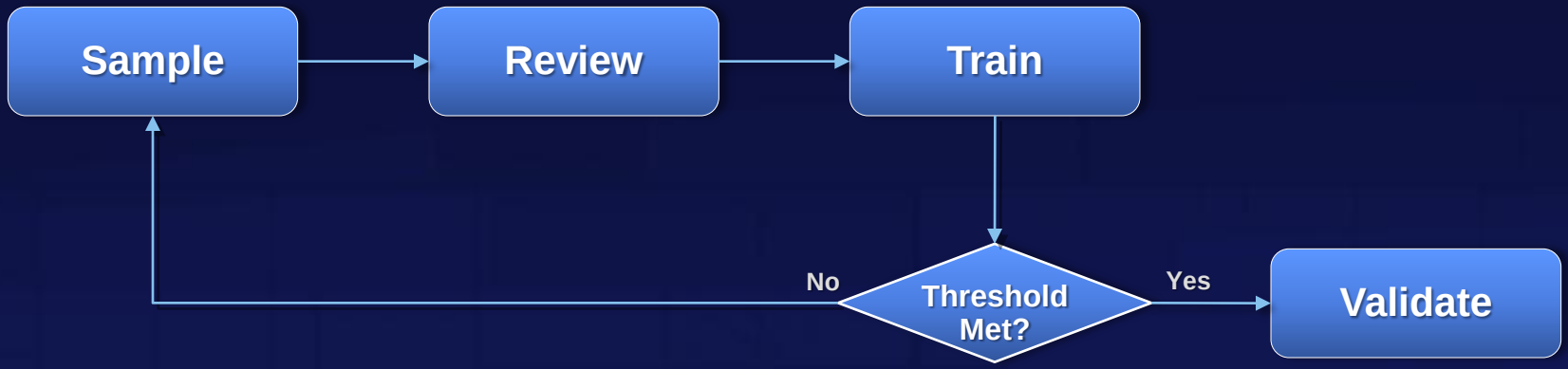
TAR – Technology Assisted Review

CAR – Computer Assisted Review

CBAA – Content Based Advanced Analytics

Predictive Coding Process

An iterative process that is designed to provide a sample set of documents to the human reviewer(s) that is most knowledgeable about the subject matter so that predictive coding software can learn from the reviewer and replicate the reviewer's judgment across the entire document population.

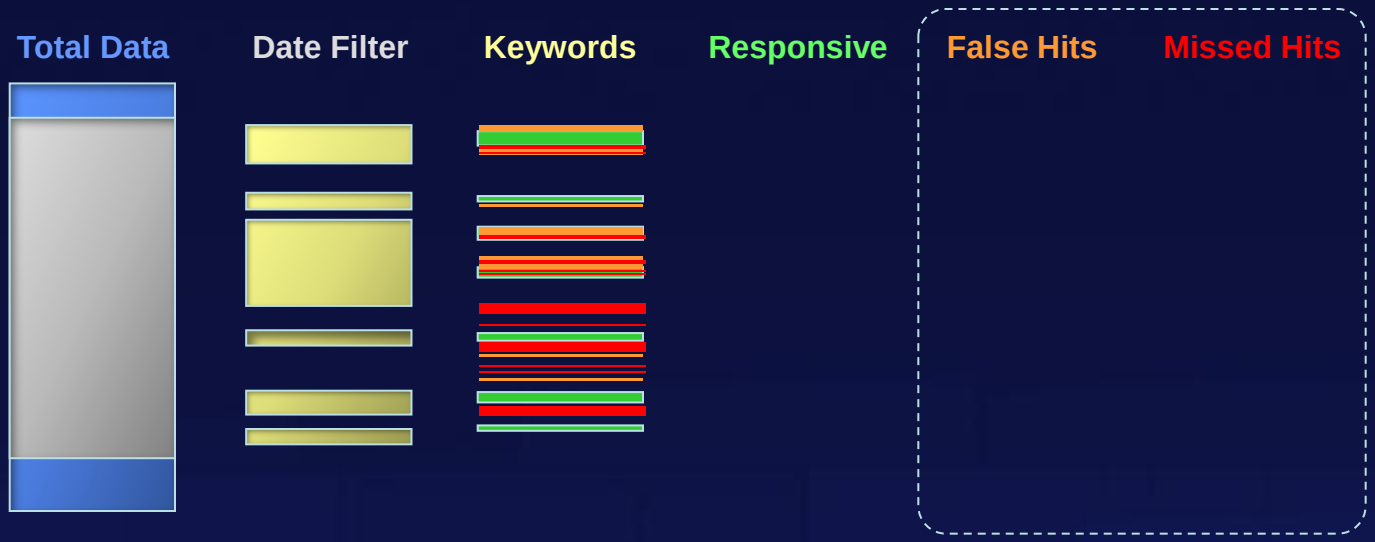


DOAR LITIGATION CONSULTING

How does this actually work?

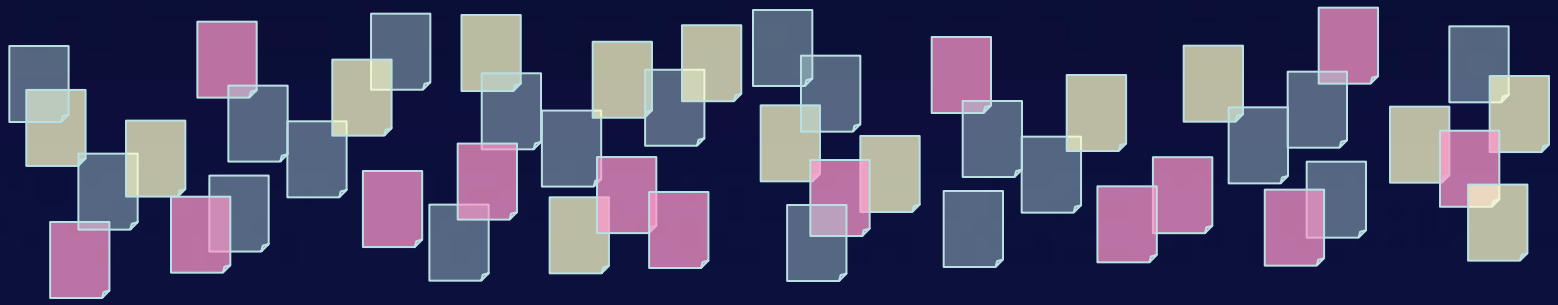
DOAR
LITIGATION CONSULTING

Traditional Culling



DOAR
LITIGATION CONSULTING

Traditional Review



Contract Attorney

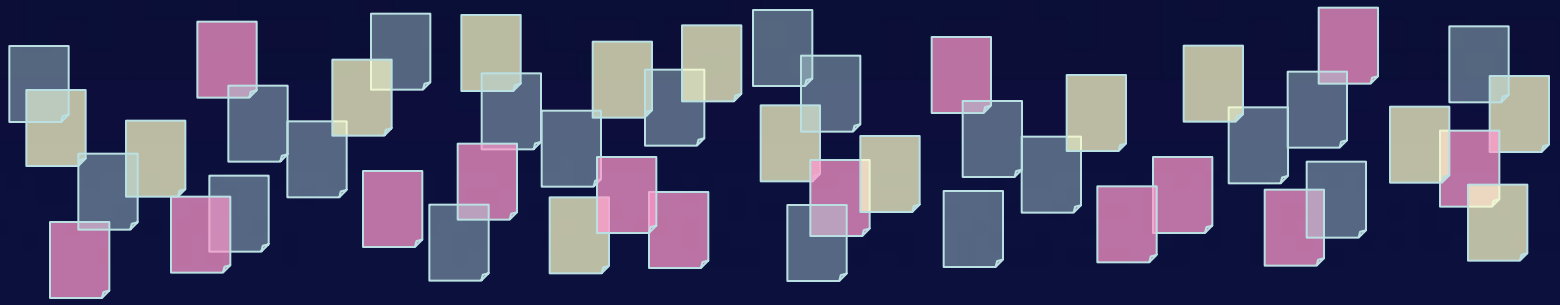
Contract Attorney

Contract Attorney

Contract Attorney

DOAR LITIGATION CONSULTING

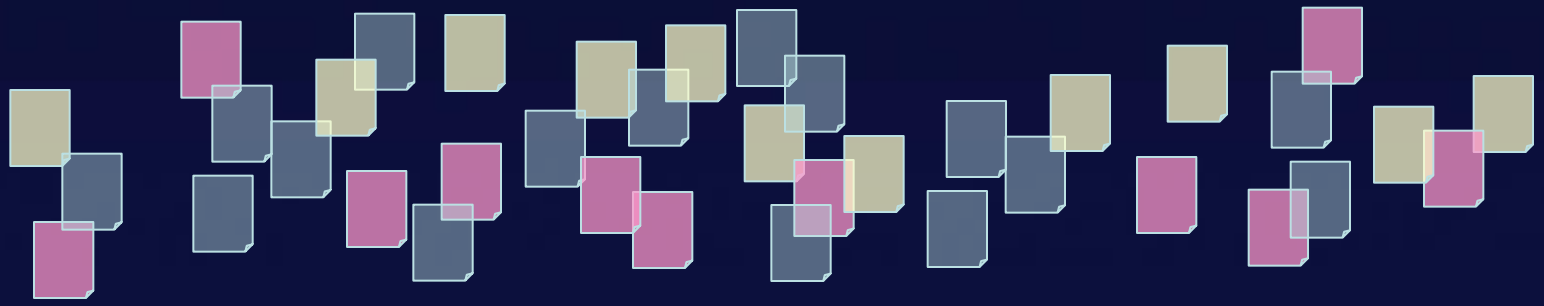
Training the System



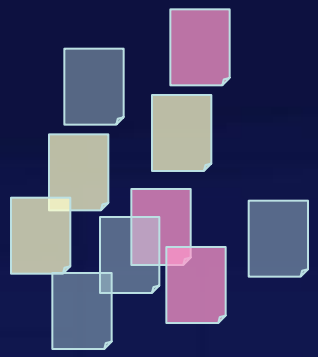
Senior
Attorney

DOAR
LITIGATION CONSULTING

Training the System



Irrelevant



Senior Attorney

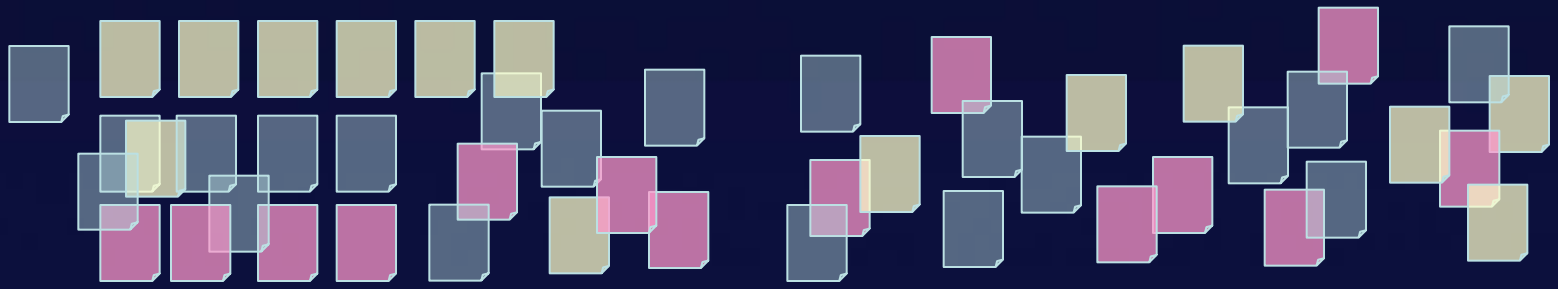
Responsive

Privileged

Train

DOAR LITIGATION CONSULTING

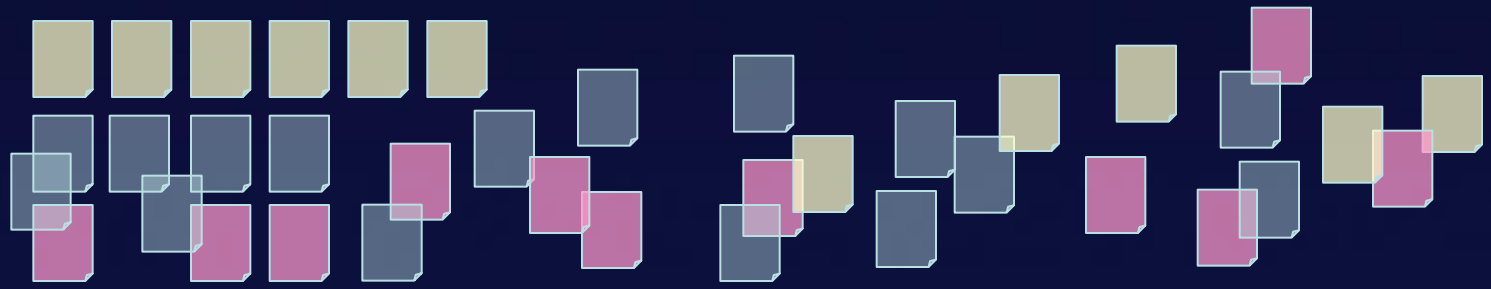
Training the System



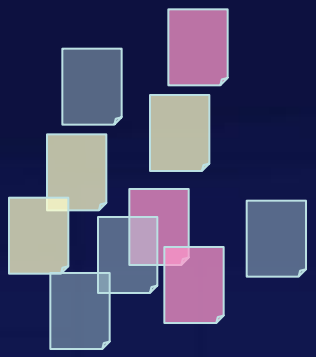
Senior
Attorney

DOAR
LITIGATION CONSULTING

Training the System



Irrelevant



Senior Attorney

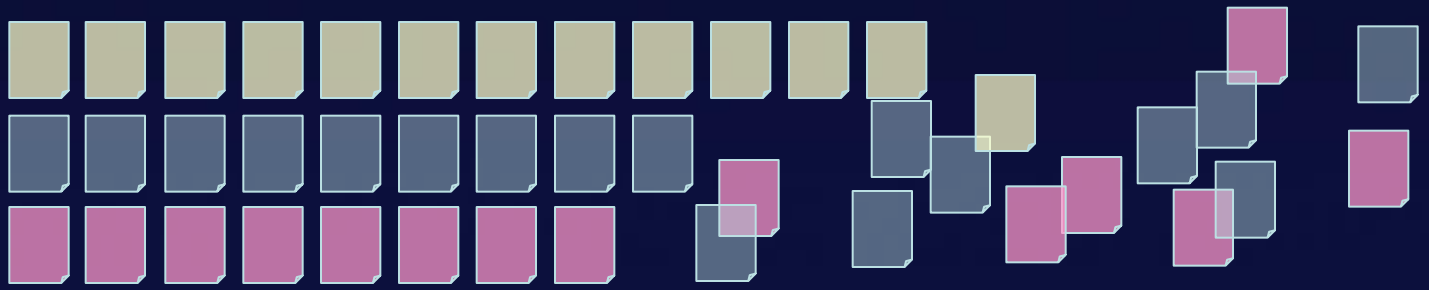
Responsive

Privileged

Train

DOAR LITIGATION CONSULTING

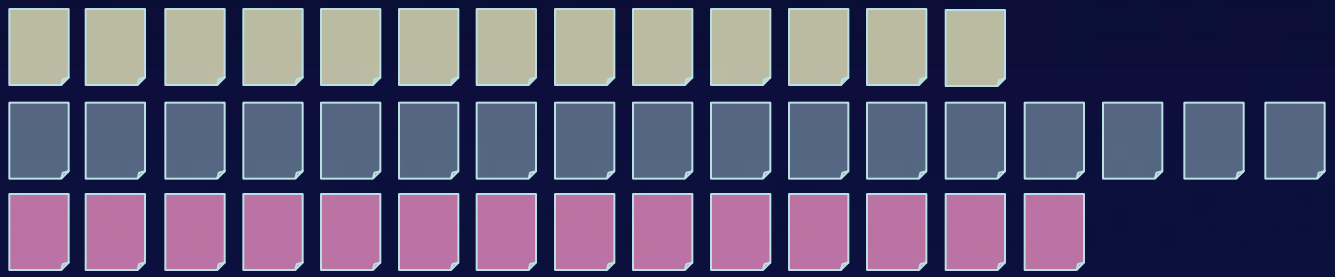
Training the System

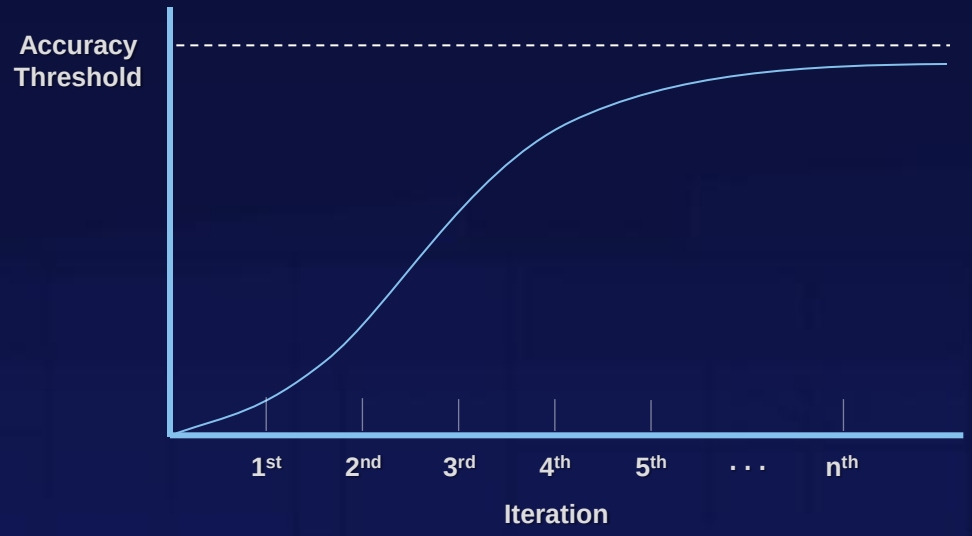


Senior
Attorney

DOAR
LITIGATION CONSULTING

Training the System – nth Iteration





Equilibrium is reached when the system predicts the same coding for documents as the human reviewer X% of the time.

X = Preset Threshold

DOAR LITIGATION CONSULTING

Industry Terminology

Yield Calculation

Initial random sample that is reviewed to determine a baseline for responsiveness. This is used to compare to overall system performance at the end of the process.

Seed Sets

Exemplar documents known to be responsive that are used to train the system.

Industry Terminology

Iterative Training

Random – Completely random selection of documents from the entire universe.
(NOTE: Data normalization is important)

Suggested – System derived document groups based on concepts and similar documents based on seed sets and previous iterations.

Training the System

Sample Calculation[1]

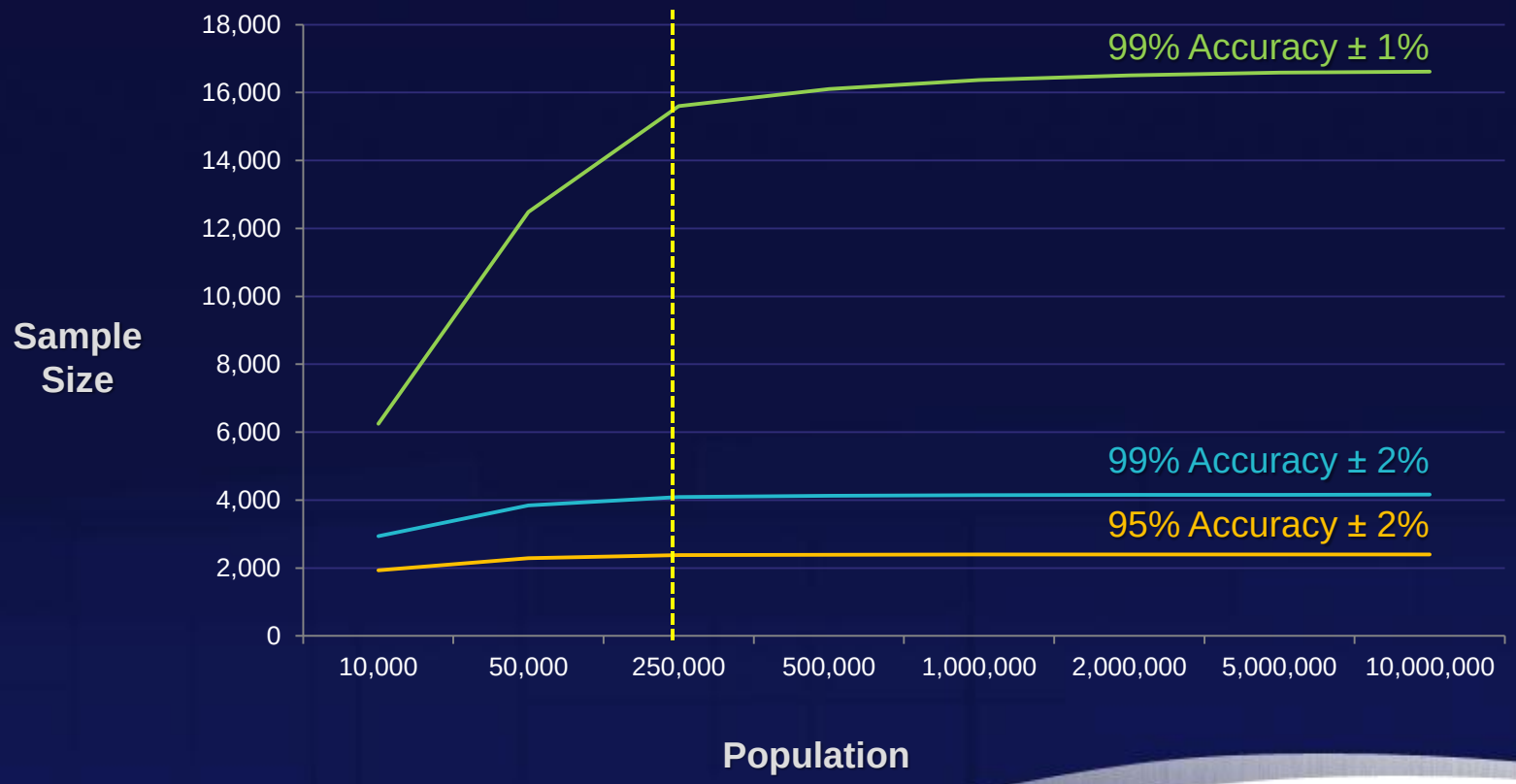
$$\text{Sample Size} = \frac{Z^2 \bullet p\,(1 - p)}{c^2}$$

Z = Z value (e.g. 1.96 for 95% confidence level)
p = percentage picking a choice, expressed as decimal (.5 used for sample size needed)
c = confidence interval (margin of error), expressed as decimal (e.g., .04 = ±4)

[1]http://www.surveysystem.com/sscalc.htm

Training the System

Sample Size v. Population



DOAR LITIGATION CONSULTING

Measuring Accuracy

The elements of a method's level of accuracy in the Information Retrieval context are Recall and Precision.

$$\textbf{Recall} = \frac{\text{Number of actually responsive documents retrieved}}{\text{Number of responsive documents overall}}$$

Out of the total number of responsive documents that exist in the population, how many did the process retrieve properly?

Measuring Accuracy

The elements of a method's level of accuracy in the Information Retrieval context are Recall and Precision.

$$\textbf{Precision} = \frac{\text{Number of responsive documents retrieved}}{\text{Number of documents retrieved}}$$

How accurate was the process in classifying responsiveness in what was retrieved?

Measuring Accuracy

F-Measure (or F_1 Score)

The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score.

$$\mathbf{F_1\ Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

A type of average that is calculated to determine how well a specific content retrieval method performs.

Measuring Accuracy

Example:

100 Relevant

900 Irrelevant

1,000 Documents

Calculations

System Retrieved

Relevant: 100
Irrelevant: 0

Recall = $\frac{100}{100}$ = 1.0

Precision = $\frac{100}{100}$ = 1.0

F_1 Score = $2 \cdot \frac{1.0 \cdot 1.0}{1.0 + 1.0}$ = 1.0

Measuring Accuracy

Example:

100 Relevant

900 Irrelevant

1,000 Documents

Calculations

System Retrieved

Relevant: 100
Irrelevant: 50

Recall = $\frac{100}{100}$ = 1.0

Precision = $\frac{100}{150}$ = 0.667

F_1 Score = $2 \bullet \frac{0.667 \bullet 1.0}{0.667 + 1.0}$ = 0.4

Measuring Accuracy

Example:

100 Relevant

900 Irrelevant

1,000 Documents

Calculations

System Retrieved

Relevant: 40
Irrelevant: 10

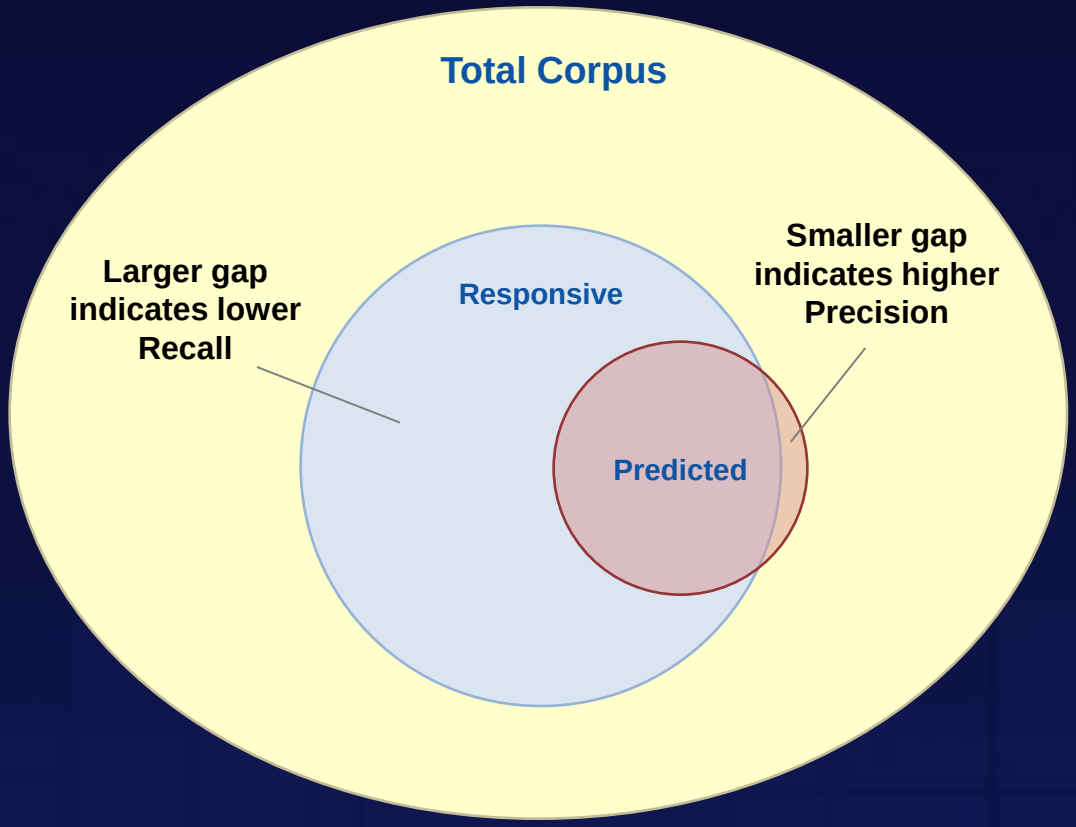
$$\textbf{Recall} = \frac{40}{100} = 0.4$$

$$\textbf{Precision} = \frac{40}{50} = 0.8$$

$$\textbf{F}_1\ \textbf{Score} = 2 \bullet \frac{0.8 \bullet 0.4}{0.8 + 0.4} = 0.46$$

DOAR LITIGATION CONSULTING

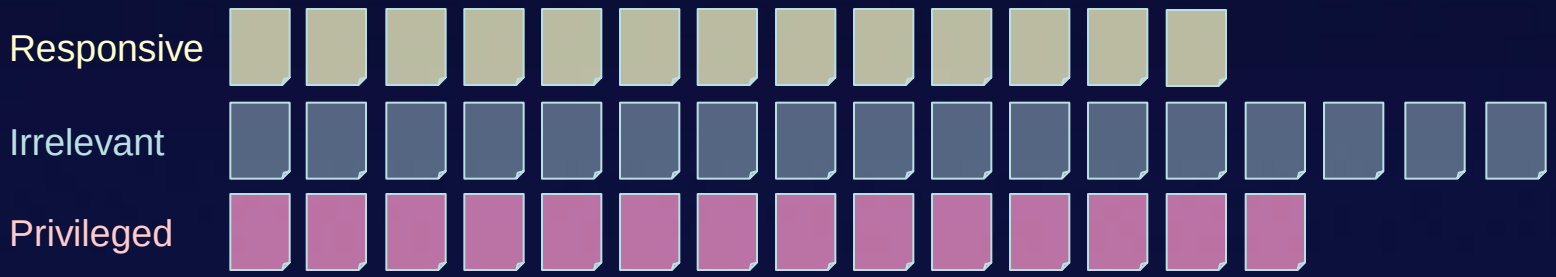
Measuring Accuracy



- Irrelevant Documents Correctly Identified as “Irrelevant”
- Relevant Documents Correctly Identified as “Relevant”
- Relevant Documents Incorrectly Identified as “Irrelevant”
- Irrelevant Documents Incorrectly Identified as “Relevant”

DOAR LITIGATION CONSULTING

Validation

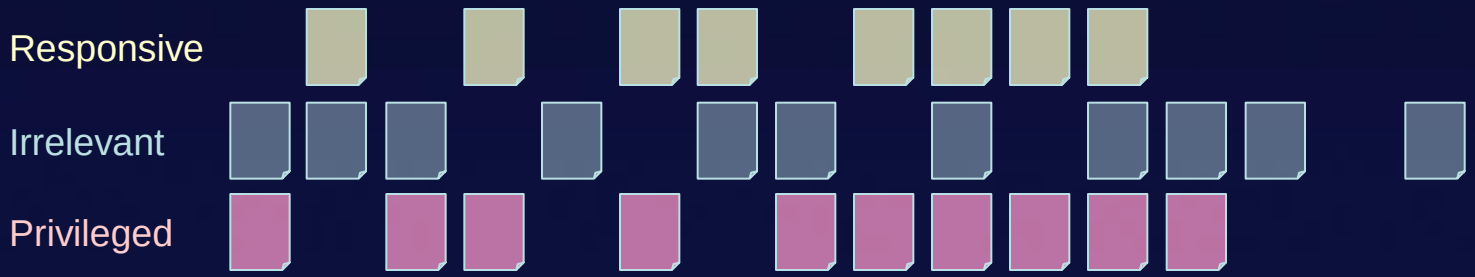


Random Sample

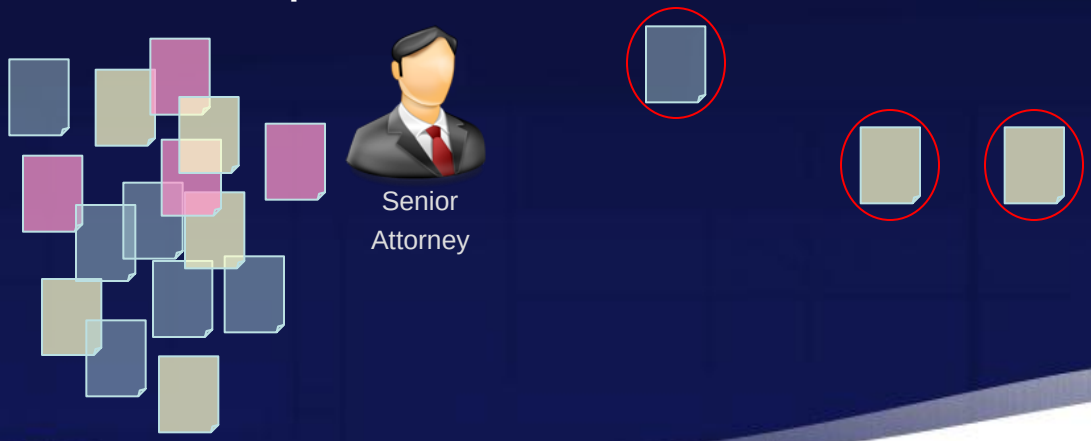
Senior
Attorney

DOAR
LITIGATION CONSULTING

Validation



Random Sample

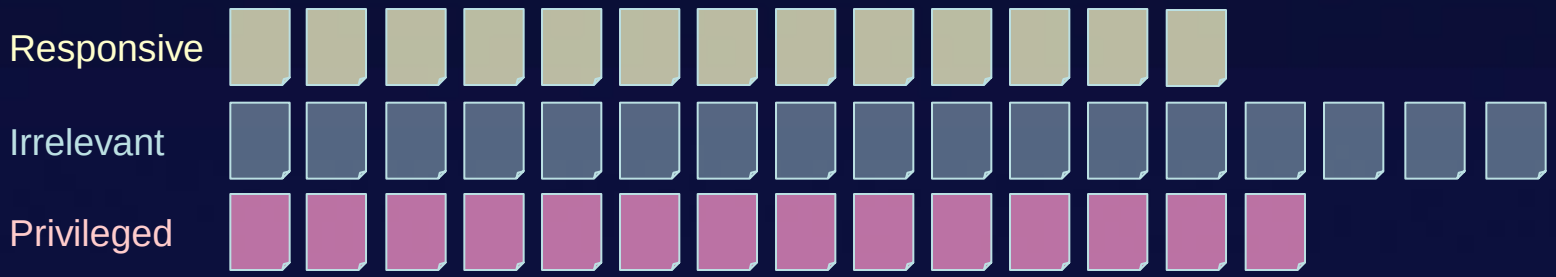


Recall & Precision

Want to have High Precision and Reasonable Recall

DOAR LITIGATION CONSULTING

Validation

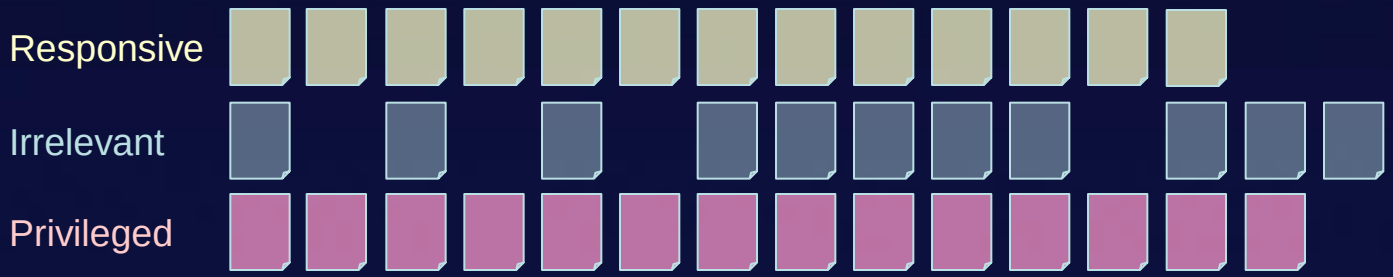


Random Sample

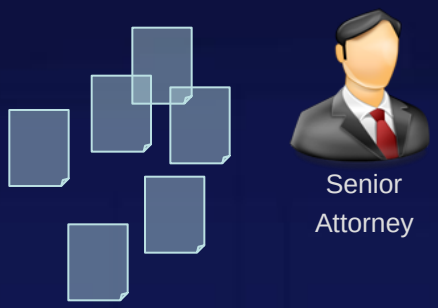
Senior
Attorney

DOAR
LITIGATION CONSULTING

Validation



Random Sample



Elusion

A zero acceptance test would require that NO responsive documents be found in the Irrelevant sample.

DOAR LITIGATION CONSULTING

System Validation

The final analysis of the result set predicted to be both responsive & non-responsive through measurement of a sampled set to ensure that the process meets the criteria for accuracy.

Recall & Precision

The calculated average between Recall & Precision that indicates that the vast majority of responsive documents were located accurately

Elusion

The percentage of the rejected documents that are actually responsive. This would allow both parties to choose what is an acceptable level.

Z-test

The comparison of responsive documents across a randomly sampled set at the start of the process compared to another randomly sampled set at the end of the process.

The Current State of Predictive Coding

- Rapidly evolving segment of the electronic discovery market
- Gaining momentum in law firms and corporations
- Fragmented market of providers
- ~~On the edge of judicial approval~~

Seminal Cases

DA SILVA MOORE, et al. v. PUBLICIS GROUPE PLC, et al. SDNY CASE NO. 11-CV-1279

- Joint ESI protocol filed by the parties (with objections)
- First opinion approving defendants' use of predictive coding issued by M. Judge Peck
- Objections filed by plaintiffs regarding the lack of measurability and inability to validate the process

Seminal Cases

KLEEN PRODUCTS LLC, et al. v. PACKAGING CORPORATION OF AMERICA, et al., ND ILL. CASE NO. 1:10-CV-5711

- Plaintiffs attempting to force defendants to use CBAA
- Multiple defendants with multiple ESI plans
- Hearing before M. Judge Nolan held on February 21, 2012 focusing on measurability and validation
- August 21, 2012 - Plaintiffs withdrew their demand

Seminal Cases

GLOBAL AEROSPACE INC., et al., v. LANDOW AVIATION, L.P., et al.
(Nos. CL 61040, CL 61991, CL 64475, CL 63795, CL 63190, CL 63575, CL 61909, CL 61712, 71633)

- Plaintiffs submitted motion for protective order to disallow the Defendant's use of Predictive Coding
- Defendant's protocol contained transparency and referenced precision, recall and steps for validation and measuring quality.
- Judge James H. Chamblin issued order approving the Defendant's use of Predictive Coding
- Case settled shortly thereafter

Seminal Cases

Actos (Pioglitazone) Products Liability Litigation, United States District Court for the Western District of Louisiana, MDL No. 2299

- Order issued embodying agreed upon protocol for a "Search Methodology Proof of Concept" process to test predictive coding output:
  - Random sample seed set
  - Required met and confer conferences to resolve issue
  - Iterative Process
  - Elusion ("Test the Rest") Validation
- Order issued
  - Case Management Order issued; includes predictive coding protocol

Seminal Cases

Federal Housing Finance Agency v. SG Americas, et al. (S.D.N.Y)

- JPMC and Plaintiffs Unable to Agree on Use of Predictive Coding
- Defendants Approached J. Cote Regarding Use of Predictive Coding (7/20/12)
- Draft Protocol Submitted by JPMC Proposing Use of PC After Search Terms
- Plaintiff Demanded Review of ALL Documents Predicted to be Non-Responsive
- Court Ordered All Parties to Meet and Confer to Attempt Agreement
- Parties Have Agreed to Seed Set and Continue to Negotiate PC Process

Seminal Cases

TSI, Incorporated v. Azbil BioVigilant, Inc. (D.C. AZ)

- Judge David Campbell Orders Party to Consider Predictive Coding
- Parties File Joint Submission (8/24/12)
- Heavy Reliance on Da Silva Moore
- Scope of ESI Too Small Due to Court-Imposed Discovery Limits
- Cost & Time Too Much with Predictive Coding
- Use of Search Terms More Effective

Key Considerations

- Lack of a definable standard
- Inconsistency of methodologies & technologies
- Use requires new levels of transparency by producing party
- Validation of results can only occur once costs are sunk
- Darwinian effect is inevitable
- Focus on validation of precision & recall


DOAR
LITIGATION CONSULTING