DESI V WORKSHOP 2013

[Similar Document Detection and Electronic Discovery: So Many Documents, So Little Time]

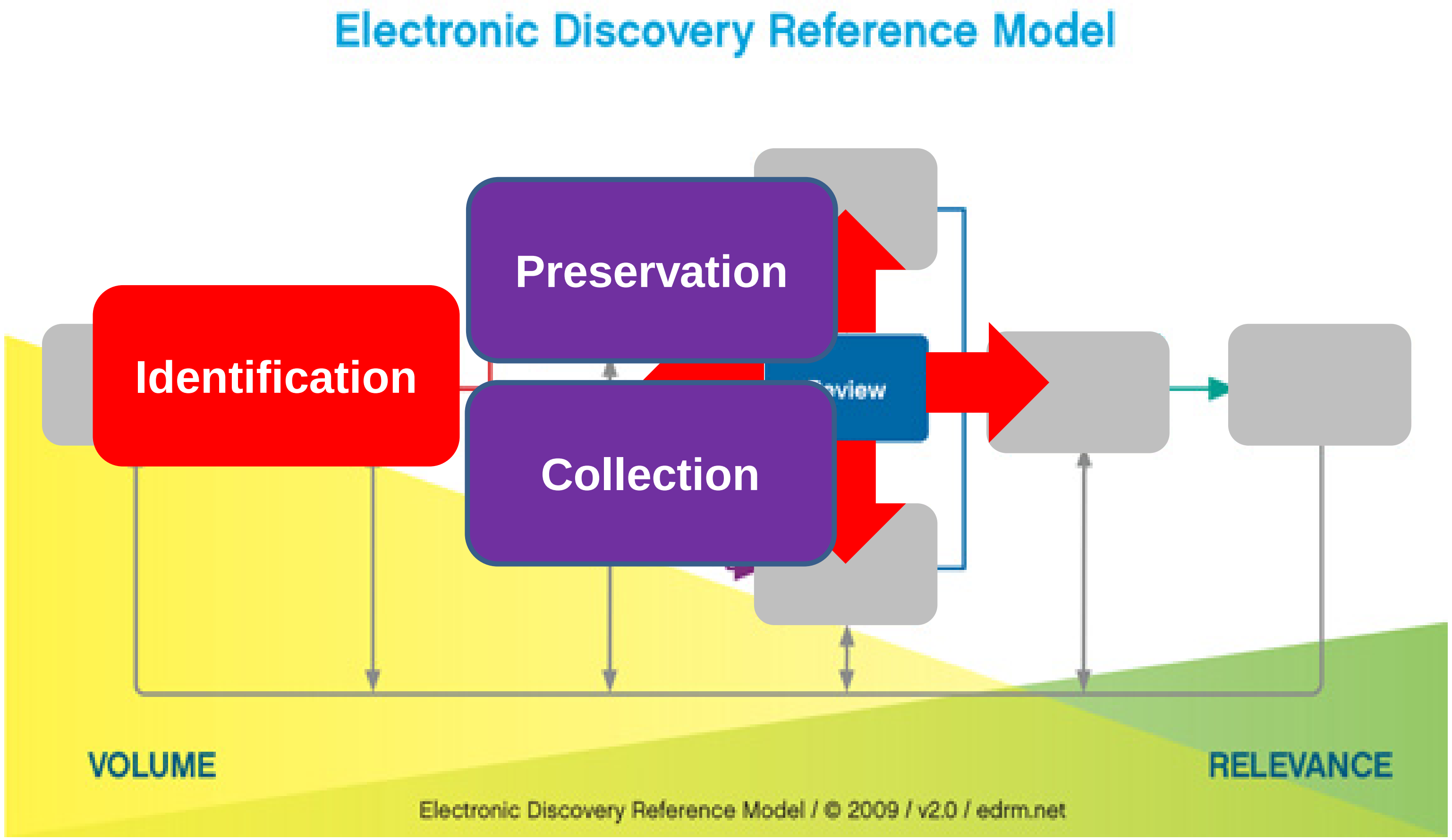
Michael Sperling, Rong Jin, Illya Rayvych, Jianghong Li and Jinfeng Yi

[Predictive Coding: Turning Knowledge into Power]

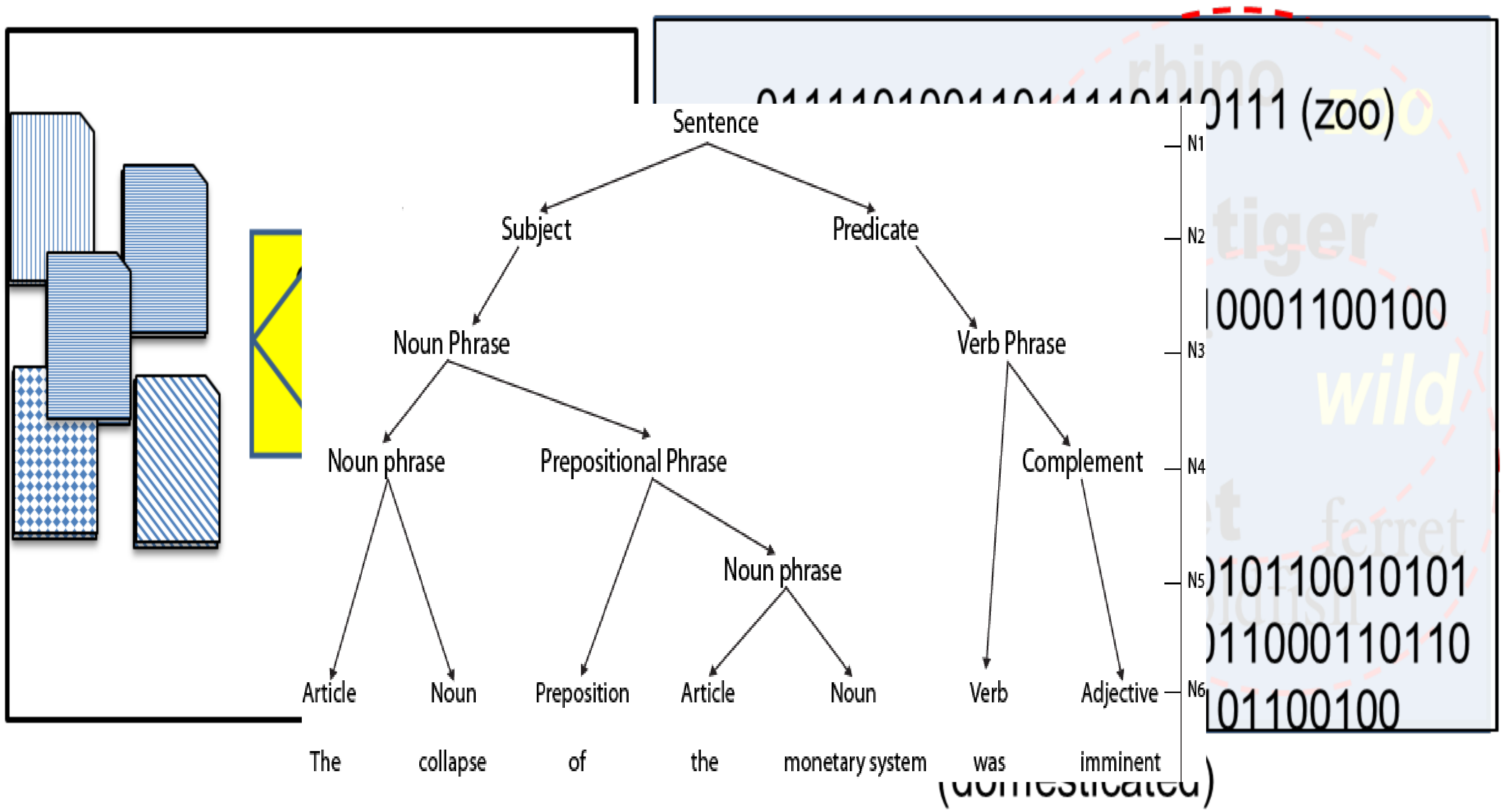
Thomas I. Barnett and Michael Sperling


Overarching theme:
EXPANSION
Process level
Analytical level

Process level expansion



Analytical level expansion



DESI V WORKSHOP 2013

Similar Document Detection and Electronic Discovery: So Many Documents, So Little Time

Michael Sperling, Rong Jin, Illya Rayvych, Jianghong Li and Jinfeng Yi

Predictive Coding: Turning Knowledge into Power

Thomas I. Barnett and Michael Sperling

הַמֵּבִין יָבִין

THEOREM 1. *Let* $\mathbf{u}$ *be a vector randomly sampled from* $\mathcal{N}(0, I/d)$. *With a probability* $1 - \delta - \frac{c \ln d}{d^3}$, *we have*

$$\sup_{\mathbf{x} \in \mathcal{D}(r,\mathbf{q})} |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}| \leq \frac{r}{\sqrt{d}} \left(C_1 \ln \frac{2m(r,\mathbf{q})}{\delta} + C_2 \sqrt{\ln \frac{2m(r,\mathbf{q})}{\delta}} \right)$$

where

$$C_1 = 6K_2, \quad C_2 = \sqrt{6K_2 + \frac{c \ln d}{d^2}} \qquad (1)$$

THEOREM 3. *Assume* m *is sufficiently large, i.e.,*

$$m \geq 64K_1 \left(C_1 \ln \frac{2}{\delta} + C_2 \sqrt{\ln \frac{2}{\delta}} \right)$$

where C_1 *and* C_2 *are defined in (1). Then, with a probability* $1 - (m+1)\delta - \frac{mc \ln d}{d^3}$, *we have*

$$\max_{1 \leq k \leq m} |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k| \geq \frac{|\mathbf{x} - \mathbf{q}|}{2\sqrt{d}}$$

COROLLARY 2. *Let* $\mathbf{u}_1, \ldots, \mathbf{u}_m$ *be* m *vectors randomly san pled from* $\mathcal{N}(0, I/d)$. *With a probability* $1 - m\delta - \frac{c \ln d}{d^3}$, *we have*

$$\sup_{\mathbf{x} \in \mathcal{D}(r,\mathbf{q})} \max_{1 \leq k \leq m} |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k| \leq \frac{r}{\sqrt{d}} \left(C_1 \ln \frac{2m(r,\mathbf{q})}{\delta} + C_2 \sqrt{\ln \frac{2m(r,\mathbf{q})}{\delta}} \right)$$

where C_1 *and* C_2 *are defined (1).*

THEOREM 4. *Let* $\mathbf{u} = \frac{1}{\sqrt{d}}(u_1, \ldots, u_d)$ *be a random vector with* u_i *drawn from a Bernoulli distribution* $\Pr(u_i = 1) = \Pr(u_i = -1) = 1/2$. *Then, with a probability* $1 - \delta$, *for a fixed data point* $\mathbf{x}$, *we have*

$$\sup_{\mathbf{x} \in \mathcal{D}(r,\mathbf{q})} |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}| \leq \frac{r}{\sqrt{d}} \left(2 \ln \frac{2m(r,\mathbf{q})}{\delta} + \sqrt{2 \ln \frac{2m(r,\mathbf{q})}{\delta}} \right)$$

THEOREM 5. *Let* $U = \frac{1}{\sqrt{d}}(\mathbf{u}_1, \ldots, \mathbf{u}_m)$ *be random variables with* $U_{i,j}$ *having equal probability to be* $+1$ *and* -1. *With a probability at least* $1 - 2m/[d^3]$, *we have*

$$\sup_{\mathbf{x} \in \mathcal{D}(r,\mathbf{q})} \max_{1 \leq k \leq m} |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k| \leq \frac{r}{\sqrt{d}} \left(2 \ln \frac{2m(r,\mathbf{q})}{\delta} + \sqrt{2 \ln \frac{2m(r,\mathbf{q})}{\delta}} \right)$$

When m *is sufficiently large, i.e.,*

$$m \geq 64K_1 \left(2 \ln \frac{2}{\delta} + \sqrt{2 \ln \frac{2}{\delta}} \right)$$

Then, with a probability $1 - (m+1)\delta$, *we have*

$$\max_{1 \leq k \leq m} |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k| \geq \frac{|\mathbf{x} - \mathbf{q}|}{2\sqrt{d}}$$

THEOREM 6. *(Talagrand's inequality) Let* $X_1, \ldots, X_m$ *be independent random variables in* $\mathcal{X}$. *For any class of functions* $\mathcal{F}$ *on* $\mathcal{X}$ *that is uniformly bounded by a constant* $U > 0$ *and for all* $\delta > 0$, *with a probability* $1 - \delta$, *we have*

$$\left| \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n f(X_i) \right| - \mathrm{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n f(X_i) \right| \right| \leq K_1 U \ln \frac{K_1}{\delta} + \sqrt{K_1 \sigma^2 \ln \frac{K_1}{\delta}}$$

where K_1 *is an universal constant and* σ^2 *is defined as*

$$\sigma^2 = \mathrm{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n f^2(X_i)$$

PROOF. Since

$$\max_{1 \leq k \leq m} |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k| \geq \sqrt{\frac{1}{m} \sum_{k=1}^m |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k|^2}$$

it is sufficient to bound $\frac{1}{m} \sum_{k=1}^m |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k|^2$. Using the Telegrand inequality, we have

$$\sup_{\mathbf{x} \in \mathcal{D}} \left| \frac{1}{m} \sum_{k=1}^m |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k|^2 - \mathrm{E}\left[\frac{1}{m} \sum_{k=1}^m |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k|^2 \right] \right| \leq K_1 U \ln \frac{K_1}{\delta} + \sqrt{K_1 \sigma^2 \ln \frac{K_1}{\delta}}$$

where

$$U = \sup_{\mathbf{x} \in \mathcal{D}} \sup_{1 \leq k \leq m} |(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k|^2$$

$$\sigma^2 = \mathrm{E} \sup_{\mathbf{x} \in \mathcal{D}} \sum_{k=1}^m \frac{|(\mathbf{x} - \mathbf{q})^\top \mathbf{u}_k|^4}{|\mathbf{x} - \mathbf{q}|^4}$$

Current Approaches

Two Basic Elements:

1. Vector representation of document (e.g., *n*-grams, vector space model)
2. Mapping vector representation to perform search

The Problem

- Inefficiency
  - Costly in compute time and storage (due to *heavy* representation of documents)
  - Slower than desired processing time
- Lack of flexibility
  - Static model for data flow doesn't match real world
  - Static centroid document doesn't allow adaptation to specific data set characteristics

Issues with Static Clustering

- ✓ Well Separated Document Clusters
  - – A well separated document cluster is a set of documents such that any document in a cluster is closer to every other document in the cluster than to any point not in the cluster.
  - – Challenges
    - Diversity of document population
      - – Individual documents are not highly focused
    - Documents arrive in waves
      - – Adding to cluster with closest centroid degrades clusters
- ✓ Threshold for "similarity" cannot be dynamically adjusted – *it's set at cluster creation*

Why Similar Doc Detection in a world of Predictive Coding?

- ✓ Combining analytical approaches can improve results in appropriate cases
- ✓ Quality control of training set
  - – Check for consistency of responsive and nonresponsive Are any near duplicates of responsive documents tagged as non-responsive?
  - – Especially important when multiple reviewers are independently tagging training docs
  - – In our case, 312 docs in training set violated this constraint. Retraining without them significantly improved model

Why Similar Doc Detection in a world of Predictive Coding?

- ✓ Highlighting subtle changes between documents, especially drafts (Examples from Enron corpus)
  - Predictive coding will not pick up these differences
  - Terms of contract:
    - *with the first such installment being due and payable upon the issuance and activation of the initial password and user ID*
    - *with the first such installment being due and payable* ***within five business days*** *after issuance* ***or*** *activation of the initial password and user ID*
  - Comments on Electricity Competition and Reliability Act
    - **Initial draft** – *Cinergy violated East Central Area Reliability Coordination Agreement by improperly drawing power it did not own from the interchange to meet its own supply obligations*
    - **Final document** - *Cinergy* ***apparently*** *violated East Central Area Reliability Coordination Agreement by improperly drawing power it did not own from the interchange to meet its own supply obligations*

Requirements

- ✓ Minimal resource consumption
  - Lightweight representation – storage conservation
  - Rapid preprocessing – no delay in making documents available for review within total processing time and
  - Almost instantaneous retrieval of near duplicates – reviewers are the most expensive resource
- ✓ Accuracy – high recall and precision
- ✓ Dynamically vary "near" threshold : "nearness"
- ✓ Requirement varies with different doc populations
- ✓ Deal properly with new docs – doc arrival not controllable: need to analyze entire corpus, not just new wave

Our approach

- ✓ Lightweight document representation – 62 tuple vector for counts of Capitals, Lowercase, and Numerals + total character count + vector length
- ✓ Dynamic search for similar documents, rather than static clusters (*short-form* vector)
  - Implemented as a sequence of one-dimension range searches
  - Use random projections to reduce vector dimensionality
  - Verify retrieved documents at end using 62 tuple representation
- ✓ We prove mathematically and show experimentally the soundness of this approach

Experimental Results

- ✓ Corpus
  - 13,228,105 documents drawn from an actual e-discovery project
  - Contained diverse content typical of e-discovery
- ✓ Sufficiency of lightweight representation
  - We show 62 tuple representation close => documents close
- ✓ Efficacy of sequential range searches and 8 random projections
  - Recall / Precision
    - Recall of .999
    - Precision of .912
  - Speed
    - 2.57 seconds (time for search to return results—too slow due to Oracle quirk)
  - Heuristics for Oracle implementation
    - Speed heavily dependent on the precision of first range searches performed
    - Use character count and 62 tuple vector size as first 2 range searches
    - Improves speed to .48 seconds

DESI V WORKSHOP 2013

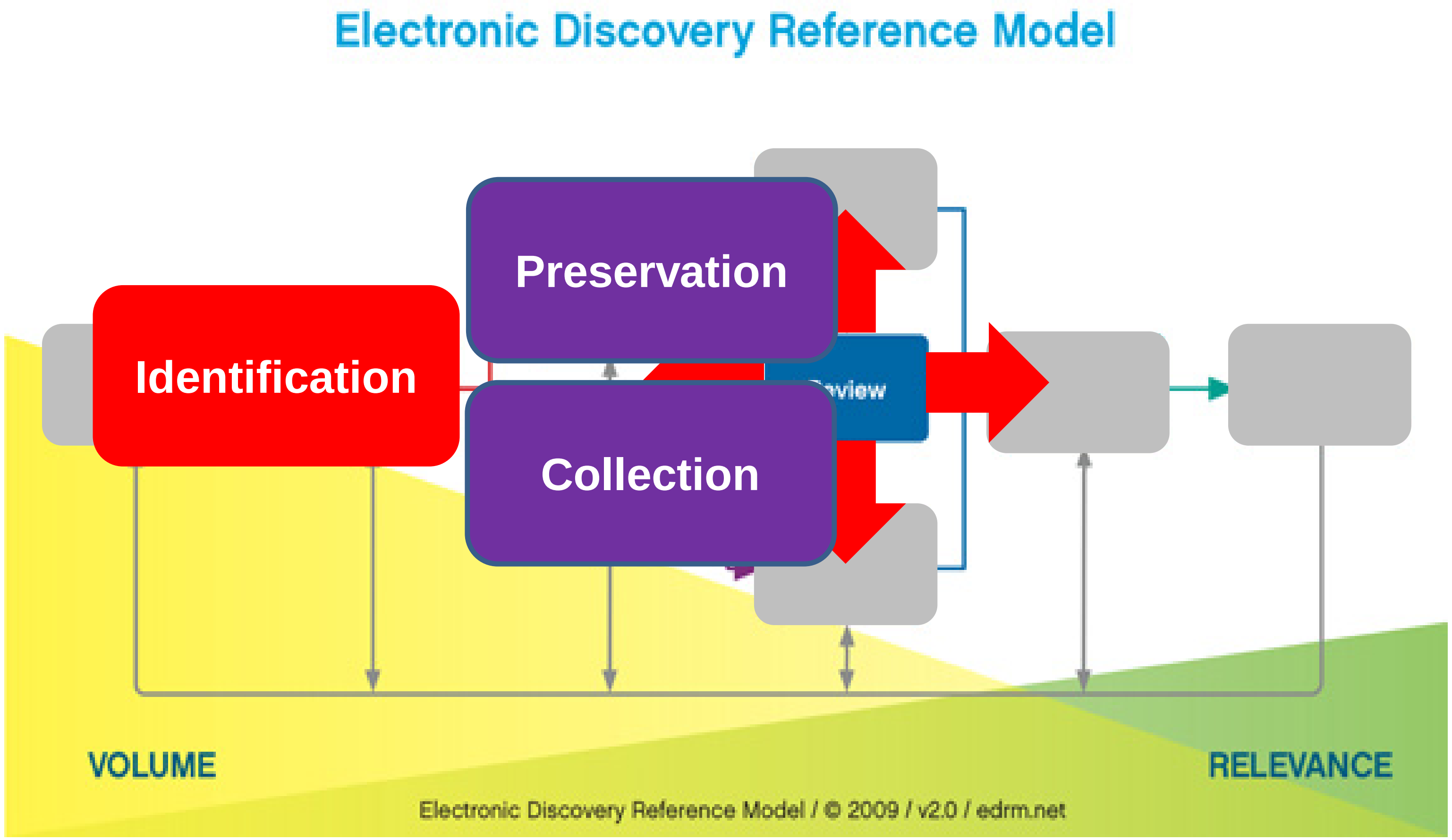
Similar Document Detection and Electronic Discovery:
So Many Documents, So Little Time

Michael Sperling, Rong Jin, Illya Rayvych, Jianghong Li and Jinfeng Yi

Predictive Coding: Turning Knowledge into Power

Thomas I. Barnett and Michael Sperling

Process level expansion



The case: a typical large class action…

Legal Obligations

Rule 26. Duty to Disclose; General Provisions Governing Discovery

(a) Required Disclosures.

(1) Initial Disclosure.

. . .

(ii) a copy—or **a description by category and location—of all documents**, **electronically stored information**, and tangible things that the disclosing party has in its possession, custody, or control and **may use to support its claims or defenses**, unless the use would be solely for impeachment;

. . .

(2) *Conference Content; Parties' Responsibilities* . . . **discuss any issues about preserving discoverable information; and develop a proposed discovery plan**.

. . .

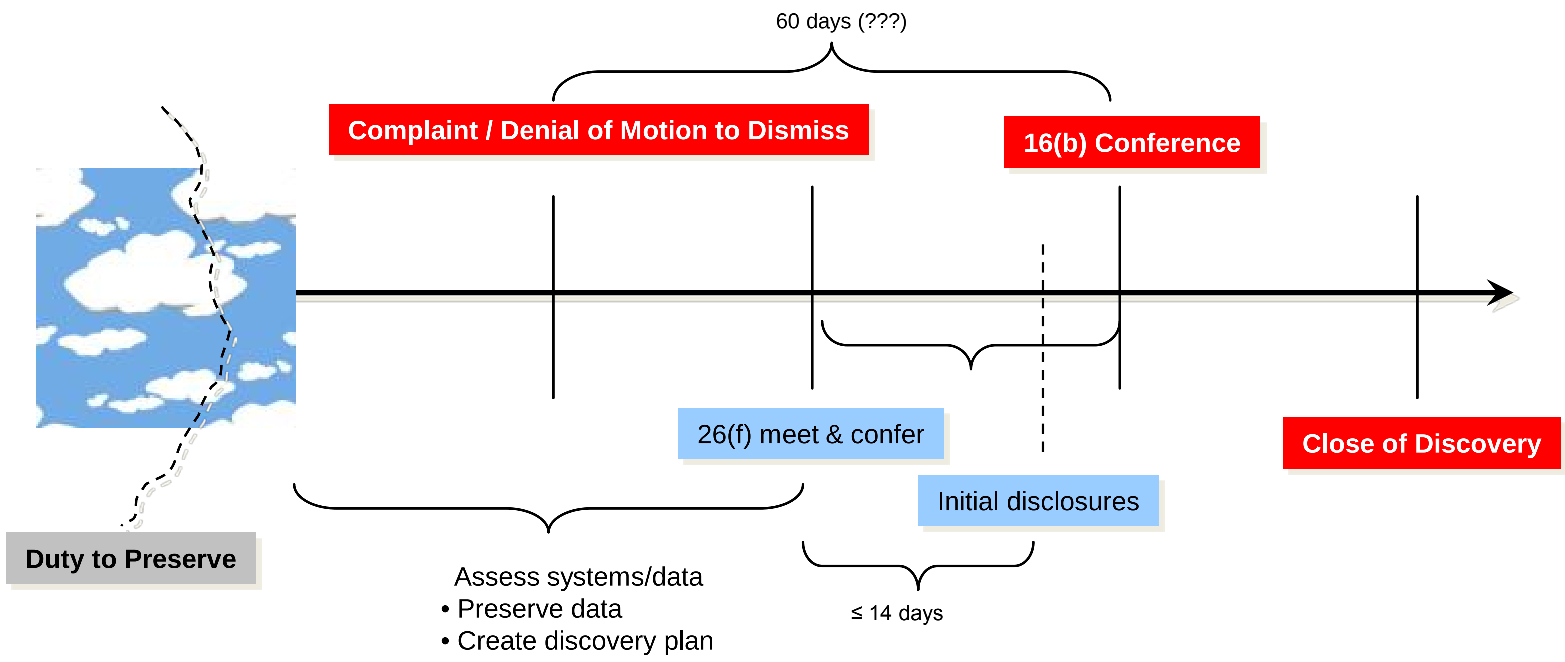
Rule 37. Failure to Make Disclosures or to Cooperate in Discovery; Sanctions

. . .

(f) **Failure to Participate in Framing a Discovery Plan**. If a party or its attorney fails to participate in good faith in developing and submitting a proposed discovery plan as required by Rule 26(f), the **court may, after giving an opportunity to be heard, require that party or attorney to pay to any other party the reasonable expenses, including attorney's fees, caused by the failure**.

Time Requirements

•



Typical Attorney Knowledge Base for 26(f) Conference

- ✓Estimate of number of data custodians
- ✓Partial list of possible data sources
- ✓Some preservation efforts
- ✓Some data custodian interviews

When it comes to negotiating decisions that can cost a company millions of dollars, putting aside potential penalties or liability, ***this is a very thin and indefensible knowledge base***.

Thesis of Position Paper:

Predictive coding (and other analytical tools) can and should be used to provide substantive quantifiable data upon which to negotiate scope of discovery in a meaningful way.

Available Information

- ✓ Supportable estimate (not perfect) of how much data will actually need to be reviewed (i.e., time and cost)
- ✓ Supportable estimate of likely percentage of responsive data
- ✓ Defensible information as to relative value of data sources/custodians
- ✓ Actionable information that can be used to substantively challenge unnecessarily broad requests

Conclusion and future work

- ✓ The emphasis on "coding" as in "coding for production" is misguided and unnecessarily limiting.

- ✓ There are many ways to apply analytical approaches to this multifaceted problem called *data discovery* and they go well beyond simply responsiveness or issue coding.

- ✓ There is an opportunity to develop work flows using different combinations of analytical approaches and get beyond the highly limited and limiting world of litigation support technology.

- ✓ There is a whole world of advanced analytical tools and processes beyond those dreamt of in most lawyers’ philosophies.

THANK YOU