

# Text Classification in Electronic Discovery

Dave Lewis, Ph.D.  
David D. Lewis Consulting, LLC  
[www.DavidDLewis.com](http://www.DavidDLewis.com)

Slides for lecture via Skype on February 23, 2012 in LBSC 708X/INFM 718X  
(Seminar on E-Discovery, Univ. of Maryland).

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

1

## Disclaimer

- I am not a lawyer
- Nothing here should be considered legal advice
- If you need legal advice, consult a lawyer
- Repeat as needed

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

2


## Disclaimer 2

- I have a variety of interests in this area
  - Co-founder of TREC Legal Track
  - Publishing and public speaking in attempt to influence evolving legal standards
  - Consulting expert and expert witness on legal cases involving e-discovery
  - Algorithm designer for an e-discovery vendor

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

3

## Outline

- What is e-discovery?
- IR technologies in e-discovery
  - Text retrieval
  - Text classification 
  - Effectiveness evaluation
- Summary

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

4

## E-Discovery in Black and White

- A document is
  - Preserved or not
  - Reviewed or not
  - Listed in privilege log or not
  - Turned over as responsive or not
  - Presented in court or not
    - As evidence of a particular fact or not
- Law tries to define actions unambiguously
  - That means *classification*

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

5

## Text Classification

- Deciding which of several predefined classes (groups) a text belongs to
- Crudest form of understanding text
- BUT...current technology often can do well
- AND...many tasks can be viewed as classification
  - Particularly tasks of organizing information

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

6

## Advantages of Viewing a Task As Classification

- Classifiers are good plug-in modules
  - “Text in/class out” a simple interface
- Finite set of outputs a good basis for
  - Conditioning action on output
  - Counting, correlation, etc.
- Automated applying and learning of classifiers
  - Software reusable across classification tasks
- Clarifies nature of task, successes, failure modes
  - Straightforward numerical measures of quality

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

7

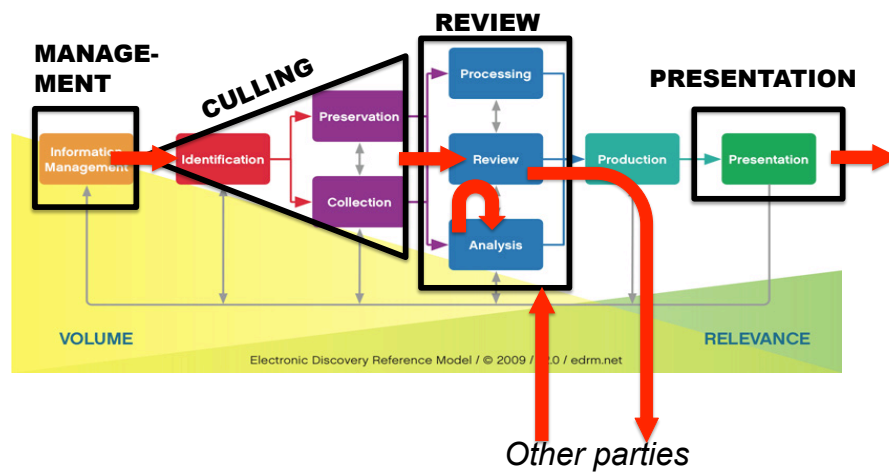
## Organizing Sets of Classes

- Binary classification
- Three or more classes
  - Multiclass
  - Multilabel
  - Ordinal
  - Hierarchical
- Probabilistic and graded classification

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

8

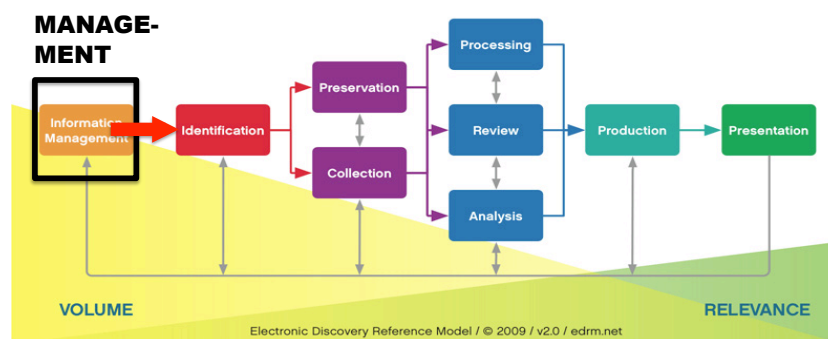
## classification in e-discovery



Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

9

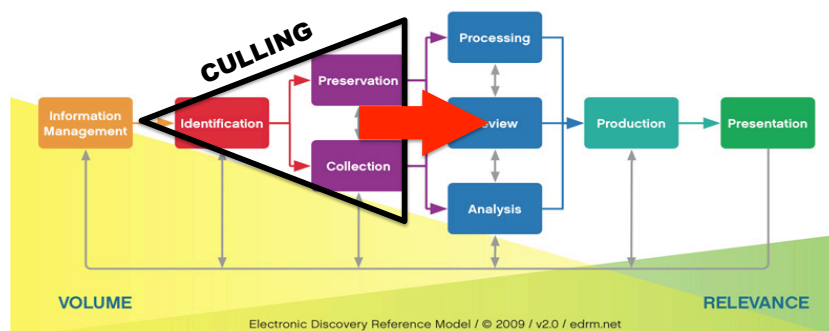
## which documents to keep?



Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

10

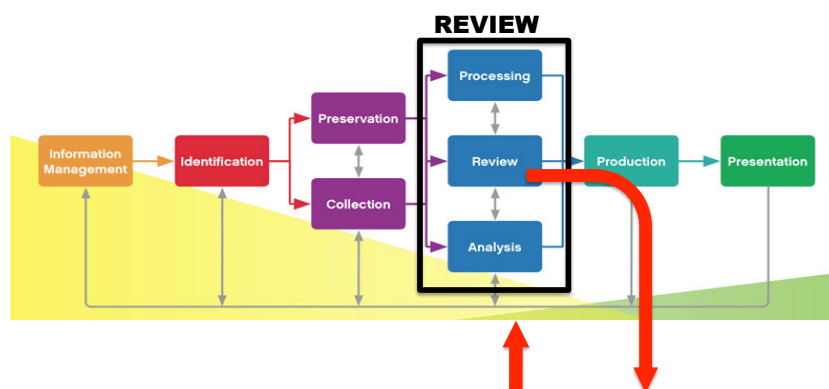
which documents to review?



Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

11

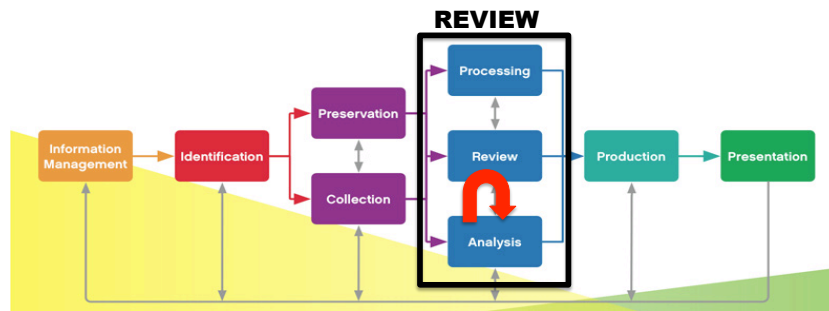
which documents must  
we turn over?



Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

12

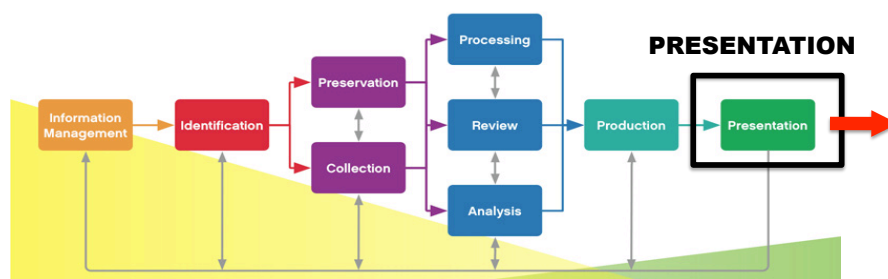
what types of documents do we have?



Copyright 2001-2010, David D. Lewis Consulting. All rights reserved.

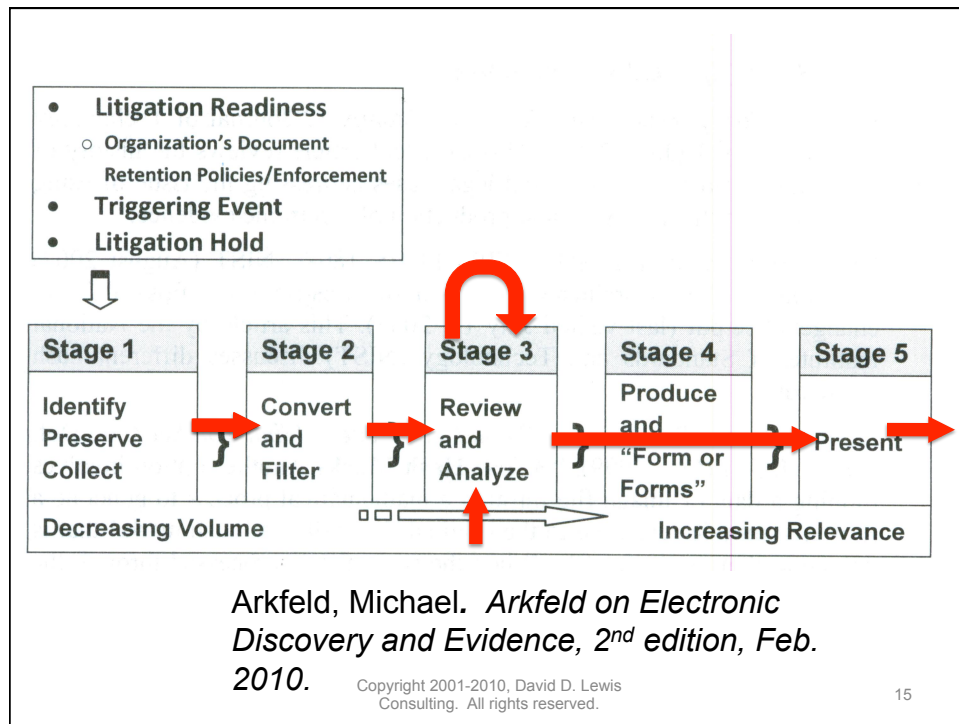
13

which documents do we show in court?



Copyright 2001-2010, David D. Lewis Consulting. All rights reserved.

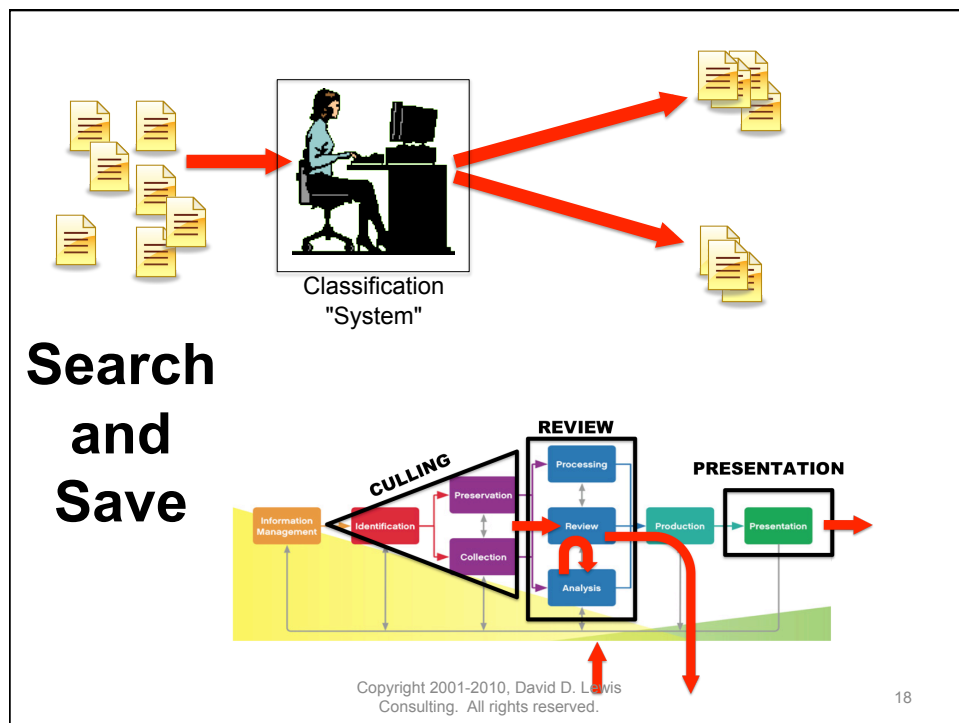
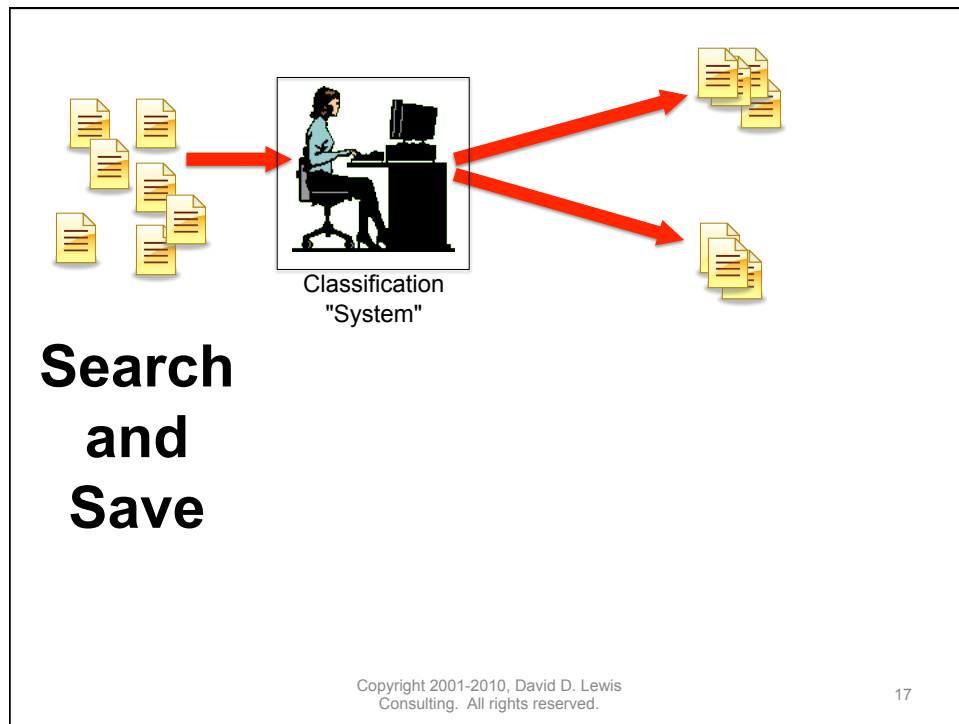
14



## Approaches to Classification

- Interactive manual classification
  - Search and save documents
- Interactive manual classifier creation
  - Search and save query
- Supervised learning of classifier
  - Review documents
  - Computer creates query



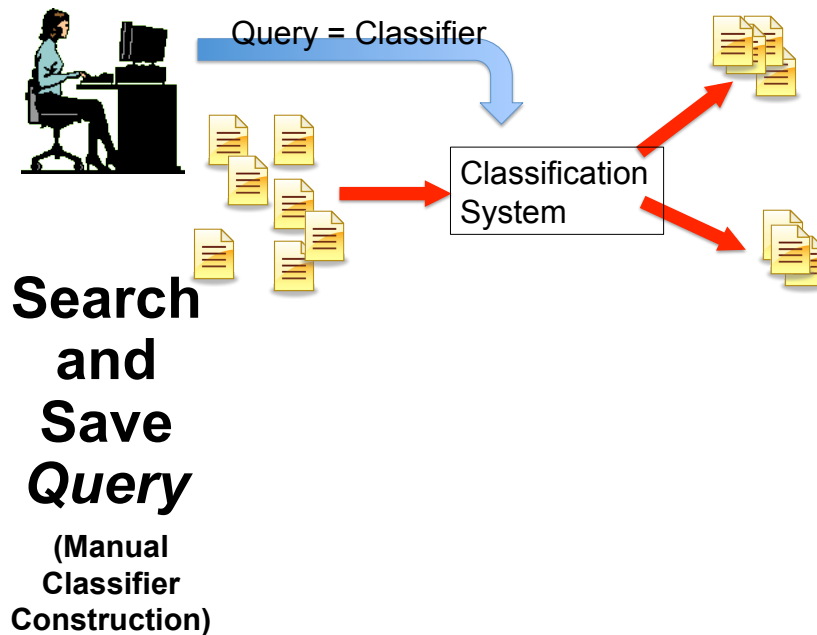


## Search and Save

- Strengths:
  - See interesting docs first
  - Leverage search algorithms and interfaces
- Weaknesses
  - Personnel must have both case knowledge and search skills
  - Documents arrive in bursty fashion
    - Constantly need these highly trained people
  - Human variability
  - Can't show later why doc was or wasn't saved

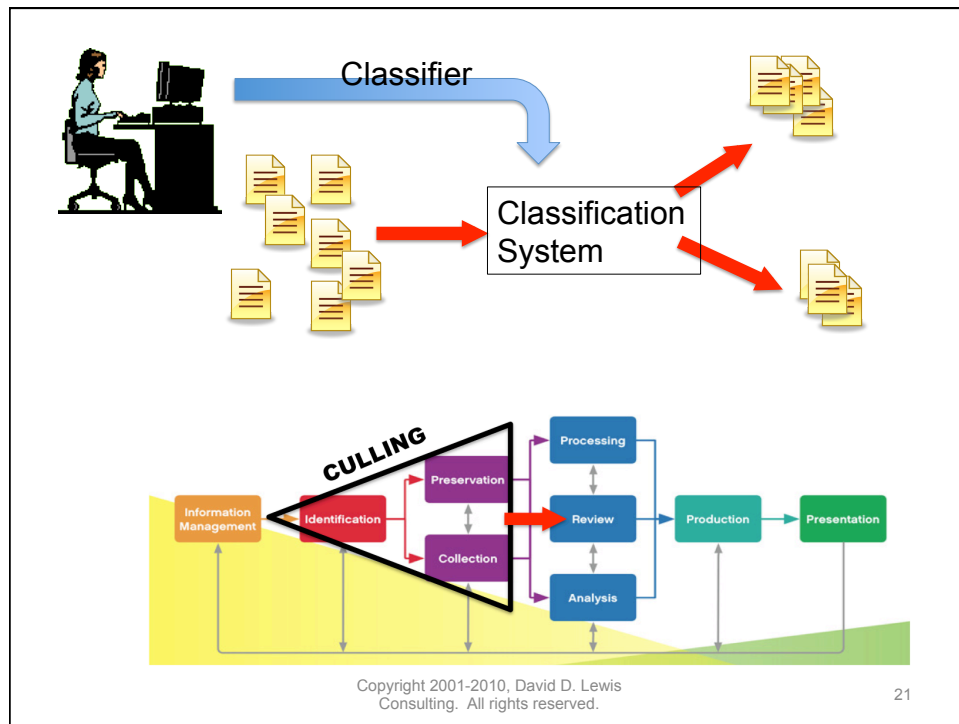
Copyright 2001-2010, David D. Lewis Consulting. All rights reserved.

19



Copyright 2001-2010, David D. Lewis Consulting. All rights reserved.

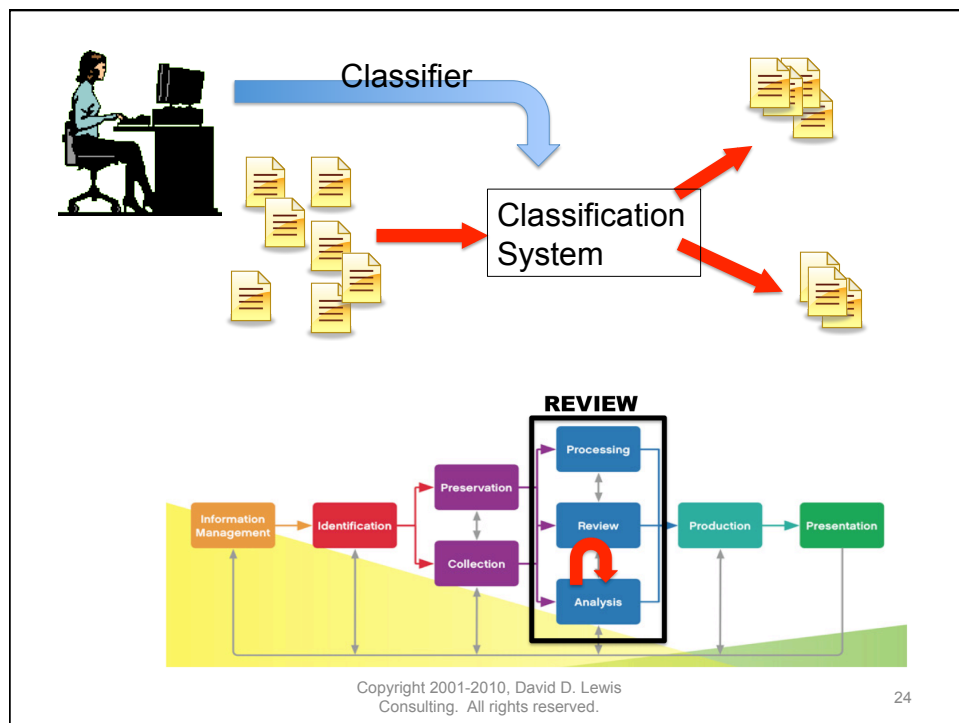
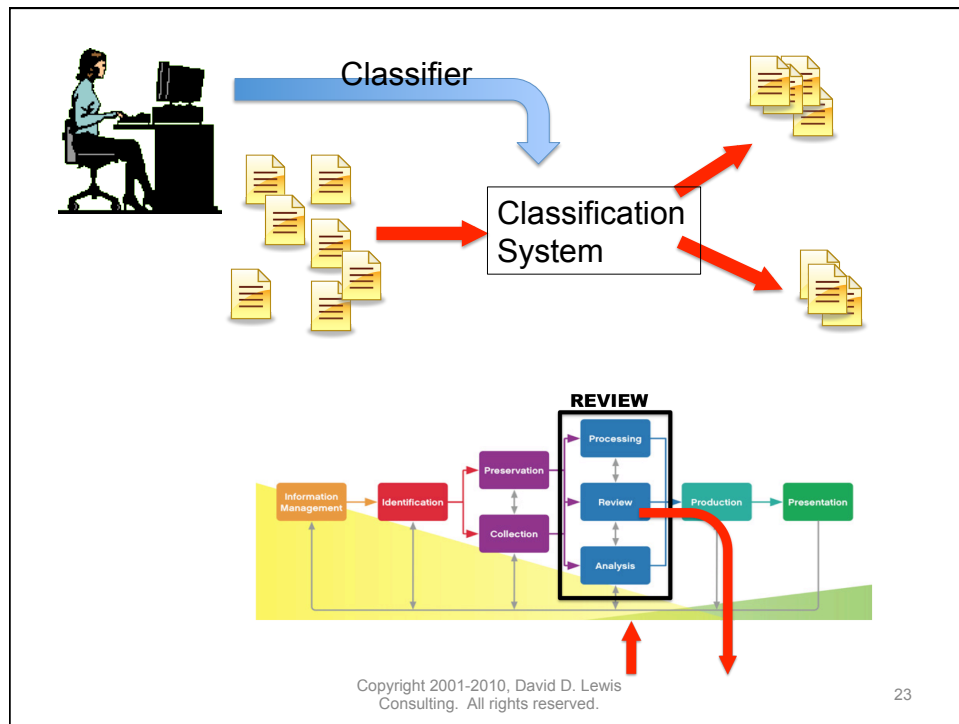
20

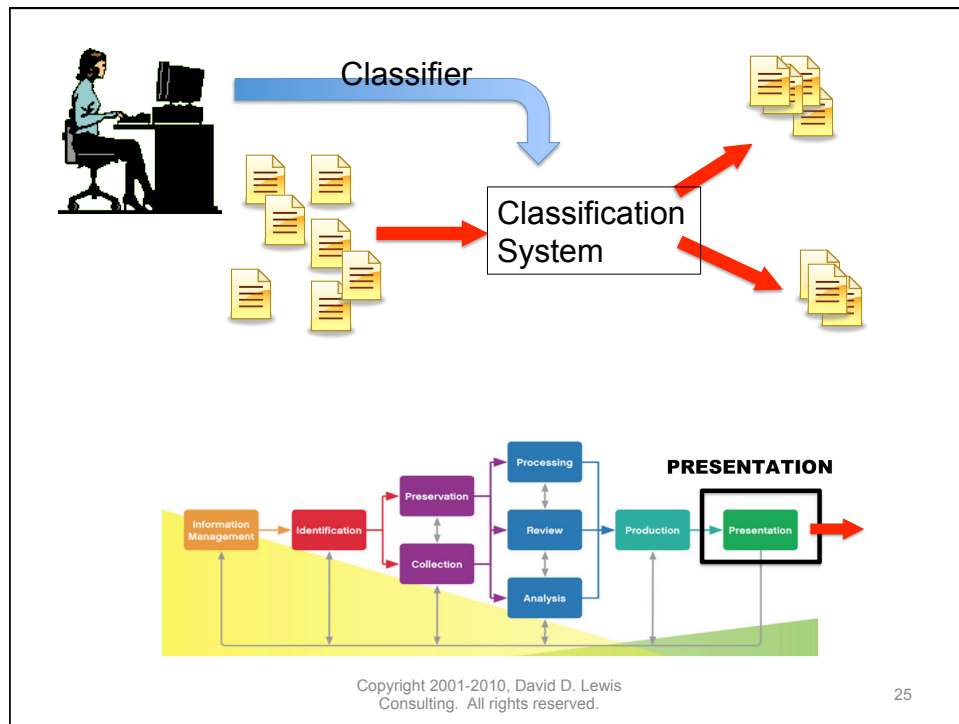


## Manual Classifier Construction in Culling

- 1-2 orders of magnitude reduction typically quoted
- Traditionally simple Booleans
- Developed iteratively
- Sometimes statistical evaluation of effectiveness

*Things are moving very rapidly in this area*





## Research Questions in Manual Classifier Construction for E-Disco

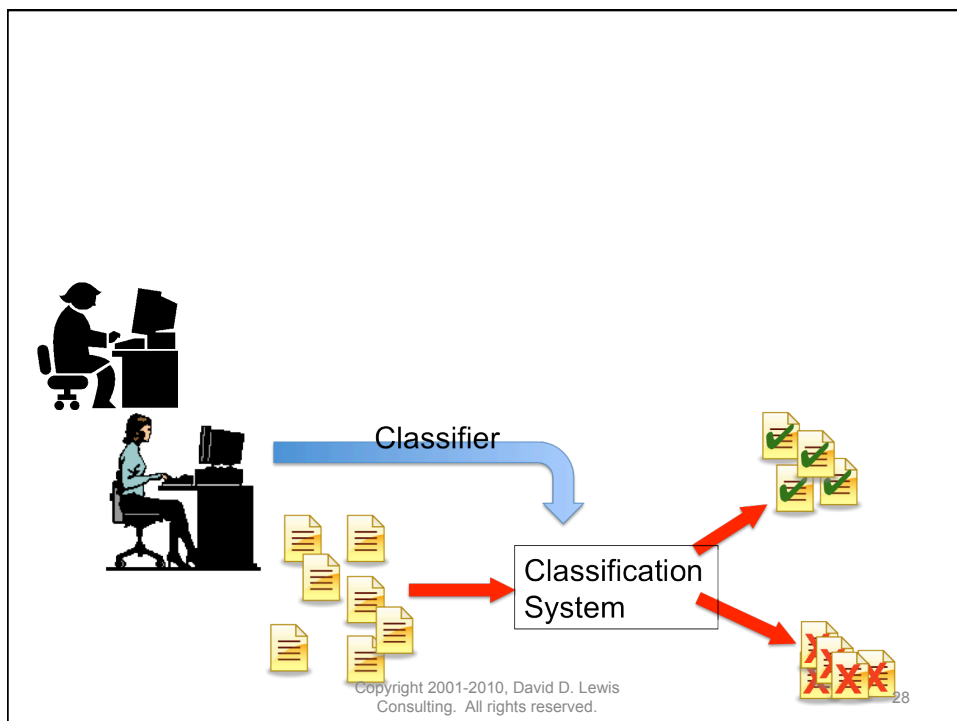
- Culling:
  - Optimizing large-grained selection
  - Detecting redundancy
  - Summary representations
  - How good are people at this?
  - How can we help them?
- Much from distributed search applies!

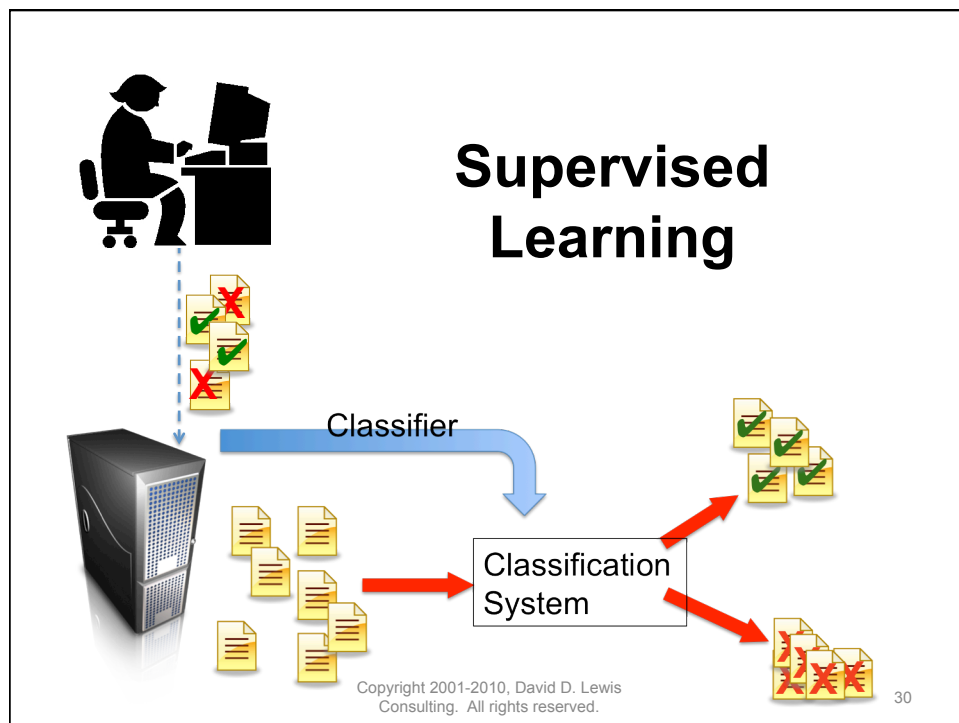
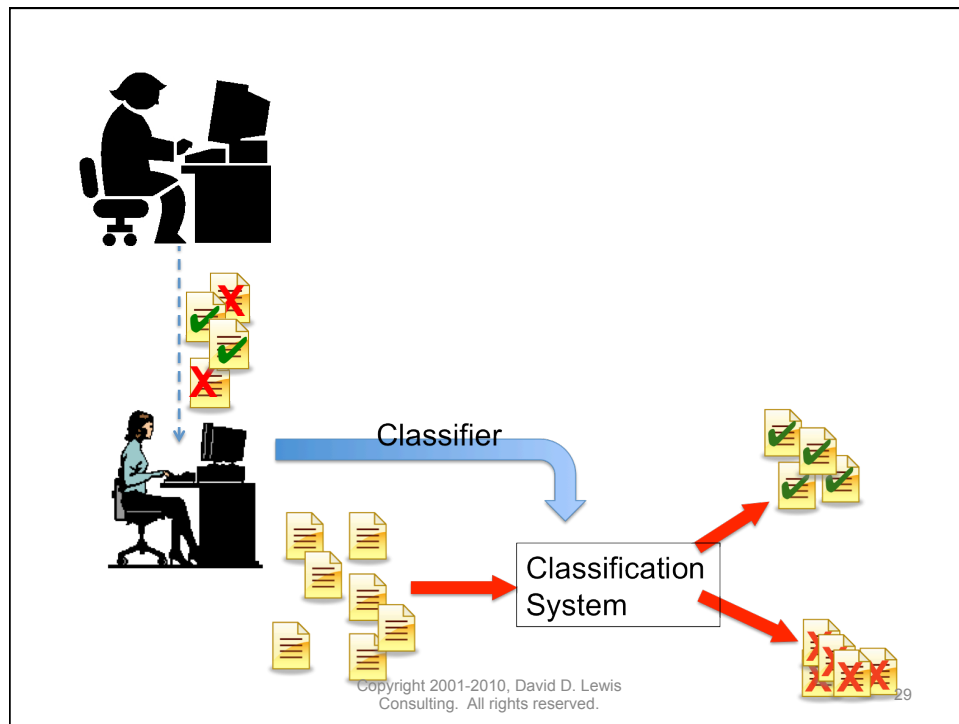
## Manual Classifier Construction

- Strengths
  - Classifier can be run as new docs arrive
  - Explicit justification of decisions
- Weaknesses
  - Must justify why classifier built way it was
  - New docs may differ from old ones, requiring ongoing modification of classifier
    - Haven't escaped need for single search/case expert

Copyright 2001-2010, David D. Lewis Consulting. All rights reserved.

27



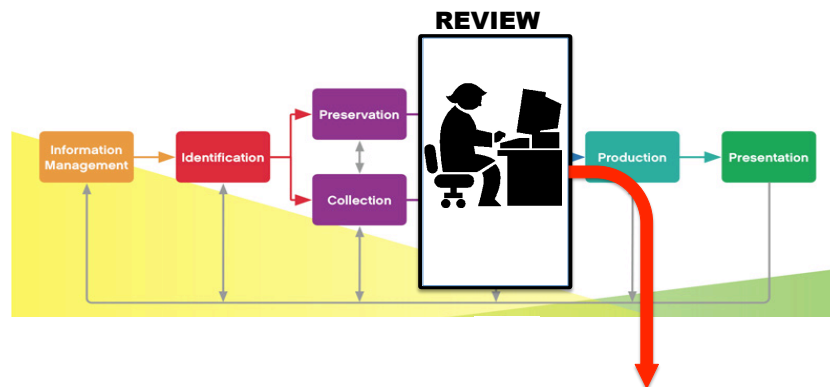


## Supervised Learning for Text Classification

- People manually classify documents
- Computer searches for a rule that
  - Agrees with most manual classifications
  - Is not too "complex"
- Uses that rule to classify future documents
  - Rules typically associate numeric weights with words and other document features

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

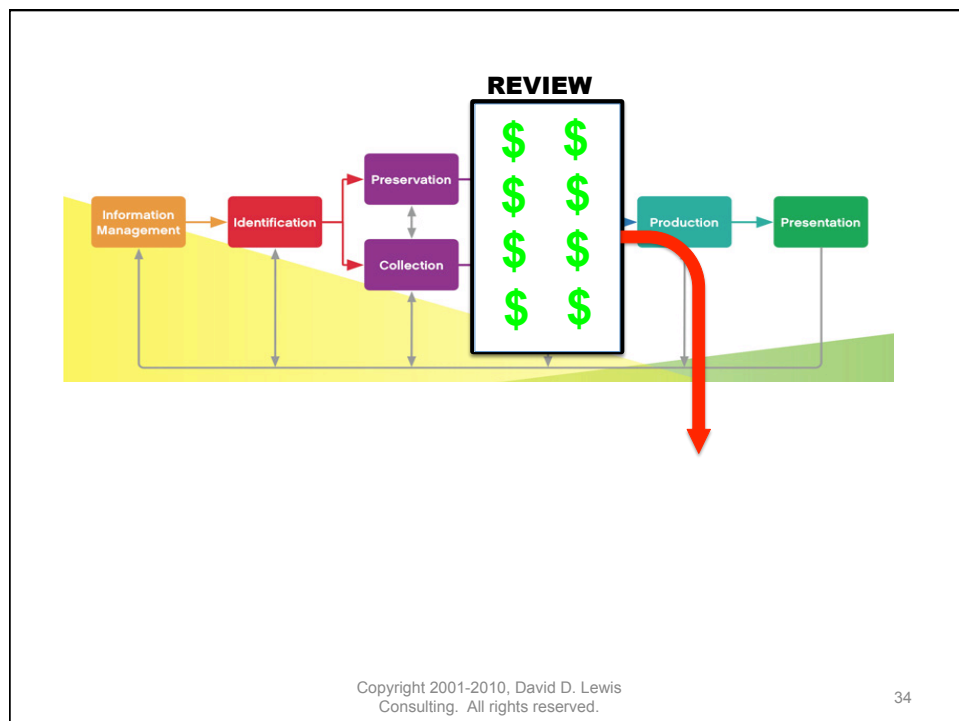
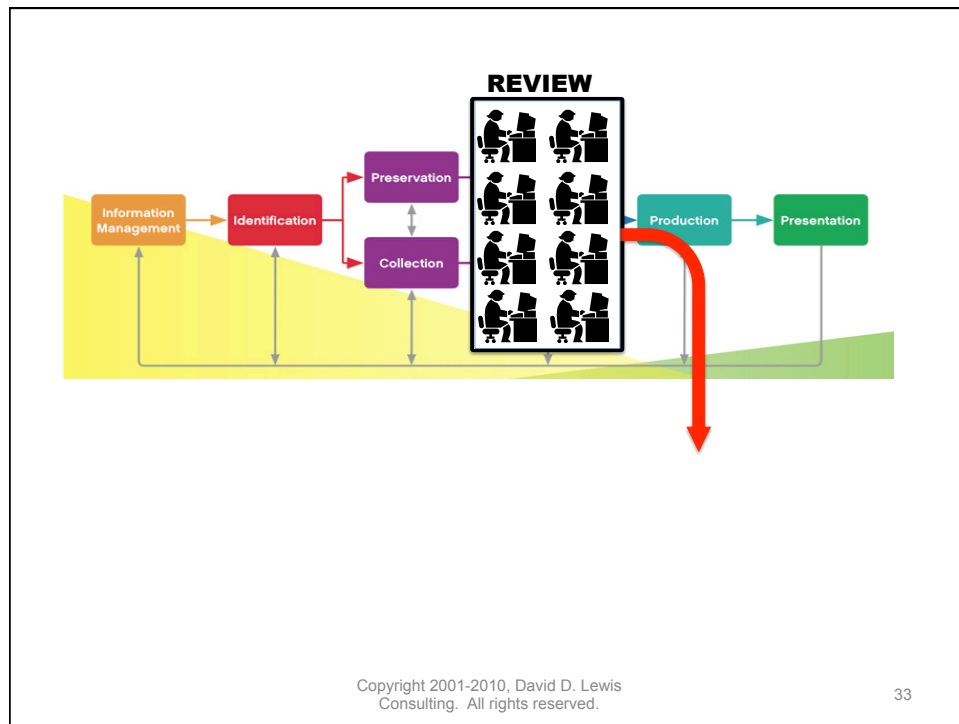
31

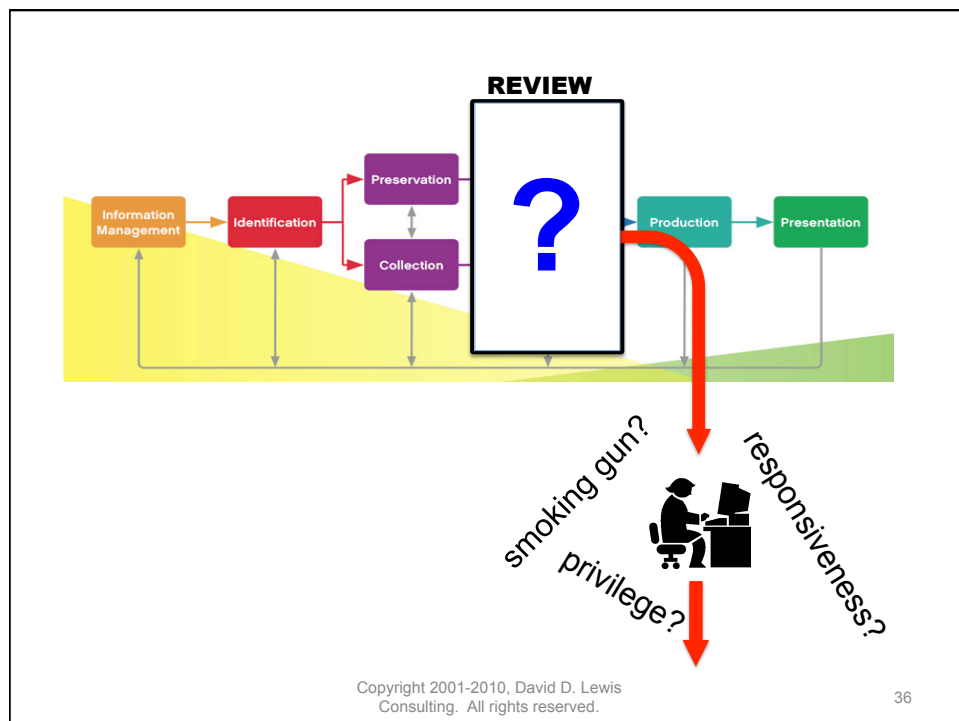
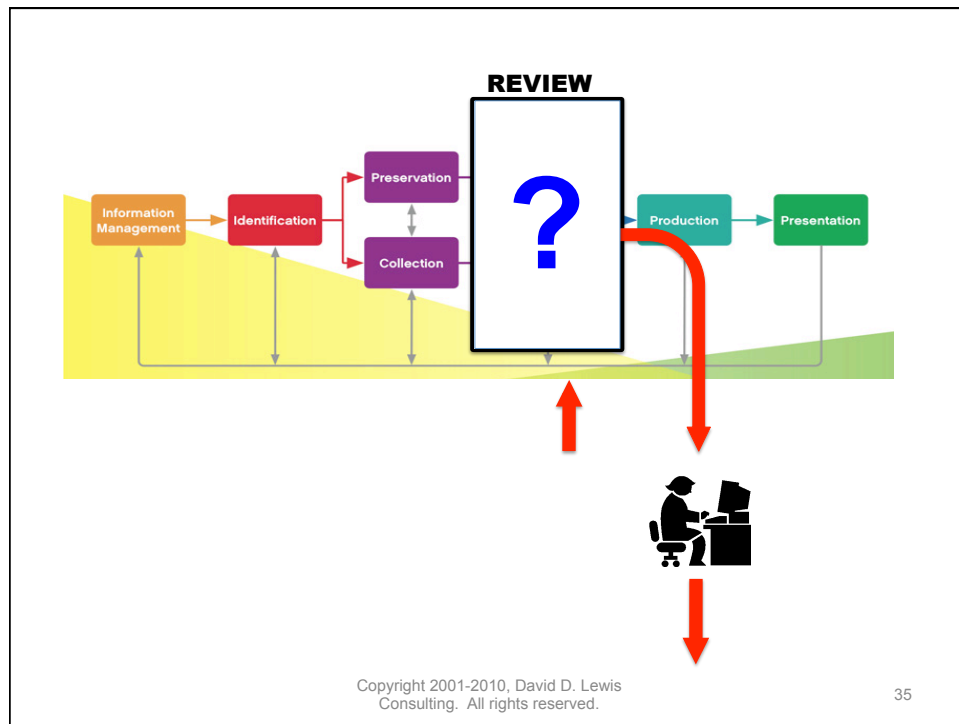


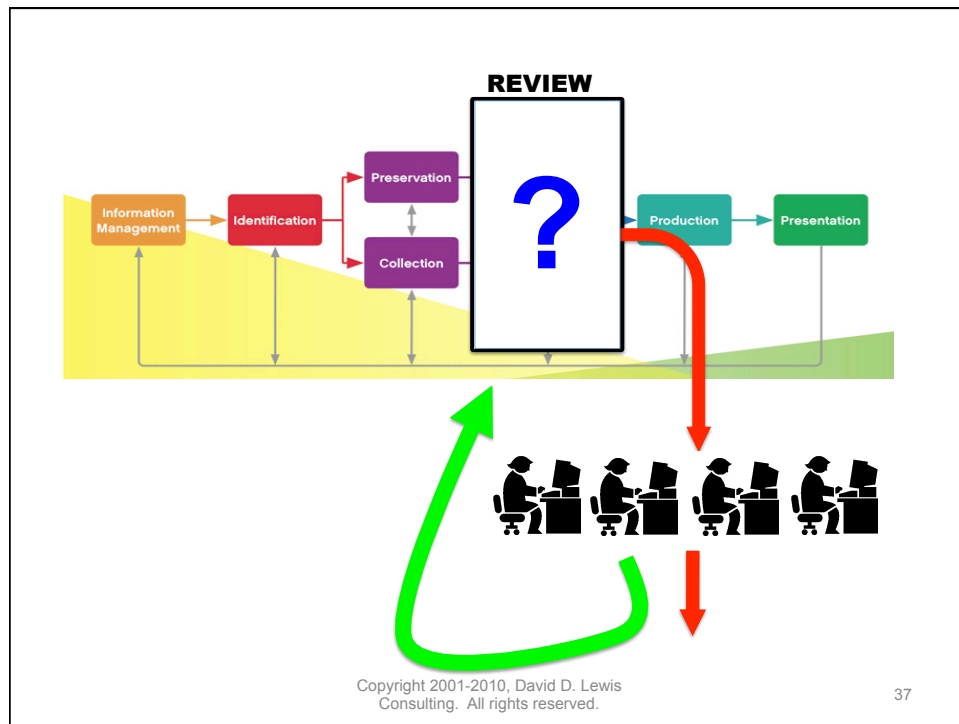
Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

32



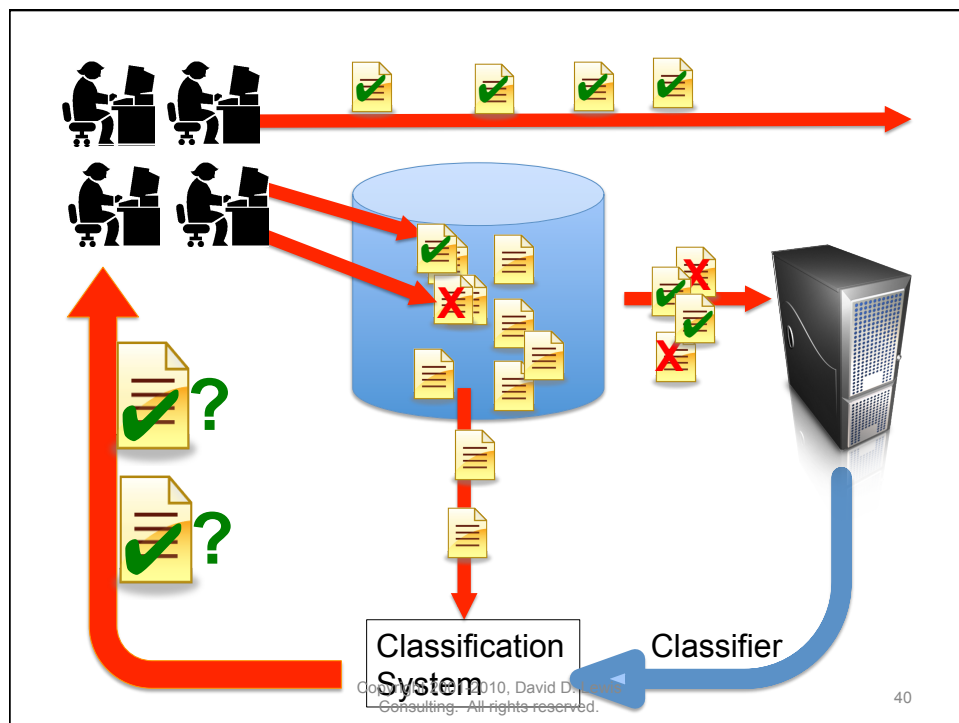
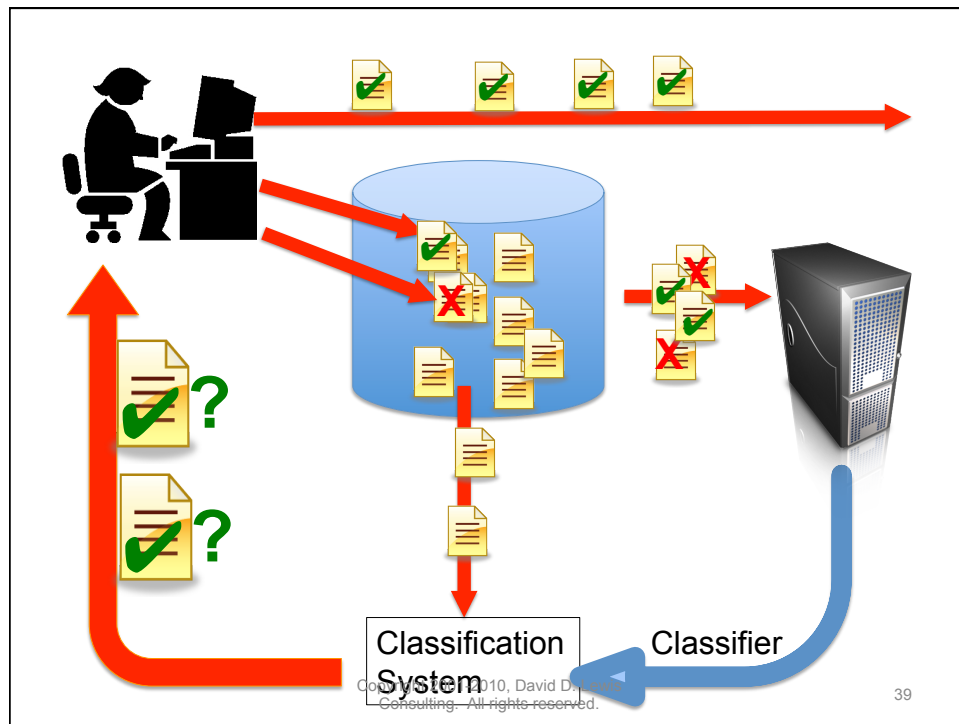


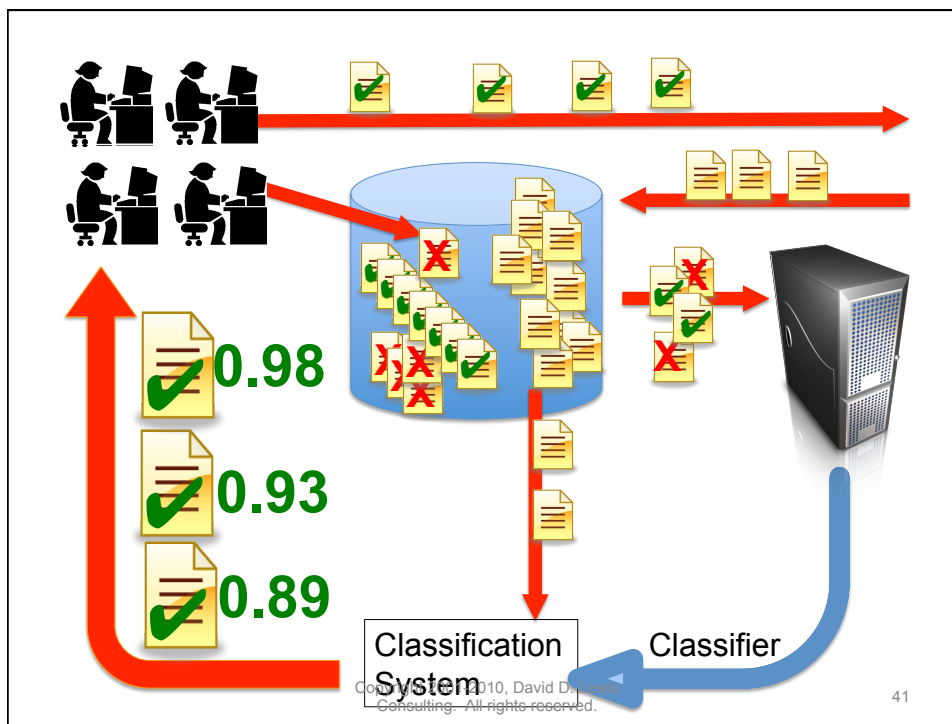




## Other Favorable Factors for Supervised Learning

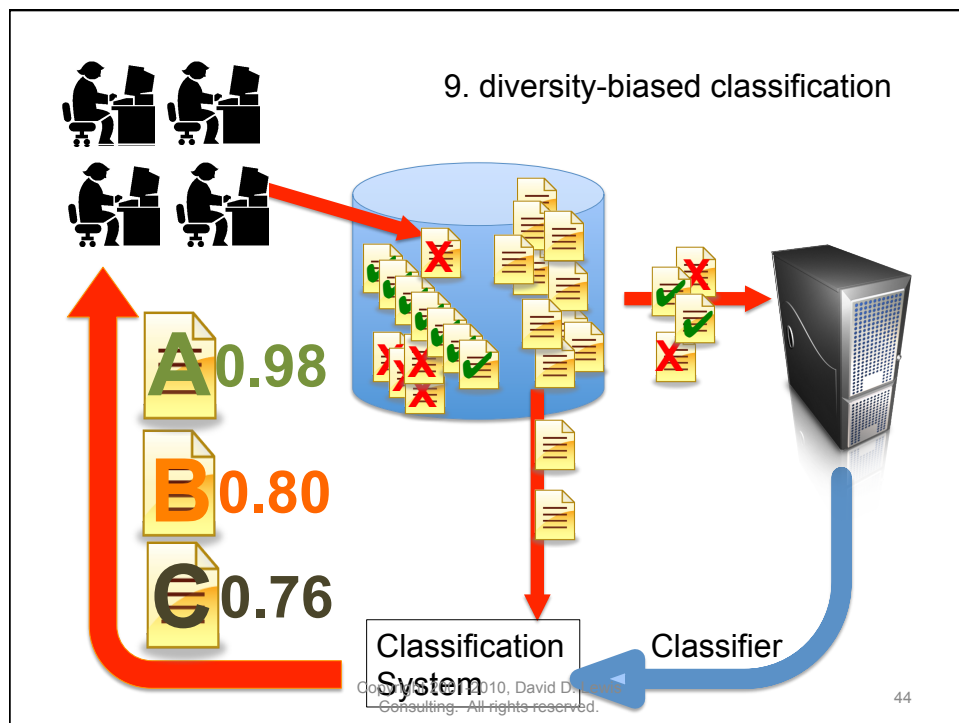
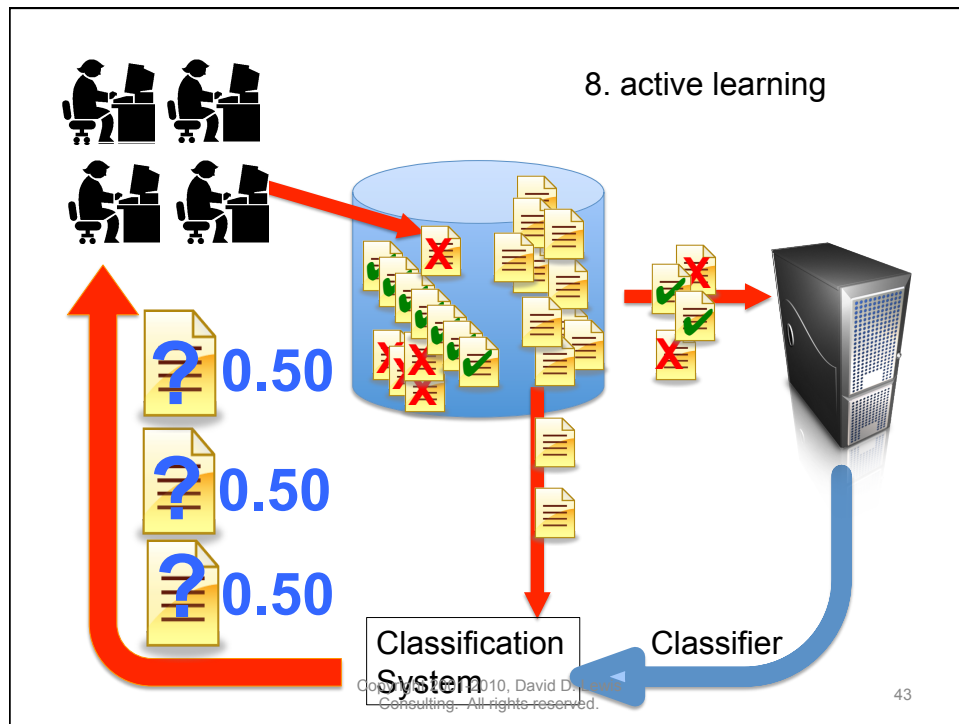
- Review is a recall-oriented task
  - Classifier likely to need thousands of words
  - Very hard to create by hand
- Many, weak, disparate predictors
  - Text, custodian, time, place, organizational structure, file type, message headers,...
  - Again very hard to manually use
- Objective reason the classifier is way it is
  - "The learning algorithm said so"

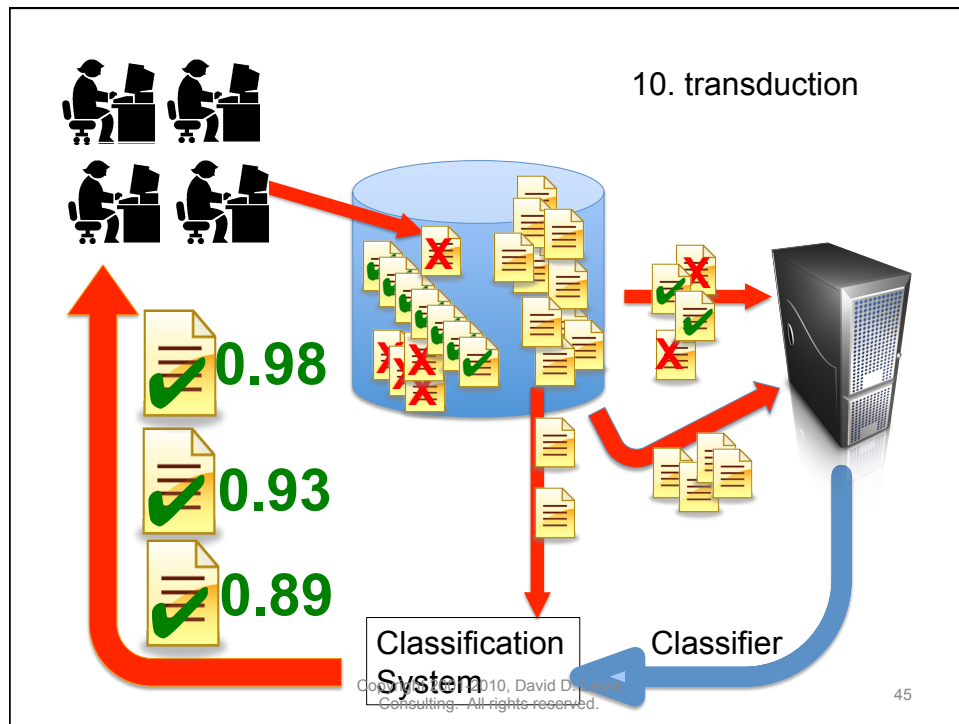




## Unusual Aspects of E-Discovery for Supervised Learning

- 1 Ultra high recall, moderate precision
  - 2 Extremely diverse document sets
  - 3 Duplicate documents
  - 4 Document families
  - 5 Multiple assessors
  - 6 Some categories conditional on others
  - 7 Goal of assessment is review, not training
    - May be unwilling to evaluate documents for categories if not responsive
- All of these are pains / research opportunities*





## Outline

- What is e-discovery?
- IR technologies in e-discovery
  - Text retrieval
  - Text classification
  - Effectiveness evaluation
- Why I love e-discovery ←

## The Joy of E-Discovery

- All that SIGIR/TREC/etc. IR geek stuff...
  - Supervised learning from huge amounts of training data
  - High recall search
  - Careful statistical measurement of effectiveness
  - Obsessed tuning to get effectiveness up
- ***Matters, works, and saves people great drudgery and cost***

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

47

## Summary

- E-discovery is an application where IR really matters
  - Current technology can help a lot
  - Advances would help even more
  - Rigor matters: write, program as if you might have to testify about it

Copyright 2001-2010, David D. Lewis  
Consulting. All rights reserved.

48



## Thanks!

- I'm always happy to answer questions:

**umd20120223@DavidDLewis.com**

- Sign up if you'd like to be on list to receive updated copies of this talk