DrinkerBiddle

The Future of Email Archiving: Four Propositions

Digital Preservation Coalition Briefing Day
London, U.K.
January 24, 2018

Jason R. Baron
Drinker Biddle & Reath LLP
Washington, D.C.

DrinkerBiddle

Lesson -- Email is the long time, undisputed world champion 800 lb. gorilla of information governance problems, scandals, controversies, etc.

DrinkerBiddle

Growth in Presidential and Federal Records in Electronic Form

- William Clinton White House 32 million emails (presidential and federal records)
- George W. Bush White House 200+ million emails
- Barack Obama White House ~ 300+ million emails expected by Jan. 2017 + records on social media
- And after 2019: billions of emails and all forms of electronic records from hundreds of Executive Branch agencies being accessioned into the National Archives
- **HOW MANY OF THESE RECORDS WILL BE ACCESSIBLE TO THE PUBLIC AND HOW SOON?**

DrinkerBiddle

DrinkerBiddle

Tomorrow: Exponentially Growing Amounts of Digital Information

DrinkerBiddle

The Future of Email Archives: Four Propositions

Proposition 1

- ***Manual approaches to preserving e-mail result in noncompliance with recordkeeping obligations, and should be abandoned in favor of adopting automated rules which minimize end user involvement.***

DrinkerBiddle

20th Century Records Schedules

REQUEST FOR RECORDS DISPOSITION AUTHORITY
(See Instructions on reverse)

LEAVE BLANK (NARA use only)
JOB NUMBER: N1-60-92-1
DATE RECEIVED: 10/28/91

TO: NATIONAL ARCHIVES and RECORDS ADMINISTRATION (NIR)
WASHINGTON, DC 20408

1. FROM (Agency or establishment): Department of Justice
2. MAJOR SUBDIVISION: Environment and Natural Resources Division
3. MINOR SUBDIVISION:
4. NAME OF PERSON WITH WHOM TO CONFER: Ann Sloan
5. TELEPHONE: 514-3411

NOTIFICATION TO AGENCY

In accordance with the provisions of 44 U.S.C. 3303a the disposition request, including amendments, is approved except for items that may be marked "disposition not approved" or "withdrawn" in column 10.

DATE: 1/3/92
ARCHIVIST OF THE UNITED STATES: James W. Moore

6. AGENCY CERTIFICATION

I hereby certify that I am authorized to act for this agency in matters pertaining to the disposition of its records and that the records proposed for disposal on the attached _____ page(s) are not now needed for the business of this agency or will not be needed after the retention periods specified; and that written concurrence from the General Accounting Office, under the provisions of Title 8 of the GAO Manual for Guidance of Federal Agencies,

[X] is not required; [ ] is attached; or [ ] has been requested.

DATE: 10-24-91
SIGNATURE OF AGENCY REPRESENTATIVE: Bernard W. Bergland
TITLE: Records Officer
Systems Policy Staff
Justice Management Division

7. ITEM NO.	8. DESCRIPTION OF ITEM AND PROPOSED DISPOSITION	9. GRS OR SUPERSEDED JOB CITATION	10. ACTION TAKEN (NARA USE ONLY)
1.	Duplex-numeric Classification 90-1-23, Actions Against the United States Involving Real Property. Disposition: Transfer to WNRC one year after close of case. Destroy 20 years after close of case.	N1-60-88-12, Item 1B(23)	

Copies sent to NN-W, NNT, NCF 1/15/92

115-109
NSN 7540-00-634-4064
PREVIOUS EDITION NOT USABLE
STANDARD FORM 115 (REV. 3-91)
Prescribed by NARA
36 CFR 1228

Figure 5-8. Completed and Approved Standard Form 115

DrinkerBiddle

Problems with the paradigm

- Traditional record schedules work when….
  - Hard copies are managed by assistants & secretaries
  - There are a well-known, semi-structured set of categories based on either format or content of the document
  - Staff can manageably handle conditional trigger dates

DrinkerBiddle

Problems with the paradigm

- Records schedules don't work
  - When everything is born digital, and end-users are expected to perform their own "recordkeeping" functions
  - When end users must print out or drag and drop objects based on pre-populated (and often highly granular) rule-sets
  - When digital objects come in all varieties
  - WHEN THERE ARE TOO MANY RECORDS CATEGORIES TO CHOOSE FROM!

DrinkerBiddle

Failed Email Preservation Paradigms

- Print to paper
- Backup tapes
- Idiosyncratic e-recordkeeping in user folders (e.g., .pst files)
- DoD 5015.2*

*not quite “failed,” but under-utilized and with an Achilles heel

THE WHITE HOUSE

Office of the Press Secretary

For Immediate Release November 28, 2011

November 28, 2011

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

SUBJECT: Managing Government Records

Section 1. Purpose. This memorandum begins an executive branch-wide effort to reform records management policies and practices. Improving records management will improve performance and promote openness and accountability by better documenting agency actions and decisions. Records transferred to the National Archives and Records Administration (NARA) provide the prism through which future generations will understand and learn from our actions and decisions. Modernized records management will also help executive departments and agencies (agencies) minimize costs and operate more efficiently. Improved records management thus builds on Executive Order 13589 of November 9, 2011 (Promoting Efficient Spending), which directed agencies to reduce spending and focus on mission-critical functions.

When records are well-managed, agencies can use them to assess the impact of programs, to reduce redundant efforts, to save money, and to share knowledge within and across their

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
WASHINGTON, D.C. 20408

August 24, 2012

M-12-18

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES AND INDEPENDENT AGENCIES

FROM: **Jeffrey D. Zients**
Acting Director
Office of Management and Budget

David S. Ferriero
Archivist of the United States
National Archives and Records Administration

SUBJECT: Managing Government Records Directive

On November 28, 2011, President Obama signed the Presidential Memorandum – Managing Government Records. This memorandum marked the beginning of an Executive Branch-wide effort to reform records management policies and practices and to develop a 21st-century framework for the management of Government records. The expected benefits of this effort include:

DrinkerBiddle

Archivist/OMB Directive

- M-12-18, Managing Government Records Directive, dated 8/24/12:

  1.1 By 2019, Federal agencies will manage all permanent records in an electronic format.

  1.2 By 2016, Federal agencies will manage both permanent *and* temporary email records in an accessible electronic format.

  http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-18.pdf

DrinkerBiddle

The Future of Email Archives: Four Propositions

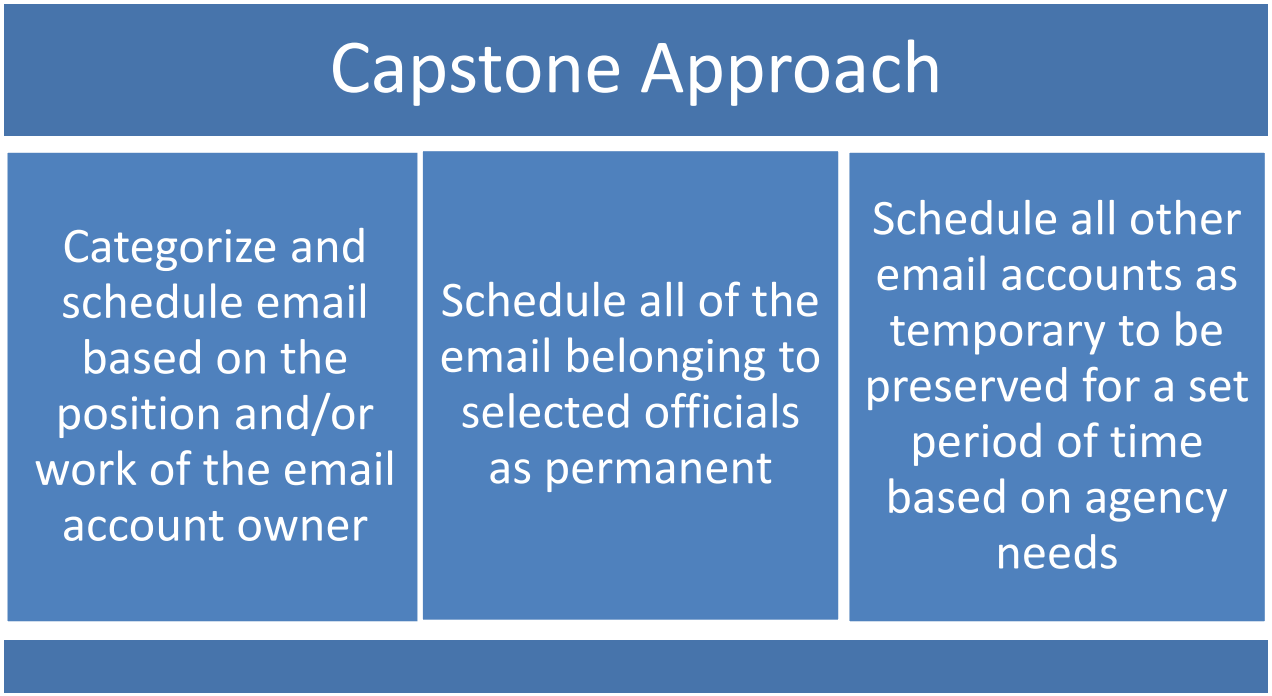
Propositions 1 & 2

- *Manual approaches to preserving e-mail result in noncompliance with recordkeeping obligations, and should be abandoned in favor of adopting automated rules which minimize end user involvement.*
- ***Capture technologies should be implemented to lower enterprise risk of loss and for optimum preservation of a historical record, including for business purposes.***

DrinkerBiddle

Capstone Approach

- Addresses challenges in managing email by reducing preservation to a binary rule, namely:
- Preserve permanently valuable email and dispose of temporary email



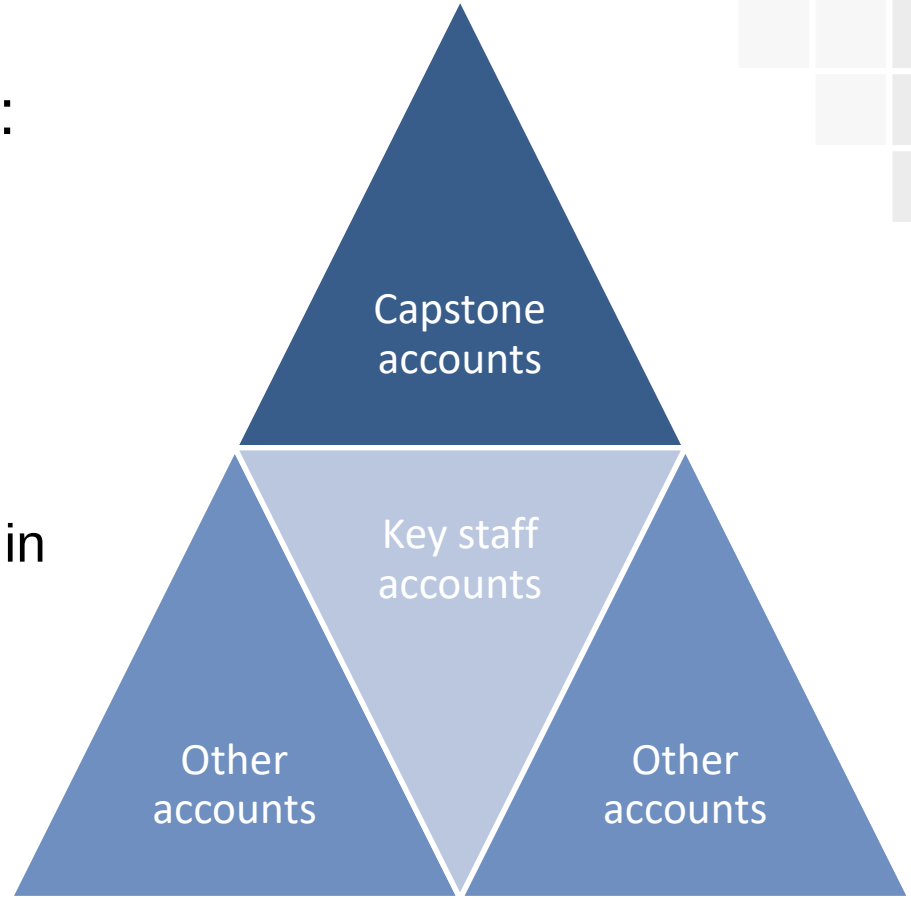
DrinkerBiddle

Capstone Officials

Capstone officials may include:

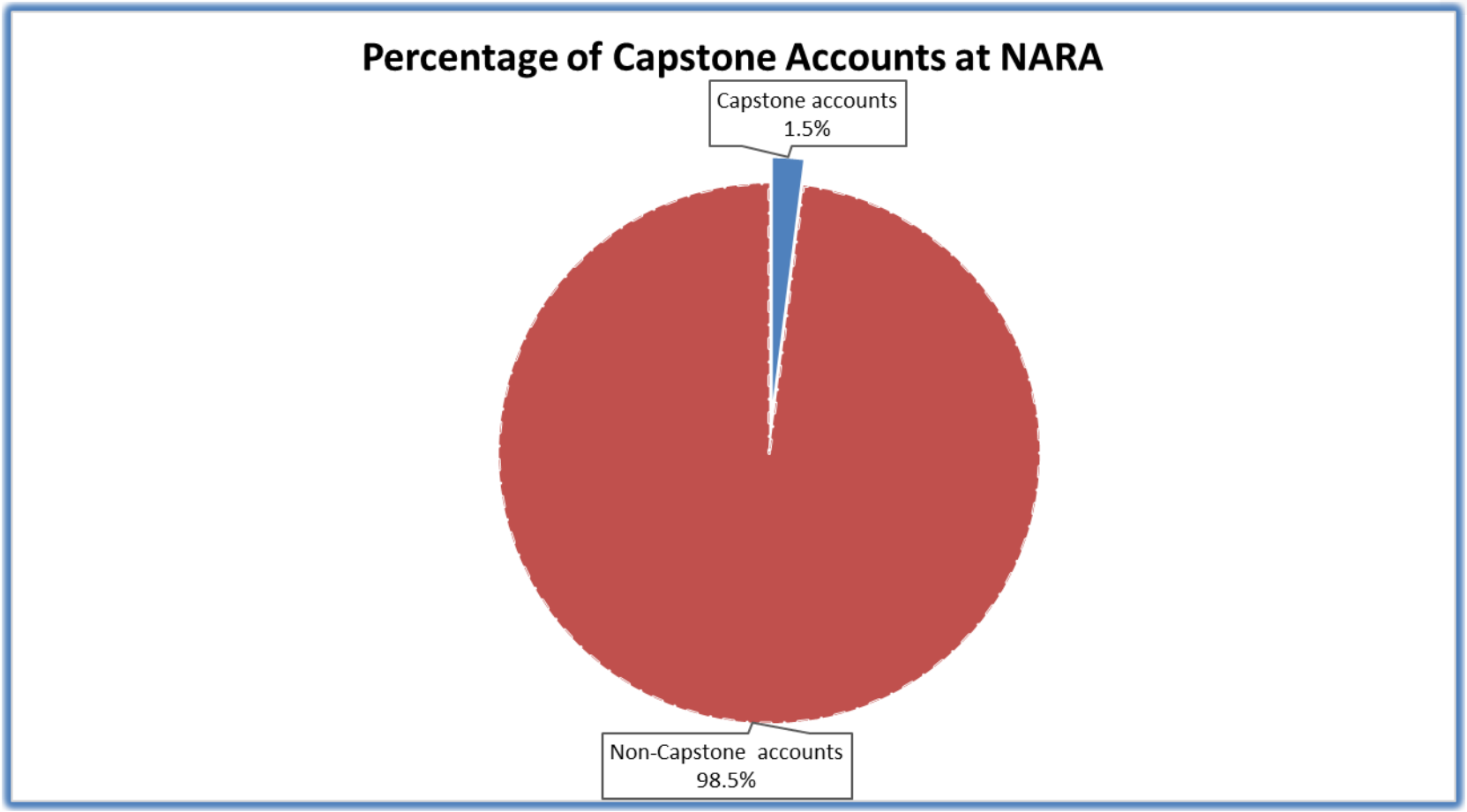
- Officials at or near the top of an agency or an organizational subcomponent

- Key staff members that may be in positions that create or receive presumptively permanent email records



DrinkerBiddle

One example of % Capstone Accounts at an Agency



DrinkerBiddle

Responsibilities

When using the Capstone approach, agencies must continue to:

- Ensure email records are scheduled
- Prevent the unauthorized access, modification, or deletion of declared records
- Ensure all records in the repository are retrievable and usable
- Consider whether email records and attachments can or should be associated with related records
- Capture and maintain required metadata

Drinker Biddle

From the 2016 Federal Agency Records Management Annual Report:

"Our analysis of the SAORM [Senior Agency Officer for Records Management] reporting data shows that M-12-18 is changing Federal records management from paper-intensive analog-based methods to a digital government as intended."

Source: https://www.archives.gov/files/records-mgmt/resources/2016-federal-agency-records-management-annual-report.pdf

DrinkerBiddle

Beyond Capstone: Automated Categorization of Records into Records Schedule Categories

DrinkerBiddle

Bigger Buckets / Fewer Categories

DrinkerBiddle

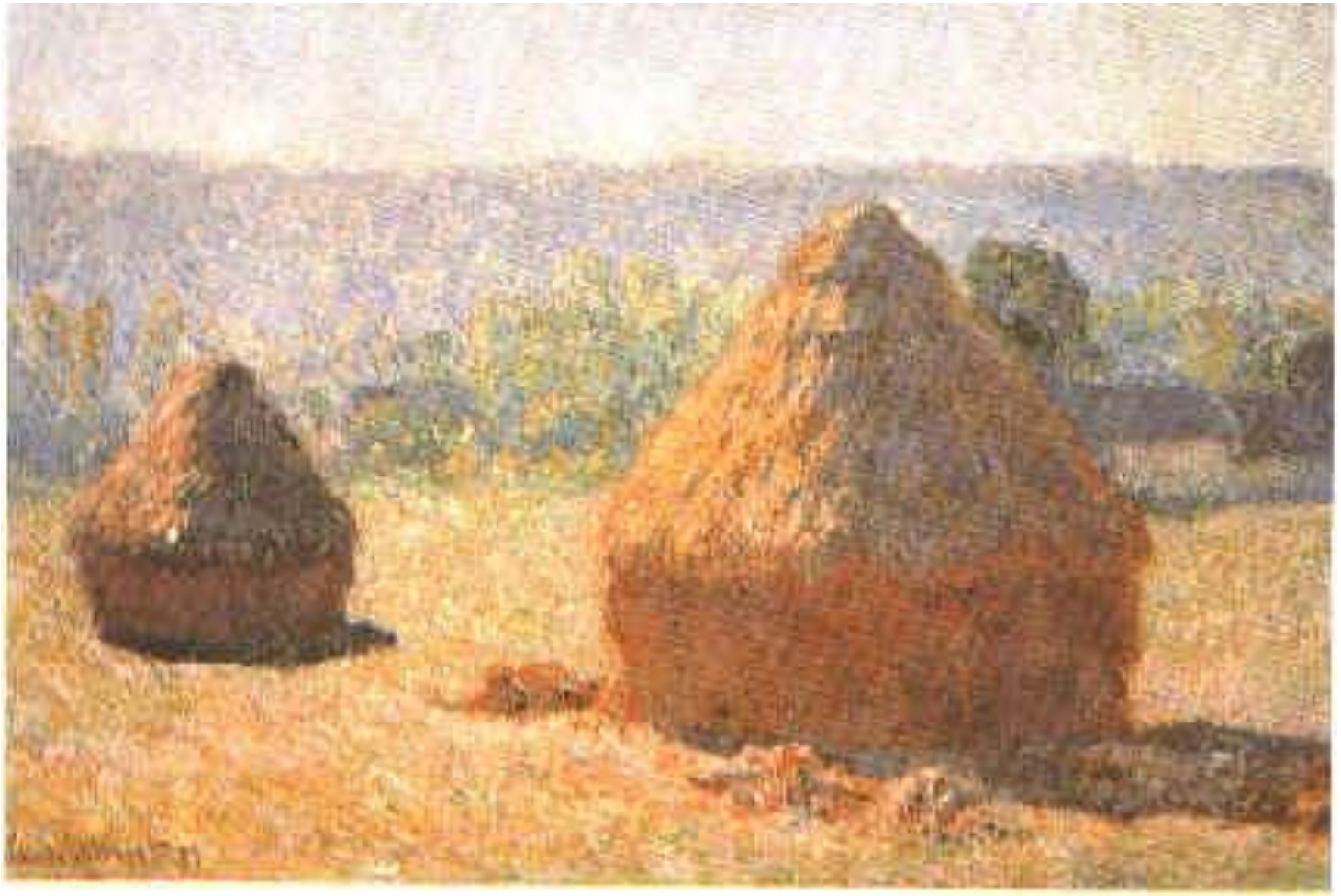
The Future of Email Archives: Four Propositions

Propositions 1, 2 & 3

- *Manual approaches to preserving e-mail result in noncompliance with recordkeeping obligations, and should be abandoned in favor of adopting automated rules which minimize end user involvement.*
- *Capture technologies should be implemented to lower enterprise risk of loss and for optimum preservation of a historical record, including for business purposes.*
- ***Deployment of advanced search methods (e.g., machine learning) will greatly aid in providing more expeditious access to relevant materials in ever more voluminous collections.***

DrinkerBiddle

An information retrieval task: searching the Haystack….



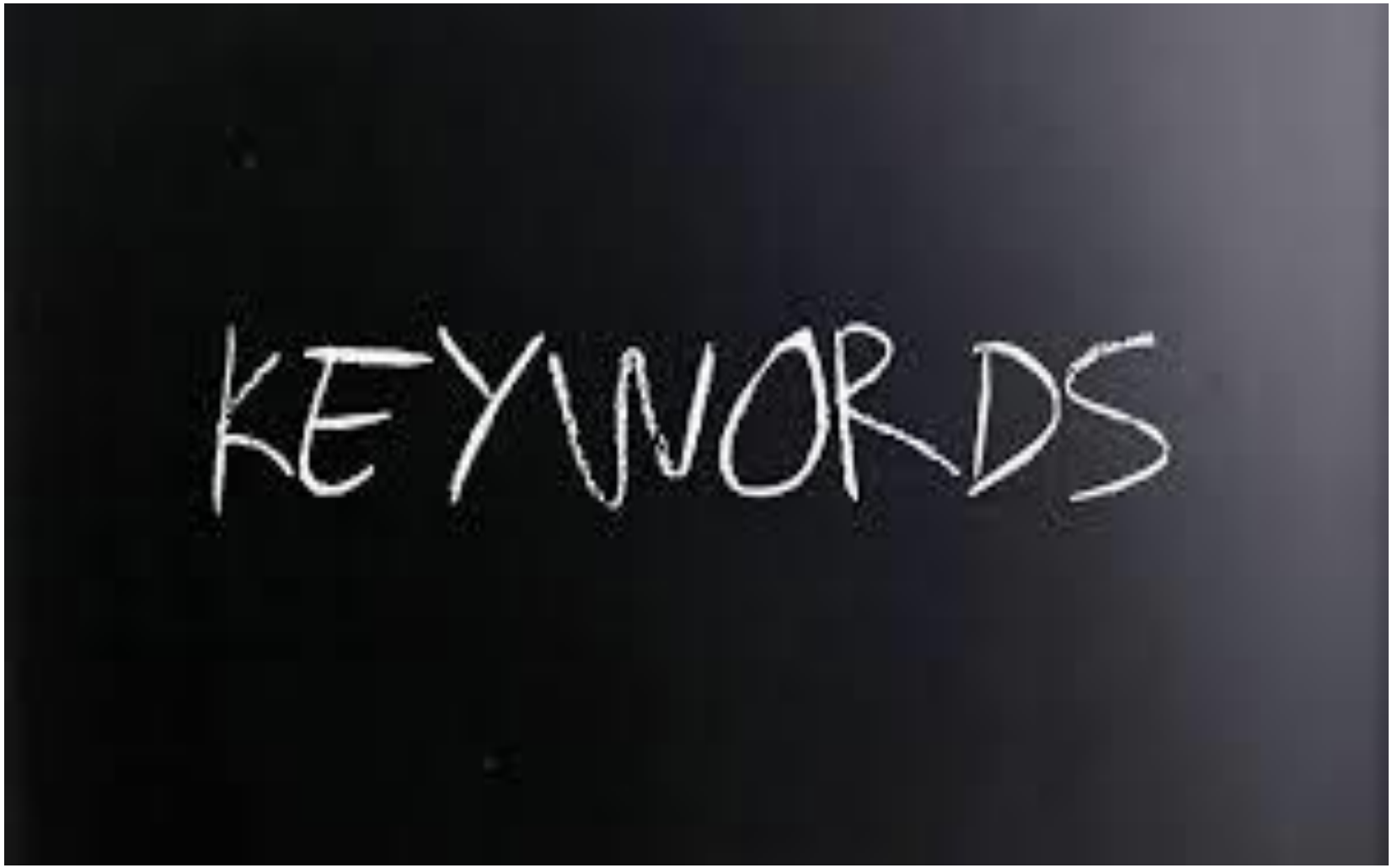
DrinkerBiddle

to find relevant needles…

DrinkerBiddle

But not just one or a few needles…

A flawed, inefficient approach to searching ESI…..

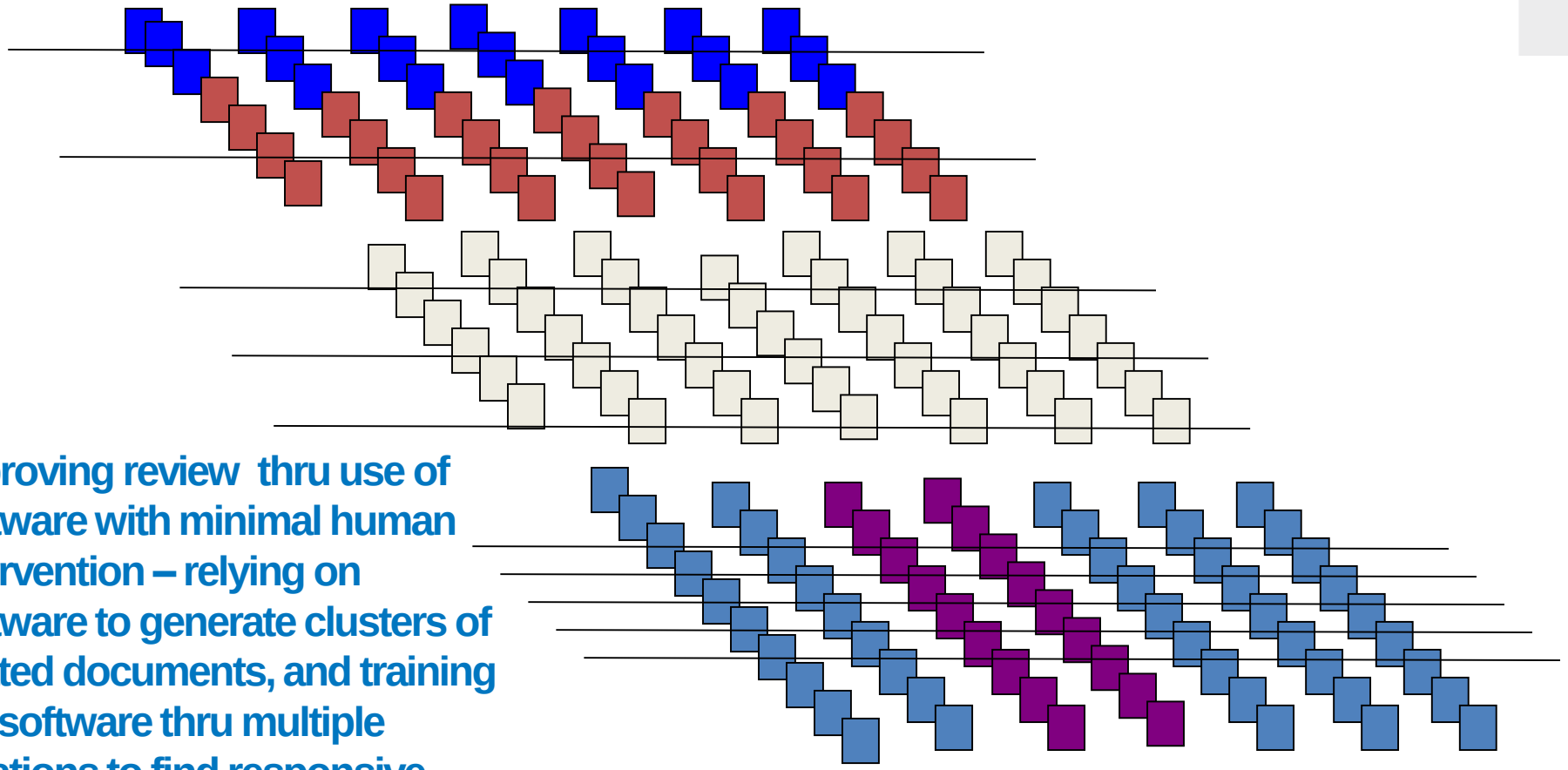


In Search of A Better Way for Lawyers To Search ESI


Automated
Intelligent
Tagging
Coding
Advanced
Prioritization
Ranking
Document
Privilege
Presumptive
Transparent
Assisted
Based
Computer-Assisted
Predictive
Technology
Classification
Adaptive
Automatic
Meaning
AutoSuggest
Review

DrinkerBiddle

E-discovery Strategies: "Predictive Coding"

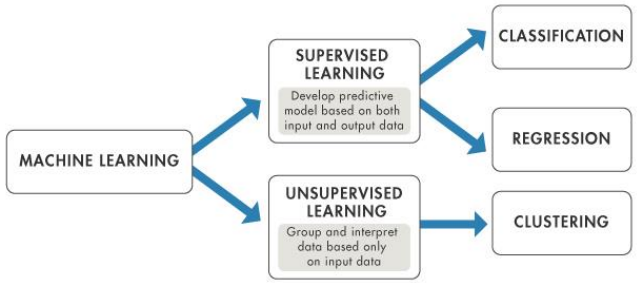


Improving review thru use of software with minimal human intervention – relying on software to generate clusters of related documents, and training the software thru multiple iterations to find responsive records

DrinkerBiddle

Supervised Learning

- Supervised learning is the machine learning task of inferring a function from labeled training data.
- In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal).
- A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.



See Grossman, M.R., Cormack, V., "A Tour of TAR," chap. 3 in J.R. Baron et al., eds., PERSPECTIVES ON PREDICTING CODING And Other Advanced Search Methods for the Legal Practitioner (ABA 2016).

Hastie, T., Tibshirani, R., Friedman, J. THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE AND PREDICTION. Springer 2017)

Acknowledgement: Slide from Nathaniel Payne, Ph.D. candidate, Univ. of Br. Columbia, joint presentation at 2017 IEEE Big Data Conference, Computational Archives Workshop, Boston, MA (Dec. 2017).

DrinkerBiddle

Judicial endorsement of predictive analytics in document review by Judge Peck in *da Silva Moore v. Publicis Groupe* (SDNY Feb. 24, 2012)

This opinion appears to be the first in which a Court has approved of the use of computer-assisted review. . . . What the Bar should take away from this Opinion is that computer-assisted review is an available tool and should be seriously considered for use in large-data-volume cases where it may save the producing party (or both parties) significant amounts of legal fees in document review. Counsel no longer have to worry about being the 'first' or 'guinea pig' for judicial acceptance of computer-assisted review . . . Computer-assisted review can now be considered judicially-approved for use in appropriate cases.

DrinkerBiddle

The Future of Email Archives: Four Propositions

- *Manual approaches to preserving e-mail result in noncompliance with recordkeeping obligations, and should be abandoned in favor of adopting automated rules which minimize end user involvement.*
- *Capture technologies should be implemented to lower enterprise risk of loss and for optimum preservation of a historical record, including for business purposes.*
- *Deployment of advanced search methods (e.g., machine learning) will greatly aid in providing more expeditious access to relevant materials in ever more voluminous collections.*
- ***Sensitivity review needs to be embedded in email archiving workflows, to enable timely future access by researchers and civil society.***

DrinkerBiddle

NARA 1601 on Screening Records (2002)

- Screen records if there is a reasonable chance that they may contain information about a living individual that reveals details of a highly personal nature, which if released would constitute a clearly unwarranted invasion of privacy. Withhold such information from files before disclosure if it has not been officially released previously or if it relates to individuals less than 75 years old or events that occurred less than 75 years before the date of screening.

Source: https://www.archives.gov/foia/directives/nara1601.pdf

DrinkerBiddle

Anticipating the need to filter sensitive content of all types

DrinkerBiddle

Categories of Sensitive Info in Records (non-exhaustive)

- Personally identifiable information (PII)
  - Names as metadata fields; social security numbers; telephone numbers; driver's license information; taxpayer information; bank or financial account information'; credit card numbers; vehicle registrations; dates of birth, height and weight characteristics; asset information
- Personal health information (PHI)
- Arrest records
- National security & law enforcement investigations
- Intellectual property
- FOIA exempt material (in U.S., nine categories)
- Material covered by a legal privilege

DrinkerBiddle

Types of sensitive content

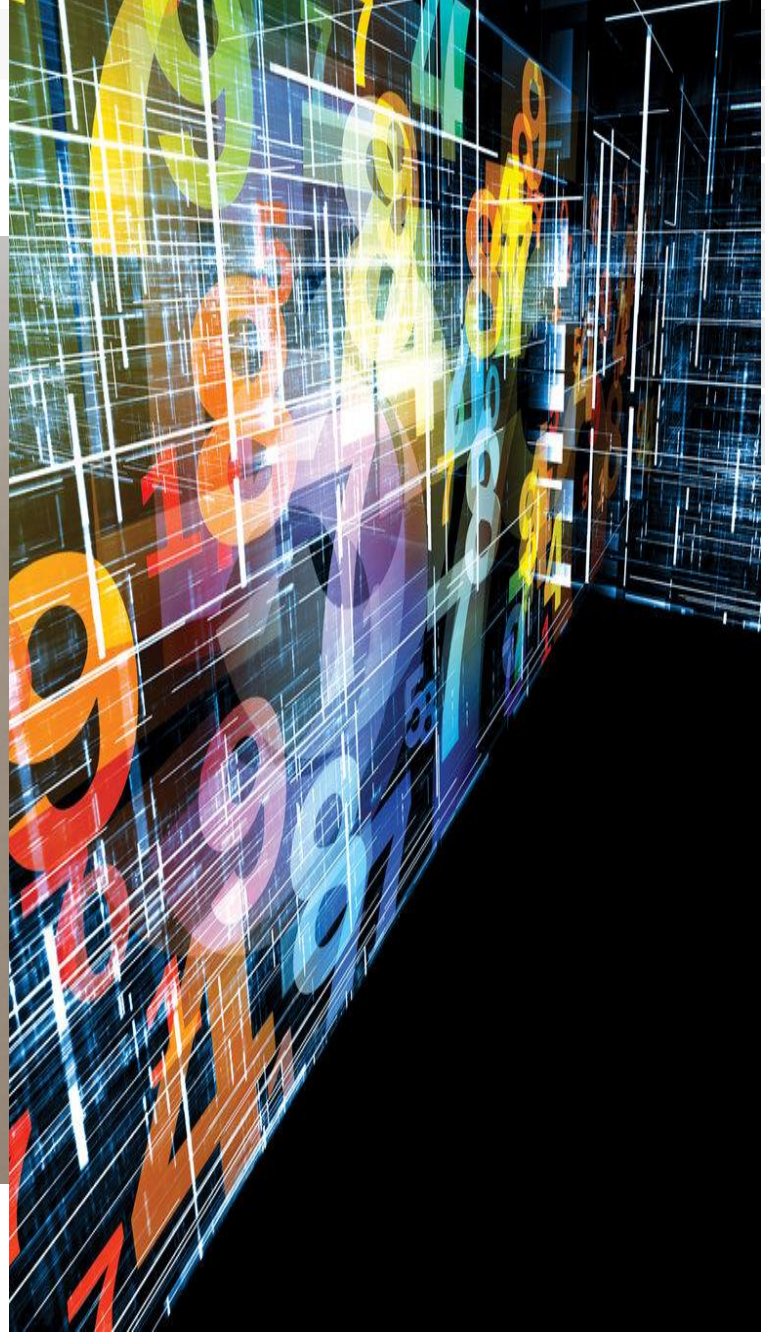
- Content susceptible to deletion through the use of “regular expressions”
- Textual content
- Nontextual content (images, fingerprints, biometric data, geospatial imaging)

DrinkerBiddle

Research agenda

- Signal to noise ratio: analyzing Capstone's success in capturing (and only capturing) e-mail records from senior officials that are of permanent value;
- Appraising existing archival autocategorization strategies for distinguishing between record series & in filtering sensitive content; and
- Evaluating how well total capture of email has increased the right of citizen access to these records, through informal and formal means.

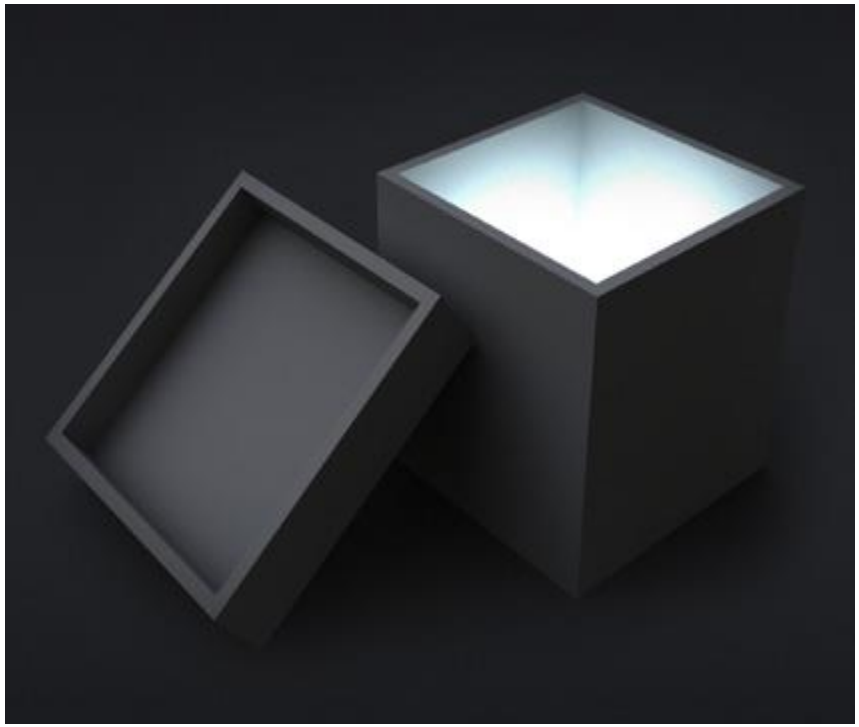
DrinkerBiddle



Faith in Analytics

DrinkerBiddle

Black boxes: how do algorithms separate wheat and chaff?



DrinkerBiddle

To Advance the Cause of Access to Future Archives, Lawyers & RM & IT & Engineers & Others Need To Cross Intellectual Boundaries

DrinkerBiddle

Culture change is possible (even for large institutions, public and private)

DrinkerBiddle

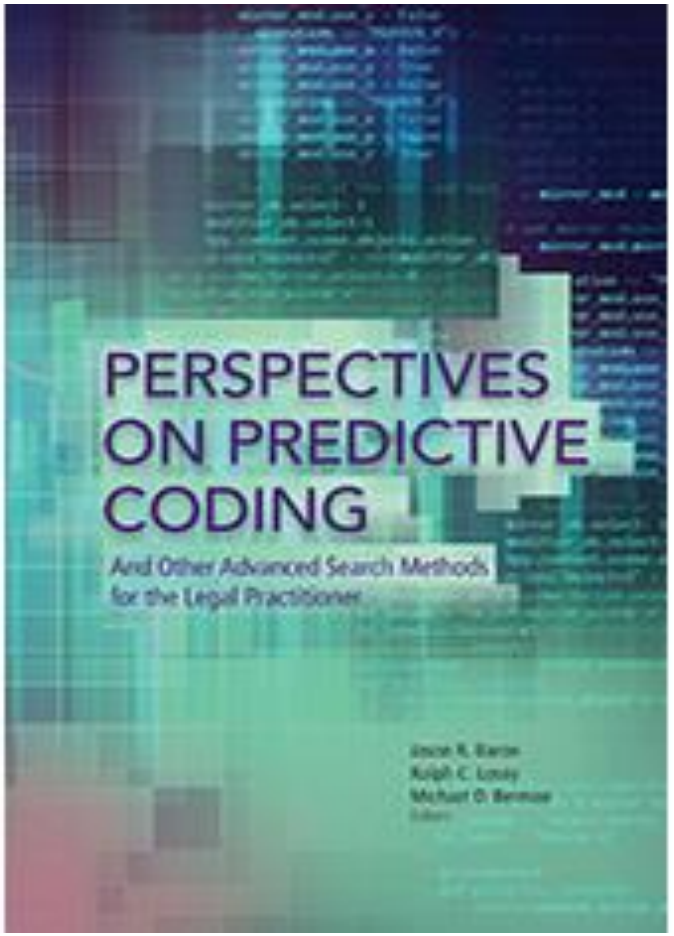
The Coming Age of Dark Archives (i.e., the inability to provide access unless we have smart ways of extracting signal from noise, including use of privacy filters)

DrinkerBiddle

Windows for Citizen Access to Dark Archives….

DrinkerBiddle

Perspectives on Predictive Coding



Jason R. Baron
Drinker Biddle LLP
jason.baron@dbr.com
202.230.5196
Twitter: @jasonrbaron1