eDiscovery and Data Analytics
History and Current Legal Applications
Jay Brudz, Chair, Information Governance and eDiscovery

Northeast General Counsel Seminar
September 19-20, 2016
Philadelphia
History of eDiscovery

1938: FRCP established: emphasis on discovery

1971: 1st ARPANET email sent

1991: MS Mail for PC

1999: fen-phen – early large case with emails at issue

2002: Arthur Andersen charged with obstruction for deletion of documents

2004: Zubulake V: Attorneys must oversee preservation of emails

2006: FRCP Amendments: Rule 26(f) Conference, definition of ESI

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding

1971: 1st ARPANET email sent

1999: fen-phen – early large case with emails at issue

2004: Zubulake V: Attorneys must oversee preservation of emails

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding

1938: FRCP established: emphasis on discovery

1971: 1st ARPANET email sent

1991: MS Mail for PC

1999: fen-phen – early large case with emails at issue

2002: Arthur Andersen charged with obstruction for deletion of documents

2004: Zubulake V: Attorneys must oversee preservation of emails

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding

1938: FRCP established: emphasis on discovery

1971: 1st ARPANET email sent

1991: MS Mail for PC

1999: fen-phen – early large case with emails at issue

2002: Arthur Andersen charged with obstruction for deletion of documents

2004: Zubulake V: Attorneys must oversee preservation of emails

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding

1938: FRCP established: emphasis on discovery

1971: 1st ARPANET email sent

1991: MS Mail for PC

1999: fen-phen – early large case with emails at issue

2002: Arthur Andersen charged with obstruction for deletion of documents

2004: Zubulake V: Attorneys must oversee preservation of emails

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding

1938: FRCP established: emphasis on discovery

1971: 1st ARPANET email sent

1991: MS Mail for PC

1999: fen-phen – early large case with emails at issue

2002: Arthur Andersen charged with obstruction for deletion of documents

2004: Zubulake V: Attorneys must oversee preservation of emails

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding

1938: FRCP established: emphasis on discovery

1971: 1st ARPANET email sent

1991: MS Mail for PC

1999: fen-phen – early large case with emails at issue

2002: Arthur Andersen charged with obstruction for deletion of documents

2004: Zubulake V: Attorneys must oversee preservation of emails

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding

1938: FRCP established: emphasis on discovery

1971: 1st ARPANET email sent

1991: MS Mail for PC

1999: fen-phen – early large case with emails at issue

2002: Arthur Andersen charged with obstruction for deletion of documents

2004: Zubulake V: Attorneys must oversee preservation of emails

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding

1938: FRCP established: emphasis on discovery

1971: 1st ARPANET email sent

1991: MS Mail for PC

1999: fen-phen – early large case with emails at issue

2002: Arthur Andersen charged with obstruction for deletion of documents

2004: Zubulake V: Attorneys must oversee preservation of emails

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding

1938: FRCP established: emphasis on discovery

1971: 1st ARPANET email sent

1991: MS Mail for PC

1999: fen-phen – early large case with emails at issue

2002: Arthur Andersen charged with obstruction for deletion of documents

2004: Zubulake V: Attorneys must oversee preservation of emails

2010: Pension Committee

2015: FRCP Amendments: proportionality, data retention policies, preservation

2012: Da Silva Moore: endorses predictive coding
Snapshot 1995

- Paper-based discovery
- eMail not always considered in scope
- “Full Contact” discovery
For 19 years, LEGALTECH presents extraordinary opportunities for lawyers and law office decision-makers to learn about the latest developments in law practice technology for the legal profession. LEGALTECH brings together hundreds of lawyers, general counsel, IT directors, CIOs, CFOs and administrators for an unparalleled two or three day educational experience designed exclusively for the legal professional. The LEGALTECH exposition features product demonstrations and presentations by the leading companies in their field. These displays boast the newest, most innovative and sophisticated products available for the legal professional.

At LEGALTECH, you'll find the conference offers the practical information you seek, while the exhibit hall provides a hands on venue to experience the latest technology solutions. The

- eMail often blown back to paper
- Coding done by paralegals
- Lots of scanning
Snapshot 2005

- ~$3200/GB processing
- Associate review common
- Important for large cases
Snapshot 2010

- $450/GB processing
- Staff attorney or contract attorney review
- 40-60 custodians common
- Over-preservation rampant
Snapshot 2016

- 10-15 custodians
- Proportionality in preservation and search
- Contract attorney or TAR approaches acceptable
- Penetrated to almost all cases
- Hot topics:
  - IoT
  - Big Data
  - Analytics/TAR
Summary of Trends

• Per unit costs have steadily dropped over the years
• Review costs have dropped to commoditized labor price points
• Volumes increased from 1995 to around 2015, traditional document volumes now dropping due to emphasis on proportionality
• Data sources besides emails and documents are becoming more important
  • Mobile
  • Social
  • Online databases
  • Big data
  • IoT
• eDiscovery is permeating almost all cases (large and small, federal and state, all areas of law)
Key: Review is difficult to manage and is still expensive and time consuming

Electronic Discovery Reference Model / © 2014 / v3.0 / edrm.net
Typical Volumes

- 10-15 Custodians
- 5-20 GB / Custodian
- 4,000-15,000 Documents / GB
- Major Variables
  - Cull rate
  - Responsiveness
  - Privilege
  - Redactions
Technology Assisted Review
TAR: Technology for better/faster/cheaper review

- Near Duping
- Threading
- Searches (fuzzy, etc)
- Clustering
- Sampling
- Dataviz
- Predictive Coding
  - Unsupervised learning
Predictive Coding

- Prioritizing or coding a collection of electronic documents
- Harnesses human judgments
- One or more subject matter experts
- Smaller set of documents
- Extrapolates judgments
Functional Overview of Predictive Coding
TAR/PC Case Study

HSR 2\textsuperscript{nd} Request – 2 Major Retailers

- Competitor sought to prevent the merger through hostile takeover
- HSR approval critical to closing deal
- Used TAR and Predictive Coding to identify, collect, and review seven million documents in seven weeks
- $9.1 BB merger approved
An AI Koan -

The Uncarved block
In the days when Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6. "What are you doing?", asked Minsky. "I am training a randomly wired neural net to play Tic-tac-toe", Sussman replied. "Why is the net wired randomly?", asked Minsky. "I do not want it to have any preconceptions of how to play", Sussman said. Minsky then shut his eyes. "Why do you close your eyes?" Sussman asked his teacher. "So that the room will be empty."

-At that moment, Sussman was enlightened.
The Seed Set

**Random selection**: documents selected at random from the collection, coded as relevant or not, and fed to the learning system

- Problem: *may miss certain types of relevant docs that are rare in the collection and therefore unrepresented in a random sample*

**Judgmental selection**: documents are selected by an expert, using prior knowledge, or ad hoc searches, or other advanced tools, and then fed to the learning system

- Problem: *may miss certain types of relevant documents that are overlooked by the expert, or not identified by the search and analytic tools*

Source: M. Grossman ABA Slideset 2012
Summary: TAR/Predictive Coding

- Court acceptance is increasing
- Approaches vary
- Differing levels of engagement required
- Generally save money and effort
- Can be applied in non-litigation contexts
- Can be incorporated into traditional workflows
Analytics
Information is Everything

- Analytics:
  - Gets to an answer faster
  - Knowing your position resolves your case faster
  - You can prove a negative
  - Dramatically lower cost
An Analytics Primer

- Analytics identifies patterns in data; and thus patterns in human conduct.
- Descriptive analytics looks back in time to tell you what happened.
- Predictive analytics determines the probability of a future event, or the probability that a new thing is like a thing already seen in the data.
Structured Data Analytics - Example

- Class action suit for alleged SMS text messages in violation of the TCPA.
- Multiple SQL databases ETL challenge
- Advanced data analytic and modelling techniques to accurately categorize and identify text messages falling within class period
- Settled
Where are we going?
Predictive Compliance
Predictive Compliance – Objectives

- Identify patterns associated with misconduct
- Detect these patterns in email and unstructured data
- Develop models to be used in an early warning system to detect and prevent misconduct as it develops
Text Mining
Predictive analytics to improve information governance

- RM: defensible disposal of low value information
- Regulatory compliance
- Risk mitigation – segregating sensitive materials
  - (PII, proprietary, etc.)
- Business intelligence
- E-discovery
- Collaboration across enterprise
- Providing access to dark data & archives
IoT - Hypothetical

• You manufacture heavily censored ATSDD’s- data gathered includes:
  – Water quality, chemistry, and temperature; Number, speed and \( \Delta v \) of dunks; provenance of tea from barcode on teabag label; Voltages, resistances and temperatures of all electrical parts

• From this data, you are able to derive many useful and marketable inferences, such as:
  – What combination of tea, water, temperature, and dunking style produces the best cup of tea as measured by consumer “likes” correlated with product lot #
  – How often does maintenance need to be done on the ATSDD – lubricating, changing motor brushes, etc.

• Marketing decides to aggregate and monetize this data by selling it to your “Platinum Program” customers.
IoT – Hypothetical Issues

- Analytics reveals that a certain combination of water, tea and dunk style, cuts motor brush life in half, well short of your standard maintenance window. This can cause a massive explosion.
- Do you have to give this information to non “Platinum Program” customers?
- Other issues?
  - How do you identify, preserve, review and produce this data in litigation?
  - How do you produce a trained algorithm?
Questions

Jay Brudz
Chair, Information Governance and eDiscovery
Drinker Biddle & Reath LLP
1500 K Street, N.W.
Washington, DC 20005-1209
(202) 230-5194 office
Jay.Brudz@dbr.com