



PREDICTIVE CODING

Bringing Text Analytics to the Courtroom

15.071 – The Analytics Edge

Enron Corporation



- U.S. energy company from Houston, Texas
- Produced and distributed power
- Market capitalization exceeded \$60 billion
- *Forbes*: Most Innovative U.S. Company, 1996-2001
- Widespread accounting fraud exposed in 2001
 - Led to bankruptcy, the largest ever at that time
 - Led major accounting firm Arthur Andersen to dissolve
- Symbol of corporate corruption

California Energy Crisis

- California is most populous state in United States
- In 2000-2001, plagued by blackouts despite having plenty of power plants
- Enron played a key role in causing crisis
 - Reduced supply to state to cause price spikes
 - Made trades to profit from the market instability
- Federal Energy Regulatory Commission (FERC) investigated Enron's involvement
 - Eventually led to \$1.52 billion settlement
 - Topic of today's recitation

The eDiscovery Problem

- Enron had millions of electronic files
- Leads to the *eDiscovery* problem: how we find files relevant to a lawsuit?
 - In legal parlance, searching for *responsive* documents
- Traditionally, keyword search followed by manual review
 - Tedious process
 - Expensive, time consuming
- More recently: *predictive coding (technology-assisted review)*
 - Manually label some of the documents to train models
 - Apply models to much larger set of documents

The Enron Corpus

- FERC publicly released emails from Enron
- > 600,000 emails, 158 users (mostly senior management)
- Largest publicly available set of emails
- Dataset we will use for predictive coding
- We will use labeled emails from the 2010 Text Retrieval Conference Legal Track
 - *email* – text of the message
 - *responsive* – does email relate to energy schedules or bids?

Predictive Coding Today



- In legal system, difficult to change existing practices
 - System based on *past precedent*
 - eDiscovery historically performed by keyword search coupled with manual review
- 2012 U.S. District Court ruling: predictive coding is legitimate eDiscovery tool
- Use likely to expand in coming years