

Document Analysis Methods for Fraud Detection

Marc B. Vilain

781-271-2151 • mbv@mitre.org

John D. Burger

781-271-8784 • john@mitre.org

Conrad Chang

781-271-5584 • conrad@mitre.org

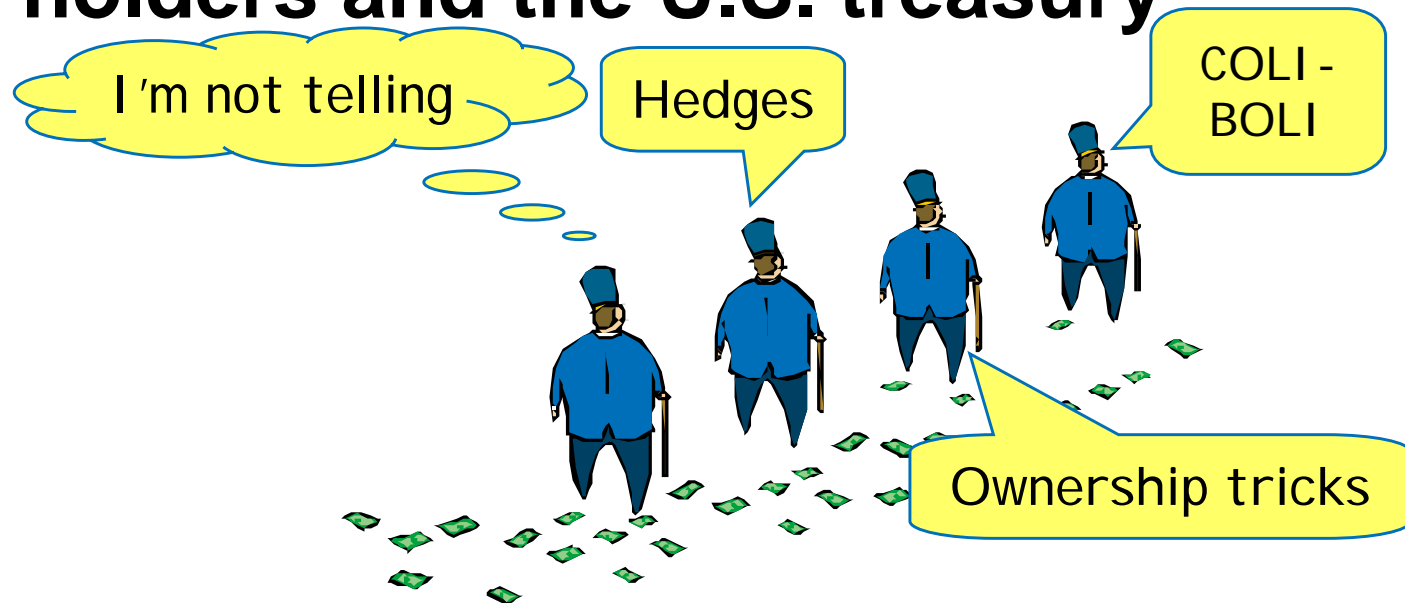
MITRE Sponsored Research

**MITRE
Technology
Program**

MITRE

Problem

- Key fact: annual uncollected tax revenue due to corporate fraud: **\$75-85 billion**
- Steady supply of new schemes defraud both share holders and the U.S. treasury



- Need financial “all-source” analysis

Background

- Investigation of financial improprieties is resource limited: few investigators, highly technical work, specialized skills
- Staggering volumes of data: thousands of schedules, hundreds of subsidiaries, and tables that gloss key transactions
- But: open source financial documents may offer clues to fraud and non-compliance



Objective

- **Practical aim: find indicators of financial fraud and non-compliance in financial texts**
- **Research approach: exploratory document exploitation**
- **Method: attack the hard problems with proven human language technologies**
- **Evaluation: establish baseline performance of known techniques, identify and address open needs, assess value to end user**

Activities

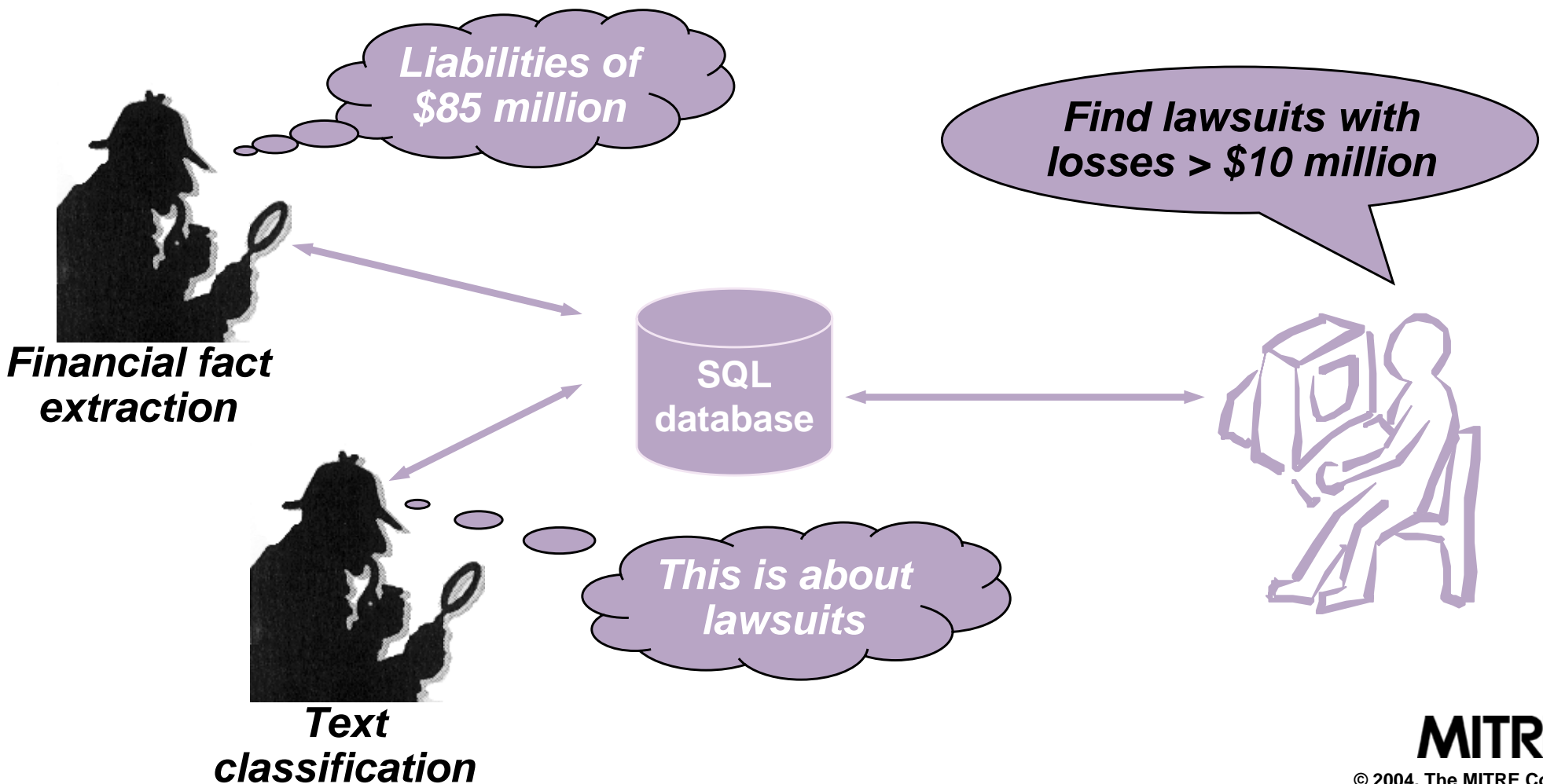
Main goals for Q1/Q2:

- Establish testbed
- Baseline performance

- **Experimental testbed: a document exploitation prototype that integrates analysis methods**
- **Text classification: identify subject matter of text segments through word signatures**
- **Fact extraction: delineate financial reportables by understanding the meaning of financial text**
- **End-user needs: assess applicability of methods to actual fraud issues**

Highlight: Demo prototype

- Language-enabled processors add to DB
- Web-based interface for search and retrieval



Highlight: Classification

- **Goal: identify subject area of text segment: debt, derivatives, law suits, mergers, etc.**
- **Method: corpus-based training, based on hand-labeled data at the paragraph level**
 - **Collect “word vector” statistics, e.g., (interest=4 rate=5 swap=3 ...) -> DERIVATIVE**
 - **Model probability distribution with support vector machines (among others)**
- **Good preliminary performance, F=88**
 - **Precise subjects are best, e.g., derivatives**

Impacts

- **Science:** advance foundation of fact detection, text categorization, document analysis, and related human language technologies
- **Application:** prototype tools that support investigation of financial fraud and abusive tax sheltering schemes
- **Spin-off:** transition technological innovations to non-financial problem areas with like needs
- **Community:** disseminate lessons learned, best practices, and methods to sponsors and academe

Future Plans

- **Text classification: refine text categories, extend training data for low-performance categories, tweak algorithms/features**
- **Information extraction: refine repertoire of financial facts, introduce automated training, create training corpus**
- **Document structure modeling: introduce training methods for document segmentation**
- **Evaluation: validate approach through end-user feedback and application tailoring**