

# Document Analysis for Fraud Detection

**Marc Vilain**

781-271-2151 • [mbv@mitre.org](mailto:mbv@mitre.org)

**Conrad Chang**

781-271-5584 • [conrad@mitre.org](mailto:conrad@mitre.org)

MITRE Sponsored Research



# Problem

- **Corporate tax fraud: steady supply of new schemes that defraud the U.S. treasury**
  - **Key fact: annual uncollected tax revenue due to corporate fraud: \$75–85 billion**
- **International terrorism: increasing use of criminal financial methods (money laundering)**
  - **Key fact: “the oft-repeated assertion that bin Laden was funding al Qaeda from his personal fortune was in fact not true”**  
*- 9/11 report*

# Background

- Investigation of financial improprieties is resource limited: few investigators, highly technical work, specialized skills
- Staggering volumes of data



**MITRE**

© 2005, The MITRE Corporation

# Objective

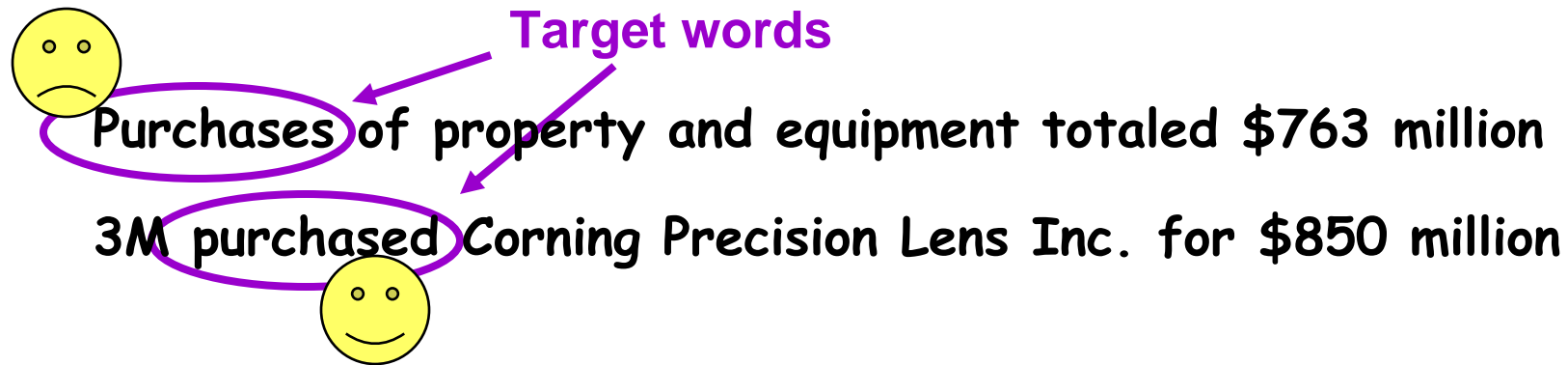
- **Practical aim: find indicators of financial fraud and non-compliance in financial texts**
- **Application approach: mutually leverage human language technologies and data mining through interactive exploration**
- **Scientific contribution: identify and address open technology needs, drawing motivation, data, and examples from key financial tasks**
- **Evaluation: determine baselines for current techniques, measure performance of new methods, and assess value to end-user**

# Activities

- **Maytag prototype: create integrated Web-based interface to SQL database; provides uniform framework for invoking text mining and language processing**
- **Query profiling: model financial tasks by creating cases “on the fly” based on Maytag queries**
- **Trainable fact extraction: apply machine learning techniques to the identification of financial relations and events**
- **Adaptation: increase the ability of language processing algorithms to apply to new tasks**

# Highlight: Training acquisition extraction

## ■ Classify target words: is it an actual acquisition?



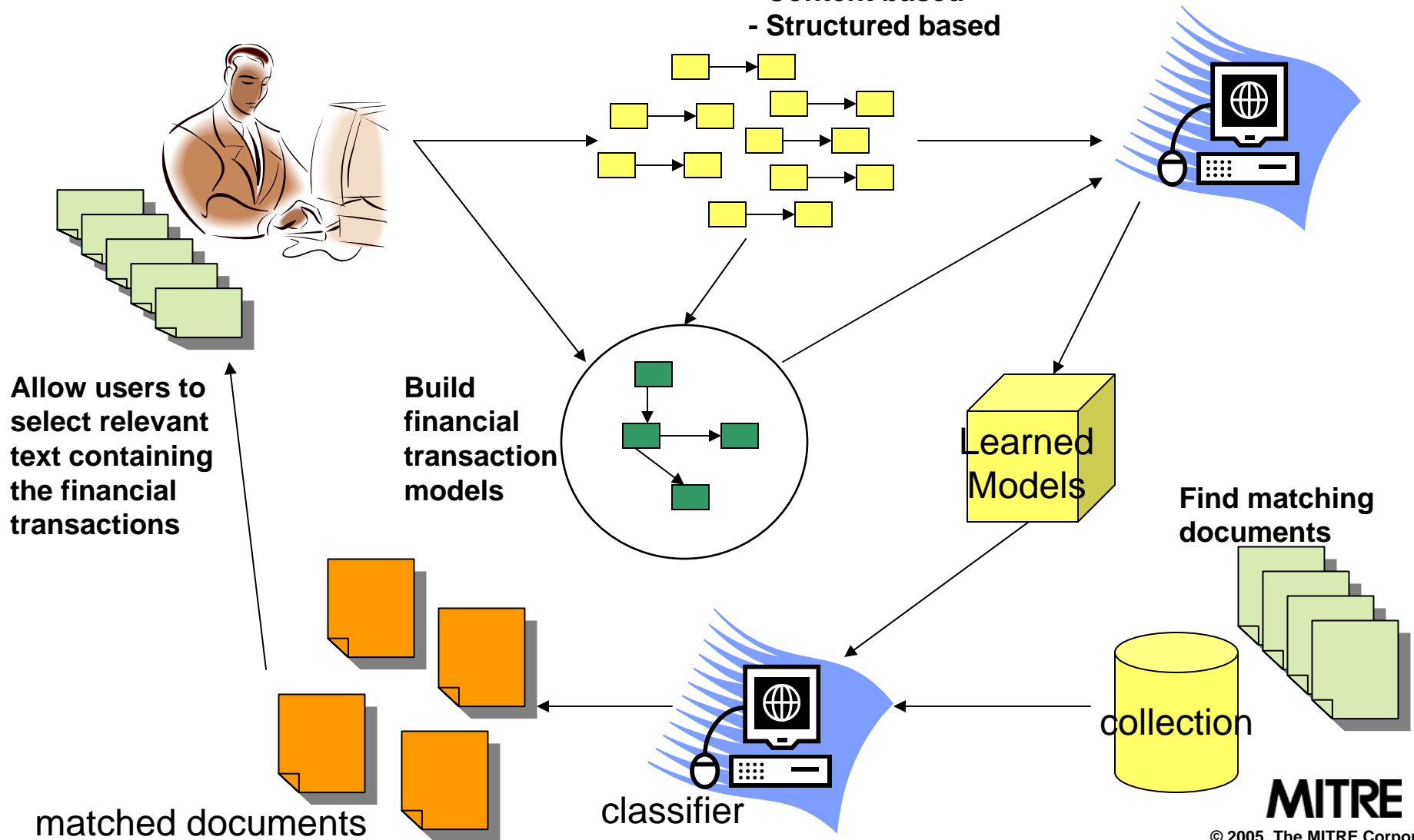
## ■ Classify phrases: which parts of the sentence fill which acquisition role?



# Demonstration: User-Directed Analysis (Query by Profiling, Financial Model Builder)

Extract components related to the transaction

- Content based
- Structured based



**MITRE**

© 2005, The MITRE Corporation

# 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

- **Trainable fact extraction: validate on multiple tasks and data sets; incorporate more refined models of syntax and semantics**
- **Query profiling: validate approach through end-user experiments; refine use of machine learning methods**
- **Adaptation: further experiment with statistical modeling**
- **Tasks and domains: refine through further interactions with sponsoring government agencies; transition to sister projects**