

# Text Extraction and Mining

Marc Vilain

781-271-2151

[mbv@mitre.org](mailto:mbv@mitre.org)

Zohreh Nazeri

703-983-5841

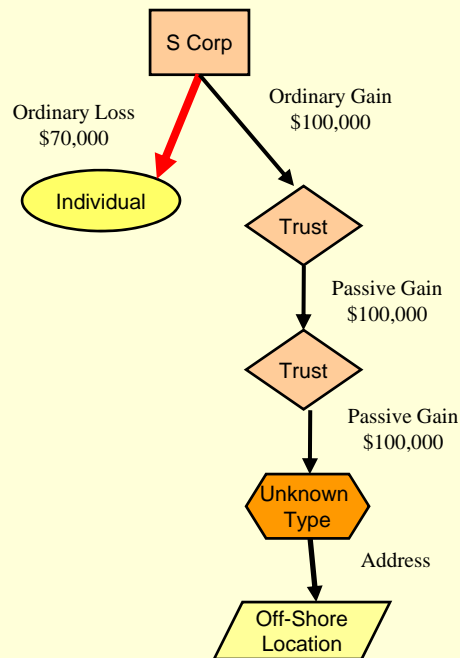
[nazeri@mitre.org](mailto:nazeri@mitre.org)

CEM IR&D

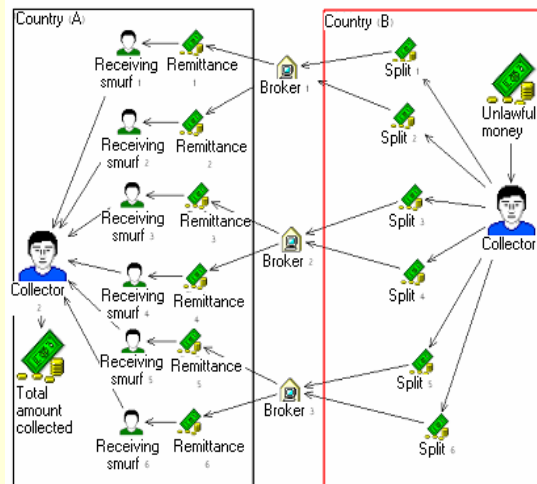


# Problem

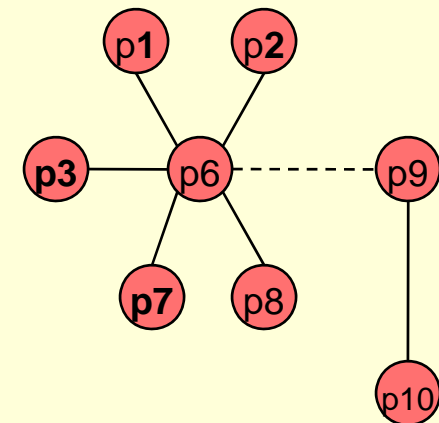
- Many MITRE sponsors are faced with the challenging task of linking together the information embedded in large and disparate sources of data, in a variety of domains.



Tax Evasion



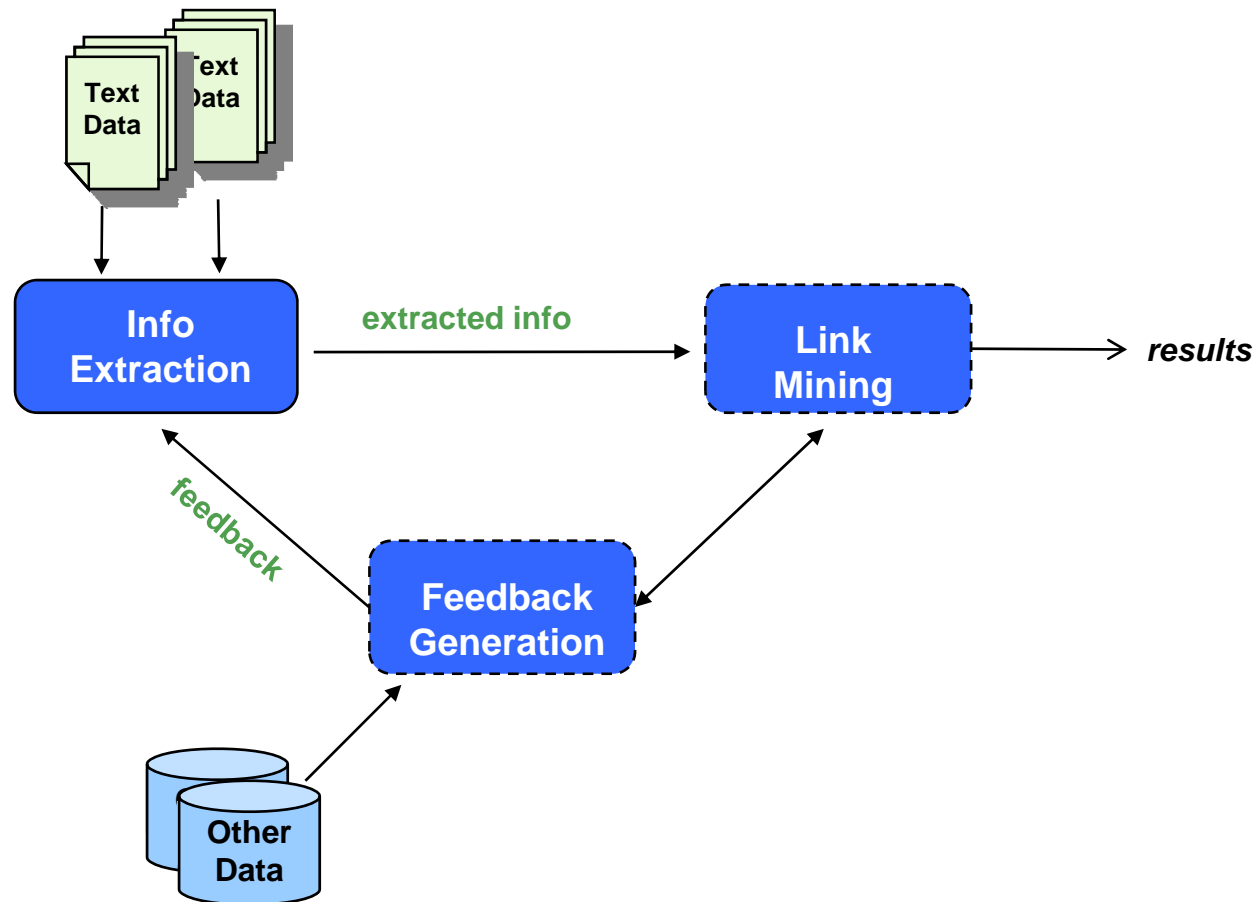
Money Laundering



Social Network Analysis

# Background

- Text Extraction and Mining (TEAM) is a system based on a prototype developed by the Closed-Loop Link Mining MSR.



# Objectives



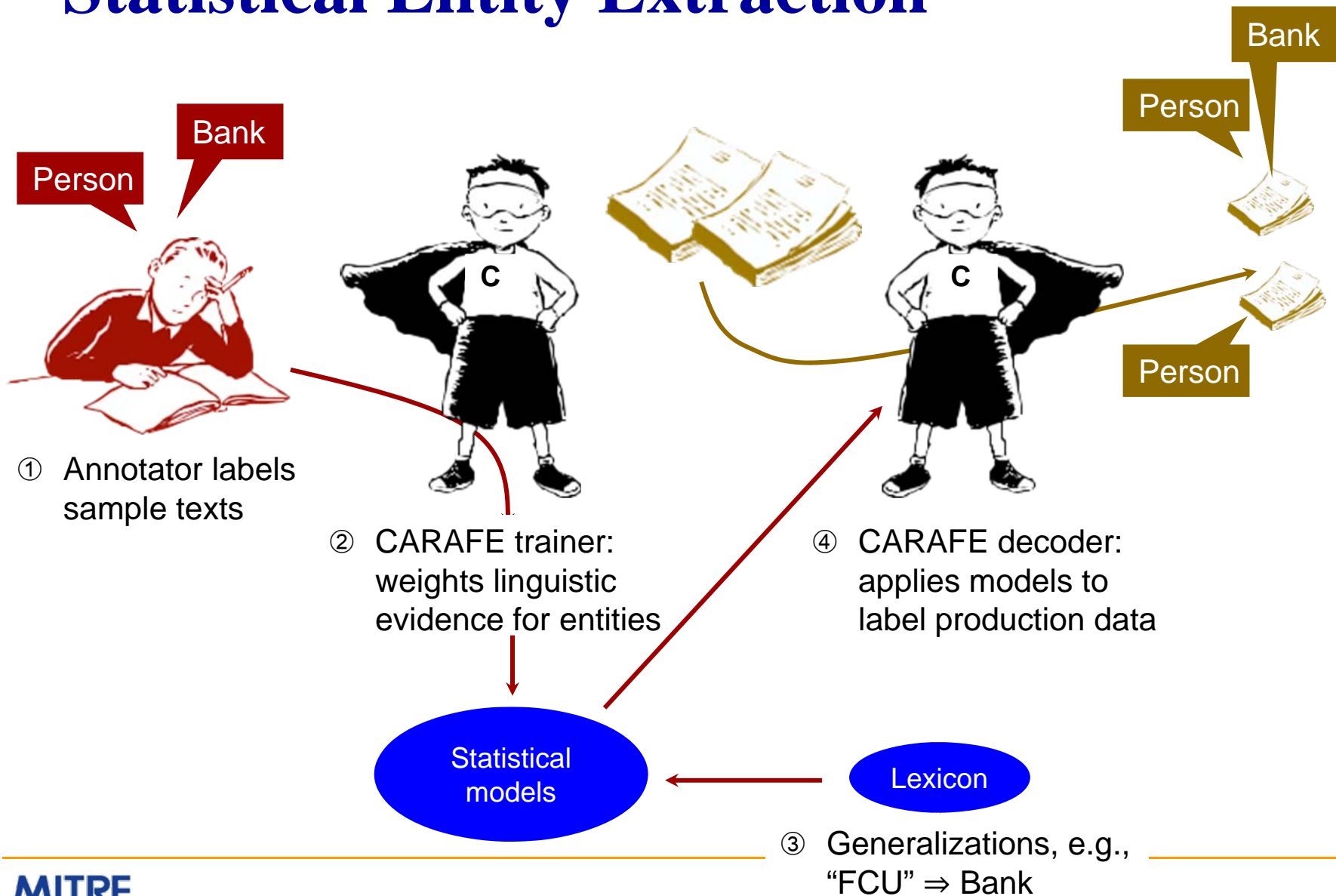
- **Further enhance prototype developed by previous MITRE research (Closed-Loop Link Mining MSR).**
  - **Extraction of entities and events from unstructured text**
  - **Link discovery and mining**
  - **Integrated system.**
  
- **Develop a stand-alone demonstration package**
  - **Sanitized data set**
  - **Vehicle for sponsor outreach.**
  
- **Provide demonstrations to key sponsors.**

# Activities



- **Extended entity extraction for BSA suspicious activity reports (SARs)**
  - **Broadened coverage to SEC SARs**
  - **Annotated additional training data**
  - **Cross-evaluated performance of banking SAR model.**
  
- **Stand-alone package for out-of-lab demonstrations**
  - **Link mining on sanitized extraction database.**
  
- **Outreach**
  - **Demonstrations to MITRE leadership, SEC**
  - **MITRE TEM (classified information extraction)**
  - **CEM TEM (projects interested in link mining).**

# Highlight: Statistical Entity Extraction

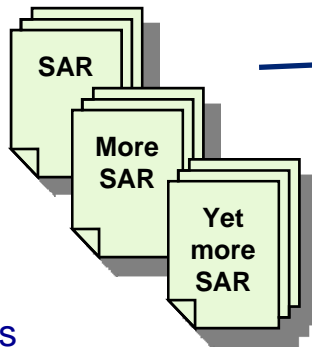


# Demonstration: TEAM prototype



“...Bank of Elbonia act # 123-456-7890 in name of Priscilla Pseudonym...”

“John Q. Scumbuckett makes frequent cash deposits followed by wire transfers ... Wires go to Lenny Launderer, Bank of Elbonia act # 123-456-7890”



Unstructured text

Entities and events

Link structures



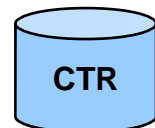
John Q. Scumbuckett

Act# 123-456-7890  
Bank of Elbonia

Lenny Launderer

Priscilla Pseudonym

\$ 1/16  
\$ 1/18



Structured databases

Lenny Launderer	DEP	1/16
Lenny Launderer	DEP	1/18

# Impacts



- **Sponsor missions**
  - **Treasury Dept: state-of-the-art extraction methods capture entities and events in BSA suspicious activity reports**
  - **Securities and Exchange Commission: initial capability for processing SEC-filed SARs**
  - **Internal Revenue Service: new approaches to prosecuting tax fraud through BSA data.**
- **MITRE's broader sponsor base**
  - **Broad need to analyze disparate sources of data in order to discover links between embedded information**
  - **The TEAM system can help with this challenging task.**

# Future Plans



- **Short-term objectives achieved**
  - **Information extraction modules: largely stable**
  - **Graph mining: configured to key tasks**
  - **Prototype: stand-alone out-of-lab demo.**
  
- **Follow-on work**
  - **Shift from R&D only to mission support backed up by R&D**
  - **Primary sponsor missions: SEC, IRS, Treasury**
  - **Potential MITRE IR&D role: refine extraction models, simplify code base.**