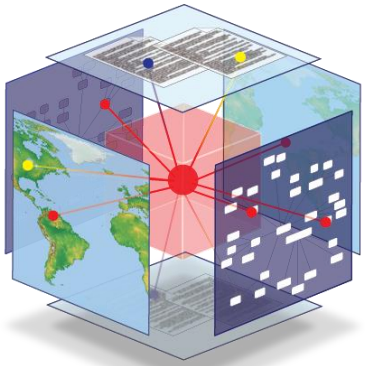


VizLinc

Kara Greenfield, William Campbell, Joel Acevedo-Aviles

GraphEx 2014

8/21/2014

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

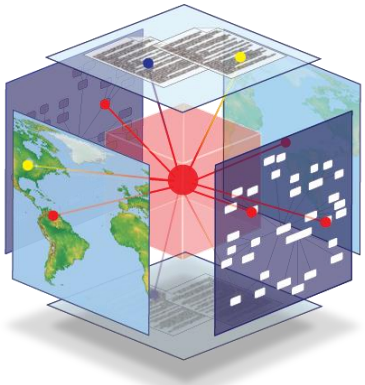
VizLinc

Software suite that integrates automatic information extraction, search, graph analysis and geo-location for interactive visualization and exploration of a text collection

- Designed for unstructured text
- Composed of two desktop applications:
  - **Ingestion Tool** : Java-based application to pre-process *unstructured* text documents
  - **Graphical User Interface:** *Gephi* plugin that provides visualization, search, geo-location, graph analytics…
- Standalone
  - Can operate completely offline
  - No specialized hardware
- Open Source
  - User Interface: https://github.com/mitll/vizlinc
  - Ingestion tool https://github.com/mitll/vizlinc_ingester

VizLinc: Goals

- **Data Characterization**
  - Understanding the type of information the data set under study contains
- **Making patterns and connections between entities evident**
- **Narrow down the corpus**
  - Ideally: to a small fraction that users can quickly read



VizLinc

System for Content + Context Analysis

Documents

Twitter

Newswire

Reddit

Wikipedia

Text Entity Extraction

Pedro worked near *Rio Grande*

Topics

Love

Religion

Holidays

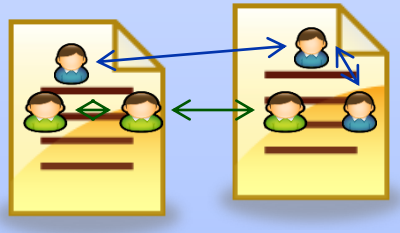
Money

Location Extraction/ Geocoding

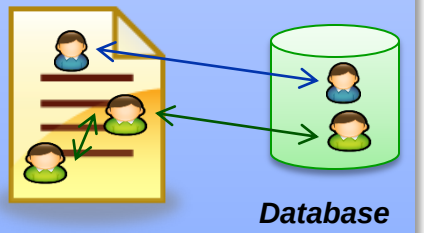
Single-Message Content Analytics

Social Networks

Entity Coref

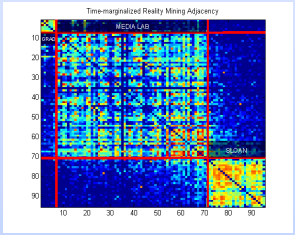


Entity Linking



Knowledge-Base Construction

Community Detection



Leadership Prediction

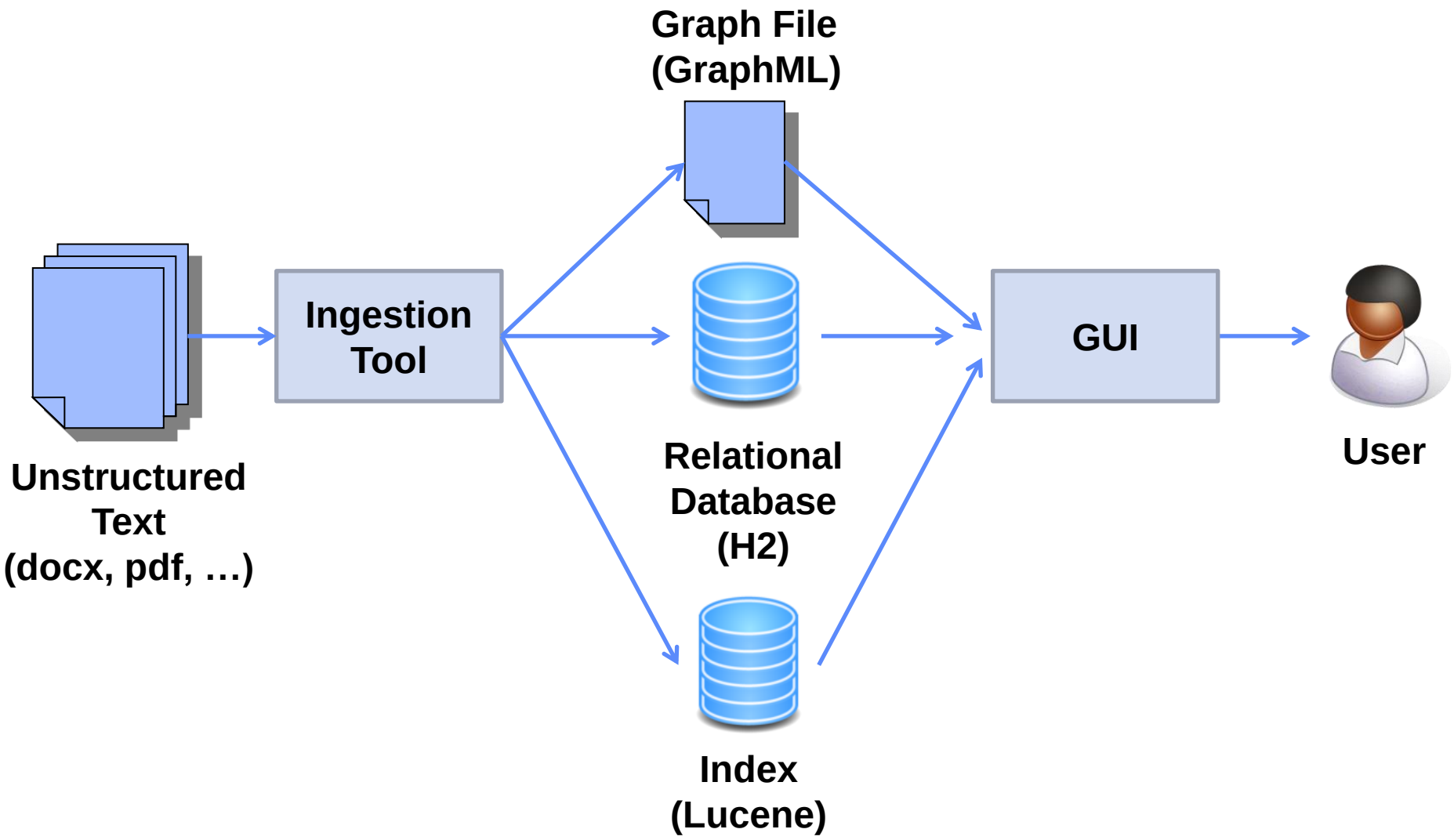
Efficient Search

High-Level Analytics

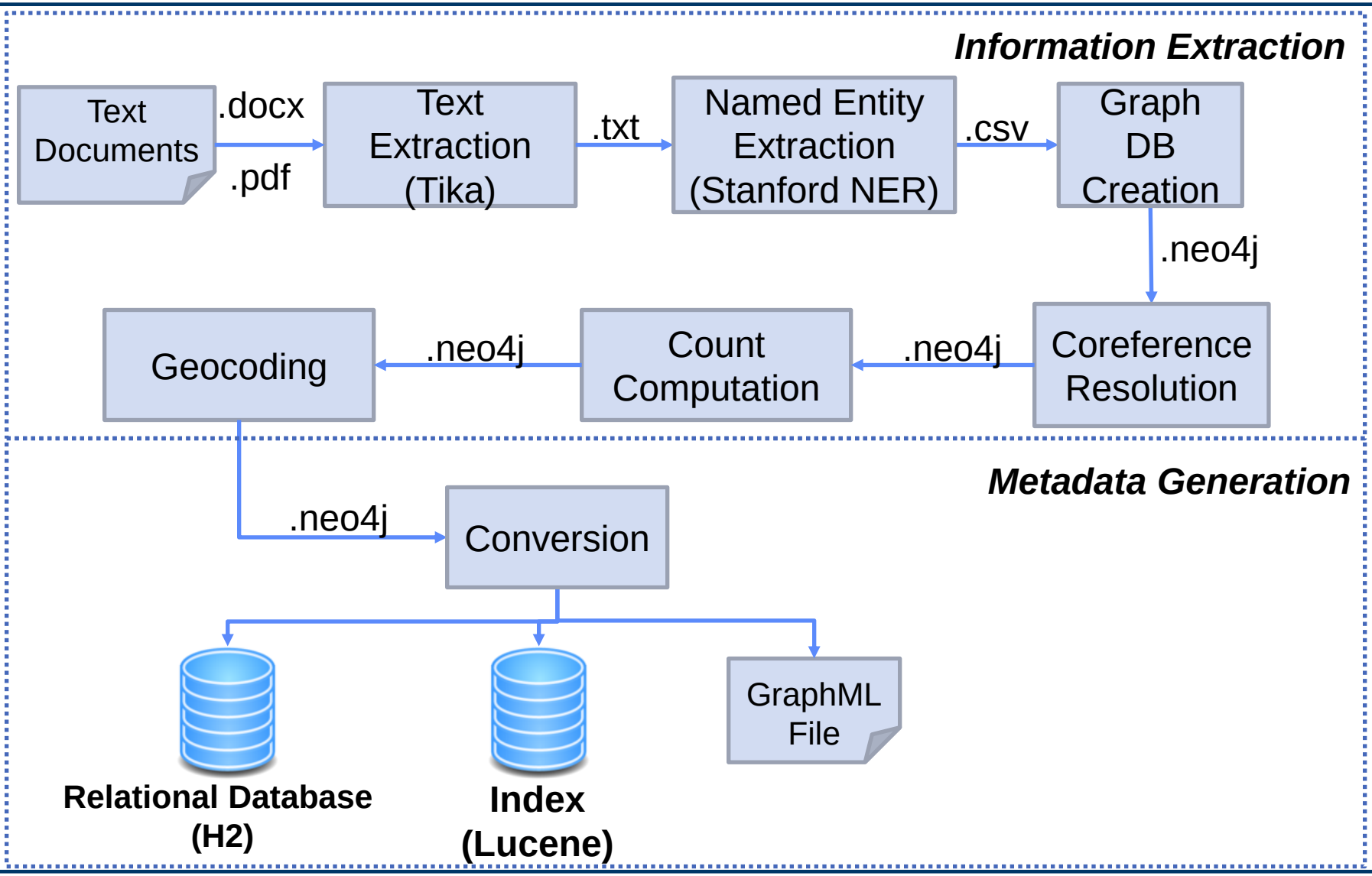
User

http://www.github.com/mitll/vizlinc

System Overview



Ingestion Tool



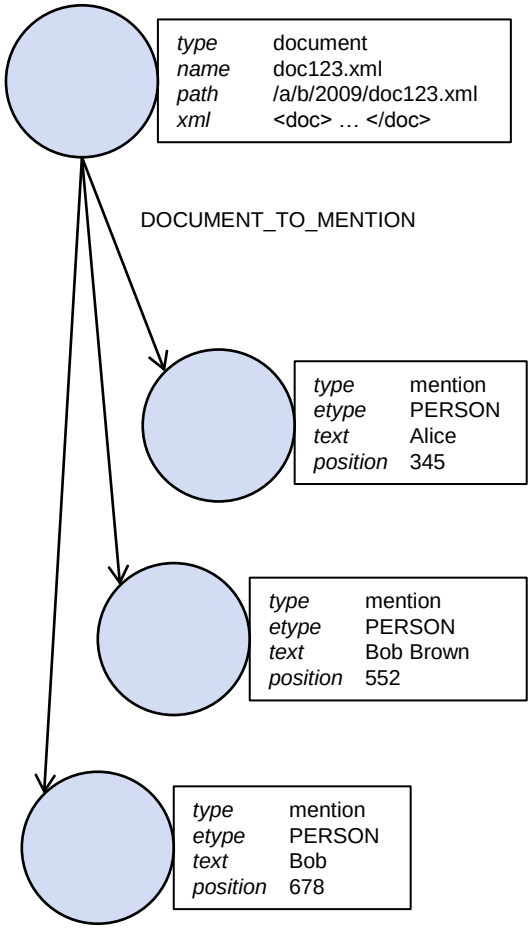
Database Implementation for VizLinc Tool

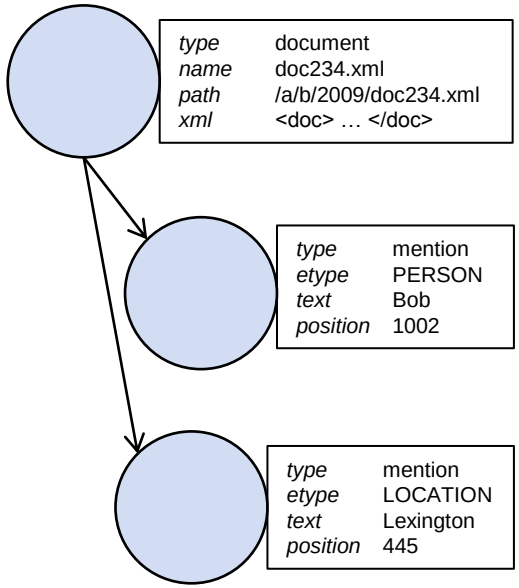
	Search Time	Flexible Schema	Storage Scalability	Graph Ops
SQL DB	Y	Y	G	Y
Graph DB	R	G	G	Y
In-Memory Data Structure	G	R	R	G

- Goals for Database storage and operations:
  - Standalone operation—no enterprise structure
  - Responsive user queries
  - Moderate size data set

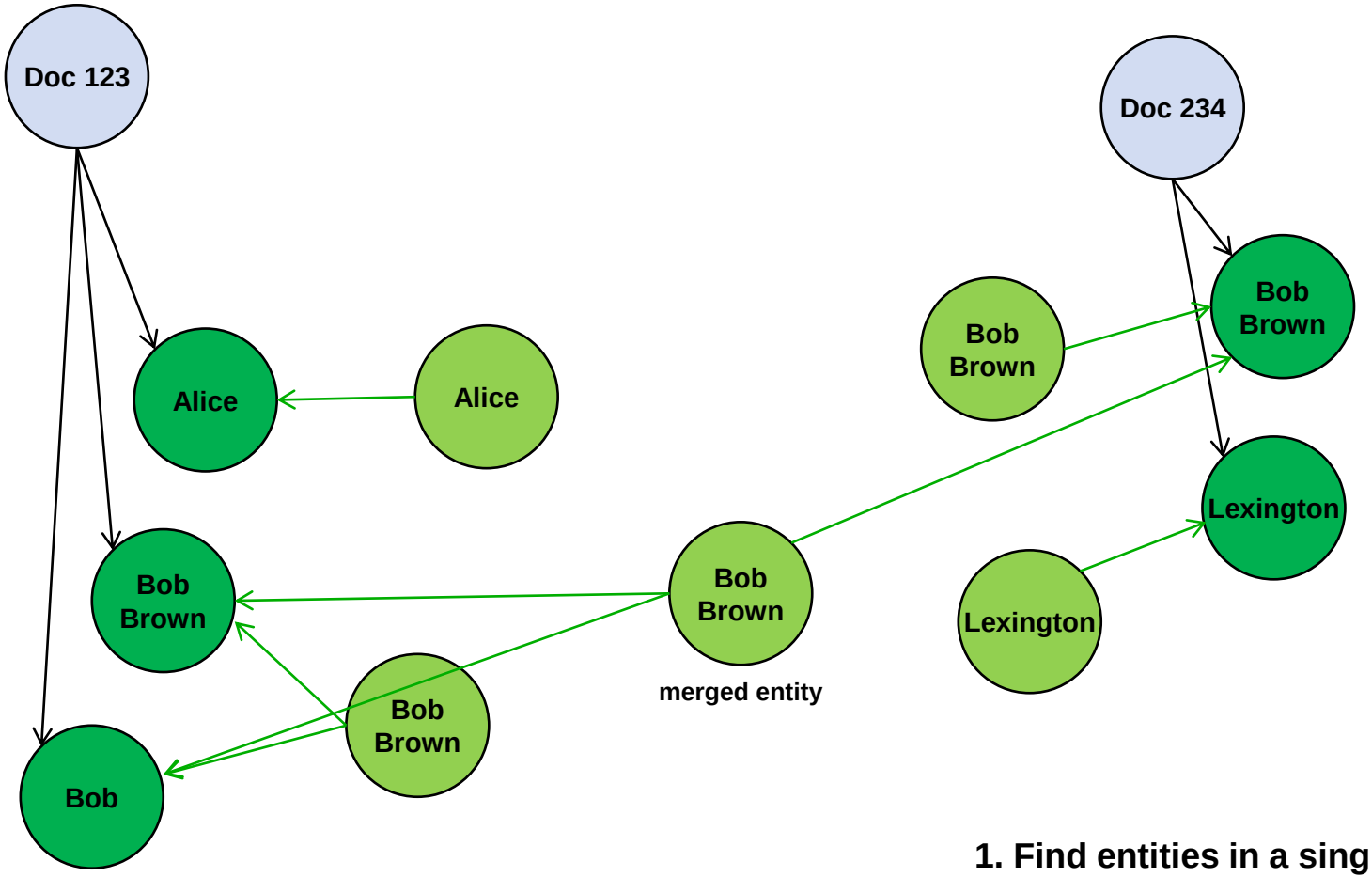
Wide variability in database performance on different tasks—no silver bullet solution

Graph Database Schema

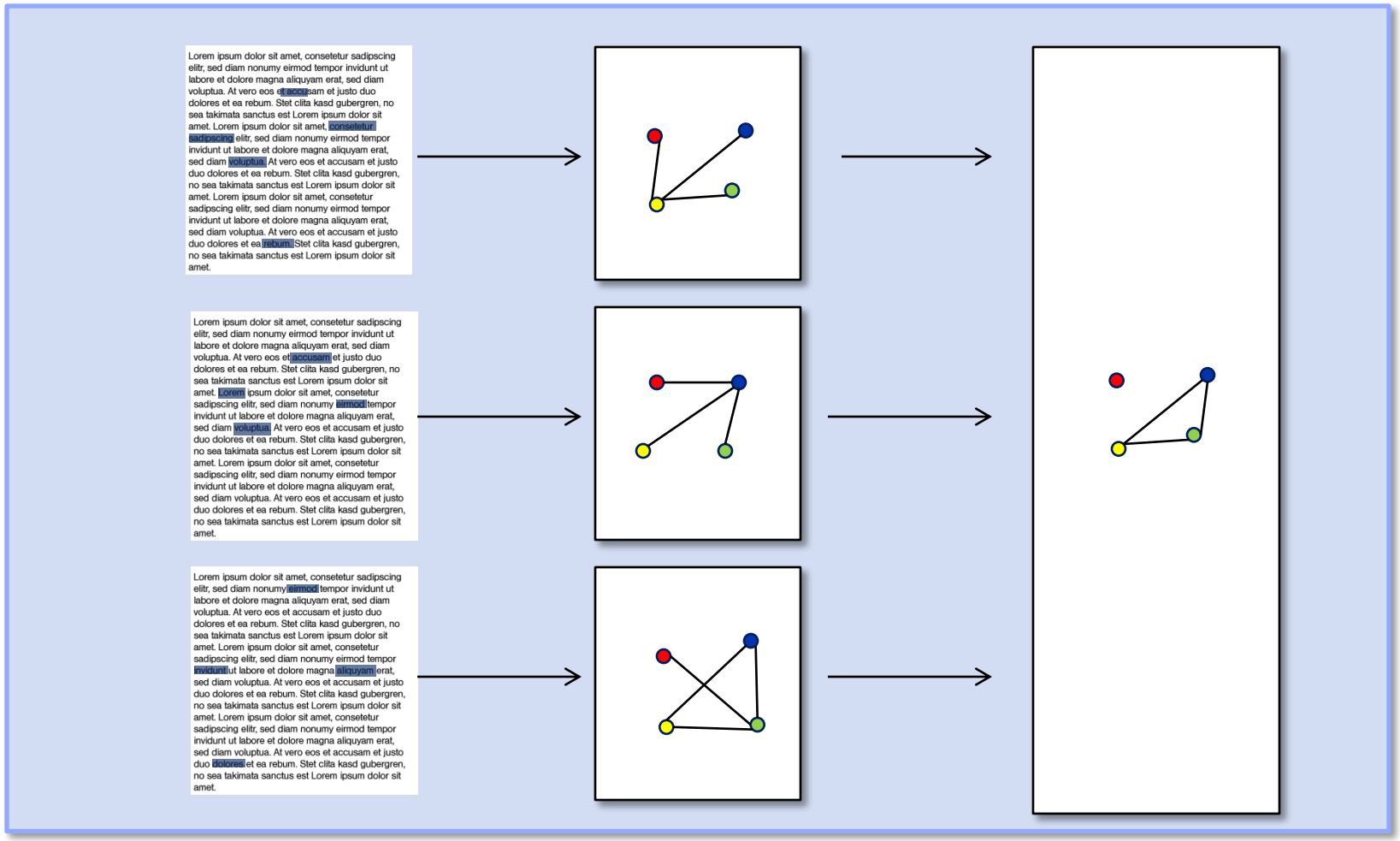




Coreference Resolution

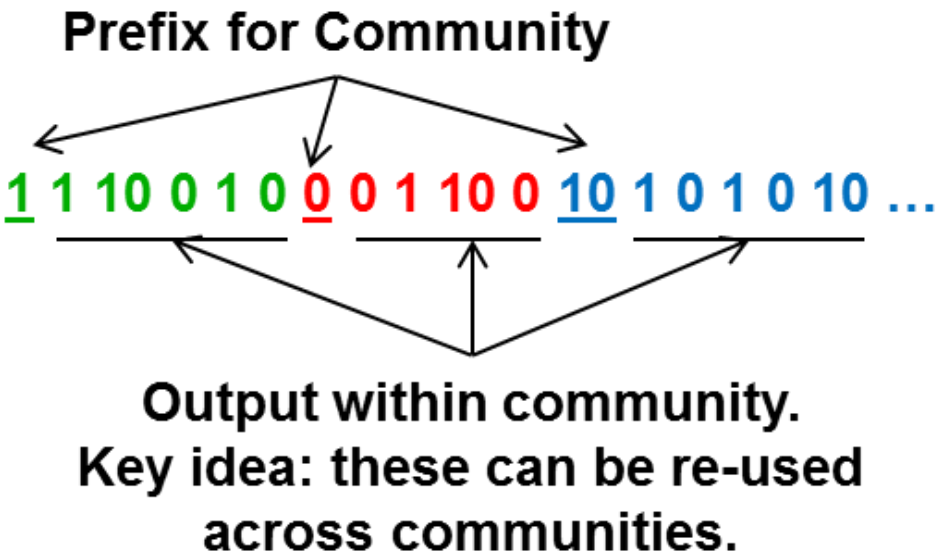


Social Network Construction

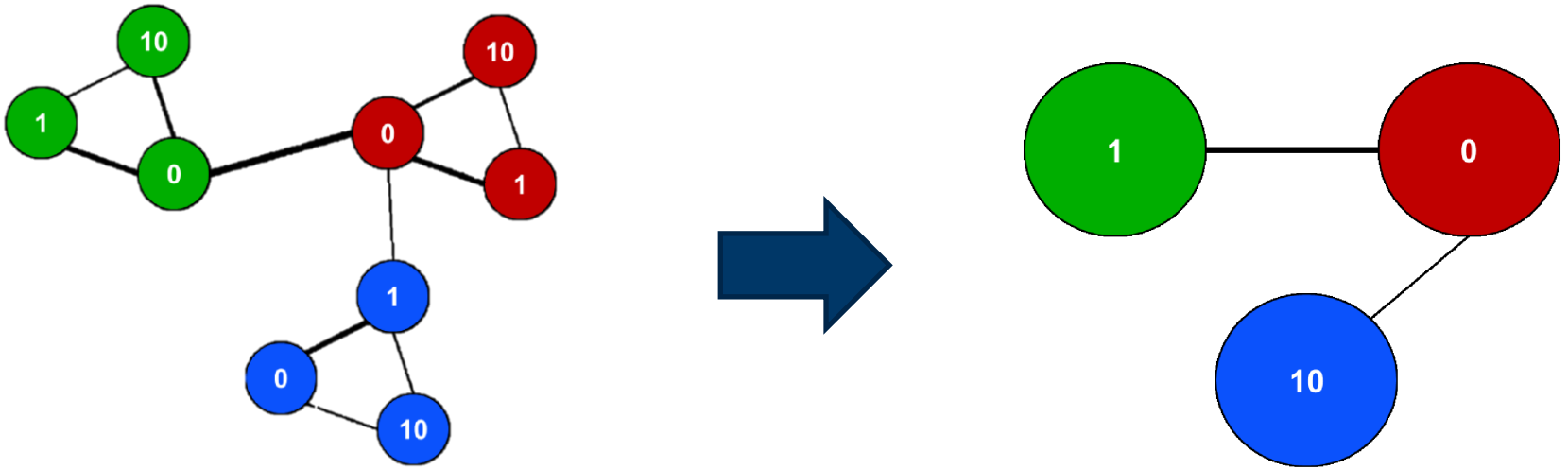


Community Detection with Infomap

- **Look at random walks on the graph—label each node**
- **Key idea: Compress random walk with a two-level code:**



Community Detection with Infomap



Leadership Prediction

- **Various centrality metrics**
  - **Page rank**
  - **Betweenness**
  - **…**
- **Typically a negligible difference in results on operational data**

VizLinc User Interface

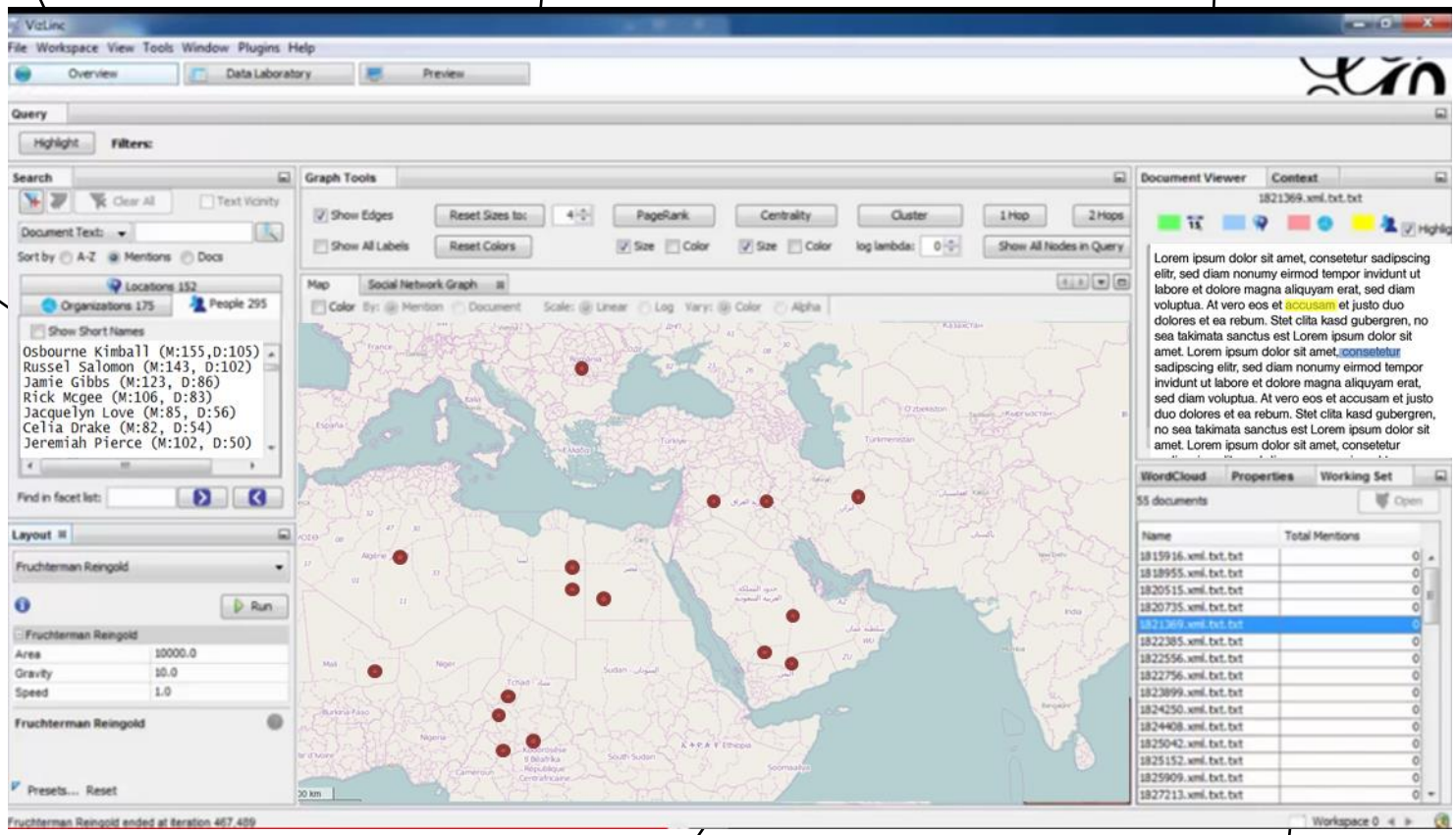
Text Search

Social Network (not displayed)
Shows people mentioned and their links

Document Content
Highlights search terms/ entities extracted

Entity Search
by person, place and/or organization

Social Network Analytics



Map
Shows all locations in working document set

Document List

VizLinc on Your Data

- **Both the VizLinc UI and the VizLinc Ingester are open source**
- https://github.com/mitll/vizlinc

Questions