

VizLinc

**Kara Greenfield, William Campbell,
Joel Acevedo-Aviles**

GraphEx 2014

8/21/2014





VizLinc

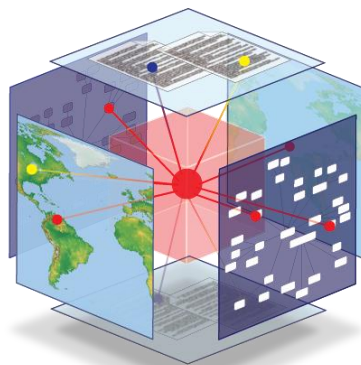
Software suite that integrates automatic information extraction, search, graph analysis and geo-location for interactive visualization and exploration of a text collection

- 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

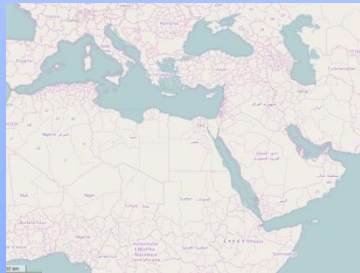
Text Entity Extraction

Pedro worked near
Rio Grande

Topics

Love
Religion
Holidays
Money

Location Extraction/ Geocoding

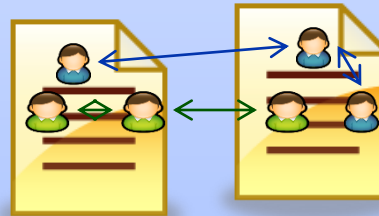


Single-Message
Content Analytics

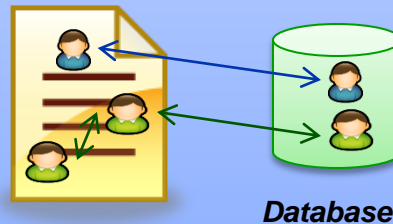
Social Networks



Entity Coref

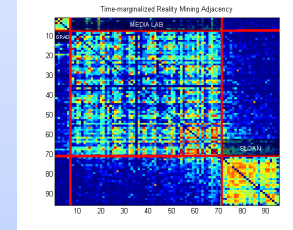


Entity Linking



Knowledge-Base
Construction

Community Detection



Leadership Prediction



Efficient Search

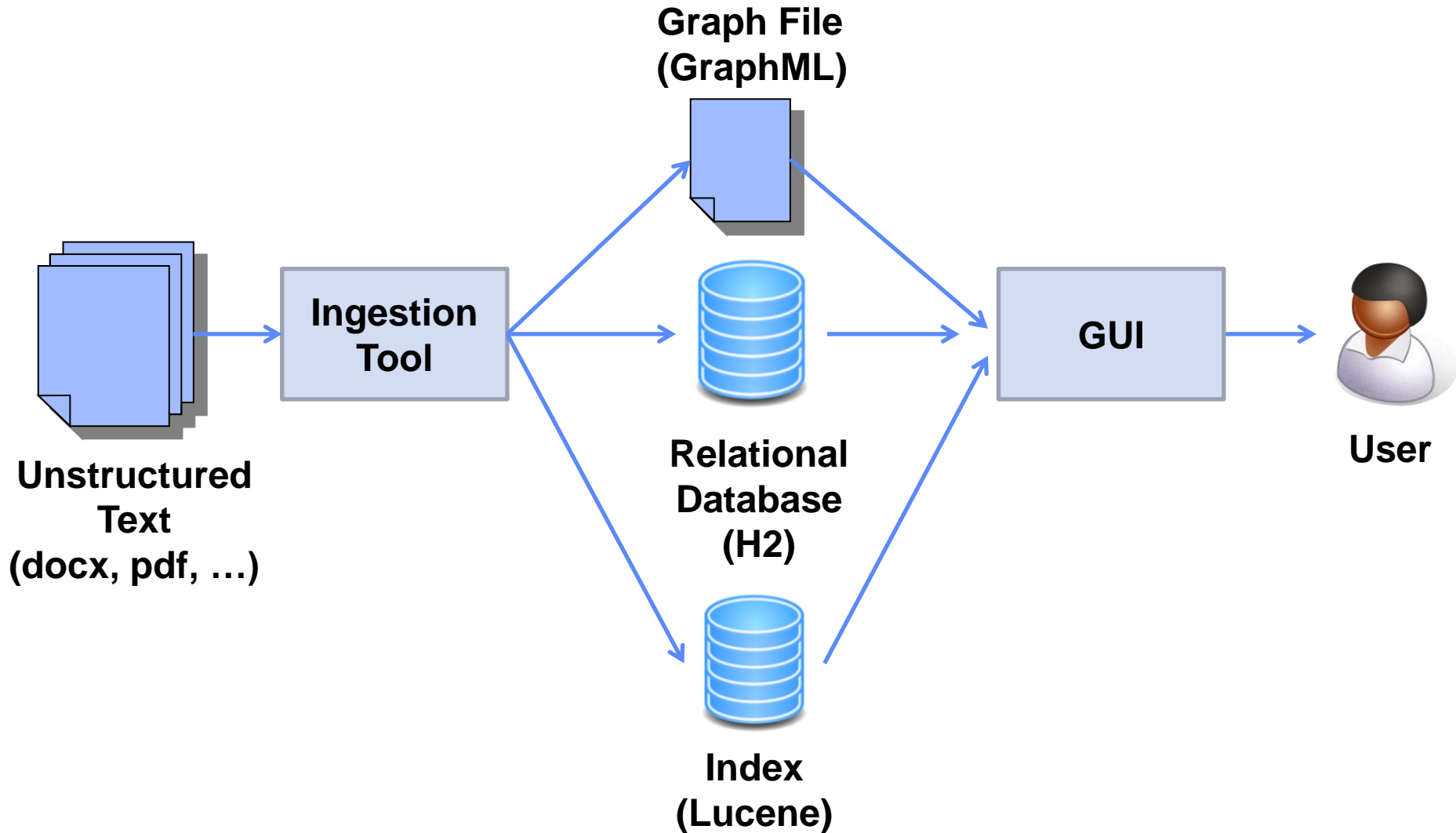


High-Level
Analytics

U
s
e
r

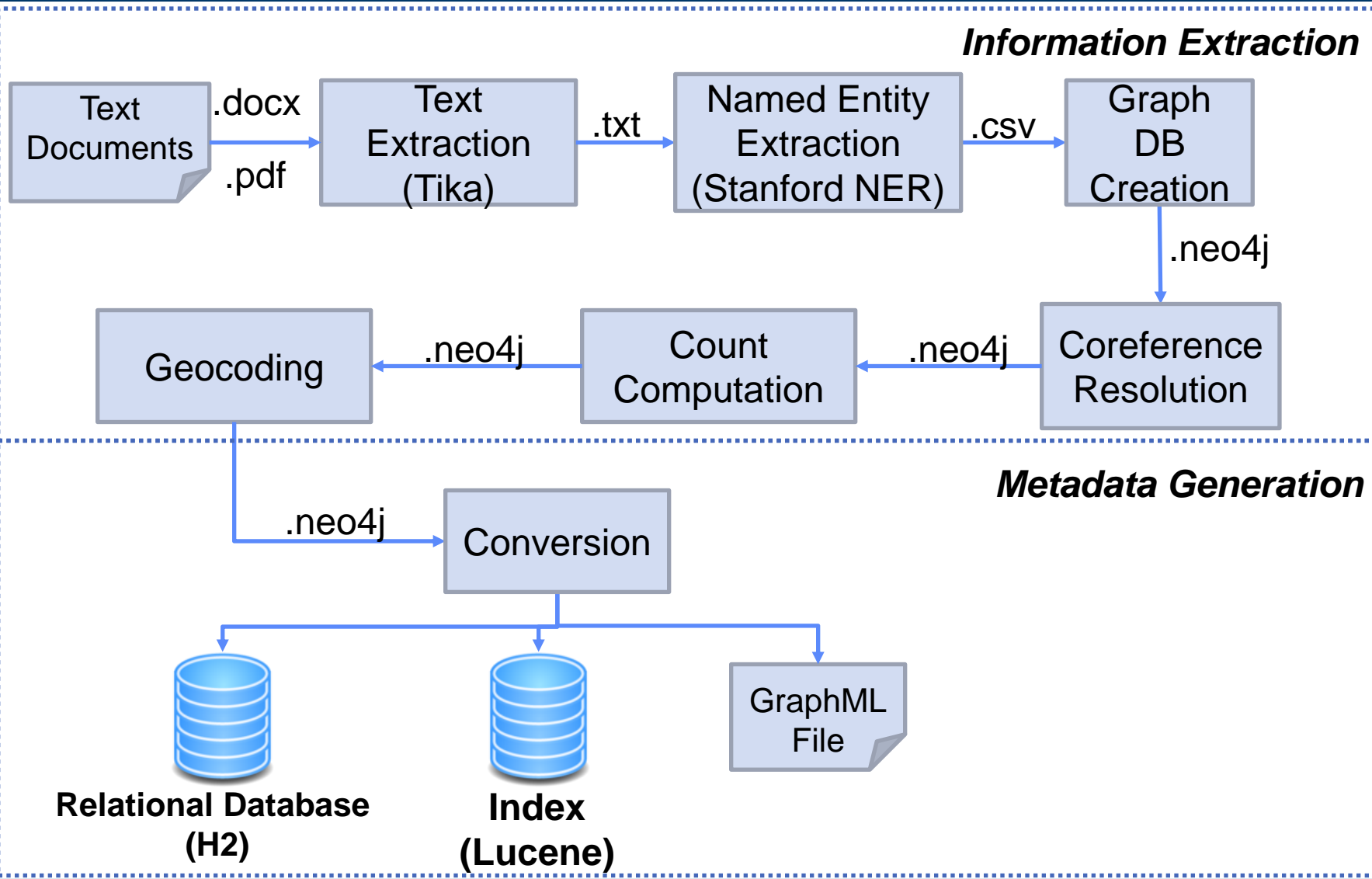


System Overview





Ingestion Tool





Database Implementation for VizLinc Tool

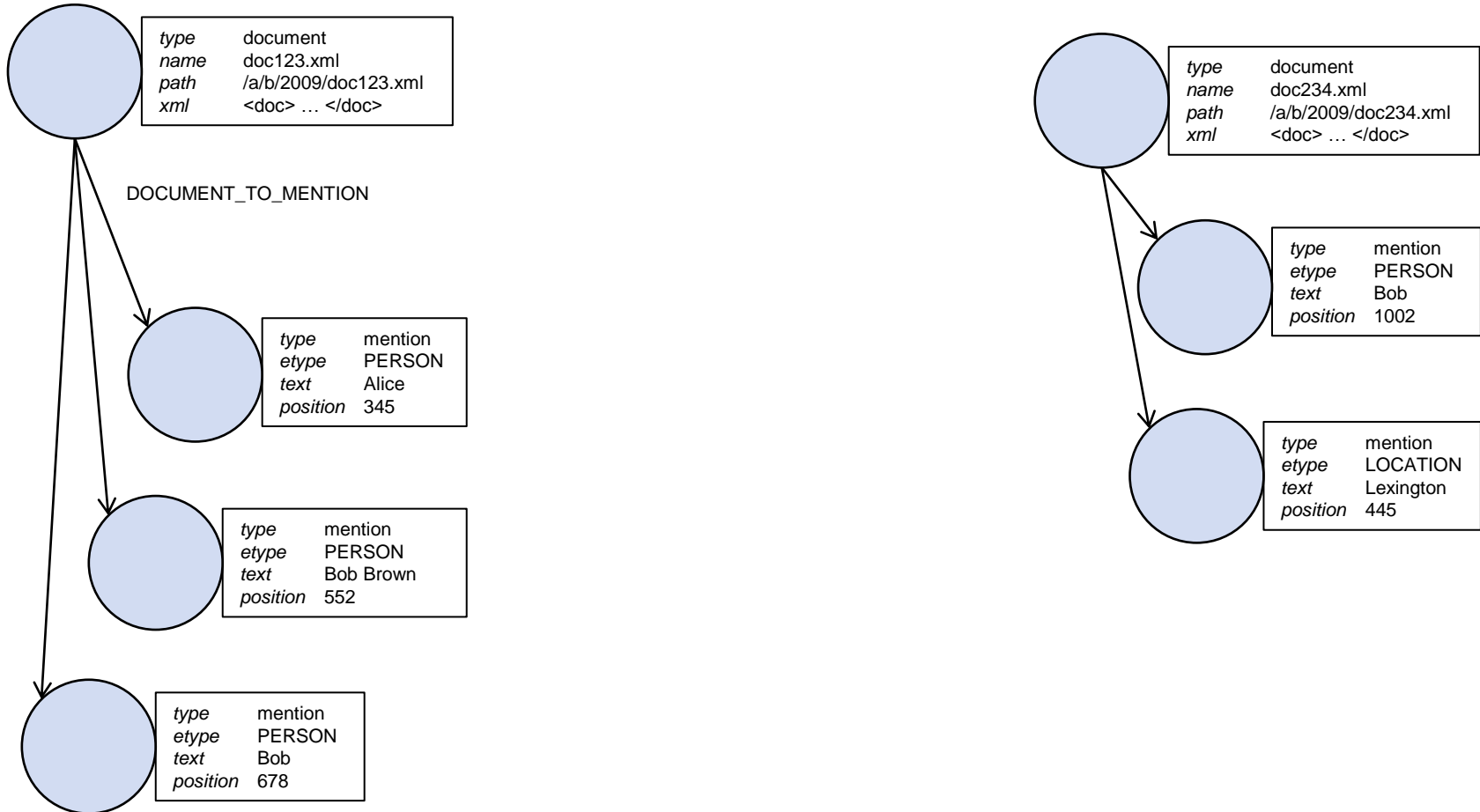
	Search Time	Flexible Schema	Storage Scalability	Graph Ops
SQL DB	Y	Y	G	Y
Graph DB	R	G	G	Y
In-Memory Data Structure	G	R	R	G

- **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

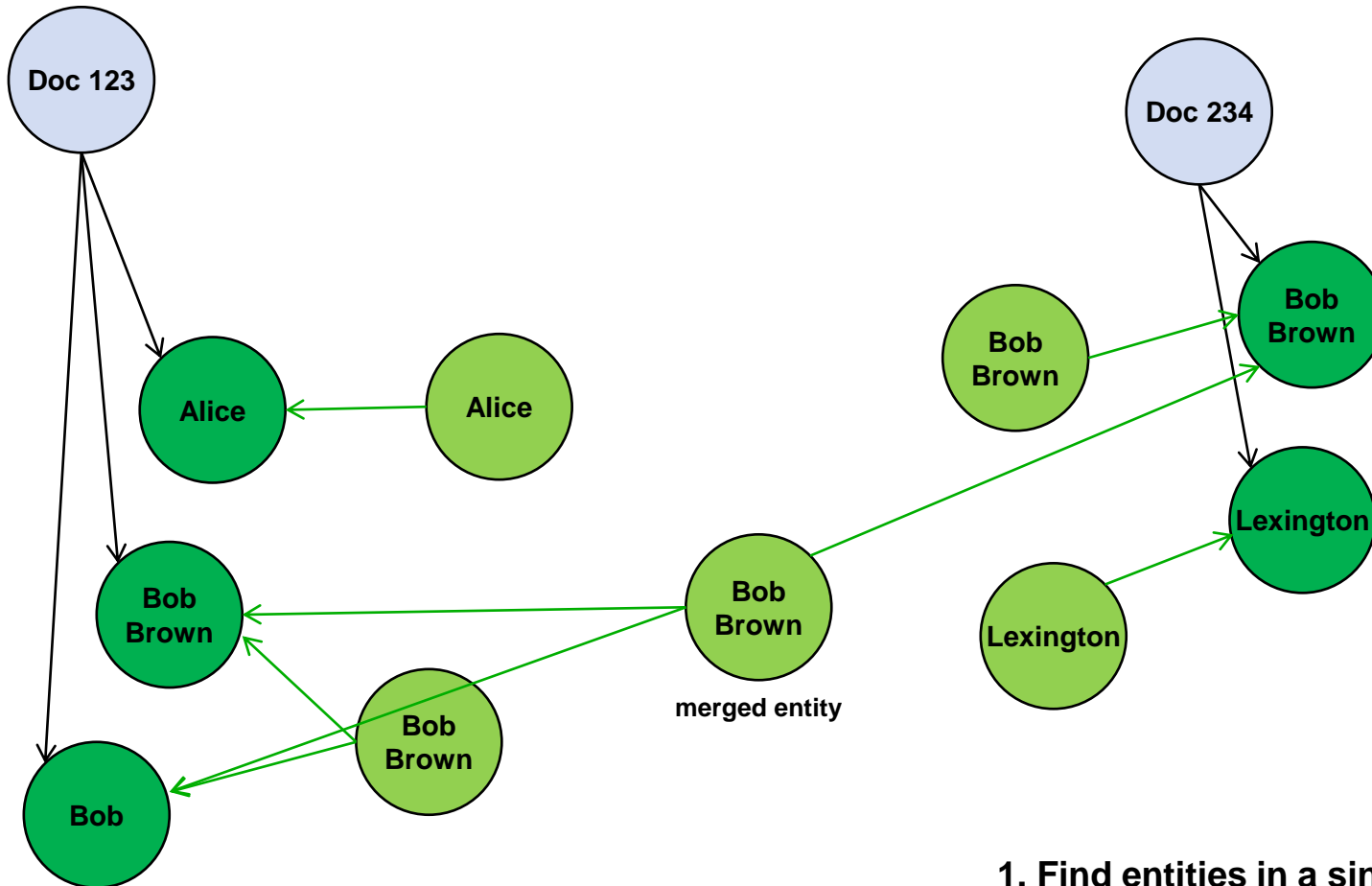


Graph Database Schema





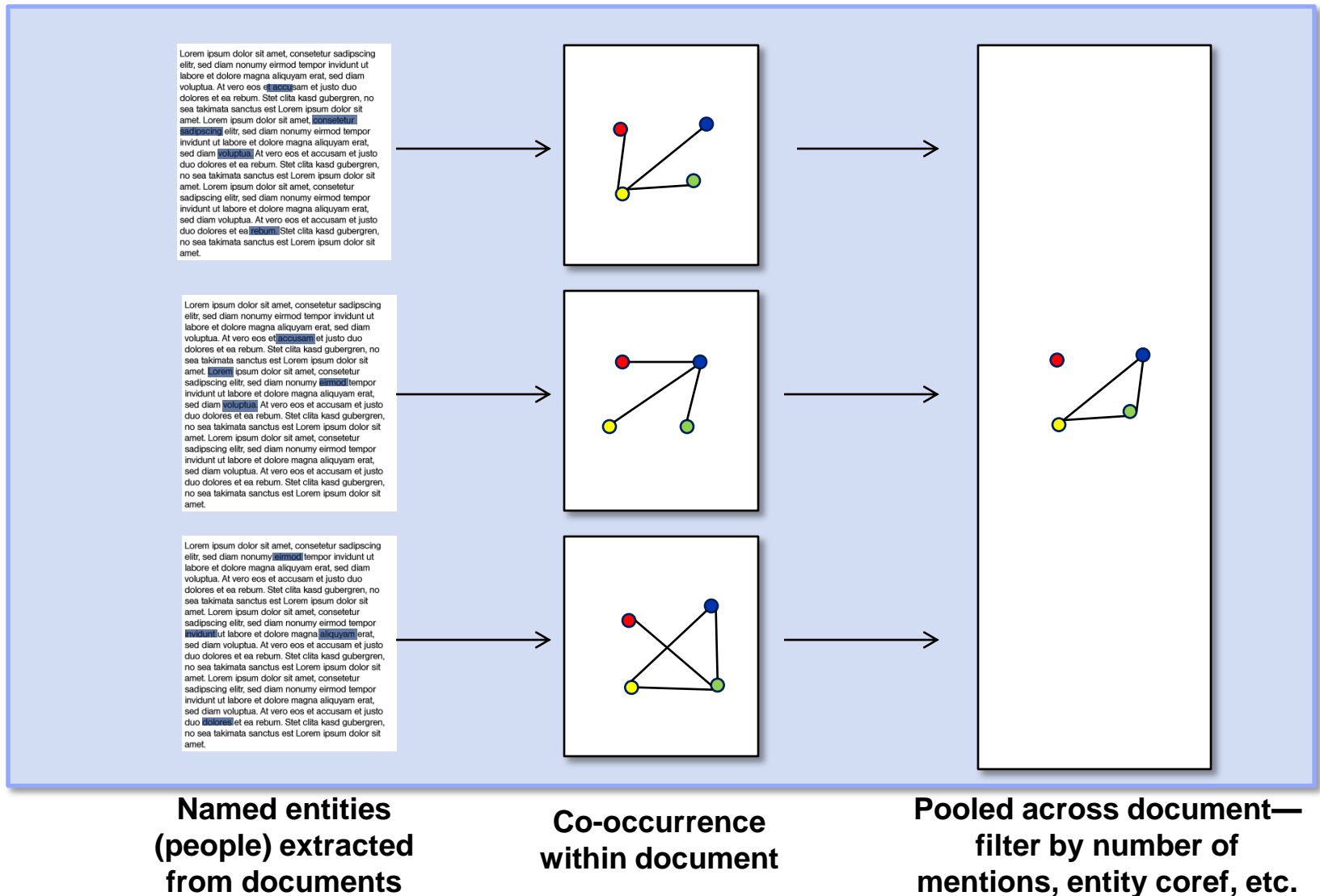
Coreference Resolution



1. Find entities in a single document
2. Merge entities across documents



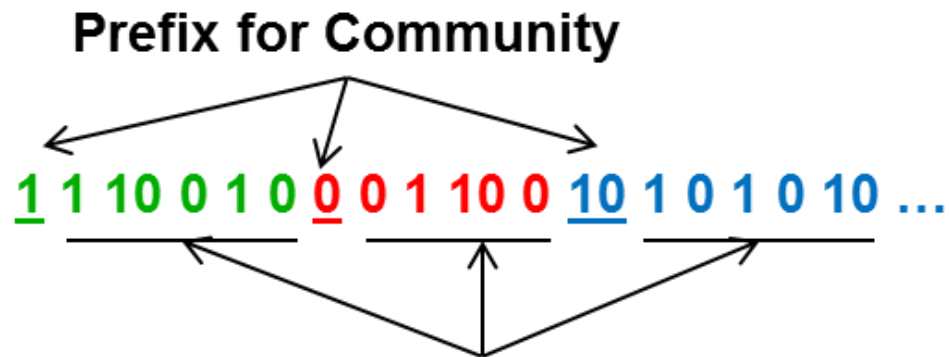
Social Network Construction





Community Detection with Infomap

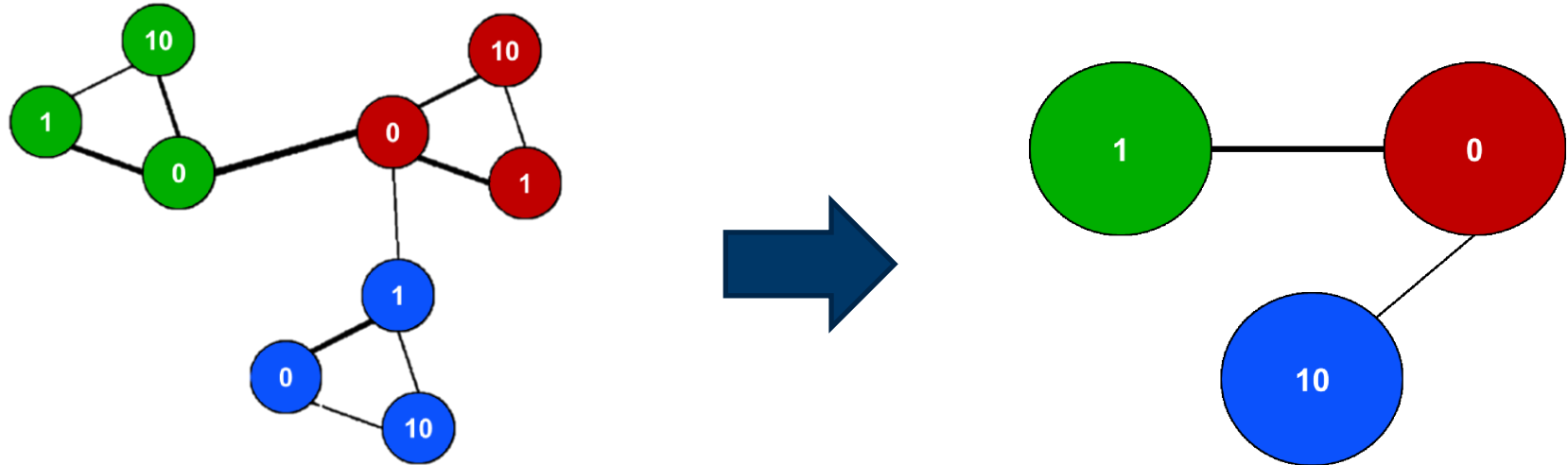
- Look at random walks on the graph—label each node
- Key idea: Compress random walk with a two-level code:



Output within community.
Key idea: these can be re-used
across communities.



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**

The screenshot shows the VizLinc application interface. At the top, there are menu options: File, Workspace, View, Tools, Window, Plugins, Help. Below the menu are tabs for Overview, Data Laboratory, and Preview. The main interface is divided into several panels:

- Query Panel:** Includes a search bar, filters, and a 'Run' button.
- Search Panel:** Shows search results for 'Locations 152', 'Organizations 175', and 'People 295'. A list of names and their mention/document counts is displayed, such as 'Osbourne Kimball (M:155, D:105)'. A 'Find in facet list' search bar is also present.
- Graph Tools Panel:** Contains options for 'Show Edges', 'Reset Sizes', 'PageRank', 'Centrality', 'Cluster', '1 Hop', '2 Hops', 'Show All Labels', 'Reset Colors', 'Size', 'Color', 'log lambda', and 'Show All Nodes in Query'.
- Map Panel:** Displays a map of the Middle East region with red dots representing locations. The map is titled 'Social Network Graph' and includes a legend for 'Color By: Mention, Document' and 'Scale: Linear, Log'.
- Document Viewer Panel:** Shows a document titled '1821369.xml.txt.txt' with highlighted text. The text includes phrases like 'Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.'.
- WordCloud Panel:** Shows a list of 55 documents with columns for 'Name' and 'Total Mentions'. The document '1821369.xml.txt.txt' is highlighted.
- Layout Panel:** Shows 'Fruchterman Reingold' settings for 'Area' (10000.0), 'Gravity' (10.0), and 'Speed' (1.0).

At the bottom of the screenshot, a status bar indicates 'Fruchterman Reingold ended at iteration 467.489'.

Map
Shows all locations in working document set

Document List



VizLinc on Your Data

- Both the VizLinc UI and the VizLinc Ingestor are open source
- <https://github.com/mitll/vizlinc>





Questions

