D2D Text Analytics Program: Overview of Tasks

Contextual Understanding

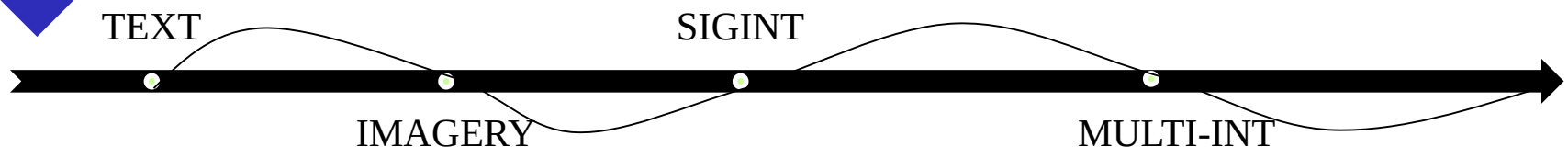
- Discovery of Events
- Motivation
- Discovery of Shared Beliefs and Sentiment Analysis
- Semantic relationships and subgroup clusters
- Social networks and relationship type/strength

Event Prediction

- Extract, characterize, monitor social networks over time
- Evolve Visual Analytics / Semantic Analysis at scale
- Extract dynamic temporal trends
- Identify key actors and supported relationships
- Detect bridging nodes to uncover hidden sub-networks & resource flow

Machine Trans-Proc

- Intelligent, adaptive and ontology-based search engines
- Improved data mining, cognitive aids and decision support tools
- Understanding of information at scale
- Improved information extraction and display
- Improved automated summarization
- Translation on non-English text for summarization, extraction, and discovery



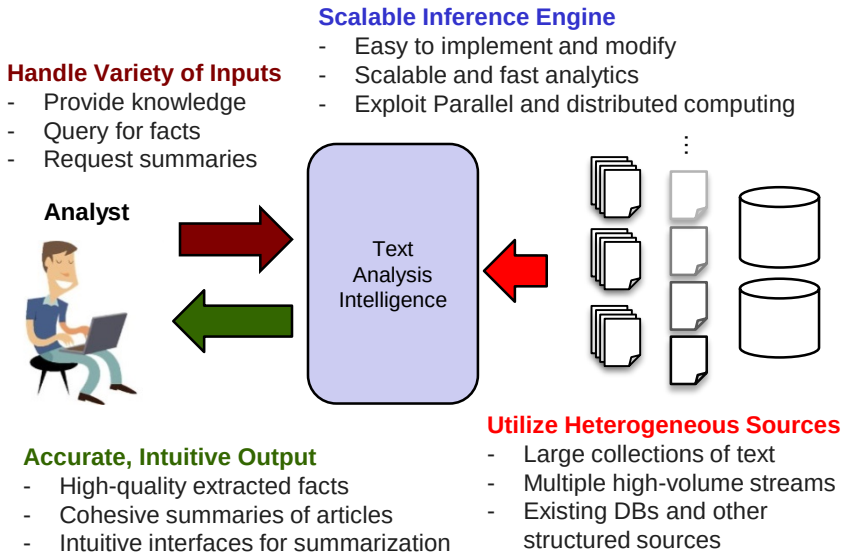
Accurate, Intuitive, Scalable Text Analytics

Vision

- Project will explore **joint models** for information extraction that allow analyst to inject alternate forms of supervision via **interactive interfaces**.
- Project will provide tools for multi-document summarization that reduce the information load on the analyst by providing **cohesive summaries** and utilizing **hierarchical presentation**.
- Project will investigate new paradigms for machine learning on **high-volume text streams**, allowing scalable analytics using accurate approaches.

Value Propositions to the Warfighter

- Reduction of error in automatically extracted knowledge from text, improving quality of decisions.
- Enable analyst to inject domain knowledge into the automation, resulting in domain-specific extractors and reduced debugging efforts
- Reduce information load when reading automatically generated multi-document summaries
- Intuitive interface for navigating text summaries
- Scalable and efficient extraction of knowledge, significantly reducing 'data to decisions' time



Outcomes

- Resulting software will provide capability to extract entity-relation facts from a text corpus and external knowledge sources.
- Components to summarize collections of text articles, and present using a hierarchical interface, will be provided.
- Will introduce paradigms for scalable analytics on multiple high-volume streams of textual data.

Statistical Challenges

- **High Dimensionality**
  - 10s of millions of features (words, phrases, relationships)
- **Sparsity**
  - Each instance of dimensionality only has a few features that fire, resulting in noise in features (e.g., useless words and patterns)
- **Lack of Labeled Data**
  - No labeled datasets exist to train the system for our problems. Those that exist (Wall St Journal) are different from African news articles (different vocabulary and style). Most existing systems work with noisy, automatically generated data—created by aligning text instances to entities in a knowledge base (like Wikipedia), and assuming all instances have the same label.
- **Structured Output**
  - The output label usually lies in the combinatorial space (predicting spans and types of multiple entities in a sentence, while predicting the relationships between them. This is a computational issue (search is enormous) and a statistical issue (amount of data needed to estimate the parameters).

Integrated Text Extraction & Analysis System (iTEAS)

Problem: Explain and forecast group-level and sub-national violence in AFRICOM AOR.

Technical Tasks: Use automated text extraction techniques to collect provincial- & group- level data :

- Data measuring patterns of interaction between groups and gov't orgs within sub-national spatial units.
- Data measuring popular sentiment toward opposition groups and gov't orgs.
- Apply several statistical modeling techniques to test and evaluate patterns that forecast violence and instability with greater than 80% accuracy.
- Develop and evaluate methods to combine and aggregate various forecasts from individual models to produce the best possible forecast.
- Develop software and visualizations to provide drill down capabilities and deeper analysis of why models are forecasting various outcomes.
- Data measuring social, political, economic, and demographic trends

Statistical Challenges

- **Source coverage/bias:** aggregating multi-source data improves modeling/forecasts compared to single source
- **Sample sizes & Representative Samples:**
  - Are reports reliable? (hard to say, assume media is reliable).
  - To answer this question→model data and forecast out of the sample, then compare to ground truth.
  - Do automated methods result in data that is representative of what is occurring?
- **Types and sources of data (First v Third person)**
  - Combine 1st & 3rd person. Factual events link A & B interactions, and take discourse directly from websites, media, etc., to get a sense of what/how people interact.
- **Quantifying qualitative data:** multi-factor constructs. Example: Repression: how to measure? Frequency of human rights abuses, # of rubber bullets shot by police, # of clashes between police and dissident groups, # of curfews implemented, # of checkpoints, etc.
- **Independence Assumptions:** highly unlikely. Forecasting and testing models out of sample (split training and test sets) guard against over-fitting.
- **Models:** Appropriate theoretical and statistically sound mathematical models to forecast social phenomena of interest, models vary parameters by unit (repression may have a different effect in Nigeria vs. Columbia) and time (repression in Nigeria varies over time).