AutoMap:
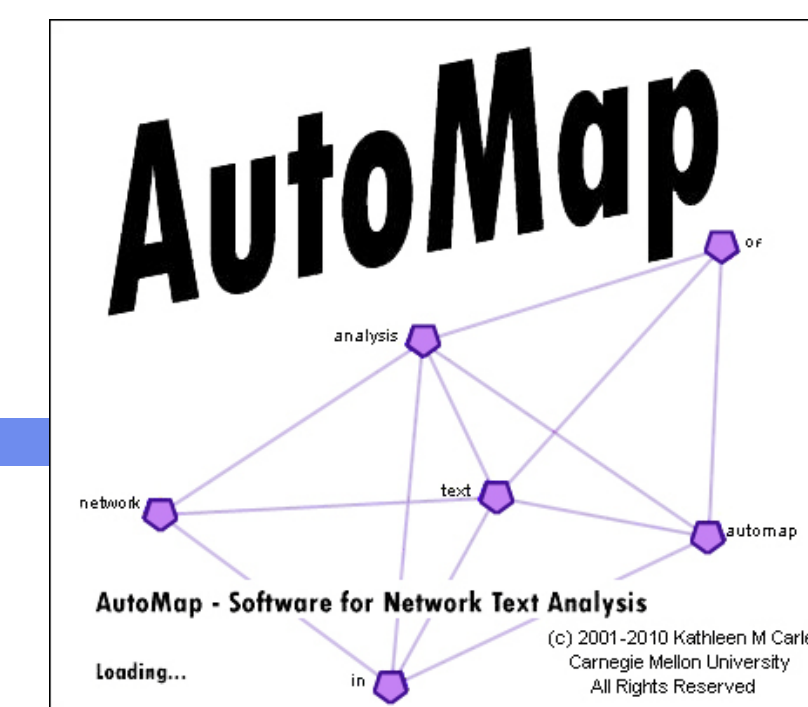## Extracting usable information from unstructured texts

Michael W. Bigrigg   **Prof. Kathleen M. Carley**
bigrigg@cs.cmu.edu   kathleen.carley@cs.cmu.edu

Frank Kunkel   **Jana Diesner**   **Todd Eisenberg**   **Daniel Chieffallo**   **Dave Columbus**
fkunkel@cs.cmu.edu   diesner@cs.cmu.edu   teisenbe@andrew.cmu.edu   dlc85@cs.cmu.edu   columbus@cs.cmu.edu
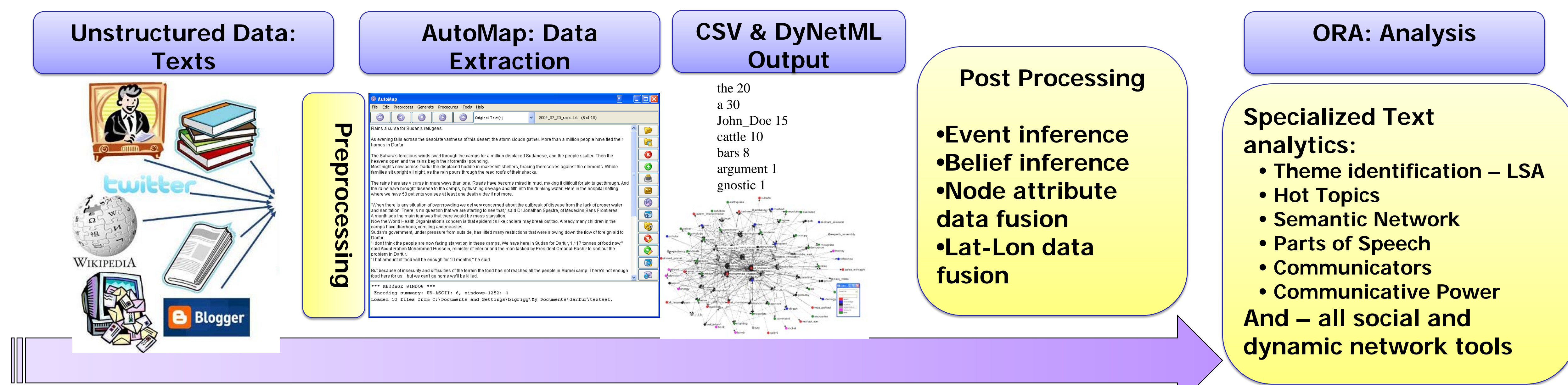
AutoMap is a text mining tool that enables the extraction of network data from texts. AutoMap can extract four types of information:
• content (concepts, frequencies and meta-data such as sentence length)
• semantic networks (concepts and relationships)
• meta-networks (ontologically coded concepts and relationships – named entities and links)
• sentiment and node attributes (attributes of named entities)
Parts of speech tagging, machine learning techniques and proximity analysis enable computer-assisted Network Text Analysis (NTA). NTA encodes the links among words in a text and constructs a network of the linked words.

*ORA

AutoMap is available as a stand-alone application as well as a module in *ORA & SORASCS.

Unstructured Data: Texts

Preprocessing

AutoMap: Data Extraction

CSV & DyNetML Output

the 20
a 30
John_Doe 15
cattle 10
bars 8
argument 1
gnostic 1

Post Processing

•Event inference
•Belief inference
•Node attribute data fusion
•Lat-Lon data fusion

ORA: Analysis

Specialized Text analytics:
• Theme identification – LSA
• Hot Topics
• Semantic Network
• Parts of Speech
• Communicators
• Communicative Power
And – all social and dynamic network tools

Extraction of the network is done using AutoMap. Additional Information added with Post processing Data fusion Routines. Analysis and visualization of networks is done using *ORA.

The GUI Interface

Quick Start Buttons
Filename Box
Preprocess Order Window
Pull-Down Menus
File Navigation Buttons
Text Display Pane
Message Pane

ScriptRunner Interface
Drag & drop editor to adjust workflow

Multiple tabs for different phases of processing

Palette of available commands.

Canvas of the current commands. Drag and drop new items here.

Preprocessing text modification prior to data extraction:
•Remove Symbols   •Bigrams
•Convert to Lowercase   •Named entity extraction
•Stemming
•Delete List   •Key word in context
•Generalization

Support tools for generation and editing of thesauri, ontologies, and delete lists exist

•Texts can be analyzed real-time
•User identifies desired manipulations in the GUI
•Manipulations can be saves as a workflow script
•This script can be run in batch mode

•The ScriptRunner can read AutoMap scripts
•Using Scriptrunner the user can:
    •Add in webscrapers
    •Add post-processing options e.g. add lat-lon from gazateers
•The SuperScript module utilizes all processors in a multi-core environment, paralellizing the text processing

www.casos.cs.cmu.edu