CASOS Summer Institute 2008

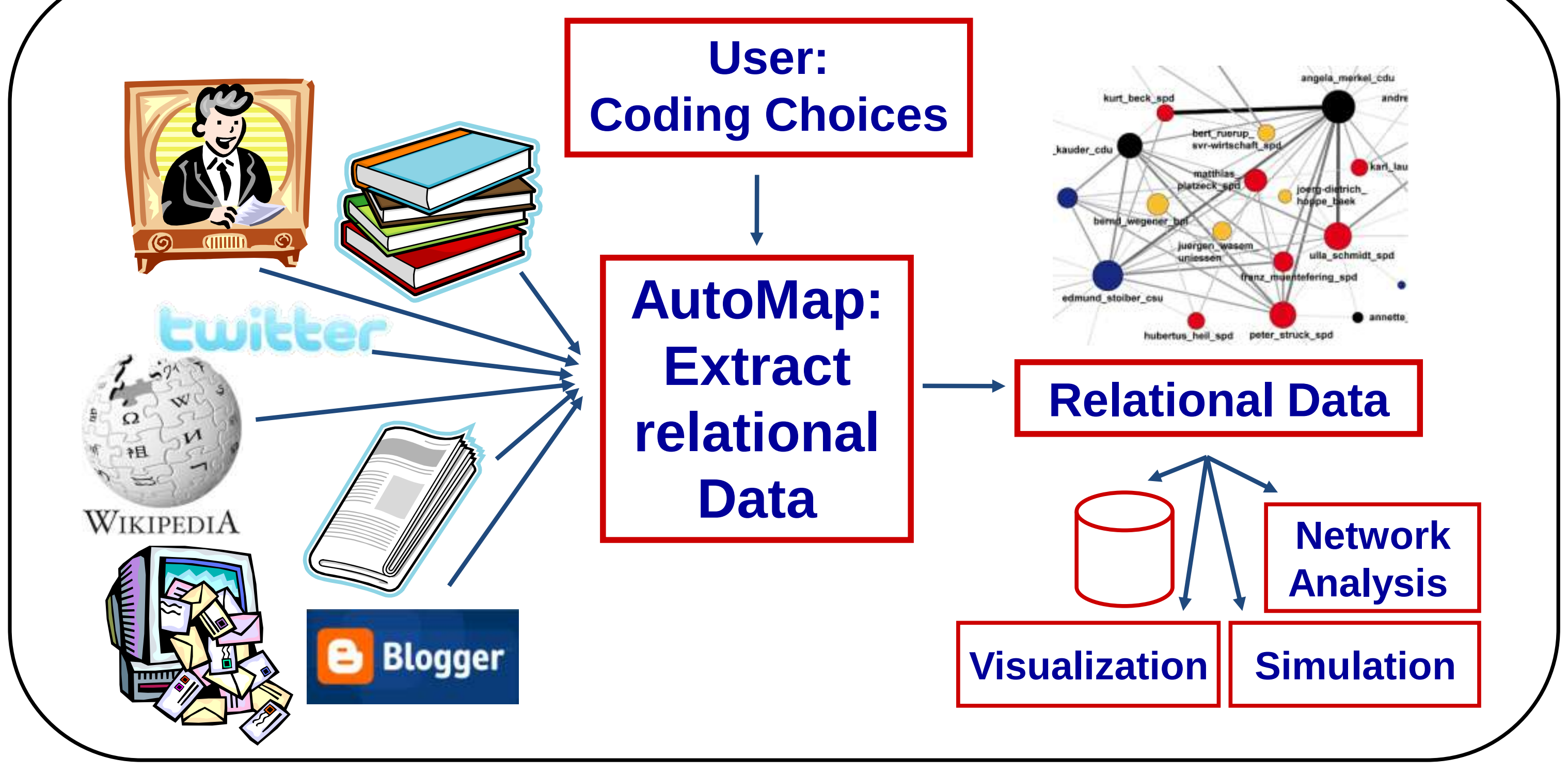
Carnegie Mellon

Ph.D. Program in Computation, Organizations & Society

From Texts to Networks

Jana Diesner
diesner@cs.cmu.edu

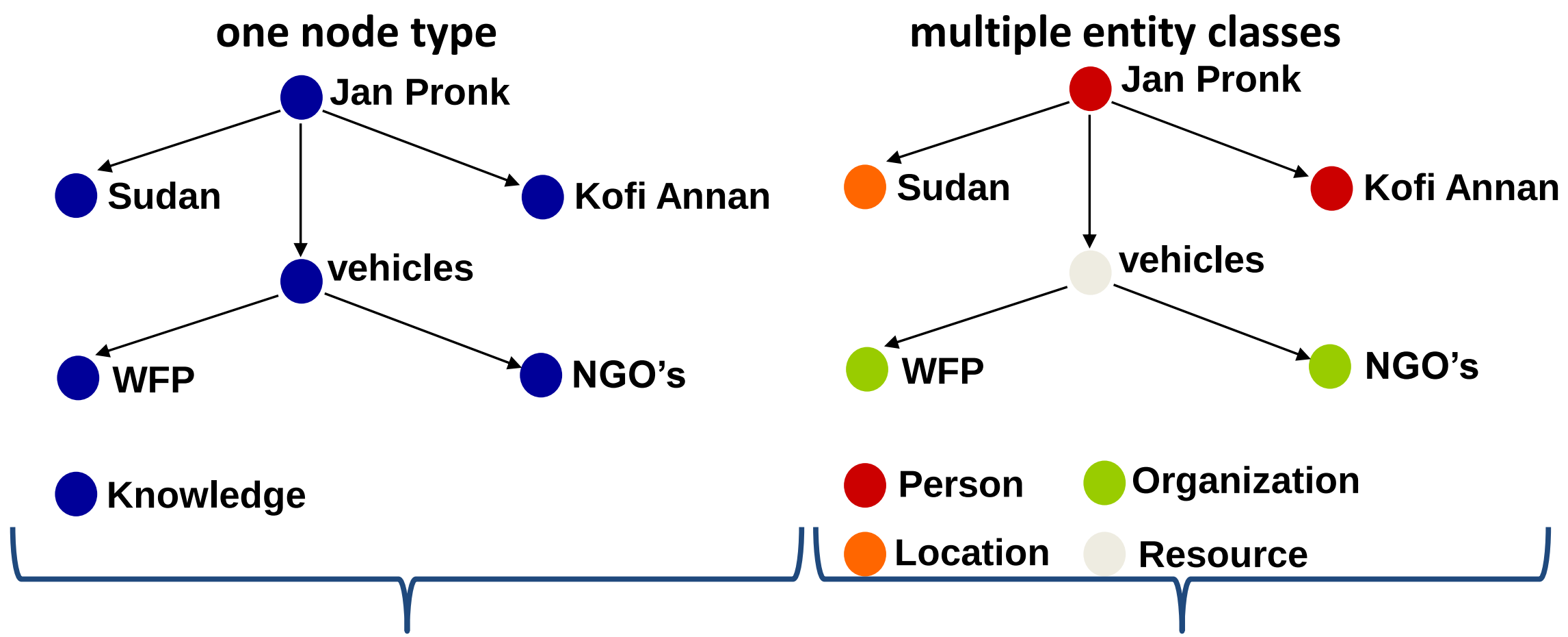
Prof. Kathleen M. Carley
kathleen.carley@cs.cmu.edu



Illustrative Toy Example:

"Jan Pronk, the Special Representative of Secretary-General Kofi Annan to Sudan, today called for the immediate return of the vehicles to World Food Programme (WFP) and NGOs." (from UN News Service, New York, 12-28-2004):

proximity-based extraction of relational data :

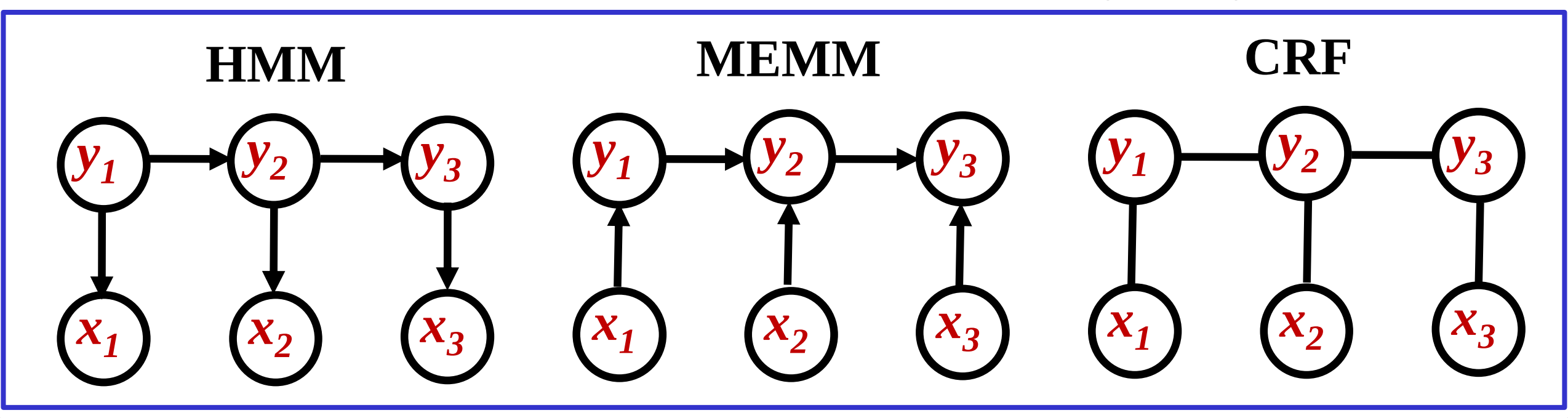


Identification:
For relational data with at least one node type: **Locate/ identify** relevant nodes (may be multi-word units)

Classification:
For ontologically coded networks: **Classify** relevant nodes according to an ontology or taxonomy

Development of Computational Solutions

- Utilize machinery from Machine Learning and Artificial Intelligence
- Deploy and develop supervised and semi-supervised **sequential stochastic learning techniques** in order to train classifiers and build models that generalize to new data
- Construct a classifier *h* that for every sequence of ***(x, y)*** (joint probability) (where *x* = words per sequence and *y* = corresponding category) or ***(x|y)*** (conditional probability) predicts a sequence ***y = h (x)*** for any sequence of *x*, incl. new and unseen data
- We work with Generative (aka discriminative) models: *P(x,y),* such as Hidden Markov Model (HMM), and Conditional models: *P(y|x)*, such as Maximum Entropy Markov Models (MEMM) and Conditional Random Fields (CRF)



Natural Language Processing and Relational Data Extraction Routines in AutoMap

- **Stemming**: Converts words into their morphemes.
- **Reduction and Normalization**:
  - Negative filters such as delete lists, removal of symbols and formatting, removal of numbers
  - Positive filters such as thesauri, spelling correction, synonym sets, antonym sets
- **Part of Speech Tagging**: Assigns a single best grammar classifier or lexical category to every word.
- **Anaphora Resolution**: Converts personal pronouns into the entity or entities that the pronouns refer to.
- **Named Entity Extraction**: Identifies relevant types of information that are referred to by a name, such as people, organizations, and locations.
- **Ontological Text Coding**: Classifies relevant types of information according to an ontology or taxonomy. User-defined categorization schemata can be applied.
- Identification of and reasoning about **node and edge attributes**, such as demographic data, beliefs, and types of relationships.
- **Email Data Analysis**: Extracts and combines different types of networks, such as social networks and knowledge networks, from emails.
- **Feature Identification**: e.g. term weights, TF*IDF
- **Entropy Assessment**: Determines the variability or heterogeneity of a text document or corpus with respect to its vocabulary.
- Classical **Content Analysis**.
- Read and write data and processing material from and to a default or user-specified **database**.

Example: Conditional Random Fields for Entity Extraction and Ontological Text Coding

- Identify and classify words that represent instances of entity classes of models or ontologies that **deviate** from classical set of Named Entities.
- Crucial step for coding texts as social-technical networks according to domain-specific ontologies and for advanced modeling of complex and dynamic real-world organizations or networks.
- Model relationship among y_i and y_{i-1} as Markov Random Field conditioned on *x*
- Conditional distribution of entity sequence *y* given observation sequence *x* computed as normalized product of potential functions $M_{i:}$

$$M_i(y_{i-1}, y \mid x) = (\exp\left(\sum_{\alpha} \lambda_\alpha f_\alpha(y_{i-1}, y_i, x) + \sum_{\beta} \mu_\beta g_\beta(y_i, x)\right)$$

$$p_\theta(y \mid x) = \frac{\prod_{t=1}^{n+1} M_i(y_{i-1},_{yi} \mid x)}{\prod_{i=1}^{n+1} M_i(x)_{start\ ,stop}}$$

- Conditional probability of label sequence *P(y|x)*, where both *x* and *y* are arbitrarily long vectors (consider arbitrarily large bag of features (> 10,000) and any property of *x*, such as long-distance information)

References:

- Carley, K.M. 2002. Smart Agents and Organizations of the Future. Ch. 4 In: L. Lievrouw and S. Livingstone (Eds.), The Handbook of New Media Sage, Thousand Oaks, CA, pp. 206-220.
- Diesner, J., & Carley, K.M. (2007). Conditional Random Fields for Entity Extraction and Ontological Text Coding. Proc of North American Association for Computational Social and Organizational Science (NAACSOS) 2007 Conference, Atlanta, GA. Best Student Paper Award.
- Diesner, J., & Carley, K.M. (2006). Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis. Chapter 4 in: V.K. Narayanan, & D.J. Armstrong (Eds.), Causal Mapping for Information Systems and Technology Research. (pp.81-108). Harrisburg, PA: Idea Group Publishing.
- Diesner, J., & Carley, K.M. (2004). AutoMap1.2 - Extract, analyze, represent, and compare mental models from texts. Carnegie Mellon University, School of Computer Science, Institute for Software Research International, Technical Report CMU-ISRI-04-100. URL: http://reports-archive.adm.cs.cmu.edu/isri2004.html
- Dietterich, T. G. (2002). Machine Learning for Sequential Data: A Review Paper Joint IAPR International Workshops SSPR 2002 and SPR 2002, August 2002, Windsor, Ontario, Canada.
- Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. 18th International Conf. on Machine Learning.
- McCallum, A. (2005). Information extraction: distilling structured data from unstructured text. ACM Queue, 3(9), 48-57.

Acknowledgements:

- This work was supported in part by the National Science Foundation under grants: No. ITR/IM IIS-0081219, NSF 0201706 doctoral dissertation award, and NSF IGERT 9972762 in CASOS. Additional support was provided by CASOS and ISRI at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, or the U.S. government.

This work is part of the Dynamic Networks project in the center for Computational Analysis of Social and Organizational Systems (CASOS) of the School of Computer Science (SCS) at Carnegie Mellon University (CMU). Support was provided, in part, by the National Science Foundation (NSF) Integrative Graduate Education and Research Traineeship (IGERT) program, 9972762, the Army Research Lab, and the Army Research Institute .The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Researc Lab, the Army Research Institute, the National Science Foundation, or the U.S. government.

www.casos.cs.cmu.edu

institute for SOFTWARE RESEARCH