
Synthetic Data for Data Mining to Support Epidemiological Modeling

Madhav V. Marathe

Network Dynamics and Simulation Science Laboratory
Virginia Bio-Informatics Institute & Dept. of Computer Science
Virginia Tech
marathe@vt.edu

NDSSL-TR-06-20
Web Site: <http://ndssl.vbi.vt.edu>

These slides are a version of the short talk describing the challenge data sets given on April 22nd as a part of Workshop on Spatial Data Mining held as a part of the Annual SIAM Data Mining Conference, April 2006.

Organizers: Professors Naren Ramakrishnan, Virginia Tech & Chris Bailey-Kellogg, Dartmouth University

Venue & Date: Washington DC April 22nd 2006.

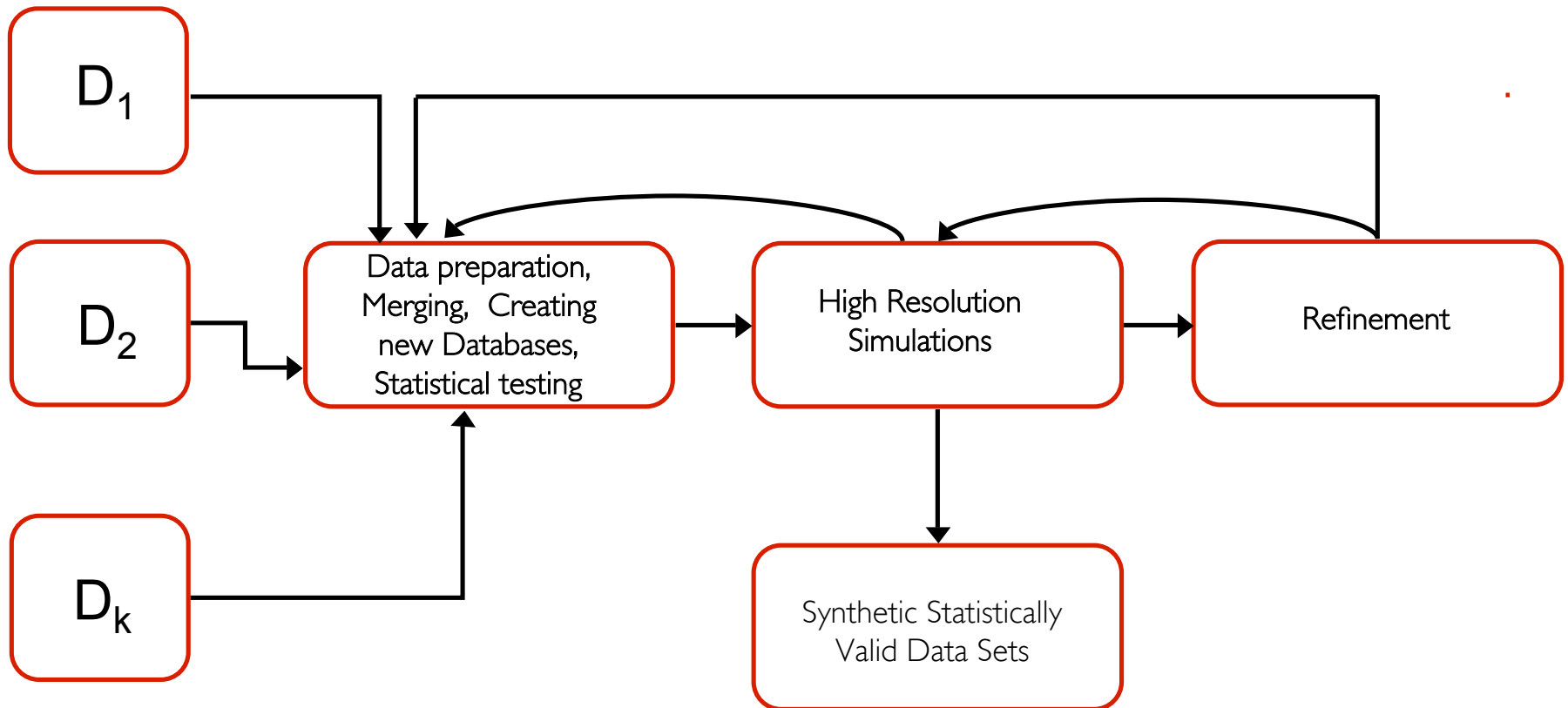
Acknowledgements

Virginia Tech: Members, Network Dynamics & Simulation Science Laboratory, VBI
Karla Atkins, Christopher L. Barrett, Richard Beckman, Keith Bisset, JiangZhou Chen,
Stephen Eubank, V.S. Anil Kumar, Achla Marathe, Henning Mortveit, Paula Stretz,

Overview

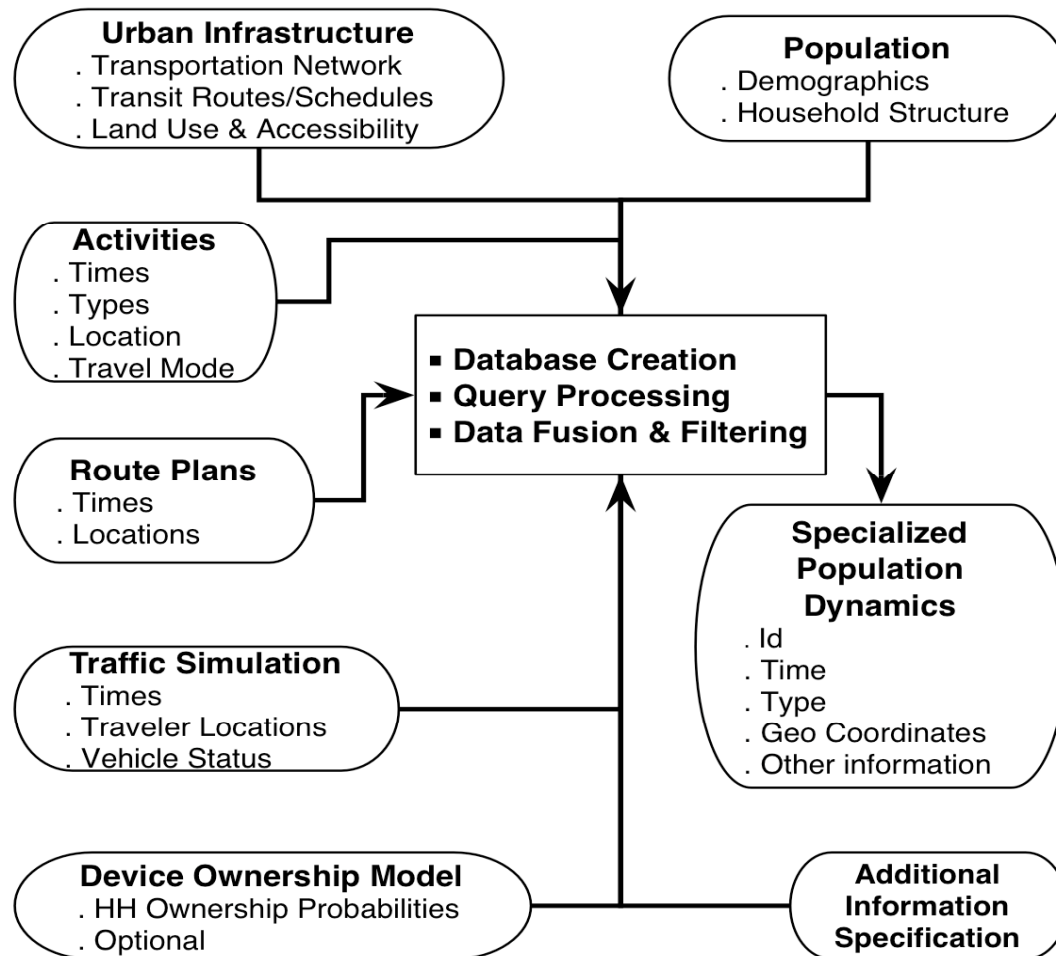
- Goal: Create high quality synthetic data sets
 - Available freely for academic purposes
 - Created using Simfrastructure: a modeling tool to be described later today
- Expect additional data sets (higher resolution, different and larger areas and new attributes) to be released on a regular basis
 - An individual calling schedule (caller, callee, their locations, the duration)
 - Mobility information for use in ad hoc network research
 - a demand pattern for electricity during a normative day set of
 - Digital device ownership at an individual level
- Enhanced versions of these data sets used in studies to support disaster planning and recovery

Overall Scheme for Producing Synthetic Data



Data sets from surveys,
and other non-
traditional sources

Data Fusion from Multiple Sources



Data Sets Currently on the Web

- **Synthetic individuals** in the city of Portland, USA,
 - small number of demographic attributes for each synthetic individual
- **Aggregated activity locations**, two per roadway link in the city of Portland
 - Each has (x, y) coordinate and a tag telling the activity that is performed
- **Daily activities for each synthetic individual** representing a normative day.
- **Time varying social contact network**, based on the daily activities.
- a description of **disease evolution** over the social network when a specific set of individuals are infected.
 - It informs when, where, and from whom a person became infected and the disease state of other individuals at each location.

Why Synthetic Data Sets

- Protects privacy of individuals, proprietary & Sensitive information
- Integration of heterogeneous data sets
 - No single data base has this information
 - Simple merging or joins will not yield statistically valid data set - Not aligned in time
- Impossible to collect all the data generated by the simulations,
 - E.g. time space plot for each vehicle in the city.
- Sampled measurements are used to ascertain their statistical validity
- Generate data for places where no data is available based
 - This data set can then be refined as more information is available
 - E.g. The Similarity based methods for data transportability