Visualization through Textualization

Ye Zhao

Kent State University

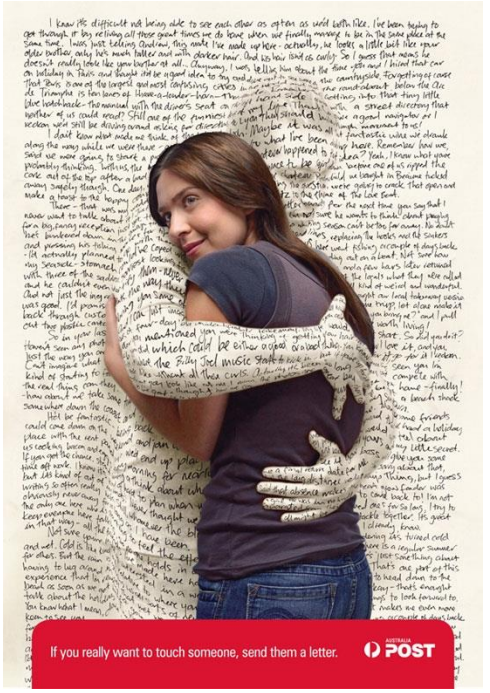
KENT STATE UNIVERSITY

Motivation

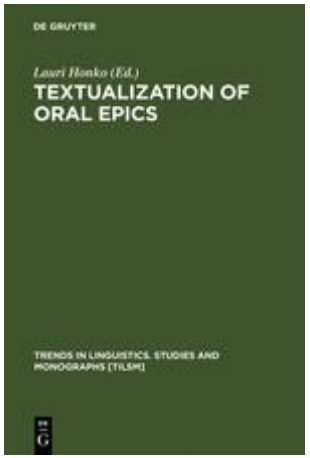
- Gap exist between what users want and what information visualization can provide
  - Visual metaphors unfamiliar to novice users
  - Cluttered views
  - Complicated interactions
- Interactive visual exploration of large, complex datasets can only be performed by a small number of visualization experts and domain experts collaborating with them

Power of Text

- Text is an intelligent tool
  - Text is used to refer to something that carries its interpretation within itself
  - it can render abundant information with its placement, font, color, feeling, and style
- Text helps people convey information and form opinions promptly



Textualization



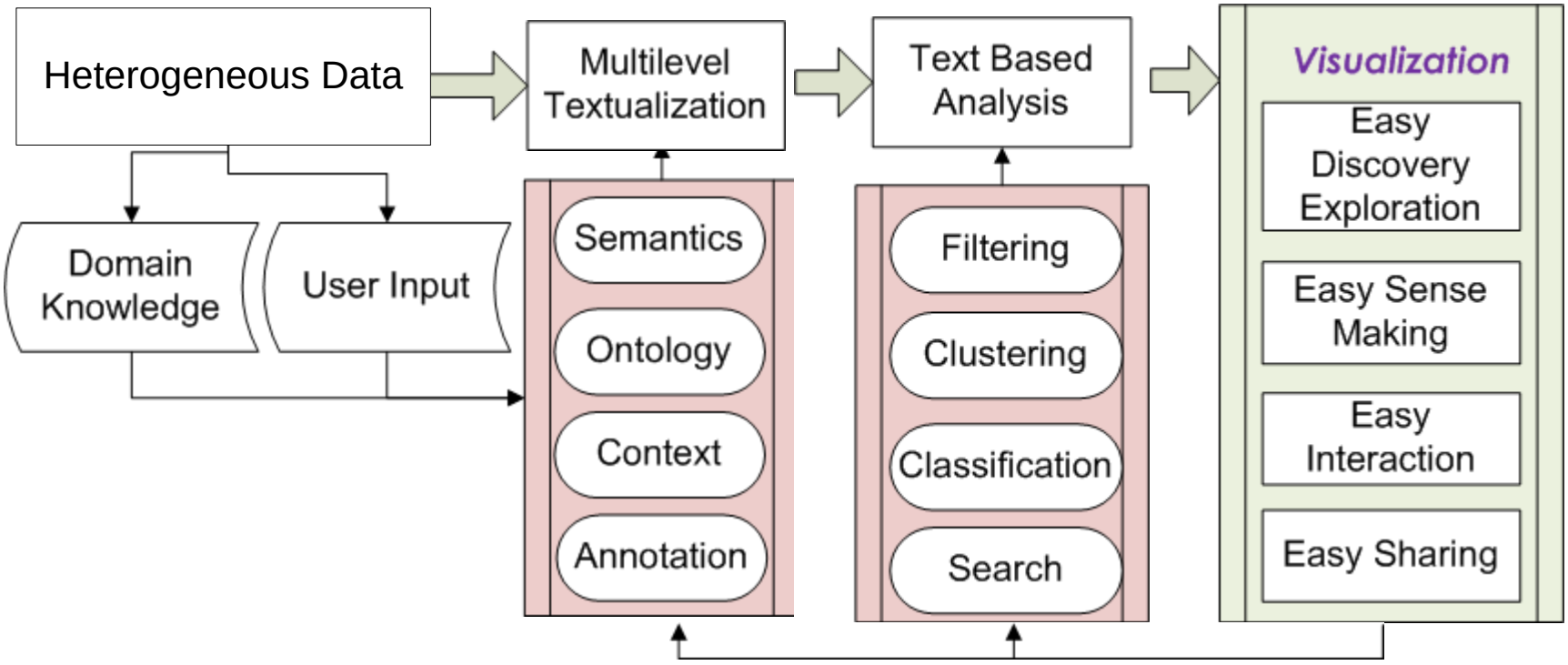
- “to put into text ” of oral epics, blogs, emotions …
- Textualization: externalizing and refining associated contextual information of the original data
- A variety of abstract data can be processed and analyzed in semantic-rich forms

KENT STATE UNIVERSITY

Visualization through Textualization

- Develop visualization techniques to fully utilize the power of text
  - Expose semantics and contextual information associated with the original data
  - Discover patterns and trends of the data with text mining techniques
  - Explore discoveries with interactive visualization

Our Scheme



KENT STATE UNIVERSITY

Stage 1: Data processing

- A multilevel textualization scheme to manifest the contextual and semantic information of the original data
  - Semantic transformation: convert data values into textual descriptions from domain knowledge
  - User Input: incorporate users' input of annotation, tagging, comments
  - Domain ontology: use the vocabulary to denote given data with its types, implicit properties, and interrelationships

Multilevel Textualization: An example

- Taxi trajectory data: two GPS samples

T1

	T1	T2
DATE/TIME	06/17/12 08:19:11	06/17/12 20:59:33
STATUS	1	0
SPEED	50KM/H	2KM/H
LATITUDE	22.533	22.649
LONGITUDE	114.044	113.814

T2

	Level 1 Street Address	Level 2 Point of Interest	Level 3 Semantics	Level 4 Annotation
T1	Xinsha Road, Futian District, Shenzhen	Shenzhen Golf Club, Sports/ Entertainment	Morning, Slow Motion, Local Road, Passenger Occupied	*Safe Place, Residential, Food/Shopping*, 9 Holes, New Driver
T2	Konggang 7 Road, Bao'an District, Shenzhen	Shenzhen Airport, Transportation	Evening, Creeping Forward, Airport Waiting, Empty	*Long Wait, Metro/Shuttle/ Ferry*, Extra Charge, Experienced Driver

KENT STATE UNIVERSITY

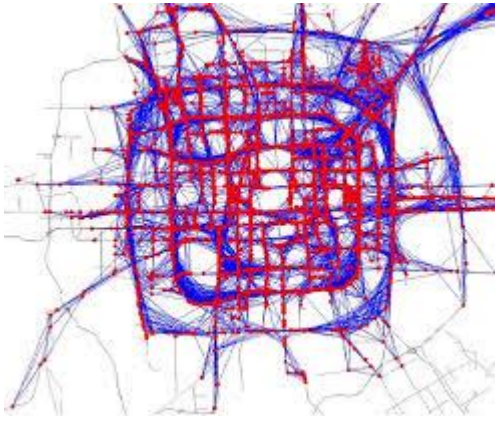
Challenge: Handling Different Data Types

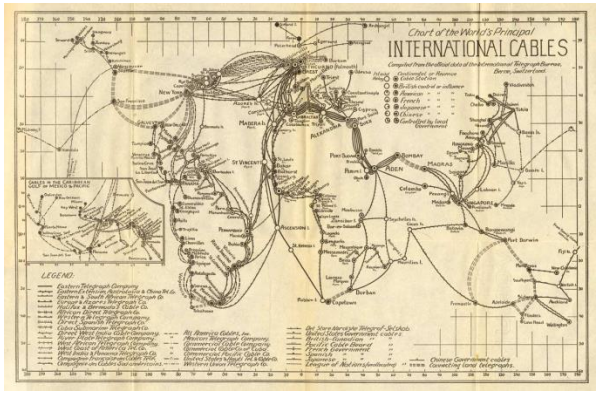
- *How can we handle different types of data with the textualization scheme?*
- *How can we preserve the data features and relationships?*

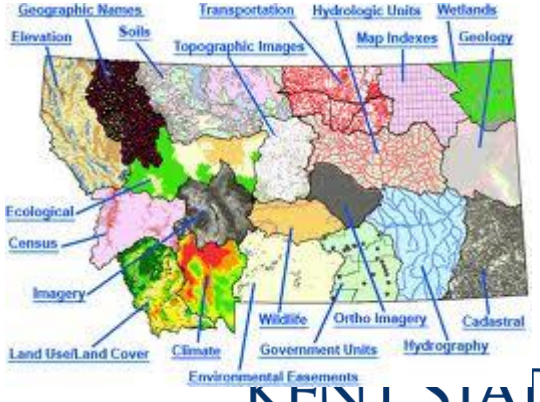
Geospatial data

- A variety of “georeferenced” information
  - demographic (e.g. census and real estate)
  - environmental (e.g. weather and climatological records)
  - geology (e.g. land features)
- Geospatial-temporal data
  - severe weather systems
  - population movement of plant and animal species
  - epidemics of flu
  - human and vehicle mobility trajectories
- Map geospatial positions to a meaningful text representation

Images from the Web
Copyright belongs to original owners

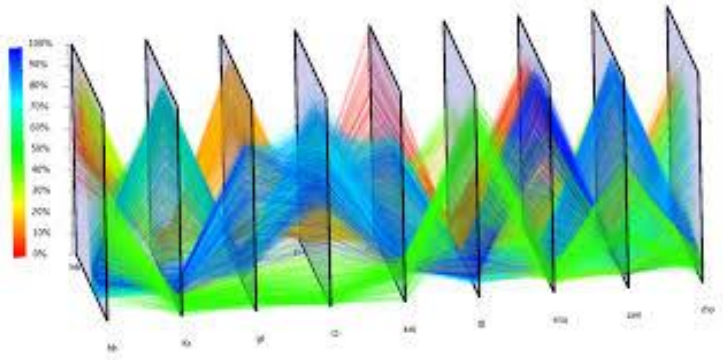


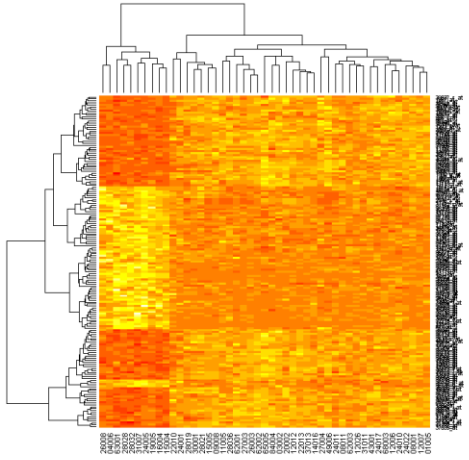




Tabular and Relational Data

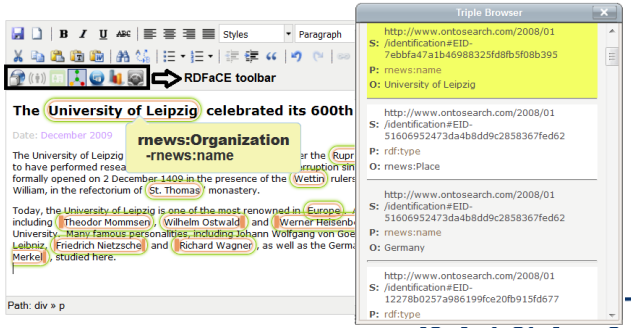
- A fundamental type incorporates numeric, ordinal, categorical, and textual variables, as well as unstructured metadata
  - Each data record can be turned into a document
  - A value in the record is converted into a keyword
- Relational datasets may contain a set of tables
  - Tables can be processed through controlled denormalization





Challenge: Interactively Incorporating User Input

- *How can we employ user input for a large dataset?*
- *How can we design an effective interface for users?*
- Users tag or comment over instance data, then apply to remaining data
- Systems developed in semantic web and information extraction
  - Mostly dedicated to unstructured text
- A good visual interface needed



Challenge: Organizing Data

- *How can we store and organize data?*
- *How can we make the organization efficient for text analysis and interactive visualization?*
- To support:
  - fast query and retrieval by text
  - query and retrieval by original data items
  - fast computation promoting interactive visualization

Stage 2: Text Data Processing

- A set of techniques in natural language processing, statistical, and machine learning that extract and analyze the information of textual data
  - Filtering
  - Clustering
  - Classification
  - Query and Search

KENT STATE UNIVERSITY

Studying Textualized Data

- Find information patterns from textualized data
  - Utilize the term (keyword) vector model to represent data items and datasets
- Fast computing and easy user interaction

Visualization

- Significant discoveries can be expressed using intuitive visualizations and textual explanations written in natural language
- Challenge 1: Expand existing text visualization techniques
- Challenge 2: Combine text visualization with abstract data visualization

Case study: Explore Taxi Trajectory with Semantic Transformation

- Convert each trajectory as a *document* consisting of the taxi-traversed streets
- Enable analysis of massive taxi datasets as document corpora with text mining tools
- Use LDA Topic modeling to infer hidden patterns of moving taxi populations
- Visualization based on the *taxi topics*

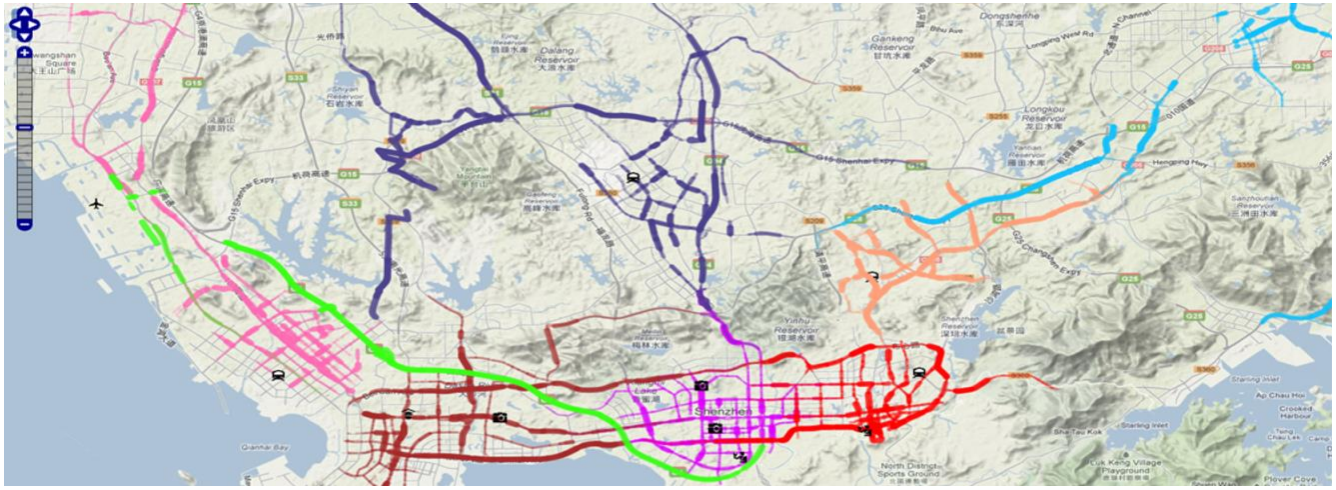
Shenzhen Data

- Daily trajectories of 21,360 taxis in Shenzhen
  - A big city in southern China bordering with Hong Kong
  - Fifteen million residents in a condensed area
  - taxis are a major means of passenger transportation
- Each taxi reports nearly three thousand GPS sample positions per day
- Each sample consists of taxi plate, time, status, speed, direction, and latitude and longitude
- A total of 59,087,230 samples recorded in one day.

KENT STATE UNIVERSITY

Taxi Topic

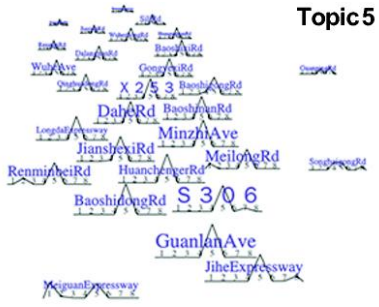
- Reveal typical traveling patterns of city cabs

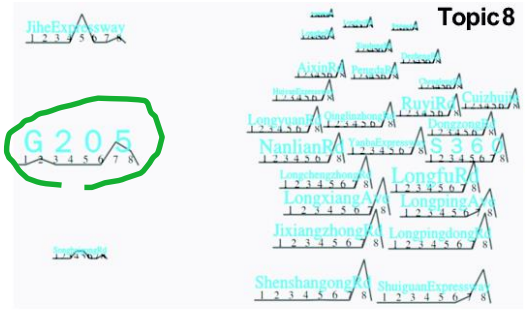


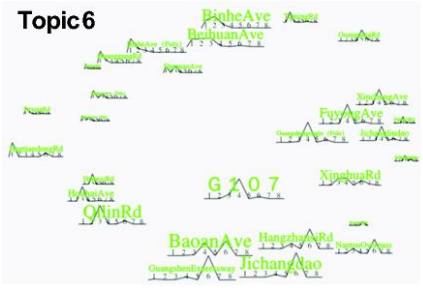
- The topics approximate the city's functional regions
  - A large portion of taxis can accomplish their movement inside a district
- Topics are more than geometrical divisions
  - An airport highway (Green) is an important component of several topics, connecting different regions

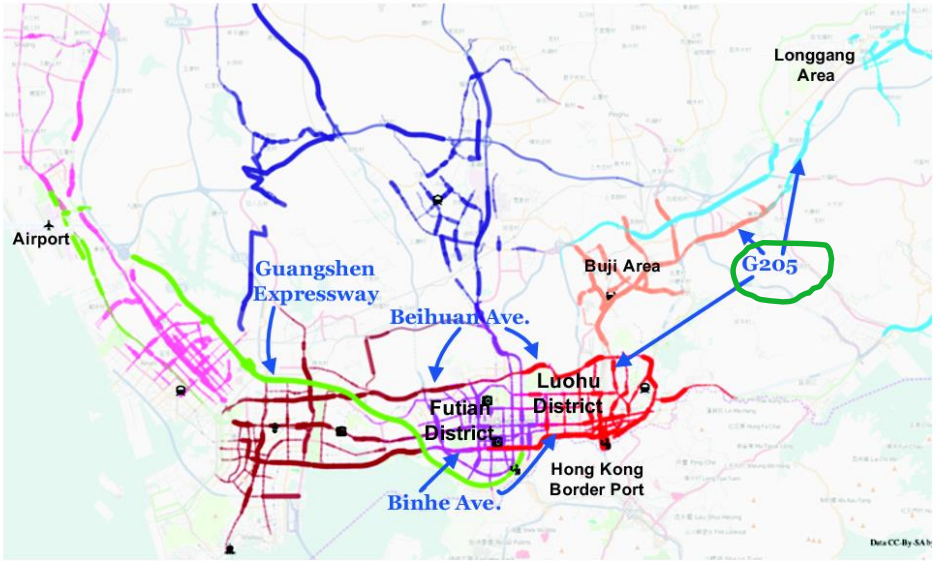
KENT STATE UNIVERSITY

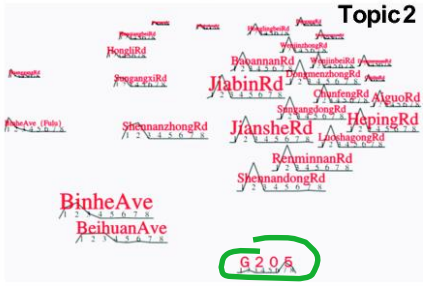
Eight Topics

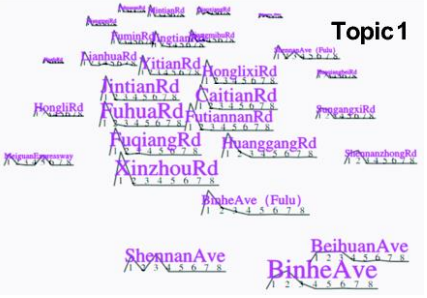










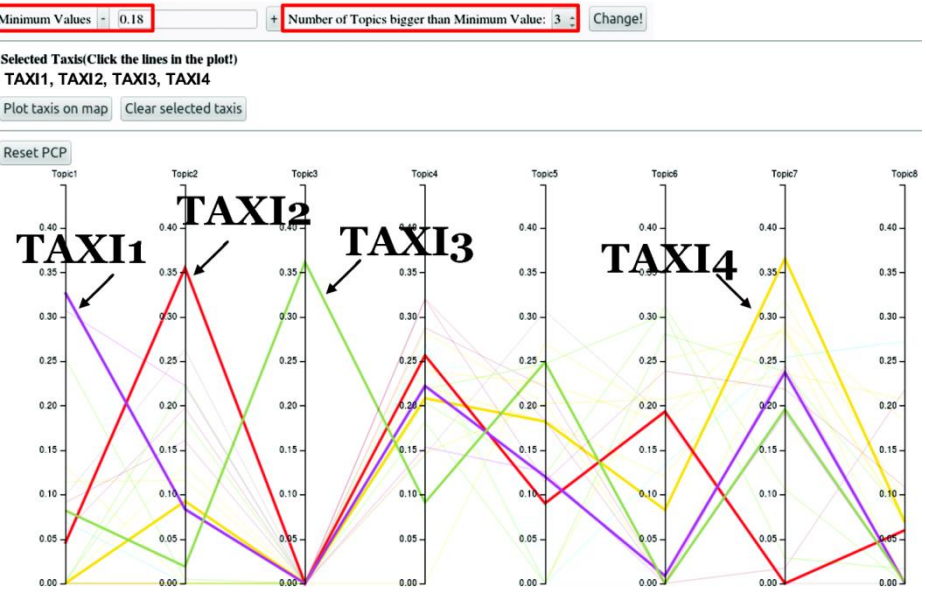


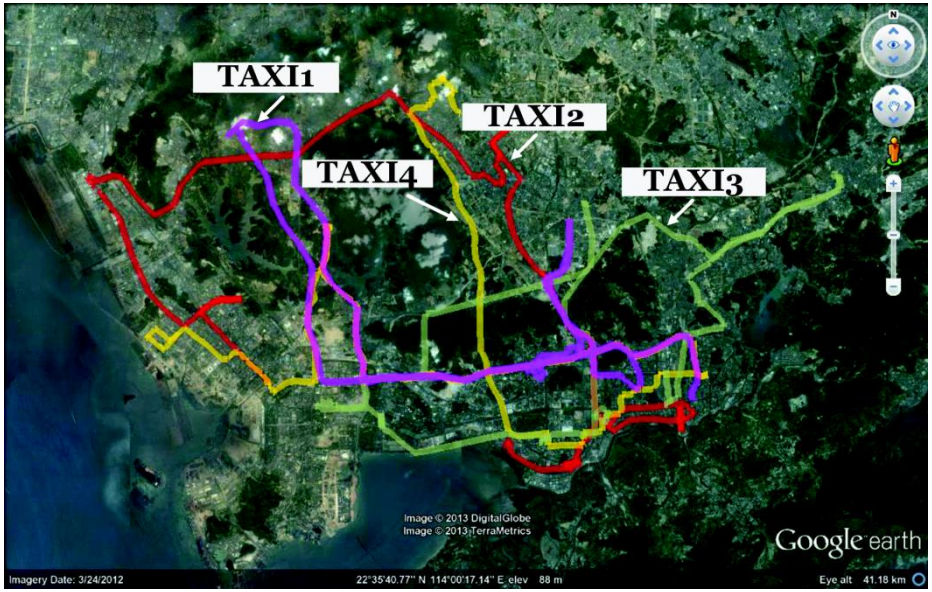
KENT STATE UNIVERSITY

PCP-based Analysis of Topic 2

KENT STATE UNIVERSITY

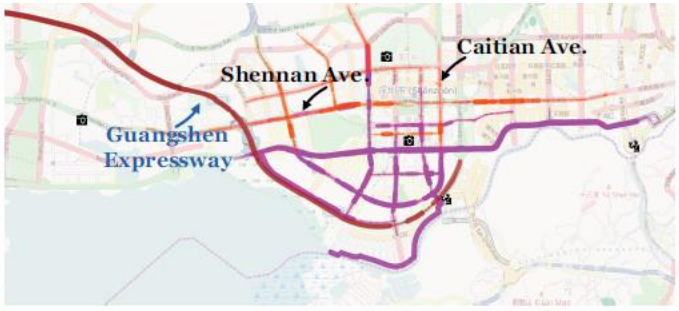
Topics and Trajectories





KENT STATE UNIVERSITY

Visualize Street Changes



(b) Topic2: 3pm-6pm

(c) Topic2: 6pm-9pm

- Visualizing disappearing (brown) and emerging (orange) streets

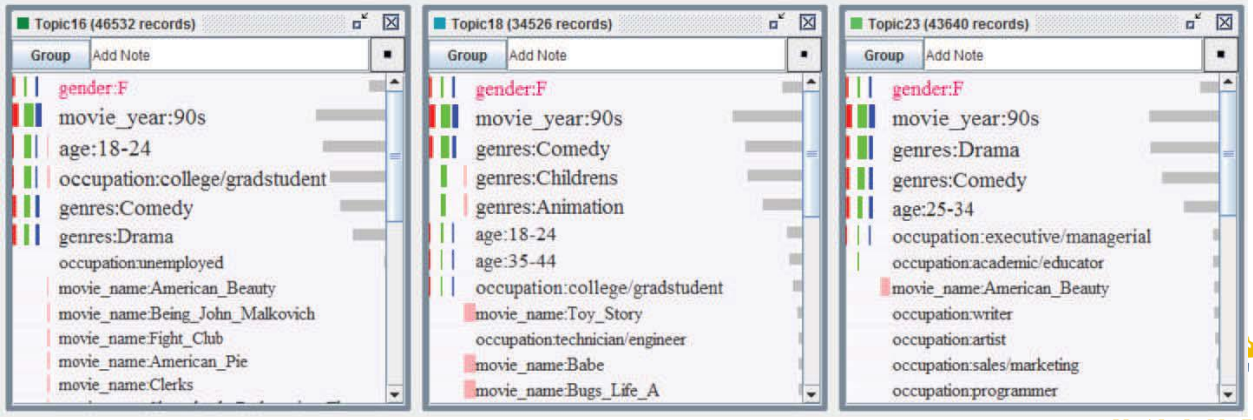
KENT STATE UNIVERSITY

Case Study: Explore Categorical Datasets as Documents

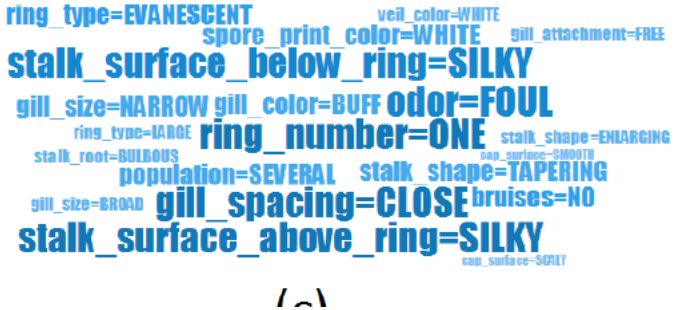
- Map each record in a categorical dataset to a document represented by a bag of categories
- Convert a categorical dataset into a document corpus
- Apply text-based cluster analysis (LDA) to discover subspace clusters of textual category values
- Use associate rule mining discovers optimal risk rules describing multivariate relationships in the textualized topical subspaces

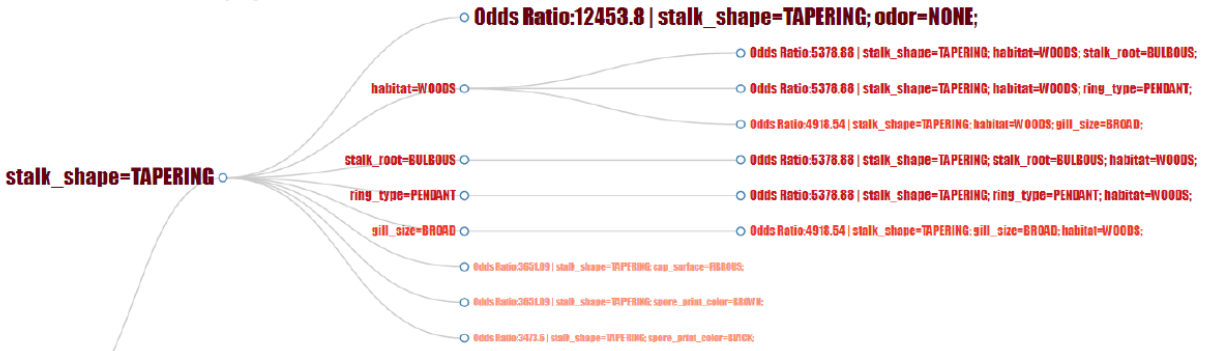
MovieLens dataset

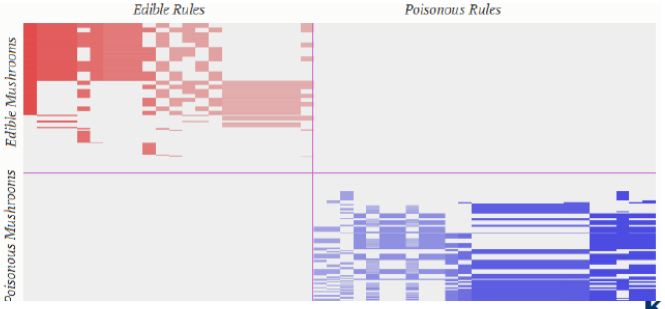
- 1,000,209 data items representing rankings from 6,040 users for 3,883 movies
- Use word clouds, word tree, and fingerprinting, are then used to visualize the rules and data items for interactive knowledge discovery
- Topic Cards: https://www.youtube.com/watch?v=W2Kt7WKIMTI



Mushroom Data





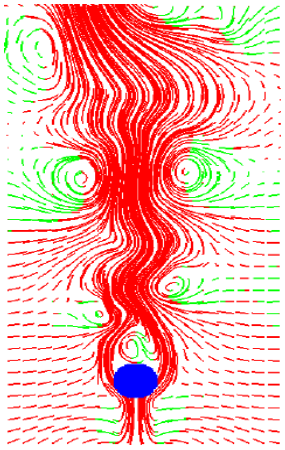


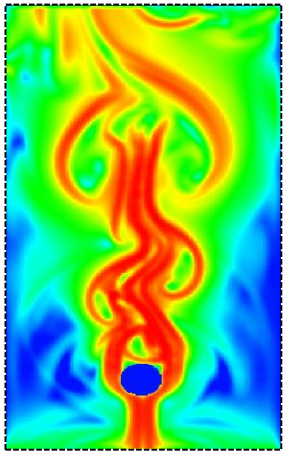
KENT STATE UNIVERSITY

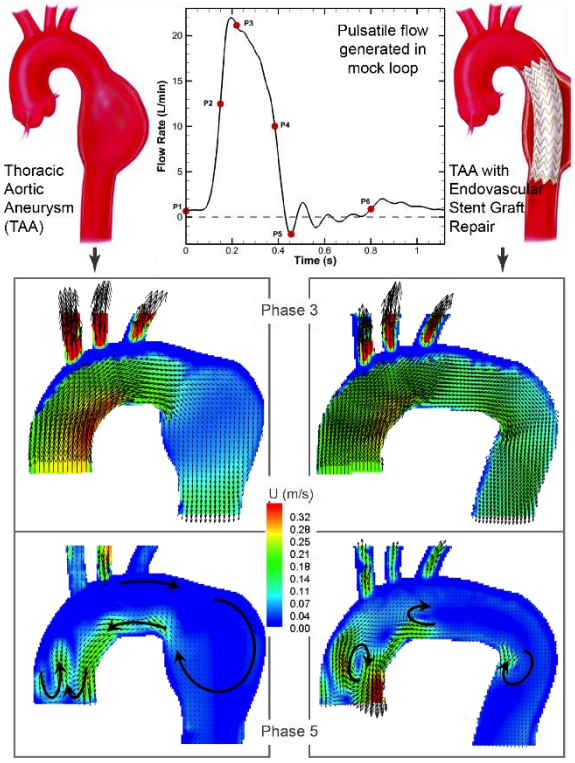
Working on Evaluation

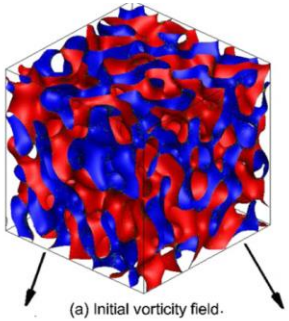
- The effectiveness and efficiency of algorithms of textualization methodology?
- Compare them with other visualizations?
- User study over different applications

Use in More Visualization Tasks?

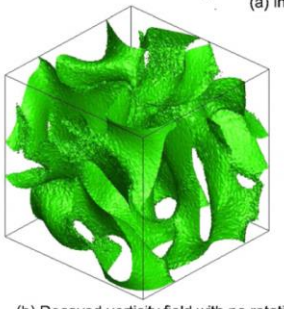




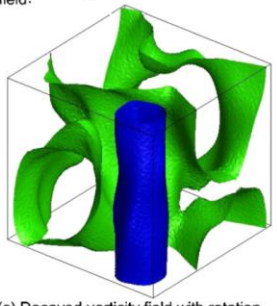




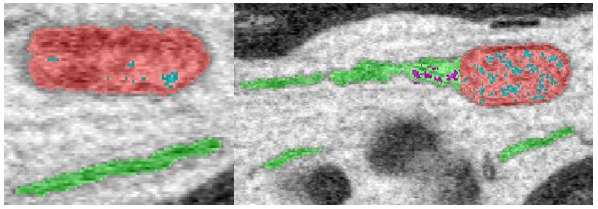
(a) Initial vorticity field.



(b) Decayed vorticity field with no rotation.



(c) Decayed vorticity field with rotation.



KENT STATE UNIVERSITY

Conclusion

- Text can be of interest in abstract data analysis and visualization
- Text analysis tools are enabled
- Text visualization to be compactly integrated with existing visualizations
- Specific approaches for different data types and applications

Thanks!

- Collaborator
  - Jing Yang, UNC-Charlotte
- Acknowledgements
  - Wei Chen, Zhejiang University
  - George Chen, Maogong Zheng, Shenzhen Institute of Advanced Technologies
  - Blake Stringer, Aeronautics Program, Kent State U.
  - Xiaoling Pu, Dept. of Finance, Kent State U.
  - Davis Sheets, Ding Chu, Xiaoke Huang, Wendy Wu, Shamal Aldohuki, Farah Kamw, Yang Chen, Yueqi Hu, Chong Zhang, Scott Barlowe

KENT STATE UNIVERSITY