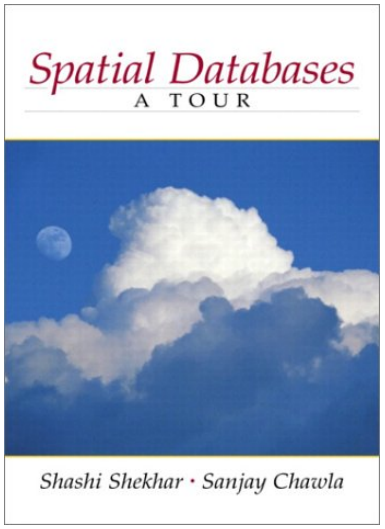
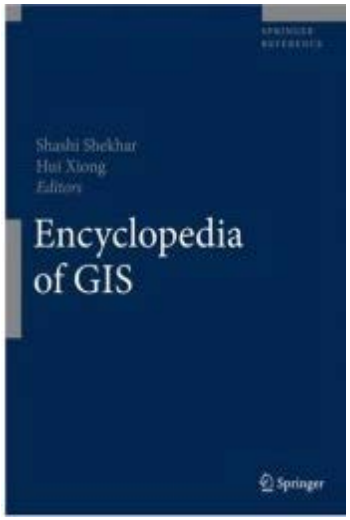
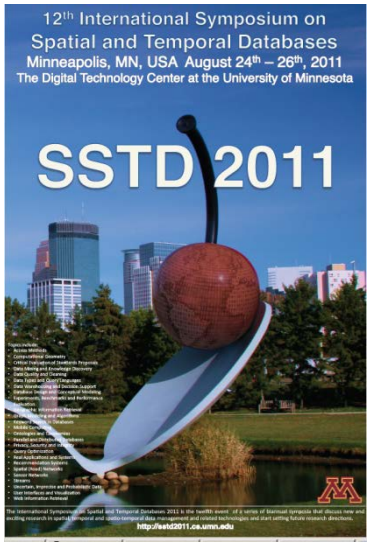
Spatial Big Data Analytics: Challenge & Opportunities

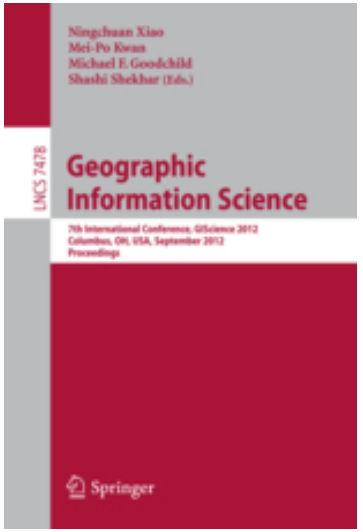
NSF Workshop on Big Data and Urban Informatics
March 28-29, 2013.

Shashi Shekhar
McKnight Distinguished University Professor
Department of Computer Science and Engineering
University of Minnesota
www.cs.umn.edu/~shekhar



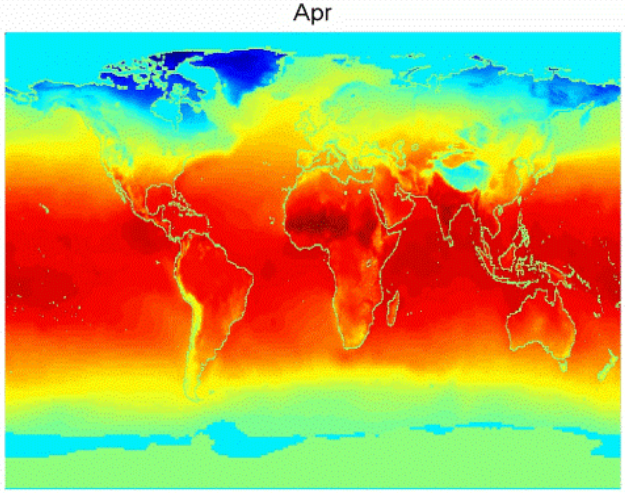






Spatial Big Data (SBD) Overview

- Emerging "Big" Datasets
  - Cell-phone trajectories, Geo-social media, climate projections, …
  - What is so"Big" about these datasets? : Volume, Velocity, Variety, Combinatorics, …
- Complex Societal Science Questions
  - Ex. Sustainable urban (re-)design despite climate change
  - Held back in past due to paucity of data, (understanding and methods)
- These exceed traditional system capacity
  - To analyze data to make decisions and take actions (with reasonable effort)
  - Platforms: Cloud Computing, Map-Reduce, Pregel, .
  - Analytics, e.g., Data Mining, Spatial Statistics, …
  - Scientific Methods, e.g., Hypothesis testing, …



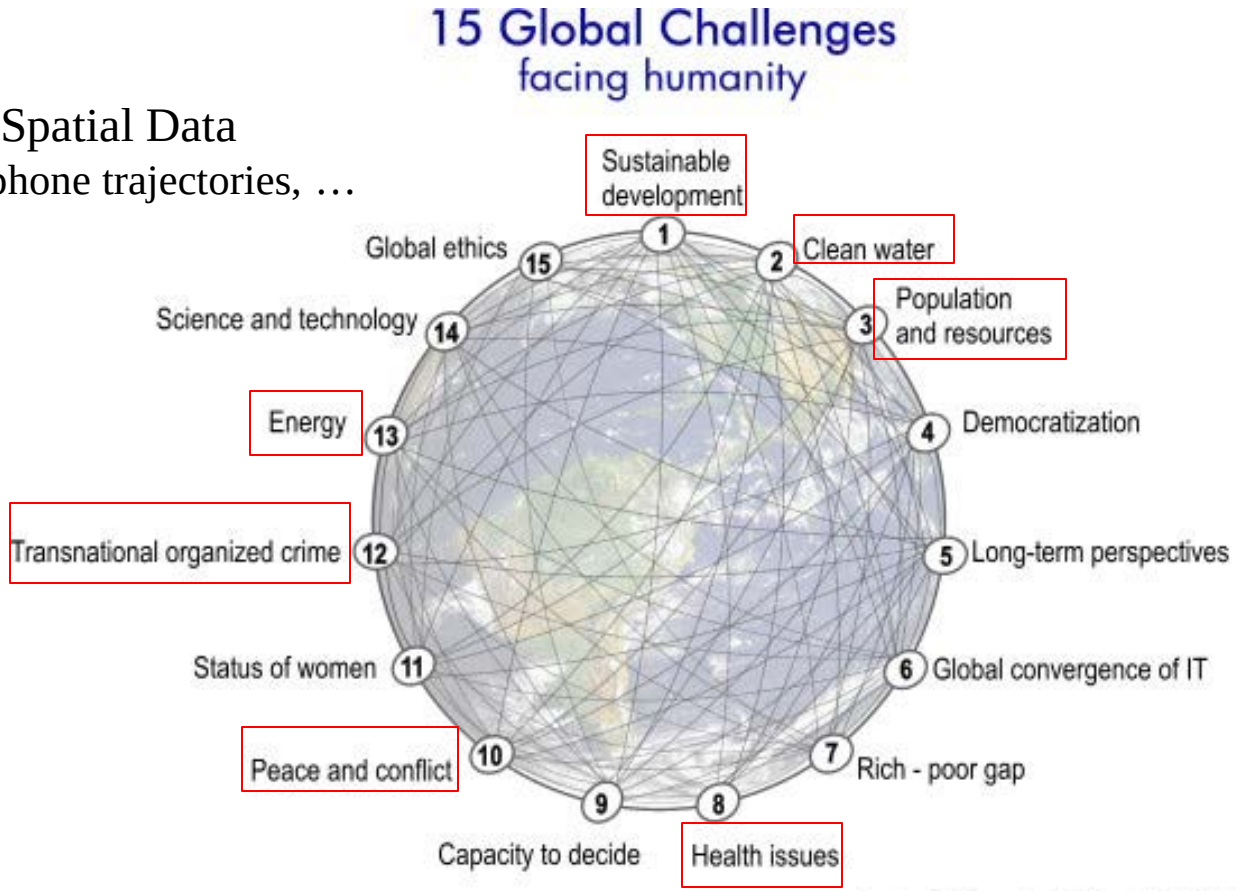
Vector SBD from Geo-Social Media

- Point reports:
- Road center-lines, …
- Persistent Surveillance
  - Outbreaks of tornadoes, , Disaster, Unrest, Crime, …
  - Emerging hot-spots, Spatio-temporal patterns

• Even before cable news outlets began reporting the tornadoes that ripped through Texas on Tuesday, a **map** of the state began blinking red on a screen in the Red Cross' new social media monitoring center, alerting weather watchers that something was happening in the hard-hit area. (AP, April 16th, 2012)

Societal Questions and Spatial Big Data (SBD)

- Global Challenges facing humanity
  - Sustainable cities – urban re-development
  - Energy efficiency, e.g., eco-routing
  - Healthy Cities, e.g., Exposomics
  - …
- Many may benefit from Big Spatial Data
  - Google Earth/Maps, cell-phone trajectories, …



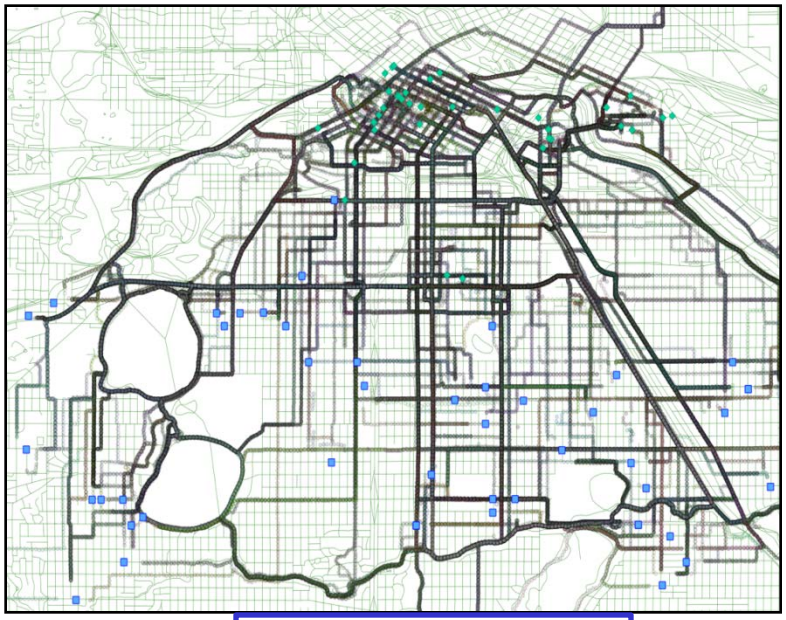
Green Corridor Planning with GPS Data

Federal funding for bike routes pays off in Twin Cities

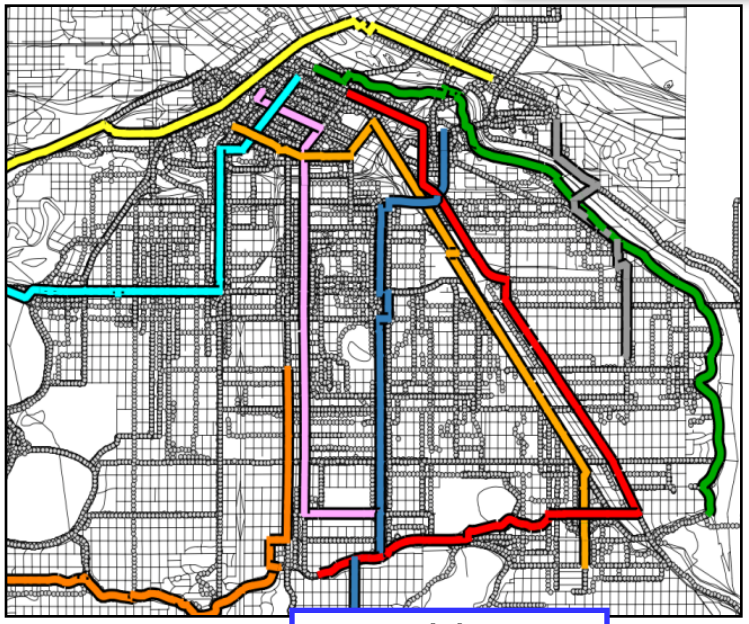
Article by: JOSEPHINE MARCOTTY , Star Tribune | Updated: May 4, 2012 - 2:57 PM

Four areas that shared $100 million saw growth in biking and walking, with less pollution. spent on health care

- Bike corridors
- Bus routes, Light-rail lines, New or express Highway lanes



819 tracks, 49 volunteers
128,083 GPS Points
57,644 nodes map



8 corridors
from PAM K-Main Corridor Algorithm

Big Data and Science

Nature, 7209(4), September 4, 2008
"Above all, data on today's scale require scientific and computational intelligence. Google may now have its critics, but no one can deny its impact, which ultimately stems from the cleverness of its informatics. The future of science depends in part on such cleverness again being applied to data for their own sake, complementing scientific hypotheses as a basis for exploring today's information cornucopia."

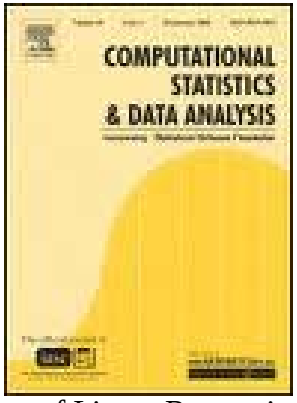
Big Data Translates into Big Opportunities... and Big Responsibilities

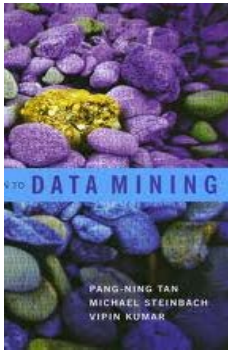
Sudden influxes of data have transformed researchers' understanding of nature before — even back in the days when 'computer' was still a job description.

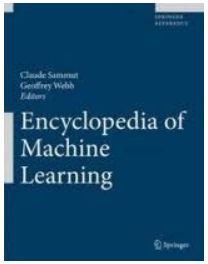
Unfortunately, the institutions and culture of science remain rooted in that pre-electronic era. Taking full advantage of electronic data will require a great deal of additional infrastructure, both technical and cultural

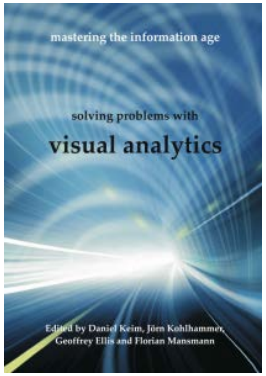
Scientific Methodologies: Fourth Paradigm

Models	Traditional (Manual)	Computer Assisted
Forward	(Fasifiable) **Theory**, Differential Equations (D.E.)	Computational **Simulation**s using D.E.s, Agent-based models, etc.
Backward	Controlled **Experiments**, Test Hypothesis(Gallelio, 1638) Random sampling (1890), Correlation (Pearson, 1880), Regression (Galton,1877), …	**Bayesian:** resampling, local regression, MCMC, kernel density estimati… generalized additive m… **Frequentist:** frequent… generation, Model ens… **Exploratory Data An…** visual analytics, geographic information science, **spatial data mining**, …



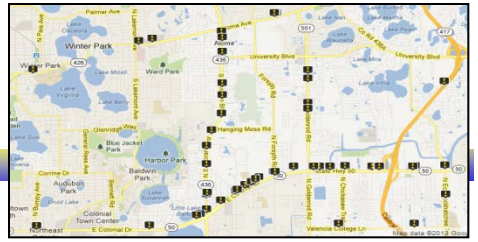




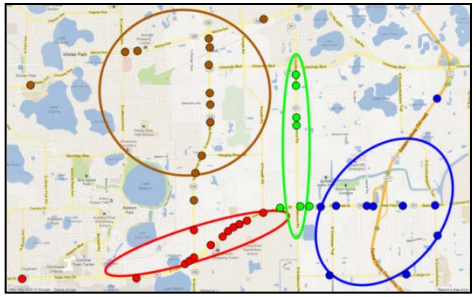


J. Stanton et al., Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors, Journal of Statistical Education, 9(3), 2001.

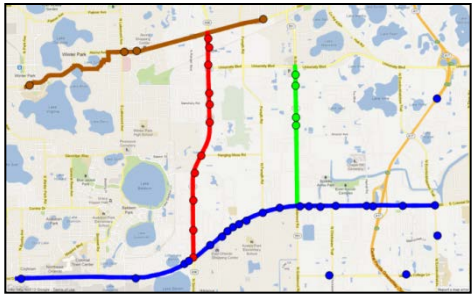
Summarization & Big Data



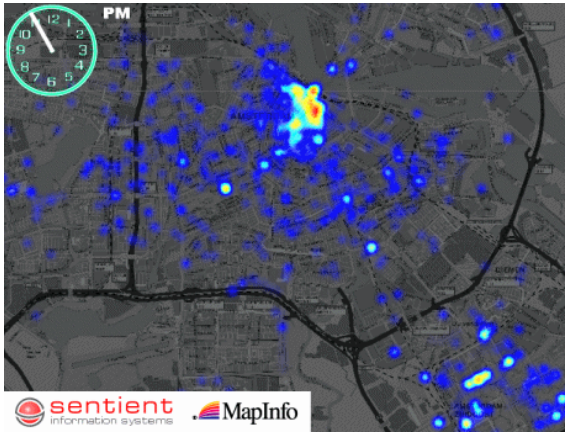
- Example: Unusally high spatial concentration
  - Cancer clusters, crime hotspots, pedestrian fatalities

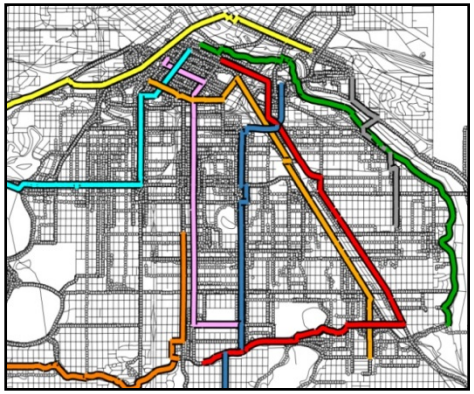


- Focus: Point Data, e.g., Tweets
  - Maps, e.g., Red Cross tweet map
  - K-Means, E.M. ellipsoids



- Big Data Opportunities: reduce semantic gap
  - Spatial Networks: route based summary
  - Summarize GPS tracks
  - Dynamics, e.g., emerging hot-spots

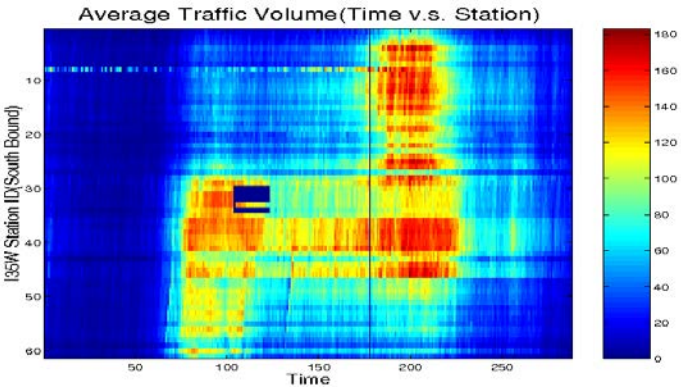




Anomalies & Big Data

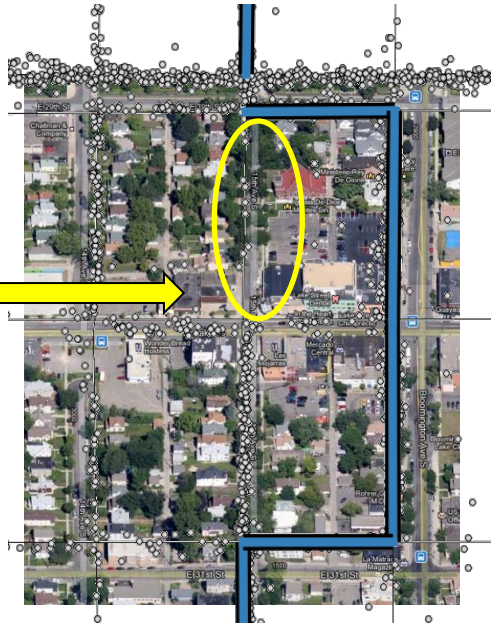
- Geo-anomalies
  - Location different from neighbors
  - Anomalous trajectories
  - Flow discontinuities

- Big Data Opportunities
  - Exceptions to patterns of life



Why is a cycling route not straight?

Why is a block Avoided by cyclists?



Association, Colocation, Co-occurrence

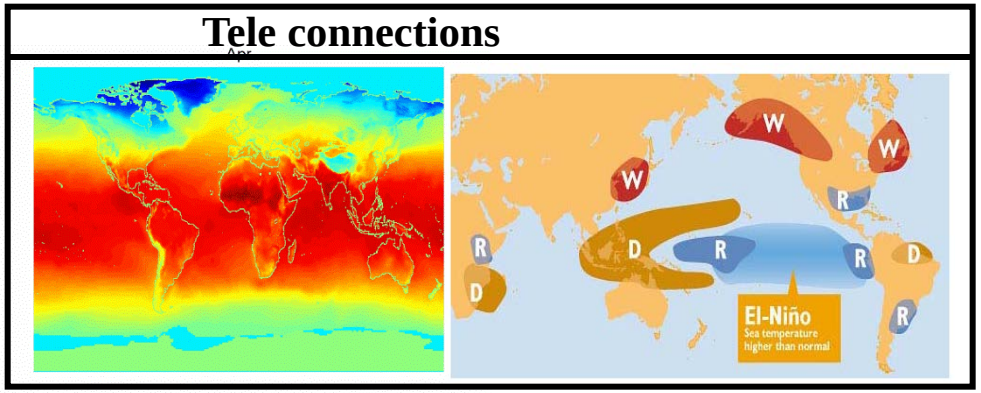
November 14, 2004 The New York Times

What Wal-Mart Knows About Customers' Habits

- 1990s: Which items are bought together?
  - correlation was too expensive on 460 Tbytes
  - Alternative: Association rule

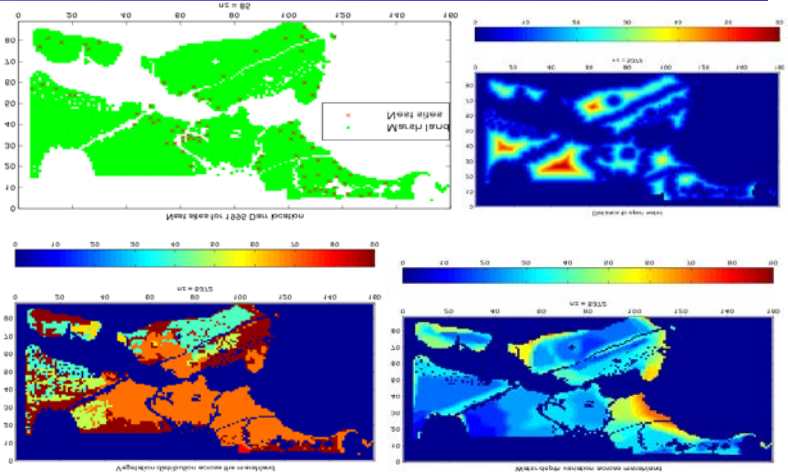
- 2000s: Which items are bought for hurricanes?
  - Transactions not natural for continuous space
  - Spatial Neighbor Graph
  - Colocation Patters, Ripley's K-function

- Future: Spatial Big Data Issues
  - Beyond Tobbler's law: Tele-connections
    e.g., impact of El Nino of a city
  - Spatio-temporal co-occurrence



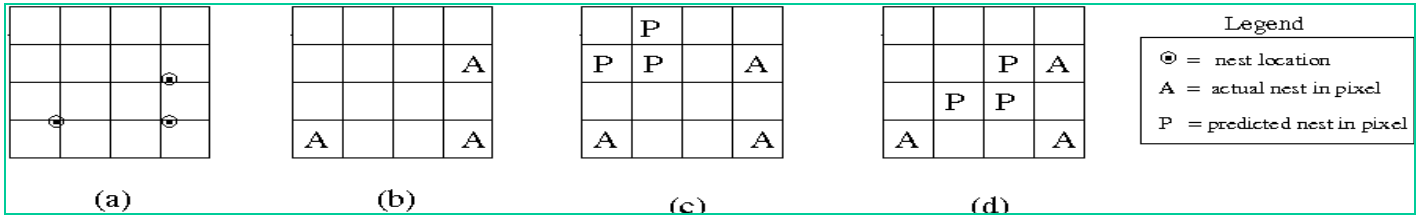
Space/Time Prediction

- Predict location, time, path, …
  - Nest sites, minerals, Hurricanes, tornadoes, …
- Data violates Independent Identical Distribution (I.I.D.)
- Auto-correlation: Tobbler's Law, Markov Model
- Heterogeneity, e.g. geo. weighted regression (GWR)
- Big Data Opportunities
  - Post Markov: Estimate W matrix from big data
  - Beyond GWR: place based ensemble of SVM …
  - Beyond I.I.D. error => Model spatial error!
  - Scalable algorithms for parameter estimation
    - But map-reduce too slow for iterative methods
    - Determinant(large)is hard!



Name	Model
Linear Regression	$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
Spatial Auto-Regression	$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$$\ln(L) = \ln\left|\mathbf{I} - \rho\mathbf{W}\right| - \frac{n\ln(2\pi)}{2} - \frac{n\ln(\sigma^2)}{2} - SSE$$



SBD Potential

McKinsey Global Institute

The study estimates that the use of personal location data could save consumers worldwide more than $600 billion annually by 2020. Computers determine users' whereabouts by tracking their mobile devices, like cellphones. The study cites smartphone location services including Foursquare and Loopt, for locating friends, and ones for finding nearby stores and restaurants.

But the biggest single consumer benefit, the study says, is going to come from time and fuel savings from location-based services — tapping into real-time traffic and weather data — that help drivers avoid congestion and suggest alternative routes. The location tracking, McKinsey says, will work either from drivers' mobile phones or GPS systems in cars.

The New York Times

Published: May 13, 2011

New Ways to Exploit Raw Data May Bring Surge of Innovation, a Study Says