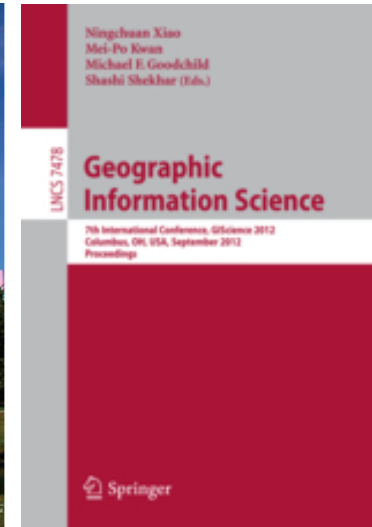
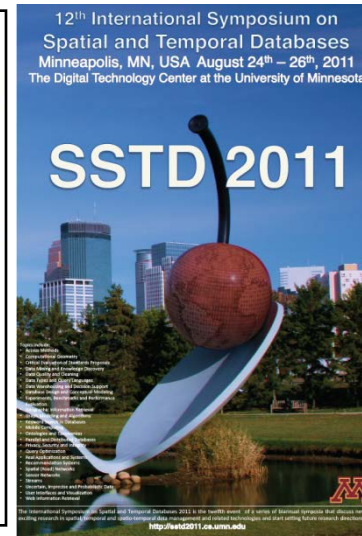
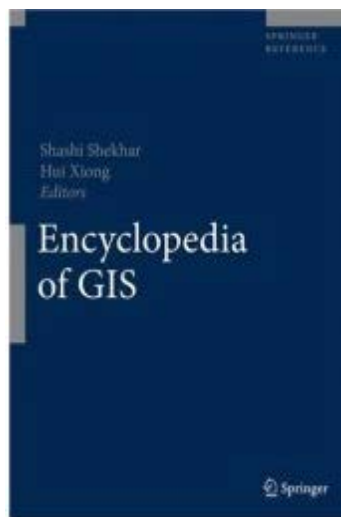
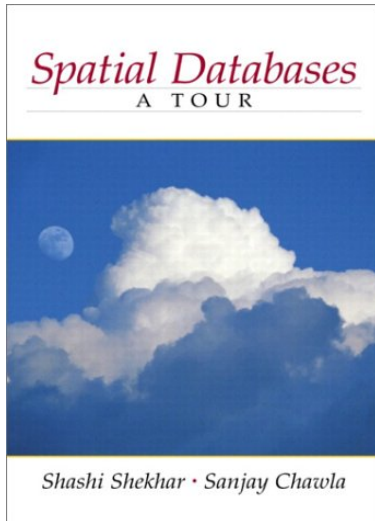


Spatial Big Data: A Perspective

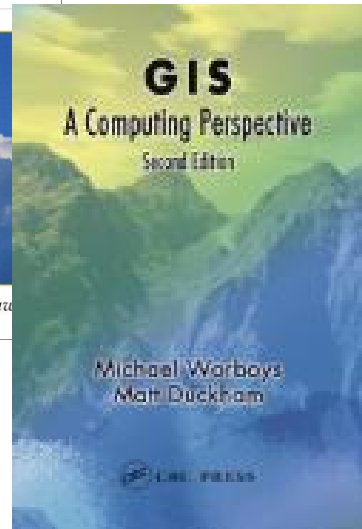
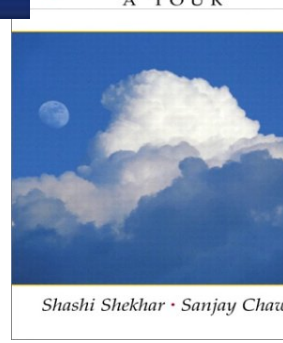
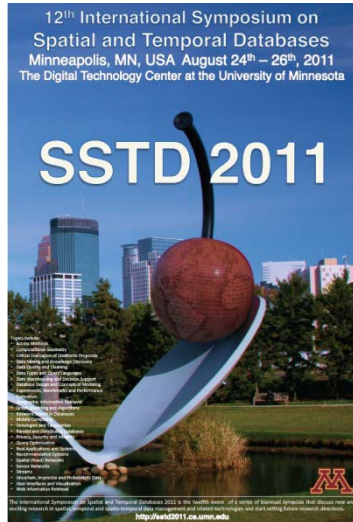
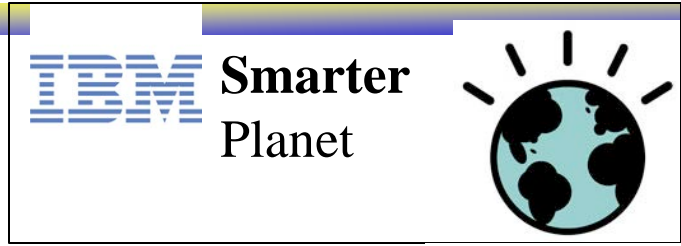
Computing, Informatics and Decision Systems Eng. Distinguished Lecture Series
Arizona State University
May 16th, 2013.

Shashi Shekhar

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



Spatial Computing



CCC Workshop: Spatial Computing Visioning (9/10-11/2012)

http://cra.org/ccc/spatial_computing.php



Computing Community Consortium

We support the computing research community in creating compelling research visions and the mechanisms to realize these visions.

HOME

ABOUT

YOUR VISION

ACTIVITIES

RESOURCES

CONTACT

GO

Funded Visioning Activities

Disaster Management SEES IT HealthIT Interactive Tech Architecture XLayer Robotics Learning Tech
Open Source Cyber Physical Systems Global Development Theoretical CS Big Data Computing NetSE
Spatial Computing

From GPS and Virtual Globes to Spatial Computing-2020

About the workshop

This workshop outlines an effort to develop and promote a unified agenda for Spatial Computing research and development across US agencies, industries, and universities. See the original workshop proposal [here](#).

Spatial Computing

Spatial Computing is a set of ideas and technologies that will transform our lives by understanding the physical world, knowing and communicating our relation to places in that world, and navigating through those places.

The transformational potential of Spatial Computing is already evident. From Virtual Globes such as Google Maps and Microsoft Bing Maps to consumer GPS devices, our society has benefitted immensely from spatial technology. We've reached the point where a hiker in Yellowstone, a schoolgirl in DC, a biker in Minneapolis, and a taxi driver in Manhattan know precisely where they are, nearby points of interest, and how to reach their destinations. Large

Logistics

Date: Sept. 10th-11th, 2012

Location: Keck Center

Hotel: Liaison Hotel

Steering Committee

Erwin Gianchandani

Hank Korth

Organizing Committee

Peggy Agouris, George Mason University

Walid Aref, Purdue University

Michael F. Goodchild, University of California - Santa Barbara

CCC Workshop: Spatial Computing Visioning

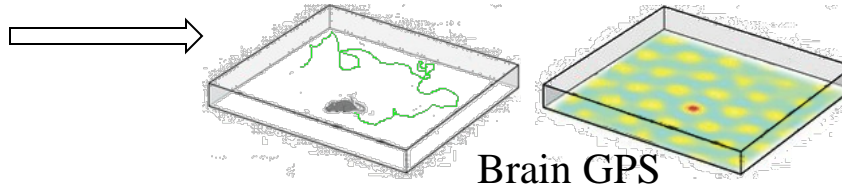
• Trends

- Everyone uses spatial computing
- Everyone is a mapmaker and many more phenomena are observable
- Every platform is location aware
- Expectations are rising and so are the risks

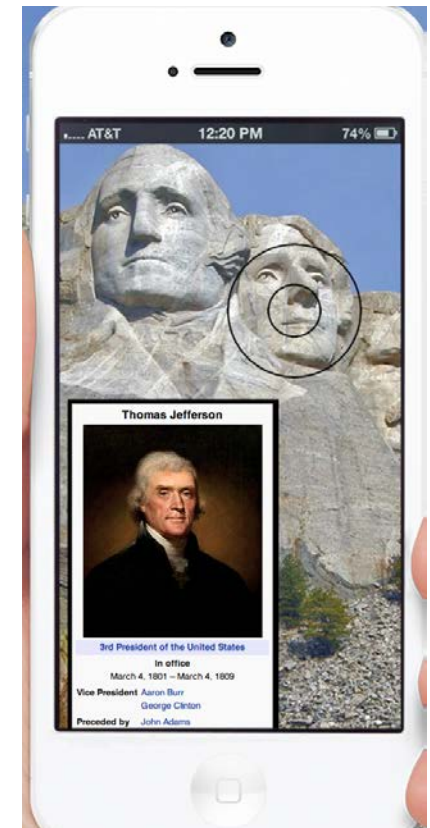
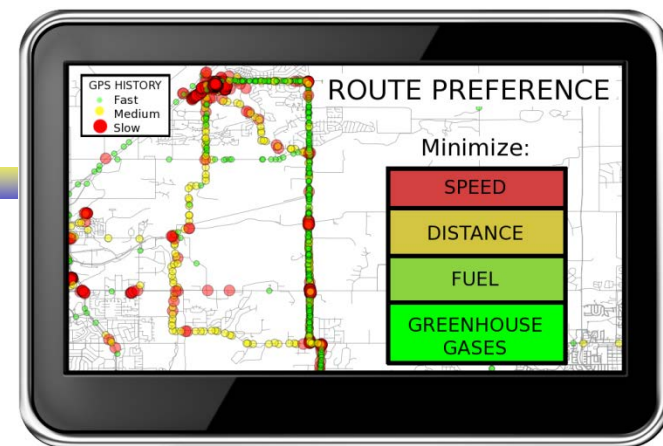
4

• Ten Promising Directions

1. Spatial Abilities Predict STEM Success
2. **Emerging Spatial Big Data**
3. Augmented Reality Systems
4. Time-Travel in Virtual Globes
5. Spatial Predictive Analytics
6. Persistent Environment Hazard Monitoring
7. Geo-collaborative Systems, Fleets, and Crowds
8. Localizing Cyber Entities
9. GPS Deprived Environment
10. Beyond Geo



Brain GPS



Big Data - Motivation

The New York Times

Published: May 13, 2011

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

Mining and analyzing these big new data sets can open the door to a new wave of innovation, accelerating productivity and economic growth. Some economists, academics and business executives see an opportunity to move beyond the payoff of the first stage of the Internet, which combined computing and low-cost communications to automate all kinds of commercial transactions.

Estimated Value >Usd 1 Trillion per year by 2020

Location-based service: usd 600 B

Health Informatics: usd 300 B

Manufacturing:

...

McKinsey Global Institute



Big data: The next frontier for innovation, competition, and productivity

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

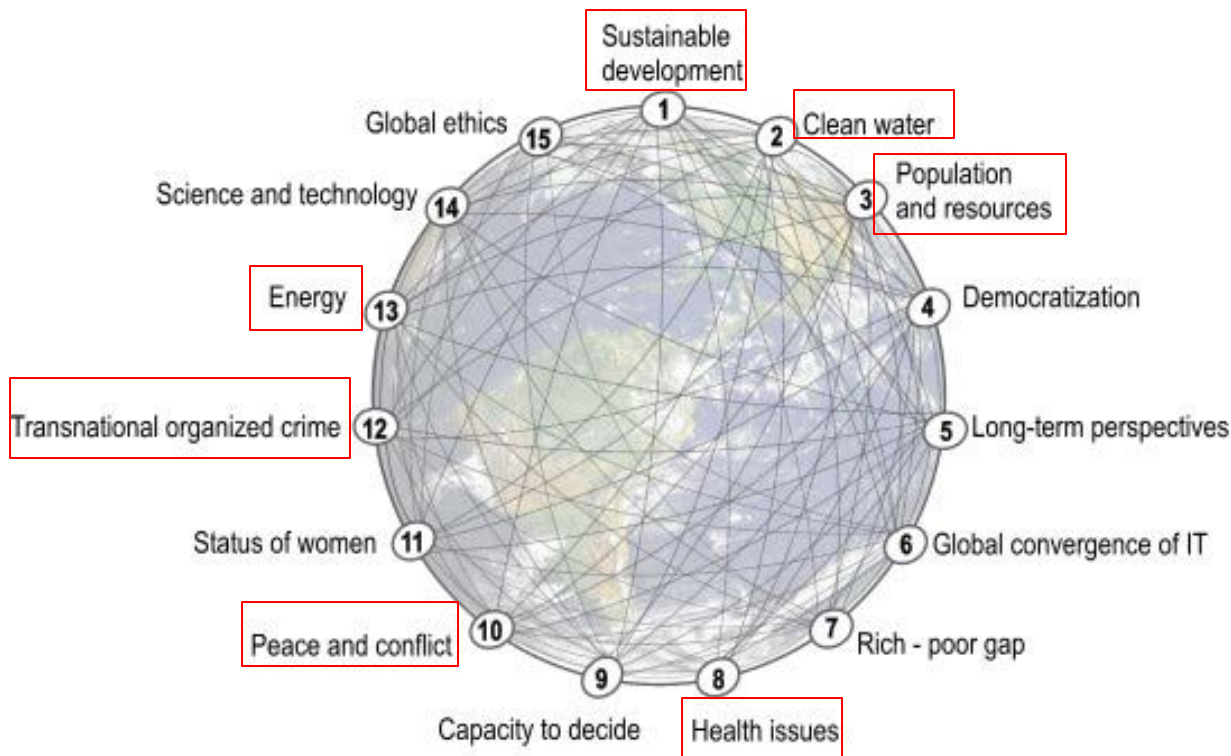
New Ways to Exploit Raw Data May Bring Surge of Innovation, a Study Says



Motivation for Spatial Big Data (SBD)

- Societal:
 - Google Earth, Google Maps, Navigation, location-based service
 - Global Challenges facing humanity – many are geo-spatial!
 - Many may benefit from Big Spatial Data

15 Global Challenges facing humanity



Outline

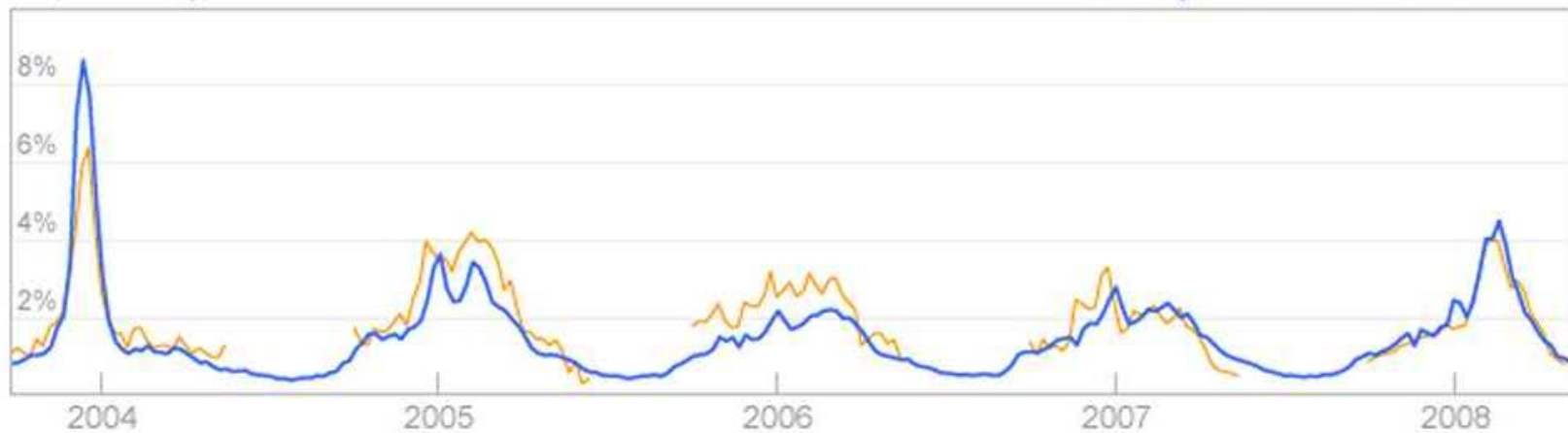
- Motivation
- **Definition**
 - Two Perspectives
 - Examples
- Infrastructure
- Analytics
- Science



Annual U.S. Flu Activity - Mid-Atlantic Region

ILI percentage

● Google Flu Trends ● CDC Data



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)

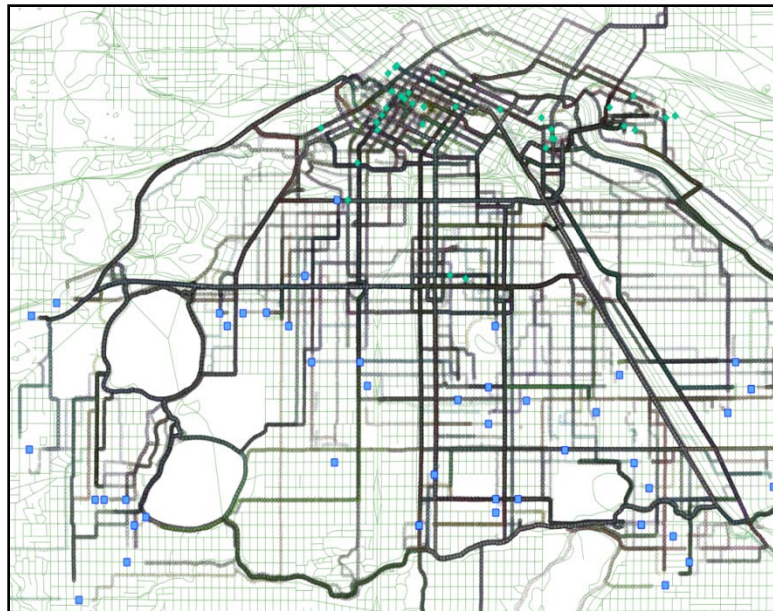
Green Corridor Planning with GPS Data

Federal funding for bike routes pays off in Twin Cities

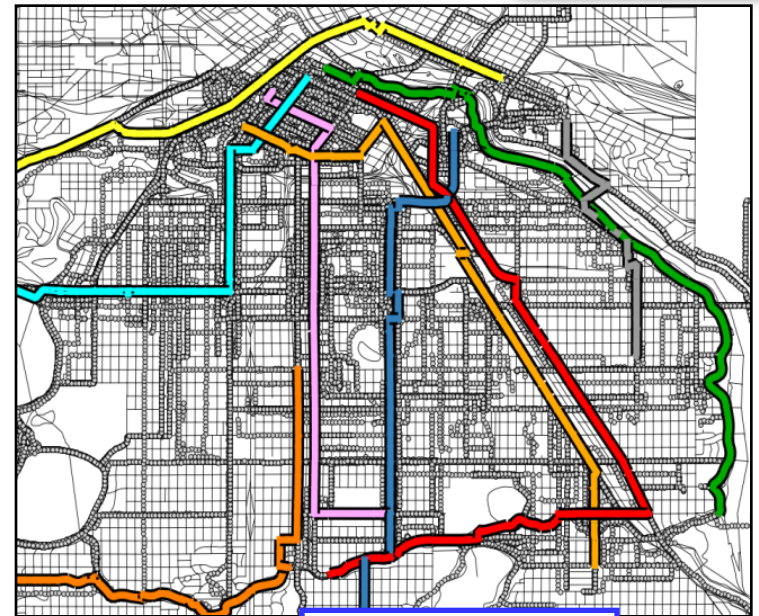
Article by: JOSEPHINE MARCOTTY, Star Tribune | Updated: May 4, 2012 - 2:57 PM

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

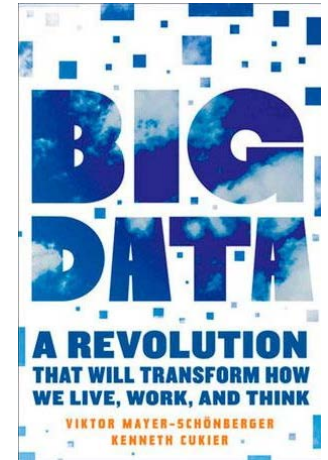


8 corridors
from PAM K-Main
Corridor Algorithm



Spatial Big Data – Data Analysis View

- Be aware of bias in Big Data
- Some time small data is better (and cheaper) – 1930s
 - Representative samples
 - Ex. Random sample, independent identical distributions (i.i.d.)
 - To estimate mean, standard deviation, etc.
- Use big data if it provides **value-added relative to small data**
 - Timeliness
 - Exploration of sub-groups
 - **Needle-in-a-haystack** patterns eclipsed by frequent patterns



Big Data – Computing View

- Datasets exceeding capacity of current computing systems
 - To manage, process, or analyze data with reasonable effort
 - Due to Volume, Velocity, Variety, Veracity, Dimensionality, ...
 - **Q? Is Census big data?**
 - **Which emerging big datasets proxy real-time census?**
- Components
 - Data-intensive Computing: Cloud Computing, Map-Reduce
 - Middleware, e.g., Web-services, Grid
 - Big-Data analytics, e.g., data mining, machine learning, ...



**WHAT IS
BIG
DATA?**

VOLUME
Large amounts of data.

VELOCITY
Needs to be analyzed **quickly**.

VARIETY
Different **types** of structured
and unstructured data.

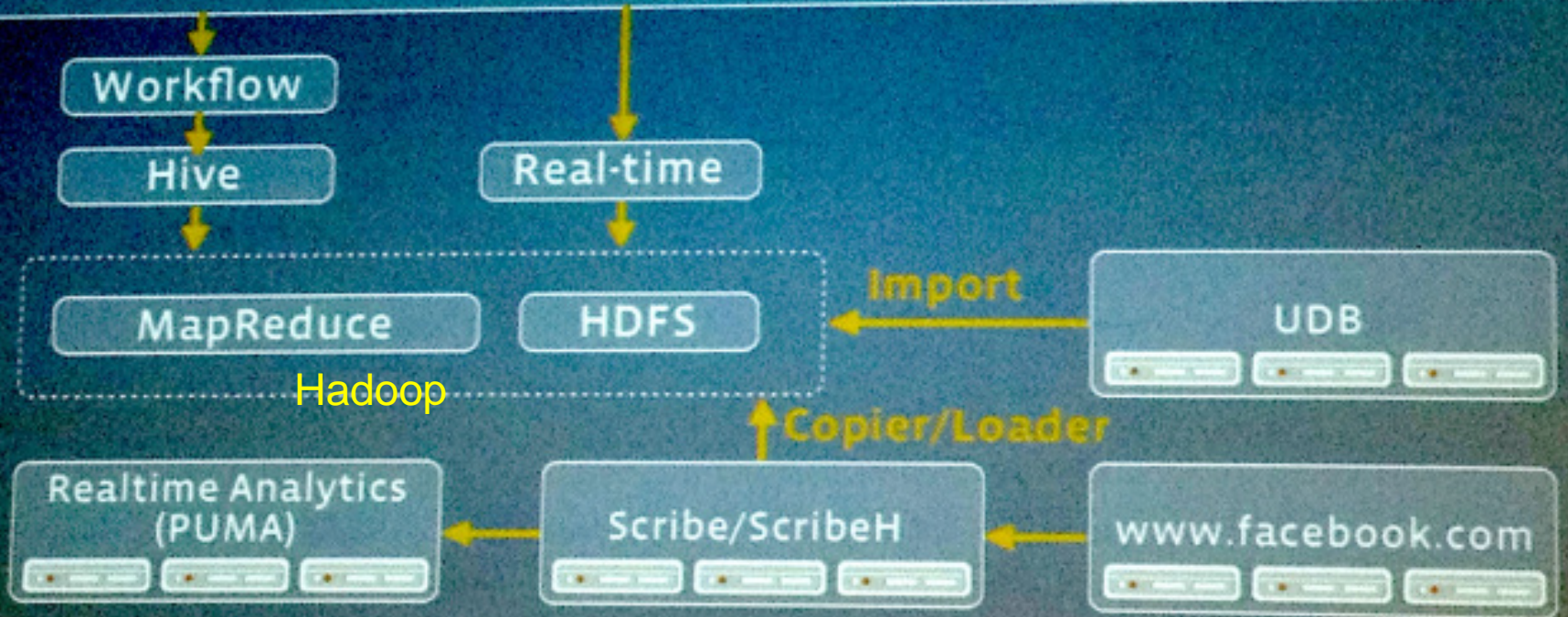
Outline

- Motivation
- Definition & Examples
- **Infrastructure**
 - Computation
 - Storage
- Analytics
- Science

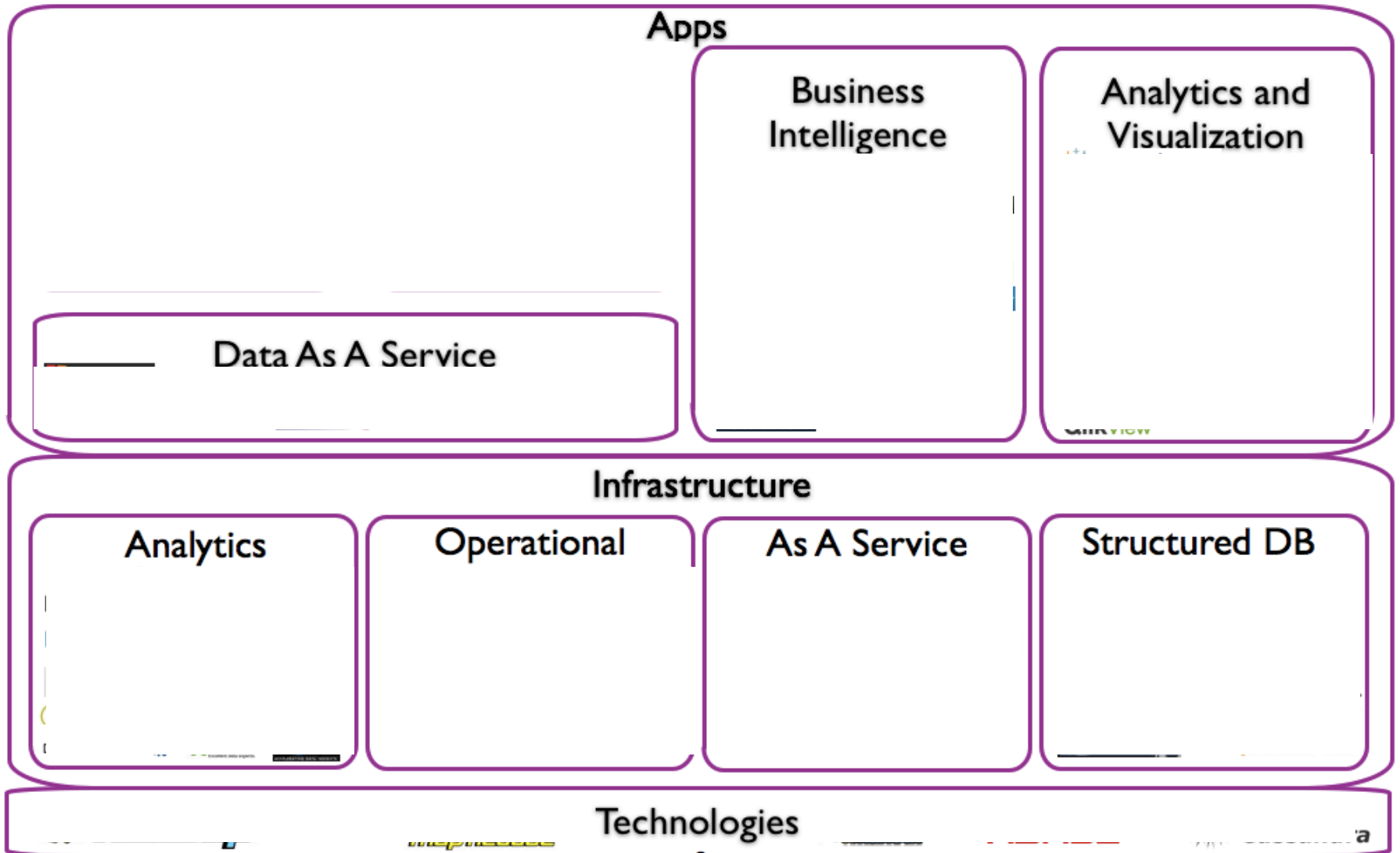


Life of data at Facebook

Data Tools



The Big Data Landscape



Parallelizing Spatial Big Data on Cloud Computing

- Case 1: Compute Spatial-Autocorrelation Simpler to Parallelize
 - Map-reduce is okay
- Case 2: Harder : Parallelize Range Query on Polygon Maps
 - Need dynamic load balancing beyond map-reduce
 - MPI or OpenMP is better!
- Case 3: Estimate Spatial Auto-Regression Parameters, Routing
 - Map-reduce is inefficient for iterative computations due to expensive “reduce”!
 - Option 1: Develop non-iterative formulations of spatial problems
 - Option 2: Alternative Platform: MPI, OpenMP, Pregel

Infrastructure Debates

- Map Reduce as an alternative to Relational Databases
 - NoSQL movement (Not Just SQL)
 - Key-value stores
- Beyond Map Reduce
 - Map Reduce speeds up “embarassingly parallel” problems
 - Not effective for iterative algorithms as “reduce” is costly
- Directions
 - Develop alternatives, e.g., Pregel, MPI/OpenMP
 - Develop non-iterative algorithms
 - Classify problems, where non-iterative algorithms are possible

Outline

- Motivation
- Definition & Examples
- Infrastructure
- **Analytics**
- Science



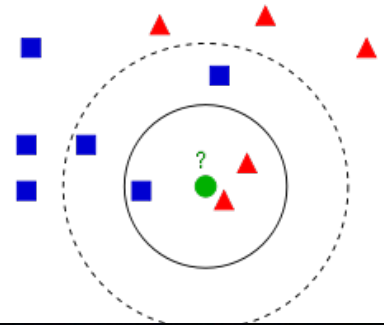
Data Analytics

• Scale-up Traditional Methods with Map-Reduce

- Statistics: Frequentist, Bayesian
- Machine Learning, Data Mining
- Visual analytics

• New Debates

- A. Extreme simplicity
 - Bigger the data, simpler the model!
 - Data as a model (No inferred model)
 - Amenable to broader audience for analytics as a service
- B. Extreme detail or diversity
 - Ensemble of traditional inferred models
 - Spatial Auto-correlation
- C. Combine best of A & B



There are always implicit assumptions behind a model and its solution method. But human beings have limited foresight and great imagination, so that, inevitably, a model will be used in ways its creator never intended. This is especially true in trading environments... but it's also a matter of principle: you just cannot foresee everything. So, even a “correct” model, “correctly” solved, can lead to problems.

The more complex the model, the greater this possibility. (Emanuel Derman 1996)

Prediction: Heterogeneity Challenge - Spatial Ensemble

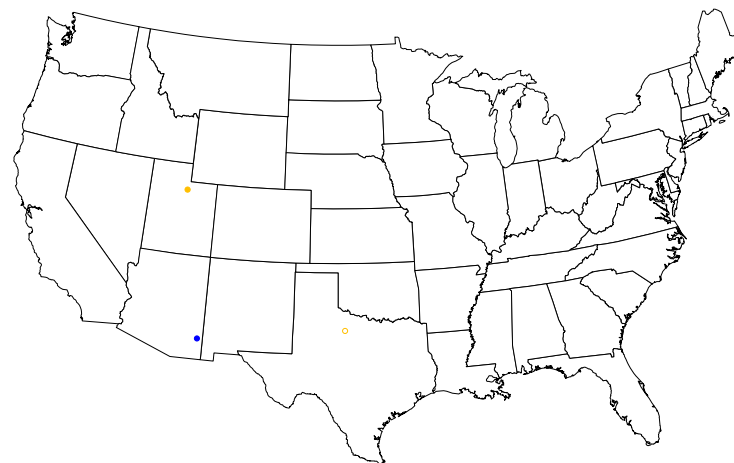
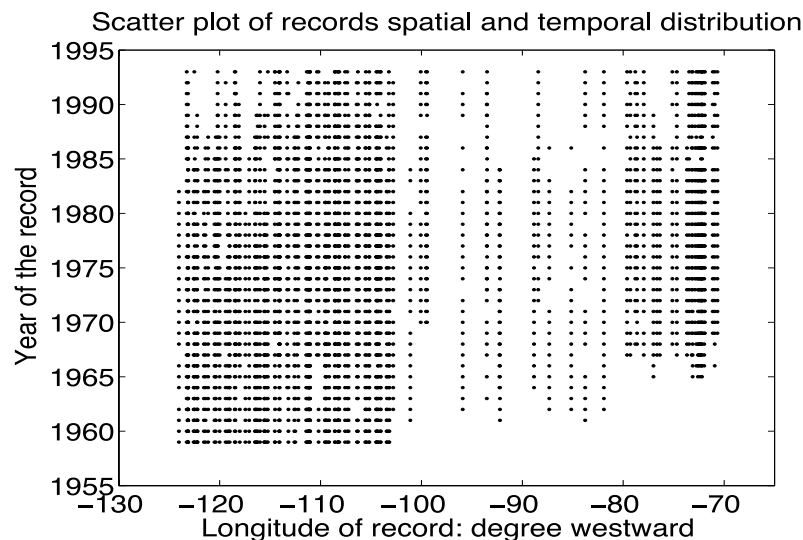
- Example: Lilac Phenology data
 - Yearly date of first leaf and first bloom
 - 1126 locations in US & Canada
- “Global” regression model shows a mystery
 - Positive Slope => blooms delayed in recent years!
- Spatial decomposition solves the mystery
 - East of Mississippi, West of Mississippi
 - Each half has Negative Slope => blooms earlier in recent years!
 - However slopes are different across east & west
 - More reports in west in recent years



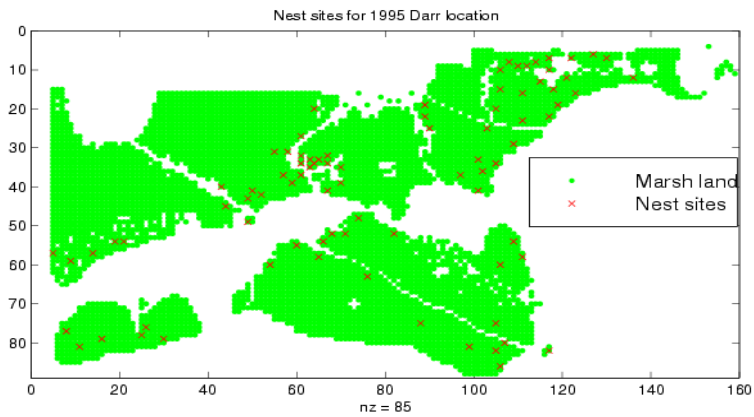
North American Lilac Phenology Data Since 1956 online



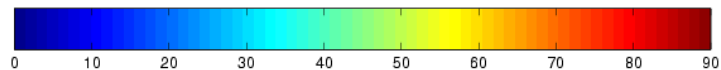
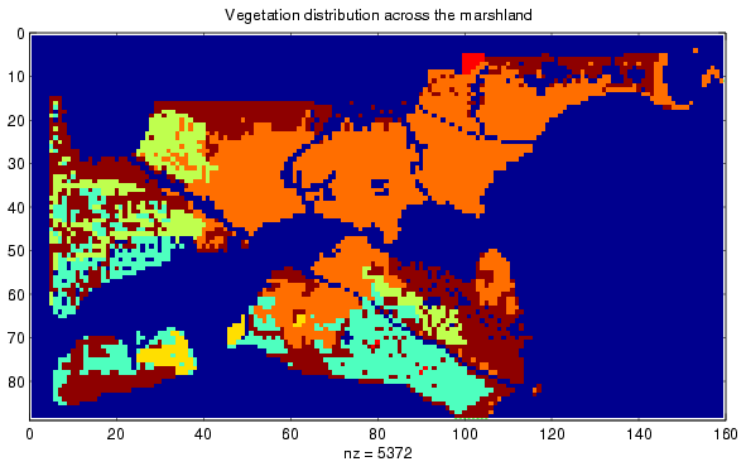
Slopes of local linear regression model at each station



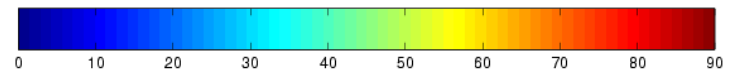
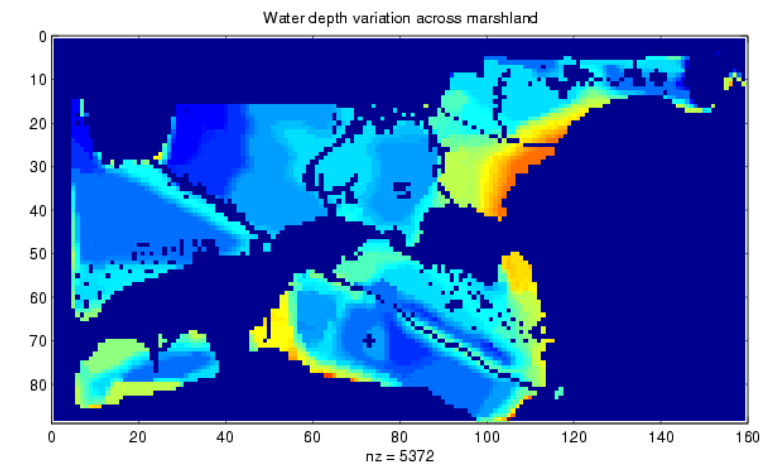
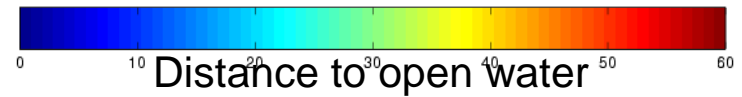
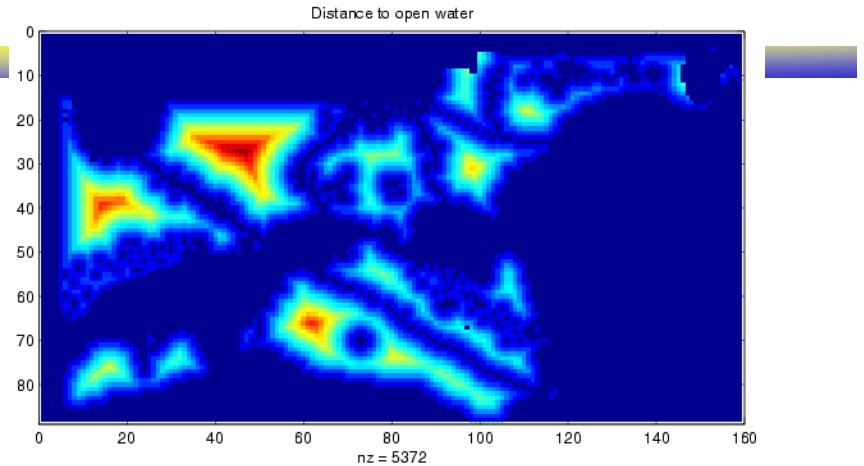
Prediction: Spatial Auto-correlation



Nest locations



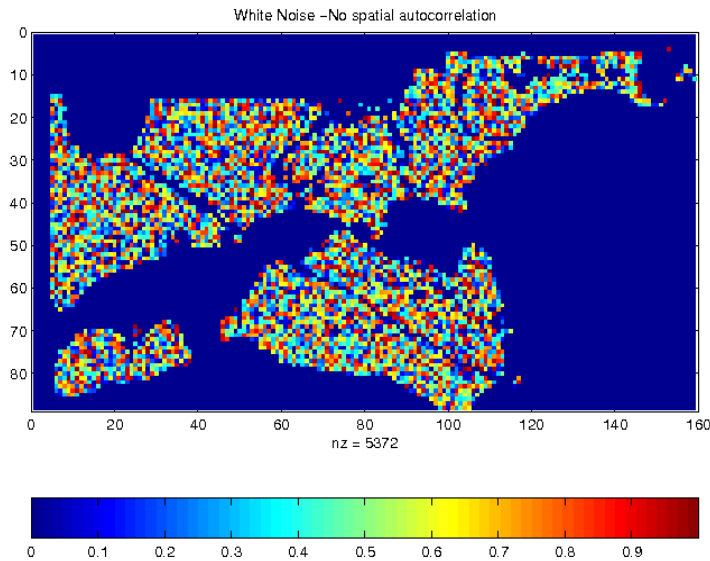
Vegetation durability



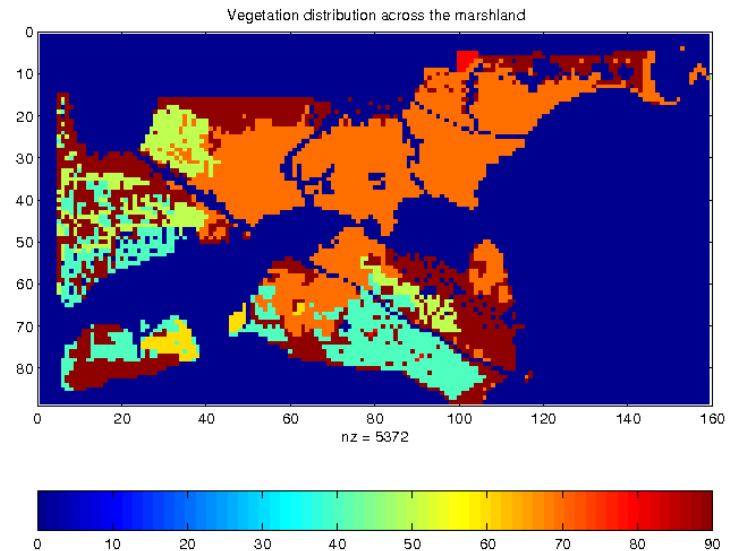
Water depth

Spatial Autocorrelation (SA)

- First Law of Geography
 - “All things are related, but nearby things are more related than distant things. [Tobler, 1970]”



Pixel property with **independent identical Distribution (i.i.d)**



Vegetation Durability with SA

- Autocorrelation
 - Traditional i.i.d. assumption is not valid
 - Measures: K-function, Moran's I, Variogram, ...

Parameter Estimation for Spatial Auto-regression

ρ : the spatial auto - regression (auto - correlation) parameter

\mathbf{W} : n - by - n neighborhood matrix over spatial framework

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






- **Maximum Likelihood Estimation**

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

- Current Limitation: Neighbor relationship (\mathbf{W}) is End-users' burden !
- SBD Opportunity 1: Post-Markov Assumption
 - SBD may be large enough to **provide reliable estimate of \mathbf{W}**
 - One may not have assume
 - Limited interaction length, e.g. Markov assumption
 - Spatially invariant neighbor relationships, e.g., 8-neighbor
 - Tele-connections are derived from short-distance relationships

Association Patterns

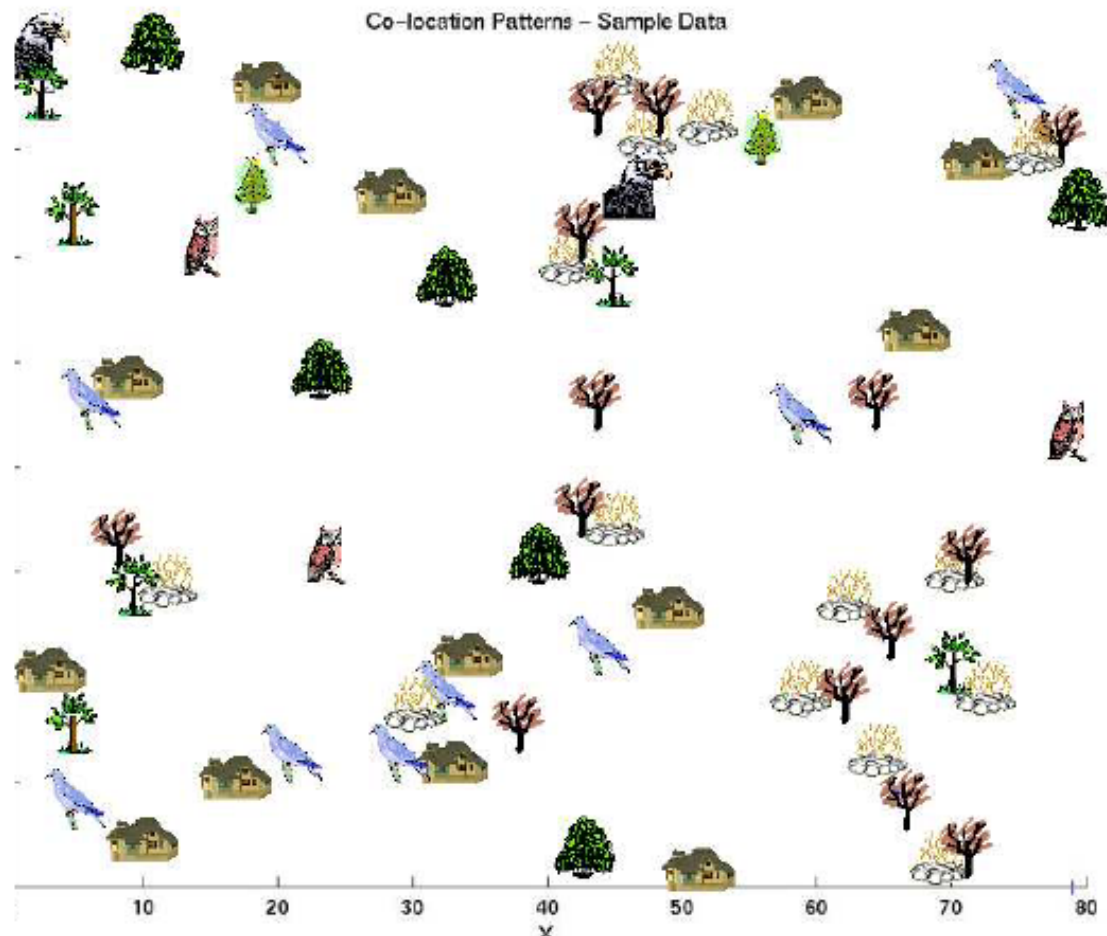
- Association rule e.g. (Diaper in T => Beer in T)

Transaction	Items Bought
1	{socks,  , milk,  , beef, egg, ...}
2	{pillow,  , toothbrush, ice-cream, muffin, ...}
3	{  ,  , pacifier, formula, blanket, ...}
...	...
n	{battery, juice, beef, egg, chicken, ...}

- Support: probability (Diaper and Beer in T) = 2/5
- Confidence: probability (Beer in T | Diaper in T) = 2/2
- Algorithm Apriori [Agarwal, Srikant, VLDB94]
 - Support based pruning using monotonicity
- Note: **Transaction is a core concept!**

Spatial Associations: Co-locations, Co-occurrence

- Given: A collection of different types of spatial events
- Find: Co-located subsets of event types
- Challenge:
No Transactions
- New Approaches
 - **Spatial Join** Based
 - One join per candidate is Computationally expensive!



Answers:



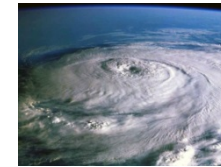
and



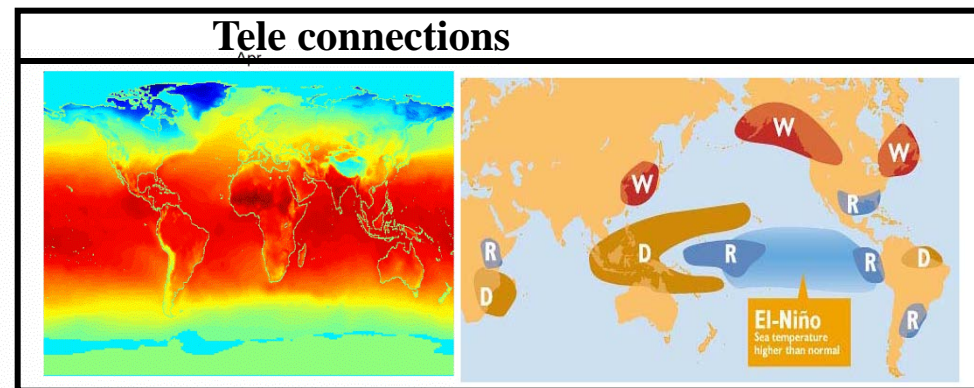
Association, Colocation, Co-occurrence

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

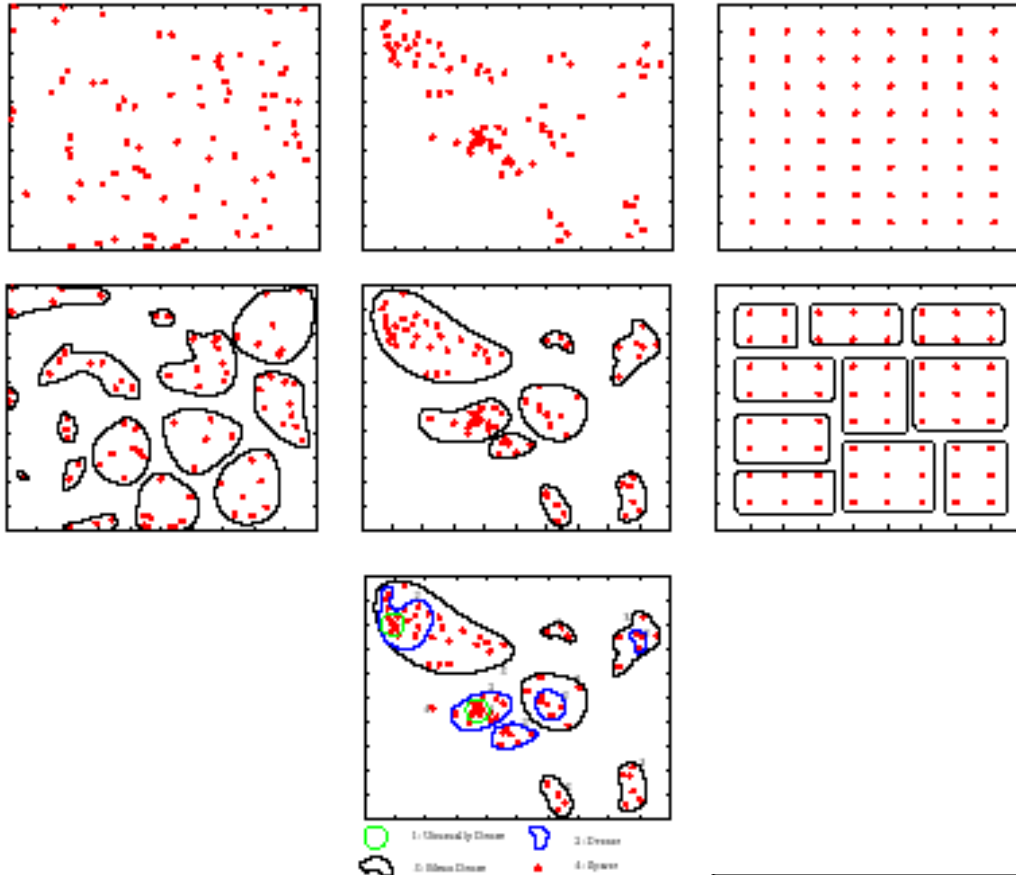


- **Future: Spatial Big Data Issues**
 - **Eclipsed Patterns**
 - Beyond Tobler's law: **Tele-connections**
e.g., impact of El Nino of a city
 - Spatio-temporal co-occurrence



SBD Summarization: Hotspots & Declusters

- Clustering: Find groups of tuples
- Statistical Significance
 - Complete spatial randomness, cluster, and de-cluster

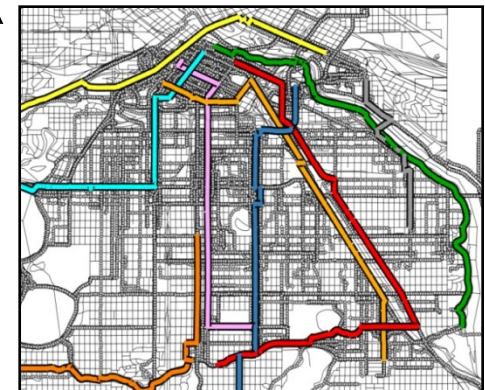
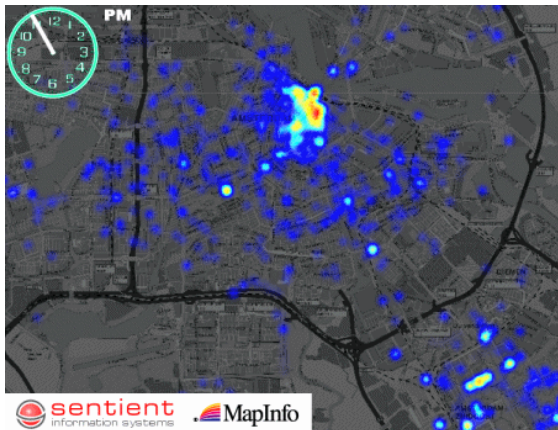
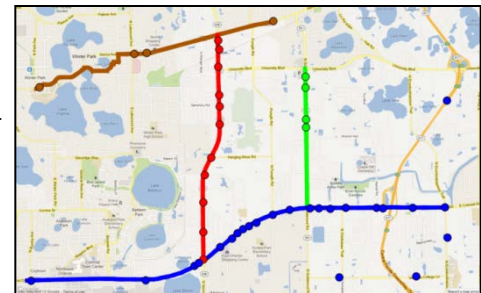
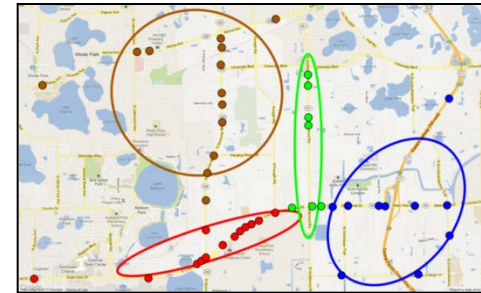
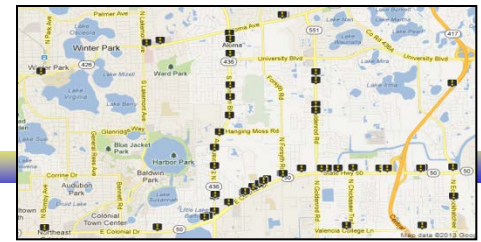


Inputs:
Complete Spatial Random (CSR),
Cluster,
Decluster

Classical Clustering
(K-means always finds clusters)

New Summarization Challenges

- Example: Unusually 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

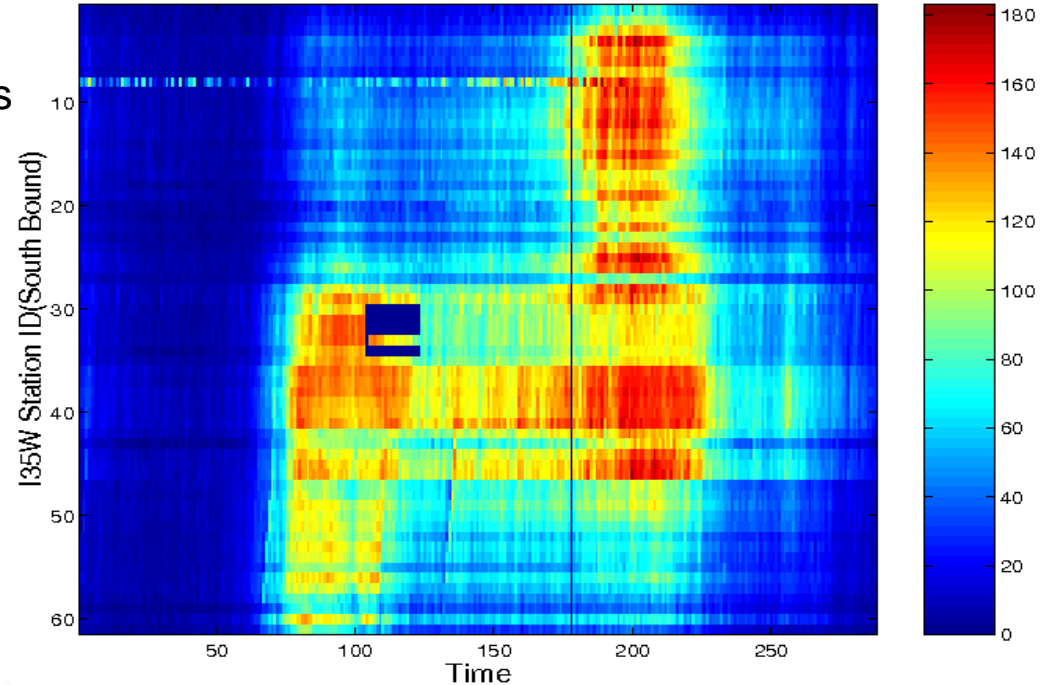


Spatial Outliers

- Spatial Outliers
 - Locations different than neighbors
 - Ex. Sensor 9
 - Source: Traffic Data Cities
- Spatial Join Based Tests



Average Traffic Volume(Time v.s. Station)



Anomalies & Big Data

- Geo-anomalies

- Location different from neighbors
- Anomalous trajectories
- Flow discontinuities



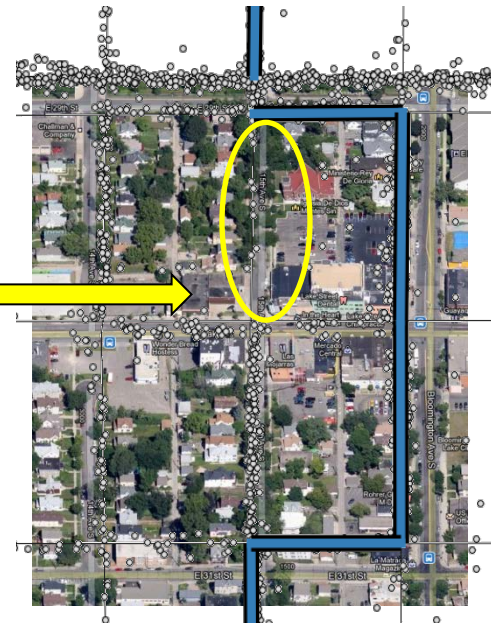
- Big Data Opportunities

- Exceptions to patterns of life



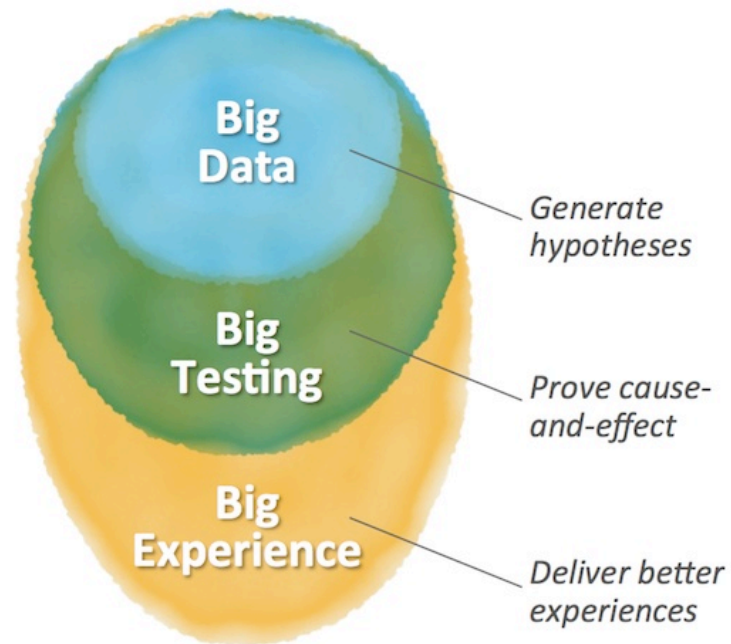
Why is a cycling route not straight?

Why is a block avoided by cyclists?



Outline

- Motivation
- Definition & Examples
- Infrastructure
- Analytics
- **Science**



Preparing Science for Big-Data

Nature, 7209(4), September 4, 2008

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

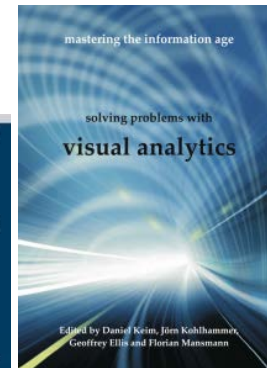
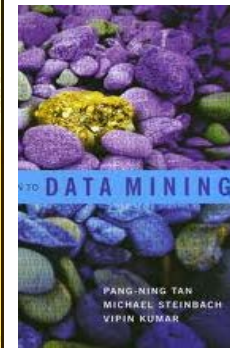
Sudden influxes of data have transformed researchers' understanding of nature before — even back in the days when 'computer' was still a job description.

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 Simulations 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 estimation, generalized additive models Frequentist: frequent generation, Model ens Exploratory Data An visual analytics, geographic information science, <u>spatial data mining</u>, ...



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.

How Did the Obama Campaign Leverage Big Data This Election?



Data Consolidation

Merging multiple databases to facilitate data sharing between separate teams

facilitated:

Data Collection Through Ongoing Experimentation

Including:

Data that provided universal insights about volunteers, donors and voters

facilitated:

Data that provided insights about key demographics in crucial regions

facilitated:

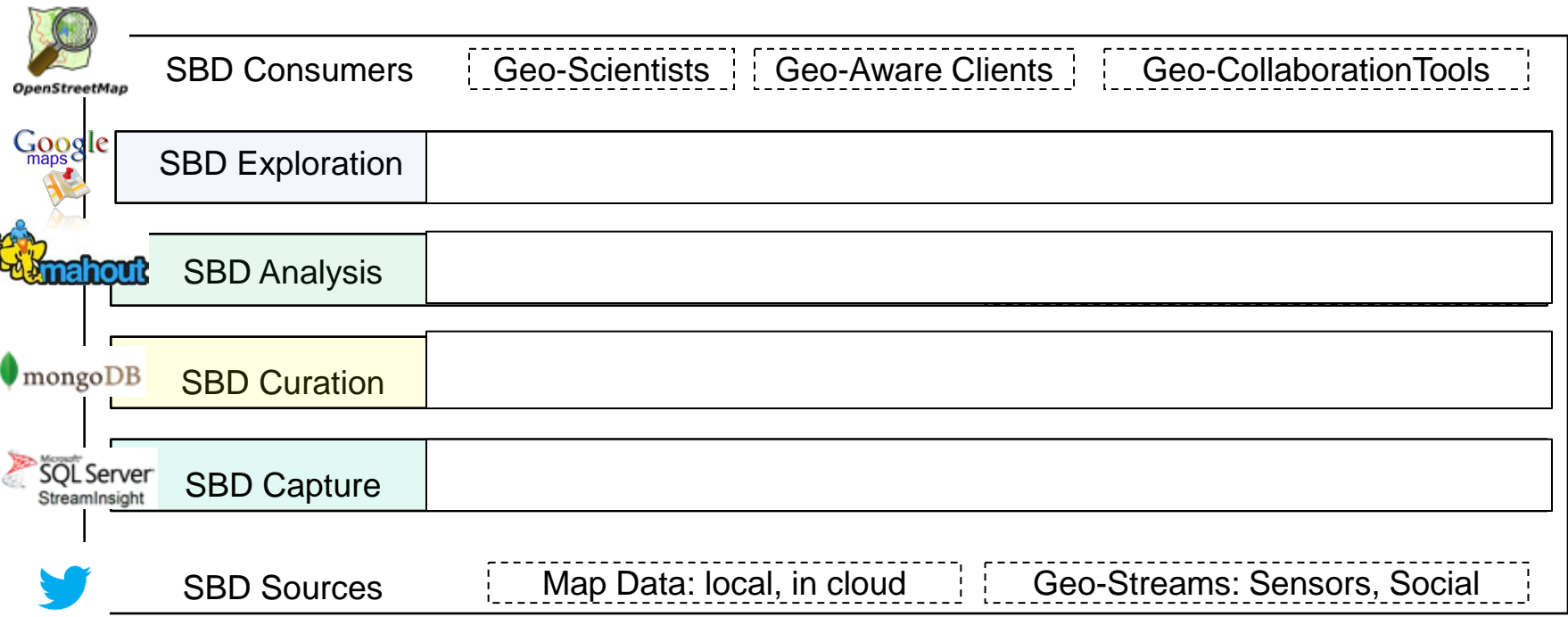
Optimizing large-scale campaign activities

e.g. sending more emails from Michelle Obama

"Getting small" by targeting specific demographics

e.g. reaching Miami-Dade women under age 35 by buying ad slots during TV series most favored by this demographic

Tools to Support Life-Cycle of Big Data in Science



Science Debates

- Quality Dimensions
 - Data Quality: Timeliness vs. Bias
 - Science Quality: Transparency, reproducibility of science with company dataset
 - Methodology: fourth paradigm for hypothesis generation
- Citizen Science via Crowd-sourcing
 - Potential: Ground Truth, e.g., climate observation
 - How do we recruit and motivate citizens?
 - Challenges: Data quality, IRB, privacy, ...
- Small data vs. Big Data
 - Small: 0.01 percent sample (randomized)
 - Big: 60% of population (big data)
 - Patterns: Mean, Outliers, Sub-populations, ...

Science Debates: What big data can not do?

- Big Data is not a panacea!
- Modelers' Hippocratic Oath (wilmott.com, 1/8/2009)
 - I will remember that I didn't make the world, and it doesn't satisfy my equations.
 - Though I will use models boldly to estimate value, I will not be overly impressed by mathematics.
 - I will never sacrifice reality for elegance without explaining why I have done so.
 - Nor will I give the people who use my model false comfort about its accuracy. Instead, I will make explicit its assumptions and oversights.
 - I understand that my work may have enormous effects on society and the economy, many of them beyond my comprehension.

WHO SAYS IRRATIONAL
EXUBERANCE IS
A ONE-TIME THING?



Outline

- Motivation
- Definition & Examples
- Infrastructure
- Analytics
- Science
- **Wrapup**

Spatial Big Data - Summary

- SBD are important to society
- Definition
 - Exceed capacity of current computing systems
 - Value beyond small data, e.g., representative samples
- Platform
 - Map-reduce speeds many applications
 - Debates– Iterative computations
- Data Analytics
 - Mahout = common techniques on map reduce
 - Debate: Bigger the data, simpler the model
 - Debate: Place based ensembles, Estimate Neighbor
- Science
 - Address data paucity in social sciences,
 - Debate: timeliness vs. bias, fourth paradigm

