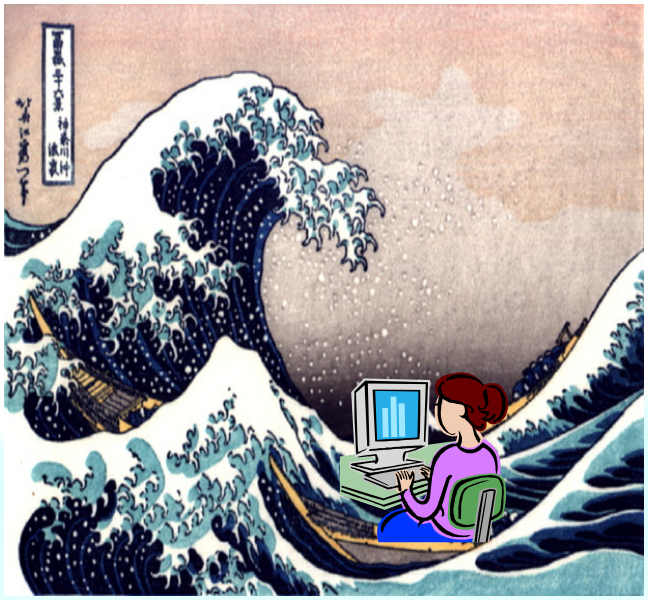
The data avalanche is here

Harvey J. Miller
Department of Geography
University of Utah
harvey.miller@geog.utah.edu

JRS 50th Anniversary Symposium, New York, NY, 23-24 April 2009

Introduction

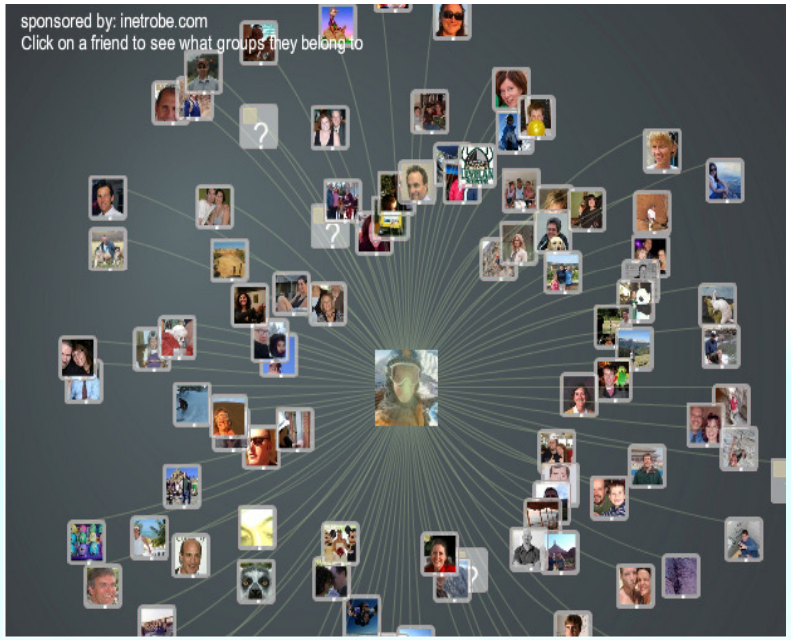
- From a data-poor environment…
  - Measurements were difficult, expensive and cumbersome
  - Science designed to tease scare information from limited observation



- … to a data-rich environment
  - Data costs have collapsed
    - Collection, storage and processing
  - We are flooded with **non-scientific** but **useful data**

Introduction

- Computational social sciences Lazer at el. (2009) *Science*
  - Based on collecting and analyzing massive databases on individual and group behavior
  - Emerging! - but at Yahoo, Google, Facebook and U.S. National Security agencies
  - Little activity in the mainstream social sciences
  - **Why?**



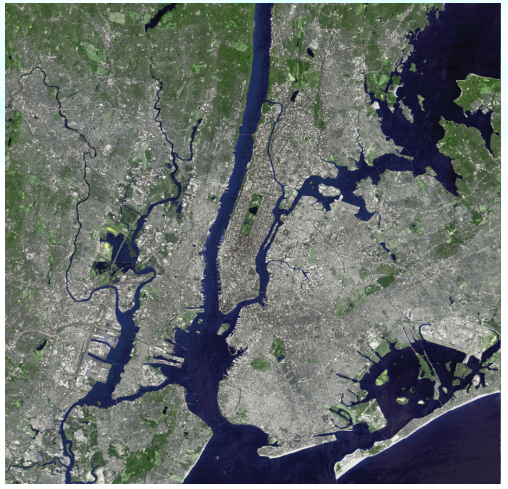
Facebook – 200 million social networks!

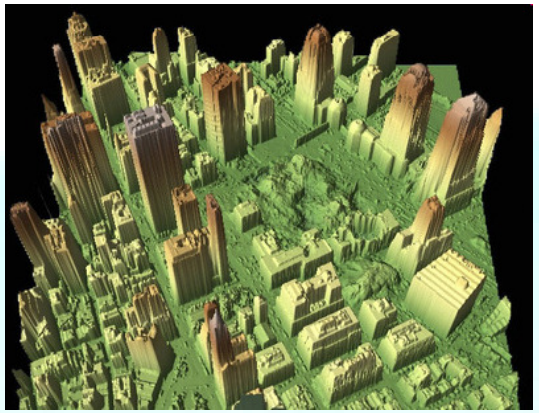
Introduction

- Regional science and Knowledge Discovery from Databases
  - What does KDD offer?
    - Individual level data on people, objects and bits
    - Powerful techniques for spatial and temporal exploration
    - New techniques for hypothesis generation (not testing)
  - What does regional science offer?
    - A rich body of theory that can guide the KDD process
- KDD is not atheoretical or anti-theoretic
  - There is a theory underlying KDD
  - KDD harmonizes with knowledge construction in (regional) science

New data sources

- Point of sale data
- Location-aware technologies
- Geosensor networks
- The Internet
- Simulation





blogs.discovery.com

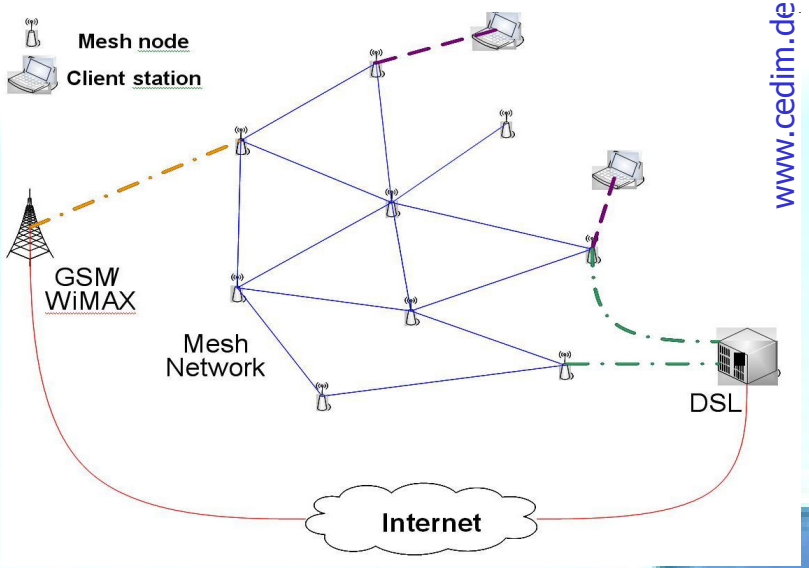
And lots of imagery too!

New data sources

- Location-aware technologies
  - Methods
    - Global Positioning System
    - Radiolocation
    - RFID chips
  - Location-based services (LBS)

lifehacker.com

- Geosensor networks
  - Connected heterogeneous data collection devices
  - Monitor environments from rooms to regions



www.cedim.de

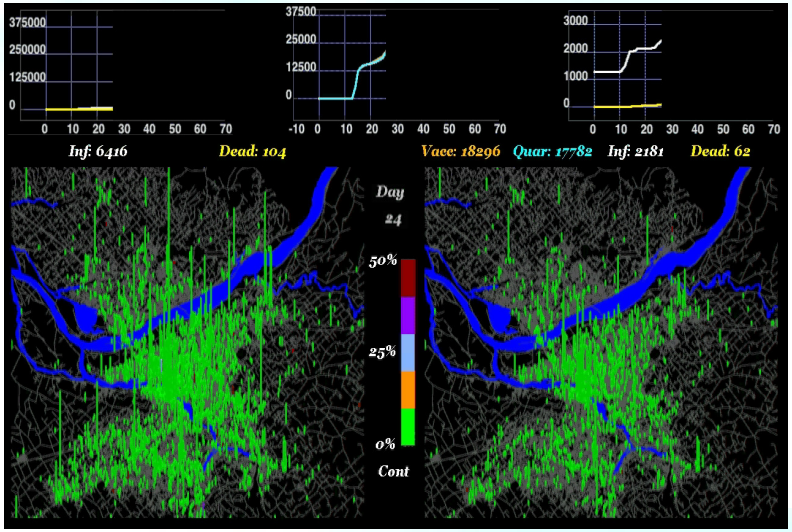
- GIS
  - Mobile object databases
  - Space-time exploratory tools

New data sources

- The Internet
  - What people are saying, buying, searching
  - Social connections
  - Techniques for natural language processing, analyzing multimedia
  - Example: Where's George?
    - Locations and times of registered bills
    - Surrogate for human travel
      - Brockman, Hufnagel and Geisel (2006) *Nature*

New data sources

- Geosimulation
  - Cellular automata
  - Agent-based modeling
- High-resolution space-time data
  - Empirical and/or synthetic
  - Geo-space and virtual space
  - Rethink theory and analysis of human behavior



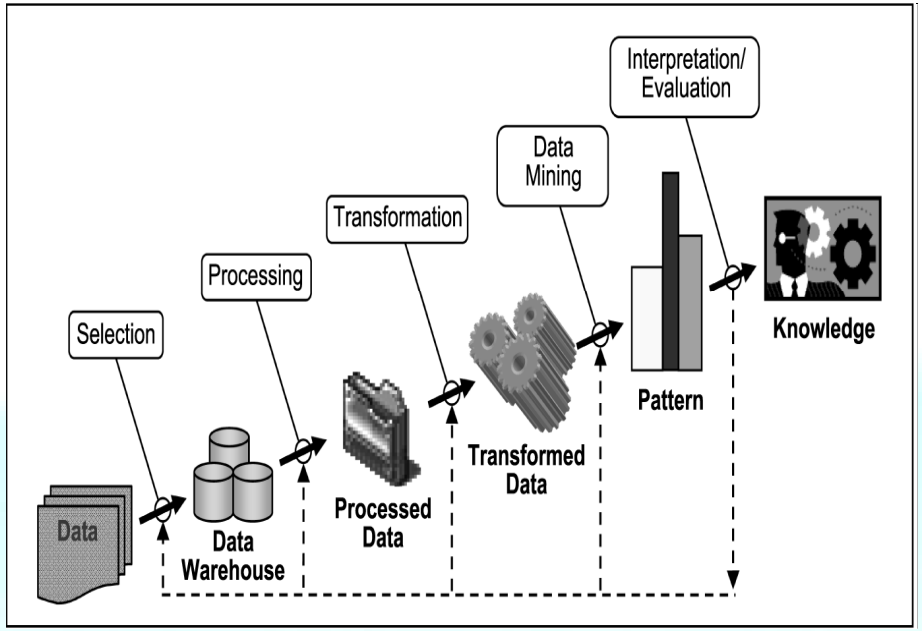
EpiSims: Individual-level simulation of disease propagation based on contacts in space and time

episims.lanl.gov

What is knowledge discovery from databases?

- KDD process
  - Complex
  - Human-centered
  - Cannot be automated
- Data mining
  - Low-level algorithms
    - Classification
    - Associations
    - Trends
    - Outliers
  - Pattern detection not model building

www.emeraldinsight.com

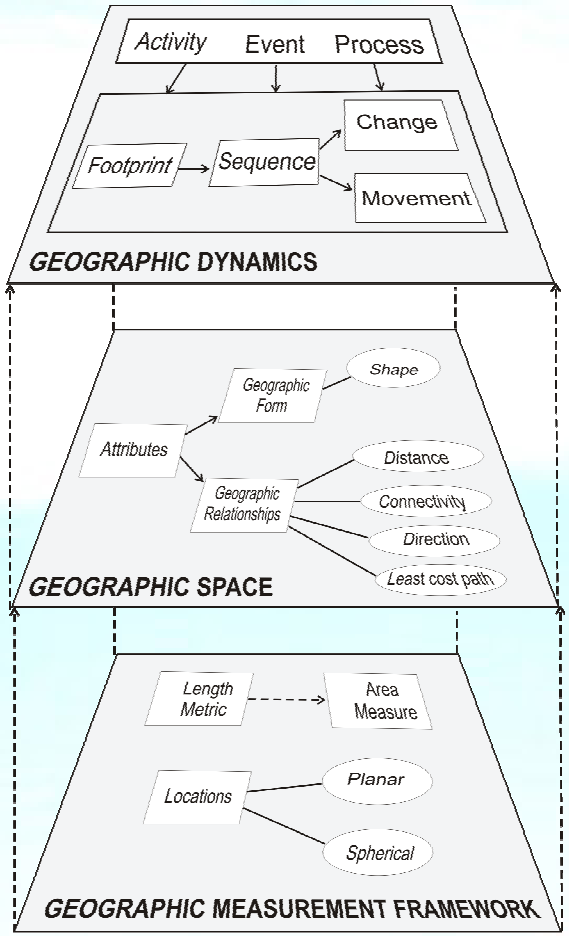


The KDD process

Why geographic knowledge discovery?

- Geo-space
  - Locations and proximity
    - Can be very general!
  - Spatial dependency and heterogeneity
- Spatial objects
  - Cannot be reduced to points
  - Size, shape, boundaries all matter
- Time
  - Change in spatial properties
  - Motion

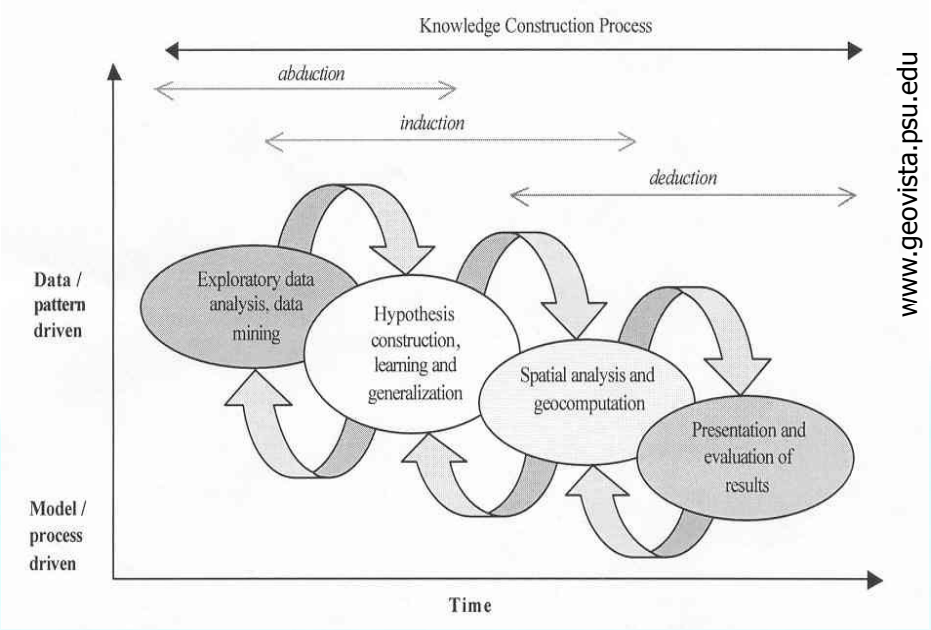
- GKD exploits the spatio-temporal properties of objects



Yuan (2009) based on Miller and Wentz (2003)

Knowledge discovery in regional science

- KDD to support theory
  - Hypothesis generation system
  - Abductive reasoning (C.S. Peirce - pragmatism)
  - Similar to a telescope - a datascope
- Theory to support KDD
  - Patterns & relations - potentially large!
  - Theory as a guide
    - Background knowledge
    - Pattern evaluation

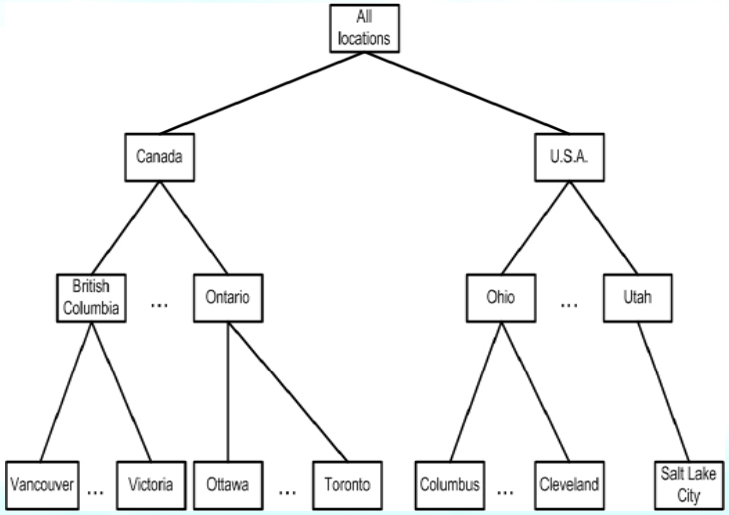


Knowledge discovery in the knowledge construction process

Mark Gahegan

Opportunities and challenges

- Background knowledge
  - Domain knowledge to guide data mining
    - Facts, experts, theory
- Regional science concepts
  - Abstract, vague, multi-level
    - Ontology?
  - Knowledge representation
    - RS: Implicit
      - Equations, algorithms, etc
    - KDD: Explicit
      - Networks, hierarchies, rules



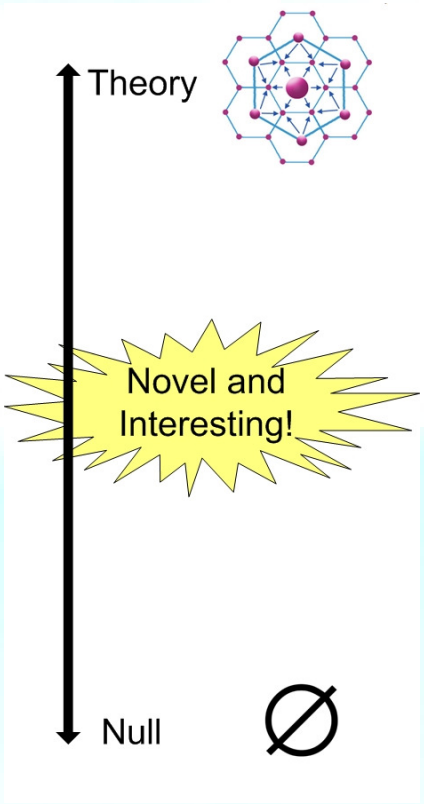
Concept hierarchy for "location"

- based on Han and Kamber (2003)

Opportunities and challenges

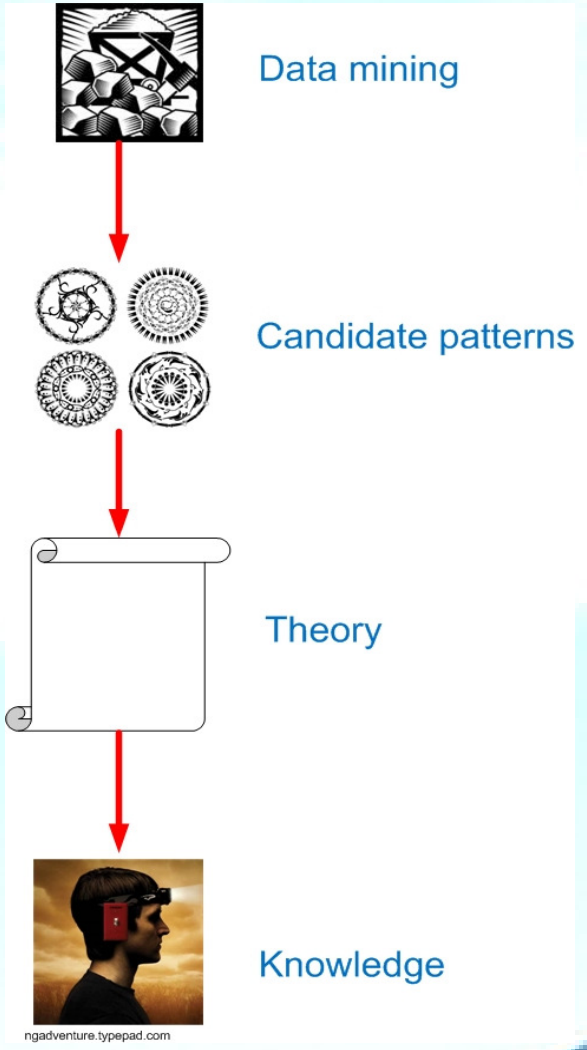
- Spatial pattern evaluation
  - Reality = theory: Interesting but not novel
  - Reality = null: Not interesting or novel
  - Between theory and null: *May* be interesting and novel

  - Problems
    - What is a good spatial null?
      - Not Complete Spatial Randomness (CSR)
    - What is the metric?
      - How do we measure spatial departures from theory and null?



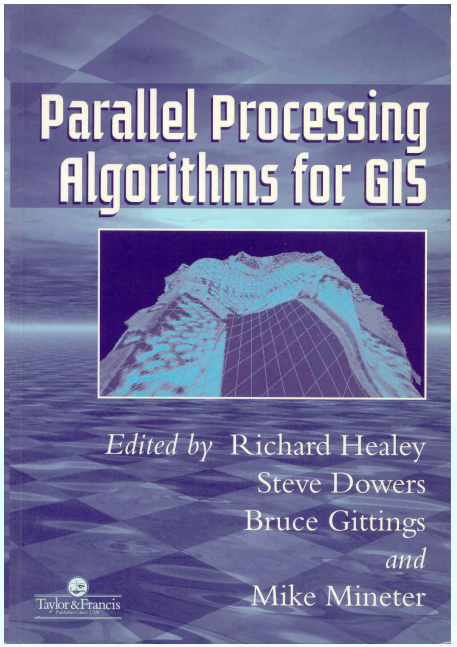
Opportunities and challenges

- Regional science as a pattern filter
  - Data mining often generates a large number of spatial and temporal patterns and relationships
    - Because there are so many possibilities!
- Meta-mining (Roddick 1999)
  - Mining the results of previous spatio-temporal mining exercises
  - Derive higher-level patterns and rules



Opportunities and challenges

- Algorithms and infrastructure
  - Geographic models can be computationally complex
    - Pairwise calculations between all locations
  - Research needs
    - Heuristics
    - Parallel, distributed and cloud-computing
- Educational challenges
  - Computational regional science?
    - Orthogonal concepts
    - Large body of knowledge to master!

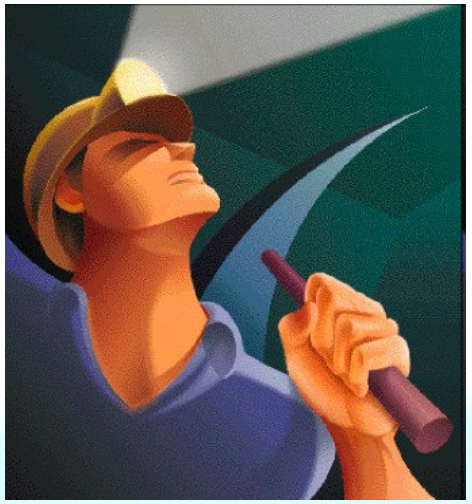


10 years old!

Conclusion

- The data avalanche is not forthcoming, **it is here**
- There are increasingly powerful tools for **discovering new knowledge** from individual level spatio-temporal data
- These technique do not replace theory, but are **complementary**
- Regional science has a **rich body of theory** that can **enhance** the discovery of new knowledge about cities, economies and societies

- **Let's start digging!**



www.liacs.nl/~edegraaf