[Spatial Data Mining for Understanding Climate Change]

Shashi Shekhar

McKnight Distinguished University Professor

Faculty of Computer Sc. and Eng., Univ. of Minnesota

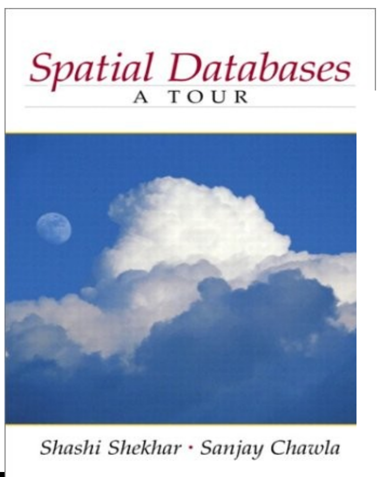
www.cs.umn.edu/~shekhar

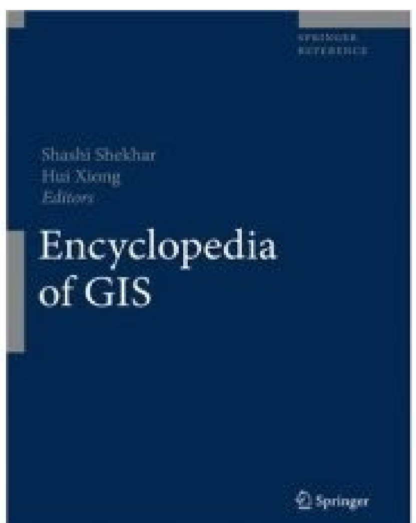
Presented at 2011 Workshop

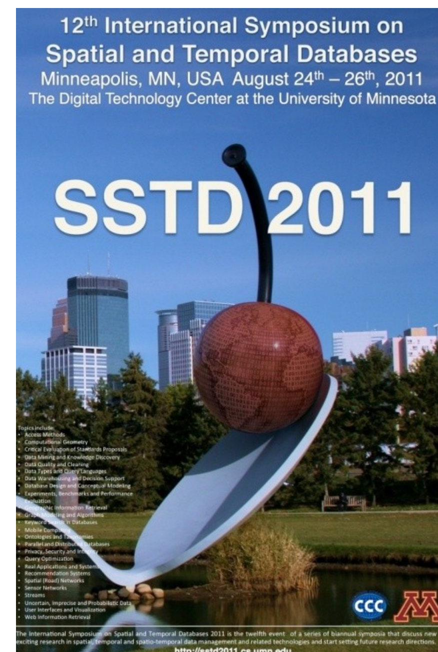
[Expedition on Understanding Climate Change: A Data Driven Approach]

August 15th-16th, 2011

Minneapolis, MN, USA







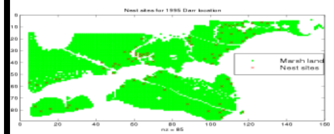
Outline

- My Background
  - Spatial Database Management Systems
  - Spatial Data Mining
- Science for Policy
- Establishing common vocabulary for interdisciplinary research
- Questions for Climate and Data Sciences
-

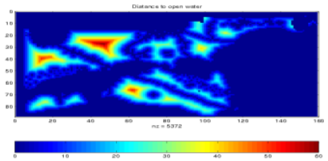
Spatial Data Mining : Representative Projects

Location prediction: nesting sites

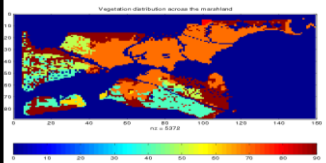
Nest locations



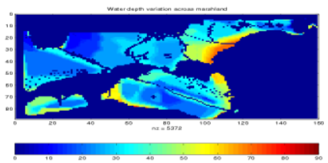
Distance to open water



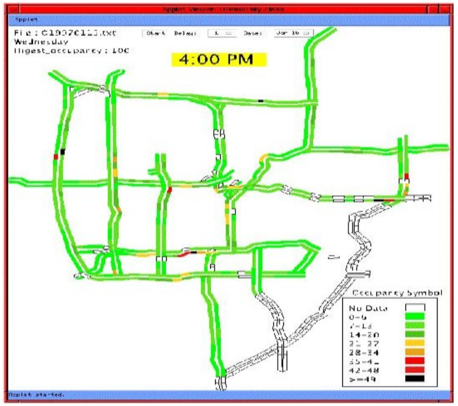
Vegetation durability

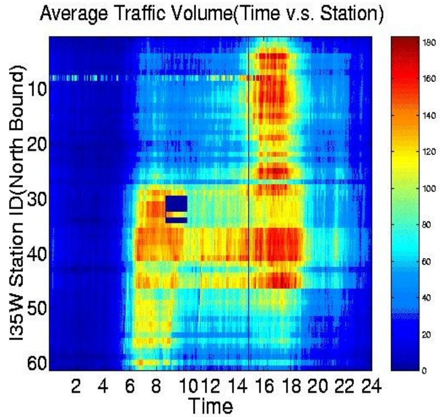


Water depth

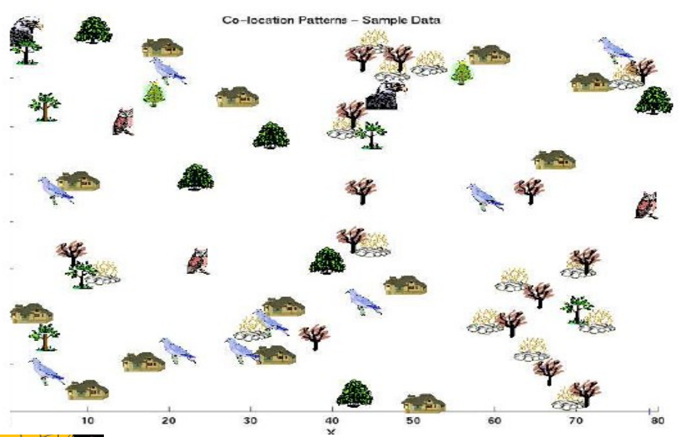


Spatial outliers: sensor (#9) on I-35

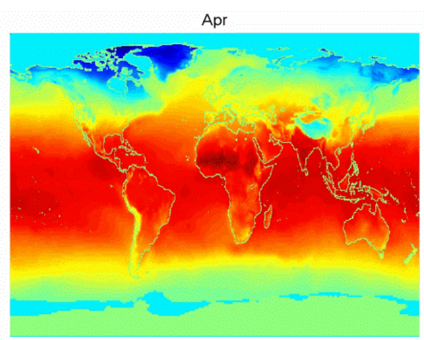


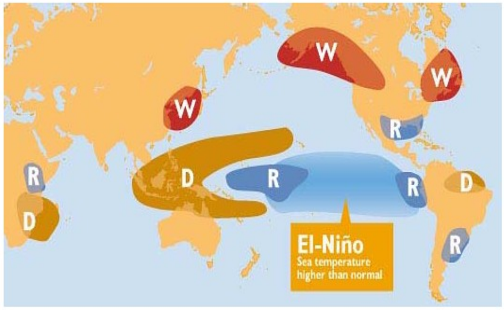


Co-location Patterns

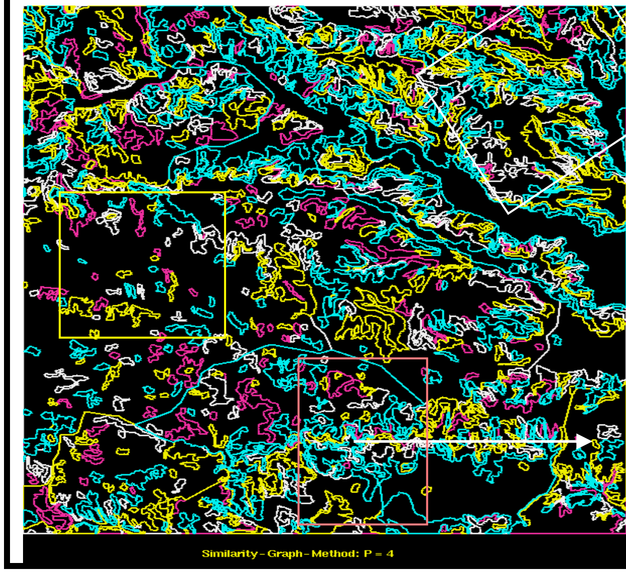


Tele connections





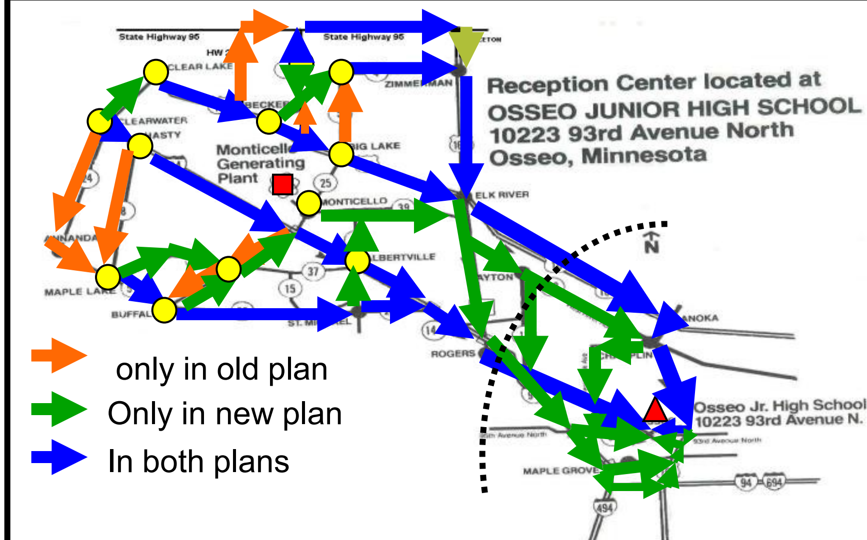
Spatial Databases: Representative Projects



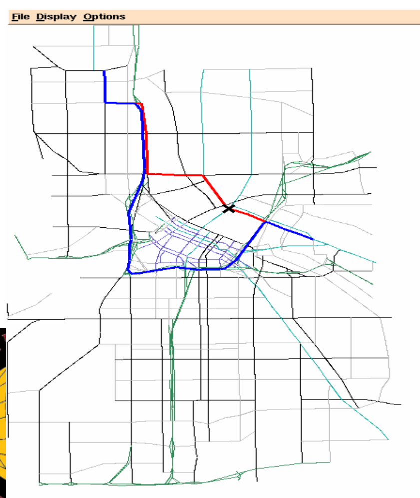
Parallelize Range Queries

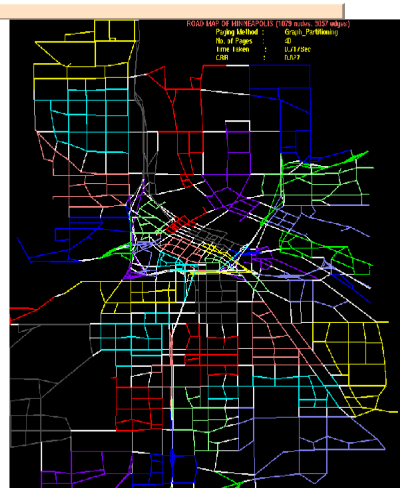
Evacuation Route Planning

Sustainable Transportation for Disasters



Shortest Paths **Storing graphs in disk blocks**





Outline

- My Background
- Science for Policy
  - Goals of Policy
  - Goals of Science
  - Science for Policy
- Establishing common vocabulary for interdisciplinary research
  - What is new in Data-intensive Science?
  - Spatial Thinking and Climate Science
- Questions for Climate and Data Sciences
-

SPATIAL DATABASE RESEARCH GROUP

Understanding Global Change: A Societal Perspective

- Climate Policy
  - Global changes is a serious concern!
  - Climate forecasts are needed by society.
  - It will be carried out whether or not science is ready!.

Decision-makers usually seek to affect how the world "ought to" or "should" be.

- Role of Science in policymaking
  - Inform policy

Science provides one source of input for making policy decisions that balance diverse considerations.

Source: Excerpted from CRS Report RL32992,

- **Science:** understand natural world
  - Subjective → Objective
  - Transparent, reproducible
  - Updating of data, which may revise hypothesis and theories

Role of Science in Policy

Box 2. Science: The Interaction with Policy

Scientific knowledge is dynamic, changing as new information becomes available. In this sense, science does not reveal "truth," so much as produce the best available or most likely explanation of natural phenomena, given the information available at the time; in many cases, analysis of data may give an estimate of the degree of confidence in the explanation. Moreover, scientific conclusions naturally depend on the questions that are asked.

The scientific method has, at its heart, two values that are strongly implied but not often stated: (1) a transparent approach in which both new and old data are available to all parties; and (2) a continuing effort to update data, and therefore modify, and even reject, previously accepted hypotheses in light of new information. Together, transparency and updating are the cleansing mechanisms that gradually sweep away scientific misunderstandings and errors—a sine qua non for scientific advancement.

Decision-makers usually seek to affect how the world "ought to" or "should" be. Science provides one source of input for making policy decisions that balance diverse considerations.

Source: Excerpted from CRS Report RL32992, *The Endangered Species Act and "Sound Science"*, by Eugene H. Buck, M. Lynne Corn, and Kristina Alexander,

Spatial Database Research Group

Congressional Research Service

Outline

- My Background

- Science for Policy

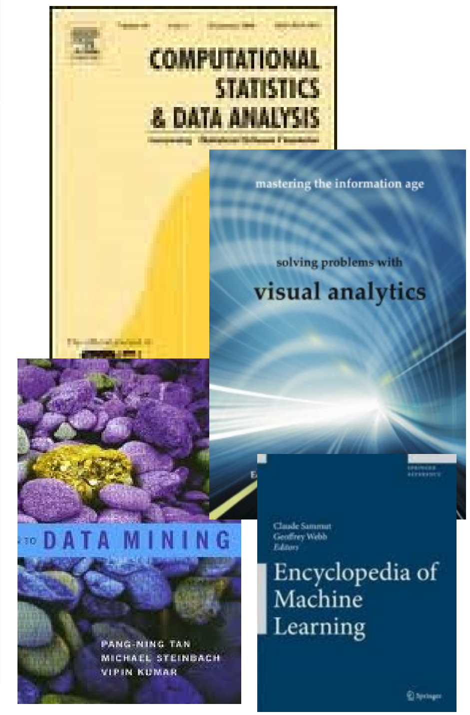
- Establishing common vocabulary for interdisciplinary research
  - What is new in Data-intensive Science?
  - Spatial Thinking and Climate Science

- Questions for Climate and Data Sciences
-

Understanding Global Change: Models in Science

- **Science:** understand natural world
  - Subjective → Objective, (transparent, reproducible)
  - Methods: Forward models, Backward models
- **Engineering:** Solve problems optimizing cost, efficiency, etc.

Models	Manual (Paper, Pencil, Slide-rules, log-tables, …)	Assisted by computers (HPCC, cyber-infrastructure, data-intensive, big-data)
Forward	Differential Equations (D.E.), Algebraic equations, …	Computational Simulations using D.E.s, Agent-based models, etc.
Backward	Parametric models, e.g. Regression, Correlations, sampling, Experiment design, Hypothesis testing, …	**Bayesian:** resampling, local regression, MCMC, kernel density estimation, neural networks, generalized additivemodels, … **Frequentist:** frequent patterns, hypothesis generation Model ensembles **Exploratory Data Analysis:** data visualization, visual analytics, geographic information science, **spatial data mining**, …



Big Data

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 Examples

- Google (late 1990s)
  - Web in your pocket, page-rank, Google earth, U-tube, …
  - Large data-centers with hundreds of thousands of computers

- IBM Watson (2011)
  - **Data:** encyclopedias, dictionaries, thesauri, newswire articles, literary works
  - **100+ models** to identify sources, <u>find</u>/ & generate hypotheses, <u>find</u> & score evidence, and merge & rank hypotheses

- Big data dimensions:
  - Variety – sensors (satellite, in-situ), social media, transactions, cell-phone gps-tracks
  - Velocity – streaming at high rate
  - Volume

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."

Science in the Petabyte Era –
Increasing Volume, Heightened Complexity, and Demands for Interoperability

Heightened Complexity, e.g. Climate

- GCMs already include 100s of phenomena & their interactions
    - Multi-phase Multi-Physics, Chemistry, Biology, Social sciences
    - Large number of constants, parameters, etc.
- What is the potential of complementing these with data-intensive paradigm?
    - Ensemble of data-intensive models to identify sources, find/ & generate hypotheses, find & score evidence, and merge & rank hypotheses

Preparing Science for Big-Data

Nature, 7209(4), September 4, 2008

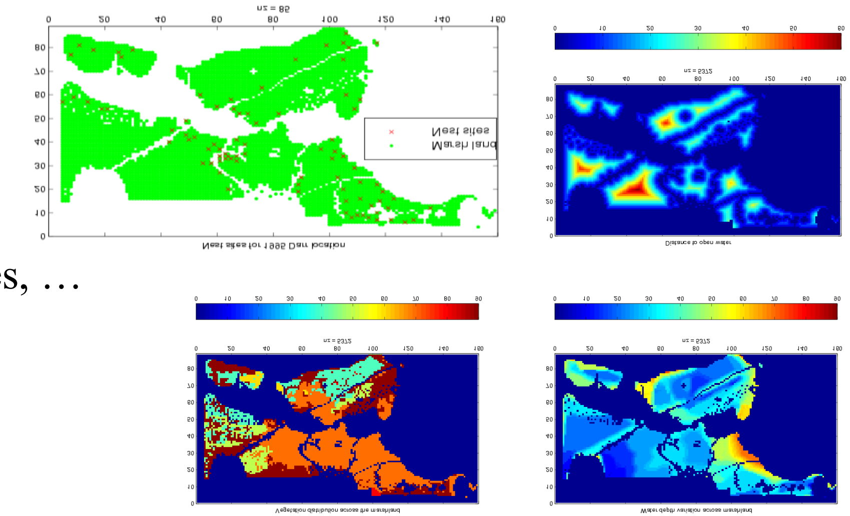
Big Data Translates into Big Opportunities... and Big Responsibilities

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

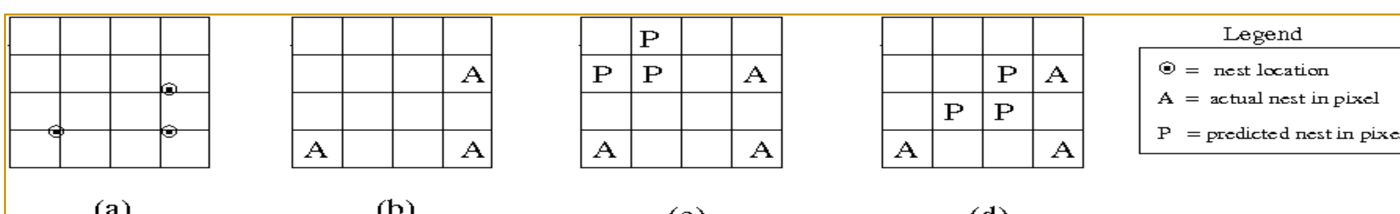
Pre-Electronic Models: An Example

- Location Prediction
  - Models to predict location, time, path, …
    - Nest sites, minerals, earthquakes, tornadoes, …
- Pre-electronic models, e.g. Regression
  - Assumed i.i.d
  - To simplify parameter estimation
  - Least squares – easy to hand-compute
- Alternatives
  - Spatial Autoregression,
  - Geographic Weighted (Local) Regression
  - Parameter estimation is compute-intensive!
- Next
  - Non-i.i.d errors: Distance based
  - Spatio-temporal vector fields (e.g. flows, motion)



$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

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



Example 2: Global vs. Local Regression

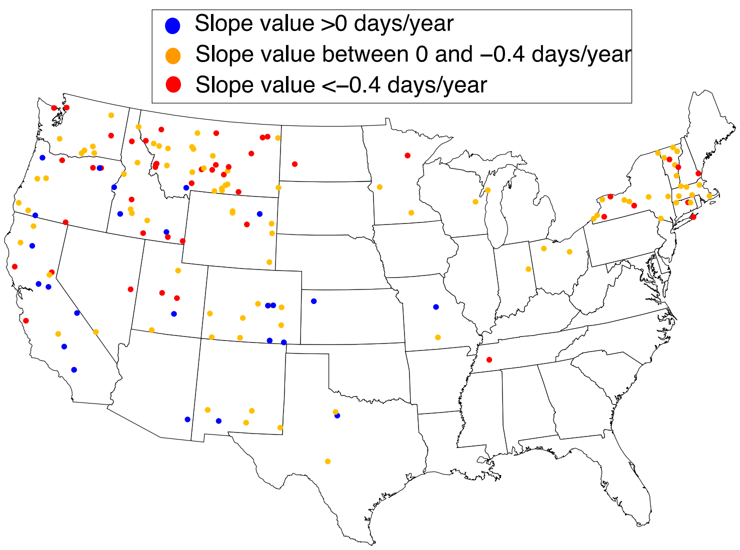
GODDARD SPACE FLIGHT CENTER

Global Change Master Directory
Discover Earth science data and services

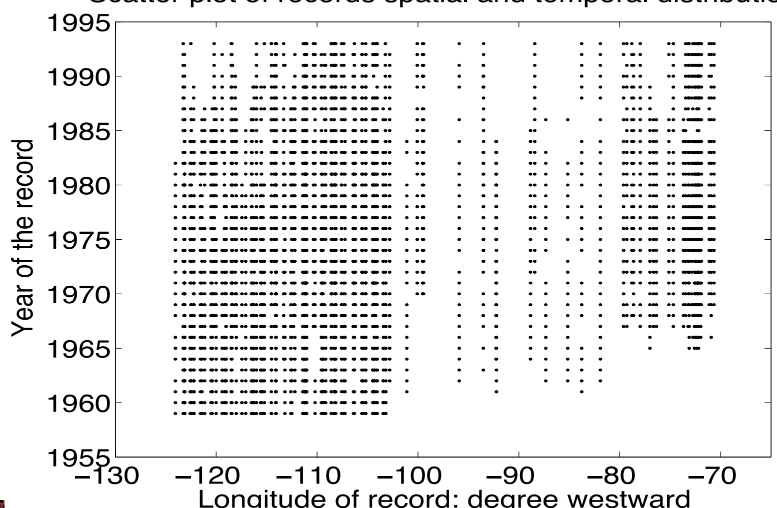
North American Lilac Phenology Data Since 1956 online

- Example: Lilac Phenology data
  - Yearly date of first leaf and first bloom
  - 1126 locations in US & Canada
- "Global" regression model shows a mystery
  - Postive 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

Slopes of local linear regression model at each station



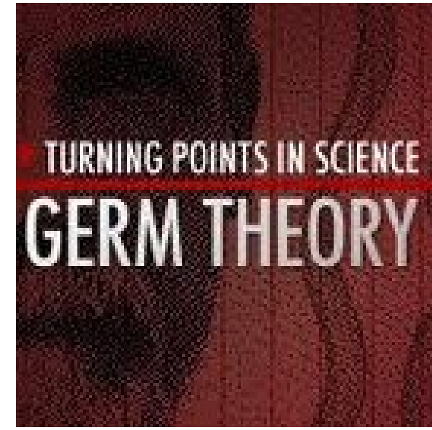
Scatter plot of records spatial and temporal distribution



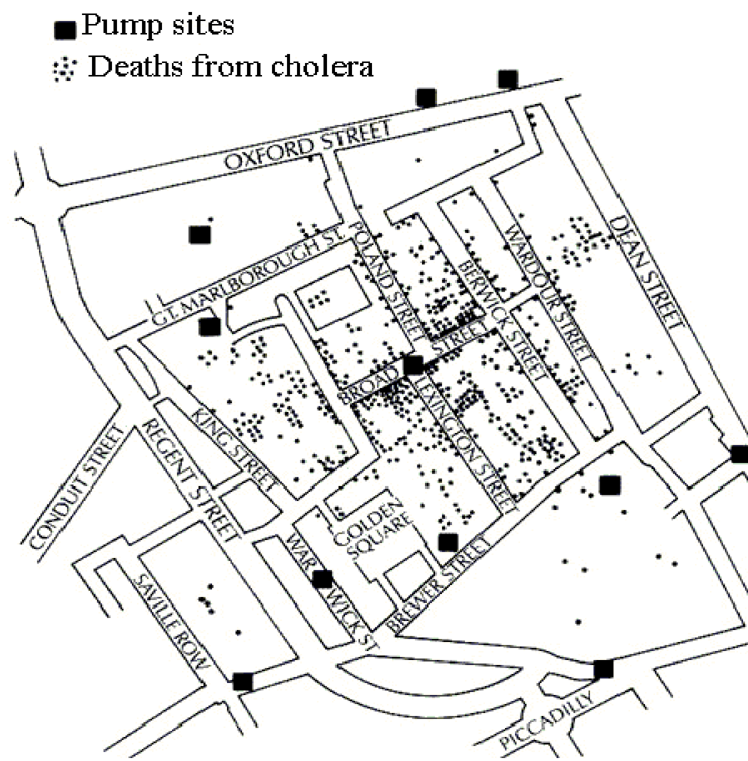
Outline

- My Background
- Science for Policy
- Establishing common vocabulary for interdisciplinary research
    - What is new in Data-intensive Science?
    - Spatial Thinking and Climate Science
- Questions for Climate and Data Sciences
-

Spatial Thinking

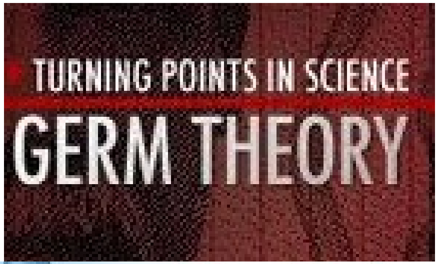


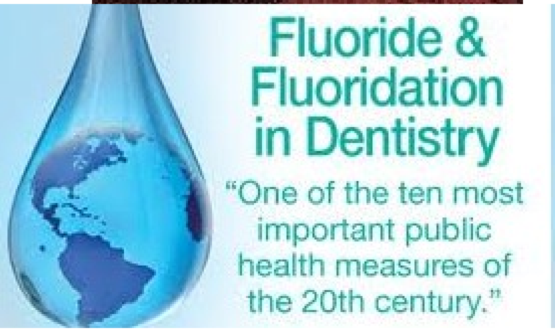
- What is it?
  - Identifying interesting, useful, non-trivial patterns
  - in large spatial or spatio-temporal datasets
    - e.g. satellite imagery, climate model output, …
    - gps-tracks, geo-sensor network, census, …
- Pattern Families
  - Hotspots
  - Spatial discontinuities
  - Co-locations, Tele-connections
  - Predicting location, trajectories, …



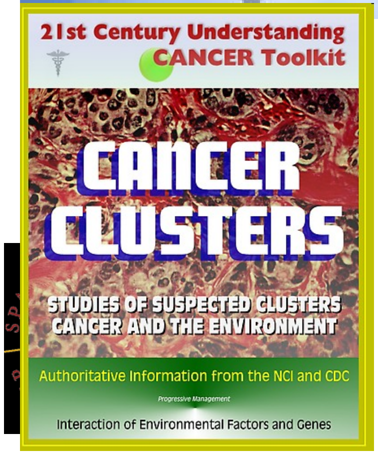
Spatial Thinking and Science: Historical Example

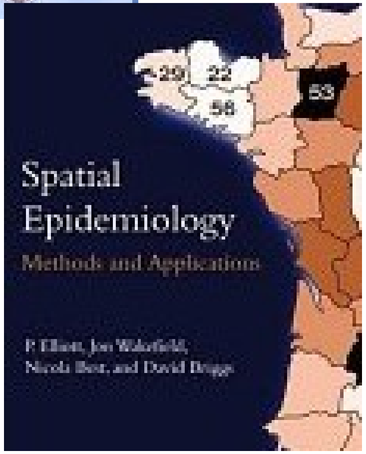
- Potential of discoveries and insights to improve human lives
  - Ex.: London Cholera Map by J. Snow → Water pump→Germ Theory
  - Ex. Colorado flourosis (1905) → water causation (1923) → Bauxite? Flouride? → 1% prevent carries (1930) → public policy (1948) …
  - Location bring in rich context to prune set of explanatory factor!
- Who regularly engages in SDM?
  - Public health: Where are cancer clusters? Environmental reasons?
  - Public safety: Where are hotspots of crime? Why?

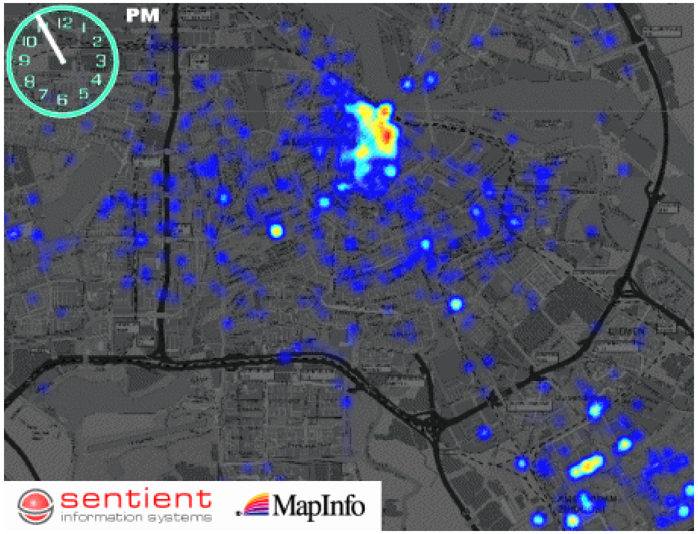


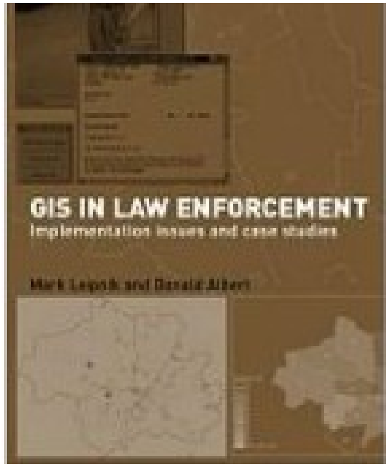


NIDCR National Institute of Dental and Craniofacial Research
National Institutes of Health
mproving the Nation's Oral Health



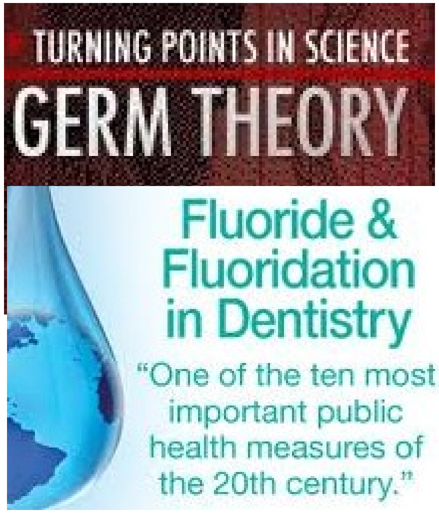


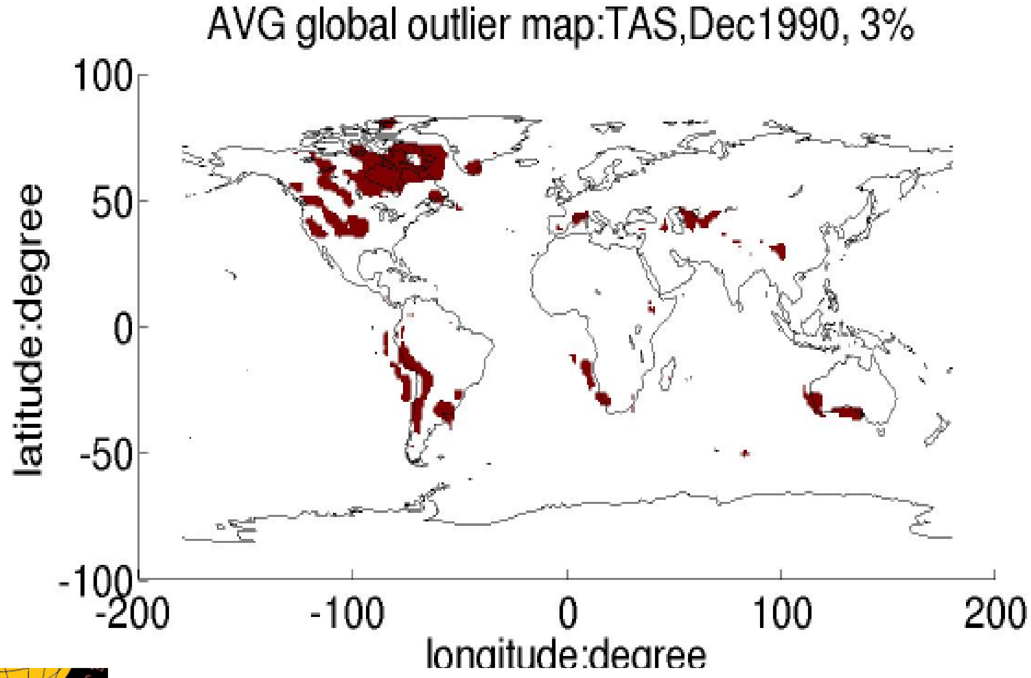




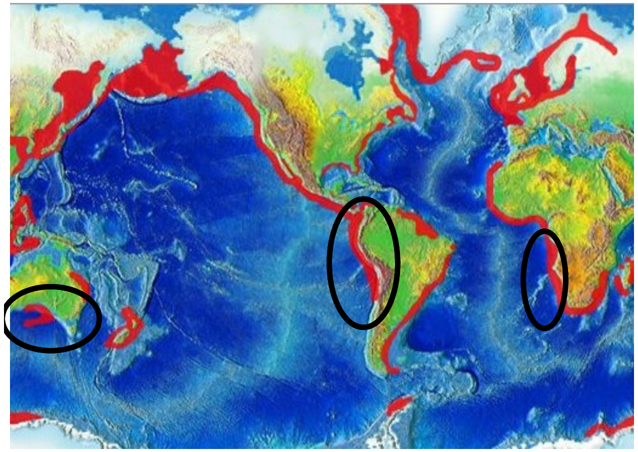
Data-Intensive Spatial Thinking: Potential for Climate Science

- Where are **hot-spots** of epistemic uncertainty?
  - Where climate models may agree, but all models are weak!
- What local phenomena, missed by GCM, may be responsible ?
- Narrow down options to refine GCMs
  - To address epistemic uncertainty
  - towards regional climate models





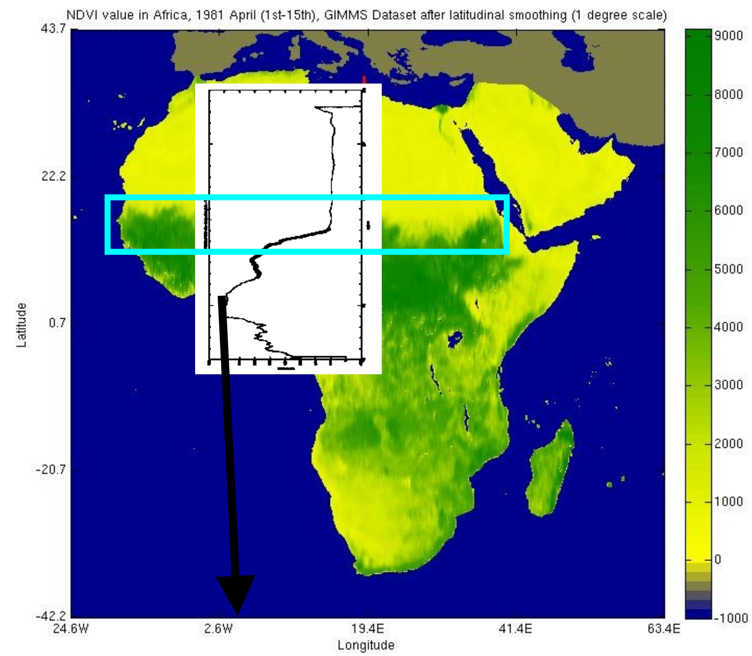
Upwelling areas map

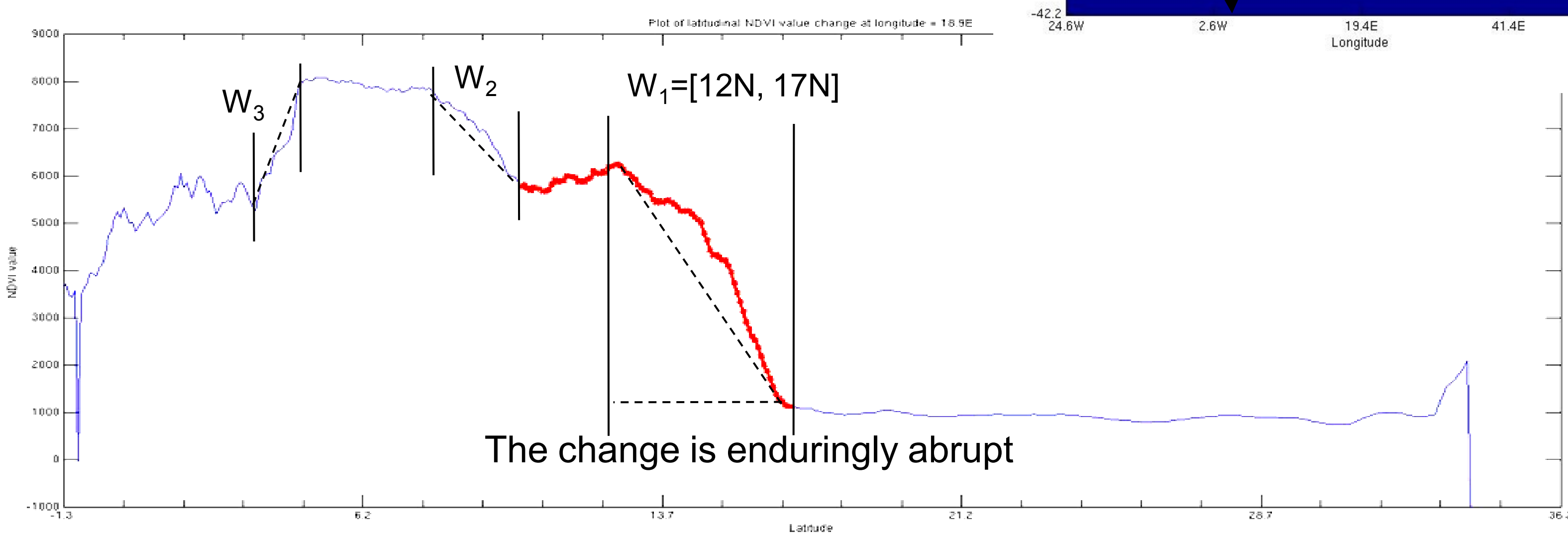


Example: Identifying Spatial Discontinuities, Sharp Changes

- Discontinuity
  - Coastline, Mountain ranges
- Sharp changes
  - Ecotones, e.g. Sahel

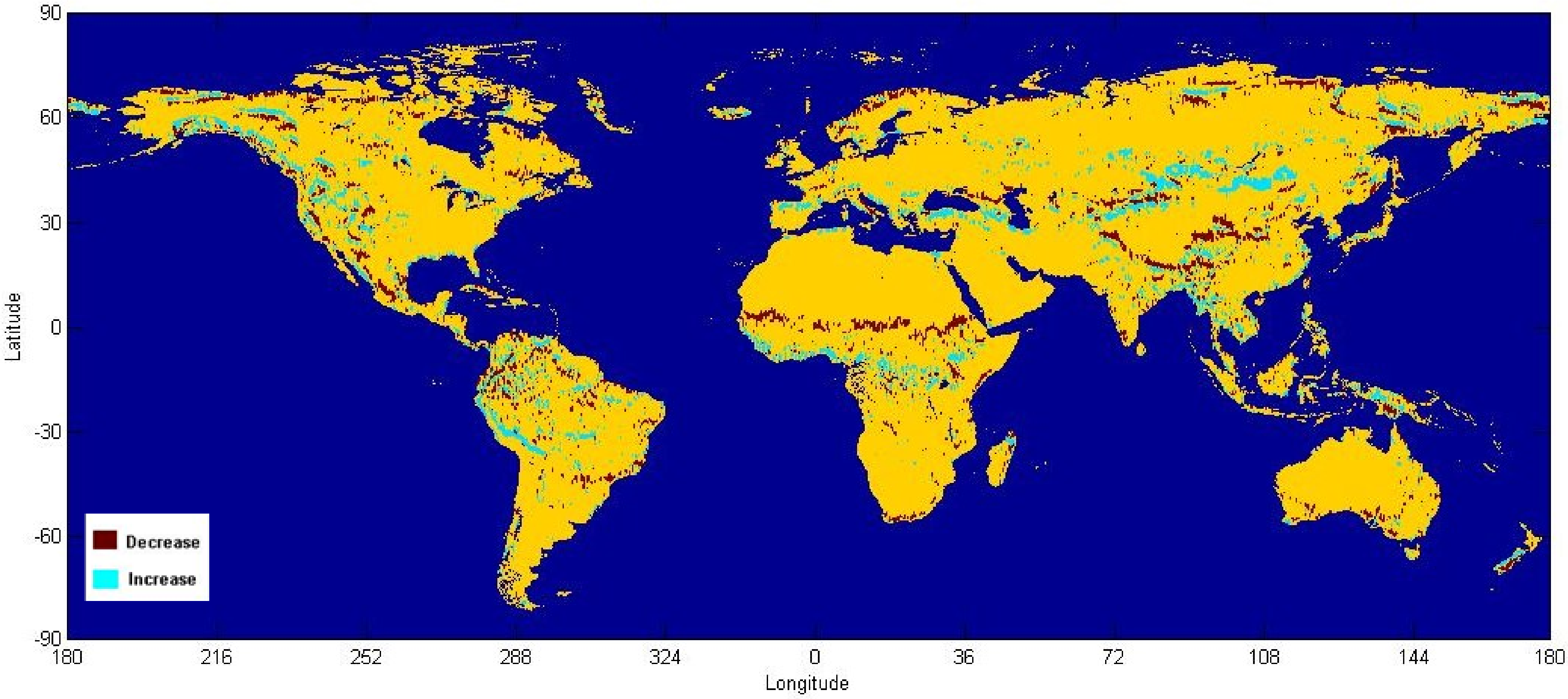
- **Climate Questions**
- Which other parts of world have sharp spatial changes?
- Are these more susceptible to global change?





Sharp Spatial Change: Entire world

- Which other parts of world have sharp spatial changes?
- Are these more susceptible to global change?
- Area of sharp spatial change
  - Blue = sudden increase (south to north)
  - Maroon = Sudden decrease (south to north)
  - Dataset used: NDVI, Aug 1-15 1981, 0.07 degree (8km) resolution



Outline

- My Background

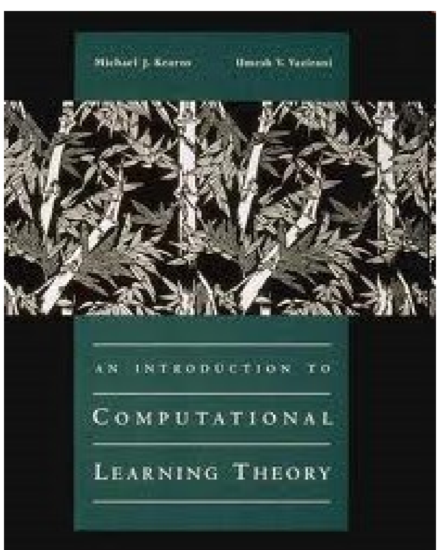
- Science for Policy

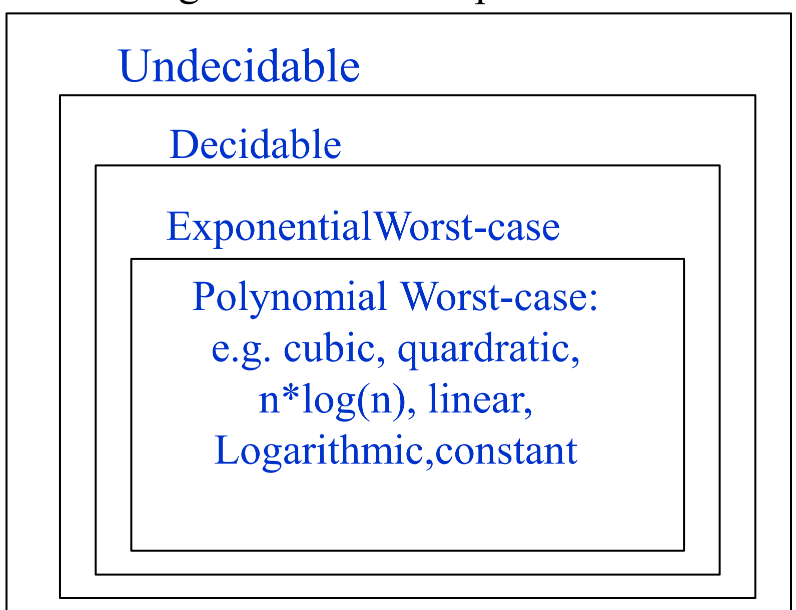
- Establishing common vocabulary for interdisciplinary research
    - What is new in Data-intensive Science?
    - Spatial Thinking and Climate Science

- Questions for Climate and Data Sciences

Questions

- Data-Intensive Climate Science
  - What would we like to analyze if old constraints on computing power were removed?
  - Model specification: Equation-based vs. data-table-based
- Computer Science: What class do climate problems fall inside?
  - What are the limits of data mining?
    - Free will, developing social consensus, randomness, chaos, computability, feature selection, confirmatory analysis for secondary datasets, …
  - Is Climate prediction undecidable ? (Dr. Kolli, WMO)
    - Could data-intensive model outperform humans?
  - Is climate projection decidable?
    - What are scalable algorithms for this problem?





Goal: Computational system outperform humans

Goal: CS efficiently solve common cases

Goal: CS efficiently solve all cases