

# Data Mining in Vehicular Sensor Networks: Technical and Marketing Challenges

---

Hillol Kargupta

Agnik & University of Maryland, Baltimore County

<http://www.cs.umbc.edu/~hillol>

<http://www.agnik.com>

# Roadmap

- Motivation
- Mining vehicular sensor network
- Building MineFleet
- Challenges
- Some Algorithmic Solutions
- Discussion

# Vehicles: Source of High Volume Data Streams



- Vehicles generate tons of data
- Hundreds of different parameters from different subsystems
- High throughput data streams
- So what?

# Why Mine Vehicle Data?

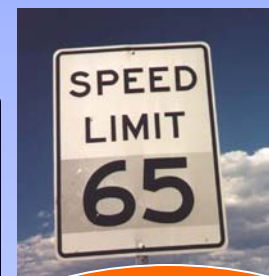


High gas prices

- Fuel consumption analysis
- Fleet analytics
- Vehicle benchmarking
- Predictive health-monitoring
- Driver behavior analytics



Breakdowns cost thousands of dollars



Bad driving costs money--- fuel, brake shoe, insurance, law-suits

# Fuel Subsystem: Sample Attributes

## Fuel Subsystem

- Air Fuel Ratio
- Fuel Level Sensor (%)
- Fuel System Status Bank 1 [Categ. Attrib.]
- Oxygen Sensor Bank 1 Sensor 1 [mV]
- Oxygen Sensor Bank 1 Sensor 2 [mV]
- Oxygen Sensor Bank 2 Sensor 1 [mV]
- Oxygen Sensor Bank 2 Sensor 2 [mV]
- Long Term Fuel Trim Bank 1 [%]
- Short Term Fuel Trim Bank 1[%]
- Idle Air Control Motor Position
- Injector Pulse Width #1 (msec)
- Manifold Absolute Pressure (Hg)

## Operating Condition

- Barometric Pressure
- Calculated Engine Load(%)
- Engine Coolant Temperature (°F)
- Engine Speed (RPM)
- Engine Torque
- Intake Air Temperature (IAT) (°F)
- Mass Air Flow Sensor 1(MAF) (lbs/min)
- Start Up Engine Coolant Temp. (°F)
- Start Up Intake Air Temperature (°F)
- Throttle Position Sensor (%)
- Throttle Position Sensor (degree)
- Vehicle Speed (Miles/Hour)
- Odometer (Miles)

# Product Concept: MineFleet

## Optimize Fuel Economy by

- Modeling fuel consumption behavior
- Identifying factors that are causing poor fuel economy
- Benchmarking fuel sub-system

## Predictive Health Monitoring

- - Automatically execute built-in library of tests for checking the health-status of the vehicle
- Predictive modeling of the vehicle sub-systems

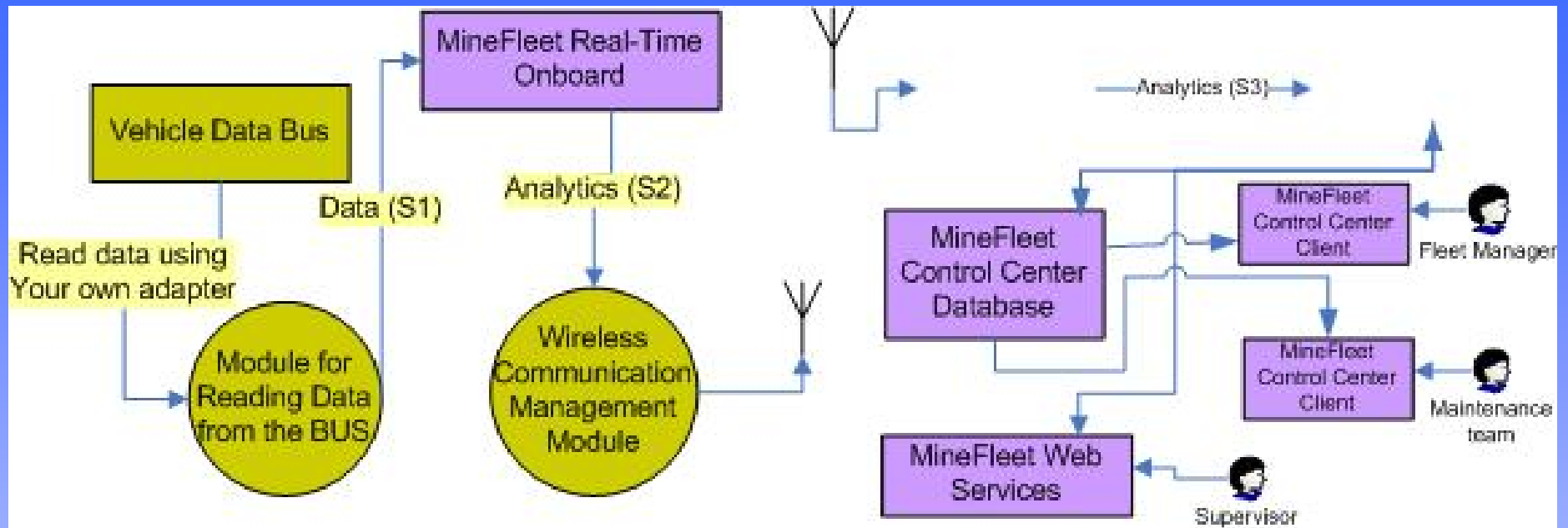
## Driver Behavior Monitoring

- Policy-based driver behavior monitoring
- Identify the effect of driver behavior on fuel economy
- 

## Minimize Wireless Communication

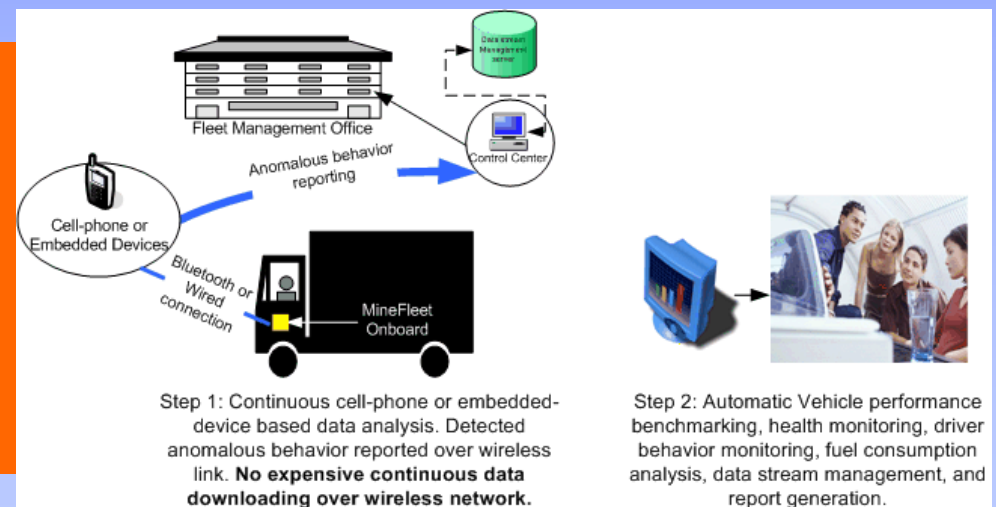
- Onboard data stream mining
- Send alerts and analytics only when problems occur
-

# MineFleet Architecture

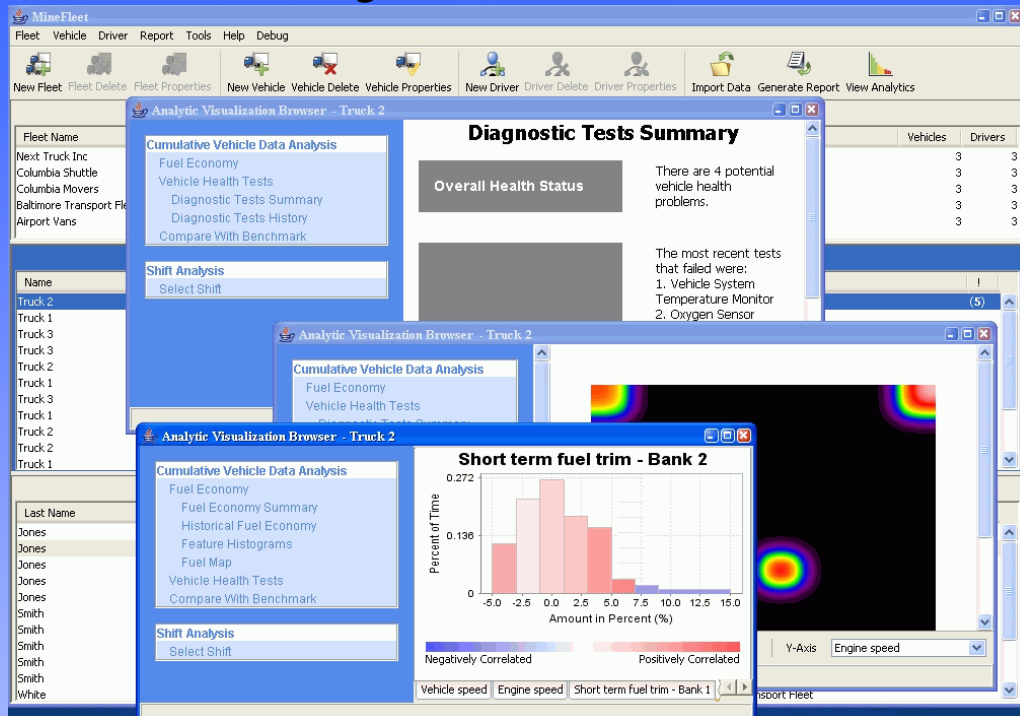


## Main Components

- ❑ Onboard Module
- ❑ MineFleet Control Center Server
- ❑ MineFleet Client Modules
- ❑ MineFleet Web Services



# MineFleet System



MineFleet Control Center



MineFleet Onboard



# Challenges: Accessing Data

- Vehicles generate data for hundreds of attributes
- But manufacturers provide open access to only about 20 of those that are needed for emission checks
- Off-the-shelf devices were designed for off-line monitoring by mechanics

# Onboard Computing Platform



Circa 2001



Circa 2005



Circa 2007



- First prototype -- PDA-based platform
- Other choices:
  - Cell phones and
  - Low-cost, less powerful embedded devices

- ┌ Market Entry Point
  - Location management companies
  - M2M companies
- ┌ Low Cost Embedded GPS Devices
- ┌ Resource constrained
- ┌ 3-4K run time memory
- ┌ 250K footprint
- ┌ Resource sharing with GPS program

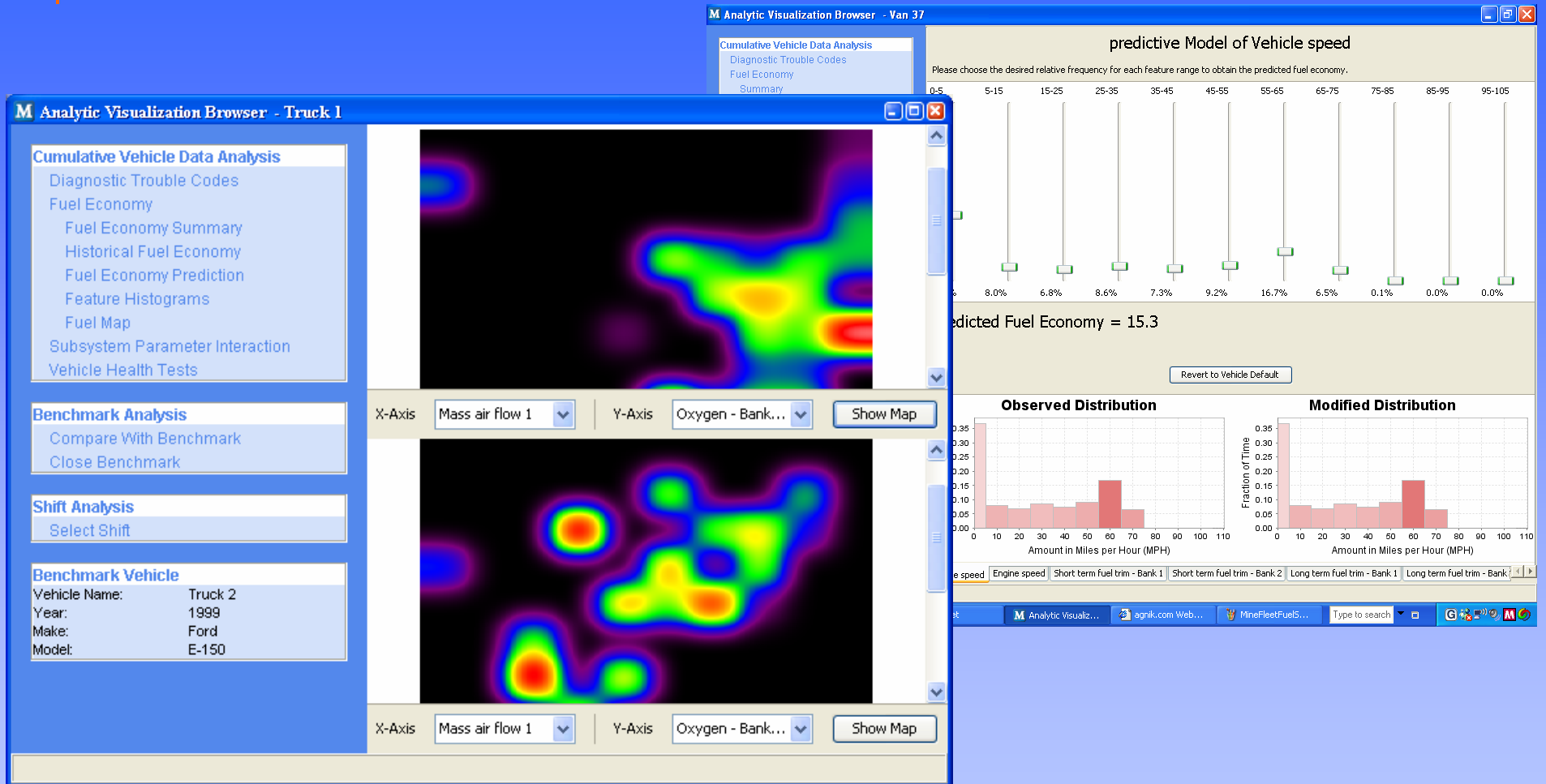
# Fuel Economy: Impact of Vehicle Condition and Driver Behavior

- Quantify the effect of vehicle condition on fuel consumption. Example:
  - ❑ Effect of air-intake subsystem behavior on fuel economy
  - ❑ Effect of fuel subsystem on fuel economy.
- Quantify the effect of driver behavior on fuel consumption.
  - ❑ Effect of speeding on fuel economy
  - ❑ Effect of acceleration on fuel economy
  - ❑ Effect of braking on fuel economy
  - ❑ Effect of idling on fuel economy
- Build predictive models of the fuel economy as a function of vehicle and driving parameters for optimizing the performance



**Poor vehicle components and bad driving reduces gas mileage**

# Fuel Heat Map & Predictive Modeling



Fuel heat maps show the vehicle operating points that offer high fuel economy. Red color represents high fuel economy and blue represents poor.

# Fuel Consumption Summary Panel & Savings Calculator

The screenshot displays a web application interface with a blue header and a white main content area. The interface is divided into several sections:

- Summary Panel (Right):** A vertical list of blue buttons on the left, each corresponding to a text box on the right. The buttons are: Average Fuel Economy, Ideal Speed for Best Fuel Economy, Ideal Acceleration for Best Fuel Economy, Ideal Engine Speed for Best Fuel Economy, Predicted Fuel Economy at Ideal Speed, and Effect of Idling. The text boxes provide detailed information for each category.
- Summary Text:**
  - Average Fuel Economy:** The average fuel economy for this vehicle from the recorded data is 15.6 miles per gallon.
  - Ideal Speed for Best Fuel Economy:** The best fuel economy for this vehicle was obtained at speeds between 55 and 65 Miles per Hour (MPH).
  - Ideal Acceleration for Best Fuel Economy:** The best fuel economy for this vehicle was obtained at an acceleration between 0 and 1 Feet per Second Squared (ft/sec<sup>2</sup>).
  - Ideal Engine Speed for Best Fuel Economy:** The best fuel economy for this vehicle was obtained at engine speeds between 1500 and 2000 Rotations per Minute (RPM).
  - Predicted Fuel Economy at Ideal Speed:** The predicted fuel economy for this vehicle driven at speeds between 55 and 65 Miles per Hour (MPH) is approximately 19.3 miles per gallon.
  - Effect of Idling:** This vehicle spends 36.9% of its time idling. Reducing the percentage of time spent idling by half will improve the fuel economy from approximately 15.3 to 15.9 miles per gallon.
- Summary Link:** A blue link labeled "Calculate Potential Savings" is located below the "Effect of Idling" text.
- Navigation Menu (Left):** A vertical list of blue buttons under the heading "Cumulative Vehicle Data Analysis": Diagnostic Trouble Codes, Fuel Economy Summary, Historical Fuel Economy, Fuel Economy Prediction, Feature Histograms, Fuel Map, Subsystem Parameter Interaction, and Vehicle Health Tests. Below this are "Benchmark Analysis" (Compare With Benchmark) and "Shift Analysis" (Select Shift).
- Fuel Savings Calculator (Bottom Left):** A separate window with a blue title bar. It contains a text area with instructions: "Enter the following information to estimate savings by changing the behavior of Feature Oxygen - Bank 2 - Sensor 2." Below this are input fields for "This vehicle is driven approximately 50 miles driven per per day" and "Fuel is estimated to cost approximately 2.89 per gallon". A table shows calculated savings: Savings per day: \$1.67, Savings per month: \$50.66, Savings per year: \$607.91. Buttons for "Compute" and "Close" are at the bottom.
- Taskbar (Bottom):** A Windows taskbar with a "start" button, several "MineFleet" icons, and an "Analytic Visualization ..." icon. The system tray on the right includes a search box, a volume icon, and a network icon.

# Predictive Vehicle Health Management

Detect problems using physics-based model and inductive techniques.

The screenshot displays a software interface for fleet-level analytics. The main window is titled "Fleet-Level Analytics - BWI Transporter". It features a sidebar menu with categories like "Vehicles Analytics", "Cumulative Vehicle Data Analysis", "Benchmark Analysis", and "Shift Analysis". The central pane shows a detailed report for a "Long Term Fuel Related Combustion Efficiency" test. The report includes a "Test Description" explaining the combustion process and the test's purpose, a "Test Failed" status, and a "Recommendation" section with specific advice on checking fuel pressure, injectors, and oxygen sensors. To the right, a "Vehicle Diagnostic Tests" list includes items like "Combustion Temperature Inequality Monitor" and "Air Intake Volume Inconsistency". Below this is a "Flagged Vehicles" table listing failed tests for various trucks.

Failed Diagnostic Test	Name
Long Term Fuel Related Combustion Efficiency	Truck 2
Long Term Fuel Related Combustion Efficiency	Truck 4
Long Term Fuel Related Combustion Efficiency	Truck 6
Long Term Fuel Related Combustion Efficiency	Truck 1
Long Term Fuel Related Combustion Efficiency	Truck 3
Long Term Fuel Related Combustion Efficiency	Truck 5
Air Intake Volume Inconsistency	Truck 2
Air Intake Volume Inconsistency	Truck 4
Air Intake Volume Inconsistency	Truck 6
Air Intake Volume Inconsistency	Truck 1
Air Intake Volume Inconsistency	Truck 3
Air Intake Volume Inconsistency	Truck 5

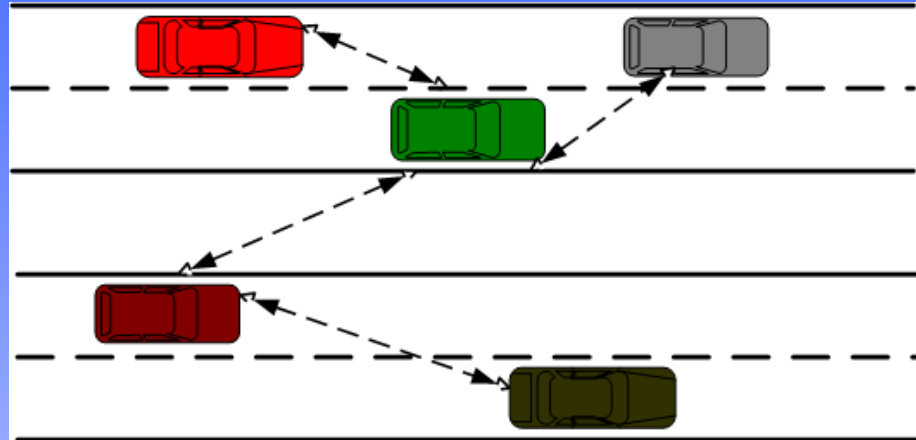
Detailed description of a specific test that the vehicle passed

Identifying all the vehicles in a fleet with a specific problem

# MineFleet VANET Project



Mobile information sources



A VANET

- Developing a mobile data stream management system for quick indexing and retrieval of information from the device onboard the vehicle.
- Distributed indexing and clustering techniques

# Algorithmic Challenges

- Ensemble-based Approach
- Exact vs. Approximate techniques
  
- Approximate monitoring of statistical properties
  - For example, Correlation Matrix
- Approximate sequence comparison
- Approximate modeling
  
- Similarity preservation, approximation and orthogonality
  - Fourier, Wavelet, Eigenvectors, Random vectors



# Correlation Matrix Computation & Monitoring

- Given data matrix  $X$
- Naïve computation: Compute  $X^T X$
- Compute in the frequency domain (take Fourier transformation)
- StatStream (Zue and Shasha, 2002)
  
- Our Approach Exploits
  - Divide and Conquer
  - Approximate orthogonality of random vectors
- Identify the region of the matrix that contain significantly changed coefficients

# Testing A Group of Correlation Coefficients Together

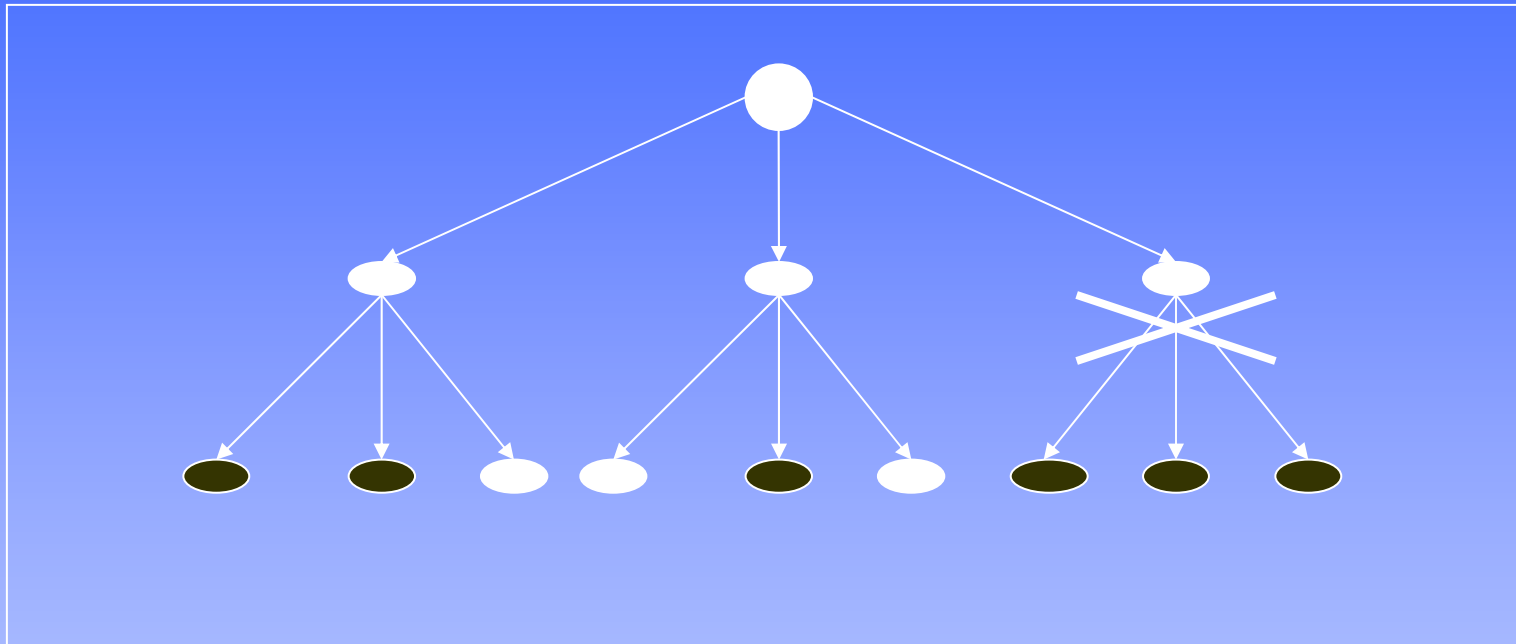
Given a subset of attributes:  $\bar{x} = \{x_1, x_2, \dots, x_k\}$

A random vector  $\bar{\sigma} = \{\sigma_1, \sigma_2, \dots, \sigma_k\}$

Compute  $s = \bar{x} \bar{\sigma}^T$

$$\frac{1}{r} \sum_{p=1}^r \text{Variance}(s)^2 \approx \sum_{l,q} \text{Correlation}(x_l, x_q)^2$$

# Divide-and-Conquer Search for Significant Correlation Coefficients



- Impose a tree-structure:
  - ❑ Leaf node: a unique correlation coefficient
  - ❑ Root of a sub-tree: set of all coefficients corresponding to the leaves in that sub-tree

# Variational Approximation

- Formulate as an optimization problem
- Introduce approximations
- Example: Finite Element Technique

Solve  $-u''(x) = f(x), \quad x \in (a, b), u(a) = u(b) = 0$

Equivalent to minimizing

$$J(u) = \int_a^b (u^{*'}(x) - u'(x))^2 dx$$

# Continued

- Introduce approximation using locally decomposable representation
- Example:  $u(x) = \sum_i \alpha_i \gamma(x)$
- Plug-in the approximation in the objective function

# Regression: Variational Formulation

$$\text{Minimize } J(w) = \frac{1}{2} (w^* - w)^T C (w^* - w)$$

$$\text{where } C_{j,k} = \sum_{i=1}^n x_{i,j} x_{i,k}$$

$$\text{Can be reduced to minimizing } J(w) = -w^T b + \frac{1}{2} w^T C w$$

$$w_j = \frac{b_j - \sum_{k \neq j} C_{j,k} w_k}{C_{j,j}}$$

Inner Product Computation

$$C_{j,k} = \sum_{i=1}^n x_{i,j} x_{i,k}$$

# Approximate Inner Product Computation

- Deterministic Techniques
  - Orthogonal Transformations
- Probabilistic Techniques
  - Random vectors

# Approximate Inner Product Computation

- Egecioglu and Ferhatosmanoglu, 2000

$$u.v \approx (b_1 \Psi_1(u) \Psi_1(v) + b_2 \Psi_2(u) \Psi_2(v))^{1/2}$$

$$\Psi_p(u) = \sum_{i=1}^n u^p \text{ for } p = 1, 2$$

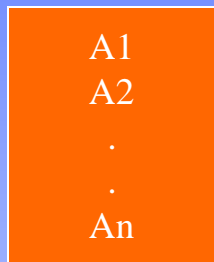
- b1 and b2 can be found by minimizing the error

$$\int ((u.v)^2 - (b_1 \Psi_1(u) \Psi_1(v) + b_2 \Psi_2(u) \Psi_2(v))^{1/2}) dudv$$



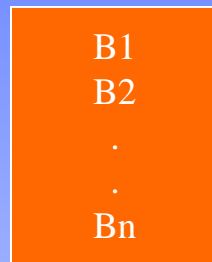
# Approximate Inner Product Computation

Vector 1



$Z_{1,k}$

Vector 2



$Z_{2,k}$



- Node 1 computes  $Z_{1,k}$ 
  - $Z_{1,k} = A_1 \cdot J_1 + \dots + A_n \cdot J_n$
  - $J_i \in \{+1, -1\}$  with uniform probability
- Node 2 calculates  $Z_{2,k}$ 
  - $Z_{2,k} = B_1 \cdot J_1 + \dots + B_n \cdot J_n$
- Compute  $z_{1,k} \cdot z_{2,k}$  for a few times and take the average

# Discussion

- Need for light-weight algorithms for real-time embedded applications
- Data intensive sensor networks may have different needs
- Distributed data stream mining

# Announcement

- National Science Foundation Symposium on Next Generation Data Mining
- [www.cs.umbc.edu/~hillol/NGDM07/](http://www.cs.umbc.edu/~hillol/NGDM07/)