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

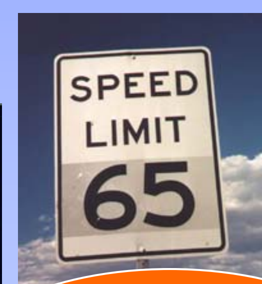
- So what?

Why Mine Vehicle Data?

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

High gas prices

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

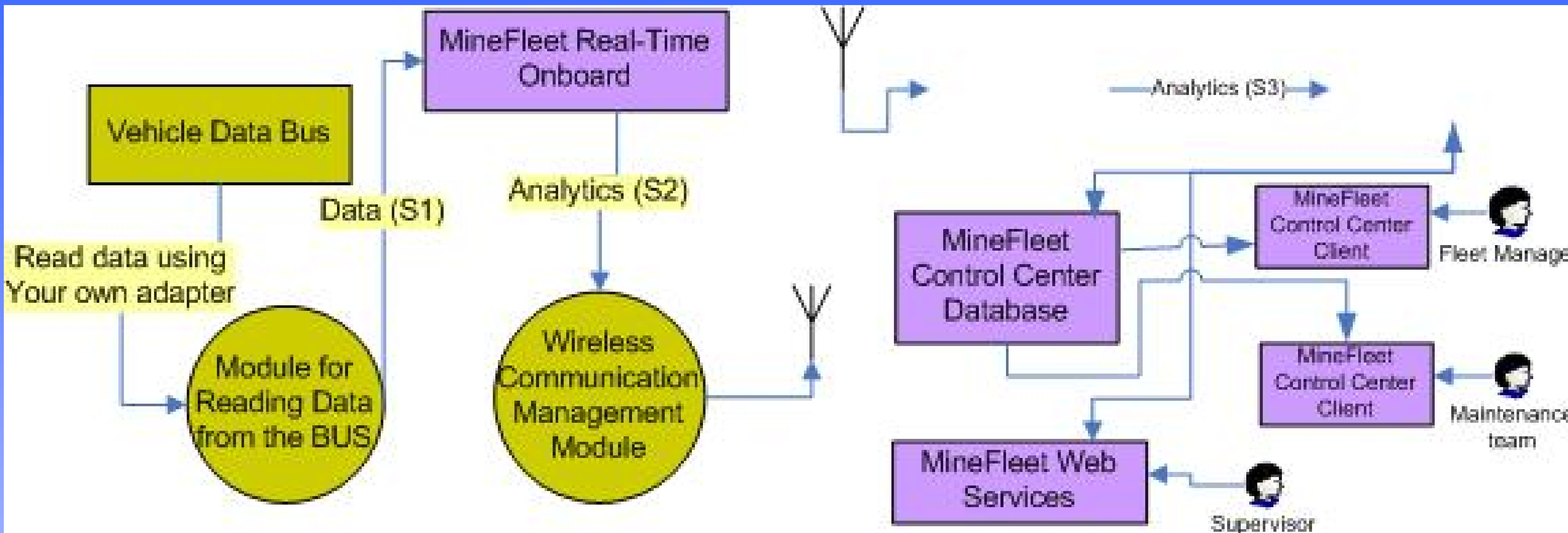
Optimize Fuel Economy by
- Modeling fuel consumption behavior
- Identifying factors that are causing poor fuel economy
- Benchmarking fuel sub-system

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

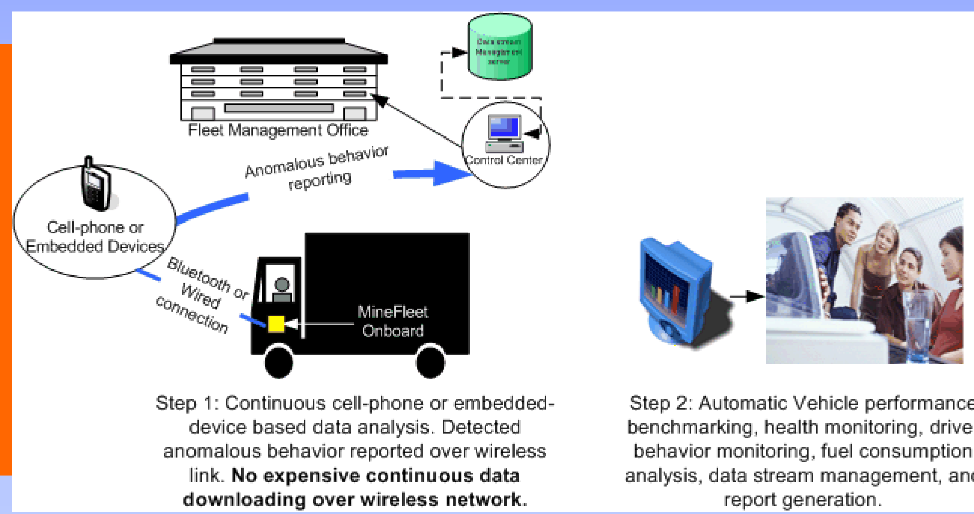
Driver Behavior Monitoring
- Policy-based driver behavior monitoring
- Identify the effect of driver behavior on fuel economy
-

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

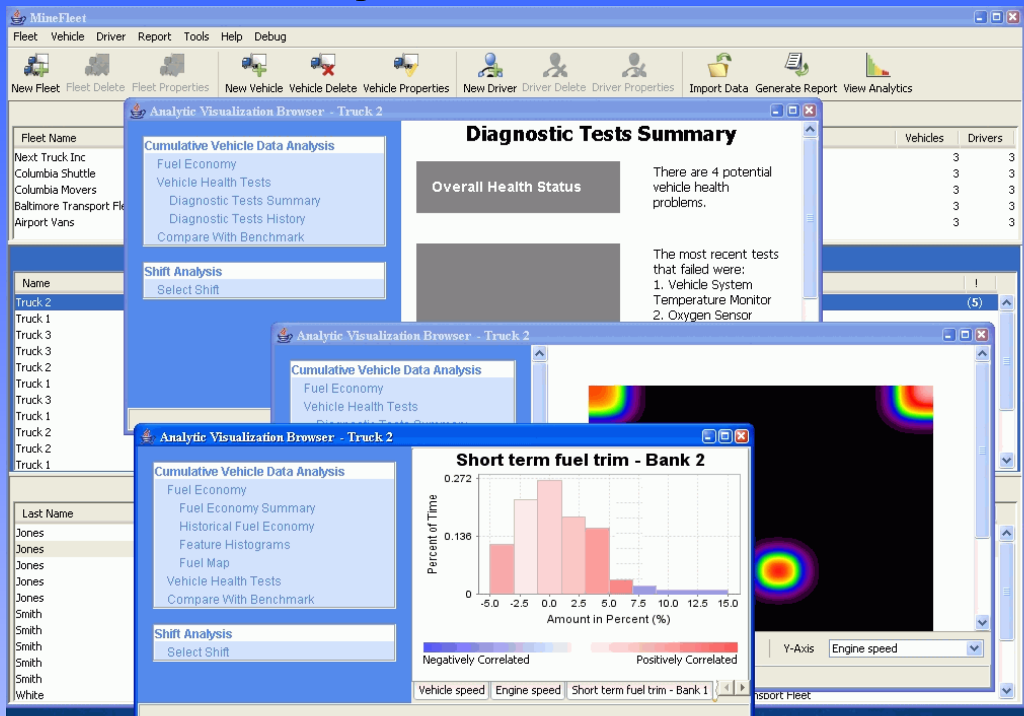
MineFleet Architecture



- 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

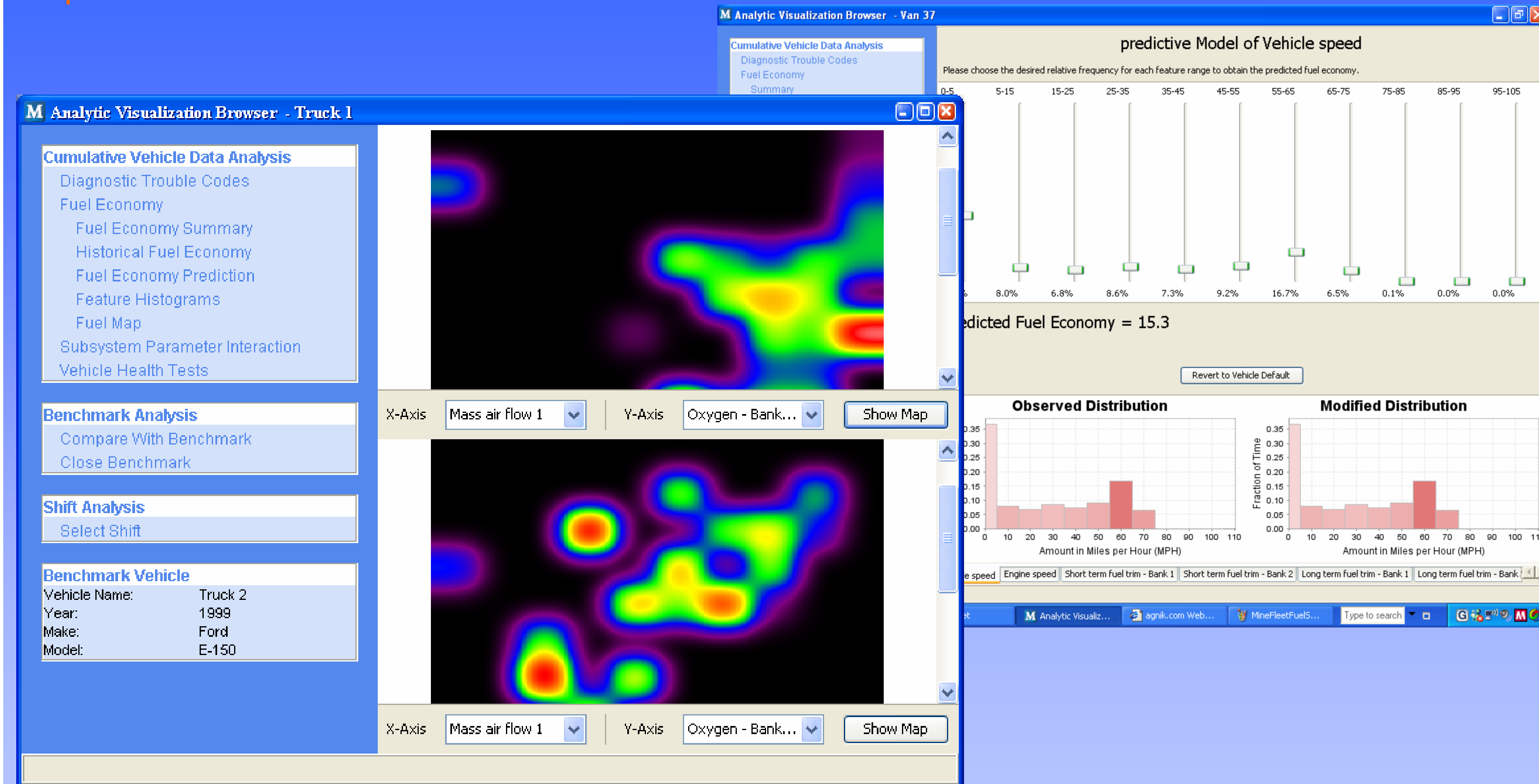
Fuel Economy: Impact of Vehicle Condition and Driver Behavior

- 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

Poor vehicle components and bad driving reduces gas mileage

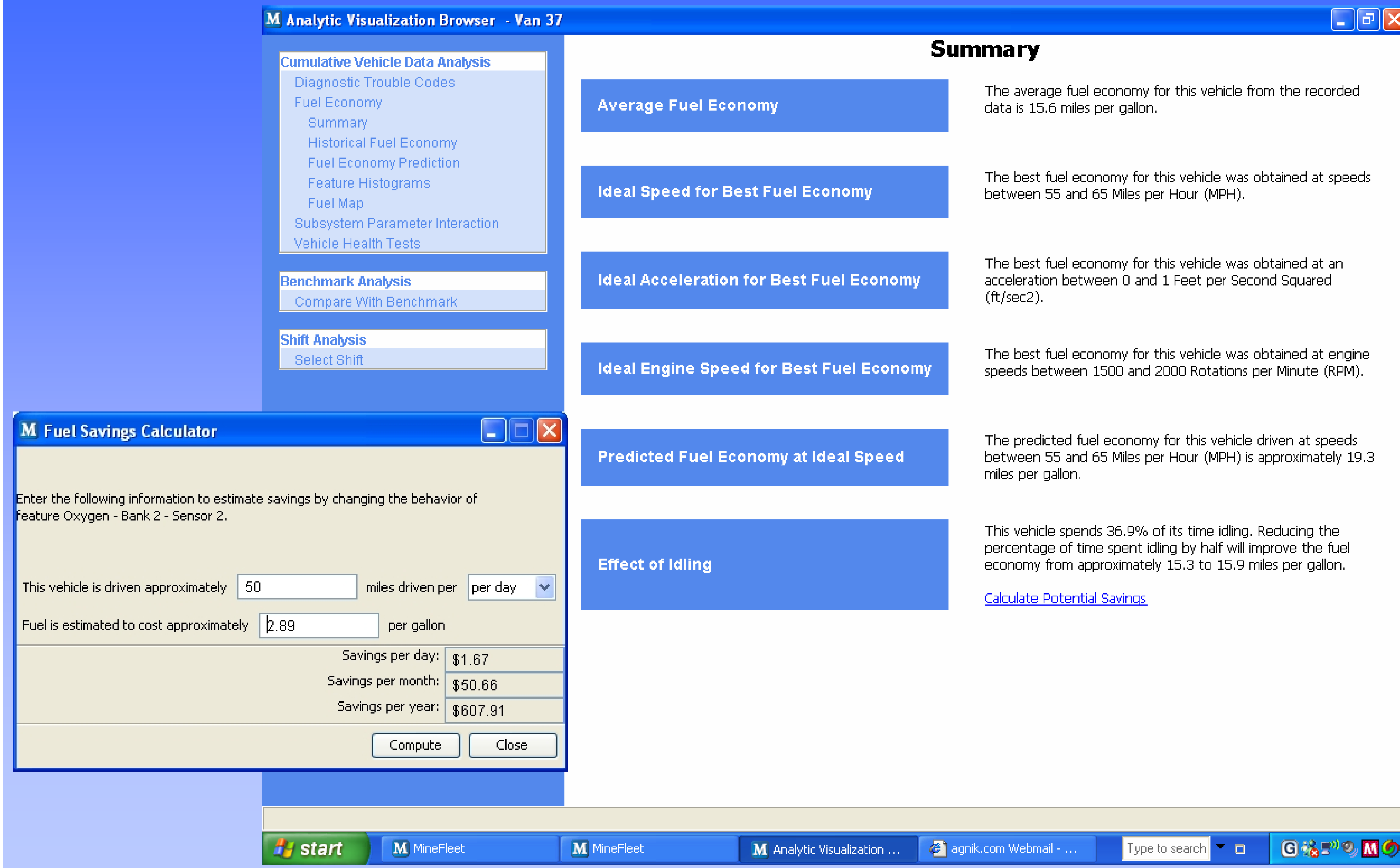
- Build predictive models of the fuel economy as a function of vehicle and driving parameters for optimizing the performance

Fuel Heat Map & Predictive Modeling



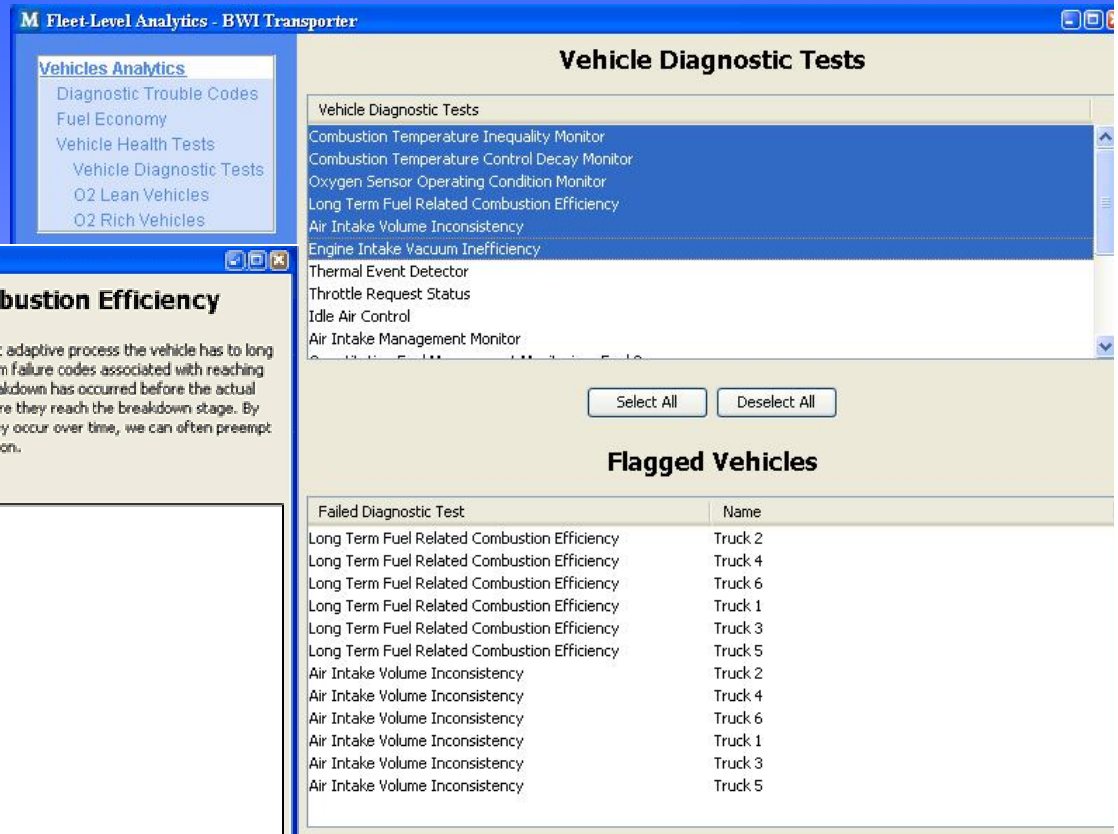
Fuel heat maps show the vehicle operating points that offer high fuel economy. Red color represents high fuel economy and blue represents poor.

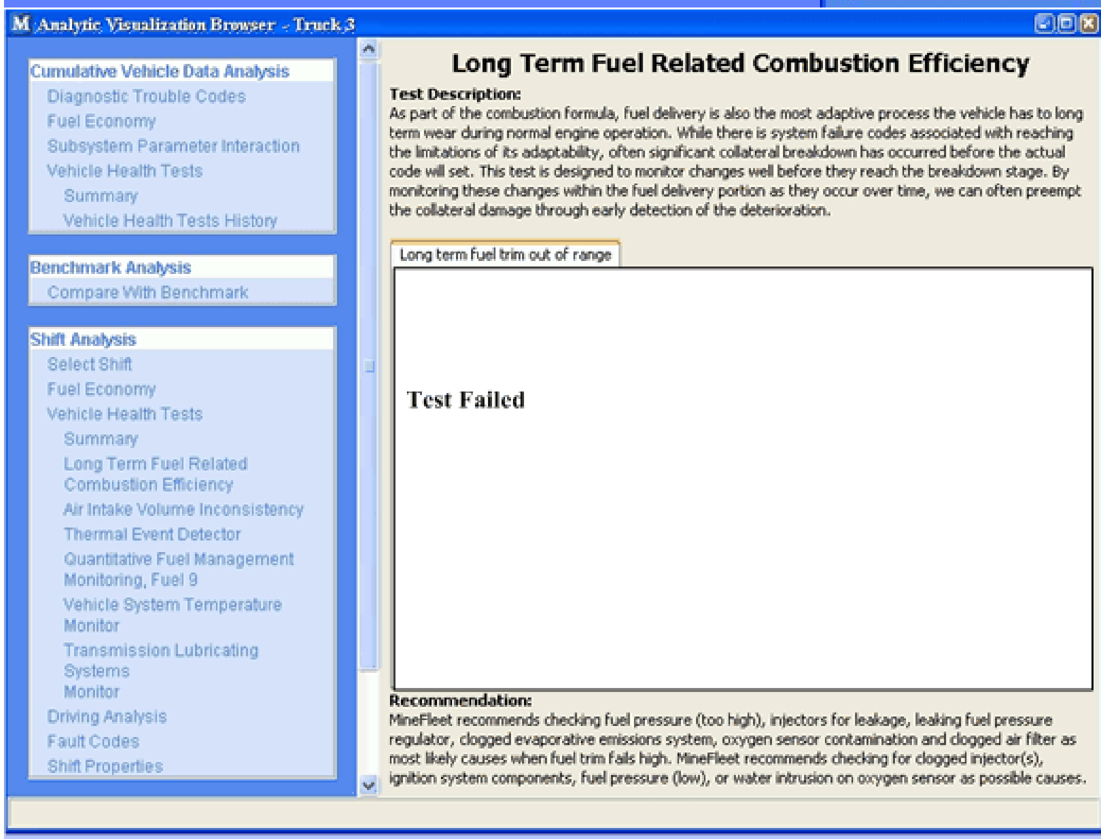
Fuel Consumption Summary Panel & Savings Calculator



Predictive Vehicle Health Management

Detect problems using physics-based model and inductive techniques.



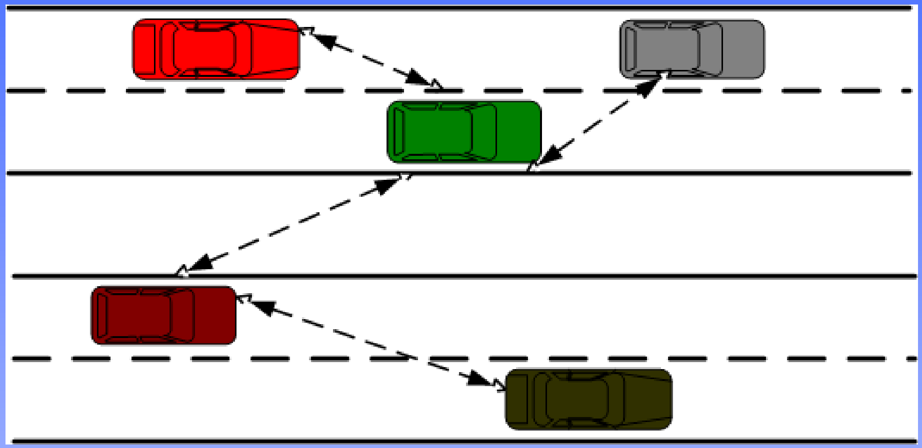


Identifying all the vehicles in a fleet with a specific problem

Detailed description of a specific test that the vehicle passed

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

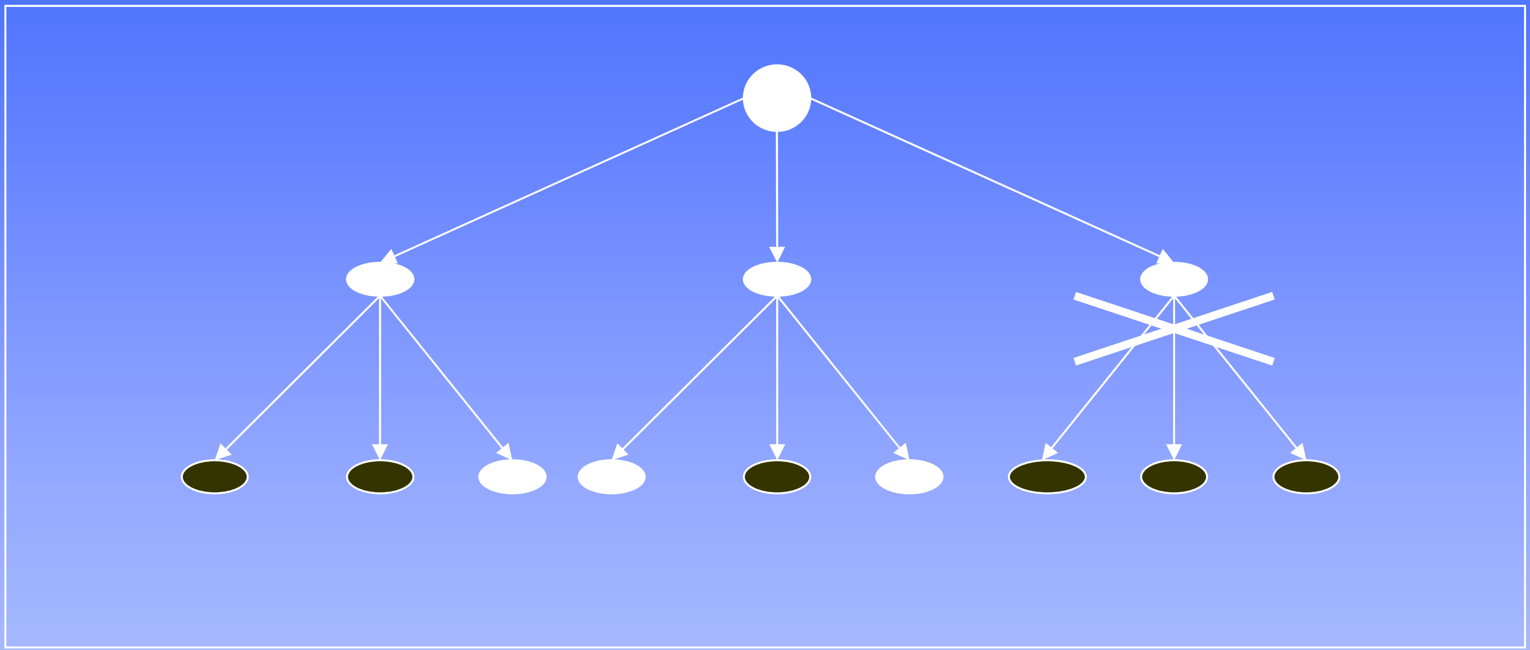
Given a subset of attributes: $\bar{x} = \{x_1, x_2, \ldots\ldots x_k\}$

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

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

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

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

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

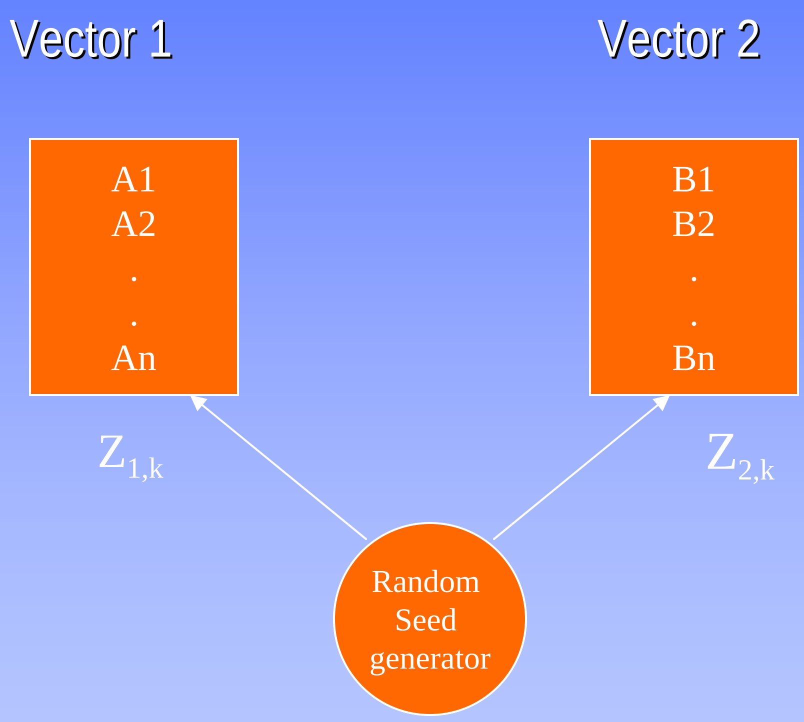
$$u.v \approx (b_1\Psi_1(u)\Psi_1(v) + b_2\Psi_2(u)\Psi_2(v))^{1/2}$$

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

- 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



- Node 1 computes $Z_{1,k}$
  - $Z_{1k}=A1.J_1+..+An.J_n$
  - $J_i \in \{+1,-1\}$ with uniform probability
- Node 2 calculates $Z_{2,k}$
  - $Z_{2k}=B1.J_1+..+Bn.J_n$
- Compute $z_{1,k}.z_{2,k}$ 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/