Data Mining Applications in Ubiquitous Environments: From Theory to Practice and the Vice Versa

Hillol Kargupta
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, MD 21250, USA
http://www.cs.umbc.edu/~hillol
hillol@cs.umbc.edu
&
AGNIK, LLC
Columbia, MD 21045
http://www.agnik.com
hillol@agnik.com

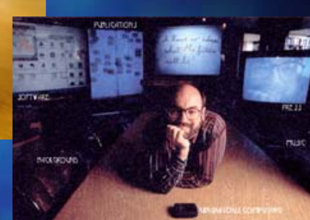
Roadmap

- Introduction
- What is Ubiquitous Data Mining?
- From Cars to Cell-phones: Vehicles for Ubiquitous Data Mining
- MineFleet project
- Conclusions

Ubiquitous: The Word

- Literal meaning: “existing or being everywhere at the same time” --- Webster’s Dictionary

Ubiquitous Computing: Early Work



- Mark Weiser, Xerox Palo Alto Research Center
- “Tabs", "pads", and "boards" built at Xerox PARC, 1988-1994
- Apple Classroom of Tomorrow (ACOT) Project (1985)
- Newton MessagePad, Apple (1993)

Ubiquitous Computing

- Embedding computation into the environment

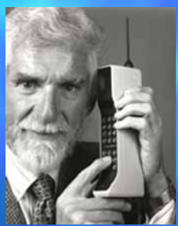
What is Ubiquitous Data Mining?

- Embedding data mining into an environment
- Data mining is not an isolated process anymore
- Rather an integral part of an ensemble of objects and processes in an environment

Data Mining

- Data Mining: Scalable analysis of data by paying careful attention to issues in
  - computing,
  - storage,
  - communication,
  - human-computer interaction.

Evolution of Cell-phones



Cell-phone in 1973

Cell-phone in 2006

- Cell-phones are playing an increasingly important role in making computing ubiquitous.
- Computing getting pushed by the core need of mobile communication.

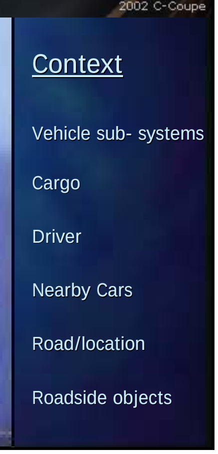
Cell-phones and Cars

Car phone system, 1970

First Cell-phone Call By Martin Cooper of Motorola, 1973

Vehicle and Its Environment

- Data
- Computing
- Communication
- Human-computer interaction.



Context

Vehicle sub- systems

Cargo

Driver

Nearby Cars

Road/location

Roadside objects

Data Sources in a Vehicle

- Vehicle sub- systems generating data using hundreds of sensors:
  - Operating conditions
  - Fuel sub-system
  - Ignition sub-system
  - Transmission sub-system
  - Exhaust sub-system
  - Tire system
- Cargo data:
  - Supply chain data
  - RFID
- Driver behavior data
- Nearby cars generating information
  - Social networking
  - Advertisement
- Road/location
  - Terrain information
  - GPS location
- Roadside objects
  - Service advertisement
  - Traffic information

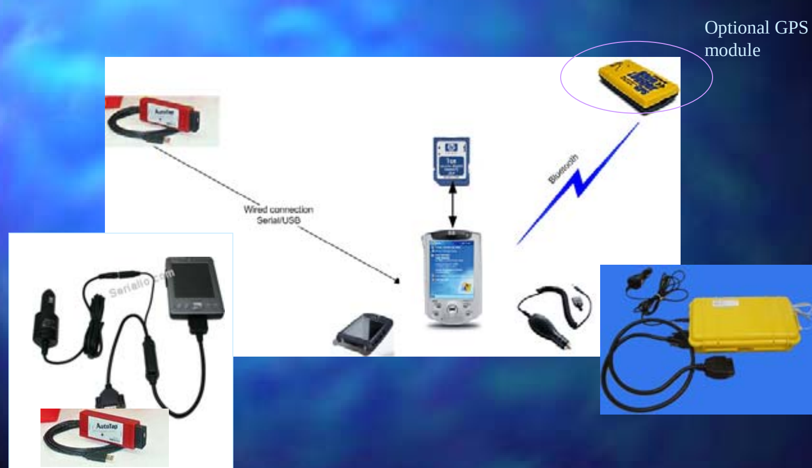
Vehicular Computing and Communication

- Computing:
  - Onboard embedded systems
  - Cell-phones and PDAs
  - Remote machines connected through wireless network
- Communication
  - Network of embedded systems in the vehicle
  - Land or satellite-based wireless network
  - Vehicular Ad hoc Network (VANET)
  - Personal area network inside the vehicle

Human Interaction

- Driver
- Traffic and drivers in a VANET
- Road-side individuals
- Remotely located individuals interacting with the vehicle through a wireless network

Tapping into the Vehicle Sub-System Data



Vehicle Data Stream Mining

- **Vehicle Health Monitoring and Maintenance:**
  - Model and data driven fault-tests
  - Detecting unusual behavior for a subsystem and accessing the data producing this behavior
- **Fuel Consumption Analysis:**
  - Is the vehicle burning fuel efficiently? Identify influencing factors and optimize
  - Detect influence of driver behavior on gas mileage and eliminate inefficient driving practices
- **Driver Behavior Monitoring:**
  - Route monitoring: Fixed and variable routes
  - Direct Cost Issues: e.g. Idling, braking habits
  - Safety Issues: e.g. speeding, trajectory monitoring (e.g. stopping, turns)

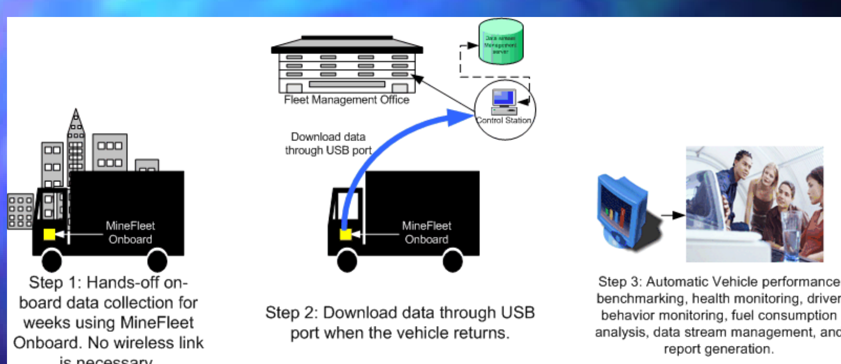
Building a Novel Application: The NABC Mantra

- Need
- Approach
- Benefit
- Competition

MineFleet Just-in-Time

- Modeling, benchmarking, and monitoring of vehicle health, driver behavior, fuel-consumption, and fleet characteristics.

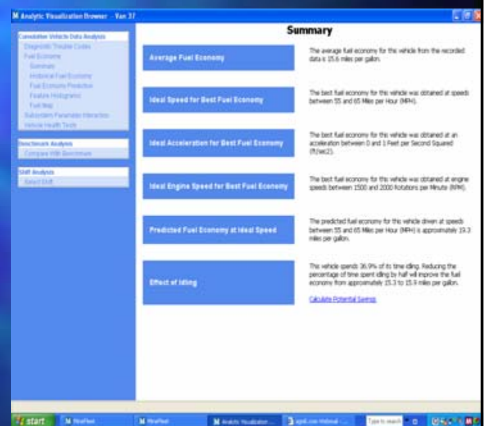
Just-in-time Mode



- Also works with third-party hardware for collecting onboard data

Fuel Economy: Impact of Driver Behavior

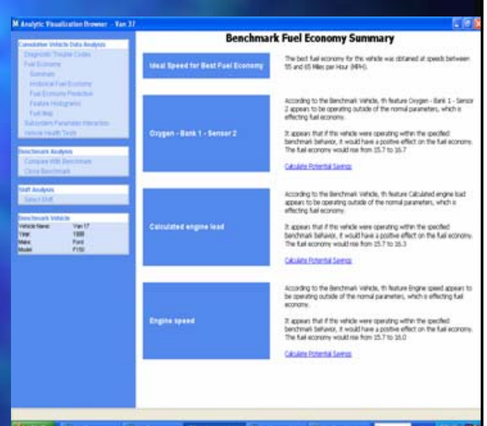
- Quantify the effect of driver behavior on fuel economy. Examples:
  - Effect of speeding
  - Effect of acceleration
  - Effect of braking
  - Effect of idling



Summary of the fuel economy analysis in MineFleet

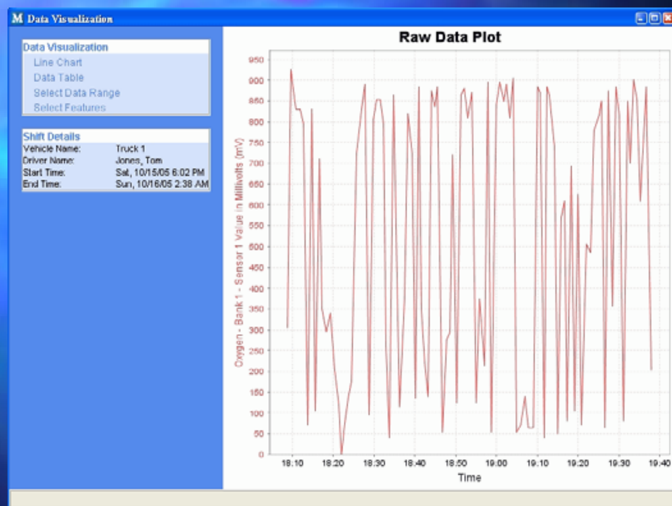
Fuel Economy: Impact of Vehicle Condition

- 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.



Summary of the fuel economy analysis in MineFleet

Oxygen Sensor 1

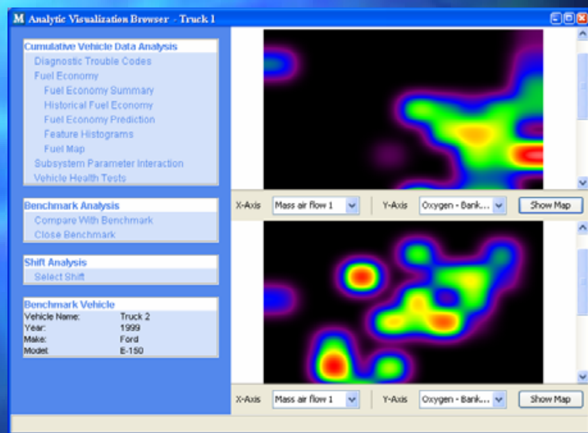


- Measures the amount of oxygen in the exhaust before it goes through the catalytic converter

Fuel Economy: Predictive Modeling

- Build a predictive model of the fuel economy as a function of vehicle and driving parameters for optimizing the performance
- Predictive modeling allows detecting the effect of any specific vehicle or driver parameter on fuel economy.

Fuel Heat Map



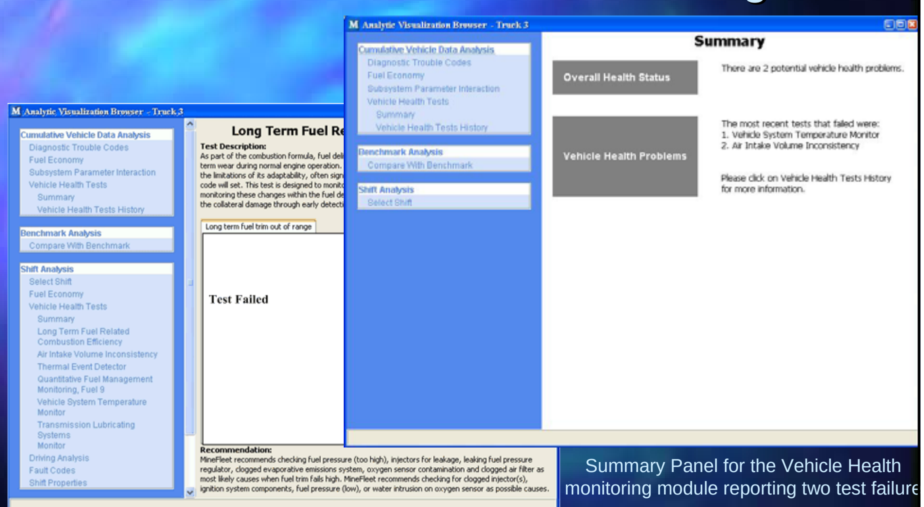
Fuel heat maps show the vehicle operating points that offer high fuel economy. Red color represents high fuel economy and blue represents poor. Fuel heat maps of different vehicles are used in this figure for comparative performance analysis.

Find anomalous vehicle behavior and avoid expensive breakdowns

Predictive Modeling for Vehicle Health Analysis

- Detect problems using model and data driven fault detection tests well before Vehicle DTC codes show up.
- Generate alerts when MineFleet detects unusual behavior of a subsystem and access the data producing this behavior.

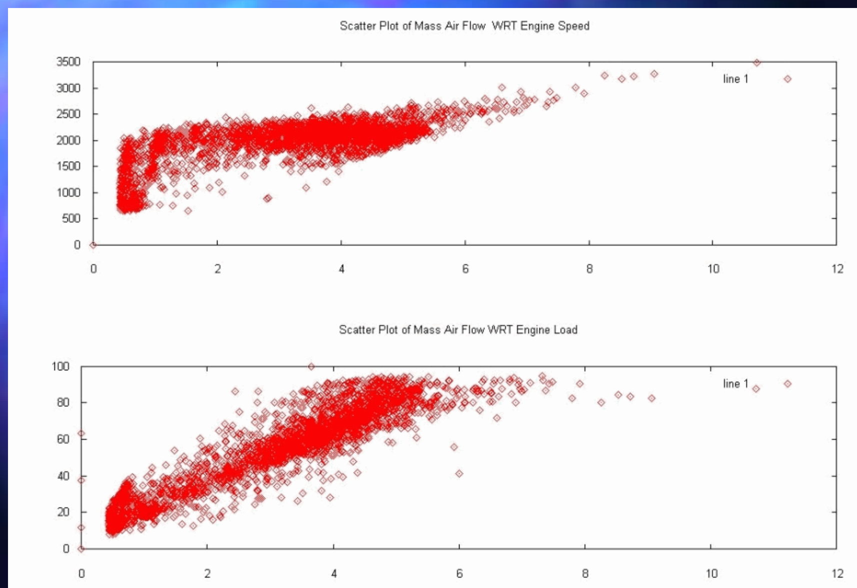
Screen Shots: Vehicle Health Management



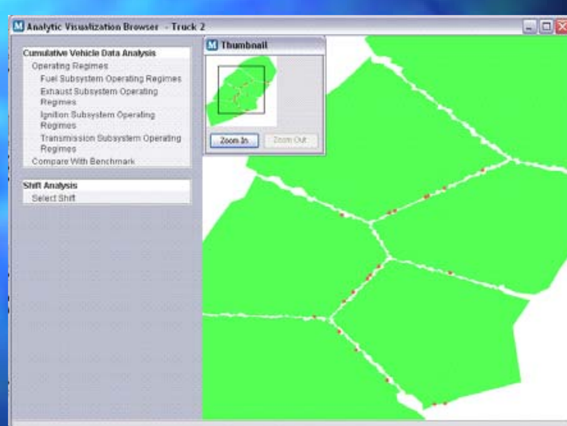
Summary Panel for the Vehicle Health monitoring module reporting two test failure

Detailed description of a specific test that the vehicle passed

Air Intake Behavior



More Screen Shots



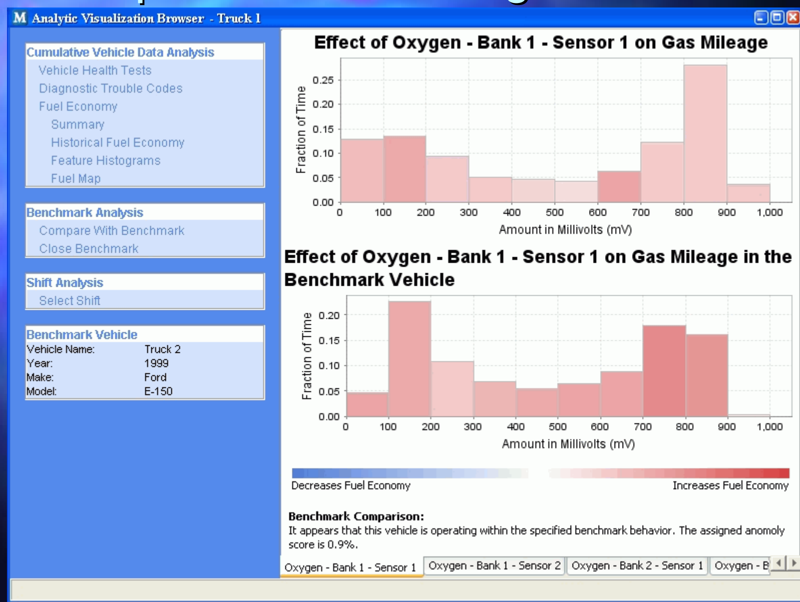
Construct the vehicle operating regimes using advanced mathematical projection techniques and identify abnormal behavior.

More Examples of Vehicle Health Monitoring Tests

- **Fuel System**
  - Oxygen Sensor Operating Condition Monitoring.
  - Long Term Fuel related Combustion Efficiency Monitoring
  - Air Intake Volume Inconsistency Monitoring
  - Engine Intake Vacuum Inefficiency Monitoring
  - Engine Thermal Event Detection
  - Throttle Request Status Monitoring
  - Idle Control Monitoring
  - Intake Air Management Monitoring
  - Quantitative Fuel Management Monitoring
  - Vehicle System Temperature Management Monitoring
  - Quantitative Fuel System Management monitoring
- **Exhaust System**
  - Combustion Temperature Inequality Monitoring
  - Combustion Temperature Control Decay Monitoring
- **Ignition System**
  - Vehicle Ignition System Voltage Monitoring
  - Spark Control Monitoring
  - Vehicle Operating System Voltage Monitoring

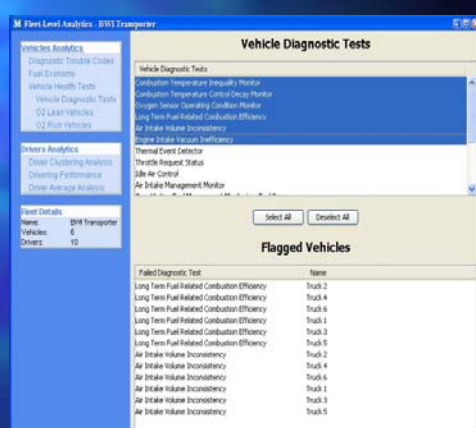
Benchmark vehicles and vehicle-subsystems to identify poorly performing ones

Example: Benchmarking O2 Sensor

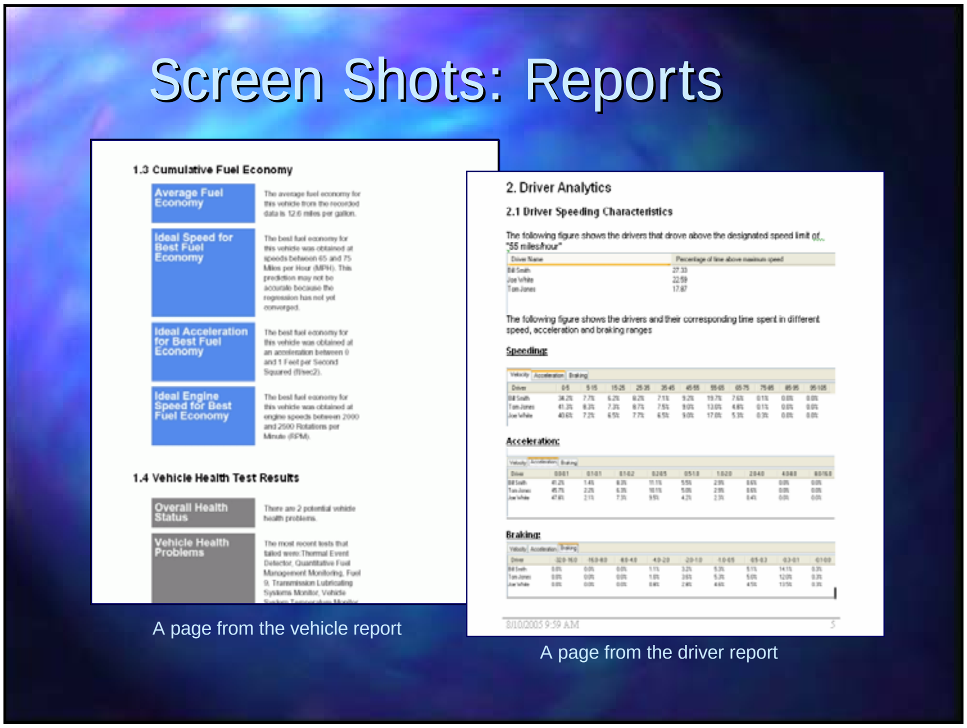


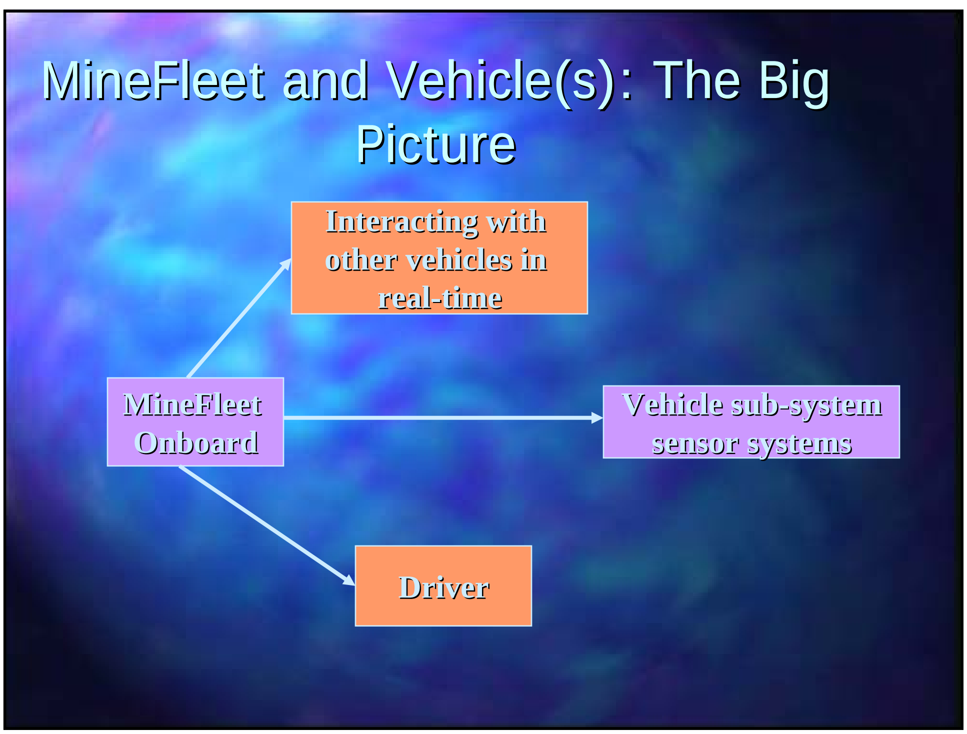
Fleet Analytics

- Compare and benchmark vehicles in the fleet. Example:
  - Compare the fuel subsystems of all similar vehicles in the fleet and identify the unusually behaving ones.
- Identify all vehicles showing specific type of problems. Example:
  - Identify all the vehicles that are running O2 lean or rich



Identifying all the vehicles in a fleet with a specific problem (e.g. O2 Lean, miss-fires, Long-term fuel related combustion efficiency test-failure)

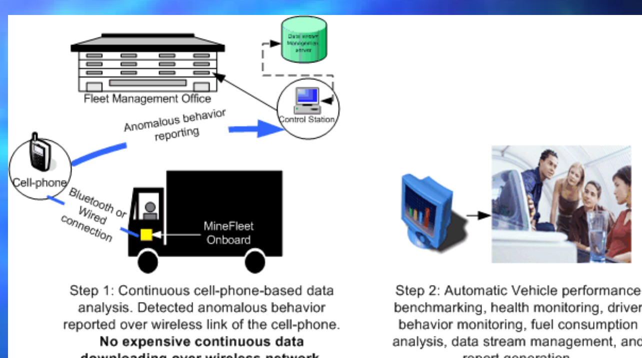

Screen Shots: Reports
1.3 Cumulative Fuel Economy
Average Fuel Economy
Ideal Speed for Best Fuel Economy
Ideal Acceleration for Best Fuel Economy
Ideal Engine Speed for Best Fuel Economy
1.4 Vehicle Health Test Results
Overall Health Status
Vehicle Health Problems
A page from the vehicle report
2. Driver Analytics
2.1 Driver Speeding Characteristics
Speeding:
Acceleration:
Braking:
A page from the driver report


MineFleet and Vehicle(s): The Big Picture
Interacting with other vehicles in real-time
MineFleet Onboard
Vehicle sub-system sensor systems
Driver

Further Embedding MineFleet in the Vehicle

- MineFleet *Real-Time*: A cell-phone-based version for onboard modeling and monitoring of the vehicle data streams.
- MineFleet *VANET*

MineFleet Real-Time



- Real-time monitoring onboard the vehicle using a cell-phone
- Bluetooth connection with the data-bus

A Counting Problem

- *Count the number of Engine Misfires in last 6 months*
- Abstract Problem: Count the number of 1-s from a moving window in a binary stream.

  . . . . .1 0 0 0 1 1 | 1 0 1 0 1 0 0 0 1 |

- Need to account for the expiring bits.
- Naïve solution takes O(n) space. Expensive.

An Approximate Solution

- Store the counting information among a set of buckets of known counts.
- Time-stamp of a bucket = time stamp of the most recent entry in the bucket.
- Track the buckets.
- When the time-stamp of a bucket expires, through away the bucket.
- Error in oldest bucket only.

Continued

- Exponential histograms: Buckets of exponentially increasing size.
- Bucket sizes: 1, 2, 2^2, 2^3, ... 2^h .
- Need only O(log N) buckets.
- A bucket size can take at most log N bits.

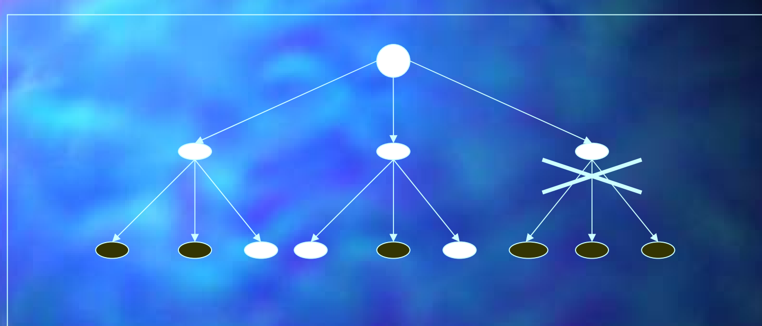
Correlation Matrix Computation

- Given data matrix X
- Naïve computation: Compute X^TX
- Compute in the frequency domain (take Fourier transformation)
- StatStream (Zhu and Shasha, 2002)

Resource Constrained Change Detection in the Correlation Matrix

- Kargupta, Puttagunta, Klein, 2005
- Efficiently detect changes in the correlation matrix
- Identify the region of the matrix that contain significantly changed coefficients

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

Does a Sub-tree Contain Any Significant Coefficient?

Given a subset of attributes: $\{i_1, i_2, \ldots\ldots i_k\}$;

Is there any significantly correlated pair of attributes?

The j-th row of the data matrix X: $x_j = [x_{j,1} x_{j,2} \ldots\ldots x_{j,n}]$

Entries from the j-th row x_j corresponding to attributes in G

$[x_{j,i_1} x_{j,i_2} \ldots\ldots x_{j,i_k}]$

Continued

Consider a random vector $\sigma_p = [\sigma_{i_1,p}, \sigma_{i_2,p}, \ldots\ldots \sigma_{i_k,p}]$

$\sigma_{j,p} \in \{-1,1\}$ with uniform probability

$$s_{j,p} = \sum_{l=i_1,i_2,\ldots i_k} x_{j,l} \sigma_{l,p}$$

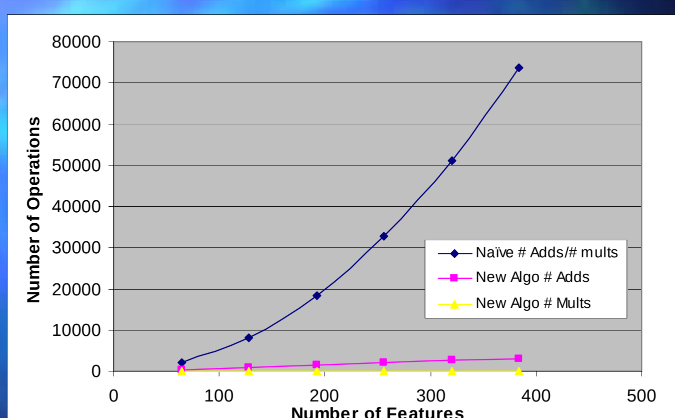
$$S_{\{i_1,i_2,\ldots i_k\},p} = [s_{1,p} s_{2,p} \ldots s_{m,p}]^T$$

The Test at Every Node

$$\frac{1}{r}\sum_{p=1}^{r} Var(S_{\{i_1,i_2,\ldots,i_k\},p})^2 \approx \sum_{l_1,q_1} Corr(x_{l_1}, x_{q_1})^2$$

- Compute the left hand side at every node and proceed only if it is greater than a threshold.

Detecting No Changes



- **Number of multiplications and additions performed by the naïve and the proposed algorithms for correctly detecting no significant changes in the correlation matrix.**

Eigenstate Monitoring: Using Bounds from Matrix Perturbation Theory

- (λ_1,v_1) (λ_2,v_2): Most significant eigenvalue and the corresponding eigenvector of Cov_t and Cov_{t-1} .

$$\Delta = Cov_t - Cov_{t-1}$$

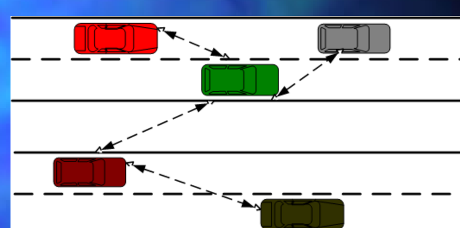
$$\| v_1 - v_2 \|_2 \le \frac{4 \| \Delta \|_F}{\delta - \sqrt{2} \| \Delta \|_F}$$

$$| \lambda_1 - \lambda_2 | \le \sqrt{2} \| \Delta \|_F$$

$$\text{Frobenius norm} \| \Delta \|_{\mathrm{F}} = (\sum_i \sum_j \Delta_{ij}^2)^{1/2}$$

MineFleet VANET Project

Advertisement on a bus



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

Some of the Challenges

- Network topology is extremely dynamic
- Connection time is small, 15-30 seconds depending upon the speed and the connection protocol
- Existing P2P data mining algorithms do not work.

Resources

- Distributed and Ubiquitous Data Mining Wiki:

  http://www.umbc.edu/ddm/wiki/

- Bibliography, papers, data, software.

Conclusions

NABC: Need, Approach, Benefit, Competition