



Large Scale Vehicle Performance Monitoring in Distributed Environments

University of Maryland Baltimore County
&
Agnik, LLC

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

Copyright 2009, Agnik, LLC



Road Map

- Distributed Data Mining: An Overview
 - Algorithms
 - Commercial Applications
- Anomaly Detection from Multi-Party Data
- Vehicle Performance Data Mining and Monitoring
- Summary

Data Mining and Distributed Data Mining

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



- Distributed Data Mining (DDM): Mining data using distributed resources.

Data Mining for Distributed and Ubiquitous Environments: Applications

- ❑ **Mining Large Databases from distributed sites**
 - Grid data mining in Earth Science, Astronomy, Counter-terrorism, Bioinformatics
- ❑ **Monitoring Multiple time critical data streams**
 - Monitoring vehicle data streams in real-time
 - Monitoring physiological data streams
- ❑ **Analyzing data in Lightweight Sensor Networks and Mobile devices**
 - Limited network bandwidth
 - Limited power supply
- ❑ **Preserving privacy**
 - Security/Safety related applications
- ❑ **Peer-to-peer data mining**
 - Large decentralized asynchronous environments

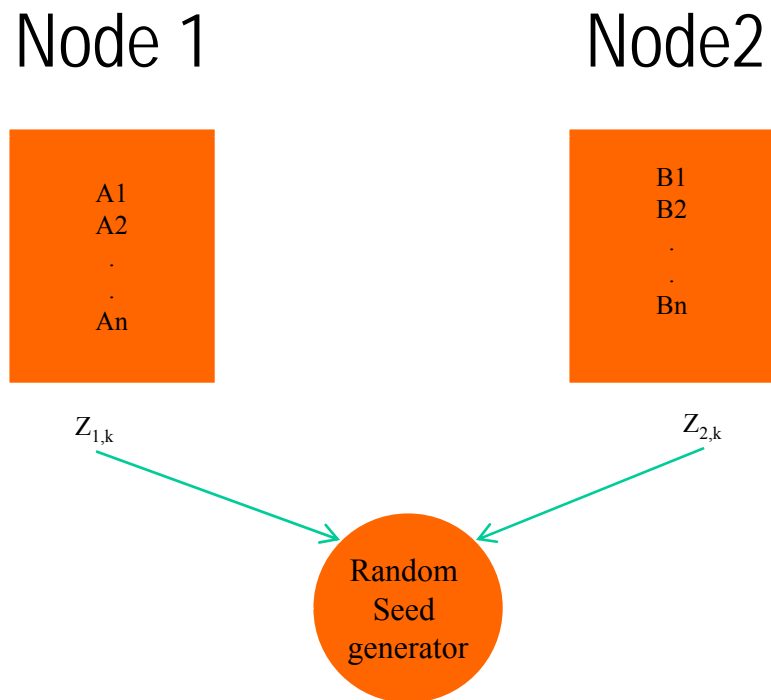
DDM Algorithms: Some Examples

- Distributed Primitive Computation
 - Probabilistic techniques
 - Deterministic exact techniques
 - Deterministic approximation techniques
- Anomaly Detection
 - Principal component analysis based approach
 - Optimization-based approach, distributed linear programming

DDM Algorithm Design: Methodology

- Distributed environment $G = (V, E)$
- Each node contains some data O_k
 - Same schema
 - Different schemas
- Compute function $f(V)$
- Construct a decomposed representation of $f(V)$ where $f(V)$ can be computed from locally computed functions $p(O_k)$
- Correctness and Scalability

Distributed Randomized Similarity Search



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

Distributed PCA & Max Sum-Square Computation

- Principal Component computation in heterogeneous environment
- Can be reduced to the distributed sum-square computation

Problem

- Site A has $a_1, \dots, a_n \in \mathbb{R}$, Site B has $b_1, \dots, b_n \in \mathbb{R}$
- Sites must compute

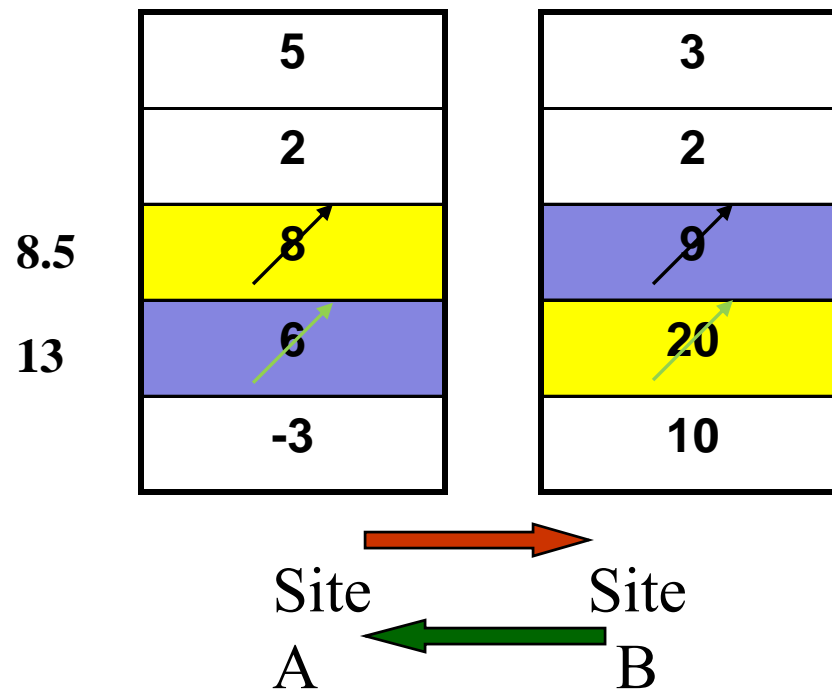
$$i^* = \text{Argmax} \{ c_1 = (a_1 + b_1)^2, \dots, c_n = (a_n + b_n)^2 \}$$

Distributed Max Sum Square

The Algorithm

- (1) Site A selects max of a_{iA} 's and sends $\langle a_{iA}, iA \rangle$ to Site B
(2) Site B selects max of b_{iB} 's and sends $\langle b_{iB}, iB \rangle$ to Site A
- Site A receives the message and replies with $\langle a_{iB} \rangle$. Site B replies with $\langle b_{iA} \rangle$
- Both sites now have a_{iA} , a_{iB} , b_{iA} , b_{iB} and the corresponding indices. If $iA = iB$ terminate and return iA .
- Otherwise
 - (1) Site A replaces a_{iA} with $(a_{iA} + b_{iA})/2$ and a_{iB} with $(a_{iB} + b_{iB})/2$
 - (2) Site B replaces b_{iA} and b_{iB} similarly

Illustration



- Communications Cost Analysis:
 - Average case $O(\log n)$
 - Worst case $O(n)$
 - Synchronous

Distributed Classifier Learning & Outlier Detection

- Linear classifier construction and outlier detection
- Can be posed as linear programming problem.

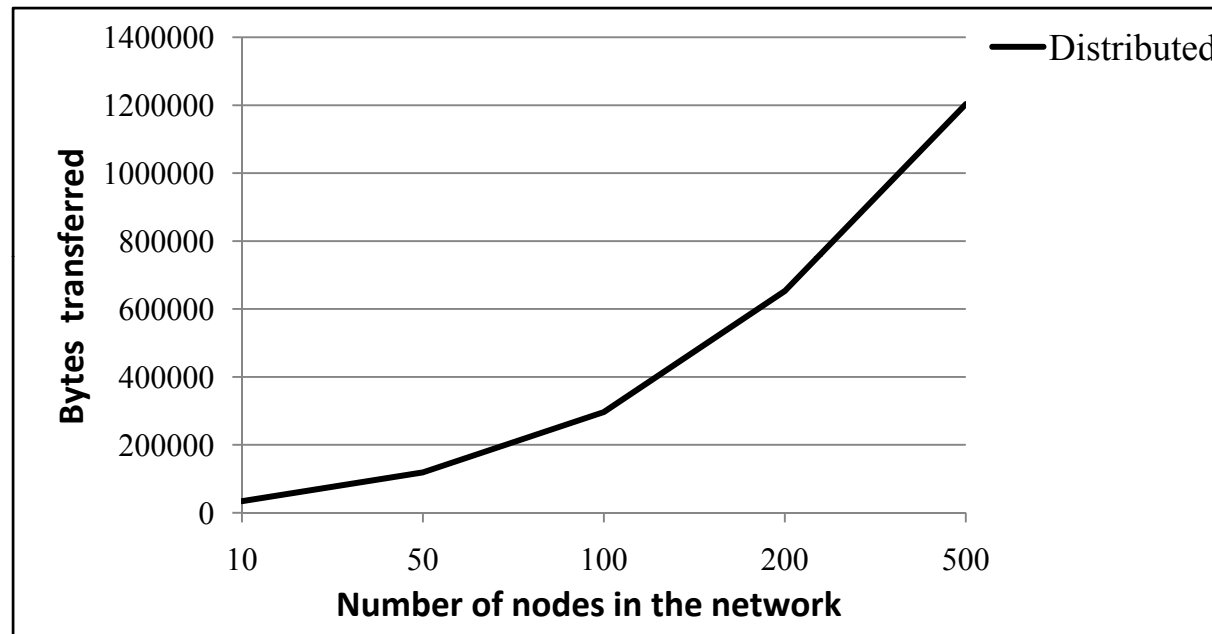
Examples:

- Minimizing the error
- Minimizing the entropy objective function by taking out the outliers
- Distributed linear programming
- Distributed simplex algorithm

Distributed Simplex

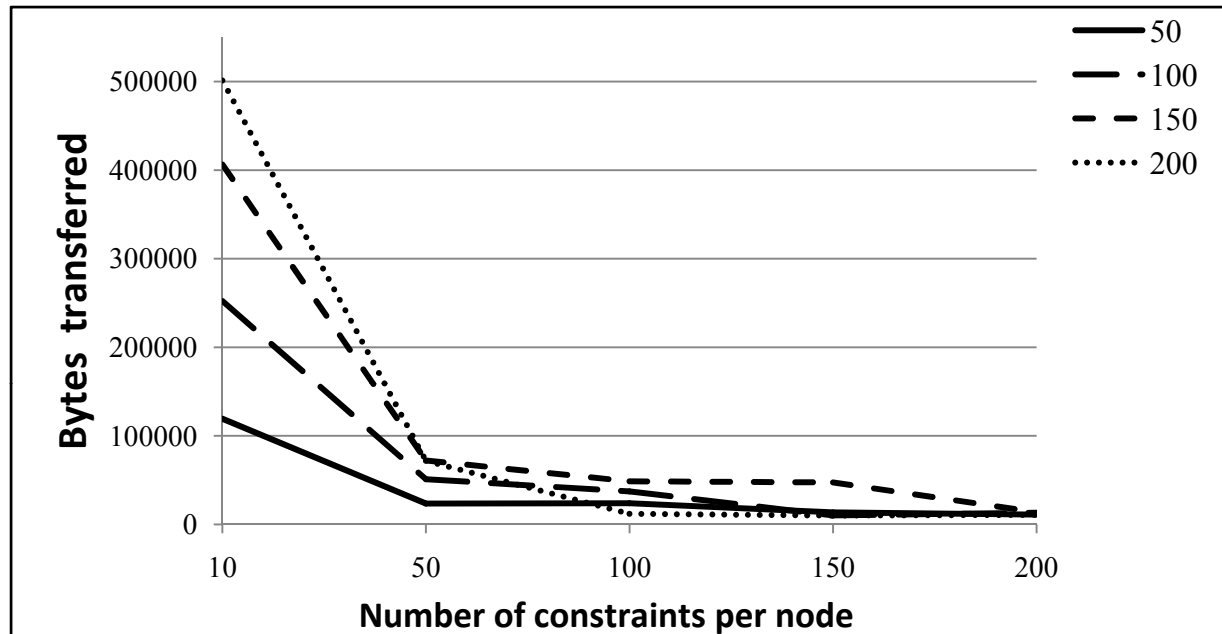
- Simplex rely upon pivot computations
- Pivot computation can be reduced to distributed min computation

Communication Cost vs. Network Size



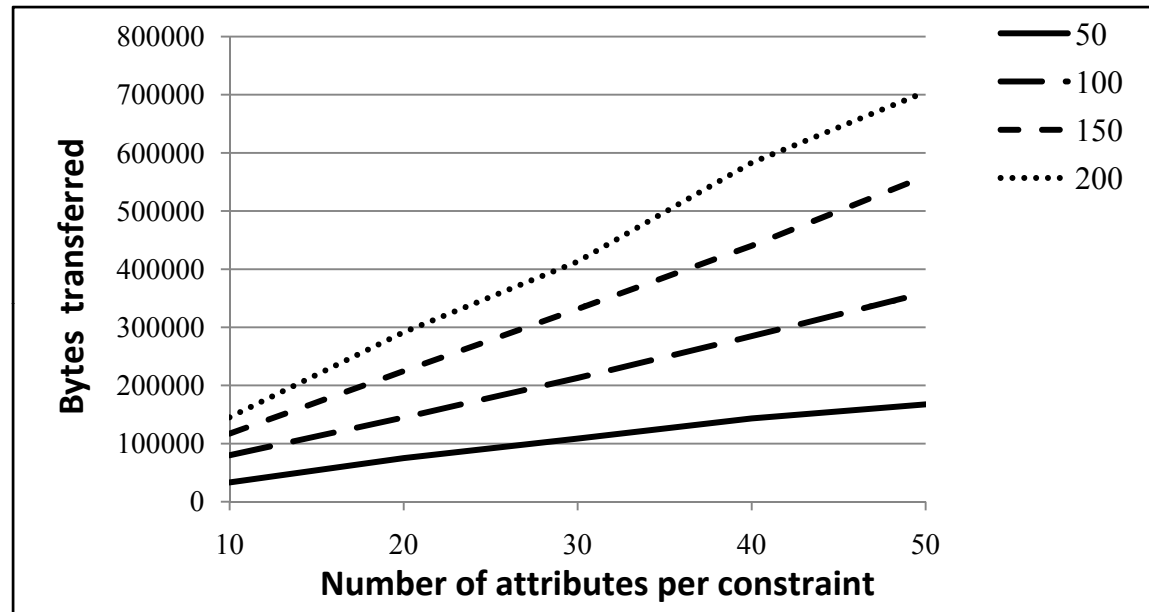
- Number of nodes in the network is varied from 10 to 500 nodes
- Number of variables in a constraint equation is kept constant at 35

Communication Cost vs. Input size



- Graph shows results for a 50 node network with 4 different topologies with number of edges varying from 50 to 200
- The number of constraints is varied from 10 to 200

Communication Cost vs. Attributes per Constraint



- Graph shows results for a 50 node network with 4 different topologies with number of edges varying from 50 to 200

Some References: P2P and Distributed Data Mining

- H. Kargupta and K. Sivakumar, (2004) Existential Pleasures of Distributed Data Mining. *Data Mining: Next Generation Challenges and Future Directions*. Editors: H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha. AAAI/MIT Press.
- H. Dutta, C. Giannella, K. Borne and H. Kargupta (2007). Distributed Top-K Outlier Detection in Astronomy Catalogs using the DEMAC system. *Proceedings of SIAM International Conference on Data Mining*.
- K. Das, K. Bhaduri, and H. Kargupta. (2008). An Ordinal Framework for Identifying Significant Inner Product Elements in a Peer-to-Peer Network. *IEEE Transactions on Knowledge and Data Engineering*, volume 19, number 3.
- J. Branch, B. Szymanski, R. Wolff, C. Gianella, H. Kargupta. (2006). In-Network Outlier Detection in Wireless Sensor Networks. *Proceedings of the 26th International Conference on Distributed Systems*, 2006.
- R. Wolff, K. Bhaduri, H. Kargupta. (2009). A Generic Local Algorithm for Mining Data Streams in Large Distributed Systems. *IEEE Transactions on Knowledge and Data Engineering*. Volume 21, Issue 4, April 2009. Pages 465-478.

Commercial Applications

Commercial Products from Agnik

- DIA: Anomalous event detection from distributed data sources (*US Missile Defense Agency*)
- PURSUIT: Network threat detection from multi-party privacy-sensitive distributed data (*US Department of Homeland Security*)
- MineFleet: Real-time vehicle performance monitoring for commercial fleets (*Commercial system adopted by many organizations*)

Academic Projects at UMBC

- PADMINI: Distributed data mining from NASA Virtual Observatories (*NASA*)
- Green Flights, Aircraft Health, and Distributed Data Stream Mining (*2008 IBM Innovation Award*)

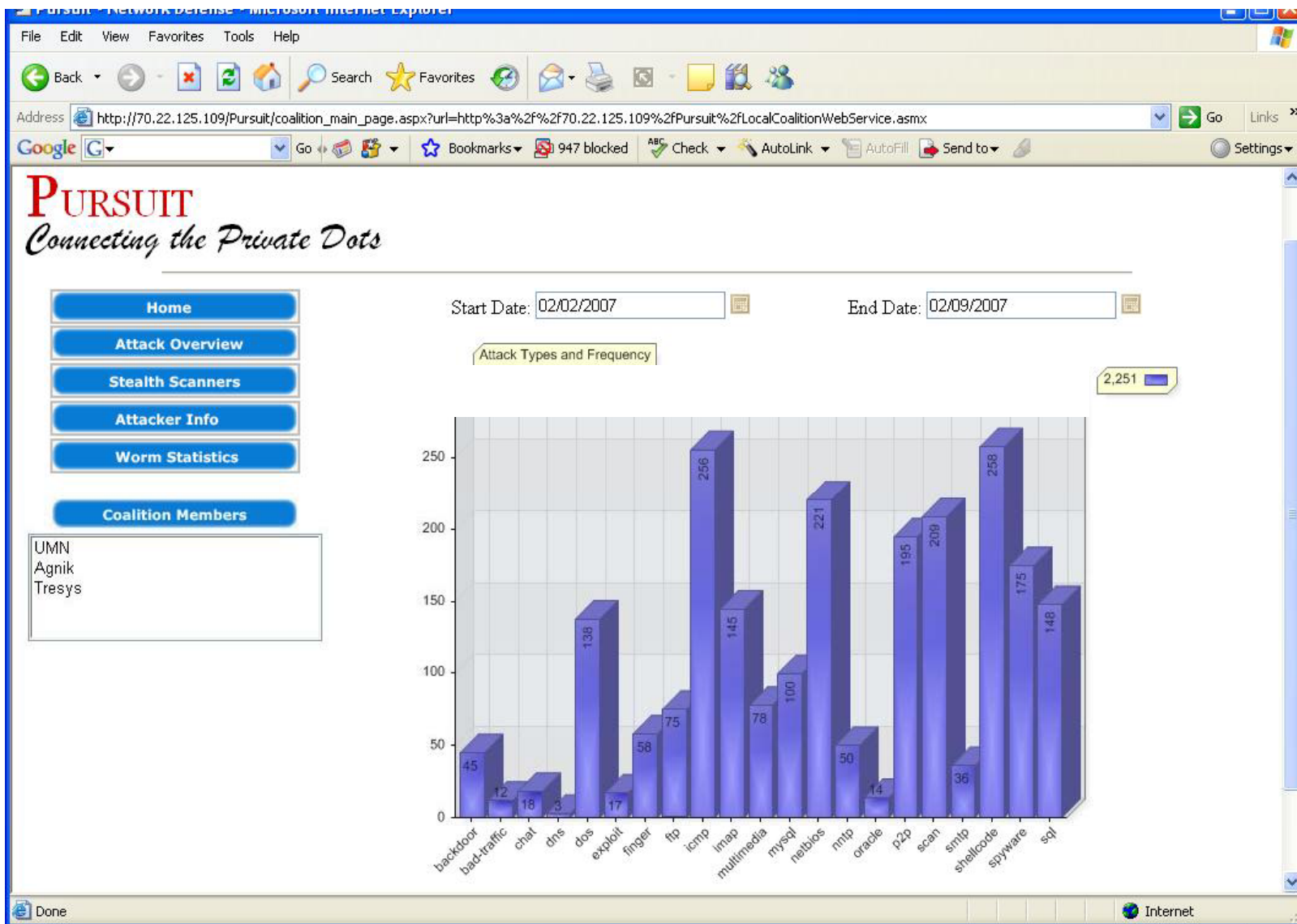
Private & Secure Data Mining from Multi-Party Distributed Data

- Compute global patterns without direct access to the multi-party raw distributed data
- Minimize communication cost
- Must come with provably correct guarantees with respect to a given privacy model
- Must be scalable with respect to
 - number of data sites
 - size of the data
- Privacy-preserving data mining
 - Blends in “pattern-preserving” transformations with data analysis

How PURSUIT Works for the User

- Need to have your own sensor such as SNORT, MINDS
- Download PURSUIT plug-in for the sensor and install
- PURSUIT plug-in offers
 - A stand-alone interface for processing your alerts from the sensor and cross-domain analysis
 - Web account for detailed cross-domain statistics
 - Optional distributed collaboration management module for managing the threats and archiving forensics

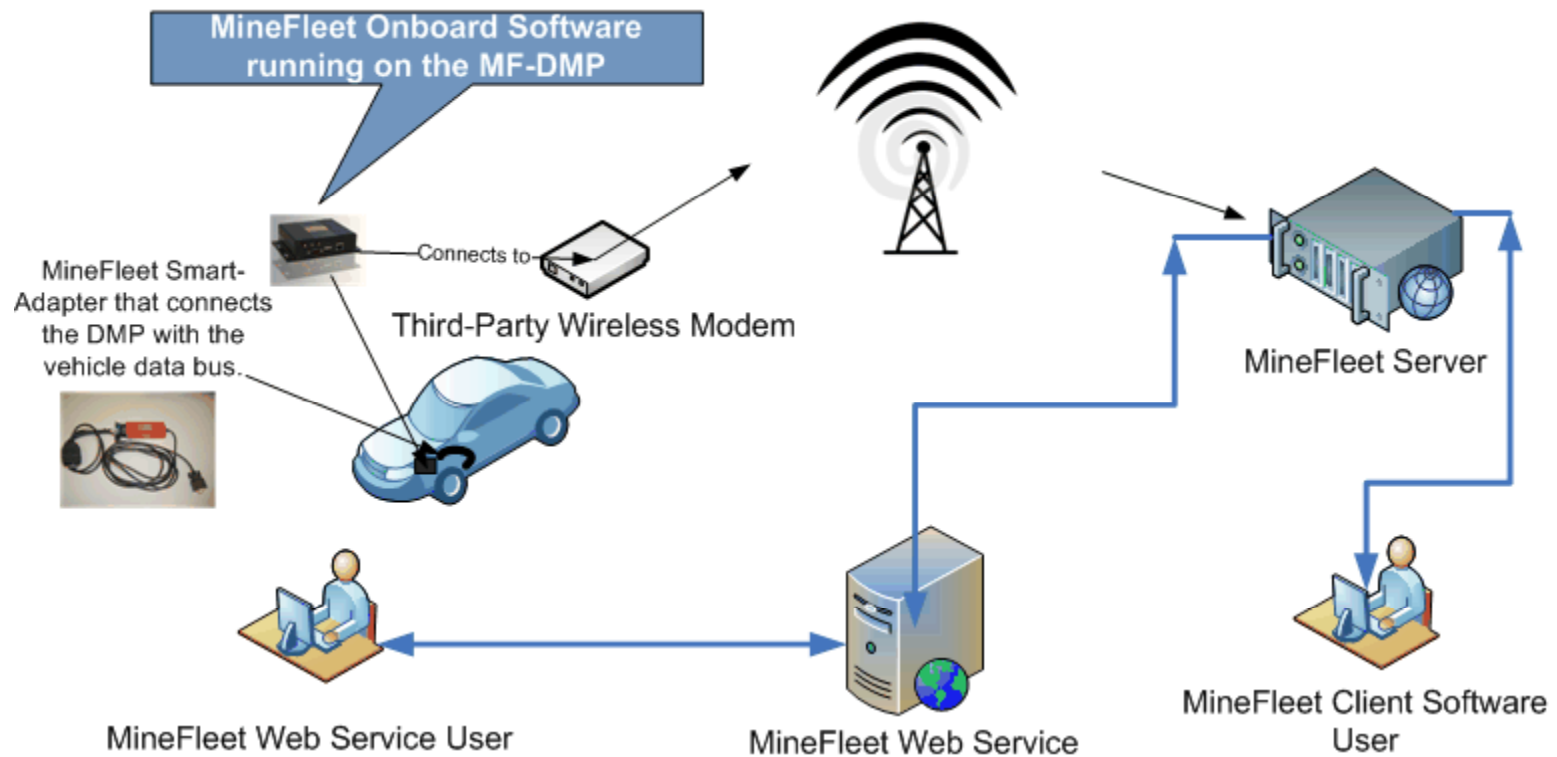
PURSUIT Web Service





**MineFleet®: Onboard Vehicle Performance
Data Mining System**

MineFleet Architecture



Need for MineFleet

- Billions of trucks and cars world-wide.
 - Poor fuel economy results from malfunctioning parts or bad driving
 - Mechanics inspect a vehicle only when there are some obvious drivability problem
 - Bad driving is expensive
 - Lack of vehicle behavior benchmarking tools--- poor depreciation analysis
 - Emerging need for "greener" vehicles

Reduce your fuel consumption

Breakdowns cost thousands of dollars

Bad driving costs money---fuel, brake shoe, insurance, lawsuits

Reduce your carbon footprint



MineFleet System

The screenshot displays the MineFleet desktop application with several overlapping windows:

- Diagnostic Tests Summary:** Shows an overall health status and lists 4 potential vehicle health problems.
- Cumulative Vehicle Data Analysis:** Lists various analysis options like Fuel Economy, Vehicle Health Tests, and Diagnostic Tests Summary.
- Short term fuel trim - Bank 2:** A histogram showing the distribution of fuel trim percentages, with a color-coded correlation scale from negatively to positively correlated.
- Vehicle List:** A table listing vehicles and drivers.

| Vehicles | Drivers |
|----------|---------|
| 3 | 3 |
| 3 | 3 |
| 3 | 3 |
| 3 | 3 |
| 3 | 3 |

MineFleet Server

The screenshot shows the MineFleet web services interface in a browser window. It features the Agnik logo and a 'Vehicle Health Trend Analysis' section. A pie chart displays the distribution of vehicle health problems. The interface includes navigation menus, search bars, and data selection options.

Vehicle Health Trend Analysis

This graph displays the vehicle health problems that have the most occurrences within the specified time span. You can click a slice of the pie chart to get more detailed information about the vehicle health problem(s).

For Fleet: Agnik Fleet | Based on: Number of Occurrences

Show Month-by-Month Analysis

Start Date: 7/15/2008 | End Date: 08/22/2008

| Problem Category | Percentage |
|---------------------|------------|
| Category 1 (Purple) | 42% |
| Category 2 (Red) | 18% |
| Category 3 (Green) | 12% |
| Category 4 (Blue) | 8% |
| Category 5 (Yellow) | 9% |
| Category 6 (Orange) | 7% |

MineFleet Web Services



MineFleet Onboard

The background image features a collage of automotive-related elements: a road with cars, a fuel pump nozzle, and a line graph with data points. The title 'Fuel Subsystem: Sample Attributes' is overlaid in white text on the left side of the image.

Fuel Subsystem: Sample Attributes

Fuel Subsystem

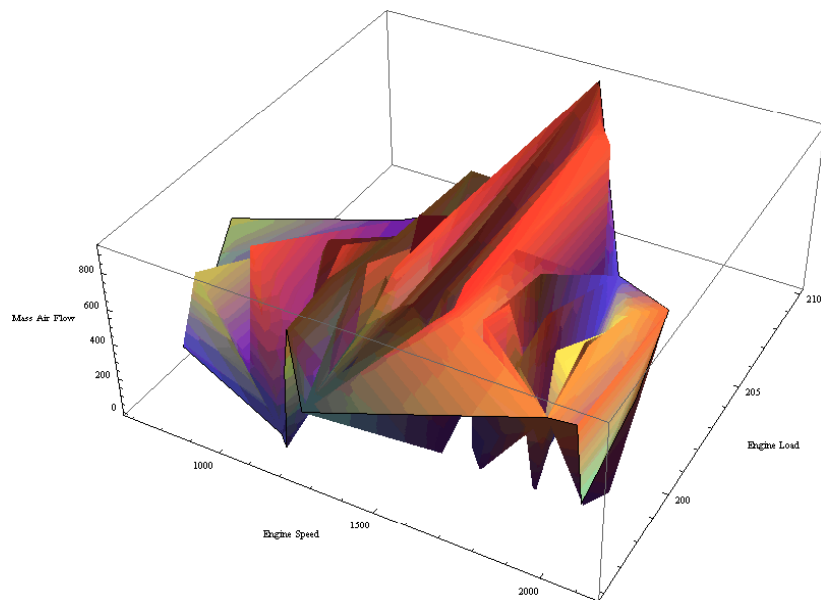
- Air Fuel Ratio
- Fuel Level Sensor (%)
- Fuel System Status Bank 1 [Categ. Attrib.]
- Oxygen Sensor Bank 1 Sensor 1 [mV]
- Oxygen Sensor Bank 1 Sensor 2 [mV]
- Oxygen Sensor Bank 2 Sensor 1 [mV]
- Oxygen Sensor Bank 2 Sensor 2 [mV]
- Long Term Fuel Trim Bank 1 [%]
- Short Term Fuel Trim Bank 1[%]
-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

Operating Condition

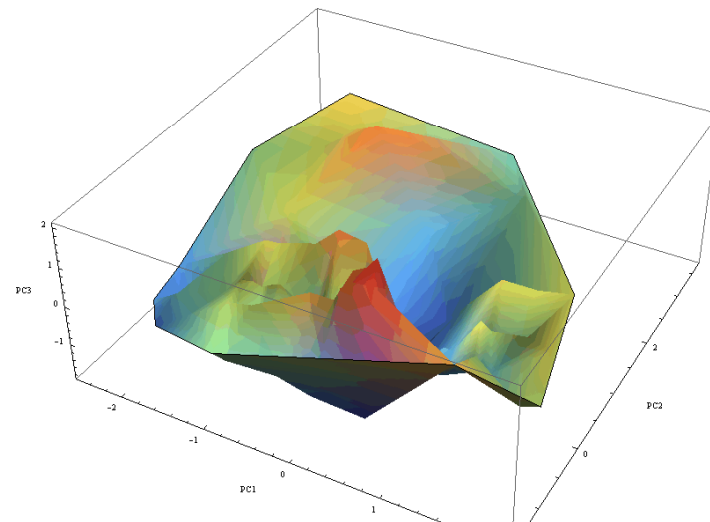
- Barometric Pressure
- Calculated Engine Load(%)
- Engine Coolant Temperature (°F)
- Engine Speed (RPM)
- Engine Torque
- Intake Air Temperature (IAT) (°F)
- Start Up Engine Coolant Temp. (°F)
- Start Up Intake Air Temperature (°F)
- Throttle Position Sensor (%)
- Throttle Position Sensor (degree)
-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

MineFleet for Advanced Onboard Data Analysis

- Advanced trend analysis, machine learning, data mining and anomaly detection algorithms for onboard statistical analysis and modeling.
- Minimizes wireless data transmission.



Variation of Mass Air Flow with respect to Engine Speed and Engine Load



Modeling through advanced engine analysis.



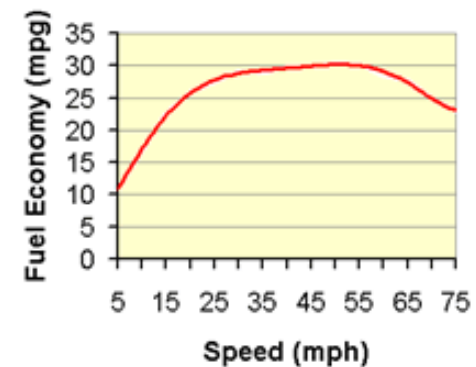
Find out Reasons Behind Poor Fuel Economy

Copyright 2009 Agnik, LLC



Data from EPA

- **Rapid Acceleration and Braking:** Aggressive driving (speeding, rapid acceleration and hard braking) wastes gas. It can lower your gas mileage by 33 percent at highway speeds and by 5 percent around town. You may save in between 5 to 33 percent in fuel economy by minimizing aggressive driving. (Savings of \$0.12-\$0.76/gallon)
- **Speeding:** Just by observing the speed limit you may save in between 7 to 23 percent in fuel economy. (Savings of \$0.16-\$0.53/gallon)
- **Idling:** Idling reduces the overall gas mileage; so minimize idling.



Fuel Economy: Impact of Driver Behavior

- Quantify the effect of driver behavior on fuel consumption and train drivers to prevent inefficient driving practices.

- Effect of speeding on fuel economy
- Effect of acceleration on fuel economy
- Effect of braking on fuel economy
- Effect of idling on fuel economy
- Many more....



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

Screen Shot: Fuel Consumption Summary Panel

M Analytic Visualization Browser - Van 37



Cumulative Vehicle Data Analysis

- Diagnostic Trouble Codes
- Fuel Economy
 - Summary
 - Historical Fuel Economy
 - Fuel Economy Prediction
 - Feature Histograms
 - Fuel Map
- Subsystem Parameter Interaction
- Vehicle Health Tests

Benchmark Analysis

- Compare With Benchmark

Shift Analysis

- Select Shift

Summary

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/sec²).

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.

[Calculate Potential Savings](#)



M MineFleet

M MineFleet

M Analytic Visualization ...

agnik.com Webmail - ...

Type to search

Windows taskbar icons including a search icon and a help icon.



Data from EPA

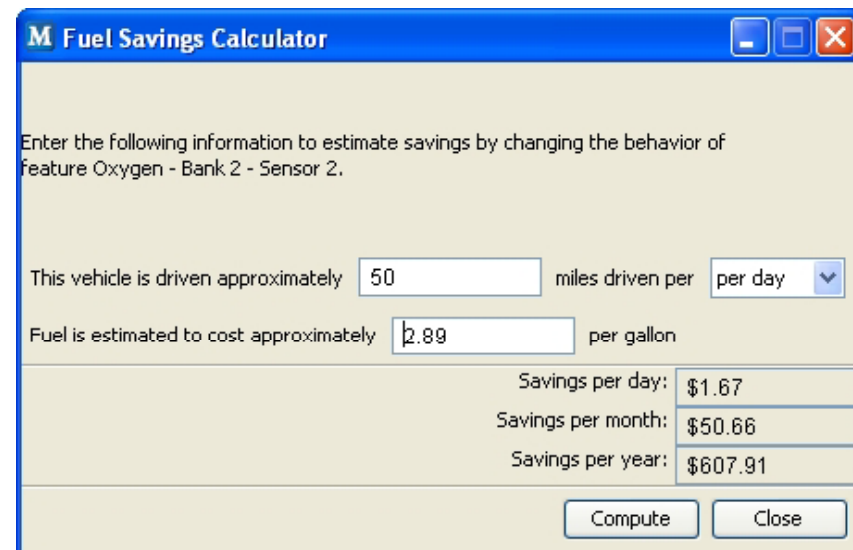
- **Faulty Oxygen Sensors:** Fixing a faulty oxygen sensor, can improve your fuel economy by as much as 40%. (Savings of \$0.9/gallon)
- **Basic Maintenance:** Fixing a car that is out of tune or has failed an emissions test can improve its gas mileage by an average of 4 percent. (Savings of \$0.09/gallon)
- **Excess Weight:** Removing excess weight may have considerable impact on the fuel economy. An extra 100 pounds in your vehicle could reduce your fuel economy by up to 2%. (Savings of \$0.02-\$0.05/gallon per 100 lbs)

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. For example, MineFleet can quantify how much your fuel economy is hurting because of a bad oxygen sensor.



M Fuel Savings Calculator

Enter the following information to estimate savings by changing the behavior of feature Oxygen - Bank 2 - Sensor 2.

This vehicle is driven approximately miles driven per

Fuel is estimated to cost approximately per gallon

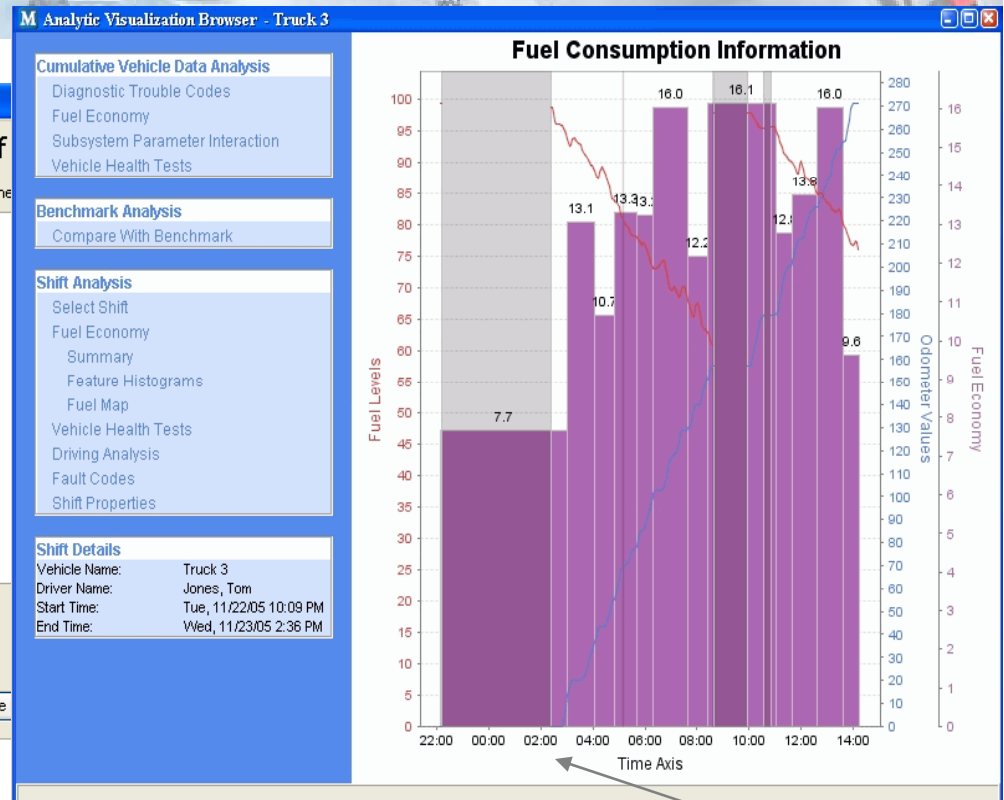
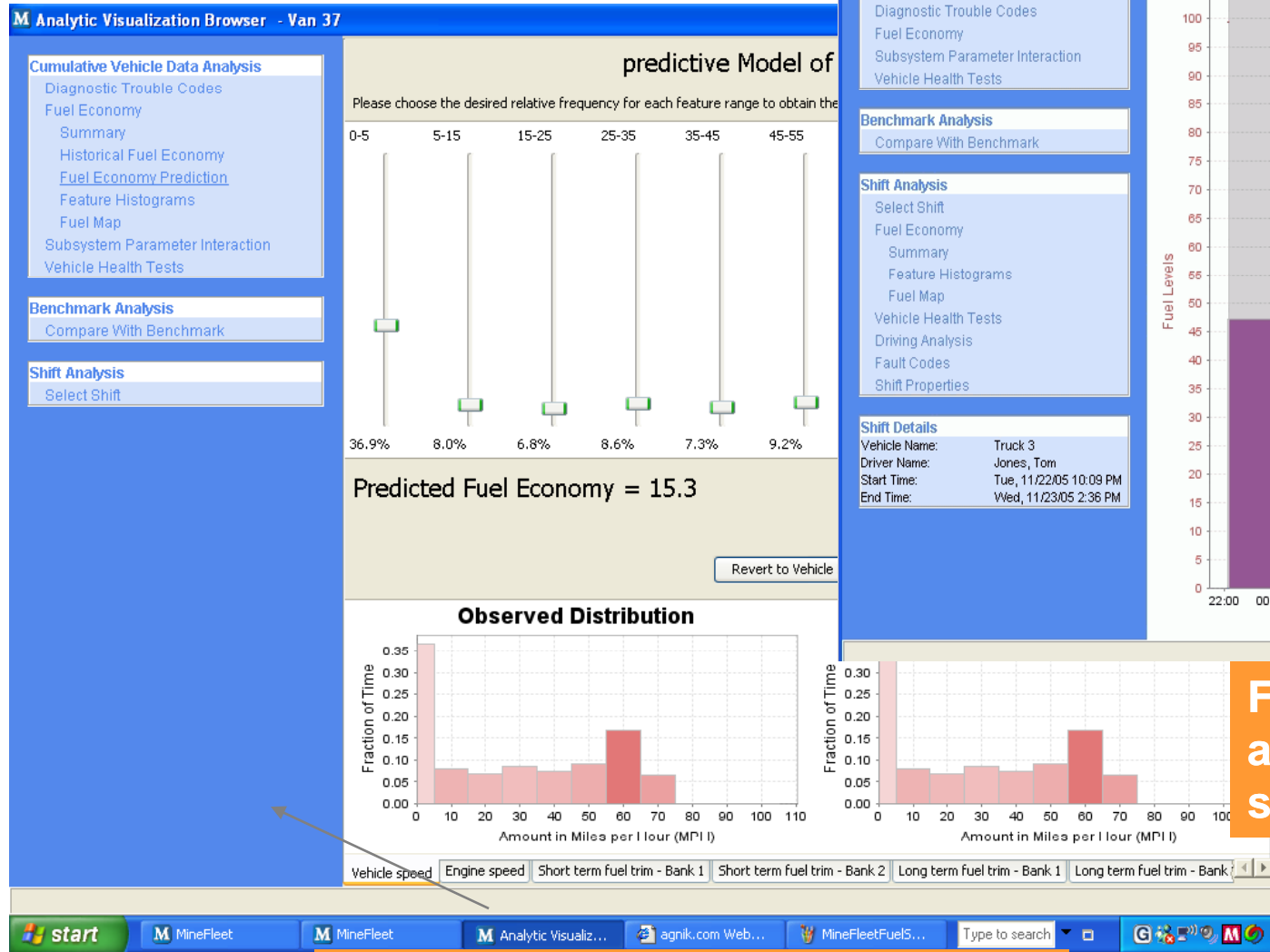
| | |
|--------------------|----------|
| Savings per day: | \$1.67 |
| Savings per month: | \$50.66 |
| Savings per year: | \$607.91 |



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.

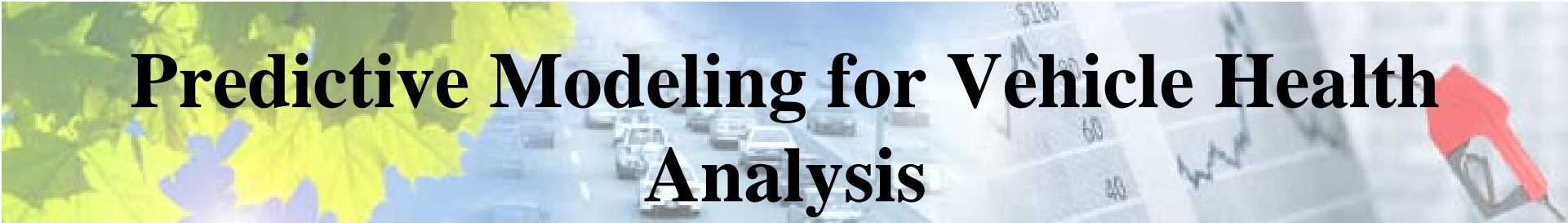
Screen Shot: Advanced Predictive Fuel Consumption Optimization



Fuel economy analysis for a specific shift of the vehicle

Predictive model for fuel economy





Predictive Modeling for Vehicle Health Analysis

- Detect problems using model and data driven fault detection tests well before DTC code shows up.
- Auto-generate alerts when MineFleet detects unusual behavior of a subsystem and access the data producing this behavior.
- Manage vehicle data and performance history.
- Track maintenance and vehicle performance history.

Screen Shots: Vehicle Health Management

The screenshot displays the 'Analytic Visualization Browser - Truck 3' interface. On the left, a navigation pane lists various analysis categories: Cumulative Vehicle Data Analysis, Benchmark Analysis, and Shift Analysis. The main content area is titled 'Long Term Fuel Related C' and shows a 'Test Failed' status. The test description explains that as part of the combustion formula, fuel delivery is also the term wear during normal engine operation. The recommendation section suggests checking fuel pressure, injectors, fuel pressure regulator, clogged evaporative emissions system, oxygen sensor contamination, and clogged air filter.

Cumulative Vehicle Data Analysis
Diagnostic Trouble Codes
Fuel Economy
Subsystem Parameter Interaction
Vehicle Health Tests
Summary
Vehicle Health Tests History

Benchmark Analysis
Compare With Benchmark

Shift Analysis
Select Shift
Fuel Economy
Vehicle Health Tests
Summary
Long Term Fuel Related
Combustion Efficiency
Air Intake Volume Inconsistency
Thermal Event Detector
Quantitative Fuel Management
Monitoring, Fuel 9
Vehicle System Temperature
Monitor
Transmission Lubricating
Systems
Monitor
Driving Analysis
Fault Codes
Shift Properties

Long Term Fuel Related C

Test Description:
As part of the combustion formula, fuel delivery is also the term wear during normal engine operation. While there is the limitations of its adaptability, often significant collateral code will set. This test is designed to monitor changes well monitoring these changes within the fuel delivery portion the collateral damage through early detection of the dete

Long term fuel trim out of range

Test Failed

Recommendation:
MineFleet recommends checking fuel pressure (too high), injectors for leakage, leaking fuel pressure regulator, clogged evaporative emissions system, oxygen sensor contamination and clogged air filter most likely causes when fuel trim fails high. MineFleet recommends checking for clogged injector(s), ignition system components, fuel pressure (low), or water intrusion on oxygen sensor as possible c

Summary

Overall Health Status
There are 2 potential vehicle health problems.

Vehicle Health Problems
The most recent tests that failed were:
1. Vehicle System Temperature Monitor
2. Air Intake Volume Inconsistency

Please click on Vehicle Health Tests History for more information.

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

Detailed description of a specific test that the vehicle passed

Screen Shot: Fuel Economy Benchmarking

M Analytic Visualization Browser - Van 37



Benchmark Fuel Economy Summary

Cumulative Vehicle Data Analysis

- Diagnostic Trouble Codes
- Fuel Economy
 - Summary
 - Historical Fuel Economy
 - Fuel Economy Prediction
 - Feature Histograms
 - Fuel Map
- Subsystem Parameter Interaction
- Vehicle Health Tests

Benchmark Analysis

- Compare With Benchmark
- Close Benchmark

Shift Analysis

- Select Shift

Benchmark Vehicle

| | |
|---------------|--------|
| Vehicle Name: | Van 17 |
| Year: | 1998 |
| Make: | Ford |
| Model: | F150 |

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).

Oxygen - Bank 1 - Sensor 2

According to the Benchmark Vehicle, the feature Oxygen - Bank 1 - Sensor 2 appears to be operating outside of the normal parameters, which is effecting fuel economy.

It appears that if this vehicle were operating within the specified benchmark behavior, it would have a positive effect on the fuel economy. The fuel economy would rise from 15.7 to 16.7

[Calculate Potential Savings](#)

Calculated engine load

According to the Benchmark Vehicle, the feature Calculated engine load appears to be operating outside of the normal parameters, which is effecting fuel economy.

It appears that if this vehicle were operating within the specified benchmark behavior, it would have a positive effect on the fuel economy. The fuel economy would rise from 15.7 to 16.3

[Calculate Potential Savings](#)

Engine speed

According to the Benchmark Vehicle, the feature Engine speed appears to be operating outside of the normal parameters, which is effecting fuel economy.

It appears that if this vehicle were operating within the specified benchmark behavior, it would have a positive effect on the fuel economy. The fuel economy would rise from 15.7 to 16.0

[Calculate Potential Savings](#)



Summary of the fuel economy benchmarking analysis





Onboard Emissions Analysis in MineFleet

- Quantitative assessment of vehicle emissions, including CO₂, CO, NO_x, and hydrocarbons.
- MineFleet Green Scoring™
- How the emission patterns are correlated with environmental and vehicle performance parameters

A horizontal banner image featuring a collage of elements: yellow and green leaves on the left, a road with cars in the center, and a red gas pump nozzle on the right. The text 'MineFleet Web Portal' is overlaid in the center in a bold, black, serif font.

MineFleet Web Portal


Copyright 2009 Agnik, LLC



MineFleet Main Page - Microsoft Internet Explorer


File Edit View Favorites Tools Help

Address <http://www.services-agnik.com/MainPage.aspx>



- Home
- Refresh
- Purchase Client Keys
- Trend Analysis
- Fleet Performance
- Reports
- Emission Analytics
- Technical Support
- Logout

Agnik Reseller Demo (227)



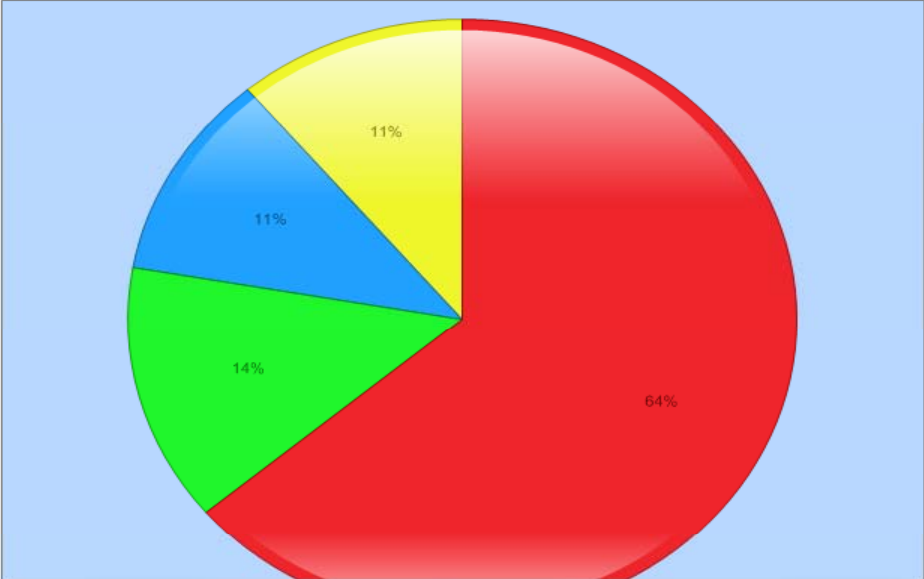
Driver Behavior

This graph displays the driver behavior problems that have the most occurrences within the specified time span. You can click a slice of the pie chart to get more detailed information about the driver behavior problem(s).

For Fleet: Agnik Demo Fleet Based on: Number of Occurrences

Show Month-by-Month Analysis

Start Date: 2/11/2009 End Date: 05/06/2009



| Behavior Problem | Percentage |
|--------------------|------------|
| Problem 1 (Red) | 64% |
| Problem 2 (Green) | 14% |
| Problem 3 (Blue) | 11% |
| Problem 4 (Yellow) | 11% |

CO Monitoring

MineFleet Main Page - Microsoft Internet Explorer

File Edit View Favorites Tools Help



Address http://www.services-agnik.com/minefleetemissions/MainPage.aspx



Home

Refresh

Create Product Key

Purchase Client Keys

Trend Analysis

Fleet Performance

Reports

Pending Request

Create Reseller

View Tickets

Emission Analytics

Technical Support

Logout

Agnik



Vehicle Emission

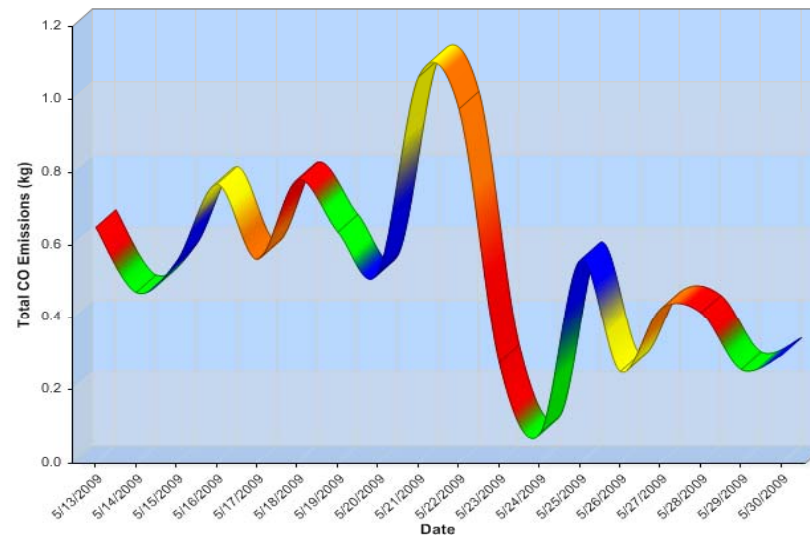
This graph displays the vehicle emission trends for the selected fleet within the specified time span.

for Fleet

Based on

Start Date:

End Date:



Done

Internet

Copyright 2009 Agnik, LLC





Vehicles and Green House Gases (GHG)

- Transportation activities are responsible for approximately 25% to 30% of total U.S. GHG emissions
- On-highway commercial truck market accounting for over 45% of transportation GHG

Emissions & Airlines Industry

- A Boeing 747 uses approximately 1 gallon of fuel every second.
- A flight from Washington DC to Los Angeles emits about 726 pounds of CO₂.
- Aircrafts generate large volume of data even from short flights (e.g. 10MB from an hour long flight depending upon the type of aircraft)





Summary

- Distributed data mining
 - Decade-long literature offering many synchronous and asynchronous distributed data mining algorithms
 - Distributed anomaly detection from vehicle performance data streams
 - Correctness, Efficiency, Scalability: Centralized vs DDM



Announcement

- National Science Foundation Data Mining Summit on Energy Crisis, Greenhouse Emission, and Transportation Challenges

<http://www.kd2u.org/NGDM09/>
Baltimore, Oct 1—Oct 3, 2009



Resources

- DDMWiki (<http://www.umbc.edu/ddm/wiki/>)
- DDMBib (<http://www.cs.umbc.edu/~hillol/DDMBIB/>)
- Recently formed nonprofit organization:
Association for Knowledge Discovery in Distributed and Ubiquitous (KDD&U) Environments
(www.kd2u.org)