Large ScaleVehicle Performance Monitoring in Distributed Environments

University of Maryland Baltimore County

&

Agnik, LLC

www.cs.umbc.edu/~hillol

http://www.agnik.com

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

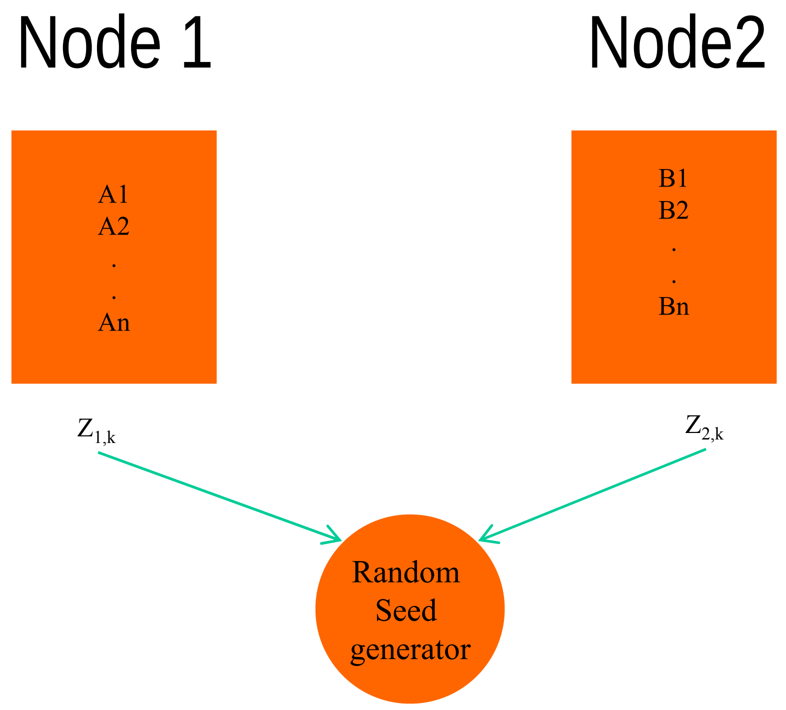
DDM Algorithms: Some Examples

- 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 $Z_{1,k}$
  - $Z_{1k}=A1.J_1+..+An.J_n$
  - $J_i \in \{+1,-1\}$ 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

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 $a_1 \ldots\ldots a_n$ Є R, Site B has $b_1 \ldots\ldots b_n$ Є R
- Sites must compute

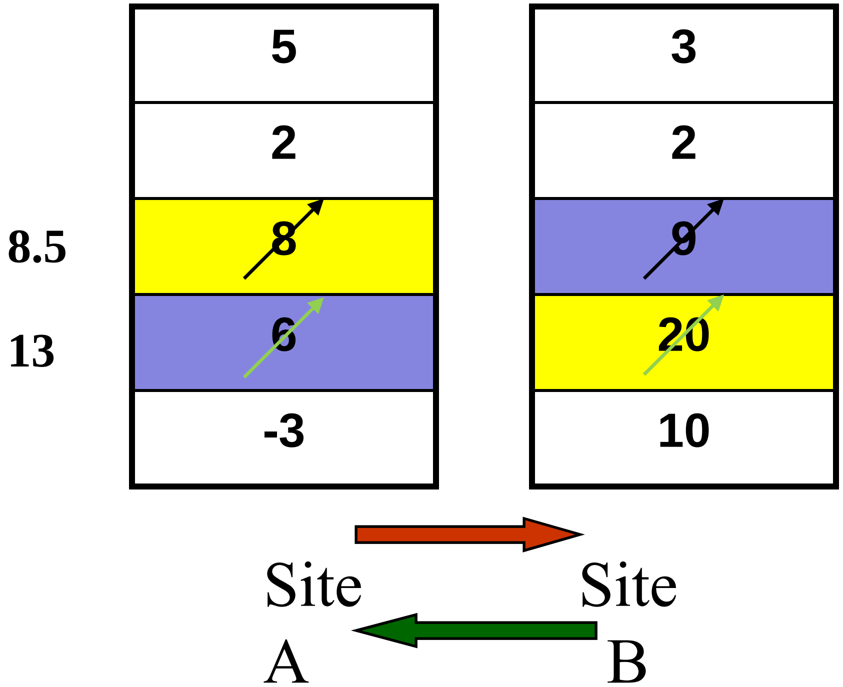
$$i^* = \text{Argmax}\{c_1 = (a_1 + b_1)^2, \ldots, c_n = (a_n + b_n)^2\}$$

Distributed Max Sum Square

The Algorithm

- (1) Site A selects max of a_{iA} ‘s and sends $< a_{iA}, iA >$ to Site B
  (2) Site B selects max of b_{iB} ‘s and sends $< b_{iB}, iB >$ to Site A
- Site A receives the message and replies with $<a_{iB}>$. Site B replies with $<b_{iA}>$
- 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

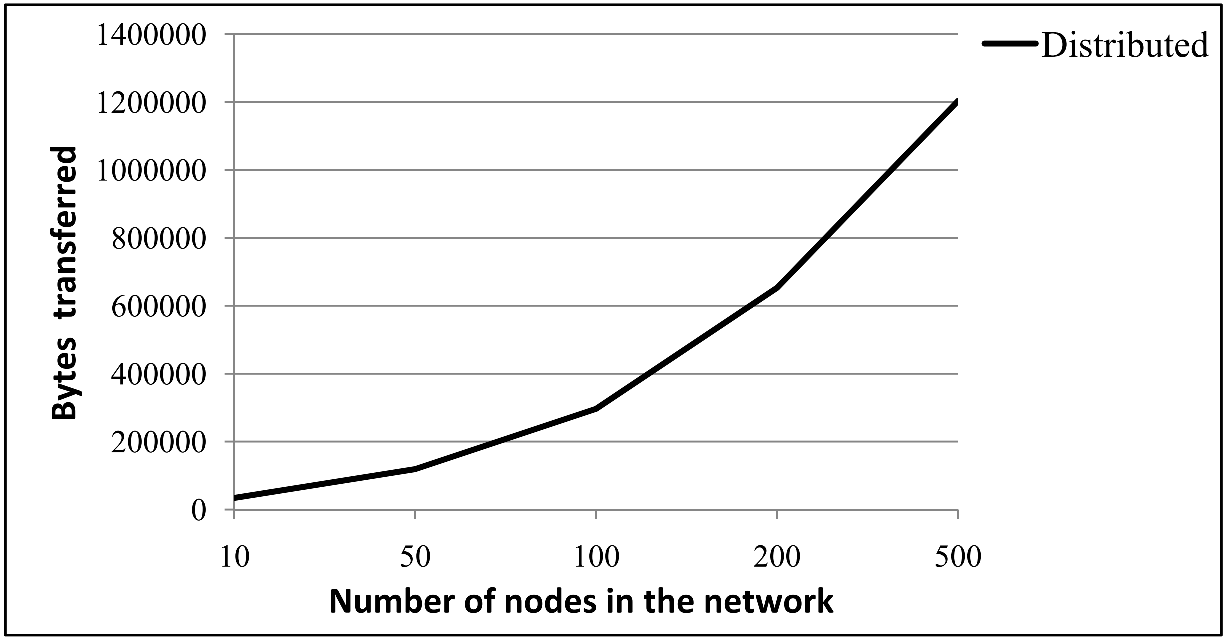
Distributed Classifier Learning & Outlier Detection

- 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

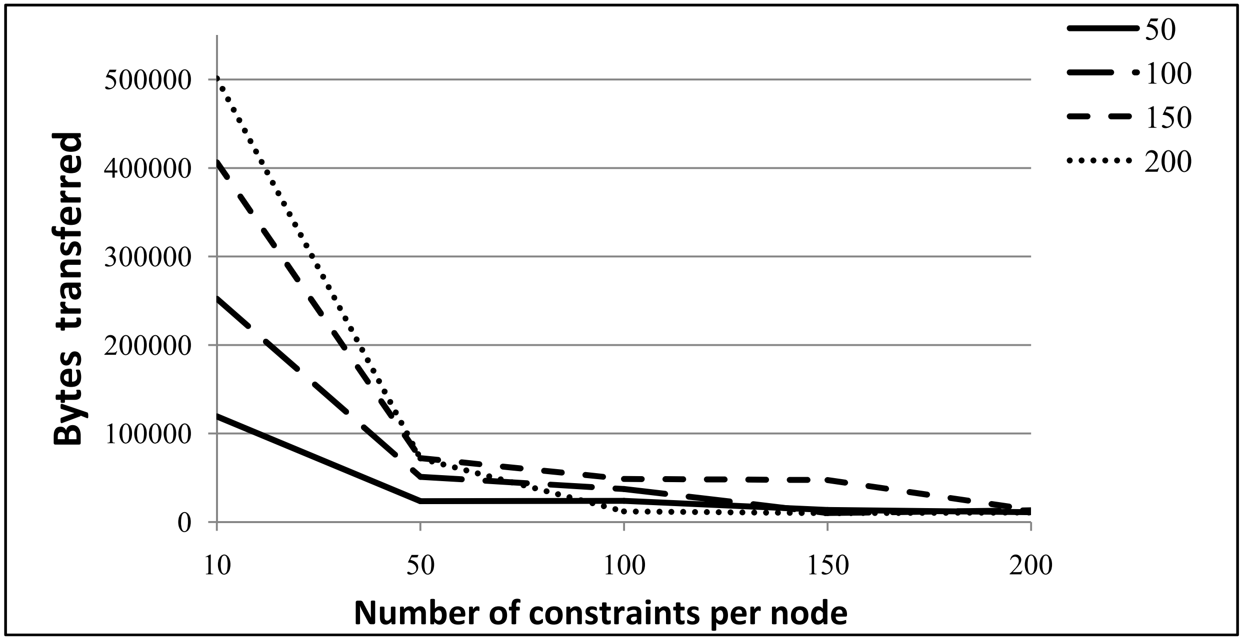
- Simplex rely upon pivot computations
- Pivot computation can be reduced to distributed min computation

Communication Cost vs. Network Size



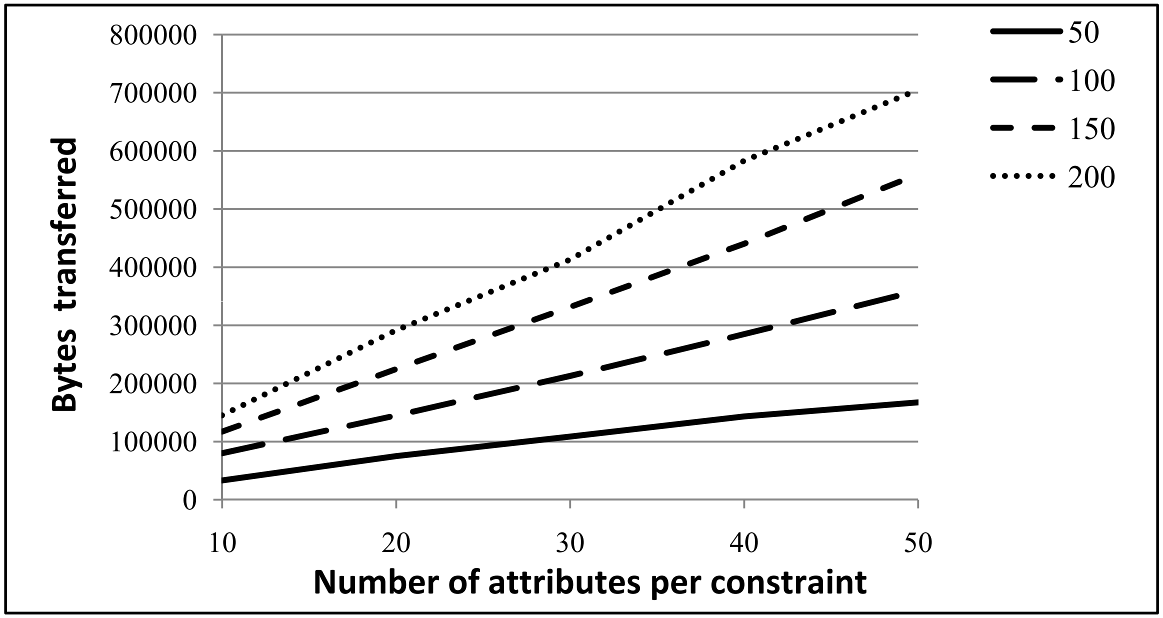
- Number of nodes in the network is varied from 10 to 500 nodes
- Number of variables in a constraint equation is kept constant at 35

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)*

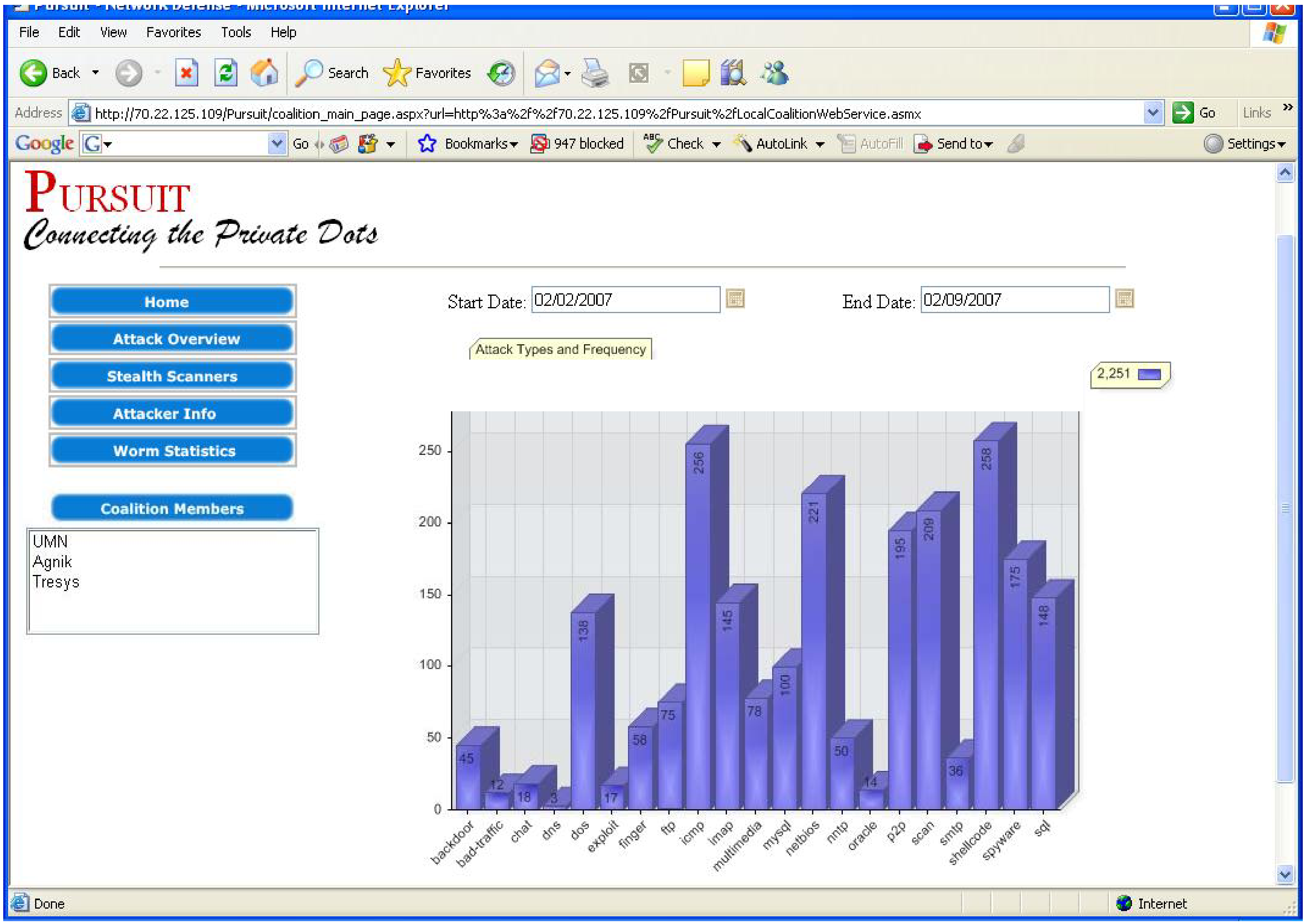
Private & Secure Data Mining from Multi-Party Distributed Data

- 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

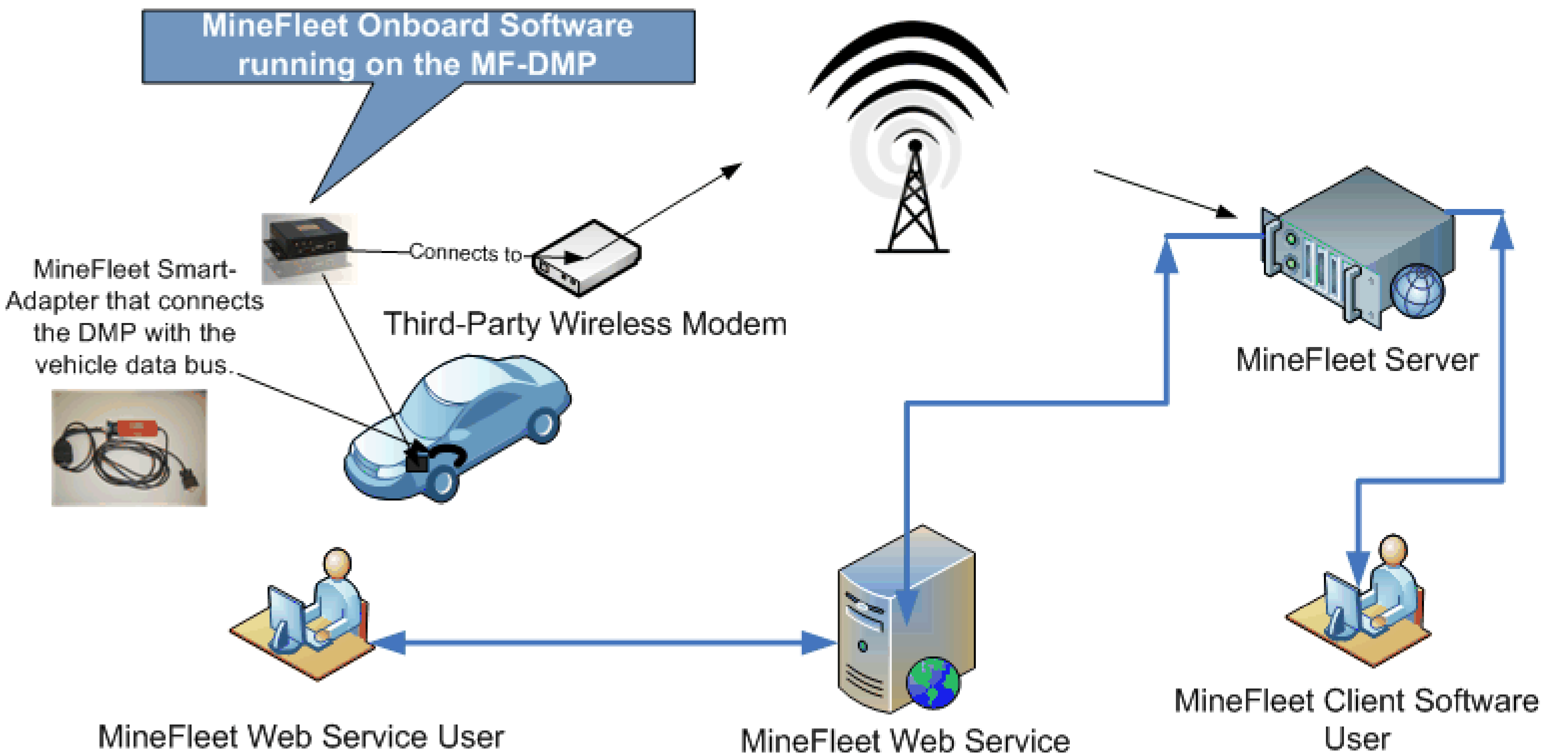
- Need to have your own sensor such as SNORT, MINDS
- Download PURSUIT plug-in for the sensor and install
- PURSUIT plug-in offers
  - A stand-alone interface for processing your alerts from the sensor and cross-domain analysis
  - Web account for detailed cross-domain statistics
  - Optional distributed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

PURSUIT Web Service



MineFleet®:Onboard Vehicle Performance Data Mining System

MineFleet Architecture



Agnik

Need for MineFleet

Reduce your fuel consumption

- 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

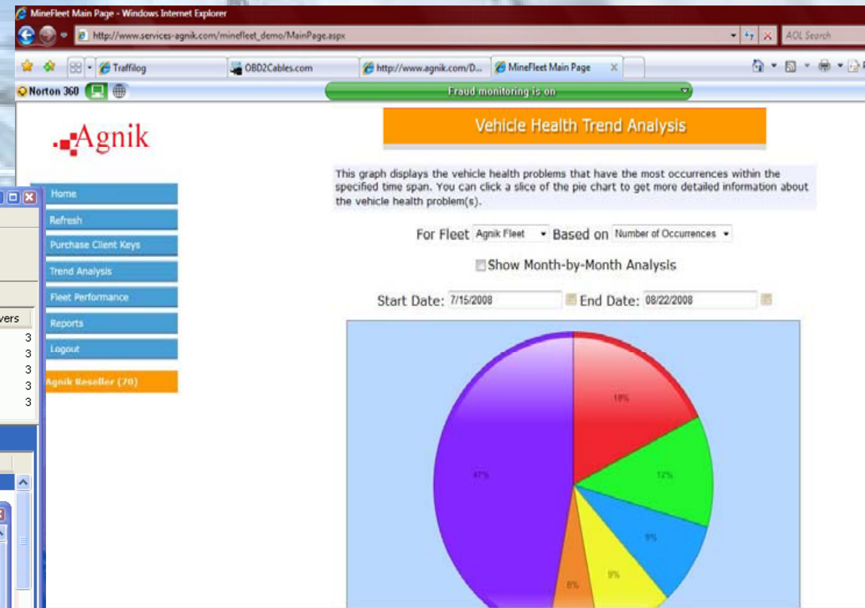
Breakdowns cost thousands of dollars

Bad driving costs money---fuel, brake shoe, insurance, la w-suits

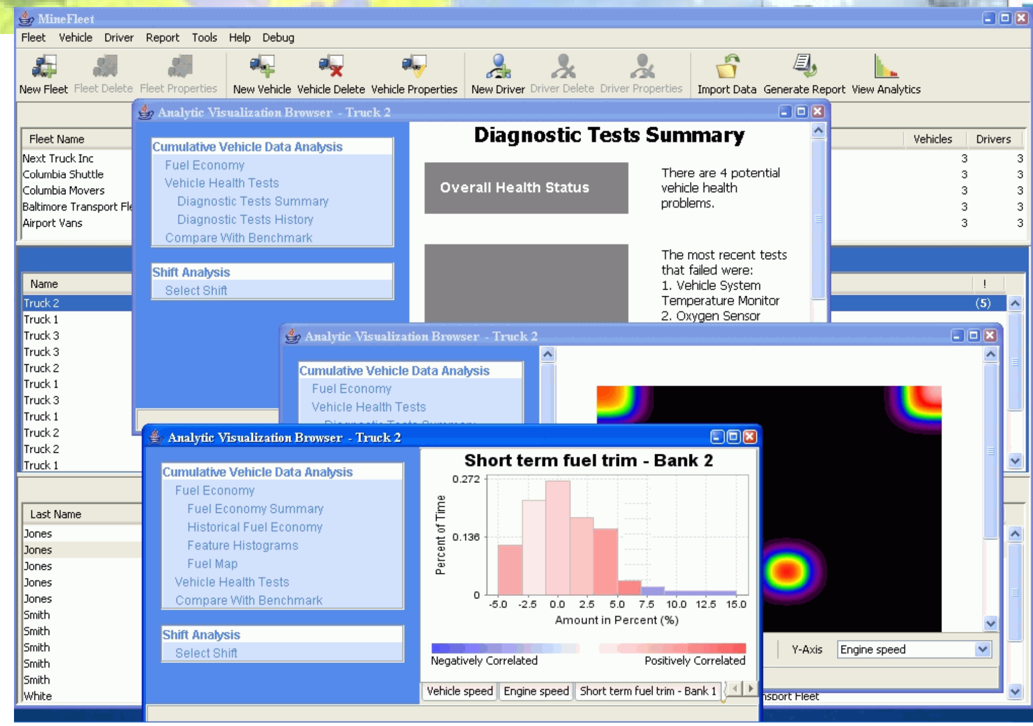
Reduce your carbon footprint

Agnik

MineFleet System



MineFleet Web Services



MineFleet Server

MineFleet Onboard

Fuel Subsystem: Sample Attributes

Fuel Subsystem

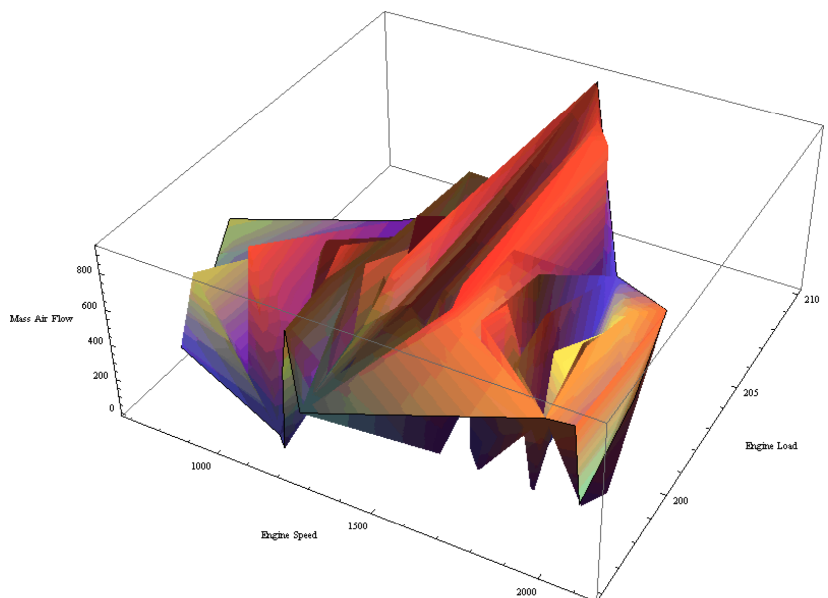
- Air Fuel Ratio
- Fuel Level Sensor (%)
- Fuel System Status Bank 1 [Categ. Attrib.]
- Oxygen Sensor Bank 1 Sensor 1 [mV]
- Oxygen Sensor Bank 1 Sensor 2 [mV]
- Oxygen Sensor Bank 2 Sensor 1 [mV]
- Oxygen Sensor Bank 2 Sensor 2 [mV]
- Long Term Fuel Trim Bank 1 [%]
- Short Term Fuel Trim Bank 1[%]
-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

Operating Condition

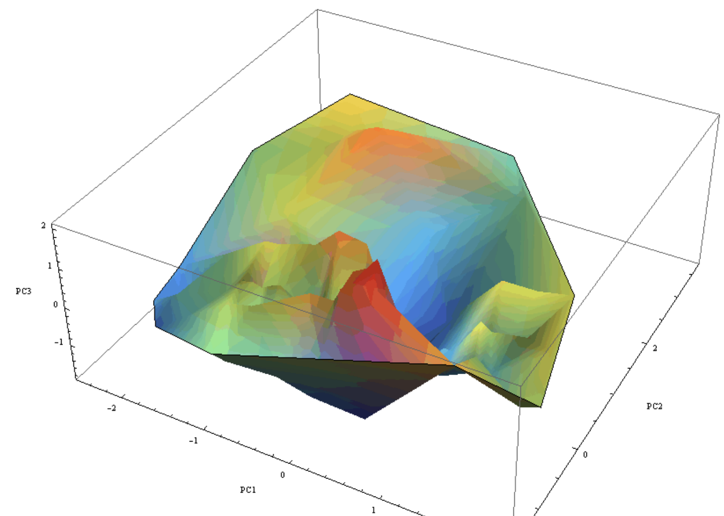
- Barometric Pressure
- Calculated Engine Load(%)
- Engine Coolant Temperature (°F)
- Engine Speed (RPM)
- Engine Torque
- Intake Air Temperature (IAT) (°F)
- Start Up Engine Coolant Temp. (°F)
- Start Up Intake Air Temperature (°F)
- Throttle Position Sensor (%)
- Throttle Position Sensor (degree)
-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

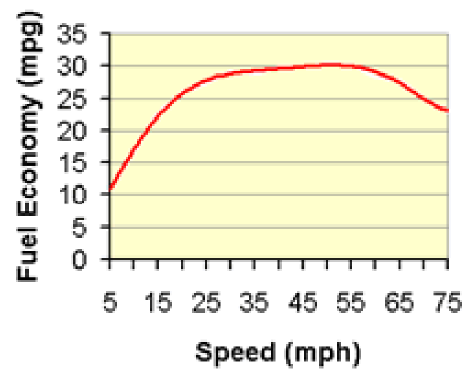
Agnik

Find out Reasons Behind Poor Fuel Economy

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.

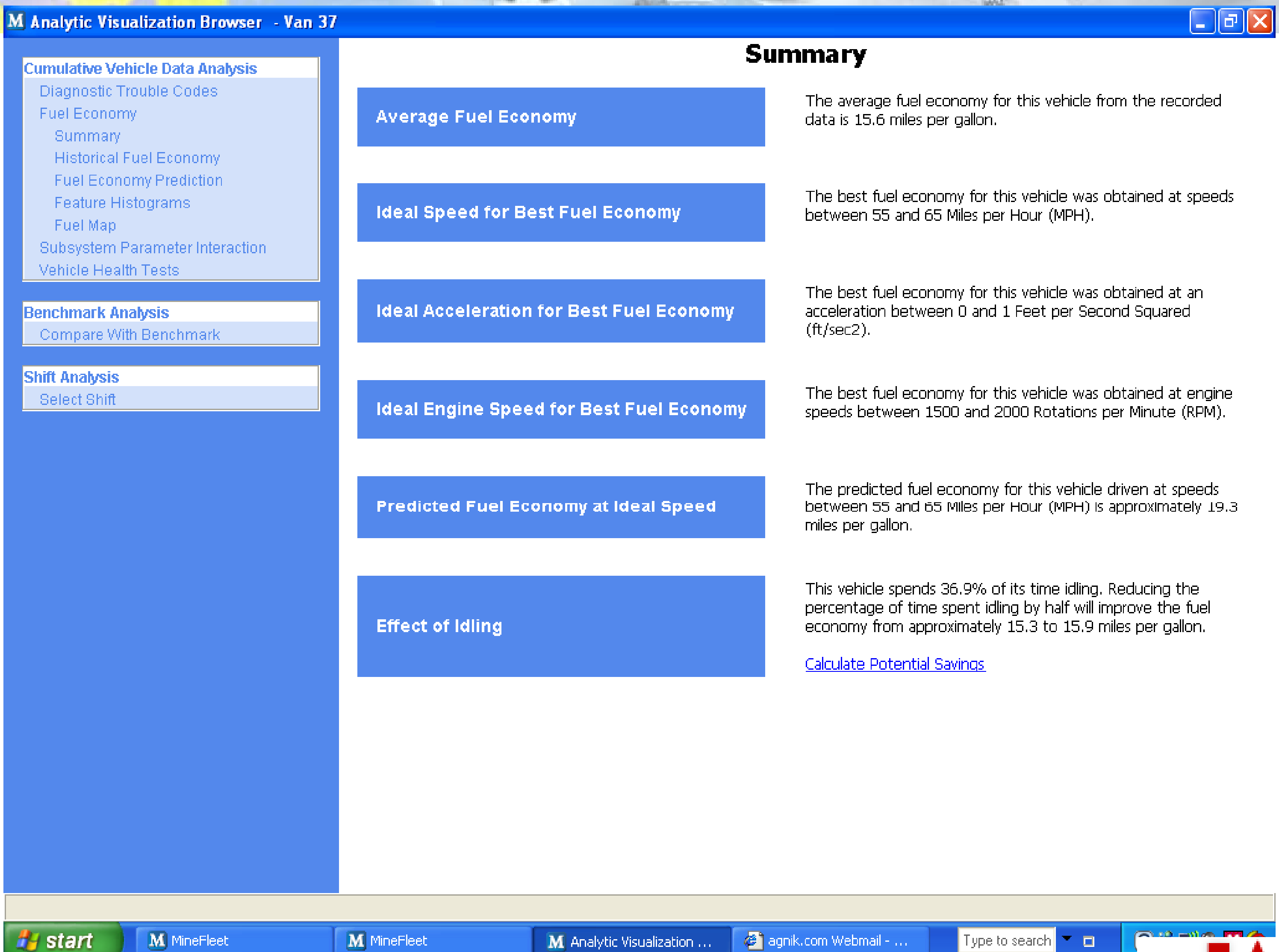
Agnik

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



Agnik

Data from EPA

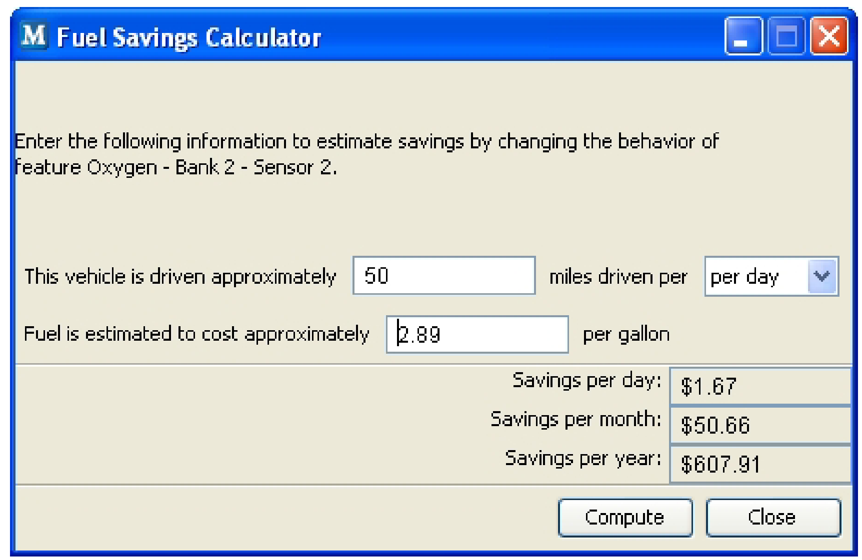
- **Faulty Oxygen Sensors:** Fixing a faulty oxygen sensor, can improve your fuel economy by as much as 40%. (Savings of $0.9/gallon)

- **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.

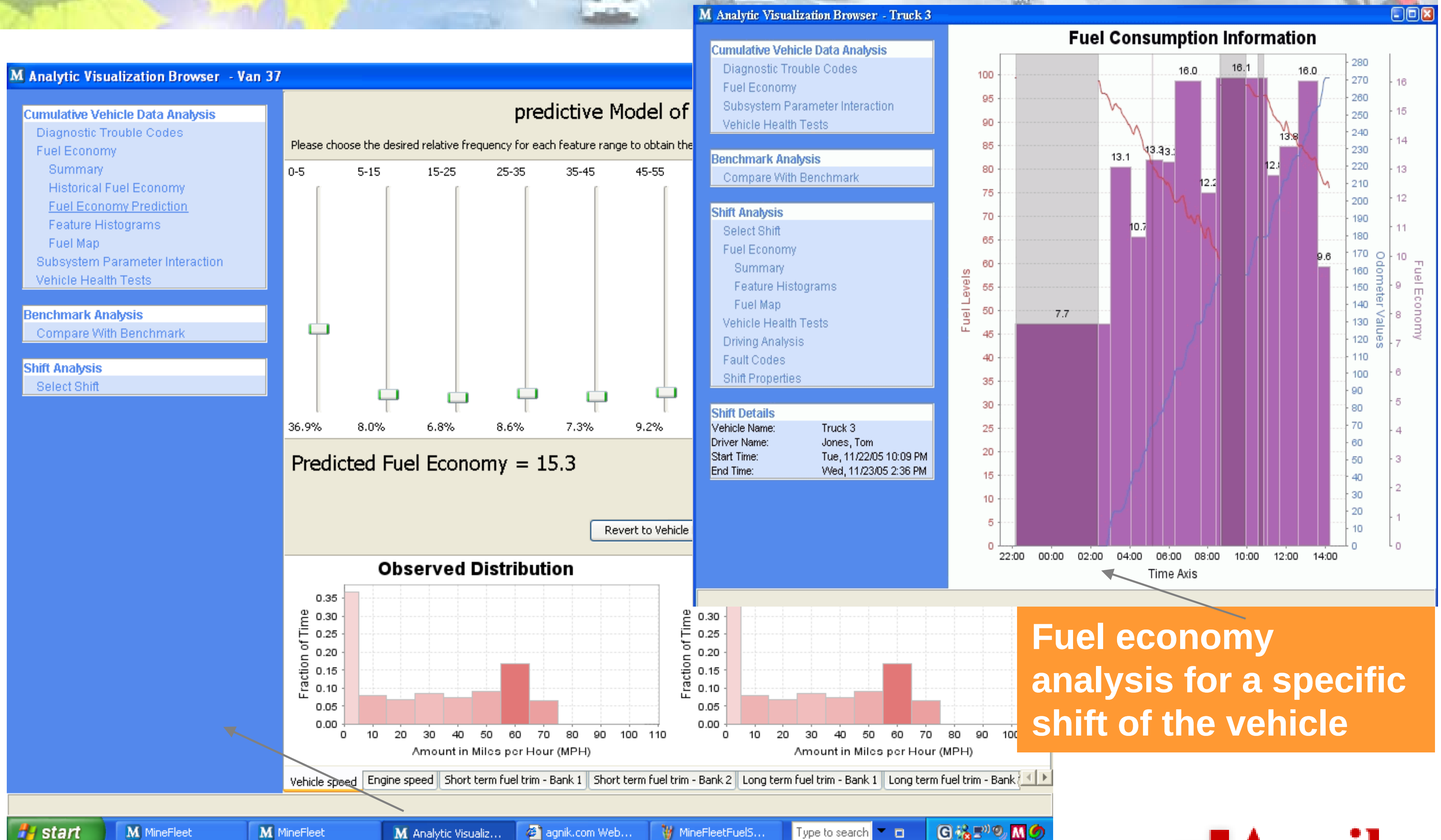


Fuel Economy: Predictive Modeling

- 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



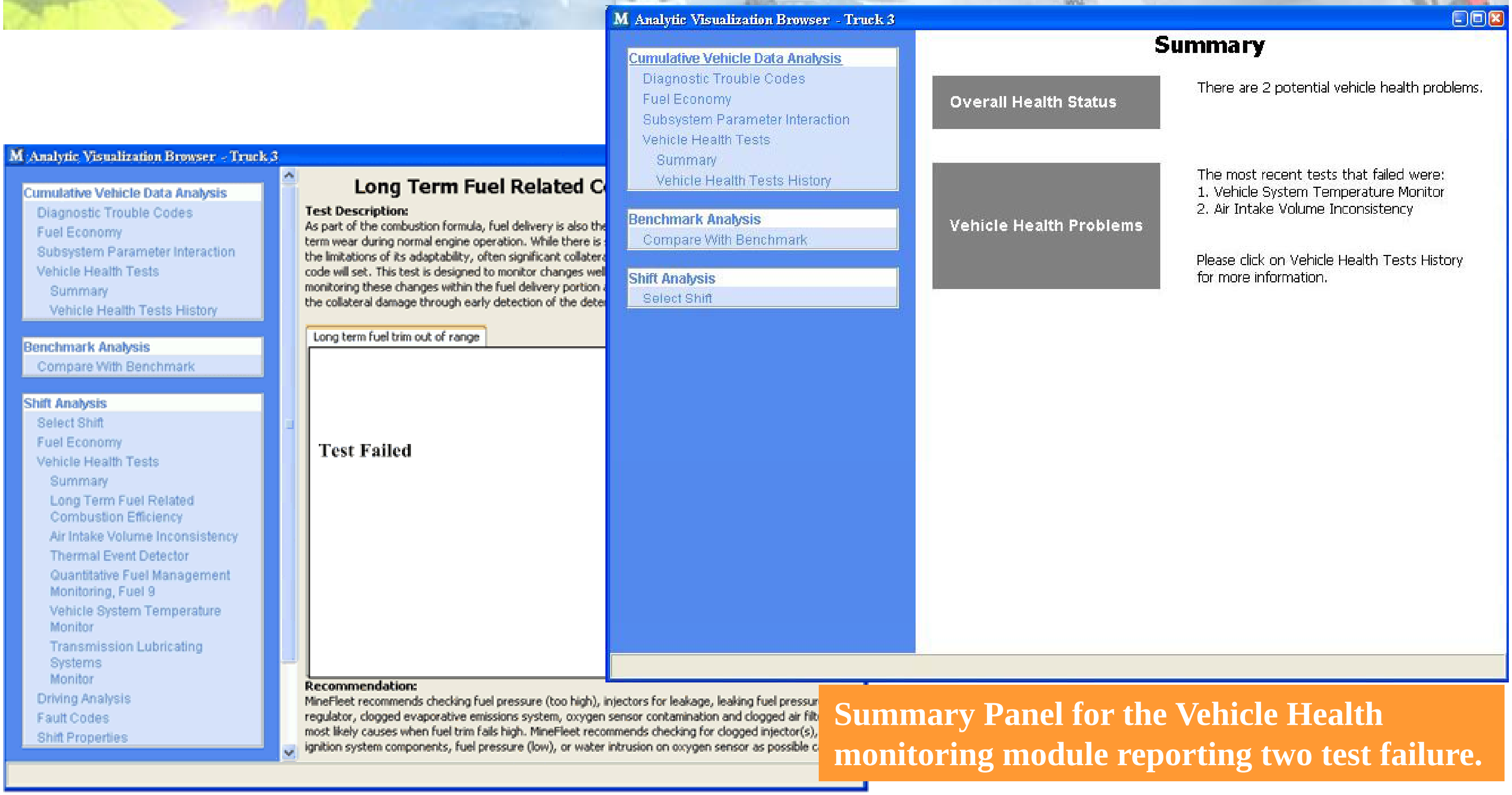
Fuel economy analysis for a specific shift of the vehicle

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

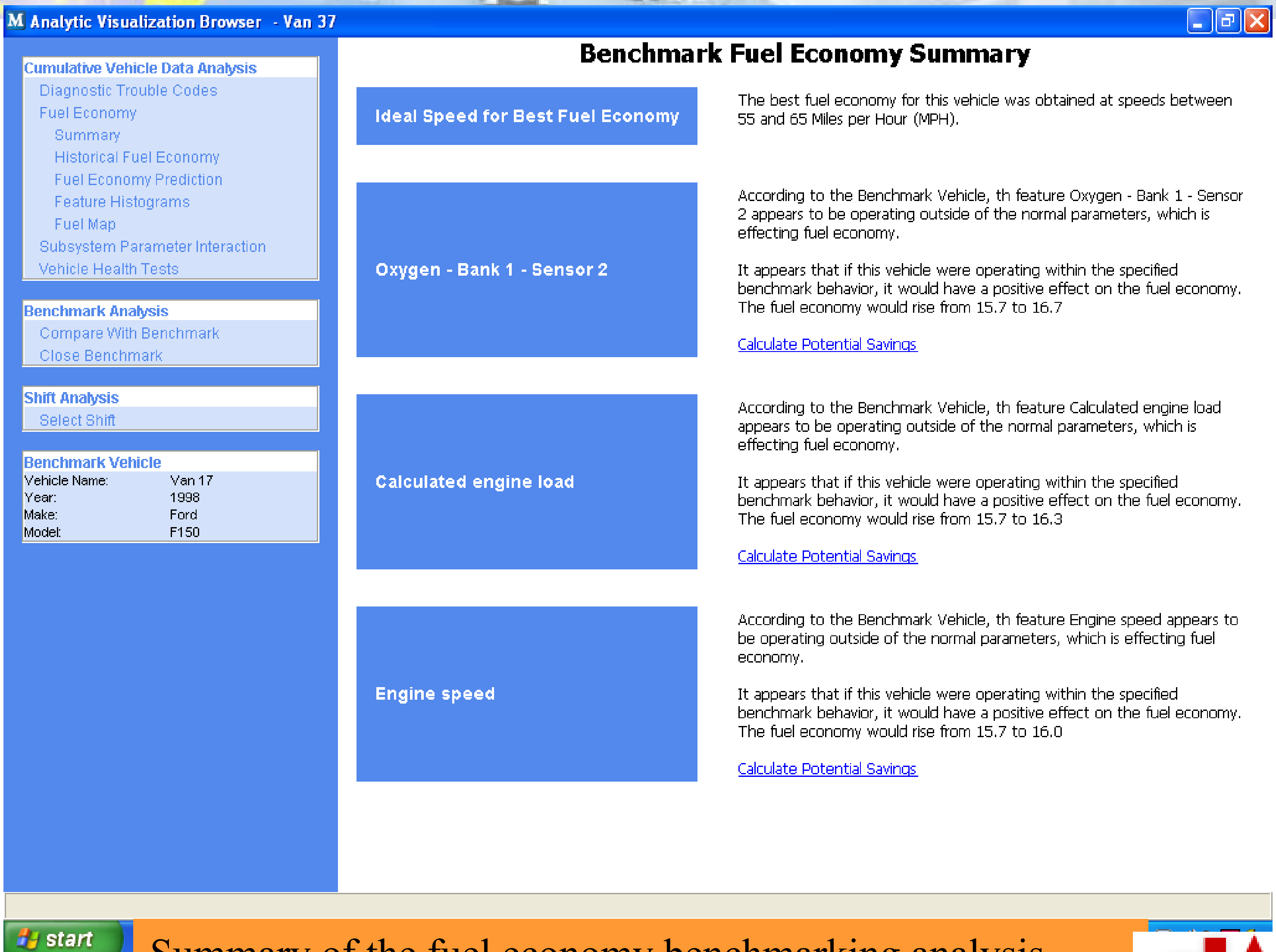


Summary Panel for the Vehicle Health monitoring module reporting two test failure.

Detailed description of a specific test that the vehicle passed

Agnik

Screen Shot: Fuel Economy Benchmarking



Summary of the fuel economy benchmarking analysis

Agnik

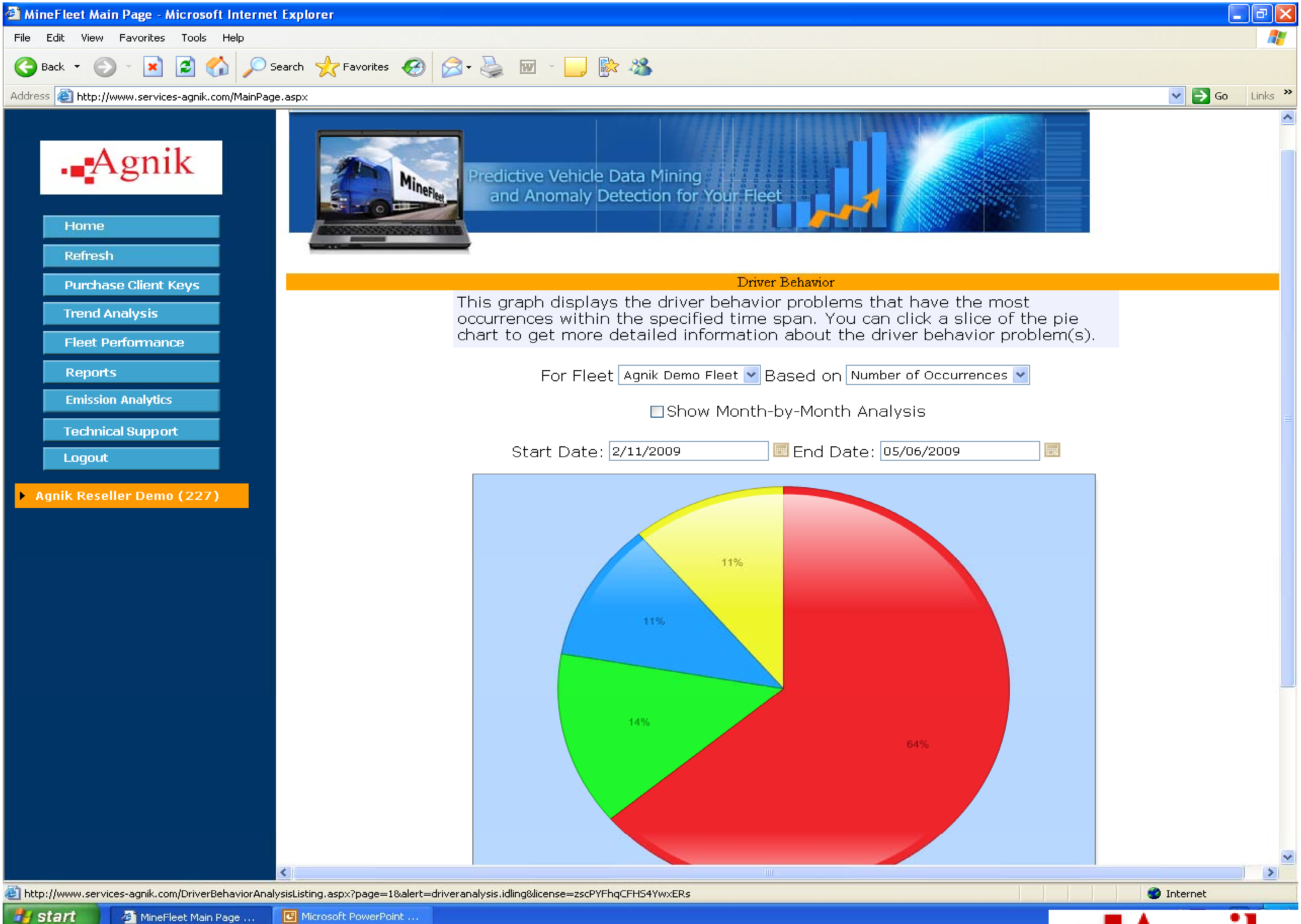
Onboard Emissions Analysis in MineFleet

- Quantitative assessment of vehicle emissions, including CO_2, CO, NOx, and hydrocarbons.
- MineFleet Green Scoring™
- How the emission patterns are correlated with environmental and vehicle performance parameters

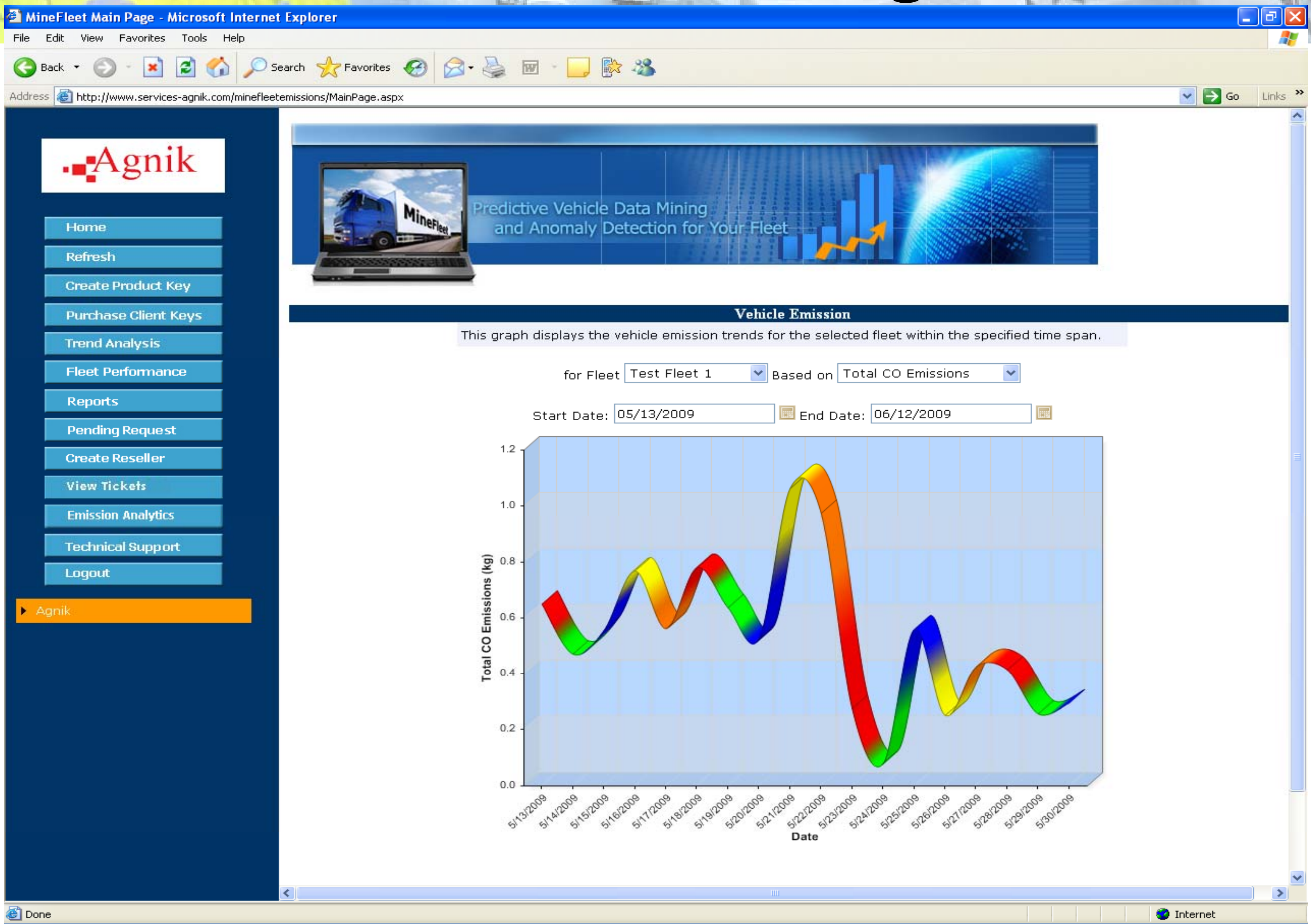
Agnik

MineFleet Web Portal

Agnik


MineFleet Main Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.services-agnik.com/MainPage.aspx
Go Links
Agnik
Home
Refresh
Purchase Client Keys
Trend Analysis
Fleet Performance
Reports
Emission Analytics
Technical Support
Logout
Agnik Reseller Demo (227)
MineFleet
Predictive Vehicle Data Mining and Anomaly Detection for Your Fleet
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
11%
11%
14%
64%
http://www.services-agnik.com/DriverBehaviorAnalysisListing.aspx?page=1&alert=driveranalysis.idling&license=zscPYFhqCFHS4YwxERs
Internet
start
MineFleet Main Page ...
Microsoft PowerPoint ...

CO Monitoring



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

Agnik

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 CO_2.
- 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)