Caterpillar Big Data Infrastructure

Big Data, Data Analytics, and Machine Learning
Caterpillar is the world’s leading manufacturer of construction and mining equipment, industrial diesel engines and gas turbines, and diesel-electric locomotives.
Solutions

Autonomy and Operator Assistance

Autonomous Haul Trucks

Non-Line of Sight
Remote Semi-Autonomy

Operator Assistance

Machine Learning on Advanced Sensor Data
Why Do We Need a Big Data Infrastructure?

Example: Machine Learning Flow
We were spending too much time on ground truth and managing training and testing data.

**Example: Machine Learning Flow**

1. **Data Pre-Processing**
2. **Separate Training & Test Data**
3. **Ground Truth Data**
4. **Determine Initial Machine Learning Algorithms to Compare**
5. **Train Algorithms on Training Data**
6. **Compare Algorithms on Test Data**
7. **Select Algorithm**
8. **Verify Algorithm**
9. **Determine Data to Acquire**
CatBigDat – Field Data Collection
CatBigDat – Web Based Ground Truth Tagging
CatBigDat – Ground Truth Metadata Database
CatBigDat – Engineering Interface Leverages Power of MATLAB
Completely Flexible and Modifiable Ground Truth Label Hierarchy - Vehicle
Completely Flexible and Modifiable Ground Truth Label Hierarchy - Personnel
General Additional Fields - Pick Lists

- Environmental Lighting
  - Sunny Day - Full day data, dawn to dusk on clear sunny day with mixed lighting (shadows and bright sunlight)
  - Cloudy Day - Full day data, dawn to dusk on cloudy day
  - Low Light
  - Night w/ Lights - Night data with vehicle lighting
  - Night w/ Lights and Incidental - Night data with vehicle and incidental lighting

- Background Environment (Construction Building, Construction Highway, Mine Surface, Commercial, Residential, Urban, Rural)

- Location (Indoor, Outdoor)

- Airborne obscurants (Dust, Fog, Smoke)

- Weather (Raining, Snowing)

- Ground Conditions (Mud/Dirt, Partial Snow, Majority Snow, On-Road, Off-Road, vegetation, gravel)

- Quality of Focus (Good, Poor, Lens Occlusion, Lens Damage)
Example Queries w/ Example Results

- Standing, un-occluded people
- Crouching, un-occluded people
- Close range, occluded people
- Negative Data (e.g. Non-People)
- Hydraulic Excavator, Side View
- Hydraulic Excavator, Rear View
- Wheel Loader, Bucket in Air
Automatic Labeling of Data
Tight integration with MATLAB Classification Learner App

- Simple queries into Caterpillar labeled data to import multi-class positive and negative data for training.
- Tight integration with MATLAB Machine Learning Backend (Classification Learner and Command Line)
Integration with Auto-Coding Tools
And 3\textsuperscript{rd} Party Machine Learning

- HDL Coder
- Caffe: Deep learning framework
- VIVADO: HLx Editions
- SDS\textsuperscript{SoC}: Environment
Using MATLAB for Continuous Improvement in our Big Data, Data Analytics, and Machine/Deep Learning Infrastructure

Because it is MATLAB, development time is short
Future Direction for the Infrastructure

Continuous Efficiency Improvement Feedback

Collect and Ground Truth Data → Auto Gather New Ground Truth → Run New Training with Pre-defined Queries on HPC → Calculate Performance Statistics of New Classifiers → Review and Select Classifier Performance → Auto Generate Code and Download to Embedded Platform → Evaluate Performance in the Field

Make it Even Easier to Find Best Classifiers to Solve a Given Problem - More Science, Less Art
Conclusions

• Developed big data and machine/deep learning infrastructure
• Web based ground truth interface
• Automatic ground-truth -- limits need for human supervision, reducing development time
• Database for storing and querying meta-data
• Engineering interface with tight integration with MATLAB products for learning, visualization, verification
• Code generation - direct to embedded real-time platforms
• Scalable in number of users, amount of data, and compute power
Thank You!

Amine El Helou

Gary Gunterman

Arvind Hosagrahara

Steve Kuznicki

Brett Shoelson

Lisa Crosier

Joe Forcash

Larry Mianzo

Dan Troniak