



Sensing Data Quality in Sensor-Based Data - A Department of Transportation Perspective

**Bureau of Labor Statistics
Data Quality for Statistical Products Workshop
December 1, 2017**

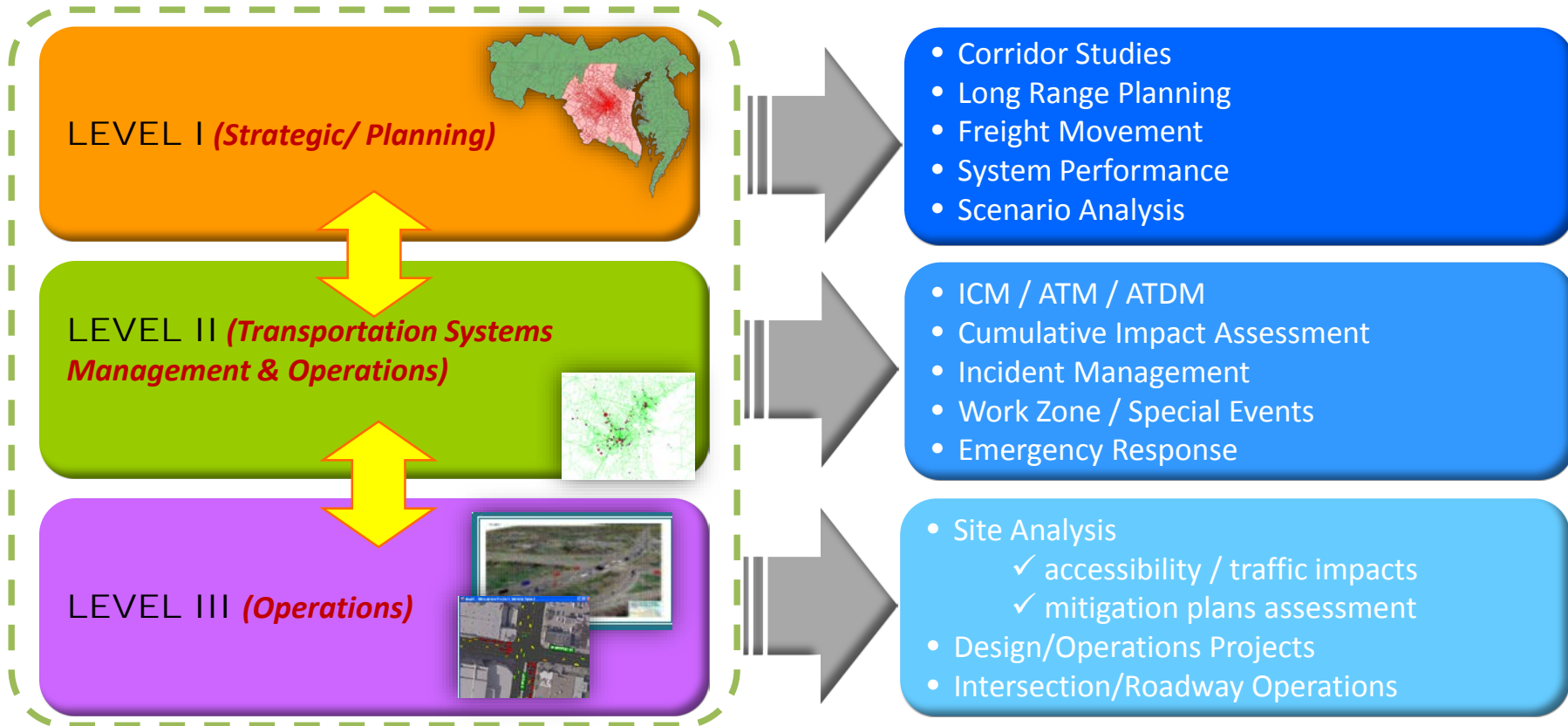
**Subrat Mahapatra
Maryland DOT State Highway Administration**

21st Century Transportation Trends

- Changing Customer Needs and Expectations
- Focus on Operations, Efficiency & Reliability
- Freight Movement/ Economy
- Technological Innovations (CV/AV, Ridesharing Apps)
- Transportation/ Environment/ Economy/ Health Linkages
- Performance Management & Communicating Performance
- Big Data Innovations

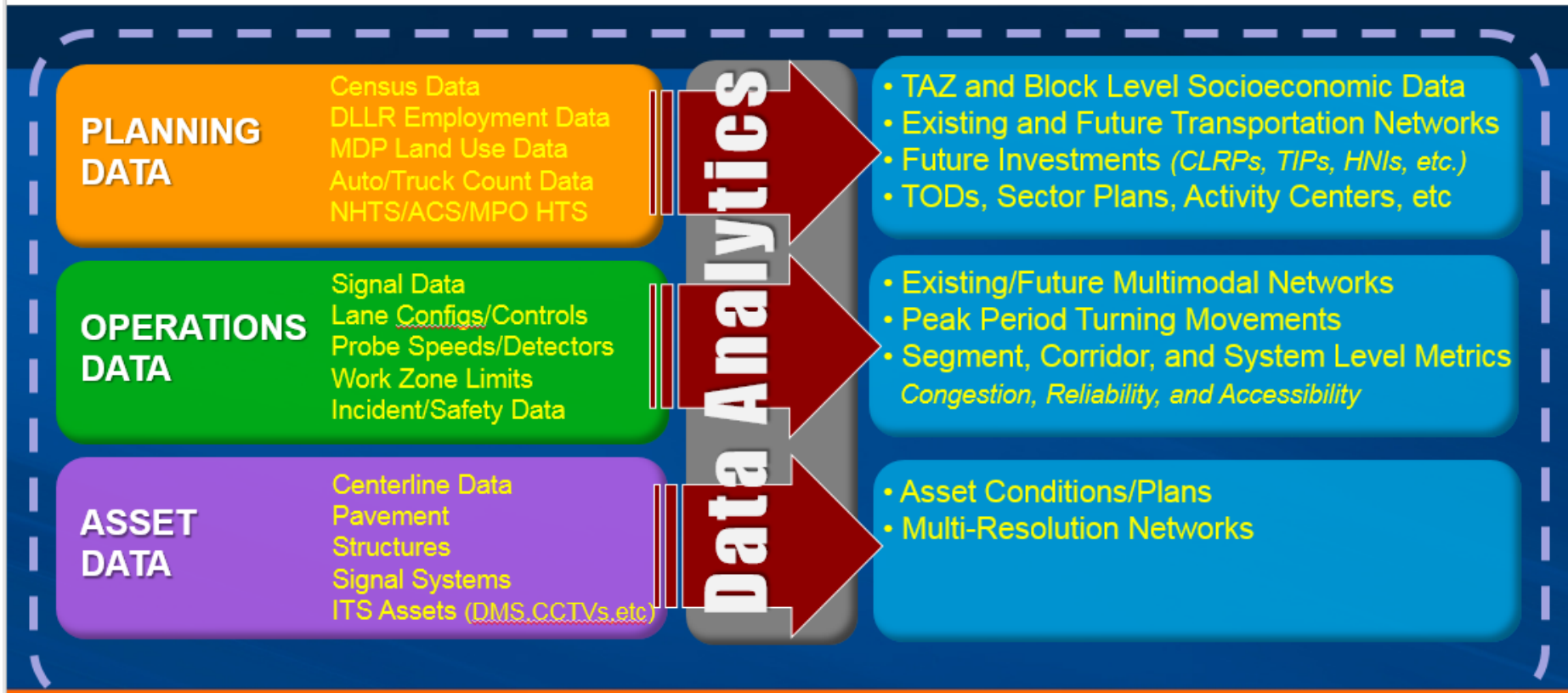


Transportation Decision-making Context



ROLE OF GOOD QUALITY DATA IS CRITICAL

Data Shaping Transportation Decision-making

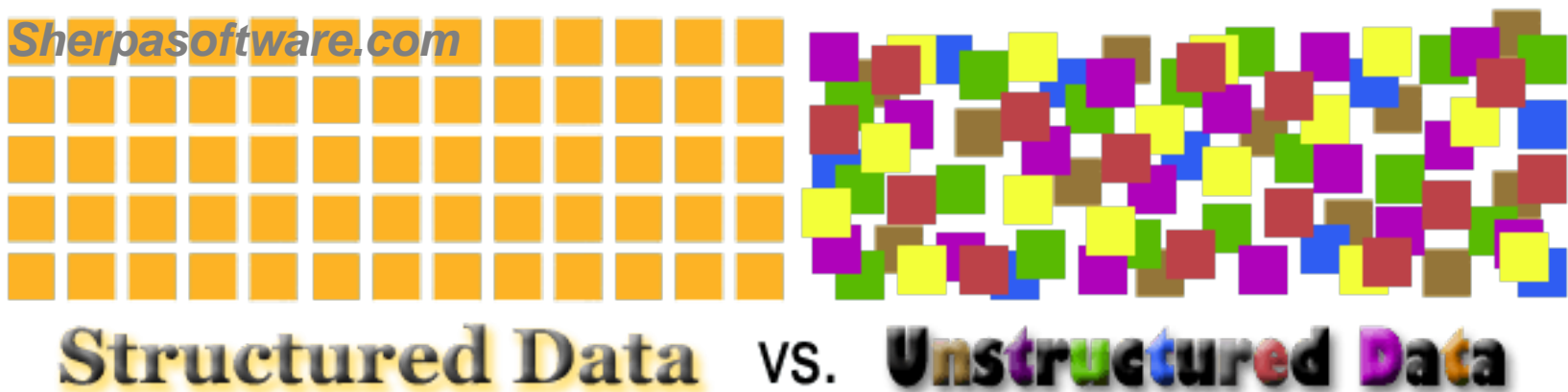


DATA GOVERNANCE PRINCIPLES

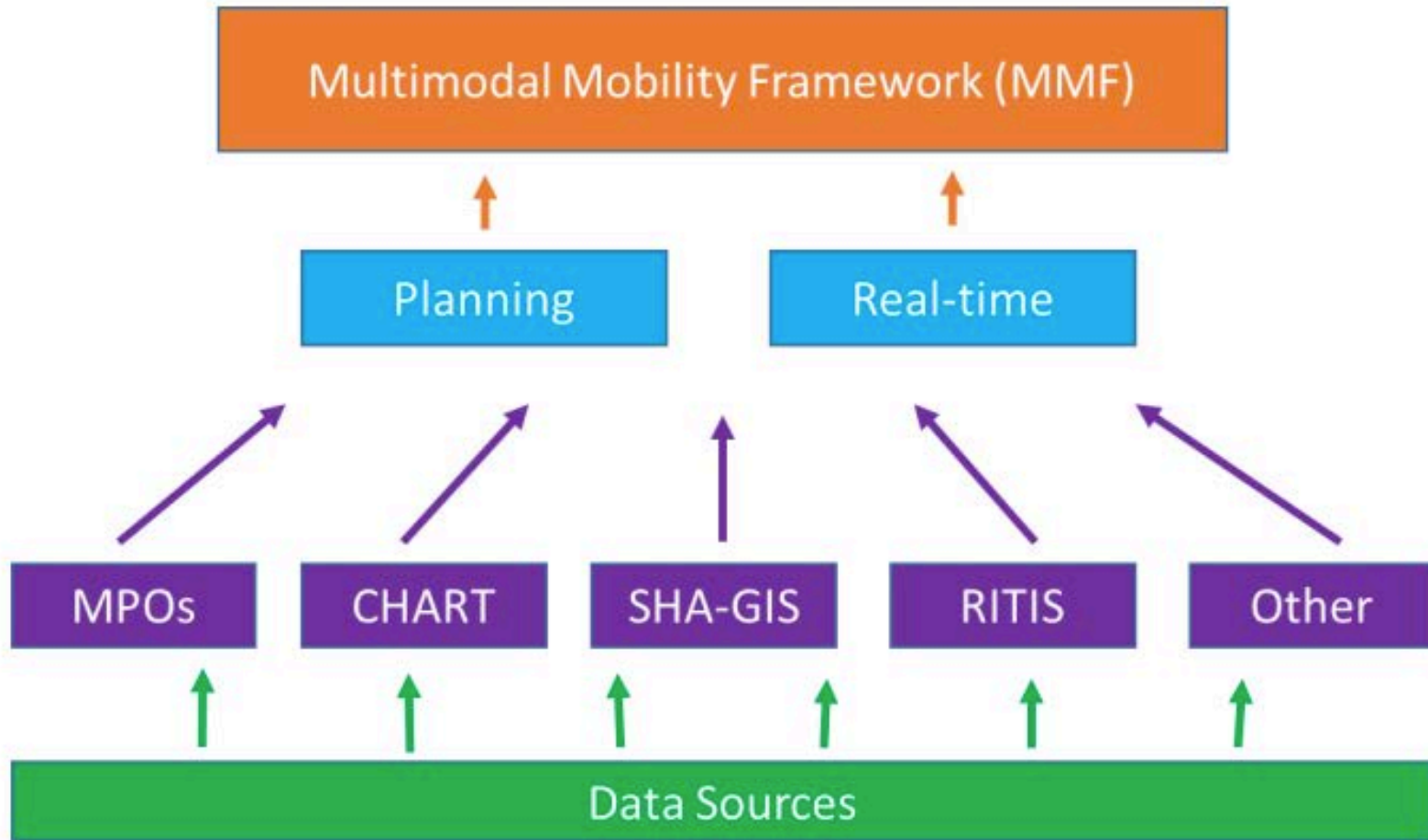
ACCURACY, RELEVANCY, TIMELINESS, ACCESSIBILITY,
COHERENCE & COMPATIBILITY

Structured vs. Unstructured Data

- >90% of the data used at MDOT – (planning, design & operations) is highly structured, meaning...
 - It is machine-readable
 - It can be easily stored in relational databases
 - It is standardized in some format
 - It follows agreed upon rules



Data Quality/ Architecture & Governance for Mobility Data

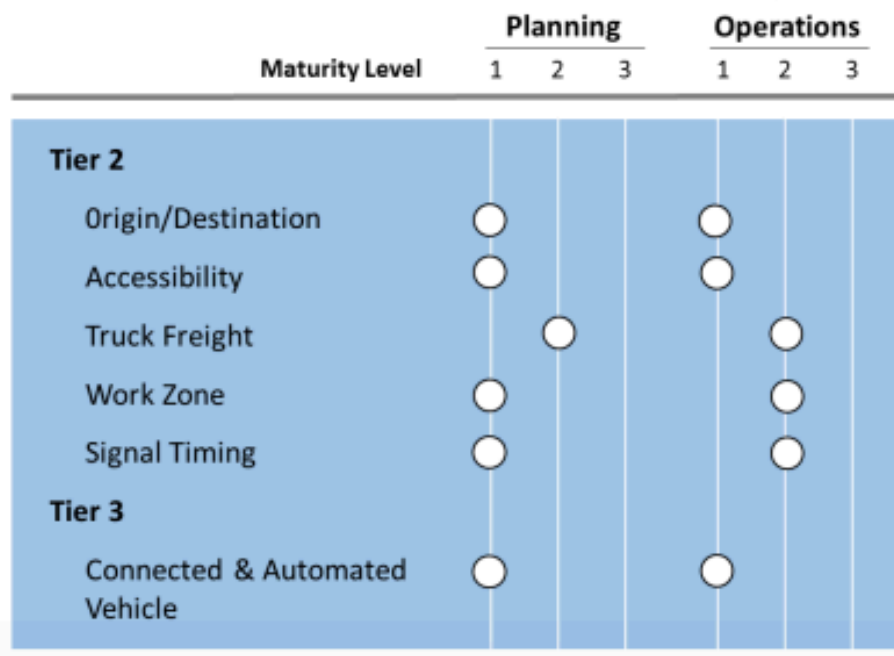
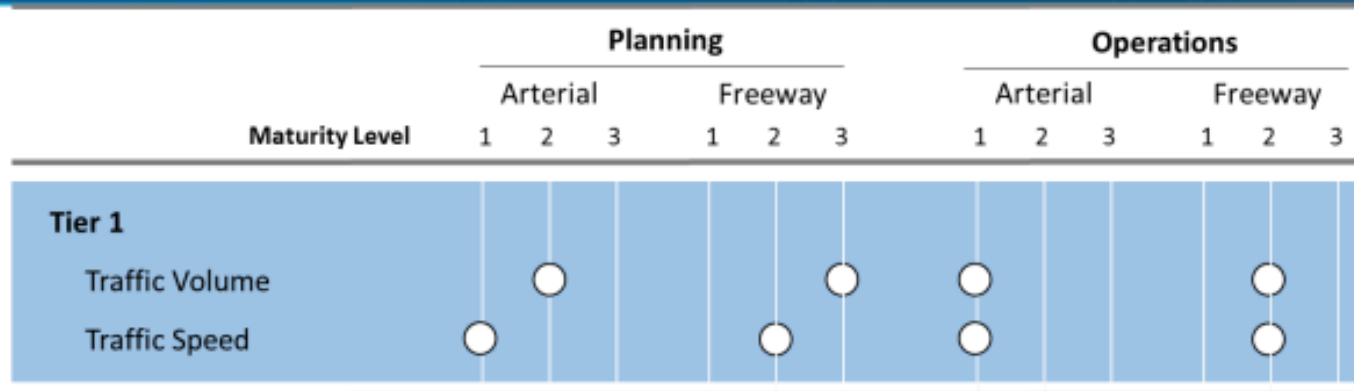


Source: FHWA Data Business Plan – MDOT Pilot

Mobility Data Maturity Framework

Capability Assessment

Structured Data Realm



Portable Sensors: Vehicular Volumes

STRUCTURED DATA MOSTLY USED FOR PLANNING/DESIGN

Type:

- Intersection Turning Movements (includes bike/ped)
- Mainline Traffic Volumes
- Truck Percentages (Vehicle Class)

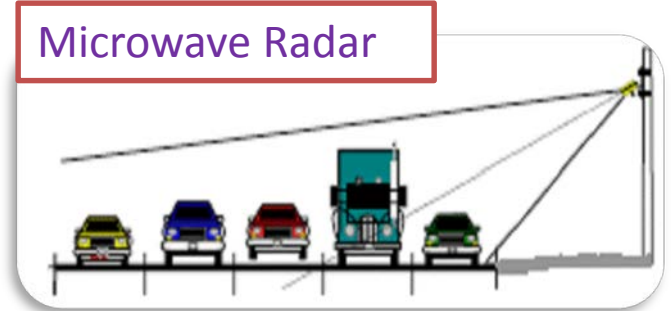
Source:

- Manual
- Pneumatic Tube counts
- Radar-based counts
- Infrared Sensors
- Video-based counts

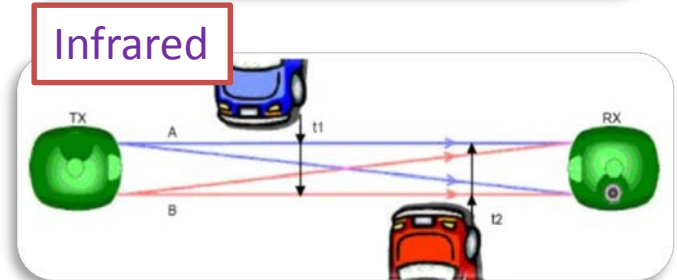
Induction Loops



Microwave Radar



Infrared



Video Image Processing



Portable Sensors: Travel Times and Speeds

STRUCTURED DATA MOSTLY USED FOR PLANNING/ OPERATIONS

Type:

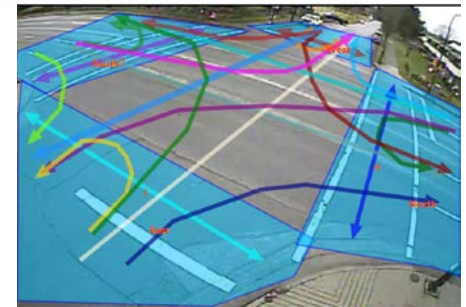
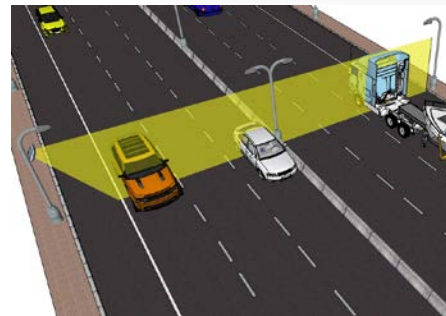
- Vehicle Travel Times & Speeds by Segment
- Point Speeds



Source:

- Floating Car Travel Runs
- Bluetooth & Wi-Fi O/D
- Hi-Def Signals
- ATRs/ Roadside sensors (e.g. side-fire)
- Private Sector Vehicle Probe Data

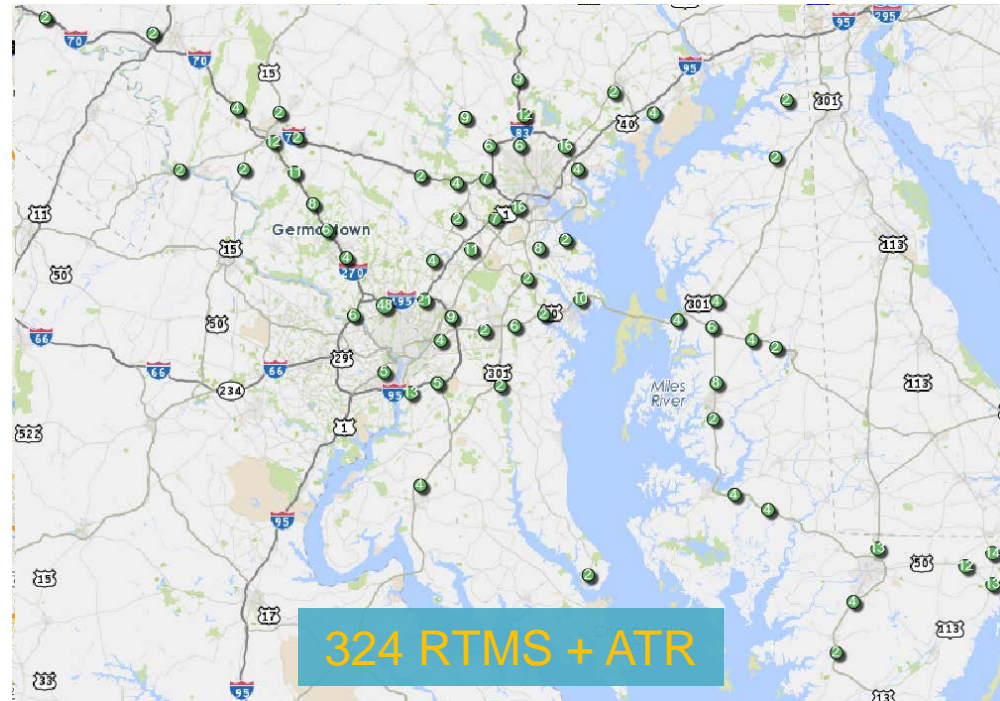
9:00 AM	9:15 AM	9:30 AM	9:45 AM	10:00 AM	10:15 AM	10:30 AM	10:45 AM	11:00 AM	11:15 AM	11:30 AM
100	100	97.43	93.92	90.01	96.40	97.13	97.76	94.27	97.54	98.01
100	100	100	93.94	100	99.64	99.39	100	89.82	96.24	93.7
13.46	13.97	15.1	14.27	17.08	15.67	19.42	33.45	55.72	94.97	89.24
18.79	33.46	15.01	22.34	18.82	17.33	37.21	32.95	28.45	83.27	89.33
19.08	20.52	21.96	19.61	19.48	21.83	29.41	38.3	33.2	80.76	94.51
35.5	35.77	31.92	37.05	48.85	53.46	65.77	63.33	52.82	98.77	94.74
38.21	40.77	36.9	49.23	62.31	63.33	75.64	75.26	63.46	92.82	96.67
57.82	54.1	49.1	46.54	76.28	100	100	97.89	91.79	100	100
68.89	45.97	42.78	60.97	77.92	100	99.17	95.69	95.42	100	98.61



Fixed Point ITS Infrastructure/ Sensors

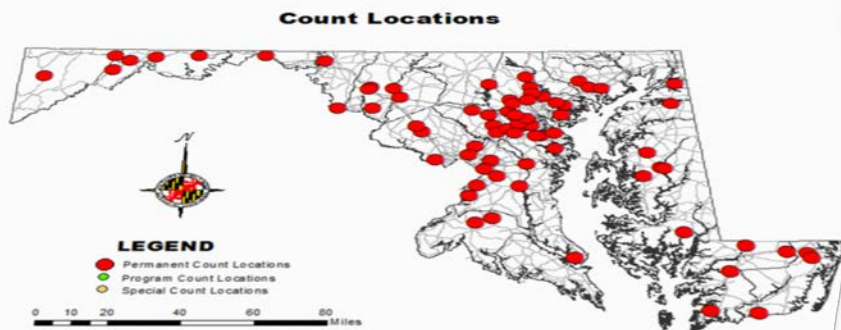
STRUCTURED/ UNSTRUCTURED DATA MOSTLY USED FOR OPERATIONS

- CCTV Cameras:
307 – controllable by CHART
800+ accessible to CHART as view-only
- Dynamic Message Signs (DMS): 218
- Remote Traffic Microwave Sensor (RTMS): 234
- Automatic Traffic Recorder (ATR): 90
- 90% State owned signals have video and remote access to video feed

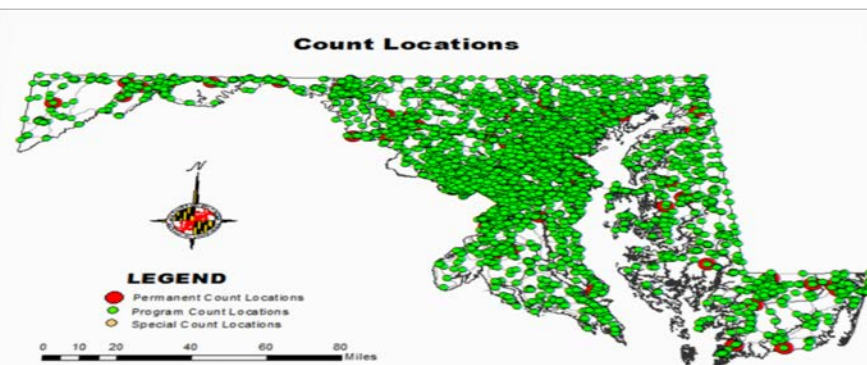


Fixed Point/ Portable Sensors for Traffic Counts

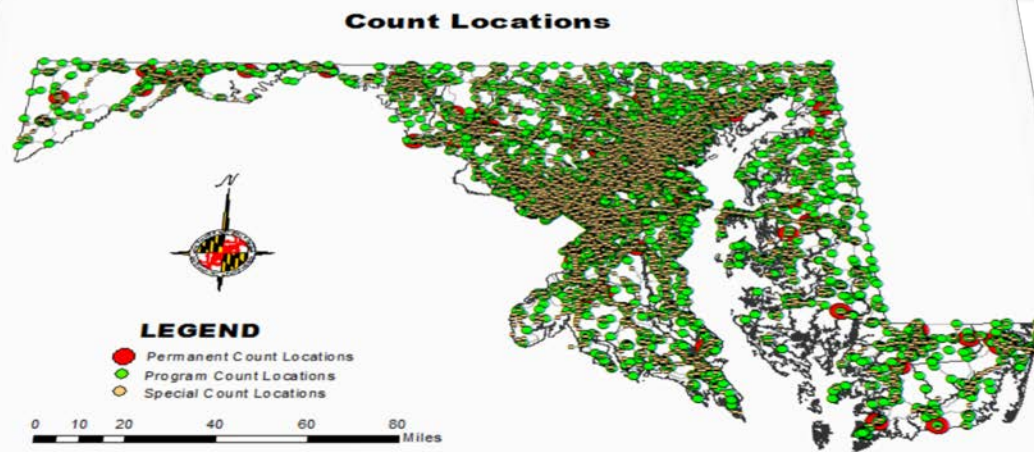
STRUCTURED DATA MOSTLY USED FOR PLANNING/ DESIGN



87 Permanent
Automatic Traffic Recorders (ATRs)

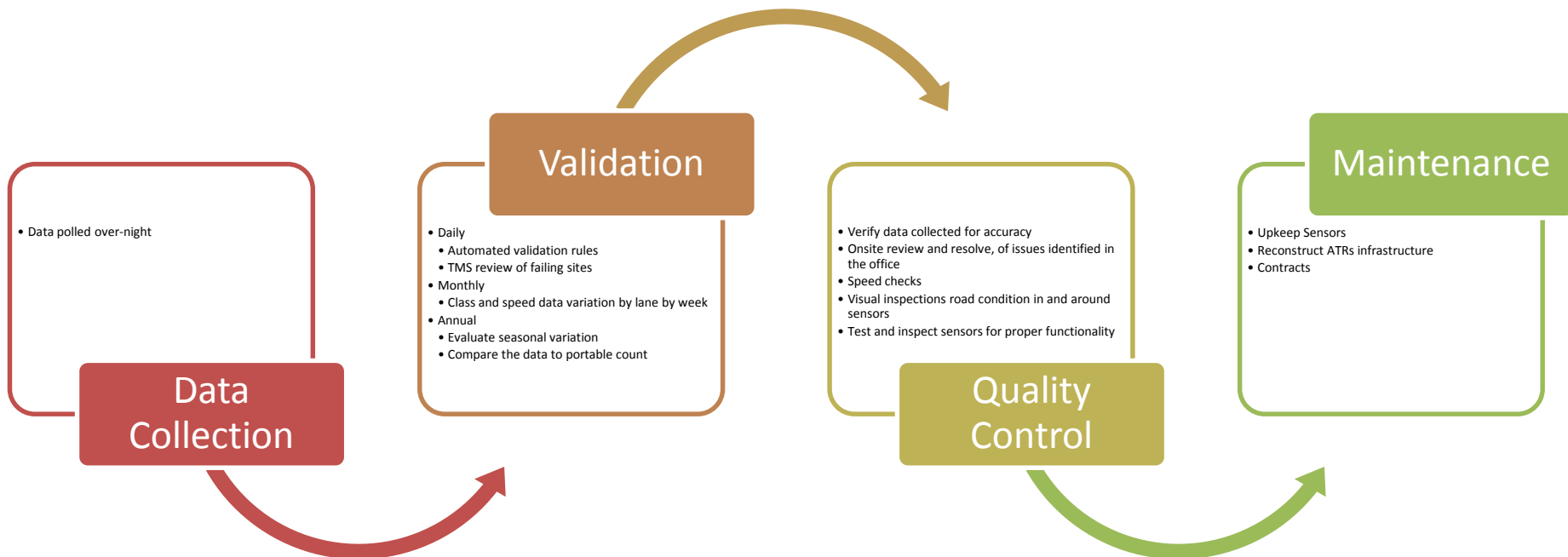


8800+ Program (Coverage) Counts on 3 & 6 year cycle



1400 Special Project Counts per year

Maintaining Data Quality of Structured Data



Automated Rule based Validations

RULE NUMBER	RULE NAME	DESCRIPTION
Rule 1	Consecutive Hours	Consecutive Hours should not have the same volume, especially 0 volumes
Rule 2	Directional Split	The directional volumes should be within a 60%-40% range
Rule 3	Standard Deviation	should be within a 60%-40% range
Rule 4	Standard hours of collection	Each data file should have complete day (24 Hours) of data
Rule 5	Lane Volume of zero	Interstate and Primary Arterial Routes should not have zero volumes for any given hour

Statistical Evaluation Criterion for Sensors (Structured Data)

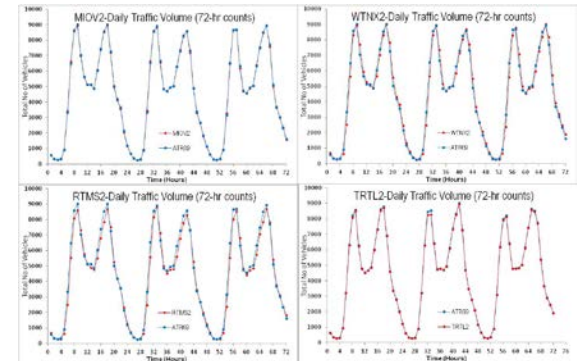
Descriptive Statistics

- ❑ Average and Standard Deviation of Daily Traffic
- ❑ Average and Standard Deviation of Hourly Traffic

$$\text{Error(Difference)} = \text{Baseline}_{\text{ATR}} - \text{Sensor}$$

Graphical Analysis

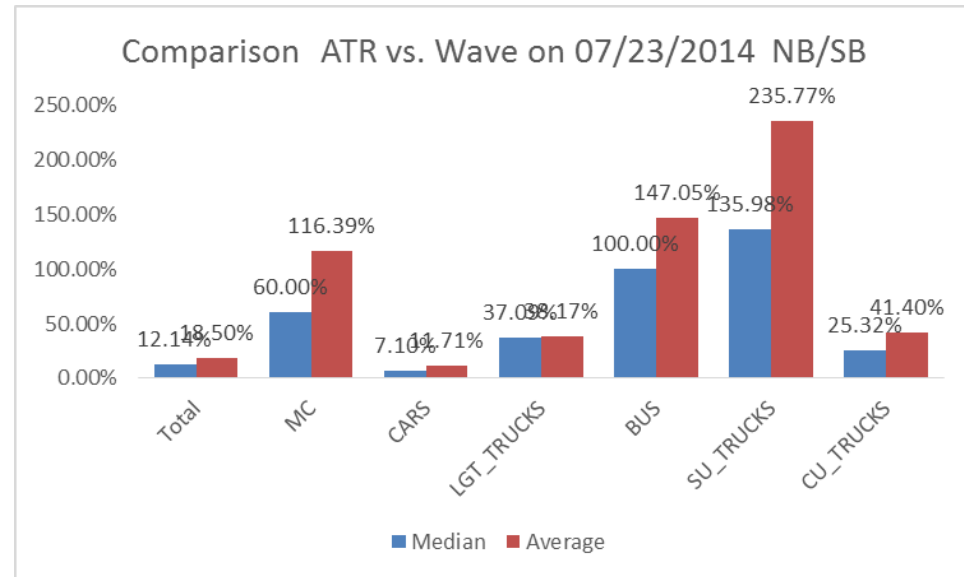
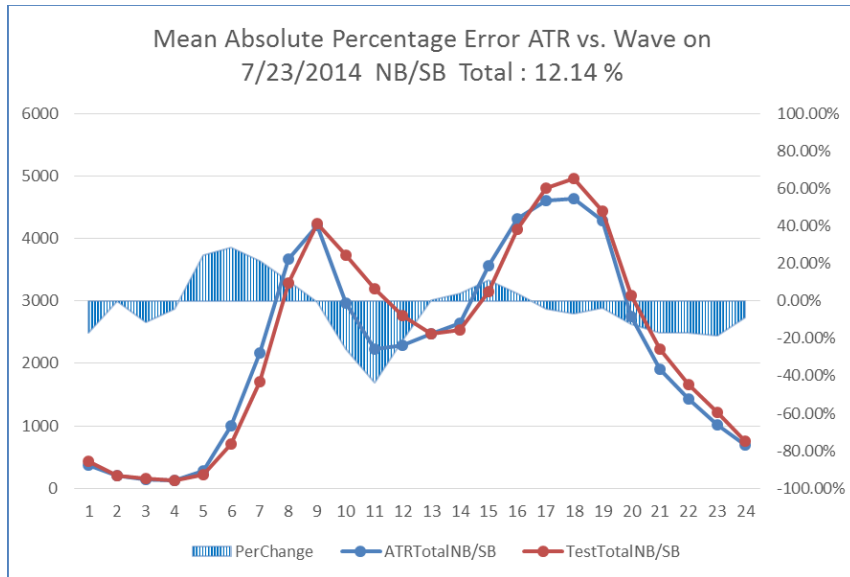
- ❑ Mean(Bias)and Std-Dev (Precision)of the Difference
- ❑ Mean Absolute Percentage Error (MAPE)
- ❑ Mean Percent Error (MAE)-Mean Deviation
- ❑ GEH Statistics



Analytical Statistics : Quantify the Difference(Error) using Linear Regression

- ❑ Root Mean Squared Error (RMSE)
- ❑ Scatter Plots
 - ❑ Pearson's Correlation Coefficient r
 - ❑ Coefficient of Determination r^2

Standardized Evaluation for Data Quality of Sensors for Structured Data



In-depth comparison of Data Collected from New Sensors Versus the Base Line

Probe Data vs. Traditional Sensors

Probe Technology

- Vehicle travel time is measured directly
- Only a sample of all vehicles is monitored
- Volume is inferred from sample size
- Speed estimates are space-mean speed
- Roadside infrastructure is minimized or eliminated
- Quality of data is based on the percent of vehicles monitored

Fixed-point Sensors

- Traffic volume and occupancy is measured directly
- Traffic speed is inferred from occupancy based on an average vehicle length
- Travel time is inferred from a network of sensors
- Equipment in the right-of-way is required
- Cost of deployment and maintenance has historically been high

Outsourced Freeway Probe Data Quality

- Accuracy of probe data on freeways for all major vendors has been extensively tested.
- Probe data can be safely used for real-time applications, planning and performance measurement on freeway network.
- Freeway validation studies show that probe data exhibits latency
 - Average latency is 4.4 minutes and varies by vendor
 - Latency may impact real-time applications



Outsourced Arterial Probe Data Quality

- Unlike freeways, quality of probe data varies by number of lanes, AADT, signal density, access points, speed limit, median access, and major junctions are important factors.
- Probe data quality **most correlated** to signal density
- Accuracy is anticipated to improve with increased probe density

Arterial Probe Data Usability

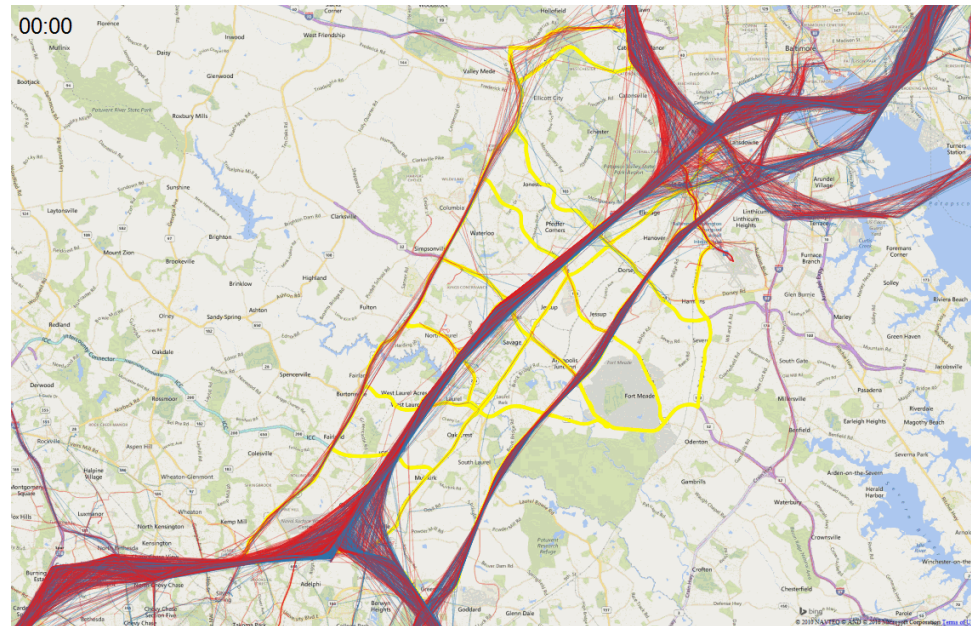
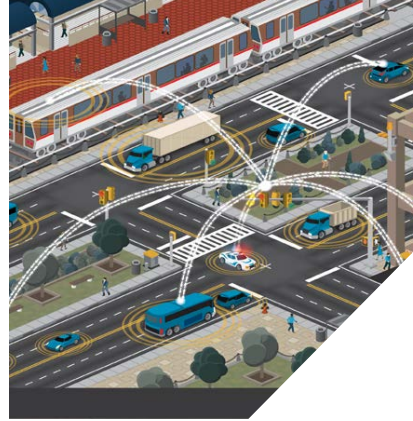
✓ RECOMMENDED	🔑 SHOULD BE TESTED	✗ NOT RECOMMENDED
<ul style="list-style-type: none"> ● <= 1 signal per mile ● AADT > 40,000 vpd (2-way) ● Limited curb cuts <p>Principal Arterials Likely to be accurate...</p>	<ul style="list-style-type: none"> ● 1 to 2 signals per mile ● AADT 20K to 40K vpd (2-way) ● Moderate number of curb cuts <p>Minor Arterials Possibly accurate, test ...</p>	<ul style="list-style-type: none"> ● >= 2 signals per mile ● AADT < 20K (2-way) - low volume ● Substantial number of curb cuts <p>Major Collectors Unlikely to be accurate...</p>

Emerging Performance Measures

- High-resolution probe and Smart Signals
 - Percent Arrivals on Green
 - Capacity Utilization at Intersections
 - Travel Time Reliability

- Origin/ Destination Data
 - Cellular Data
 - Vehicle Probes/ OEM Data
 - Smart Phone/ Wi-fi/ Bluetooth

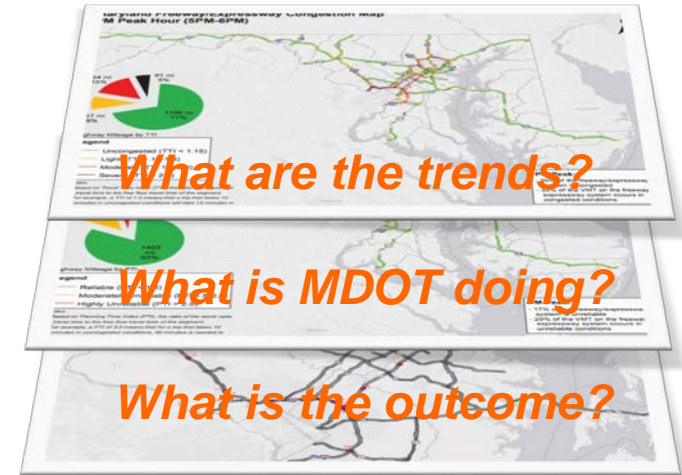
- Connected/ Automated Data



Informing Transportation Investment Decisions

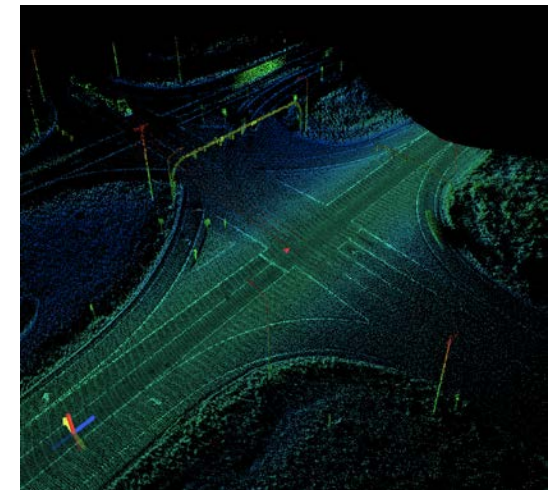
Integration of **Geo-spatial Analysis** and **Trend Analysis** of various data layers:

- Safety
- Mobility
- Asset Conditions
- Environmental
- Accessibility & Economic Opportunities

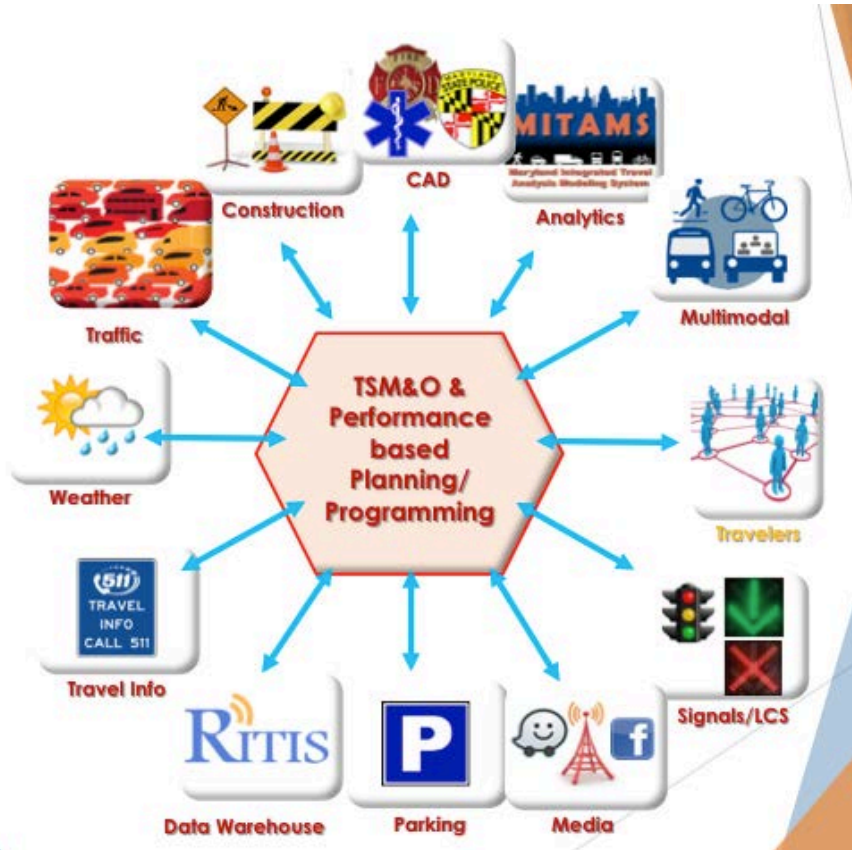


With **Analytical Tools and Applications** to inform decisions at multiple levels:

- Strategic
- Long range planning
- Corridor Studies and Project Level decisions



Unstructured Data



Examples of Unstructured Data

(not just in Transportation)

1. Writing

- reports, emails, meeting minutes, narrative weather reports, etc.

2. Social Media

- Scanning streams to detect real time information such as incidents or public sentiment about MDOT.

3. Natural Language

- Any form of audio recordings—like voicemail, interviews, radio communications, etc.

4. Photographs & Video

- Photographs of MDOT assets, real-time CCTV video, etc.

5. Communications

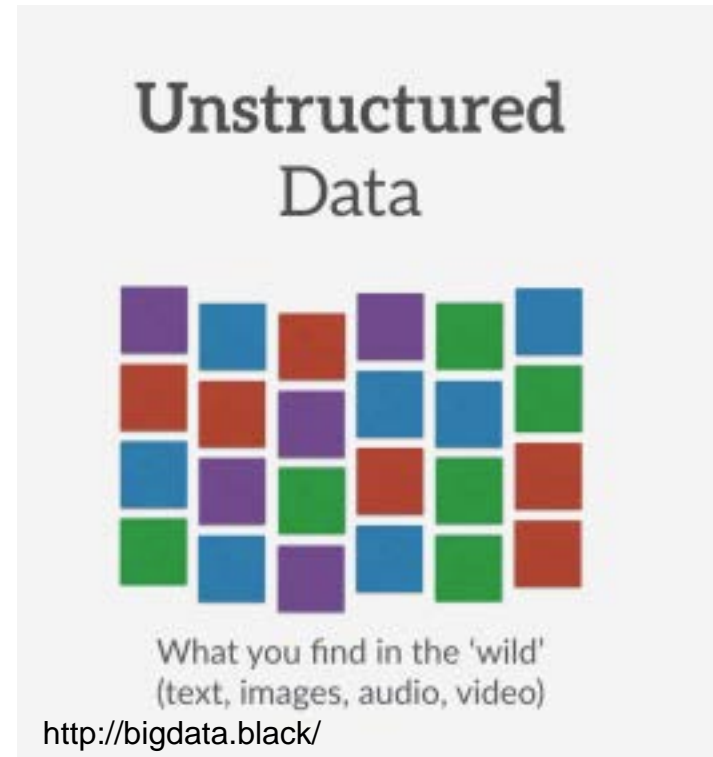
- Scanning communications such as emails to detect spam.

6. Science

- Looking for patterns in interstellar radio messages in order to discover intelligent life.

7. Health

- Analysis of x-ray images for signs of disease.



Planning & Ops Examples in DOT World

Structured

- Incident/Event Details from CHART
- Speed & Volume Sensor Measurements
- Probe-based speed data
- Crowd-sourced WAZE data
- Other ITS devices
- Digitized plans
- Asset Data – GIS
- Police Crash Reports (ACRS)
- Plow, Transit, and maintenance vehicle AVL
- Weather Radar
- RWIS weather measurement

Unstructured

- Free-text Tweets from Twitter
- CCTV video
- Radio/Scanner audio
- Free-text weather reports
- Free-text CAD notes
- Free-text operator notes/communication logs
- Emails
- Reports

Unstructured is Decreasing in DOT Ops

- Example 1:
 - The MDOT SHA CHART system used to include MANY free-text fields about
 - The location of an incident
 - Exit 23 vs. X-23 vs. Ex. 23 vs Ext twenty-three vs. Exit near McDonald's
 - The type of incident
 - Disabled Vehicle vs. Dsbld vhcl vs DV vs. Disabled
 - Which responders were notified and/or responded
 - Fire Company 23 vs. Howard County Fire vs. Ladder Truck 23 vs. F23
 - Which Lanes are closed
 - Ln 1 vs Lane 1 vs. Right lane vs. rt. Ln.
 - Today, nearly every field is automated or populated with drop-downs—making it extremely easy to run performance reports that analyze the effectiveness of CHART

Unstructured is Decreasing in DOT Ops

- Example 2:
 - Maryland Police Crash Reports used to be hand-written field reports that were later painstakingly digitized by data entry personnel.
 - This led to significant data entry errors due to misinterpretation of notes, bad hand-writing, etc.
 - This also meant that data availability and analysis were delayed by a year or more due to the amount of time it took to manually digitize all of the reports
 - The transition over to digital police crash reports with drop-down field entry through the ACRS platform has significantly reduced data lag and errors.

Some Unstructured Data Persists






- CAD notes can still contain significant text
- CHART continues to deploy fixed and mobile CCTV cameras that provide valuable insights in the moment, but are difficult to query and/or analyze by machines
- Lengthy research reports are still produced
- Audio “data” from radio systems persists

The Impact?

- Unstructured data in transportation is largely ignored for reporting and analysis purposes because of the time and cost needed to make lasting sense of it.
- Unstructured data may still be useful “in the moment” for operations and other decision making, but it is rarely archived in a meaningful/useful way.
- Structured data has been leveraged significantly by planning and operations communities for performance reporting and decision making because it is easier to work with and less prone to errors.

Next Gen Data Sources

More and more, the transportation sector is relying on data to drive decisions, and on technology to reimagine how we move people and goods.

Connected Vehicles

Vehicles that communicate are the latest innovation in a long line of **successful safety advances**.

The motor vehicle fatality rate has dropped by **80%** over the past 50 years.

Connected vehicles and new crash avoidance technology could potentially address **81%** of crashes involving unimpaired drivers.

Robotics

Advances in robotics are changing transportation operations and will impact **the future transportation workforce**.

Robots will perform vital transportation functions, such as critical infrastructure inspection.

NextGen

GPS and new technologies are leading to a **safer, more efficient U.S. airspace**.

By 2020, **one-second updates** will pinpoint the **aircraft location and speed** of 30,000 commercial flights daily.

Real-time Travelers

Mobile access to everything from **traffic data to transit schedules** informs our travel choices.

90% of American adults own a mobile phone.

20% use their phones for **up-to-the-minute** traffic or transit information.

Smartphones are regularly used for **turn-by-turn navigation**.

NEXT TRAIN IN 2 MIN

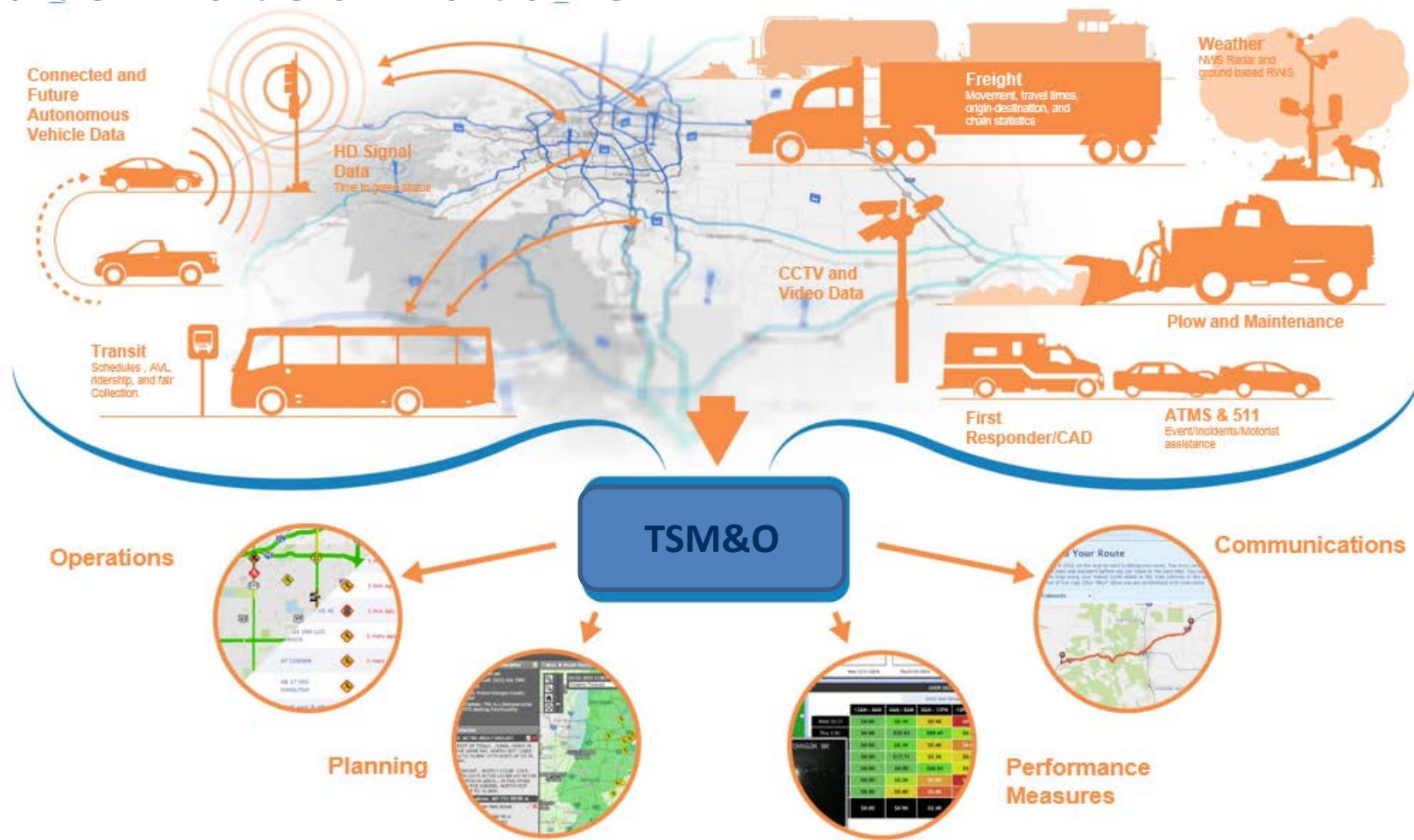
Big data

is all around us. Global data generated is projected to grow by **40%** annually.

Data enables innovative transportation options, such as **car-sharing, ride-sharing, and pop-up bus services**, and more **rapid delivery of goods**.

Source: USDOT

Road Ahead in a Connected/ Automated Future



SOLUTIONS HAVE TO ENTAIL USE OF BOTH STRUCTURED & UNSTRUCTURED DATA

Role of Unstructured Data in DOT

- DOTs recognize the value but mainstreaming use is still in infancy
- Planning and Communications team have started layering social media feeds in Story Maps
- Most of the use of Unstructured Data is ad-hoc

HUGE FOCUS ON CUSTOMER EXPERIENCE !!

UNSTRUCTURED DATA USE IS SEEN AS A KEY TO UNDERSTANDING THE CUSTOMER PERSPECTIVE.

IN SUMMARY...

- It has taken a long time but we are witnessing traction towards **standardizing structured data for planning, operations and TSM&O.**
- Data quality, collection procedures are getting better and agencies are considering **data as an asset with multi-faceted uses.**
- **Focus has been on converting unstructured data to structured data for better decision-making**
- DOTs recognize the **value of UNSTRUCTURED DATA, but this remains an UNTAPPED OPPORTUNITY AREA !!**

ROLE OF QUALITY DATA IS CRITICAL

Contact Information

Subrat Mahapatra

Chief, Innovative Performance Planning Division
Office of Planning & Preliminary Engineering
Maryland DOT State Highway Administration

707 North Calvert Street

Baltimore, Maryland 21202

SMahapatra@sha.state.md.us

410-545-5649