UNITED STATES
DEPARTMENT OF TRANSPORTATION

Data Capture and Management State of the Practice Assessment and Innovations Scan Overview

Mobility Program Summer Webinar Series

Mohammed Yousuf
FHWA Office of Operations R&D
August 17, 2011

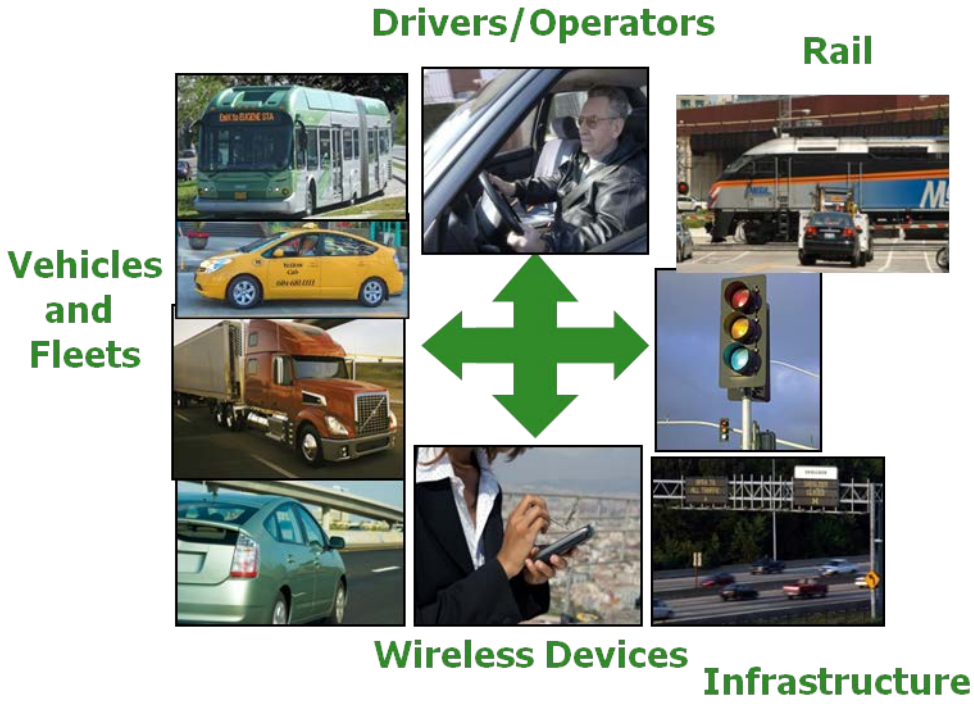
Overview of Webinar

- Brief introduction of ITS Research and Mobility Program
- Purpose of project
- Issues and Innovations
- Fundamental challenges and best practices
- Recommendations on technologies and methods with the most promise for data capture and management in a multi-source data environment
- Next steps
- Getting involved
- Discussion

ITS Research = Multimodal and Connected

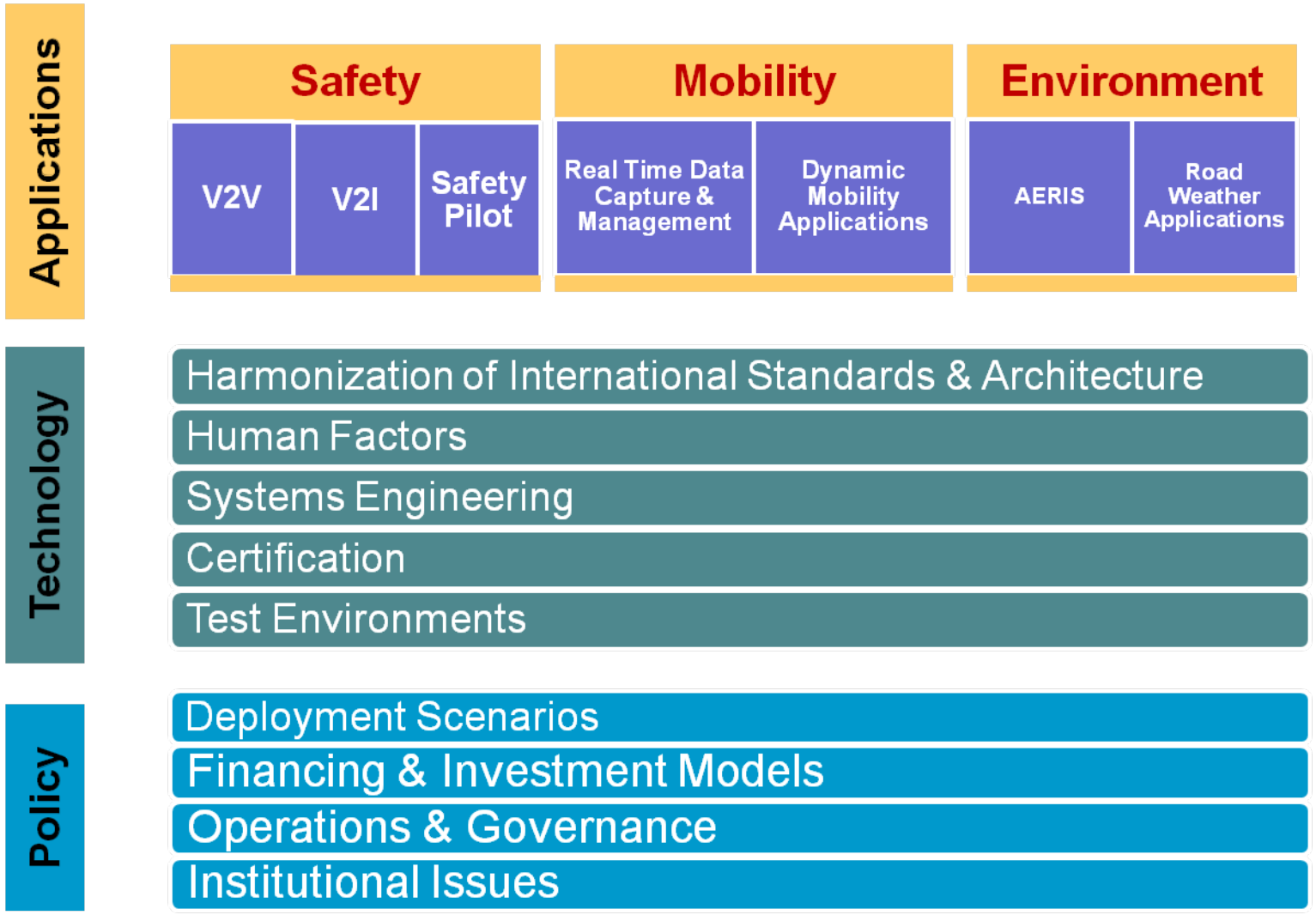
- To improve Safety, Mobility and Environment

- Research of technologies and applications that use wireless communications to provide connectivity:
  - Among vehicles of all types
  - Between vehicles and roadway infrastructure
  - Among vehicles, infrastructure, and wireless consumer devices



- FCC Allocated Spectrum at 5.9 GHz for Transportation Safety (known as DSRC)

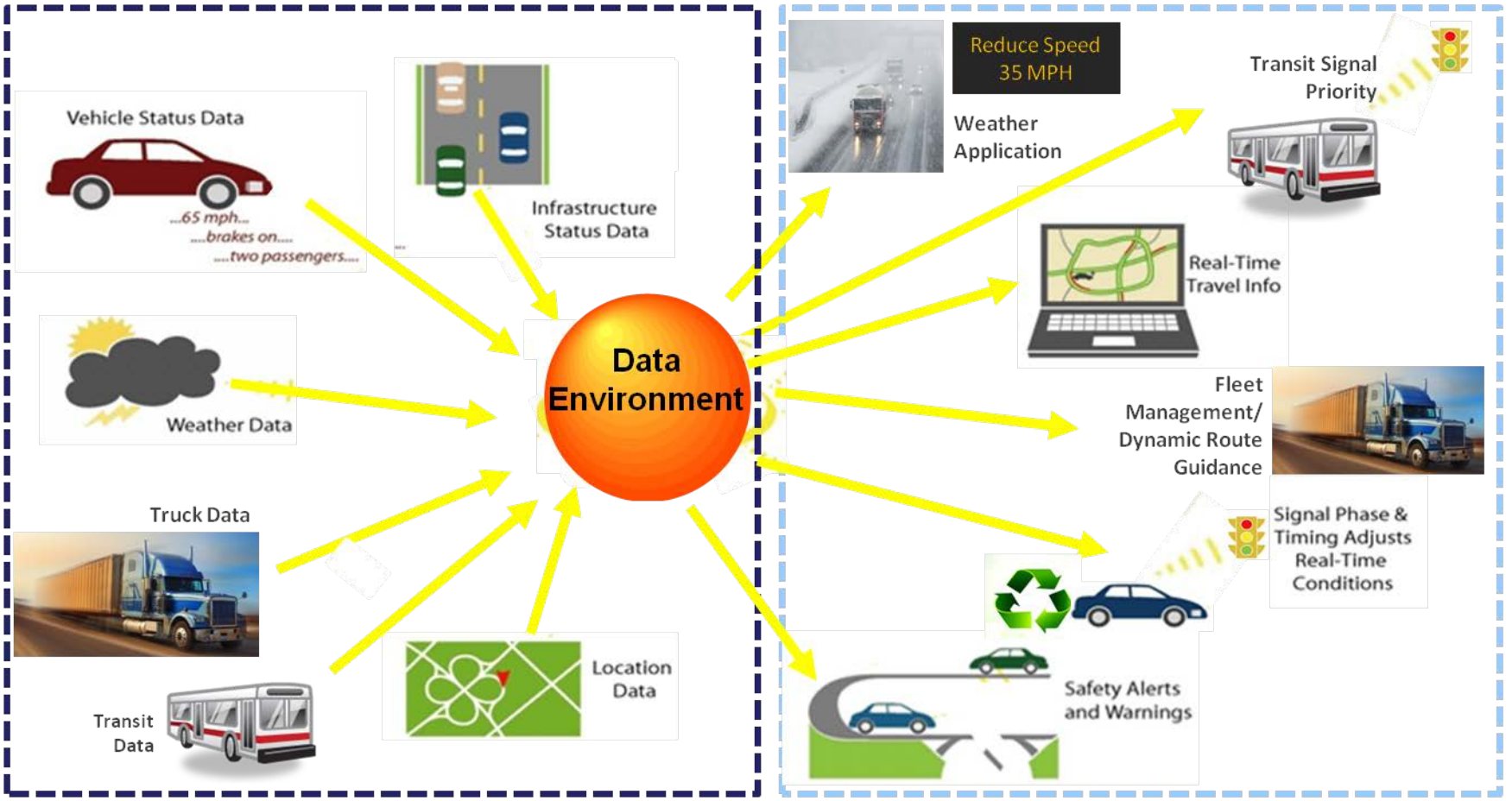
ITS Research Program Components



Mobility Program

Real-time Data Capture and Management

Dynamic Mobility Applications



Data Capture and Management (DCM) Program: Vision and Program Objectives

Vision

- Active acquisition and systematic provision of integrated, multi-source data to enhance current operational practices and transform future surface transportation systems management

Objectives

- Enable systematic data capture from connected vehicles (automobiles, transit, trucks), mobile devices, and infrastructure
- Develop data environments that enable integration of data from multiple sources for use in transportation management and performance measurement
- Reduce costs of data management and eliminate technical and institutional barriers to the capture, management, and sharing of data
- Determine required infrastructure for transformative applications implementation, along with associated costs and benefits

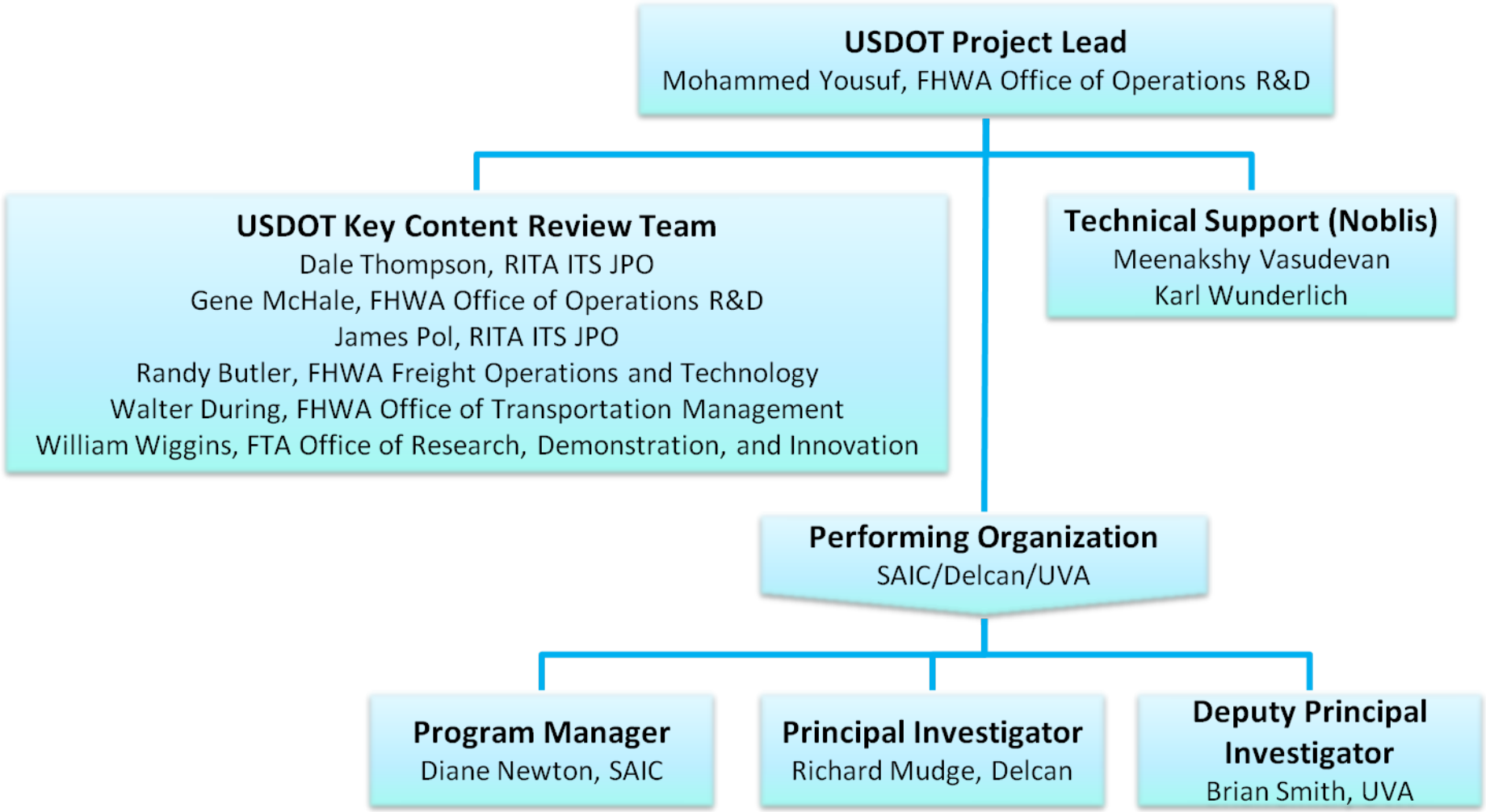
Program Partners

- ITS JPO, FTA, FHWA R&D, FHWA Office of Operations, BTS, FMCSA

Purpose of the State of the Practice Assessment and Innovations Scan Project

- Assess industry best practices in data capture and management methods and technologies that are applicable to the DCM Program
  - Industries: Aviation; Freight Logistics; Internet Search Engines; Rail Transit; Transportation Systems Management
  - Focus areas: quality assurance; access, security, and privacy; storage and backup; operations and maintenance; critical failures
- Identify emerging concepts and technologies that might potentially address issues related to the new paradigm for data capture and management
  - Industries: Information Technology; Aviation; Freight and Transit; Government; Defense; Smart Home and Infrastructure Monitoring; Banking, Finance, and E-Commerce; Health Care and Bioinformatics
- Recommend methods that have the most value for capturing and managing/reporting data in a multi-source data environment

Project Organizational Chart



Challenges and Innovations

Data Capture Challenges

Innovations

- Dynamic Interrogative Data Capture (DIDC)
- Crowdsourcing

Data Management Challenges

Innovations

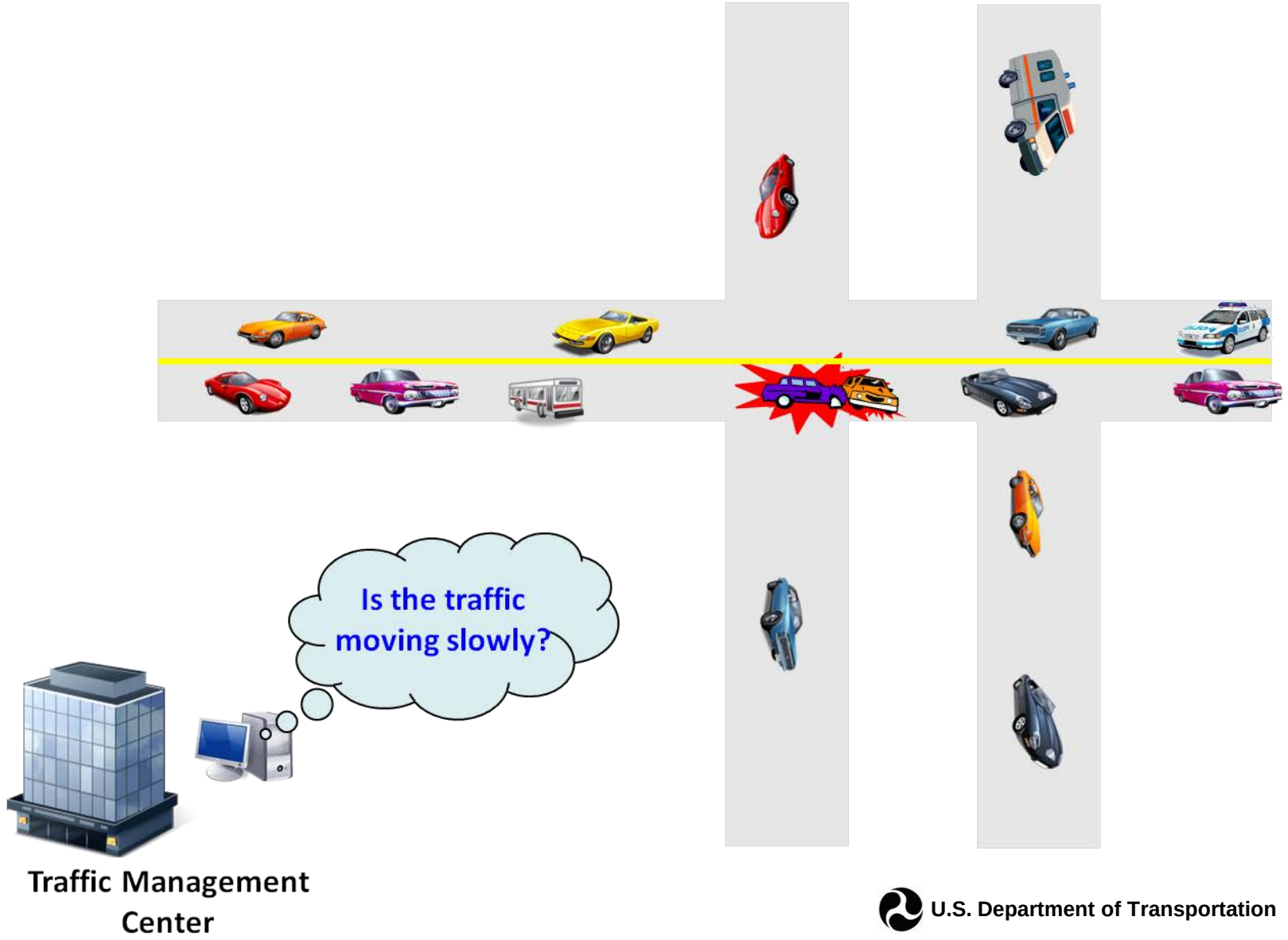
- Virtual Data Warehousing (Cloud Computing)
- Virtual Data Warehousing (Data Federation)

Data Capture Challenge: *Bandwidth Overload*

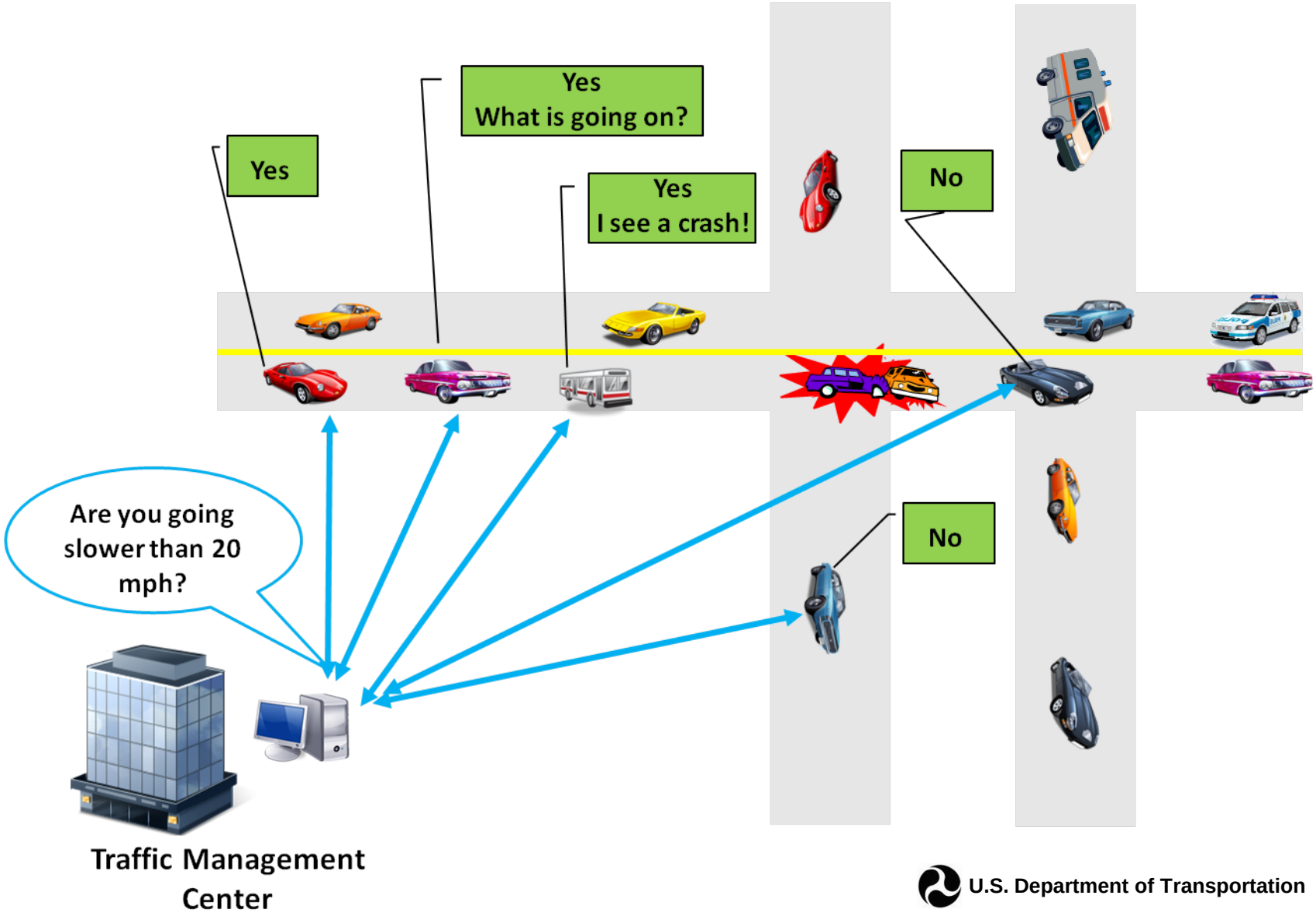
- Issue: Potential data explosion due to new forms of data will likely over-burden the computational and communication systems
  - Large volumes of data with connected vehicles, infrastructure, and travelers
    - Approximately 1.2 MB of accelerometer data generated per vehicle per mile (*Source: Cooperative Transportation Systems Pooled Fund Study on Pavement Assessment by Auburn University*)
    - Translates to 2 TB of data per day just for pavement assessment for Washington, DC
  - Capturing, transmitting, cleaning, and storing large volumes of data can over-burden the system and be cost-prohibitive

- Innovation: Dynamic Interrogative Data Capture (DIDC)

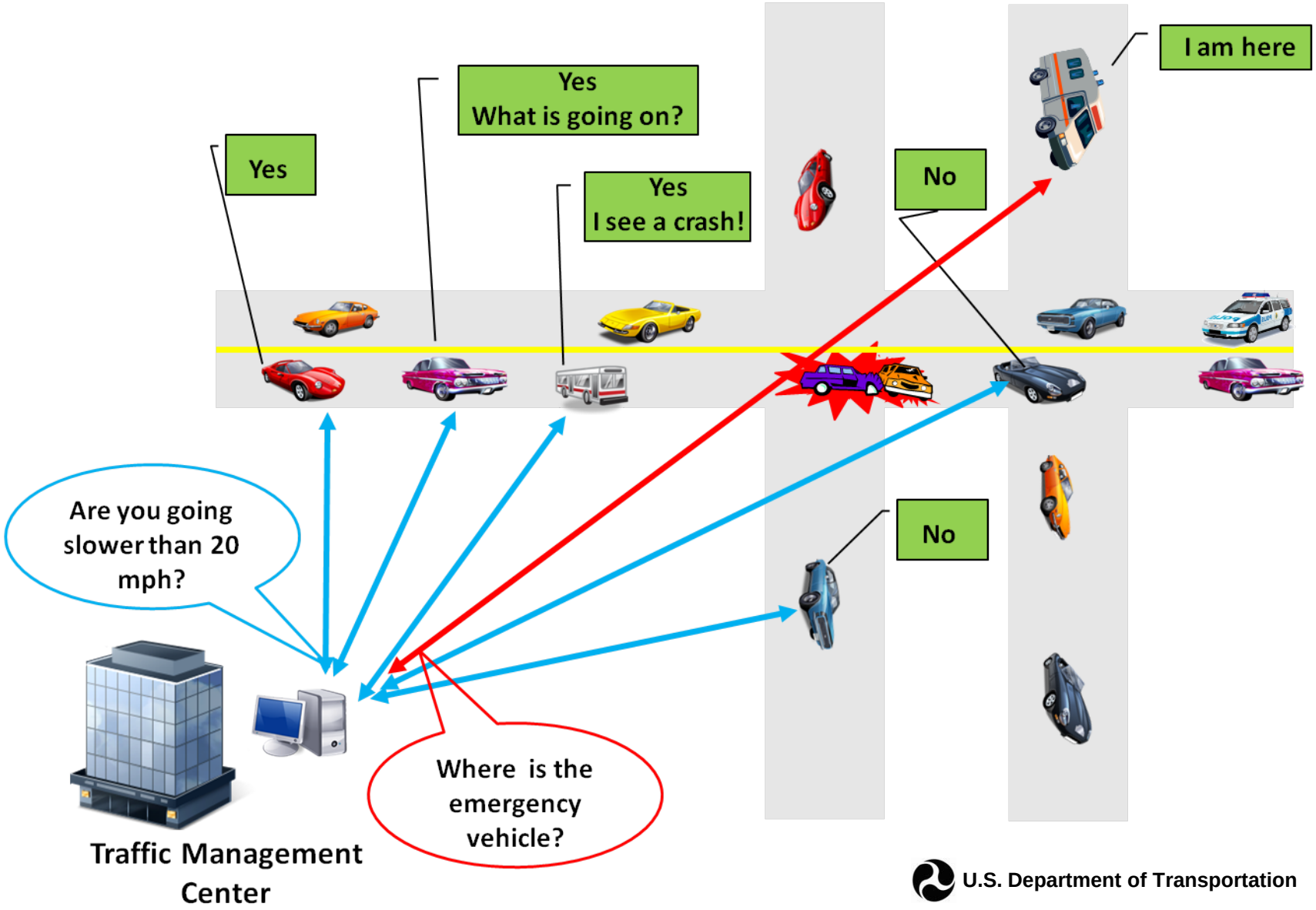
Dynamic Interrogative Data Capture Concept: *Incident Case*



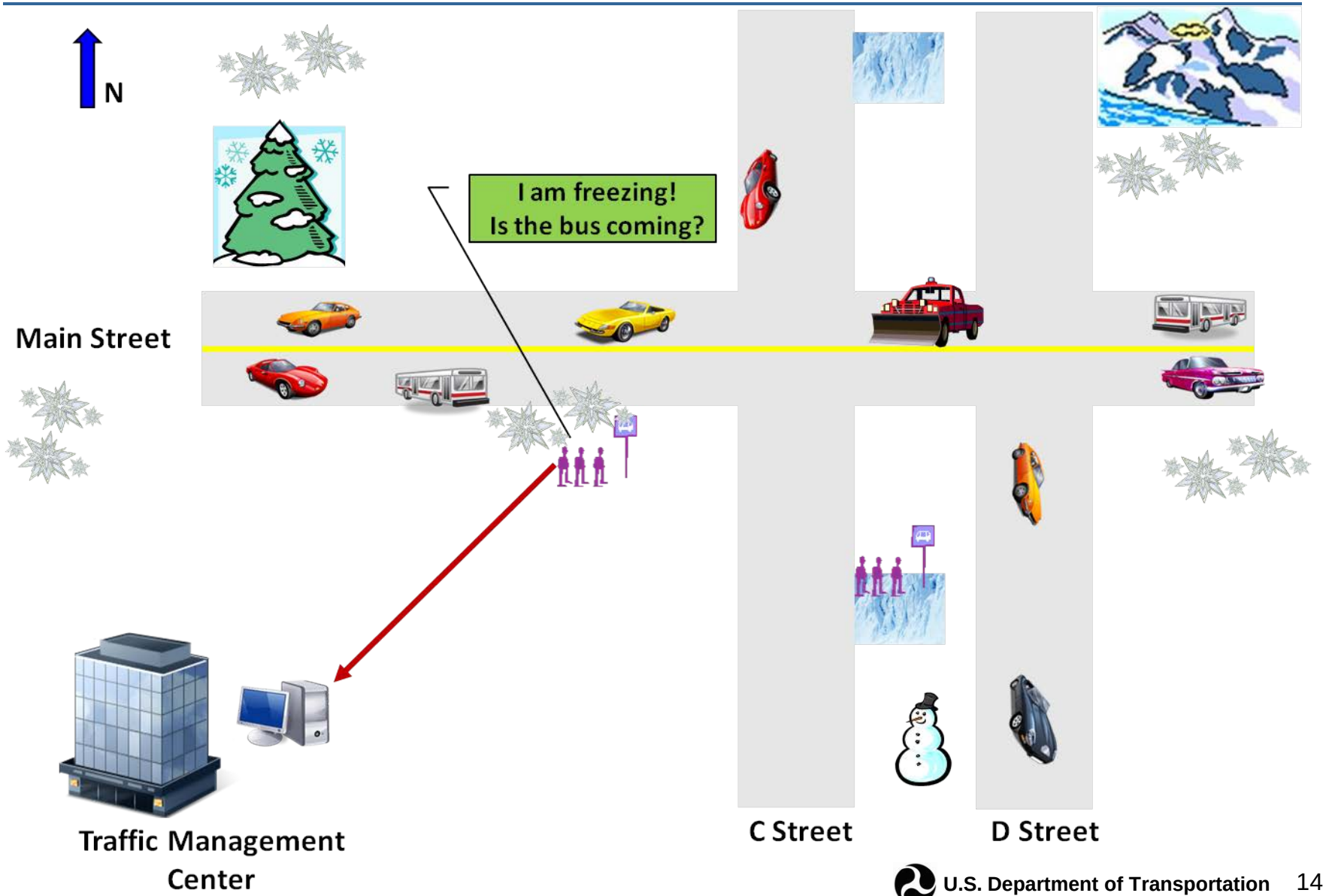
Dynamic Interrogative Data Capture Concept: *Incident Case*



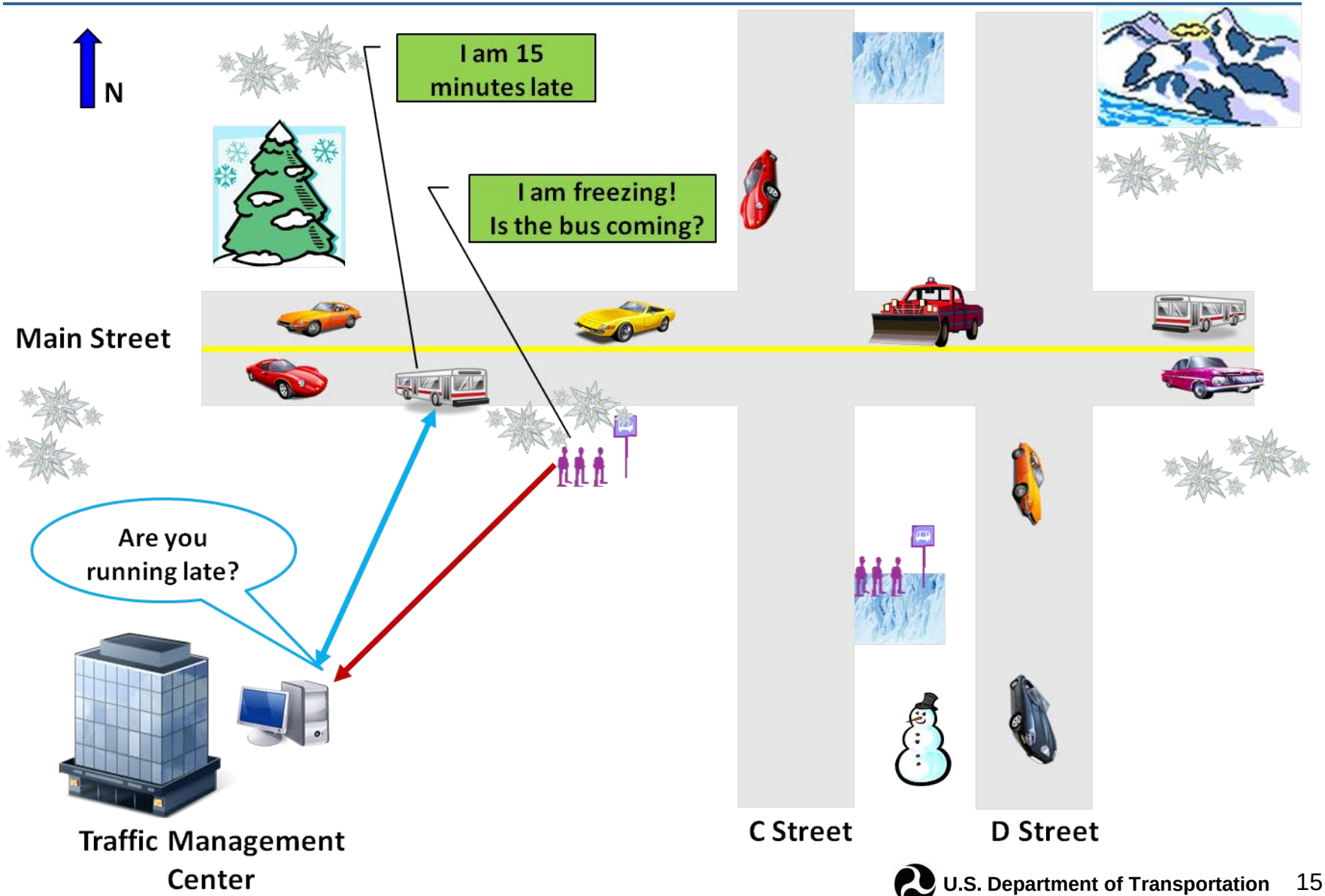
Dynamic Interrogative Data Capture Concept: *Incident Case*



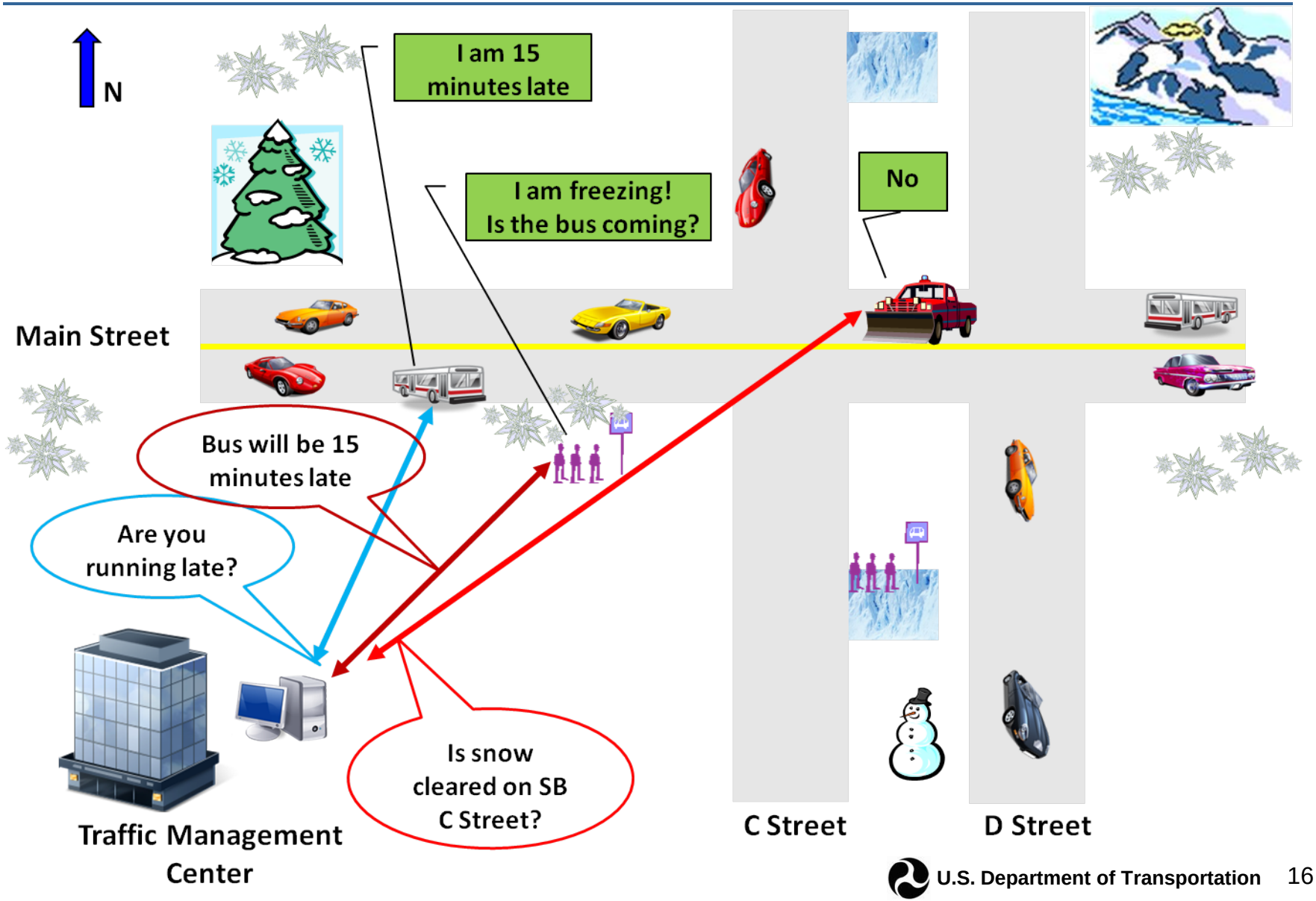
Dynamic Interrogative Data Capture Concept: *Snow Event*



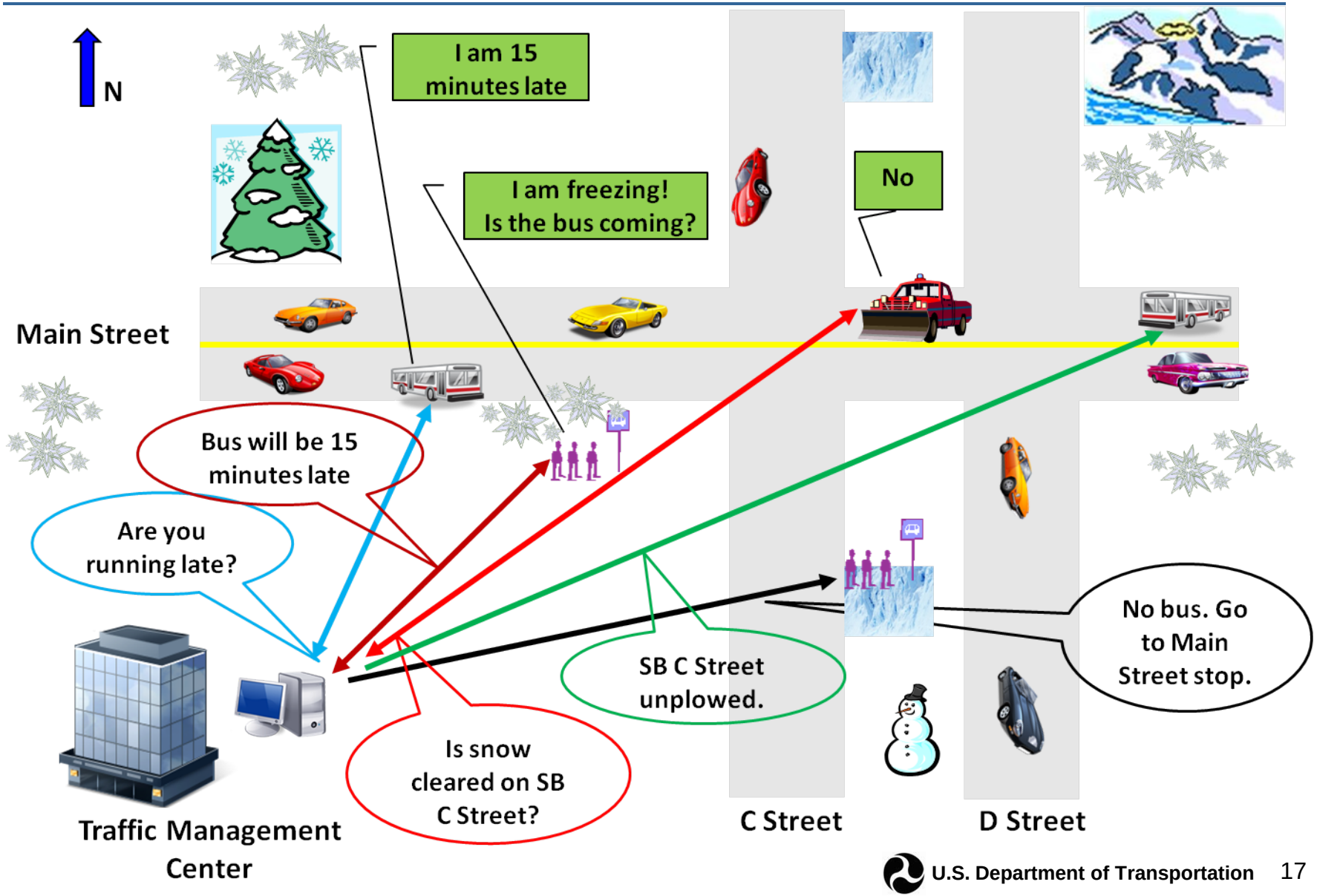
Dynamic Interrogative Data Capture Concept: *Snow Event*



Dynamic Interrogative Data Capture Concept: *Snow Event*



Dynamic Interrogative Data Capture Concept: *Snow Event*

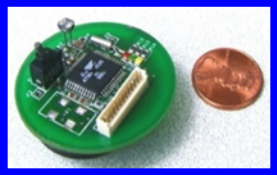


Dynamic Interrogative Data Capture

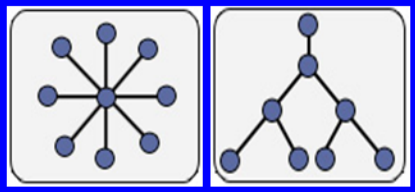
- Each device (in a vehicle, on the infrastructure, or on a person) can set and reset message priorities to different data elements
- Each device can intelligently and dynamically decide on data aggregation levels and transmission frequencies, based on its own state (local conditions) as well as the state of the network (global conditions)
- Each device can query other devices in its vicinity, depending on its data needs, and request certain data aggregation levels
- System of devices dynamically sets thresholds for transmission frequencies and data aggregation levels to control communication overload and energy costs

Example of DIDC Concept

Wireless Sensor Networks (WSNs): Adaptive management of data flow in complex systems



Mote Technology



Example WSN Topologies

Dynamic Interrogative Data Capture: *Value Proposition*

- Increased efficiency
  - Identify critical data elements
  - Collect, clean, transmit, analyze, and store only the required amount of data
- Energy and cost savings
- Increased availability of critical data sets

Dynamic Interrogative Data Capture: *Key Challenges*

- Identify critical data elements to query
  - For every data element that is stored, but not used, there is a cost associated with the capture, cleaning, transmission, and storage
- Determine when to query
  - If a data element is considered unimportant, and not captured, the data will not be available for potential future applications
- Intelligence needs to built into each device
  - Does bandwidth cost savings outweigh putting intelligence into each device?

Data Capture Challenge: *Data from Travelers*

- Issue: Envisioned transformative applications require new forms of real time and archived data that are extremely costly to obtain, or create possible privacy conflicts if required from all vehicles or travelers

- Innovation: Crowdsourcing

Crowdsourcing: *Definition*

- Practice of tapping into the collective intelligence of the public at large to complete tasks that a company would normally either perform itself or outsource to a known entity (blend of crowd and outsourcing)
- Requires motivating crowds
  - Extrinsic: financial compensation; recognition; free use of crowdsourced product (e.g., Challenge.gov)
  - Intrinsic: altruism; autonomy; self-expression; desire to learn new things; entertainment (e.g., Wikipedia)
- Highest quality achieved when intrinsic motivation exceeds extrinsic motivation

When is it beneficial?	Benefits
• Need massive amounts of real-time data • Need continuous temporal and spatial data • Create data archives • Solve challenging problem • Need innovation	• Improves productivity • Minimizes labor and research expenses • Consumers involved in creating product

Crowdsourcing Application: *Data Collection*

- Data collection
  - Most beneficial when massive amounts of data needed
  - Continuous temporal and spatial coverage
- Examples:
  - Inrix: provides traffic information using crowdsourced traffic data, traditional sensor data, and other relevant data (e.g., incidents, weather, construction, special events)
    - crowdsources data from 3 million GPS enabled vehicles and devices covering 450,000 miles of roadways
  - Waze: provides 100% crowsourced, free real-time traffic information on mobile devices
    - crowdsources data from GPS enabled vehicles of volunteers for real-time traffic information and maps (passive participation is sufficient)
    - crowdsources data for map correction (requires active participation)

Image Source: http://www.Inrix.com

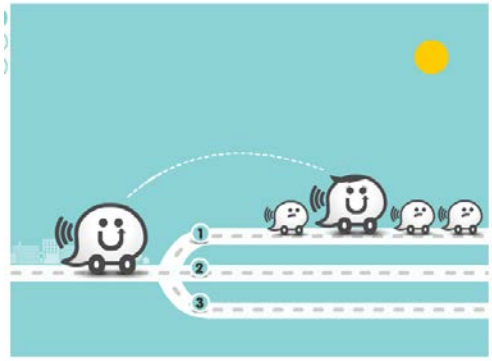


Image Source: http://www.waze.com

Crowdsourcing Application: *Other Areas*

- Data archiving
  - Create or enhance online data repositories to benefit a number of people
  - Examples:
    - Wikipedia: best known encyclopedic knowledge website
    - YouTube: popular website that allows users to upload, share, view videos
- Problem solving
  - Crowdsource solutions to specific challenges of interest mostly to the solicitant
  - Examples:
    - Challenge.gov: federal government poses specific challenges, allowing the public to respond
- Open innovation
  - Crowdsource generation, development, and implementation of ideas
  - Examples:
    - MassDOT's Developers Real-Time Challenge: developers create applications using a real-time feed of bus locations and arrival predictions to make the information available anywhere, anytime

Crowdsourcing: *Key Challenges*

- No control over crowds
  - Some problems may not get solved within the time frame of interest, or in some instances, may not get solved at all
- Little control over quality of crowdsourced product (data)
- Perception of privacy intrusion
  - Can hinder participation in crowdsourced projects
- Expectation of in-kind compensation for participation
  - Possible in kind compensation: Recognition, transparency
  - Crowdsourced data is almost always given back to users, for no cost

Data Management Challenge: *Large Volumes of Data*

- Issue: Large volumes of diverse spatial data call for new methods of data management
  - Infrastructure costs, operations and maintenance costs, storage costs, labor costs can quickly add up
  - Wide range of multi-source data needs to be widely accessible to integrate systems (e.g., signal systems, traveler information systems, transit operations)

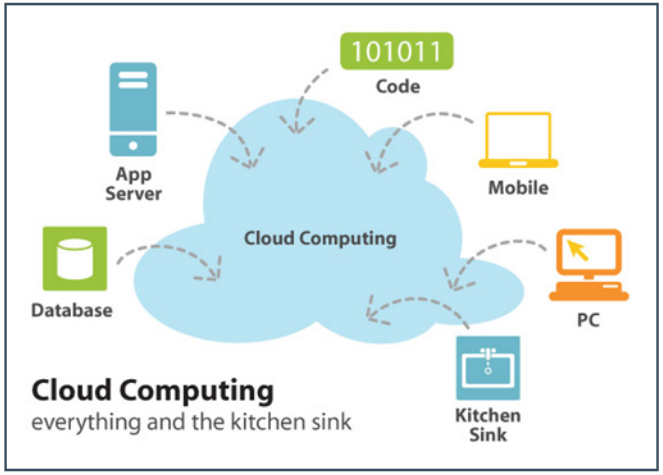
- Innovation: Virtual Data Warehousing (VDW)

Virtual Data Warehousing (VDW)

- What is Virtual Data Warehousing?
  - Functional, virtual equivalent of conventional data warehouse (e.g., CPU time, storage space, operating systems, database)
  - Allows data to be integrated dynamically from heterogeneous data sources that are housed in different locations
  - Allows for rapid sharing of large amounts of data
  - Minimizes data integrity issues
  - Requires less time and expense to develop
- Promising Innovations in VDW Technology:
  - Cloud computing
  - Data federation

Cloud Computing: *Definition*

- "Model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction" (*Source: National Institute of Standards and Technology (NIST)*)



Principal Characteristics	Benefits
• On-demand self-service • Broad network access • Resource pooling • Rapid elasticity • Measured service	• Flexibility • Cost-effectiveness • Ease of use • Reliability • Scalability • Elasticity • Location independence • Increased focus on core company business

Cloud Computing: *Primary Cloud Service Categories*

Infrastructure-as-a-Service (IaaS) ***IT Level***	System administrators obtain general processing, storage, database management, and other resources and applications through the network and pay only for what gets used.	Joyent Cloud Computing, IBM, at&t, GoGrid A ServePath Company, terremark, flexiscale™, Amazon EC2
Platform-as-a-Service (PaaS) ***Developer Level***	Developers design, build, and test applications that run on the Cloud provider's infrastructure and then deliver those applications to end-users from the provider's servers.	Google App Engine, Azure Services, RACE, BUNGEEconnect, force.com platform as a service, amazon web services™
Software-as-a-Service (SaaS) ***User Level***	Users run complete software applications delivered via the Cloud as a service rather than installed on their desktops.	Quicken ONLINE, webex, web.com, workday., Salesforce, Microsoft Office Live

Source: John Veiga, Cloud Computing: A Catalyst for Change, Technology Tuesdays Presentation, Noblis, March 2010.

Cloud Computing: *Deployment Models*

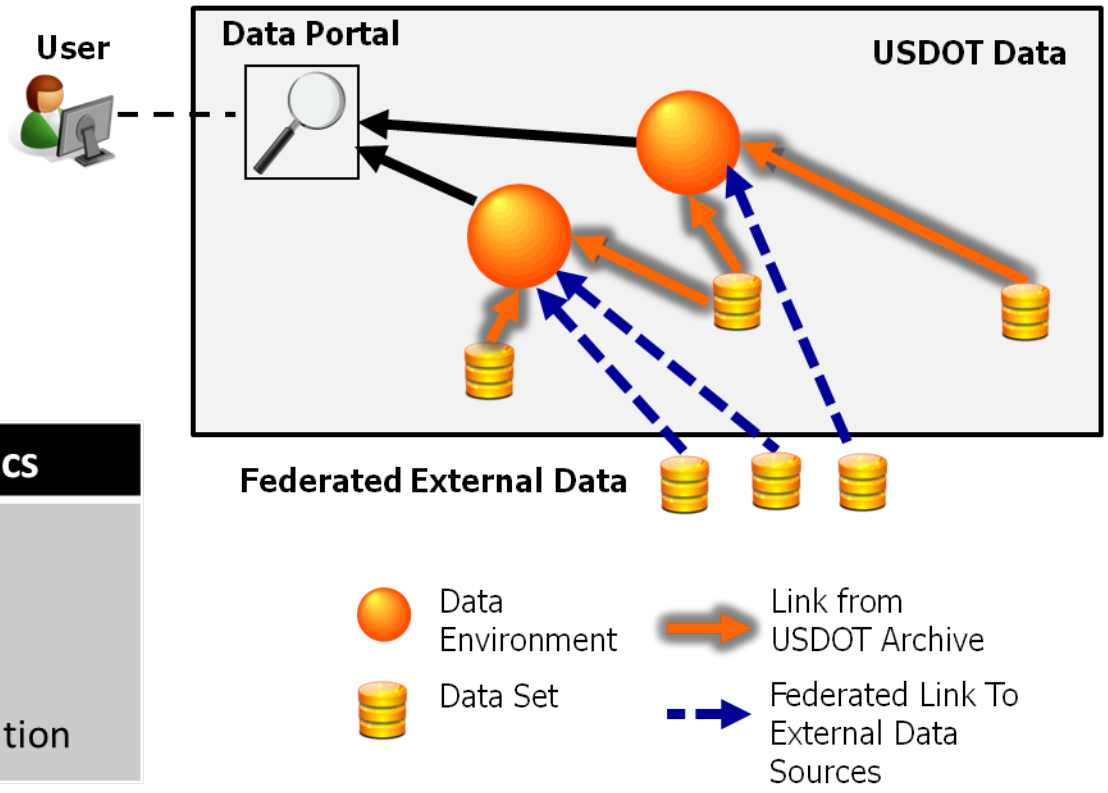
CLOUD TYPE	SUMMARY	CHARACTERISTICS
Public cloud	Cloud infrastructure owned by a third party made available to the general public	• Least secure, and least reliable • Most scalable • Requires the lowest level of management from the user
Private cloud	Cloud infrastructure owned by and operated solely by an organization	• Offers the highest security • Least scalable • Requires the highest level of management from the user
Community cloud	Cloud infrastructure that is shared by a number of partnering organizations with common needs and functions	• Offers middle ground between public and private clouds
Hybrid cloud	Cloud infrastructure comprised of two or more clouds that interoperate through technology	• Inherits features from component entities

Cloud Computing: *Key Challenges*

- Data security
  - Use encryption to protect against snooping during data transit
  - Use intrusion detection and prevention mechanisms
  - Be aware of service provider's security policies
- Reliability and availability
  - Perform periodic off-line data backups
  - Google successfully used tapes to recover data deleted inadvertently in a software roll-out in February 2011
- Data transfer bottlenecks
  - Use of private cloud physically close to the customer can reduce the problem, although at a high cost
- Legal compliance
  - Use service providers with strong security controls
- Data consistency
  - Users perceive eventual consistency as strong consistency
  - Google Apps platform; and Amazon's S3 (Simple Storage Service), SimpleDB and EC2 Elastic Compute Cloud) are successful implementers of eventual consistency

Data Federation: *Definition*

- Form of data virtualization where data from multiple, heterogeneous, autonomous data sources are made accessible to data consumers as if it is contained in one single relational database, by using on-demand data integration
  - Serves as middleware for new or existing databases
  - Stores only metadata



Principal Characteristics

- Decentralized
- Virtual
- Autonomous
- Heterogeneous
- On-demand virtual integration

Data Federation: *Value Proposition*

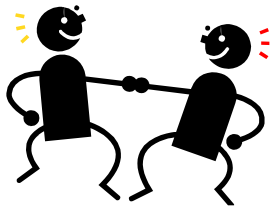
- Transparency of underlying heterogeneity
  - Consumer sees a single uniform interface
  - Consumer doesn't need to know where the data is stored or how it is stored
- Time-to-market advantage
  - Reduces development time significantly when multiple sources have to be integrated
- Reduced development and maintenance costs
  - Develop integrated view once, and leverage multiple times
  - Integrate disparate data sources without consolidating to a single location
- No consistency issues
  - Data are not replicated

Data Federation: *Key Challenges*

- Assumes data already in storage
- Does not scale well
- High management and maintenance effort and cost
- Data transfer bottlenecks
- Does not address reliability and availability issues
- No data replication
  - Be aware of storage and backups at original location
- Data security
  - Protect against snooping during data transit (e.g., use encryption)
  - Be aware of security procedures at original location

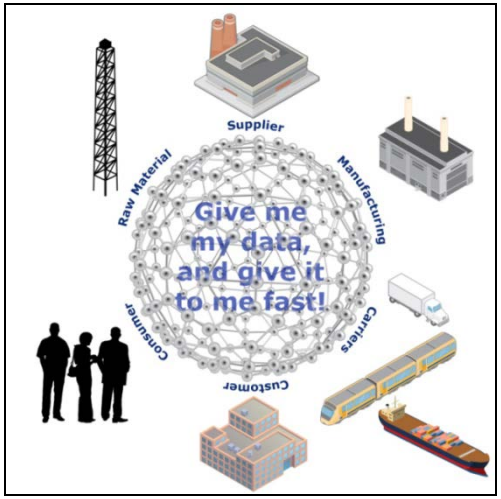
Fundamental Challenges and Best Practices

- Quality Assurance
- Access, Security, Privacy
- Storage and Backups
- Operations and Maintenance
- Critical Failures



Quality Assurance: *Key Insights*

- Collect redundant data from multiple sensors
  - Data can be combined so that false positives are filtered out
- Use standard industry reference files when possible
  - Reduces erroneous information
- Data quality is highly industry dependent
  - Choice between high-output, real-time data, and scrubbed, pseudo real-time data
  - Greater the overall veracity required, the more process-intensive it is to enforce that veracity, and the higher the delay in disseminating the data
  - Typically, fast, general data-quality analysis for real-time systems, and thoroughly scrubbed and sanitized data for historical and post real-time analysis and display



Timely data is worth nothing, if it is inaccurate

Access, Security, Privacy: *Key Insights*

- Access is most often controlled by the holder of the data
- Systems have been designed to see that the right people have access only to the data they need
- Access to the data is usually password protected
- Protection of the **source** of the data is highly protected
  - Within the internet search industry, it is so highly protected that there is no concrete evidence of exactly how it is protected
  - Access and security regarding the **ability** to search is completely disregarded since the ability to perform a search is of paramount importance in the internet search industry

Storage and Backups: *Key Insights*

- Determine what data needs to be stored and for how long
  - Once the system is out of test mode, is there any need to retain all that information?
  - Most industries do not have hard-rule on how this should be done
  - In aviation industry, data are typically kept for only a brief period of time before being discarded. If incident occurs, data are spooled off for review before being destroyed
- Frequent backups and storage off site is typical
  - Google (Search Engines) does real-time streaming backups across multiple data centers in order to ensure that searches are always available
- Perform preventative maintenance regularly
- Consider allowing a third party to handle data storage needs
  - Cost of keeping **all** traffic data may be prohibitive for the government, but profitable for a third party

Operations and Maintenance: *Key Insights*

- Start small with an implementation which addresses the most critical needs, defined either geographically or by category of information
  - Focus first on known critical data-elements first to ensure that you have the capacity, ability, and availability of those items, before expanding
  - Take a lesson from the search engine industry – great systems are built slowly over time from small, hardened implementations
  - Ensure that core competencies are answered before trying to do everything
- Build for scalability
  - Avoid situation where a system is built to perform very well for a test setup, but does not scale well in the real-world
  - Leverage technologies such as clustered databases, virtual warehousing, virtual servers, etc.
- Use multiple servers to distribute load for real time databases
  - Databases can grow quite large very quickly
  - Easier to solve the problem once approximate data sizes and elements have been defined

Operations and Maintenance: *Key Insights (cont.)*

- Determine level (granularity), amount, and transmission frequency of data are needed
  - Data overload will negate the usefulness of the data
  - Data overload can cause critical messages to be overlooked
  - Avoid an overreaction or early reaction based on small sample sizes
- Determine what is critical to communicate
  - In airline industry, alert systems do not collect all continuously streamed data; only data needed to alert an operator of a problem
- Make data available as soon as feasible
  - Even if more processing needs to occur in the background, providing a real-time feed for current data is an attractive option to users

Critical Failures: *Key Insights*

- Do not be dependent on a single person to rectify a critical failure
  - Flight-Plan System Crash in 2009 was due to failure of a single part, which was easily replaceable; but there was only **one** technician who was qualified to do it, and it took over six hours for him to arrive and make the repair
- Systems that need to be highly available will necessitate elevated labor costs
  - If any failure of the system can be catastrophic, then it is necessary to keep round-the clock staff to fix issues

Recommendations for Promising DCM Innovations

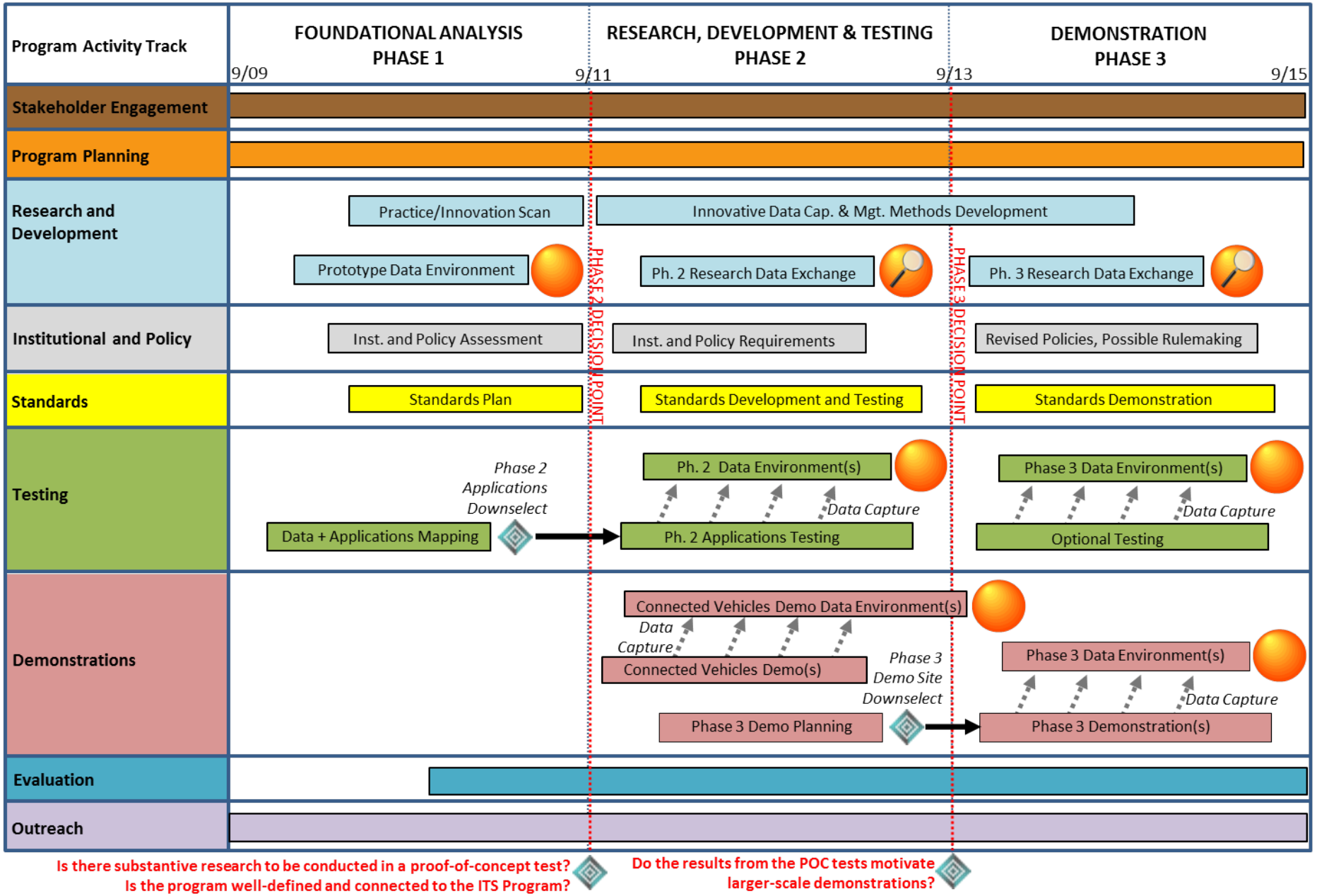
- Explore DIDC concept
  - Develop one or more prototype DIDC applications to capture data from mobile sources (vehicles, travelers, etc.)
- Examine crowdsourcing
  - Development of transformative mobility applications (e.g., multi-modal traffic signal system, transit signal priority, queue warning, speed harmonization)
  - Data collection (e.g., travel times, queues)
- Investigate Cloud Computing
  - Examine strengths/benefits of cloud types (public, private, etc.)

Next Steps

- Assess recommendations for innovative concepts and methods, and downselect promising ideas

- Issue solicitations for developing and testing selected innovations

Data Capture and Management Program: High-Level Roadmap



High-Level Roadmap v1.3 (5/9/2011)

LEGEND: Data Capture Data Feed; Decision point; Data Environment; Research Data Exchange; Program Activity

Getting Involved

- Got an interesting concept or method for data capture and management?
  - contact Mohammed Yousuf

- Respond to upcoming procurements for:
  - further research and development of innovative data capture and management concepts
  - building Phase 2 data environments to enable development of mobility applications

- Participate in future stakeholder engagement activities (e.g., users needs meetings) and provide feedback on direction of the Mobility Program

For More Information

Mohammed Yousuf
FHWA Office of Operations R&D
Mohammed.Yousuf@dot.gov

Dale Thompson
US DOT ITS Joint Program Office
Dale.Thompson@dot.gov

US DOT ITS Joint Program Office Web site
www.its.dot.gov