NCHRP 17-100

Leveraging Artificial Intelligence and Big Data to Enhance Safety Analysis
Interim Report

Yinhai Wang

on behalf of the NCHRP 17-100 Research Team

for

ACS20(1) Safety Analytical Methods Subcommittee Meeting

Jan 9, 2023

Contact: yinhai@uw.edu; +1-206-616-2696

Project Objectives

- Advance the use of AI, ML, BD in safety analysis and assess their effectiveness to support safe system, priority decision-making, performance tracking
- Identify potential data sources
- Identify or develop the requisite data preparation and extraction algorithms
- Document each source's attributes
- Develop guidance for managing data using a format that can be accessed by various tools

Tasks Overview

- Phase I – planning
  - Task 1: conduct literature review
    - AI/ML algorithms, BD and unconventional data sources, practices in supporting safe system and model priority decision-making and performance tracking of road safety features
  - Task 2: identify source data needs and their attributes
  - Task 3: prepare Phase II work plan
    - Suitable framework, algorithm development, validation, and data management plan
  - Task 4: interim report
- Phase II – methodology, pilot projects and research product development
  - Task 5: execute Task 3 work plan
  - Task 6: develop a detailed process and conduct pilot projects
  - Task 7: develop a user's guide
  - Task 8: develop an approach and multi-media materials
  - Task 9: prepare and submit final deliverables

Research Team

- University of Washington – Dr. Yinhai Wang and PhD student
- Texas Tech University – Dr. Venky Shankar and PhD student
- DKS Associates – Brian Chandler and Dr. Lacy Brown
- University of South Florida – Dr. Fred Mannering
- Gates Logic – Janet Gates

Task 1 Literature Review

- 2,366 papers from Scopus.com (1975 – 2021)
  - 340 of them are related to traffic safety
- Topic Modeling
  - Identified topics

Safety Topic	Topic Description	Number of papers
0	Text analysis	20
1	Pedestrian safety at crossings and non-crossings	16
2	Neural networks/ML	122
3	Collision avoidance	29
4	Conflict analysis	16
5	Crash severity analysis	73
6	Driving scenario	19
7	Computer vision	47
8	Incident management	38
9	Rail safety	3
10	Naturalistic and simulation analysis	23
11	Driving behavior analysis	34

Task 1 Literature Review

- Topic Modeling (continued)
  - Summarize and discuss in detail
    - Text analysis
    - Object detection and sensing
    - Reinforcement learning
    - Self-supervised decision system
- Big data and traffic safety
  - Traffic safety & BD papers from TRID

Task 1 Literature Review (cont'd)

Data source	Sub-source	Spatial/ user coverage	Frequency of collection	Typical dataset ownership	Example studies
Video					
	Roadside	Intersection	Sub-second	DOTs	(Sayed et al. 2012) (City of Bellevue 2020)
	Aerial	Segment		Research	(Outay, Mengash, and Adnan 2020) (A. Y. Chen et al. 2020) (Wu et al. 2020)
	Onboard	Neighborhood or inside of vehicle		Companies/ Research	(Ke et al. 2020) (Zhu, Wang, and Hu 2020)
LiDAR					
	Roadside	Intersection	Sub-second	DOTs/Companies/ Research	(Lv et al. 2019) (Zhenyao Zhang et al. 2019)
	Onboard	Neighborhood of vehicle		Companies/ Research	

Task 1 Literature Review (cont'd)

Data source	Sub-source	Spatial/ user coverage	Frequency of collection	Typical dataset ownership	Example studies
Crowd-sourced					
	Location-based traffic data	Fleet/user population	Seconds to minutes	Companies	(Ash 2021)
	Pedestrian foot traffic	Regional	Seconds to minutes	Companies	(Juhász and Hochmair 2020)
	weather	Regional	Seconds	Companies/Public	
	Crash report/ social media	Area	Minutes to hours	Companies/Public	(Gu, Qian, and Chen 2016)
	Dangerous driving event	Regional	Seconds to minutes	Companies/ Research	(Yang et al. 2019)
	Conflicts in active mode	Regional	Seconds to minutes	Companies/DOTs/ Research	(Mattingly, Casey, and Johnson 2017)
	Roadway infrastructure health	Regional	Sub-seconds	Research	(Gupta, Khare, et al. 2020) (Gupta, Hu, et al. 2020)
Telematics and Connected Vehicle data		Fleet	Sub-second	Companies/ Research	(Yang et al. 2021) (Saldivar-Carranza et al. 2021) (Desai et al. 2021) (Hunter et al. 2021)

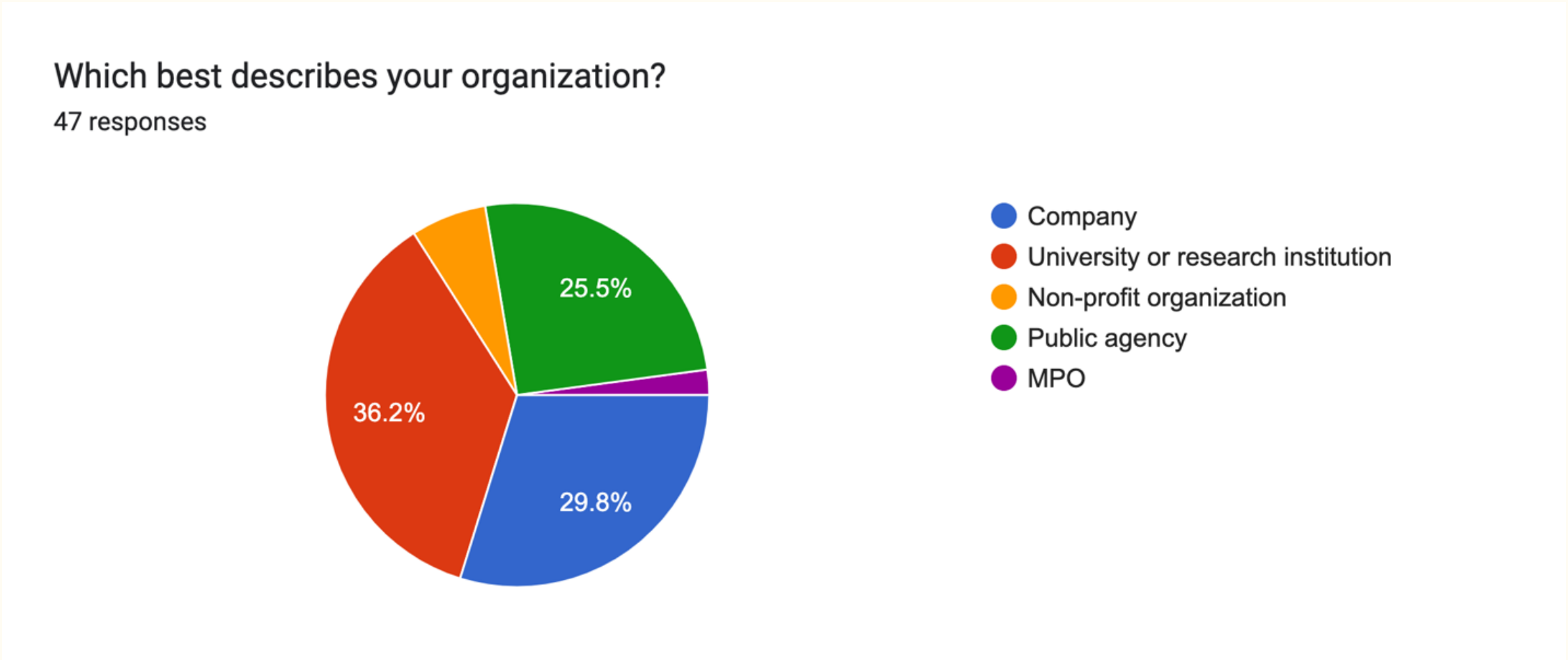
Task 2 Data Source Identification

- Data survey of 11 questions
  - 11 TRB committees
    - AED10 Statewide Data, AED20 Urban Trans Data, AED30 Information Sys & Tech, AED40 GIS, AED50 AI & AC Committee, AED60 Statistical Methods, AED70 Freight Data, ACS10 Safety Management, ACS20 Safety Analysis, ACS60 Truck & Bus Safety, AKD20 Roadside Safety
  - Two ASCE committees
    - ASCE Safety Committee
    - ASCE AI Committee
  - Two ITE committees
    - ITE Transportation Safety Council
    - ITE Vision Zero Committees

Scan this QR code to enter the survey

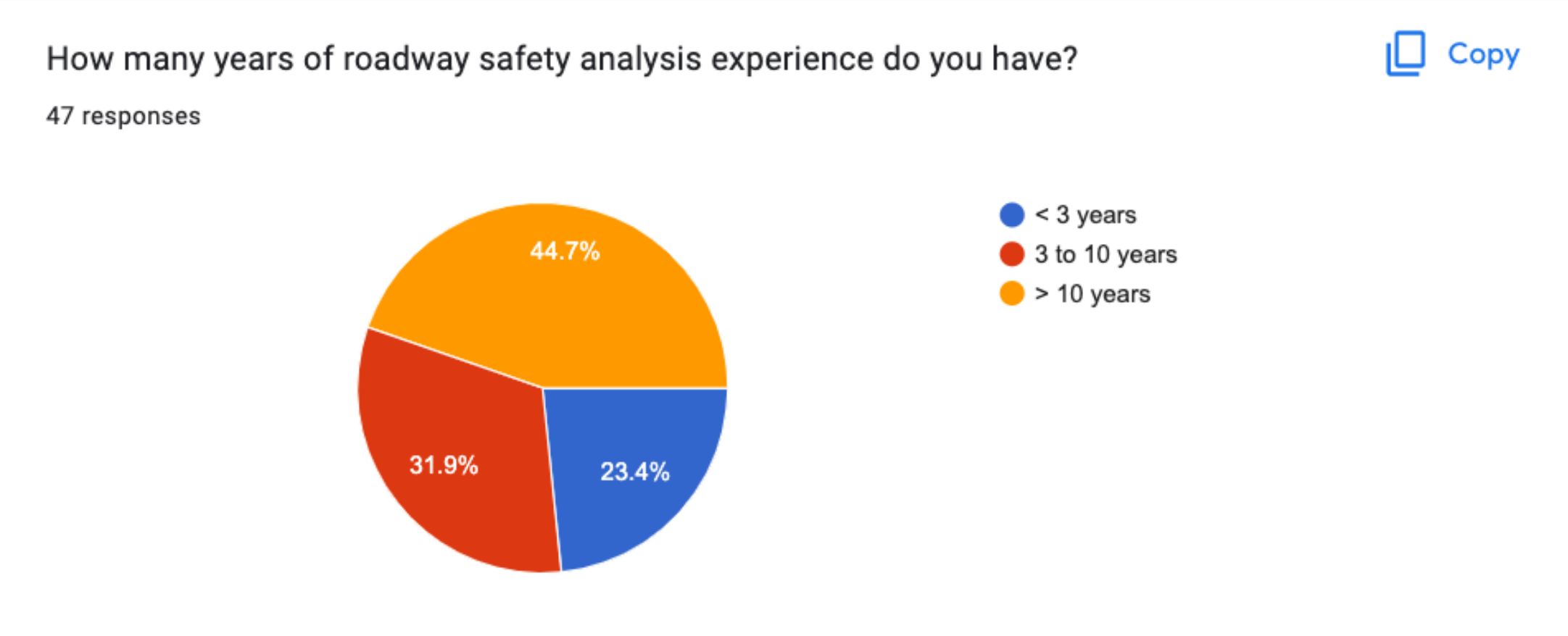
Task 2 Data Source Identification (cont'd)

- Which best describes your organization?



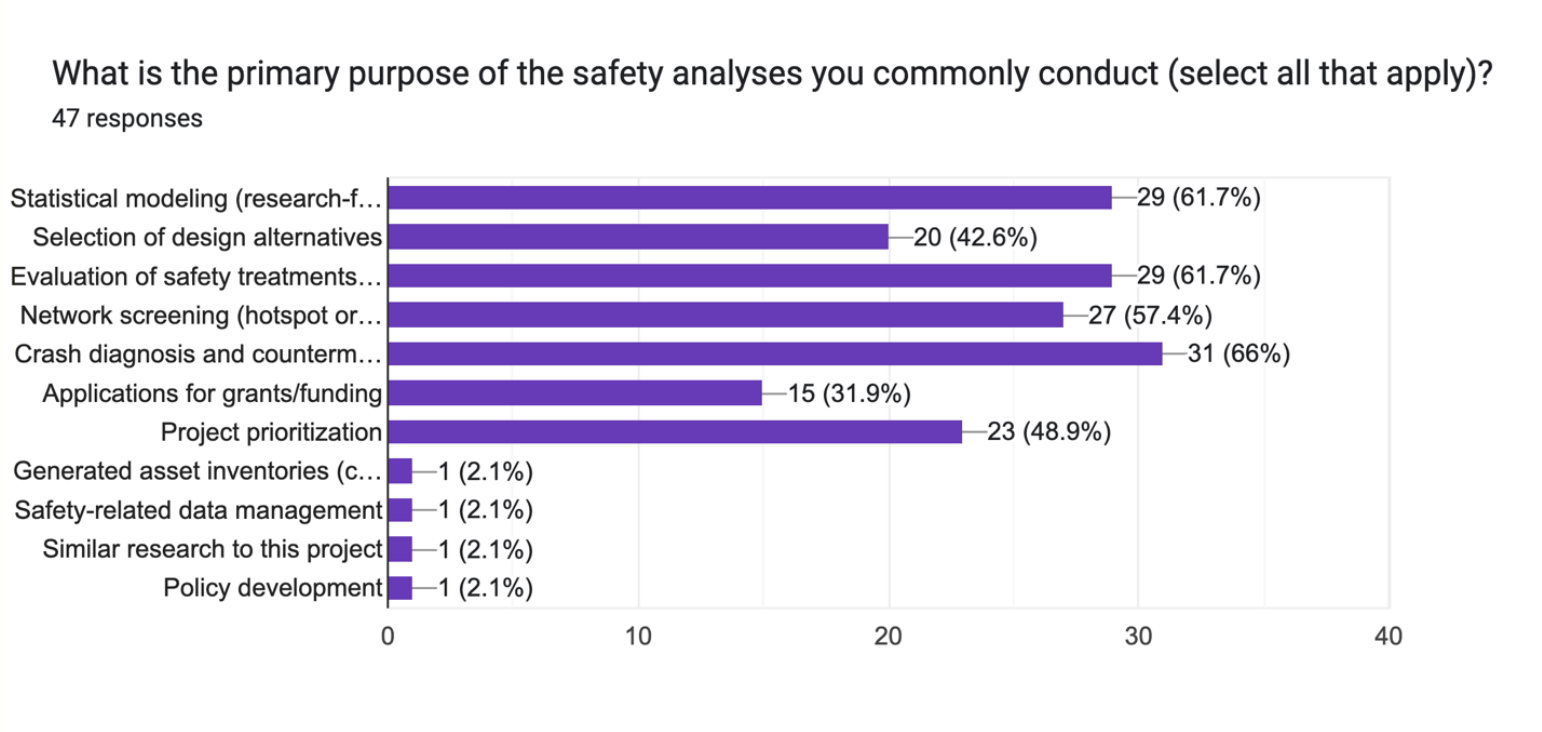
Task 2 Data Source Identification (cont'd)

- How many years of roadway safety analysis experience do you have?



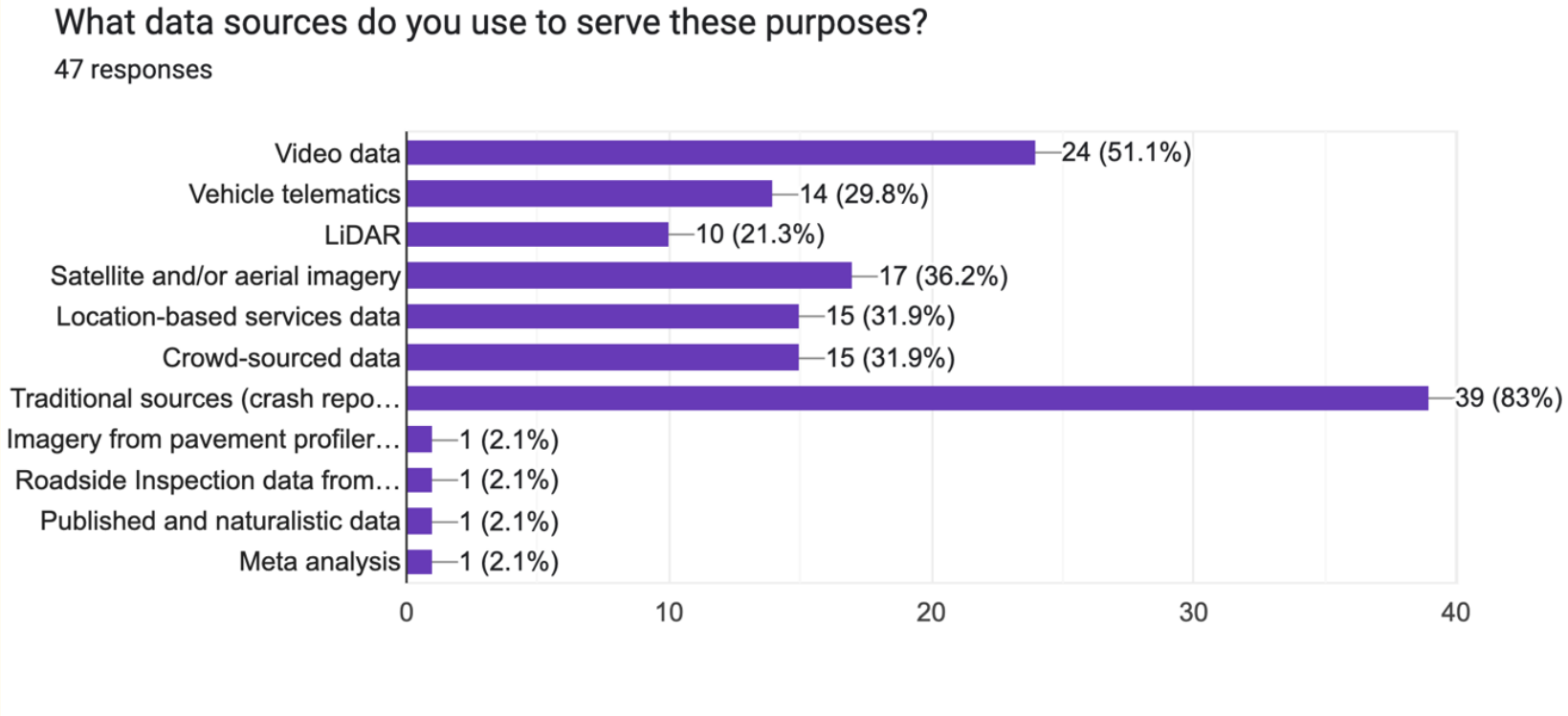
Task 2 Data Source Identification (cont'd)

- Primary purpose of the safety analyses you commonly conduct?
  - 67% – crash diagnosis and countermeasure selection;
  - 61.7% – statistical modeling (research-focused)
  - 61.7% – evaluation of safety treatments (post-installation)
  - 57.4% – network screening (hotspots or systemic)



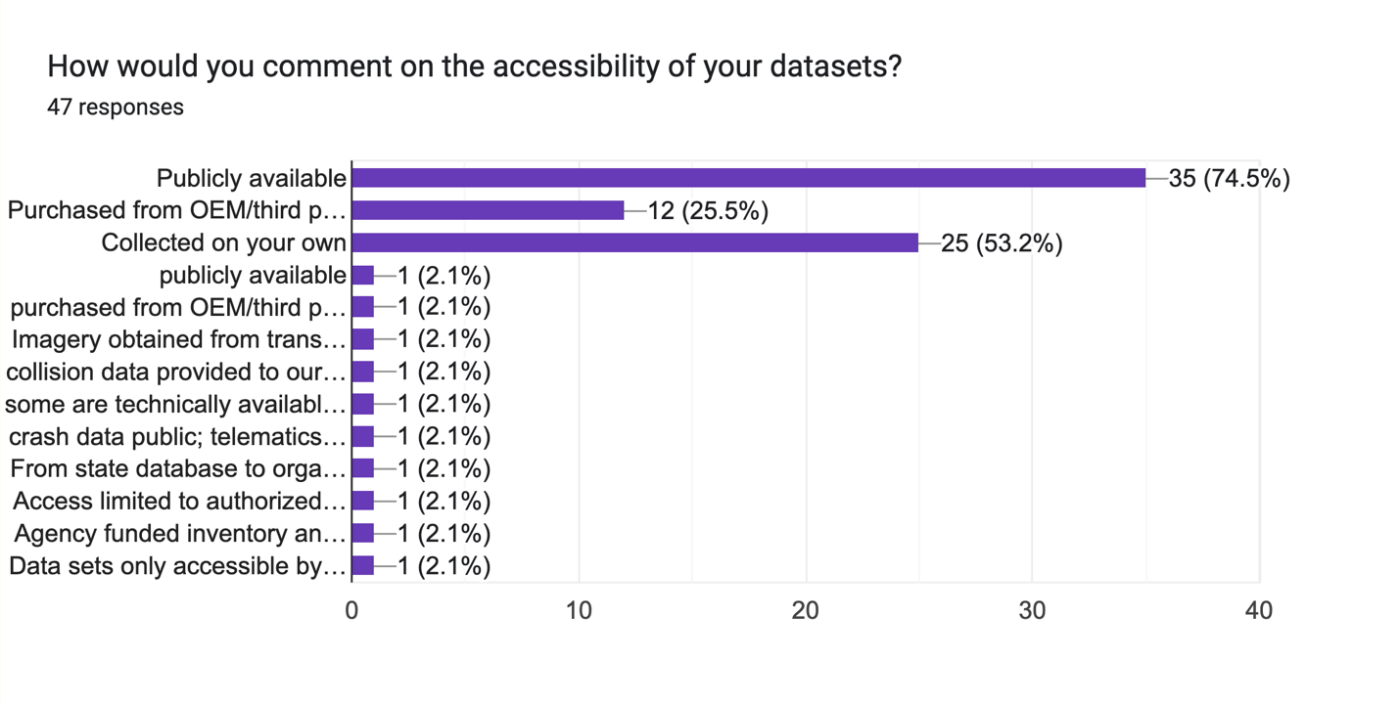
Task 2 Data Source Identification (cont'd)

- What data sources do you use to serve these purposes?
  - 80% – traditional sources (crash reports, roadway geometry features, traffic flow/counts from traditional sensors, weather conditions, land use, demographic and census data)
  - nearly 30% – already have experience using vehicle telematics, location-based services, and crowd-sourced



Task 2 Data Source Identification (cont'd)

- How would you comment on the accessibility of your datasets?
  - Not surprising to see that most datasets being used are from public sources. Traditional sources are typically collected and made available by national and state DOTs, as well as related public record systems
  - More than half respondents said they also collect data on their own, while only 25% of them said they purchased data from a third party



Task 2 Data Source Identification (cont'd)

- What are the public datasets?
  - HSIS
  - State DOT/public record system
  - FARS (Fatality analysis reporting system, yearly data)
  - CRSS (Crash Report Sampling System, "nationally representative probability sample selected from the estimated 5 to 6 million police-reported crashes that occur annually")
  - Signal4 (only Florida)
  - RITIS
  - CFM clearinghouse
  - IIHS (Insurance Institute for Highway Safety, non-profit)

Task 2 Data Source Identification (cont'd)

- If you purchased proprietary data, who was the vendor and what price did you pay?
  - Video Analytics Vendors, range from $2k-5k per intersection
  - Streetlight (for planning activities, not for safety), INRIX, HERE, Wejo
  - "Wejo for telematic, Transoft for video conflicts"
  - "StreetLight, $1.3 Million"

Task 2 Data Source Identification (cont'd)

- Are there any gaps between the data and what you are trying to solve? If yes, what are the gaps?
  - Cheaper access to connected vehicle inputs
  - Vehicle occupancy data
  - Good coverage outside of interstate and major roads
  - Actual vehicle travel speeds during collision
  - Lack of large-scale vehicle trajectory data for a long period of time
  - Lack of high accuracy data, e.g., information lost to truncation from states to USDOT
  - "Sometimes the vendor sold big data are black boxes - sample size for example. How to remedy or how the vendors remedied those problems is mystery."

Task 2 Data Source Identification (cont'd)

- What is your most desired dataset(s) that you don't currently have access to? How do you think they can improve your analyses??
  - Vehicle occupancy data, volume, and CV hard braking/hard acceleration data
  - Additional big data for predictive rather than reactive efforts
  - Inventory of roadway assets
  - Real-time travel time
  - How many apps/vehicles/devices contributed continuously to this data and the user groups
  - Pedestrian and bicycle counts

Task 2 Data Source Identification (cont'd)

- Sample data analysis
  - Connected vehicle data
    - INRIX, Wejo samples
    - Vehicle movements and driving events (e.g., abrupt acceleration changes, seat occupancy changes, ABS/AEB activation, light and signal, status)
    - Much better temporal and spatial granularity
  - Active mode users
    - ***Very hard to find low-cost, scalable, and geographically diverse dataset***

Task 2 Data Source Identification (cont'd)

- Conventional sources
  - Crash reports, roadway geometry features, weather conditions, land use, demographic and census data, traffic flow/counts from traditional sensors
  - HSIS, state crash reporting system, RITIS, DRIVE Net
- Connected vehicle data
  - Examples: Wejo, INRIX, StreetLight
- Video analytics
  - Examples: Cameras at urban intersections
- LiDAR
  - Examples: roadside LiDAR
- Social media
  - Examples: Twitter for crash identification

Task 3 Phase II Work Plan

- Available datasets
  - Safety datasets from HSIS and public record systems
  - Traffic flow and counts from roadway sensors, e.g., DRIVE Net and RITIS
- Candidate datasets
  - Connected vehicle datasets, e.g., Wejo, INRIX & GM
- Phase II work plan (at a glance)
  - Task 5 Execute Task 3 Work Plan
  - Task 6 Develop a Detailed Process and Conduct Pilot Projects
    - WSDOT and Maryland DOT
  - Task 7 Develop a User's Guide
  - Task 8 Communicating the Research Products to Decision Makers

Task 3 Phase II Work Plan (cont'd)

- Task 5 Detailed Work Plan
  - Setting up data repository and populating data structure
    - Roadway inventory, traffic sensor, connected vehicle (CV) data, weather, crash
  - Analyzing and processing collected data
    - Quality control: trajectories processing, speed benchmarking
    - Map matching of trajectories: GPS granularity, cloud computing tools
    - Extract safety events and contexts
      - Filter hard braking events from CV data for true near-miss events
      - Understand representativeness of the derived data
      - Extract contexts for these events: traffic (geometry, flow before/after the event), human factors (speed, signaling, occupancy)
    - Create new variables
      - Reaggregate these events and their contexts into segment- and intersection-based variables

Task 3 Phase II Work Plan (cont'd)

- Task 5 Detailed Work Plan
  - Developing models
    - Benchmark newly created variables of their spatiotemporal patterns and against existing metrics: unsupervised and regression models (flow metrics vs. AADT)
    - Frequency-based modeling: include both classical and newly constructed real-time variables
    - Severity-based modeling
      - Based on individual events and their contexts
      - Include near-miss as a severity level: 4 levels (fatality, injury, property damage only, near-miss) or 2 levels (crash, near-miss)
      - Supervised and econometric models

Task 3 Phase II Work Plan (cont'd)

- Task 6 Detailed Process and Pilot Projects
  - Summarize framework and associated AI/ML algorithms and models
  - Confirm data needs
  - Solicit pilot project partners and conduct pilot studies
    - WSDOT, Maryland DOT
- Task 7 Develop a User's Guide
  - Definitions and Descriptions
  - Conventional Data for Safety Analysis
  - Unconventional Data for Safety Analysis
    - Potential data sources and their attributes (requisite preparation, frequency, granularity, cost, quality, restrictions in access)
  - Procedures
  - Case studies and examples

Thank you for your attention!