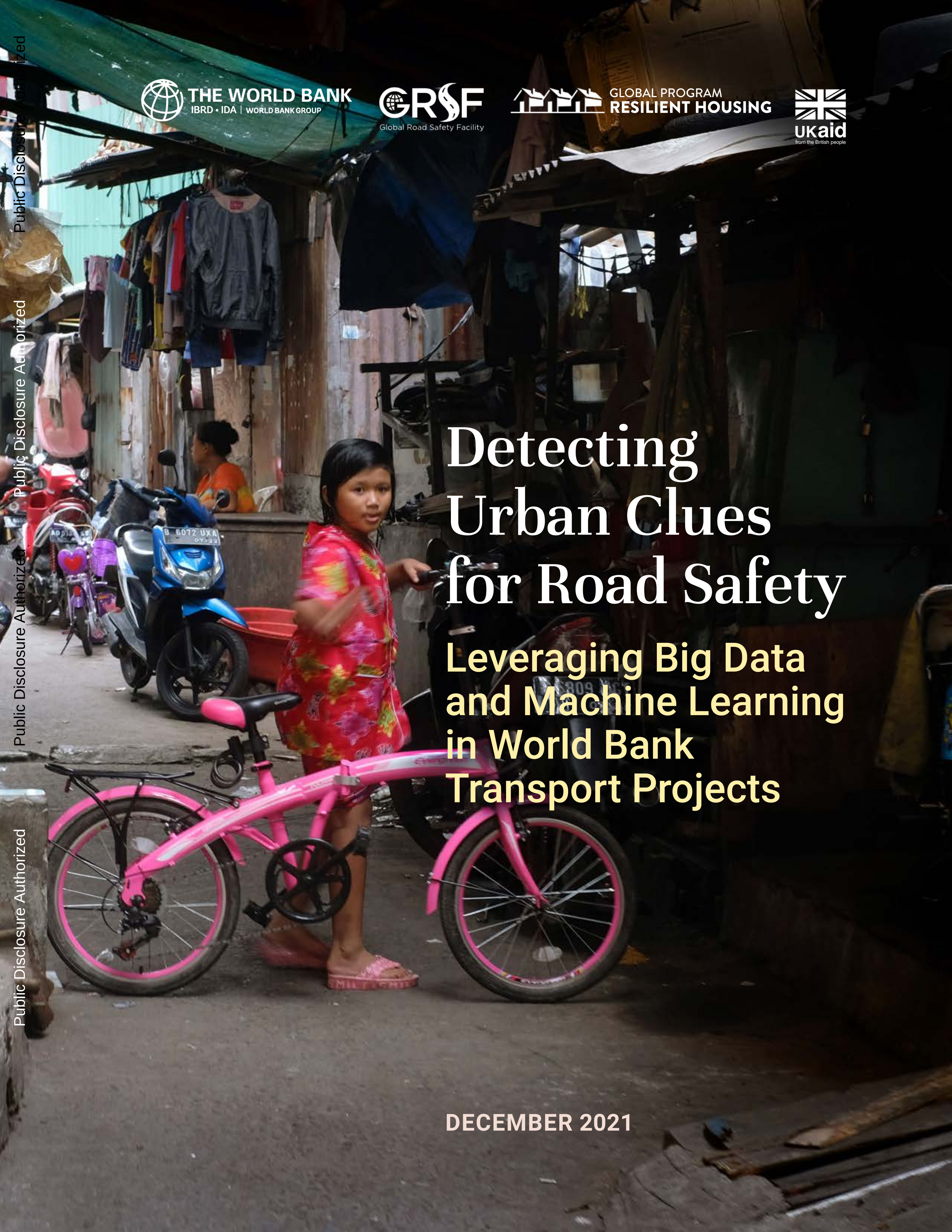
THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

GRSF
Global Road Safety Facility

GLOBAL PROGRAM
RESILIENT HOUSING

UKaid
from the British people



Detecting Urban Clues for Road Safety

Leveraging Big Data and Machine Learning in World Bank Transport Projects

DECEMBER 2021

Table of Contents

List of Figures

List of Tables

Acknowledgments

This Guidance Note was prepared by a team from the Global Program for Resilient Housing at the World Bank. The team was led by **Sarah Elizabeth Antos** (Data Scientist) and **Luis Miguel Triveno Chan Jan** (Senior Urban Development Specialist). Overall managerial support was provided by **Francis Ghesquiere** (Practice Manager, Urban EAP) and **Radoslaw Czapski** (Senior Transport Specialist).

The core team included **Jessica Gosling-Goldsmith, Charles Wang, Bushra Syed Shafat Ali,** and **Sebastian Anapolsky.**

The Global Program for Resilient Housing supports safe and resilient housing by creating new, cost-saving tools to evaluate homes from the air and the street to help identify those vulnerable to natural and health hazards. While the program focuses on housing, it developed a methodology to extract urban clues from street view imagery with multiple applications including those related to urban mobility and road safety.

The note incorporates valuable input and review from **Holly Krambeck** (Program Manager), **Said Dahdah** (Lead Transport Specialist), **Satoshi Ogita** (Senior Transport Specialist), **Veronica Ines Raffo** (Senior Infrastructure Specialist), **Li Qu** (Senior Transport Specialist), and **Glenn S. Morgan** (ESF Consultant).

During the drafting of this note several industry experts were interviewed. The team would like to express gratitude for the external inputs of: **Anthony Germanchev** (Principal Professional Leader, Advanced Technologies Lab, Australian Road Research Board), **David Hynd** (Chief Scientist, TRL), **Monica Olyslagers** (Safe Cities and Innovation Specialist, iRAP), Professor **George Yannis** (National Technical University of Athens), and **Spencer Rigler** (Account Director, TRL).

Design was done by **Xavier Conesa.**

This note would not have been possible without generous support from the **Global Road Safety Facility** and **UK Aid.**

Objective, Audience and Structure

The purpose of this Guidance Note is to provide concrete guidance on how big data and machine learning (ML) can be leveraged in road safety analysis. The document presents opportunities to use these new technologies to improve current road safety assessment procedures across the project cycle, in accordance with the World Bank's latest Environmental and Social Framework (ESF) guidelines.

This Guidance Note is for World Bank task teams who are interested in using new data sources and analytical methods for road safety analysis across various types of projects. In addition, researchers, road safety experts, data scientists, and government agencies responsible for road safety assessments, transportation management, and infrastructure development would also find this document useful to understand how these new technologies can be implemented across World Bank investment projects.

This document consists of three parts. Part 1 discusses the World Bank's current guidelines for incorporating road safety analysis across the project cycle, examines existing data and approaches and identifies opportunities to improve current methods using big data and ML. Part 2 provides an overview of these new technologies and concrete guidance on how they can be integrated into World Bank projects. Part 3 presents case studies on two regions of interest – Bogotá, Colombia and Padang, Indonesia – to demonstrate how ML can be implemented to evaluate road safety. The document concludes with recommendations for using big data and ML in road safety assessments in the future.

Abbreviations

ADB	Asian Development Bank
API	Application Programming Interface
DDP	Development Data Partnership
DL	Deep Learning
DRIVER	Data for Road Incident Visualization, Evaluation and Reporting
ESCP	Environmental and Social Commitment Plan
ESF	Environmental and Social Framework
FSI	Fatalities and Serious Injuries
GRSF	Global Road Safety Facility (World Bank)
ICR	Implementation Completion Report
IoT	Internet of Things
iRAP	International Road Assessment Programme
ITS	Intelligent Transport System
LMICs	Low- and Middle-Income Countries
ML	Machine Learning
OPTRSR	Overall Project Traffic and Road Safety Risk
OSM	OpenStreetMap
PCN	Project Concept Note
PDO	Project Development Objective
RIC	Road Information Collector
ROI	Region of Interest
RRE	Road Risk Evaluator
RSA	Road Safety Audit
RSI	Road Safety Inspection
RSIA	Road Safety Impact Assessment
RSO	Road Safety Observatory
RSSAT	Road Safety Screening and Appraisal Tool
SDGs	Sustainable Development Goals
TTL	Task Team Leader
UAV	Unmanned Aerial Vehicle

Introduction

Transportation services and infrastructure connect people, businesses, and places. They allow citizens to access opportunities, such as jobs, education, health services, recreation, and enable the movement and distribution of goods. As a result, transport services and infrastructure are key to the economic development of cities and regions.[1]

While the development of transportation systems and infrastructure is vital to economic growth, it is also important to evaluate and mitigate its potential negative externalities and costs to society.[2] According to the World Health Organization (WHO), around 1.25 million people are killed on the world's roads every year and between 20 and 50 million are seriously injured. These costs are disproportionately higher in low- and middle-income countries (LMICs), which are estimated to endure 93 percent of the world's fatalities on the road, despite having 60 percent of the world's vehicles (figure 1).[3] According to a 2019 study of select countries, road crashes cost World Bank client countries an estimated 7 percent to 22 percent of their GDP over a 24-year period.[4]

Road fatalities and injuries are predictable and preventable.[5] Research indicates that roughly 70 percent of serious crashes are due to simple and unintentional errors of perception or judgement.[6] The most vulnerable road users are pedestrians, bicyclists, and motorcyclists, accounting for more than 50 percent of reported fatalities in LMICs.[7] Effective transport planning and management that carefully considers and incorporates measures to address safety risks.[8] Speed reductions and the design of infrastructure to promote safer streets have demonstrated clear results in Colombia and India. In Bogotá, Colombia, the speed management program resulted in a 21 percent decrease in traffic fatalities compared to the average for the three preceding years (2015-18).[9] In India, Pune has become a regional leader in complete streets, in which streets are designed for all users, rather than only for cars; pedestrians, cyclists, motorists, and transit riders are given safe access with the complete streets approach.[10]

The World Bank is a key supporter of the United Nations (UN) Decade of Action for Road Safety and related Sustainable Development Goals (SDGs). These include SDG 3.6, which seeks to reduce deaths and injuries from road crashes by 50 percent, and SDG 11, which focuses on making cities and human settlements inclusive, safe, resilient, and sustainable. The World Bank is also a proponent of

[1] World Bank, *Mobile Metropolises: Urban Transport Matters: An IEG Evaluation of the World Bank Group's Support for Urban Transport* (Washington, DC: World Bank, 2017).

[2] Word Bank, *Making Roads Safer* (Washington, DC: World Bank, 2014).

[3] WHO (World Health Organization), *Global Status Report on Road Safety 2018* (Geneva: World Health Organization, 2018), 4.

[4] World Bank, *The High Toll of Traffic Injuries: Unacceptable and Preventable* (Washington, DC: World Bank, 2017).

[5] Makhtar Diop, "All Road Deaths Are Preventable. We Can Make It Happen," World Bank, accessed May 14, 2021, https://blogs.worldbank.org/transport/all-road-deaths-are-preventable-we-can-make-it-happen

[6] International Transport Forum, *Zero Road Deaths and Serious Injuries: Leading a Paradigm Shift to a Safe System* (Paris: OECD Publishing, 2016). https://doi.org/10.1787/9789282108055-en

[7] World Bank, *Good Practice Note on Road Safety* (Washington, DC: World Bank, 2019). https://pubdocs.worldbank.org/en/648681570135612401/Good-Practice-Note-Road-Safety.pdf

[8] International Transport Forum, "Best Practice for Urban Road Safety: Case Studies," *International Transport Forum Policy Papers,* no. 76 (2020).

[9] International Transport Forum, "Best Practice for Urban Road Safety: Case Studies."

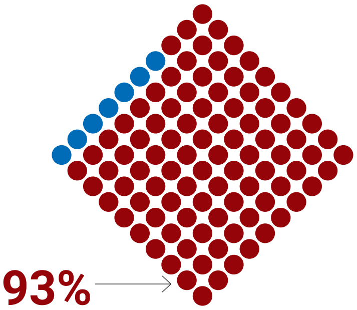
[10] Institute for Transportation and Development Policy, "Pune, India Wins 2020 Sustainable Transport Award," last modified June 27, 2019, https://www.itdp.org/2019/06/27/pune-india-wins-2020-sustainable-transport-award/

the Sustainable Mobility for All (SM4A) initiative, which highlights safety as one of the pillars of sustainable mobility.[11]

The World Bank hosts the Global Road Safety Facility (GRSF) to provide funding, knowledge, and technical assistance to help developing countries create safer roads. The Facility addresses road safety issues across a wide range of projects, from infrastructure design and vehicle safety to traffic law enforcement, post-crash response systems, data collection, and institutional strengthening. Since its inception in 2006, the Facility has disbursed a total of USD 44.6 million to improve road safety in 64 countries.

It is important, and often required, to incorporate road safety management procedures in transport projects to identify and mitigate risks in a timely manner. Governments, international development organizations, and other agencies have established various tools and systems to facilitate road safety analysis. However, the absence of valid, representative data presents significant challenges to developing a good understanding of road safety risks and reducing crash fatalities and injuries through data-driven, evidence-based interventions.[12]

FIGURE 1: Road safety is a serious concern in low- and middle-income countries



93%

of road fatalities occur in low- and middle-income countries, despite these countries having 60 percent of the world's vehicles.

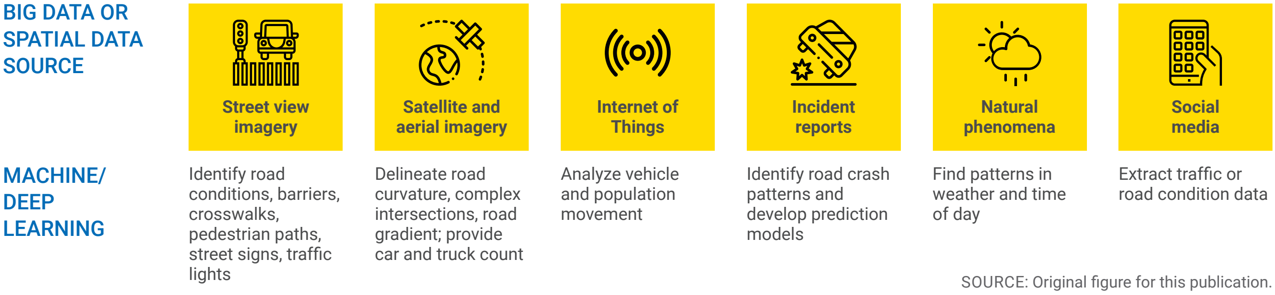
SOURCE: Original figure for this publication, based on data from WHO.

[11] World Bank, *Good Practice Note on Road Safety*, 1.

[12] World Bank, *Guide for Road Safety Opportunities and Challenges: Low and Middle Income Country Profiles* (Washington, DC: 2020). https://openknowledge.worldbank.org/handle/10986/33363

New technologies such as big data and machine learning (ML) provide promising opportunities to improve existing data sources and methods for road safety analysis. From analyzing anonymized GPS data to understand traffic flows in the Philippines to partnering with data providers that crowd-source information about crash sites in Kenya, governments, World Bank task teams, and other stakeholders are adopting innovative approaches to identify, monitor, and mitigate fatalities and injuries in high-risk areas.[13] Unsupervised learning techniques have been applied in Lima, Peru, using records of different crash types to identify safe areas along routes and safer pedestrian pathways, decreasing the likelihood of pedestrians suffering an crash.[14] The Urban Traffic Modeling and Control project at the National University of Medellín has been using deep learning (DL) techniques to classify traffic and identify motorbike usage. In Cartagena, Colombia, data mining and ML algorithms were used to analyze road records and predict the severity of traffic crashes using classification algorithms.[15] Figure 2 provides an overview of the potential uses of big data and ML in road safety analysis that will be discussed in this note.

FIGURE 2: **Potential applications of big data and ML in road safety projects**



BIG DATA OR SPATIAL DATA SOURCE	Street view imagery	Satellite and aerial imagery	Internet of Things	Incident reports	Natural phenomena	Social media
MACHINE/ DEEP LEARNING	Identify road conditions, barriers, crosswalks, pedestrian paths, street signs, traffic lights	Delineate road curvature, complex intersections, road gradient; provide car and truck count	Analyze vehicle and population movement	Identify road crash patterns and develop prediction models	Find patterns in weather and time of day	Extract traffic or road condition data

SOURCE: Original figure for this publication.

[13] World Bank, "Open Traffic Data to Revolutionize Transport," last modified December 19, 2016, https://www.worldbank.org/en/news/feature/2016/12/19/open-traffic-data-to-revolutionize-transport; Guadalupe Bedoya Arguelles, et al., "Smart and Safe Kenya Transport (SMARTTRANS)" (Washington, DC: World Bank, 2019), https://documents1.worldbank.org/curated/en/723411574361015073/pdf/Smart-and-Safe-Kenya-Transport-SMARTTRANS.pdf

[14] Jesús Lovón-Melgarejo et al., "Identification of Risk Zones for Road Safety through Unsupervised Learning Algorithms," in *16th LACCEI International Multi-Conference for Engineering, Education, and Technology: Innovation in Education and Inclusion*, http://www.laccei.org/LACCEI2018-Lima/full_papers/FP413.pdf

[15] Holman Ospina-Mateus et al., "Using Data-Mining Techniques for the Prediction of the Severity of Road Crashes in Cartagena, Colombia," in *Applied Computer Sciences in Engineering*, eds. J. Figueroa-García et al., vol. 1052 (2019): 309-20, https://doi.org/10.1007/978-3-030-31019-6_27

PART 1:
The Demand for Data to Assess Risks and Conduct Safety Assessments

1.1 Assessing Road Safety Across the Project Cycle

World Bank projects follow a project cycle to design, prepare, implement, support, and evaluate projects. The project cycle identifies six stages between project identification and project completion (see figure 3).[16] Bank staff work closely with developing country borrowers throughout the project cycle to ensure that projects meet relevant World Bank economic, financial, procurement, and environmental and social standards.

The World Bank has adopted seven pillars to identify key priorities for road safety interventions. These pillars, that aim at preventing road crashes, fatalities, and injuries across all projects include: Road Safety Management, Safer Roads and Mobility, Safer Vehicles, Safer Roads Users, Post-Crash Response, Safer Speeds, and Reduced Exposure. The first five pillars are from the UN Global Plan for Road Safety with the last two new pillars added for the Road Safety GPN.[17] Road safety objectives of World Bank projects should be aligned with these pillars and performance indicators must track progress against them.

All World Bank investment projects are required to follow the World Bank's Environmental and Social Framework (ESF), which went into effect on October 1, 2018. The ESF is a set of operational policies and procedures designed to ensure that projects are economically, financially, socially, and environmentally sound. The ESF includes protections for people and the environment from potential adverse risks and impacts that could arise from Bank-financed projects and promotes sustainable development. Within the ESF, ten Environmental and Social Standards (ESS) set out a range of responsibilities for Borrowers designed to help them manage project risks and impacts. In addition, the standards aim to improve environmental and social performance, consistent with good international practice and national and international obligations.

The World Bank's ESF calls for road safety risks to be considered in all investment projects. As relevant, Borrowers are required to undertake technical assessments and implement operational measures to avoid or minimize community exposure to project-related traffic and road safety risks. In the context of the ESF, road safety assessments are carried out as part of a project's Environmental and Social Assessment (ESA). The overall approach to ESA is defined in the standard on Environmental and Social Assessment (ESS1) that describes the requirements for project risk assessment, expectations for stakeholder engagement, and for establishing grievance mechanisms. Details describing road safety requirements are provided in the standard on Community Health and Safety (ESS4). The standard on Labor and Working Conditions (ESS2) would also apply in situations where traffic management measures are necessary to address the safety of workers and local communities in and around construction worksites. The ESF standard on Stakeholder Engagement (ESS10) will also play an important role in addressing road safety issues in most projects. The participation of road users

[16] The World Bank's Guidance Note on preparing the Project Appraisal Document for investment project finances may be useful to prepare its content.

[17] World Bank, *Road Safety Indicators for Project Monitoring* (Washington, DC: World Bank, 2021).

of all types in the planning and decision making can provide essential user perspectives, information and insights on all aspects of road safety, especially if users are expected to play an active role in implementing project activities related to monitoring, incident reporting, and grievance and dispute resolution.

ESS4 anticipates that project activities, equipment, and infrastructure can increase community exposure to risks and impacts. To manage this risk, transport or transport related projects must "identify, evaluate and monitor the potential traffic and road safety risks to workers, affected communities and road users throughout the project life-cycle." The ESF requires Borrowers to "incorporate technically and financially feasible road safety measures into the project design" to minimize road safety risks and impacts.[18] Where appropriate, the Borrower will initiate a road safety assessment for each phase of the project, monitor incidents, and prepare regular reports reviewing outcomes and observations.

The ESF standard on Stakeholder Engagement (ESS10) will also play an important role in addressing road safety issues in most projects. The participation of road users of all types in the planning and decision making can provide essential user perspectives, information, and insights on all aspects of road safety, especially if users are expected to play an active role in implementing project activities related to monitoring, incident reporting, and grievance and dispute resolution. ESS10 requires the preparation of a Stakeholder Engagement Plan which systematically identifies project stakeholders and defines approaches and methods for meaningful engagement throughout the project cycle. Different stakeholders that could be affected by road safety include: all road users; project workers involved in construction; affected communities; and vulnerable groups within those communities and user groups. ESS10 also requires the preparation of project Grievance Mechanisms which could be structured as one or more channels for raising concerns about road safety, contractor performance or overall project implementation.

A Good Practice Note on Road Safety accompanies the ESF to support its implementation and to address road safety on World Bank financed operations.[19] The World Bank's Road Safety GPN guides Borrowers and World Bank task teams in meeting the ESS4 road safety requirements by implementing the **Safe System** approach. Based on the guidelines recommended by the Global Plan for the UN Decade of Action for Road Safety, the **Safe System** approach considers risks to all types of road users, including drivers, motorcyclists, passengers, pedestrians, bicyclists, and commercial and heavy vehicle drivers. The **Safe System** framework recognizes that while a certain degree of human error and crash risk is always likely, it is possible to prevent crashes that lead to death or serious injury. The Road Safety GPN recommends strategies and technical approaches to incorporate such a holistic view of road safety that considers interactions among roads and roadsides, travel speeds, vehicles, and road users. The document's guidelines on evaluating risks across the project cycle in various types of projects, and the data requirements of these procedures are discussed in the following section.

[18] World Bank, *Environmental and Social Framework for IPF Operations, ESS4: Community Health and Safety* (Washington, DC: World Bank, 2018).

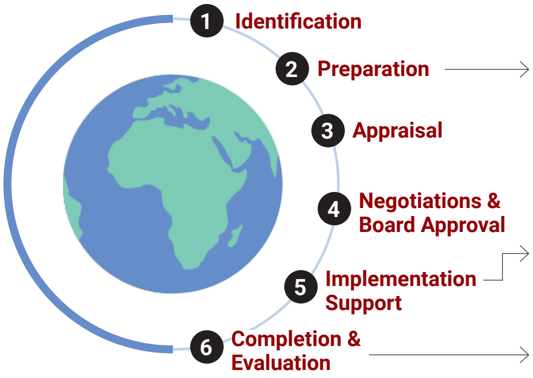
[19] World Bank, *Good Practice Note on Road Safety.*

1.2 Demand for Data to Assess Road Safety

The Road Safety GPN recommends a variety of data-driven tools and methods to evaluate road safety risks and determine mitigation measures across the project cycle. Comprehensive road safety evaluation tools and procedures require both crash and non-crash data to identify issues and measure their associated risks. The variety, quantity, and quality of data available is an important determinant of the tool for measurement and analysis of various road safety indicators.

This section provides an overview of the primary road safety assessment tools that can be used at different stages of the project cycle as well as their data requirements. Figure 3 summarizes the primary road safety activities that may need to be included in the project cycle (depending on the type of project and potential level of road safety risk). A brief description of road safety assessment procedures and tools across the project lifecycle can be found in table 1. This brief review of existing approaches informs the suggestions for improving data collection and analysis for road safety evaluation procedures through big data and machine learning (ML).

FIGURE 3: **Key road safety activities across the project cycle**



KEY ACTIVITIES
Assess project risks (tables 1 and 2)
Conduct road safety assessments and develop mitigation measures (table 3)
Prepare ESF documents
Include road safety indications in the results framework
Collect baseline data and define targets for indicators in the results framework

KEY ACTIVITIES
Prepare Implementation Status and Results (ISR) report
Regular reporting in Aide Memoire, Memos, and minutes

KEY ACTIVITIES
Prepare Implementation Completion Report: assess if targets for indicators in the results framework were achieved

SOURCE: Modified from Remote Project Supervision and Construction Management of IPF Projects. World Bank (2020).

TABLE 1: **Methods for calculating OPTRSR and identifying risk factors**

TYPE OF ASSESSMENT	WHEN TO USE (PROJECT STAGE)	WHEN TO USE (PROJECT ACTIVITY)	RELATIVE COST (HIGH, MEDIUM, LOW, DEPENDS)	DATA REQUIREMENTS (HIGH, MEDIUM, LOW, DEPENDS)	EXAMPLES OF TOOLS
Crash data-based risk assessment	Preparation, Implementation, Post-Project Operations	Pre-Planning and Design, Monitoring and Evaluation, Error Correction and Hazard Elimination	Depends, low-cost models are available	Depends	Crash frequency, crash risk factors, crash severity analysis
Road Safety Impact Assessment (RSIA)	Preparation	Pre-Planning and Design	Low	Low	
Road Safety Audit (RSA)	Preparation, Implementation	Planning and Design, Construction and Pre-Opening	Medium to High	Medium/ Depends	iRAP Road Safety Audit Toolkit, Austroads Road Safety Audit Toolkit (currently unavailable), ADB Road Safety Audit Toolkit
Road Safety Inspection (RSI)	Implementation, Post-Project Operations		High	High	iRAP
Road Assessment Program (RAP)	Preparation, Post-Project Operations	Planning and Design, Independent Assessment	High	High	iRAP, EuroRap, usRAP

SOURCE: Modified from *Remote Project Supervision and Construction Management of IPF Projects* (Washington, DC: World Bank, 2020).

Assessing Overall Project Traffic and Road Safety Risk (OPTRSR)

At the identification stage of the project, Task Teams are required to assess the Overall Project Traffic and Road Safety Risk (OPTRSR). Road safety risks arise from the interaction of many different elements, including the road and roadside design, engineering, travel speeds, the extent and type of road use, road user behavior, vehicle safety features (both active and passive), and post-crash response. The OPTRSR estimates potential traffic and road risks, and their associated risk level will inform project Preparation and help define the Borrower's responsibilities. Assessing OPTRSR also requires the identification of road safety risks that could arise as a result of project activities, for example, as a result of changing of vehicular or pedestrian traffic patterns, flows or speeds, or from the use of construction equipment or vehicles. This assessment should also identify stakeholder groups that could be affected (project workers, affected communities, or road and vulnerable road users), and institutional risks (i.e., lack of regulations, technical-knowledge, or capacity). Operational road safety risks should be addressed at this stage, not only in the context of the project implementation and construction but also the long-term project operation. The OPTRSR will identify the road safety risk level of the project as **Low, Moderate, Substantial or High.**[20]

The Road Safety GPN recognizes four different types of World Bank transport projects that require estimating the OPTRSR. Type A projects include operations which involve road construction or rehabilitation (such as urban transport projects) or any project which affects existing infrastructure or requires the creation of new transport infrastructure such as bus rapid transit lines, metro-lines, ports, railways and aviation infrastructure. Type B projects encompass other transport initiatives which do not finance transport infrastructure directly but which introduce policy changes or management measures intended to promote road safety. These may include measures such as changes to traffic speed; regulations on allowable traffic mix or volume; protections for vulnerable road users (pedestrians, bicyclists, motorcyclists); or other changes affecting vehicles, routes or facilities (e.g., vehicle import regulations). Type C projects primarily involve transport infrastructure construction with road safety impacts during the construction period only. Type D projects involve vehicle procurements, such as procurement of bus fleets or even project vehicles. OPTRSR can arise in any project as a result of the road infrastructure, operating speeds (km/h), road user behavior, vehicle standards, and/or post-crash trauma care.[21]

Different methods may be implemented for assessing the OPTRSR for each project type at the project identification stage. Based on data availability, and project type, the assessment of risk should consider all these factors: road infrastructure, operating speeds, road users, vehicle standards, and post-crash trauma care (in particular, response time and readiness of emergency care staff), three methods may be used for identifying the potential traffic and road risks and their associated level in a

[20] The principal purpose of this report is to emphasize and explain the OPTRSR risk rating and the methodologies for estimating those risks. The reader should take care to note that the OPTRSR risk rating is distinguished from the project's overall Environmental and Social risk ratings which are required for every project under the World Bank's ESF. While the overall E&S risk rating uses similar terminology, its purpose is to define the entire project risk profile taking into account all environmental and social risks and impacts. The overall project E&S risk rating takes account of the OPTRSR rating but there is not necessarily a direct correlation between them (i.e., a high OPTRSR rating may not necessarily be categorized as high E&S risk and vice versa). Each investment project will make the final determination of overall E&S risk rating and the OPTRSR rating on a case-by-case basis.

[21] According to Annex 3 in the Road Safety GPN, the Borrower and task team should ensure that the scope of the assessment is proportional to the potential risks and estimated Fatalities and Serious Injuries (FSI) for the project. This may vary for different project types. The OPTRSR process helps determine what further assessments will be relevant to the project.

project. The Road Safety GPN recommends identifying ratings and risk levels for each user group as **Low, Moderate, Substantial or High.**[22] Table 2 provides an overview of these methods.

Method I: Crash data-based risk assessments are the most reliable method for estimating the OPTRSR for Type A projects. This method effectively captures the first three criteria (infrastructure, users and speeds), and will also reflect the other two criteria (vehicle standards and post-crash trauma care). It is the go-to method when reasonable crash data from the previous three to five years is available for the road or can be estimated from data available from similar road(s) in the country and it can be used to inform the expected levels in the project. Crash data is evaluated along with an assessment of vehicle standards and post-crash trauma care to calibrate the overall risk.

Method II: When reasonable crash data is not available, and iRAP analysis of the existing road is available, iRAP results and estimated risks for other factors could be used. Dedicated to saving lives through safer roads, the International Road Assessment Programme (iRAP) provides tools and training to help countries make roads safe. iRAP Star Ratings are an objective measure of the likelihood and severity of road crashes. iRAP results are often used to deliver broad network level analysis that provide road authorities and others with risk assessment. The focus is on identifying and recording road attributes which influence the most common and severe types of crashes based on scientific evidence-based research. This approach determines the risk level of a specific road segment or network without requiring detailed crash data, which is advantageous for developing countries where data may be limited. One-star (black) roads are the least safe – a person's risk of death or serious injury is highest on these roads – while five-star (green) roads are the safest.[23]

Method III: When crash data and iRAP Star Ratings are unavailable, subjective estimates of road infrastructure risk and estimated risks for other factors should be used. In the absence of sound crash data, exposure and relative risk can be estimated especially based on WHO estimates for countries, volume by transport mode, well-established relationships between risk and operating speeds and other road design and operating features. Road infrastructure risk can also be estimated by analyzing attributes of the existing infrastructure, such as the extent of separation of pedestrians from traffic and crossing locations, extent of median separation, and presence of roadside safety barriers as well as dedicated bike, or motorcycle lanes. For both Methods II and III and for Type B, C or D projects, the OPTRSR is estimated as the weighted average of each of the identified risks.

[22] The Directive for implementing the Environmental and Social Policy for Investment Project Financing (October, 2018) Section III C defines these risks with regard to crashes as: High: "high probability of serious adverse effects to human health…"; Substantial: "there is medium to low probability of serious adverse effects to human health … and there are known and reliable mechanisms available to prevent or minimize such incidents"; Moderate: "low probability of serious adverse effects to human health"; and, Low: "if its potential adverse risks to and impacts on human populations … are likely to be minimal or negligible."

[23] iRAP (International Road Assessment Programme), *iRAP Star Rating and Investment Plan Implementation Support Guide* (London: iRAP, March 2017).

TABLE 2: **Methods for assessing OPTRSR and identifying risk factors**

METHOD	RISK FACTORS	DATA REQUIREMENTS
Crash data-based risk assessment	• FSI crashes	• Crash data from the previous 3–5 years or estimated from data available from similar roads in the country • Assessment of vehicle standards (safe vehicles) • Post-crash trauma care (response time, quality of attention)
iRAP Star Rating	• iRAP Star Rating – existing conditions for vehicle occupants • iRAP Star Rating – existing conditions for motorcyclists (if motorcycles are present on the road or likely to be present post-project) • iRAP Star Rating – existing conditions for bicyclists (if bicycles are present on the road or likely to be present post-project) • iRAP Star Rating – existing conditions for pedestrians (if pedestrians are present on the road or roadside or likely to be present post-project) • Assessment of non-infrastructure risks: operating speeds, road users, vehicle standards, and post-crash trauma care	• iRAP scores (Low, Medium, Substantial, High) • Estimates for non-infrastructure risks
Estimating road infrastructure risk without crash or iRAP data	• Extent of separation of pedestrians from traffic with provision of safe walking spaces and crossing locations (if pedestrians are present on the road or roadside or likely to be present post-project) • Extent of roadside safety barriers (omit this factor from consideration if the operating speed is <40 km/h) • Extent of median separation (omit this factor from consideration if the operating speed is <60 km/h for a rural road and <40 km/h for an urban road) • Extent of separate well-designed motorcycle lanes (if motorcycles are present on the road or roadside or likely to be present post-project) • Extent of separate off-road bicycle lane (if bicycles are present on the road or roadside or likely to be present post-project) • Assessment of non-infrastructure risks: operating speeds, road users, vehicle standards, and post-crash trauma care	• Subjective estimates of road infrastructure risk for each of the risk criteria • Estimates for non-infrastructure risks

SOURCE: Road Safety GPN.

The Road Safety Screening and Appraisal Tool (RSSAT) developed by the Transport Global Practice, is required for all World Bank transport projects (Type A) and also recommended for other projects that may involve road safety risks. Its results must be reported in conjunction with the OPTRSR. The RSSAT tool (Method IV) considers the likely fatality rate with and without the project and it is designed to undertake a quick road safety screening of World Bank projects during the concept and preparation stages. It evaluates the safety effects of different design options, and conducts a cost-benefit analysis of the project's impact on road safety, estimating change in potential Fatalities and Serious Injuries (FSI) due to the project. At the identification stage, RSSAT should be applied, and the results reported in conjunction with the OPTRSR.[24] RSSAT does not require crash data to identify likely change in FSI risk, and it is now required for all World Bank financed transport projects to estimate the economic cost of road crashes on project roads.Type A projects should demonstrate Project Safety Impact of 1 or below for all road segments before approval. Table 3 summarizes the data requirements for RSSAT.

[24] World Bank, *Good Practice Note on Road Safety.*

TABLE 3: **The World Bank Road Safety Screening and Assessment Tool**

METHOD	PROJECT SAFETY COST/BENEFIT IMPACT	DATA REQUIREMENTS
RSSAT	• Project safety impact analysis and safety impact model	Baseline and projected estimates for: • Fatalities by mode • Speeds by fleet type • Segment characteristics and road features • Traffic flows

SOURCE: Road Safety GPN.

Traffic and Road Safety Assessments

During the project Preparation stage, the Borrower may need to conduct more in-depth assessments to identify and evaluate potential traffic and road safety risks. When traffic and road safety issues are likely to be significant for the community or road users, the objective of the road safety assessment is to consider these risks in more detail to determine the most appropriate mitigation (control) measures that can be implemented in the project. The assessment should consider the **Safe System** principles to confirm that all opportunities to minimize risks have been realized. The **Safe System** approach addresses all of these interactive elements in an integrated manner and emphasizes sharing accountability with designers and users of the road network to achieve road safety targets.[25]

Assessments prepared early in the project cycle help to identify and evaluate potential traffic and road safety risks that may arise from the project activities and/or their implementation. Such assessments are intended to help the Borrower mobilize appropriate resources, analyze risks in detail, and identify and adopt the most appropriate mitigation measures. This assessment also guides the preparation of the environmental and social documents, such as the Environmental and Social Impact Assessment (ESIA), Environmental and Social Management Plans (ESMP), and the Environment and Social Commitment Plan (ESCP).

For projects with High or Substantial road safety risks, assessments should be completed before the project is fully appraised to inform project objectives, components and activities, and the results framework.[26] Type A projects, or Type B and C projects with major construction activities require more robust or detailed assessments. Substantial and High-risk projects should, as a minimum, include intermediate indicators related to traffic and road safety risk mitigation. Table 4 summarizes the different types of assessment tools (Methods V-VII) that can be used for this purpose as well as their data requirements.

One or more of these assessments may be conducted at once or at different phases of project Preparation. Road Safety Audits (RSA) and Road Safety Impact Assessments (RSIA) involve examining a traffic project, which may involve new construction or altering an existing road, to improve traffic and road safety performance. An RSA is a formal procedure to assess the crash risk potential and expected safety performance of a design for a road or traffic scheme. RSIA is a strategic assessment of the impact of different planning options. Safe System Assessments (SSA) evaluate the design against **Safe System** principles to confirm that all opportunities to mitigate risks and maximize road safety have been realized.

[25] Tony Bliss and Jeanne Breen, "Meeting the Management Challenges of the Decade of Action for Road Safety," *IATSS Res.*, 35 (2012): 48–55, https://doi.org/10.1016/j.iatssr.2011.12.001

[26] For projects with Moderate or Low safety risks, the Borrower and the Bank may agree on more flexible timelines for the completion or road safety assessments and/or mitigation or management measures. Such agreements would be specified In the project's ESCP.

These assessment procedures enable road safety engineering and crash analysis to be used for the prevention of crashes on new or modified roads. They can be conducted at different stages of the project cycle to identify key road safety challenges to guide designers, confirm that safety elements are correctly captured, check for any unsafe feature not apparent at previous stages and check that all the design details have been correctly implemented, identify deficiencies that need to be corrected, or to evaluate the road's performance with traffic and determine areas that require further attention. The earlier road safety risks are assessed within the design and development process the better to ensure that safety is fully integrated into all elements of the project's infrastructure, with minimal risk of redesign or physical rework at a later stage.

The main data needed to perform these types of assessments are FSI, traffic flows, and road features. Data analyses, modelling or estimates quantify and forecast traffic volumes and road crash FSI. Depending on data availability, these would aim to identify crash locations and crash types, at-risk individuals and groups, and key risk factors influencing exposure to risk, crash involvement, crash severity and post-crash outcomes. Even in the absence of sound crash data, exposure and relative risk can be estimated based on estimates for countries, volume by transport mode, well established relationships between risk and operating speeds, and other road design and operating features. Capacity reviews to assess the efficiency and effectiveness of road safety measures can be relevant when the project involves road safety policy change.

TABLE 4: **Overview of primary tools for traffic and road safety assessments**

METHOD	OBJECTIVES	DATA REQUIREMENTS
Road Safety Audits (RSA) (performed by an independent team of specialists)	Identify safety concerns. It audits the safety of the specific design of the chosen scheme.	Analysis of project designs and interventions: specialists assess road options, such as intersections, signs, crossings; design standards, and the relationship of this intervention to main network. Main data needed includes: • Scheme plans • Crash and FSI data • Traffic mix and volumes • Road features (e.g., design elements, such as bypasses, cycle routes, junction improvements, installation of traffic signals, roundabouts, traffic calming, bend realignment, safety fence schemes and pedestrian crossing facilities)
Road Safety Impact Assessments (RSIA) (performed by members of the project design team with road design and road safety auditing experience)	Assess the impact of each of the planning options on the safety performance of the current road network. It estimates the impact of possible schemes on safety for an entire geographic area at the strategic level.	The evaluation of each alternative is based on several factors, some of which includes: • The scheme objectives • Crash and FSI data • Traffic mix and volumes • Road features • Categorization of roads and streets of that network
Safe System Assessment (SSA)	Assess how closely road design and operation align with the Safe System objectives, and to clarify which elements need to be modified to achieve closer alignment with these objectives.	The core of the SSA approach is the "Safe System Matrix" framework, which is essentially a risk assessment. The assessment is done by scoring the risk exposure, likelihood and severity from 0–4. The Austroads approach can be used to perform this type of assessment. Data needed includes: • Traffic mix and volumes • Road features

SOURCE: Road Safety GPN.

Since the key objectives of these assessments (i.e., identifying risk elements and estimating crash exposure, likelihood, and severity for different road users) are complex and not standardized, the scoring system is subjective. This can complicate comparisons between sites, especially when these have been assessed by different individuals or teams. It is, therefore, usually most suitable for comparing options at a single site, identifying sources of risk and identifying solutions, rather than for comparing different sites.

Results Frameworks and Monitoring Plans

In addition to these assessments, a Results Framework that articulates the expected outcomes and impact of the project on road safety should also be developed before project Appraisal. A Results Framework is a management tool that presents how the development objective(s) of an operation will be evaluated, measured and monitored, based on the results chain (outputs, outcomes, and impacts). The Results Framework is based on the Project Development Objective (PDO) that indicates expected project outcomes. Depending on project design, intermediate indicators for each project component can be used to track implementation progress including the units of measurement, baselines, and final target for each indicator. Such details are typically provided in the project's Monitoring Plan.

The Results Framework and Monitoring Plan should include a road safety indicator with baseline and target values. The Transport Global Practice has committed to including a road safety indicator in all road projects and to increase the road safety focus of urban mobility projects. All substantial and high-risk projects should include at least one indicator that addresses road safety in the Results Framework or as a Disbursement Linked Indicator, as relevant.[27]

TABLE 5: **Example of indicators that may be included in the Results Framework**

EXAMPLE OF INDICATOR	TYPE OF DATA THAT CAN BE COLLECTED
Reduction of road crashes	Crash data
Speed reductions	Traffic flows
Increased use of helmet and seat belts	Number of helmet and seat belt users

SOURCE: Original table for this publication.

There are two types of indicators that should be considered. The first kind are **intermediate indicators,** which mark the progress toward fulfilling the development objectives before the final project outcomes are achieved (these may also measure progress in project outputs). Some examples of intermediate indicators that may be relevant to transport projects include the number of speed managing devices installed and safety audit compliance. The second are **outcome indicators**, which evaluate the uptake, adoption, and use of outputs by the target group within the project period. FSI is considered the most important indicator for monitoring the outcome of road safety interventions.[28] Table 5 provides some examples of indicators that can be included in the Results Framework, as well as the type of data that can be collected to monitor and evaluate them.

Change in FSI is the most frequently tracked metric for impact evaluation of projects and interventions for monitoring the outcome of road safety interventions. In cases where data cannot be obtained, other methodologies to estimate safety risks can also be used. Projects need to undertake baseline data collection to not only establish the appropriate project interventions to address road safety risks, but also as a way of assessing whether the project will improve or worsen the situation. Target values are to measure progress towards a particular indicator. For example, the number of workers killed (zero baseline because the project has not started; and zero target because the objective should always be to avoid fatalities). This indicator should be based on one or more of the World Bank's seven road safety pillars.

During the Implementation phase, the focus shifts toward executing planned activities, and monitoring and evaluating indicators. Activities that are included in the Project Appraisal Document (PAD) are to be carried out during this phase. When key information or data for indicators included in the results matrix must be collected, it is important that procurement processes and supervision activities are planned and executed in a timely fashion to achieve expected results. It is also vital that the project design includes close monitoring of the safety performance until the project closes. In some cases, impact evaluations may also be required to monitor the long-term effects of implemented interventions. For example, the results matrix would identify the extent of progress towards achieving a particular milestone, like enumerating the number of physical features to separate traffic (e.g., footpaths, cycle lines, traffic signals) installed in the project to address the safety of vulnerable group users, such as pedestrians, bicyclists, or motorcyclists.

The Implementation Completion Report (ICR) addresses the targets achieved at the completion of the project. At project completion, the ICR carries out an ex-post analysis of project interventions, and measures outcome and intermediate indicators from the results framework to assess whether targets were achieved during implementation. The ICR will collect the indicators for the results framework for the last time to evaluate whether PDO and intermediate indicators meet their targets.

[27] World Bank, *Road Safety Indicators for Project Monitoring* (Washington, DC: World Bank, 2021).

[28] World Bank, *Road Safety Indicators for Project Monitoring.*

If, for example, a PDO is to contribute to the reduction of road traffic injuries and fatalities in selected corridors, with intermediate indicators to track progress towards some interventions, like implementing a certain number of physical features to separate traffic, the ICR will quantify this at the end of the project cycle so it can be compared with baseline indicators and expected targets.

Key Challenges with Current Approaches to Road Safety Analysis

Since data is the cornerstone of all road safety assessments, the availability of high quality, reliable data is key to extracting useful, actionable insights and improving road safety conditions. Without quality information, it is difficult to estimate crash locations and crash types, at-risk individuals and groups, and key risk factors influencing exposure to risk, crash involvement, crash severity, and post-crash outcomes. Meeting data requirements for road safety assessments can be a challenge for various reasons, such as the lack of open data, or data collection costs.

There can be a lack of adequate crash data or road ratings in data scarce countries and regions for identifying risk factors (Methods I to III). Governments often lack adequate and reliable data to identify road safety risks and perform road safety assessments. In addition, road crashes tend to be underreported, especially in LMICs. There may also be significant gaps in the data in terms of geographic or temporal coverage, or the data may be missing important variables and categories. Access to data can also be limited for certain data types, or the process of obtaining the data may be too complex, costly, and time-consuming.

Collecting data on road safety attributes through manual detection or special equipment can be expensive, time-consuming, and complex.[29] Budgeting for data collection can be a challenge for both Borrowers and World Bank task teams, especially for Methods I to IV which are required at the project identification stage. In these cases, data is most often estimated through existing road designs or by local transportation agencies. For Methods V to VII, the most cost-effective method for data collection is the installation of cameras and sensors that record street imagery, speed information and other data. Images and video are then analyzed by road safety experts to identify relevant attributes, assess road conditions and identify potential risks. Commissioning equipment and hiring resources to manually collect data on road features and design may be a hindrance, especially for smaller-scale projects where the opportunity to benefit from economies of scale is low.

In addition to the quality and availability of data, preparing and analyzing road safety data can also be costly, resource-intensive, and technically demanding. Most road safety assessments require data to be combined from various sources, which often involves aggregating, cleaning and preparing the data. Additional resources and specialist expertise may be necessary for this process, and also to analyze the data and extract useful insights using methods such as clustering and developing spatial models. Conventional statistical techniques can also be limited in their ability to identify complex correlations and underlying factors that may contribute to road safety risks across various projects.

The purpose of this Guidance Note is to identify new methods for the collection and analysis of road safety data that could overcome the limitations of existing approaches, and also improve their efficacy in identifying risks and opportunities to mitigate crashes. Conducting road safety assessments is a required component of most road investment and infrastructure development projects. Advanced technologies such as big data and ML have the potential to not only supplement existing methods, but also significantly reduce costs while improving the efficacy of road safety assessments in identifying risks and opportunities to mitigate crashes.

The following section explains how big data and ML can be practically implemented by Borrowers and World Bank task teams for various road safety assessment procedures that are required by World Bank investment projects at various stages of the project cycle. It introduces these methods and provides an overview of big data sources and ML techniques that are useful for road safety assessments (tables 6 to 10). Part 2 also discusses best practices and key considerations that are vital to implementing these new methods effectively. A framework for integrating these technologies in road safety assessments is also proposed, and subsequent sections demonstrate how this framework can be applied in LMICs through two original case studies.

[29] OECD (Organisation for Economic Co-operation and Development)/ITF (International Transport Forum), *Big Data and Transport: Understanding and Assessing Options* (Paris: OECD/ITF, 2015), https://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf

PART 2:

Big Data and Machine Learning to Strengthen Road Safety in Transport Projects

The World Bank and Global Road Safety Facility are keen to use new technologies, such as big data and ML, in data collection and analysis for road safety to overcome the limitations of existing approaches. As these technologies become more sophisticated and accessible, a growing body of research indicates their potential to complement, and eventually even surpass conventional methods.

World Bank teams have demonstrated various applications of big data and ML in road safety and other transport and infrastructure projects over the past few years. For example, a task team developed an open data platform in 2015 based on a pilot in Cebu City, Philippines, which sourced data from a taxi company to generate insights for traffic management.[30] Another team has developed a "Simplified Methodology" to implement ML in video analysis to extract data on road attributes. The new tool was piloted across over 500 kilometers of road in Mozambique and Liberia in 2019.[31] The World Bank, in collaboration with the Philippine government, has also launched the Data for Road Incident Visualization Evaluation and Reporting (DRIVER) system to facilitate data sharing for road safety analysis. This free web-based, open-source platform connects traffic crash data from multiple agencies through a standardized reporting system. DRIVER also provides tools to geo-spatially analyze road crash data, predict blackspots, estimate the economic costs of crashes, and evaluate the effectiveness of various interventions to support investments and policy-making for improved road safety.[32]

World Bank teams are increasingly turning to data partnerships to obtain crash, traffic, and other types of data for road safety analysis. For example, in Kenya, the WHO estimates that up to 75 percent of crashes go unreported.[33] SmarTTrans – a collaboration between the Kenyan government and the World Bank – has worked to fill this gap by bringing together crash information both from administrative records and from bystander crash reports from Twitter.[34] In addition, the team has leveraged the Development Data Partnership (DDP) to access Waze API and Uber congestion and speed information for all 6,200 km of the city's road network. Using all data sources, the smarTTrans team is creating near real-time analytics to facilitate the identification of crash hotspots, speeding, and congestion patterns.

[30] World Bank, *Open Traffic: Easing Urban Congestion* (Washington, DC: World Bank, n.d.), https://olc.worldbank.org/system/files/WBG_BD_CS_OpenTraffic_1.pdf

[31] World Bank, *Innovative Road Safety Risk Assessment Tool with Automated Image Analysis Technology* (Washington, DC: World Bank, 2019).

[32] World Bank, *GRSF DRIVER Completion Report* (Washington, DC: World Bank, 2019), https://documents1.worldbank.org/curated/en/245151560919065747/pdf/Data-for-Road-Incident-Visualization-Evaluation-and-Reporting-Lowing-the-Barriers-to-Evidence-Based-Road-Safety-Management-in-Resource-Constrained-Countries.pdf

[33] WHO, *Global Status Report on Road Safety 2018.*

[34] Sveta Milusheva et al., "Applying Machine Learning and Geolocation Techniques to Social Media Data (Twitter) to Develop a Resource for Urban Planning," *PLoS ONE* 16, 2 (2021),
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0244317

2.1 New Data (and Big Data) in Road Safety Analysis

Big data is generally understood as extremely large datasets that are generated by a wide range of data sources, including machines, sensors and other Internet of Things (IoT) devices. Big data can also be captured over the internet through social media and other types of applications, especially those that track locational or transactional data.

The large volume of such data is one of many characteristics that make big data especially useful for road safety and other applications in transport and infrastructure development. For example, big data can be generated at immense **velocity**, especially as more such data is collected real-time and for large populations. It also occurs in a **variety** of data formats, from structured databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions. Big data is also characterized by a high degree of **variability** since data flows can change over time, depending on seasons, off-peak hours or availability of collection methods across an entire population under study. Table 6 provides a SWOT analysis of the use of big data in road safety analysis.

For transport, the increasing use of personal mobile devices and vehicle sensors to collect traffic and location data presents a significant opportunity to augment traditional sources of transport data. Annex 1 discusses the most relevant big data types for road safety analysis. It also provides guidance on the potential applications of these sources for evaluating road safety, and the advantages and disadvantages of each source. The following sections discuss how big data can be used for the various road safety assessment methods and tools discussed in Part 1.

TABLE 6: **SWOT analysis of using big data in road safety analysis**

STRENGTHS	WEAKNESSES
• Recent and broad geographic coverage allows researchers to dive deeper into transport issues and get a comprehensive and current picture of risks. • Can help obtain real-time data and track up-to-the-minute changes in traffic flows and other important variables. • May be faster and easier to obtain and process, compared to manual collection. • Can offer higher spatial and temporal resolution than conventional sources. • Can be more affordable and easier to scale. • Vast quantities of data can limit bias from outliers and other sources of "noise" since data gets aggregated across vast populations. • Can help improve data quality since often covers large geographic and/or temporal scope, also allowing for comparison against "control" datasets and scenarios.	• Requires investment in expertise, software and computing power to store, access and process big data. • Availability of data can vary significantly by geography and context. • Coverage can be inconsistent or exclude important segments of the population. • Most big data sources are not set up to support road safety assessments—it is often data that was collected for other purposes but gets repurposed for road safety analysis. This can lead to the data being biased, incomplete and/or difficult to incorporate in road safety analysis. • Need to consider the interoperability of different datasets (i.e., how easy it is to combine different datasets for complex road safety assessment models). • Changes in privacy laws and other relevant policies can impact quality, consistency and coverage of data.
OPPORTUNITIES	THREATS
• Provides an alternative approach to road safety data collection and analysis that may complement or supplement traditional approaches or datasets. For example, big data sources may be able to collect more accurate crash data. • Big data analysis can uncover new dynamics, complex behavioral patterns and relationships, and correlations that conventional statistical methods and data may not be able to detect. • Growing interest in autonomous vehicles is generating more data about road systems, vehicles, and vulnerable users that can be integrated into road safety analysis. • Rising momentum for the creation of a "big data platform" where data providers can sell or share data.	• Privacy concerns – data should be de-identified and anonymized before use. • Data providers may be reluctant to share data. • Governments, local municipalities, and other stakeholders must invest in technological infrastructure to support big data collection and analysis. • Need to enforce quality control to limit risk of data bias. • Licensing constraints – most private companies, such as Google, provide limited licenses for data use.

SOURCE: Original table for this publication.

Big data, especially when combined with ML, which is discussed in the following section, can enhance the capabilities of current systems and road safety assessment tools. The increasing use of IoT devices, which range from smartphones to vehicle sensors, as well as Intelligent Transport Systems (ITS), is making it possible to collect, access and utilize real-time data about a large range of variables that are relevant to road safety analysis. This includes traffic flows, crash sites, peak timings, travel times and road usage by pedestrians, bicyclists, and motorists. The availability of such extensive data creates new possibilities for crash risk modelling, especially to predict the outcomes of various types of road safety interventions as well as possible impacts of road infrastructure projects.

As mobile phone use rises globally, smartphones have become a prominent source of big data, though there are many other sources to consider. In addition to the location and velocity of road travelers collected passively through mobile devices, transportation projects can take advantage of street view, aerial, and satellite imagery, traffic monitoring systems, connected vehicles for road safety analysis, as well as crowdsourced data provided by the community through mobile devices.[35] Annex 2 provides an overview of the most relevant and accessible big data sources for World Bank task teams and is a useful starting point to find relevant data sources. TTLs are advised to look for relevant

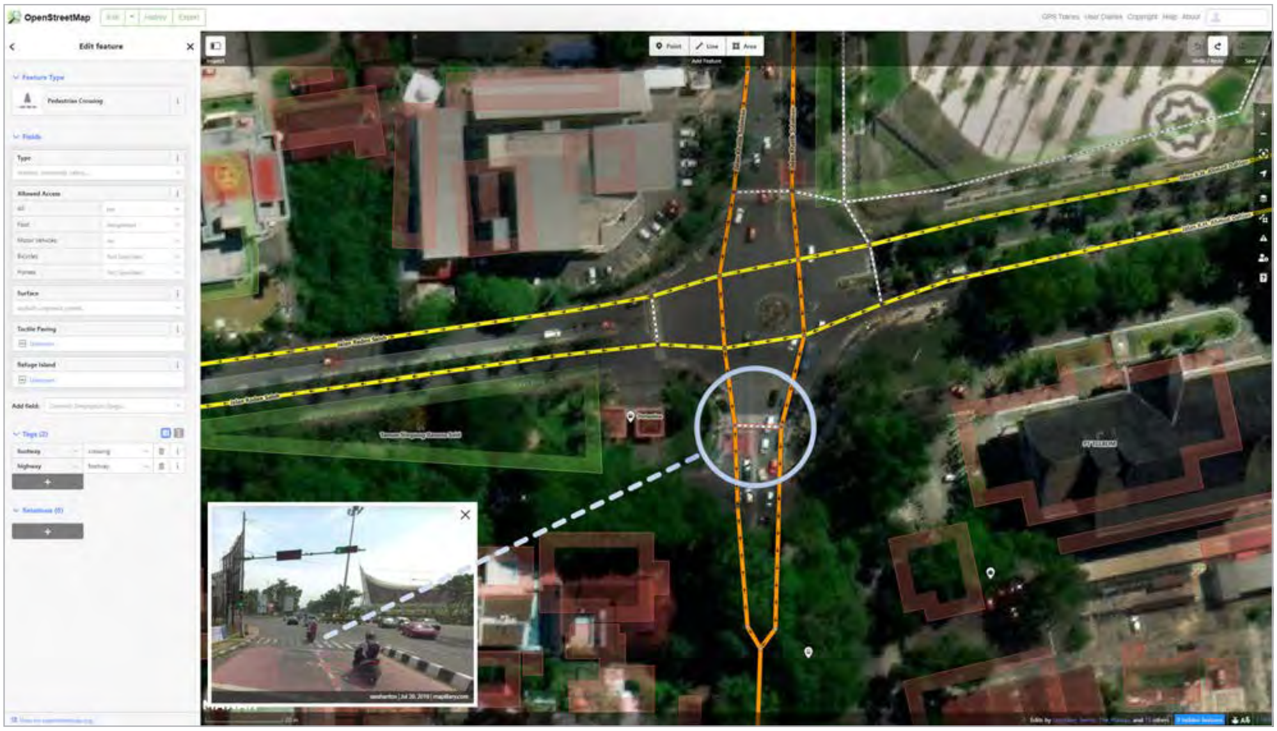
[35] Alex Neilson et al., "Systematic Review of the Literature on Big Data in the Transportation Domain: Concepts and Applications," *Big Data Res.* 17 (2019): 35-44. https://doi.org/10.1016/j.bdr.2019.03.001

local and regional data providers based on the region(s) of interest that concern their project(s). As big data infrastructure advances globally and new companies and startups begin data collection for various purposes, it is likely that the list of available big data sources in World Bank member countries will expand significantly in coming years.

Street view imagery can complement or potentially substitute manual or commissioned road surveys to collect data on road safety attributes for various types of assessments. For example, street-view imagery can help obtain baseline data for RSIA more quickly and cheaply, especially if the data is not already readily available. By applying ML algorithms to street view images, road attributes and other data can be detected that are important for road safety assessments. Similarly, there may be instances where satellite imagery or aerial imagery, those collected by an unmanned aerial vehicle (UAV) or drone, can be analyzed to detect road or road user attributes. Figure 4 shows the same crosswalk visible in satellite imagery and street view imagery using OpenStreetMap in OSM. ML is discussed in greater detail in the next section.

FIGURE 4: **Street view and OSM**

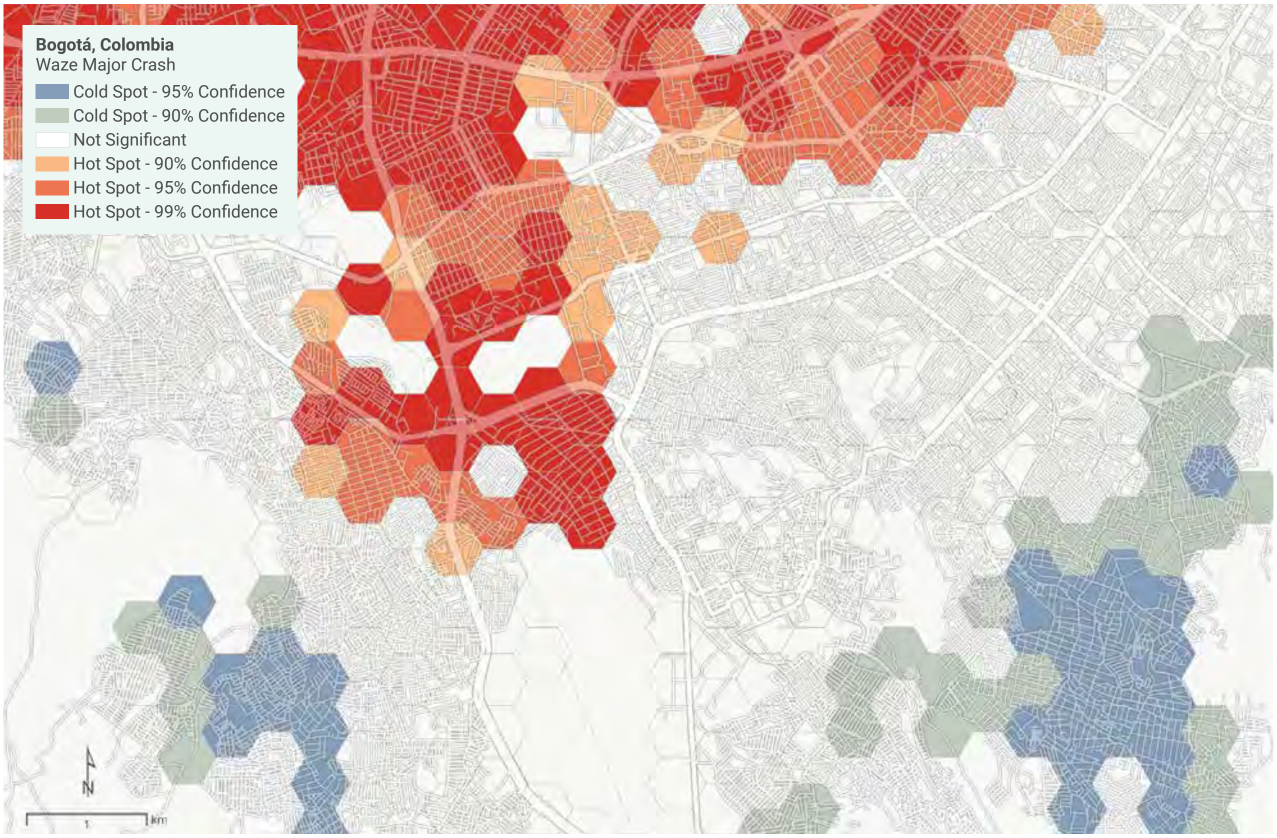
Road safety data can be extracted from images such as road markings and signs, types of road users, and designated paths for vulnerable users. Each image and relevant attributes are geolocated for further analysis. In this instance, the crosswalk identified in OSM can be verified in street view imagery.



SOURCE: Original figure for this publication derived from OSM, Mapillary, and Maxar Technologies.

Mobile applications and telematics can provide data related to vehicle movement to identify road infrastructure risks. This data includes current and historical average speeds along road segments as well as irregularities, like traffic jams and incidents. This data is useful for most proactive road safety assessment tools, including RSIA, RSA, and RSI. It can be geographically visualized and analyzed, such as through heatmaps or hotspot analysis as shown in figure 5 (see Annex 3 for additional examples and descriptions). Telematics data has also been used to assess driver behavior, facilitate the prediction of crash-prone locations and create geographic visualizations, as discussed in interviews with researchers at the ARRB and Professor George Yannis from the National Technical University of Athens. However, data privacy is an especially important concern when it comes to the use of telematics data.[36]

FIGURE 5: **Hotspot analysis of major crashes reported by Waze application users**


Bogotá, Colombia
Waze Major Crash
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
N
1
km

SOURCE: Original figure for this publication (data provided by Waze App; learn more at waze.com). Basemap provided by Esri, HERE, Garmin, METI/NASA, USGS.

Mobile applications are helping overcome underreporting of road crashes by crowdsourcing incident reports. For example, in Kenya, road crashes have been shown to be largely underreported, especially in areas where incident reporting mechanisms are lacking or underdeveloped.[37] Navigation applications such as Waze are providing a valuable new source of crash and traffic data by allowing users to report incidents through their smartphone applications. Each incident report submitted by a user is geolocated and timestamped, which allows it to be combined with other geospatial data to identify segments of a road that are experiencing major or minor crashes, light to stand still traffic jams or hazardous conditions (hazards on the road or on the shoulder, weather alerts or dangerous road surfaces). Additionally, social media platforms like Twitter are used by many people on the ground to report on crashes and traffic conditions and can be leveraged using machine learning algorithms to produce additional data on crashes, as was done by the smarTTrans team in Nairobi.[38] Lastly, mobile application data can be generated in real-time to assist with monitoring or collected and analyzed over time to develop models.

[36] Anthony Germanchev (Principal Professional Leader, Advanced Technologies Lab, Australian Road Research Board) and Professor George Yannis (School of Civil Engineering, National Technical University of Athens), in discussion with the authors, April 2021.

[37] Guadalupe Bedoya Arguelles, et al., “Smart and Safe Kenya Transport (SMARTTRANS).”

[38] Sveta Milusheva et al., “Applying Machine Learning and Geolocation Techniques to Social Media Data (Twitter) to Develop a Resource for Urban Planning.”

A growing number of countries and regions are focusing on developing a big data infrastructure to collect official incident reports. Collecting comprehensive and accurate information about road incidents is an important objective for government transportation agencies. There is growing interest in gathering and analyzing the information in big data formats to provide deeper and more comprehensive insight into road safety risks and the impact of different interventions. The collection of real-time data would also be beneficial for this purpose, for which collecting, storing, and analyzing the information as big data would be most realistic and feasible.

How to Access Big Data

Big data for road safety generally falls into two categories: public sector and private sector. Traditionally governments have collected and provided data for road safety analysis, such as police reports of crash incidents. However, alternative sources are becoming increasingly available as mobile apps are used to crowdsource reports of roadside incidents and companies aggregate traffic speeds from proprietary mobile applications. Often data quality from such sources can vary significantly by location, with certain sources being more effective, reliable, and better developed in some regions compared to others. Task teams are advised to use the list provided in Annex 2 as a starting point and find the most relevant data providers for their project's region(s) of interest.

This Guidance Note focuses on big data sources that are most easily and readily accessible to World Bank task teams. Different sources require different approaches to obtaining relevant data quickly and efficiently. It is important to understand the licensing restrictions that accompany each source. For example, even though a dataset is crowdsourced, it may have licensing restrictions. It is best to consult the World Bank Legal team and data provider to clarify terms of use when necessary.

Public sector. Governments can collect, manage, and share data relating to transport, infrastructure, and mobility. Many governments, whether at the national level or even local municipalities, are establishing open data platforms where datasets can be accessed by running a simple search query. Such platforms have already been created in the Philippines as well as in Australia and the United States.[39] In other instances, particularly where the data infrastructure is not as advanced, data may have to be requested through the relevant department. It is often possible to obtain datasets relating to crash histories or collected by road sensors from government sources which are extensive enough to be processed as big data in road safety analysis.

The World Bank's Road Safety Observatories (RSO) initiative also has the potential to become an important source of government-generated big data in the future. The Observatories provide a formal network of government representatives to share and exchange road safety data and experience in order to improve road safety throughout the region. The World Bank established its first RSO in Latin America (OISEVI), before introducing the initiative in Africa (ARSO) and Asia-Pacific (APRSO). By enhancing road safety data and information systems, the Observatories play a pivotal role in helping countries monitor, evaluate and develop more impactful road safety policies and interventions.[40]

In other cases, publicly available datasets with a global reach may be considered. A good example

[39] Australian BITRE (Bureau of Infrastructure and Transport Research Economics), "Australian Road Deaths Database (ARDD)," Australian BITRE, updated May 13, 2021, https://data.gov.au/data/dataset/australian-road-deaths-database; ODPH (Open Data Philippines), "Open Data Philippines," ODPH, accessed June 3, 2021, https://data.gov.ph/; US NHTSA (United States National Highway Traffic Safety Administration), "Data," US NHTSA, accessed May 28, 2021, https://www.nhtsa.gov/data

[40] World Bank, "Better Data for Safer Roads: The Powerful Mission of Road Safety Observatories," last modified November 5, 2020, https://www.worldbank.org/en/news/video/2020/11/05/better-data-for-safer-roads-the-powerful-mission-of-road-safety-observatories

of this is OSM, which offers freely available geographic data generated by volunteers who trace satellite images around the world to create and update the map consisting of road networks (detailing road types, bridges, tunnels, direction of traffic flow), among other features. OSM data can be combined with other datasets for road safety analysis. While OSM provides an overview of the road geometry, the recency and accuracy of the data requires validation. Due to variability in quality and coverage, OSM data would be considered a starting point and is not recommended for detailed assessments.

Private sector. Mobility datasets are generated through ride-hailing services, delivery services, social media, and other mobile applications that collect user location and movement. Companies in the transportation and logistics sector use smartphone applications to digitize their operations and take advantage of higher quality, real-time data to improve efficiency as well. Other companies provide telematics software to track vehicle movement and safety features. Companies and start-ups investing in autonomous vehicle research are providing valuable sources of big data for road safety analysis. Some companies also provide APIs that allow developers to access these datasets (often on a limited basis). However, proprietary or commercial data may have to be purchased in some instances, or data partnerships need to be established to access such data. It is also crucial to understand how the data is licensed and can be legally used for different types of analysis. For example, Google restricts digitizing and tracing information as well as using applications to analyze and extract information from street view images, although annotation and labelling is permitted.[41]

Data Partnership Agreements. World Bank task teams can apply for access to various datasets for road safety analysis through the Development Data Partnership (DDP), which is a formal collaboration of private sector companies and international organizations to use third-party data in research and international development.[42] It is accessible to all World Bank staff and partners. Upon submitting a proposal through the DDP site and signing a licensing agreement, companies provide datasets relevant to road safety, such as human movement (Orbital Insight, Unacast, and Veraset), traffic speed (Mapbox and Waze), social media (Twitter), and weather (tomorrow.io). In addition, the site shares guidance on accessing the datasets and contains a searchable inventory of Development Partner projects. DDP provides a seamless, efficient, and secure manner for World Bank teams to access data from a broad range of data providers across various regions of interest. It includes templates of data license agreements, access to multi-disciplinary teams for end-to-end support and a centralized IT architecture and processes for ingesting, storing, and pre-processing data, as well as for coding collaboration. Task teams can also benefit from extensive, up-to-date documentation that provides guidelines, code snippets and examples from data partners' products and services to facilitate their project.[43] DDP datasets are primarily intended for experimental purposes. If proven successful, governments may consider implementing a five-year agreement directly with the company to continue to use the data for road safety analysis. It is also possible to benefit from the platform by becoming a World Bank Data Fellow.

Waze for Cities is one example of a data sharing agreement that can be leveraged using the DDP platform. The program allows cities to utilize data standards designed by Waze for closure and incident reporting to reduce data fragmentation and promote transport and government data aggregation. It now has more than 500 global partners including city, state and country government agencies, nonprofits and first responders. Moovit, an app focused on public transport, offers Mobility as a

[41] Google, "Google Maps, Google Earth, and Street View," accessed May 14, 2021, https://about.google/brand-resource-center/products-and-services/geo-guidelines/

[42] Development Data Partnership, https://datapartnership.org/

[43] Development Data Partnership Documentation, https://docs.datapartnership.org/pages/documentation.html

Service (MaaS) solutions for cities, providing personalized apps, payment solutions, real-time transit information, and other analytics.

In many cases, data providers help local governments by exchanging data. For example, the city of Tokyo in Japan has partnered with a private firm to develop a smartphone compatible app, Zenryoku Annai!. The app analyzes nearly 360 million observations every second to generate real-time information on the shortest and least-congested travel routes. A similar intelligent transport system (ITS) in Denmark, Copenhagen Connecting, was implemented to promote transport sustainability through real-time digital traffic control and weather adaptation options. World Bank task teams should consider seeking the support of local governments to establish data partnership agreements, particularly if the provider is not already a part of the DDP.

Data marketplaces. Business leaders are keen to explore the value of the big data they collect as a tradable commodity. This has given rise to data marketplaces which are essentially online platforms dedicated to the buying and selling of data. These marketplaces can provide a more cost-effective source of data compared to other data mining techniques. Dedicated marketplaces for traffic and transport data have also emerged in recent years, although their coverage of LMICs tends to be low.

As part of its efforts to establish an artificial intelligence tool for road safety analysis (called Ai-RAP), iRAP is seeking to establish a data marketplace where public and private data providers can trade data for road safety analysis. The data marketplace will focus on three types of data products, according to Monica Olyslagers (Safe Cities and Innovation Specialist at iRAP), who was interviewed for this Guidance Note.[44] The first is raw datasets that need to be processed to extract relevant information. The second is datasets that have been at least partially cleaned up and processed by data

[44] Monica Olyslagers (Safe Cities and Innovation Specialist, iRAP), in discussion with the authors, April 2021.

providers or Ai-RAP and are ready to be plugged into road safety assessments. The third is prepared-for-purpose datasets that are specifically commissioned for road safety assessments in different types of projects. This data marketplace model is currently being piloted in Africa, as part of a project to set up a regional road safety observatory there in collaboration with the World Bank.

The new data marketplace will initially focus on aggregating and trading conventional datasets. However, the project team plans to bring on big data providers and incorporate ML in the Ai-RAP tool to allow for more sophisticated analysis in road safety assessment procedures. Borrowers and TTLs are advised to search data marketplaces as a lesser-cost alternative to commissioning data collection for their projects.

Key Considerations for Selecting the "Right" Big Data Source

This section provides an overview on how different big data sources can be used. The data sources covered in the tables for each method or assessment type should be viewed as guides, rather than concrete, all-inclusive lists. The most appropriate choice of data sources should eventually be determined by considering the costs and benefits of each source. A list of factors that may be useful to consider for this purpose are discussed toward the end of this section. It is also worth noting that while big data may not be a feasible alternative to conventional data for every project or assessment (if only at present), it can still complement and supplement current approaches or be used to validate their outcomes and analyses.[45]

As discussed in Part 1, assessing OPTRSR is a procedure that must be conducted at the project Identification stage to inform design and other assessments at the Preparation stage. Table 7 provides an overview of potential big data sources for the road safety assessment procedures discussed in Part 1 (Methods I-VII). Tables 8 to 10 discuss data sources that could be useful for each of the three primary methods for estimating OPTRSR, based on their respective data requirements.

[45] Holly Krambeck, Magreth Kakoko, and Mireille Raad, *Using Computer Vision to Automatically Detect Road Features for Road Safety Audits and Assessments: Inception Report* (Washington, DC: World Bank, 2019).

TABLE 7: **Overview of potential big data sources for Methods I-VII**

TYPE OF DATA REQUIRED	WHICH METHODS IT'S USED FOR	POTENTIAL BIG DATA SOURCE	EXAMPLES
Crash data from 3–5 years	Methods I, V and VI	Government	Government portal or contact
		Mobile applications and telematics	Waze
		Crowdsourced	Waze
Operating speeds	Methods II to IV	Mobile applications and telematics	Mapbox, Waze
Road features (road markings, signs, traffic calming measures, etc.)	Methods III, V, VI, and VII	Street view imagery	Mapillary
		Crowdsourced	OSM
		Aerial and satellite imagery	Maxar, UAV
Road type (urban road, pedestrian area, etc.)	Methods III, V, VI, and VII	Street view imagery	Mapillary
		Crowdsourced	OSM
		Aerial and satellite imagery	Maxar, UAV
		Mobile applications	Orbital Insight
Vehicle fleet mean speed	Methods III to VII	Mobile applications and telematics	Mapbox, Waze
Traffic flow	Methods IV to VII	Traffic imagery	Mapillary
		Aerial and satellite imagery	Maxar, UAV
		Mobile applications and telematics	Mapbox, Waze

SOURCE: Original table for this publication.

For crash data-based risk assessments (Method I), at least three years of historical crash data is required to cover three assessment criteria: infrastructure, road users, and speeds. Government data can be supplemented with data from mobile applications and telematics software, which may also have crowdsourcing capabilities, such as Waze. However, it may be a challenge to access three or more years of historical mobile or crowdsourced data. Table 8 summarizes the different data sources that can be used, although it does not include sources for two assessment criteria (vehicle standards and post-crash trauma care).

TABLE 8: **Method I — Crash data-based risk assessment**

REQUIREMENTS	DATA SOURCE	COMMENTS
Crash data from 3–5 years	Government	May be underreported; see Road Safety GPN
	Mobile applications and telematics	Companies providing mobile map apps or crash-related data within apps could be a resource for crash data
	Crowdsourced	Waze incident reports (minor or major crash)
		Incident reports from delivery drivers
		Social media text analysis, such as from Twitter

SOURCE: Original table for this publication.

If crash data is not available, Method II uses iRAP Star Ratings on the existing road to evaluate road infrastructure risk, and an assessment of the other criteria. Big data can be considered for evaluating road features, traffic flows and users' behaviour and complement iRAP Ratings. Table 9 highlights alternative big data sources that can be used to assess non-infrastructure risk. iRAP is also exploring the use of big data such as geo-located crash data to produce iRAP Risk Maps of the historical crashes per kilometer, and analyze road attributes, traffic flows, and speed data and map the safety performance and Star Rating.[46] Such a methodology would also require the use of ML, which is discussed in the next section.

[46] Omdena, "Rating Road Safety Through Machine Learning to Prevent Road Accidents," accessed May 28, 2021, https://omdena.com/projects/ai-road-safety/

TABLE 9: **Method II — iRAP Star Rating (alternative data sources using big data)**

REQUIREMENTS	DATA SOURCE	COMMENTS
Road users (behavior)		
Seat belt use for front passengers	Traffic imagery	Road surveillance images have been used to monitor front-row passengers wearing seat belts; potential to apply this to images (or video).
Child restraint and rear seat passenger seat belt use	N/A	N/A
Motorcycle helmet use	Street view imagery	Potential to identify helmet use among motorcyclists.
Operating speeds (km/h) during non-peak hours (not speed limits) for each **road type**		
Traffic video	Government data or collected by team	Video images can be used to calculate traffic flows and speeds.
Operating speeds	Mobile applications and telematics	Often provided as average speed per road segment in varying temporal resolutions.

SOURCE: Original table for this publication.

Big data can also be used to evaluate road safety risk without crash or iRAP data (Method III). Road infrastructure, operating speeds and other risks to road users may be estimated using various sources of big data (table 10). Combined with ML, clustering and other advanced analytical techniques, these data sources can also be used to model high-risk crash sites to project crash risk probability, frequency, and severity. This is discussed more in the following section.

TABLE 10: **Method III – Estimating road infrastructure risk without crash or iRAP data**

REQUIREMENTS	DATA SOURCE	COMMENTS
Road infrastructure		
Extent of separation of pedestrians from traffic with provision of safe walking spaces and crossing locations (if pedestrians are present or likely to be present post-project)	Street view imagery	Identify safe walking paths and crosswalks, traffic lights and signals
	Crowdsourced	OSM footways, intersections
	Aerial and satellite imagery	Identify walking paths and crosswalks
Extent of roadside safety barriers (omit this factor from consideration if the operating speed is <40 km/h)	Street view imagery	Identify barriers
	Crowdsourced	OSM (e.g., cable barriers or guard rails)
	Aerial and satellite imagery	Depending on image resolution and type of barrier in the ROI
Extent of median separation (omit this factor from consideration if the operating speed is <60 km/h for a rural road and <40 km/h for an urban road)	Street view imagery	Identify road medians
	Crowdsourced	OSM (e.g., cable barriers)
	Aerial and satellite imagery	Depending on image resolution and type of median in the ROI
Extent of separate well-designed motorcycle lanes (if motorcycles are present on the road or roadside or likely to be present post-project)	Street view imagery	Identify motorcycle lanes
	Crowdsourced	OSM (e.g., motorcycle lanes)
	Aerial and satellite imagery	Identify motorcycle lanes
Extent of separate off-road bicycle lane (if bicycles are present on the road or roadside or likely to be present post-project)	Street view imagery	Identify bicycle lanes
	Crowdsourced	OSM (e.g., cycleways)
	Aerial and satellite imagery	Identify bicycle lanes
Road users		
Seat belt use for front passengers	Street view imagery	Road surveillance images have been used to monitor front-row passengers wearing seat belts; potential to apply to images or video
Child restraint and rear seat passenger seat belt use	N/A	N/A
Motorcycle helmet use	Street view imagery	Potential to identify helmet use among motorcyclists
Operating speeds (km/h) during non-peak hours (not speed limits) for each road type		
Operating speeds	Mobile applications and telematics	Often provided as average speed per road segment in varying temporal resolutions
Road type (pedestrian area; urban area without pedestrians; open road, not median separated; open road, median separated)	Street view imagery	Identify pedestrian and non-pedestrian areas, open roads, and medians
	Crowdsourced	OSM roadways, footways, cable barriers or guard rails
	Aerial and satellite imagery	Pedestrian area, area without pedestrians, medians
	Mobile applications	Foot traffic, such as from Orbital Insight

SOURCE: Original table for this publication.

Projects that require reporting RSSAT results (Method IV) in addition to the OPTRSR (such as Type A projects, see Part 1) can turn to the big data sources highlighted in table 11 as an alternative or complement to traditional sources. Where existing data may be scarce or of poor quality, these sources may provide faster, more comprehensive and reliable data to estimate baseline risks.

Similar big data sources can be used for road infrastructure evaluations that involve Methods V-VII. Speed limits may be provided by the government. Roadside attributes, intersections, and mid-block attributes can be detected by ML algorithms applied to street view images.

TABLE 11: Method IV – RSSAT

REQUIREMENTS	DATA SOURCE	COMMENTS
Crash data from 3–5 years (annual fatalities, serious injury/fatality ratio; fatalities by vehicle occupant, motorcyclist, bicyclist, or pedestrian)	Government	May be underreported; see Road Safety GPN
	Mobile application	Companies providing mobile map apps or traffic-related data within apps might be a resource for crash data
	Crowdsourced	Waze incident reports (minor or major crash)
		Incident reports from delivery drivers
		Social media text analysis, such as from Twitter
Vehicle fleet mean speed	Mobile applications and telematics	Often provided as average speed per road segment in varying temporal resolutions
Segment characteristics (number of lanes per travel direction; lane width, paved shoulder width, terrain type, median type; road marking and signs; pedestrian and bicycling facilities, service road)	Street view imagery	Number of lanes, lane width, paved shoulder width, terrain type, median type, road marking and signs, pedestrian and bicycle facilities, service road
	Crowdsourced	OSM
	Aerial and satellite imagery	Number of lanes, lane width, paved shoulder width, terrain type, median type, road marking, pedestrian and bicycle facilities, service road; road signs will be a limitation
Dominant roadside object (safety barrier; minor hazards; slope; trees, poles, and fixed objects; cliff or steep drops)	Street view imagery	Safety barriers, static roadside objects, minor hazards; in some cases, cliff or steep drop may be possible
	Crowdsourced	OSM (e.g., barriers)
	Aerial and satellite imagery	Elevation for slope and steep drops; in some cases, static roadside objects, minor hazards or safety barriers
Speed management or traffic calming measures (percentage of road length)	Street view imagery	Identify physical speed inhibitors
	Crowdsourced	OSM traffic calming features by type
	Aerial and satellite imagery	Identify physical speed inhibitors
Intersection characteristics (grade separated, roundabout, signalized junction, unsignalized junction)	Street view imagery	Grade separated, roundabout, signalized junction, unsignalized junction
	Aerial and satellite imagery	Grade separated, roundabout
Pedestrian crossing (grade separated, signalized crossing, marked crossing)	Street view imagery	Grade separated, signalized crossing, marked crossing
	Crowdsourced	OSM pedestrian crossing features by type
	Aerial and satellite imagery	Grade separated, marked crossing
Traffic flow (motorized and non-motorized; both directions, per day)	Street view imagery	Static camera at a set location is preferrable
	Aerial and satellite imagery	Presents temporal limitations
	Mobile applications and telematics	Provides temporal granularity

SOURCE: Original table for this publication.

Big data sources may also be useful to monitor and evaluate indicators for the Results Framework. Table 12 provides examples of a few big data sources that could be used for the indicators covered in table 5.

TABLE 12: **Example of big data sources for road safety indicators in the Results Framework**

EXAMPLE OF INDICATOR	TYPE OF DATA THAT CAN BE COLLECTED	EXAMPLE OF BIG DATA SOURCE
Reduction of road crashes	Crash data	Government, open source data, Waze
Speed reductions	Traffic flows	Video images, telematics, mobile applications
Increased use of helmet and seat belts	Number of helmet and seat belt users	Street images and security video

SOURCE: Original table for this publication.

As a broader variety of big data sources become available, Borrowers and TTLs are advised to carefully consider the trade-offs involved when collecting data from various sources. Here is a list of factors to consider, as well as some guidance on how each of these can affect project outcomes and constraints. This is not an exhaustive list. Some factors may be more relevant to some projects than others, while additional considerations may be required for certain projects. In some cases, data from existing sources may not be available and will need to be collected using cameras, sensors, and/or other tools. The World Bank Data Lab provides resources to find, collect, manage, and gain insights from data, including access to Lab Leads who can give project-specific advice.[47]

- **It is worth noting that many of these factors are also interrelated.** For example, the types and quantity of data required could impact costs of obtaining and processing it. Costs can also vary by region, as can the availability of resources to process and analyze the data. This list may be used in tandem with Annex 2, which provides an overview of the most relevant big data sources for road safety analysis as well as their relative costs, data attributes and formats, and possible limitations.
- **Type of road safety assessment or procedure.** As discussed in Part 1, a broad range of tools and procedures are used for road safety assessments across World Bank projects. Each tool has its own specific data requirements. It is important to consider these before determining appropriate big data sources to complement analysis.
- **Context/Region(s) of Interest.** The types and variety of big data sources available can vary greatly from region to region, country to country, or even different provinces or localities within the same country. For example, Waze crowdsourced crash data is especially useful for urban regions that are more densely populated compared to rural regions.
- **Type of data required.** As more big data sources become available for road and traffic data, the task team should carefully consider which variables and data types are most relevant to their model before selecting a source. For example, Google offers a number of APIs that may be useful for road safety analysis. This includes Google Maps, Google Traffic and Google Street View. It is important to consider the quantity, duration, and extensiveness of the data required. For example, some data sources include time-series information, others do not. Some may include specific road features or road user data, while others may just be focused on traffic flows.
- **Data formats.** Big data is collected, stored, and transmitted in a wide range of formats. It is important to consider the usability of available big data formats as well as their interoperability with other types of data. Since many big data sources that are currently available are not custom designed for road safety analysis, task teams should be prepared to have some expertise and resources to extract, aggregate, clean, and convert the data into a format that can be combined with other data and/or used with analytical tools and models.
- **Cost.** Given the size of big datasets, costs can arise from accessing, storing, handling, processing, and analyzing the data. The cost may be in the form of data licenses, software licenses or

[47] World Bank Data Lab, https://wbdatalab.org/

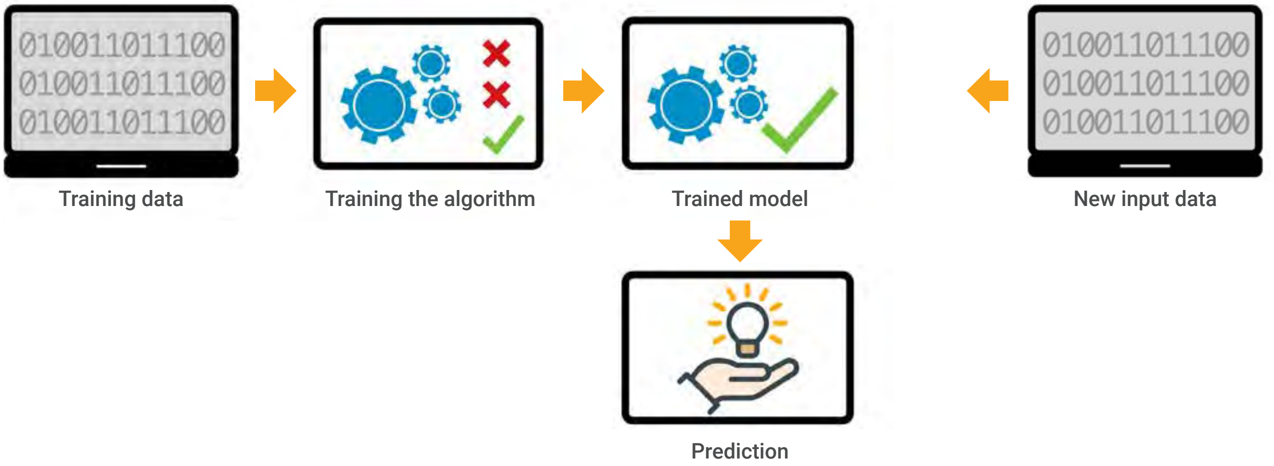
equipment (if the data is being collected specifically for the project at hand). Besides the cost of obtaining the data, it is also important to consider the cost of using it, such as by acquiring the necessary expertise, software tools and processing power for analysis. Annex 2 discusses the relative costs associated with using different big data sources.

- **Resources required to make data usable.** In addition to relevant data sources and the costs that may be associated with accessing them, other resources could also be required to utilize the data in road safety assessment and analysis. This includes technical skills and expertise required to handle and analyze the data.
- **Time constraints.** Some big data sources are faster to access and obtain data from compared to others. For example, open data platforms allow you to run a search query and instantly obtain relevant datasets. Other avenues, such as data sharing agreements, may take longer to deliver the required data. It is important to consider the project timeframe to determine which data source may be more useful for road safety analysis at a given stage.
- **Licensing constraints.** Any official and legitimate data source is accompanied by licensing regulations that outline the terms of use of the provided dataset. Big data sources are no exception. Different data sources have different licensing agreements associated with them. Some, such as open data platforms, may have minimal licensing restrictions. Others, such as APIs and datasets obtained through data partnership agreements, can have more restrictive terms of use. It is important to carefully consider these limitations before choosing a source. TTLs are advised to consult the World Bank's legal team or the data provider to fully understand licensing restrictions associated with different big data sources to avoid legal ramifications.

2.2 Machine Learning in Road Safety Analysis

ML is a branch of artificial intelligence. It involves creating algorithms that "learn" patterns, trends and behaviors from data and improve accuracy over time without further programming. As figure 6 illustrates, the lifecycle of an ML model can be typically divided into two phases: training and deployment. In the training phase, training data is fed into the algorithm to obtain a trained model. In the deployment phase, new input data is fed into the trained algorithm (or model) to predict the output.

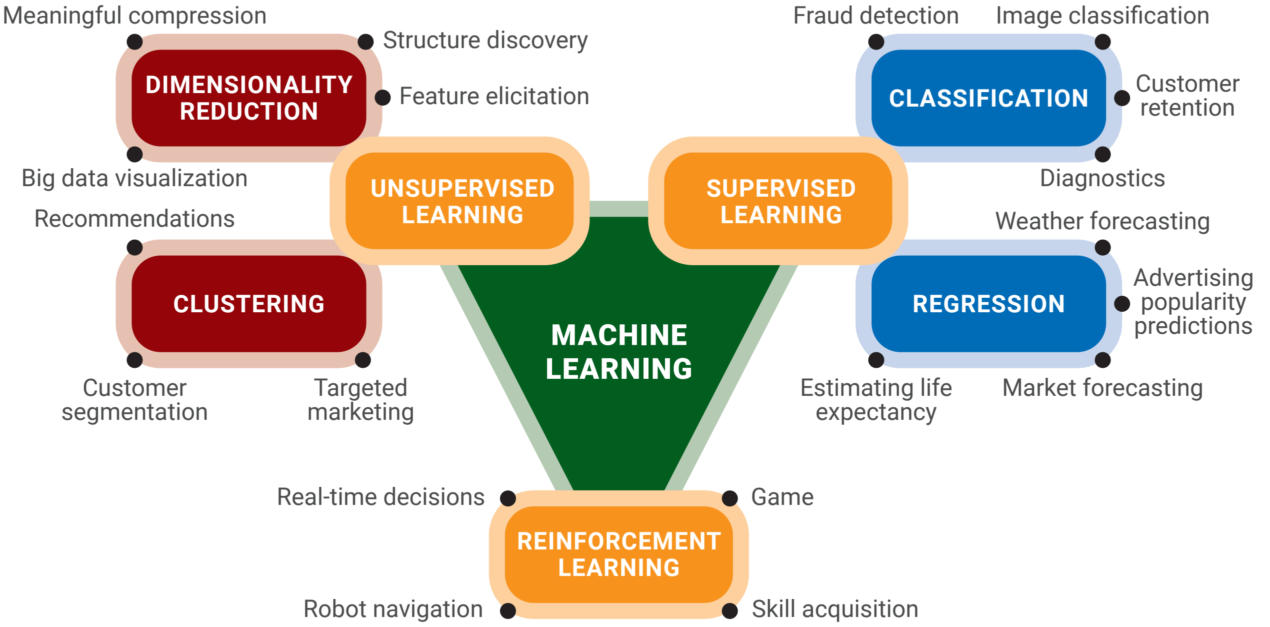
FIGURE 6: **ML lifecycle**



SOURCE: Modified from https://randomtrees.com/data-science

As shown in figure 7, ML algorithms can be divided into three categories: supervised learning, unsupervised learning, and reinforcement learning. The specific tasks they are capable of and the corresponding algorithms that are most widely used for this purpose are also listed in table 13. One significant difference between these categories is the format and the source of the training data.

FIGURE 7: Categories of ML and the tasks they can perform



SOURCE: Modified from https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d

Supervised learning is a family of algorithms that learn from previous data to map an input (X) to an output (Y). For example, a supervised learning algorithm can be used to predict the risk level or crash frequency (Y) of a road segment given its characteristics (X). "Supervised" means the training data is labelled (i.e., the training data should be pairs of X-Y, where Y is usually called labels).

Unsupervised learning algorithms find structures in a dataset in order to group or cluster data points based on their similarity. As the name suggests, these algorithms do not require "supervision" or human intervention in the training phase. This means that, unlike supervised learning, the training data for unsupervised learning algorithms has no labels (Y). These algorithms learn to group X based on similar characteristics. The most common unsupervised learning task is clustering. For example, given the characteristics of a road segment, an unsupervised learning algorithm can classify it into a group of similar segments. It does not need to understand the characteristics that the group represents to complete this task.

Reinforcement learning trains a software agent to make decisions that maximize rewards from interactions with an external environment.[48] As opposed to supervised learning and unsupervised learning, which require training data to be prepared before training, reinforcement learning generates the training data during the training phase. The data is generated when the agent interacts with the environment. For example, reinforcement learning can be used to train an agent to control traffic lights based on traffic conditions.

[48] This agent is a piece of software that makes a decision based on the environment.

TABLE 13: **Categories of ML and algorithms***

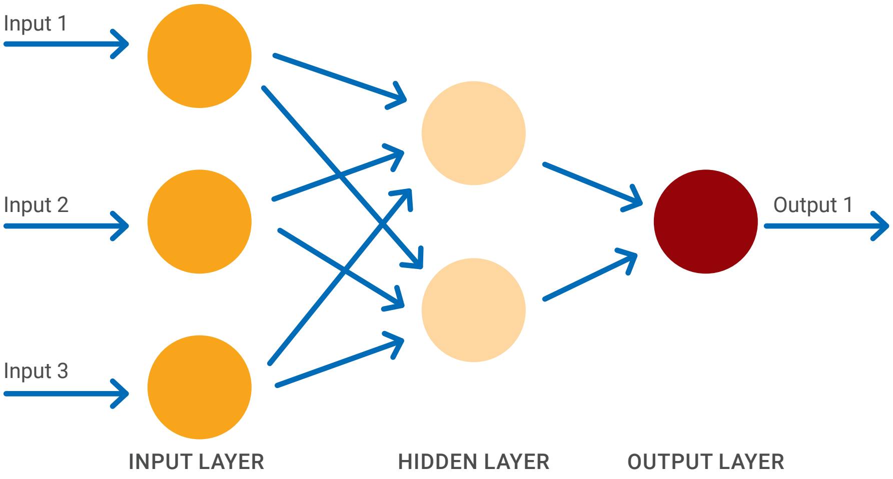
	ALGORITHMS	TASKS
Supervised Learning	SVM, DT, RF, KNN, ANN	Classification
		Regression
Unsupervised Learning	K-means, PCA, ANN	Clustering
		Dimensionality Reduction
Reinforcement Learning	Q-Learning, DQN	Robotics/Decision-making

Source: Original table for this publication.

*The algorithms listed in this table are not exhaustive.
SVM: support vector machine
DT: decision trees
RF: random forest
KNN: k-nearest neighbors
ANN: artificial neural networks
PCA: principal component analysis
DQN: deep Q-network, which includes and ANN in its algorithm

Artificial neural network (ANN) is a family of ML algorithms that have been inspired by the human brain. ANN is the most versatile ML algorithm – it can be used for supervised learning, unsupervised learning, and also reinforcement learning. As shown in figure 8, ANN structures the data and the computation in different layers. Every layer adds more depth to the algorithm; therefore, more layers indicate that it is "deeper". Such ANNs are called deep neural networks or deep ANN or DNN. ML algorithms that use deep ANN are called deep learning (DL) algorithms. Therefore, from another perspective, ML algorithms can be divided into conventional ML and DL (table 14).

FIGURE 8: **ANN structure**



SOURCE: Original figure for this publication.

TABLE 14: **ML and DL algorithms**

	CONVENTIONAL ML*	DL
Supervised Learning	SVM, DT, RF, KNN, shallow ANN	Deep ANN
Unsupervised Learning	K-means, PCA	Deep ANN
Reinforcement Learning (RL)	RL without deep ANN	RL with deep ANN

*The conventional ML algorithms listed in this table are not exhaustive.

SOURCE: Original table for this publication.

Most ML algorithms are conventional ML, such as conventional supervised learning algorithms like support vector machine (SVM), which can be used for classification or regression, for example, classifying the risk level of a road segment based on its characteristics. Conventional unsupervised learning algorithms, such as K-means clustering, automatically identify spatial patterns in datasets, which can be applied to locate clusters or areas with recurring road crashes. Conventional ML works well for small, low dimensional datasets. Meanwhile, DL is a subset of ML that learns the complex patterns from high dimensional (e.g., an image) and large quantities of data (e.g., big data). Supervised, unsupervised, and reinforcement learning algorithms that use deep ANN technique belong to the DL category. DL's first successful application is in the computer vision area. For example, image classification is a supervised learning task that utilizes deep neural networks to classify images into different classes (e.g., cars, pedestrians, etc.).

How to Use Machine Learning

The use of ML methods in road safety analyses is being widely explored.[49] As ML methods become more advanced, economical, and accessible, their potential applications in various disciplines continue to grow and become more feasible. In road safety analyses, ML has great potential to overcome the limitations of traditional statistical models in crash analysis and crash probability modeling. The applications of ML in road safety analyses are discussed under three categories: conventional ML, DL, and reinforcement learning, as listed in table 15. It should be noted that some reinforcement learning algorithms using deep ANN belong to DL, but all reinforcement techniques are discussed separately.

TABLE 15: **Frequently used ML techniques for road safety analysis***

ML CATEGORIES	SUBCATEGORIES	ALGORITHMS	TASKS	EXAMPLES
Conventional ML	Supervised Learning	SVM DT RF KNN shallow ANN	Classification	Predict risk level based on road characteristics.
			Regression	Crash frequency prediction based on road characteristics.
	Unsupervised Learning	K-means	Clustering	Group road segments by characteristics similarity; group drivers based on their driving behaviors.
		PCA	Dimensionality Reduction	Identify critical factors of road safety.
DL	Supervised Learning	CNN	Image Classification/ Object Detection/ Segmentation	Detect road features from images.
	Unsupervised Learning	GAN	Clustering/Dimensionality Reduction	Find the hidden features related to road safety from map and satellite images of the road environments.
Reinforcement Learning	N/A	Q-Learning DQN	Robotics/Decision-making	Control traffic lights based on traffic conditions.

*The algorithms and examples listed in this table are not exhaustive.
CNN: convolutional neural network, a type of deep ANN
GAN: generative adversarial networks, a type of deep ANN

SOURCE: Original table for this publication.

A growing body of research explores various ML techniques to predict the probability of road crashes and assess their severity by training on historical datasets that encompass diverse factors. Conventional ML algorithms are the most frequently used ML algorithms for this purpose. They are summarized in table 15. ML-based approaches to road safety analysis can be used to complement, supplement or even potentially substitute conventional road safety assessments.

Conventional supervised learning algorithms learn functions that take vectors of variables as input to predict the output. Most conventional supervised learning algorithms that are frequently used in data science have been used in road safety analyses, including but not limited to: decision

[49] Philippe Barbosa Silva, Michelle Andrade, and Sara Ferreira, "Machine Learning Applied to Road Safety Modeling: A Systematic Literature Review," *Journal of Traffic and Transportation Engineering* (English Edition), 7, no. 6, (2020), https://www.sciencedirect.com/science/article/pii/S2095756420301410

trees (DT), random forest (RF), support vector machine (SVM), k-nearest neighbors (KNN), and artificial neural networks (ANN).[50] It should be noted that there is no "best" algorithm. Determining which algorithm may be most appropriate for an ML-based road safety analysis is essentially a data science problem for which there are usually no set rules. One algorithm may perform well for a dataset, but badly for another. It is common practice for data scientists to try different algorithms in order to find a suitable one for a specific problem. When using the aforementioned conventional supervised learning algorithms for road safety assessments, the problem is often framed as a classification or regression problem, in which the output (Y) of the ML algorithm is either a class (e.g., risk level or severity: low, moderate, substantial or high) or a scalar (e.g., crash probability, crash frequency) and the input (X) to the ML algorithm could be any parameter (including but not limited to weather, time, road factors, human factors, etc.) that is related to the output. For example, one way to calculate OPTRSR is to frame it as a classification problem, in which the output of the model is the OPTRSR risk level, while the input is a vector of variables describing road features and typical vehicle operating speeds, or other factors that could be used for evaluating the OPTRSR risk level. Any aforementioned conventional supervised learning algorithm would be suitable for this example.

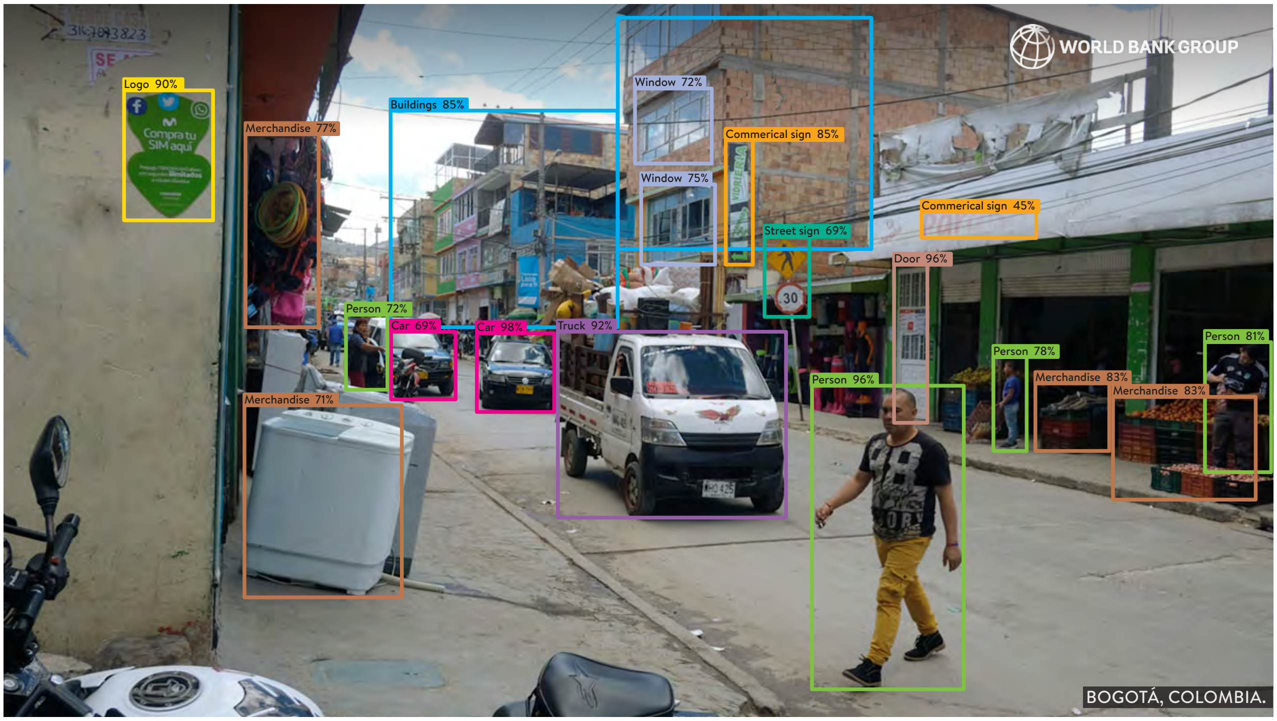
Conventional unsupervised learning algorithms are mainly used for clustering and dimensionality reduction purposes. In road safety analyses, K-means can be used for grouping tasks that help find clustering patterns in the data. For example, it can be used to group road segments by similar characteristics or group drivers based on their driving behaviors, so that dangerous road segments or drivers can be identified based on the similarity. In another example of unsupervised learning application, principal component analysis is used for reducing the dimensions of input data to identify the most critical factors that affect road safety.

DL has been applied in various disciplines and achieved impressive performance. DL technologies have progressed significantly over the past few years, especially in image analysis and computer vision, the method's first successful application. The core technique in this domain is deep convolutional neural network (CNN), which is the state-of-the-art approach for object detection, semantic segmentation, and instance segmentation of images. Object detection is a task in which, given an image, the model outputs a bounding box of detected objects (figure 9). Semantic segmentation is a task in which, given an image, the model classifies every pixel into predefined classes (e.g., road lane, traffic light, etc.). Instance segmentation is a task, in which, given an image, the model groups pixels belonging to an instance of the object.

[50] Silva, Andrade, and Ferreira, "Machine Learning Applied to Road Safety Modeling: A Systematic Literature Review."

FIGURE 9: **ML algorithms and street view**

After applying an object detection algorithm to a street view image, a bounding box surrounds each predicted object, which also contains a confidence level for each prediction.



SOURCE: World Bank Global Program for Resilient Housing.

DL-based image analysis has been successfully used in various industries for applications ranging from facial recognition to autonomous driving. It has great potential to be used in road safety analysis to automatically analyze images and infer road attributes that are relevant to road safety assessments. Large sets of images with annotations such as road lanes, traffic lights, speed limit signs, and pedestrians can be compiled for training deep CNNs so that they learn to recognize these objects through images that the models have not previously encountered. If successful, this approach should equip the model to detect road attributes at a regional scale.

The detected information can then be used for safety and risk analysis. For example, if the DL model can infer the road segment characteristics (e.g., number of lanes, terrain type, road markings and signs, and pedestrian, bicycling, and motorcycling facilities), the inferred information can readily be used as input for the RSSAT tool (Method IV). This would allow the process of detection and analysis to become fully, or at least significantly automated and scalable at a low cost.

DL can also provide a lower-risk alternative to manual detection of certain road attributes and other important variables in road safety analysis. For example, a team used imagery from Baidu Street View to provide a practical, automated alternative to the manual detection of street cracks, which can be labor-intensive, hazardous and difficult to conduct on a large scale.The authors use the Deeplabv3+ network model, a DL neural network, to develop an automated road crack identification system and demonstrate its practicality as a method to generate faster, more accurate and efficient information about road cracks at lower cost compared to manual detection.[51]

[51] Min Zhang et al., "Research on Baidu Street View Road Crack Information Extraction Based on Deep Learning Method," *Journal of Physics: Conference Series*, no. 1616 (2020). https://iopscience.iop.org/article/10.1088/1742-6596/1616/1/012086/pdf

Reinforcement learning is widely used to design intelligent control and decision-making systems. In road safety and traffic management, reinforcement learning is most commonly employed to develop intelligent signal control algorithms. A typical reinforcement learning-based traffic light system makes divisions based on specific input traffic parameters, such as the length of time for which vehicles wait at the intersection, the cumulative delay caused by waiting at the intersection, the length of time for which the light stays green for each signal head, etc. The output of the system would be the next color of the light and length of time for which it should remain switched on. Designing traffic systems using reinforcement learning helps save time and improve safety standards.

Key Considerations for Using Machine Learning

Road safety can be evaluated explicitly using rule-based reasoning systems, such as iRAP star score and RSSAT. However, developing such systems can be complex if there are many input variables. Compared with rule-based evaluation systems, ML algorithms are data-driven and don't require developing rules; therefore, they are relatively inexpensive to implement. ML algorithms are more suitable for high dimensional inputs. As a broader spectrum of ML algorithms become available, TTLs are advised to carefully consider the trade-offs involved when applying them to road safety analysis. This section discusses various factors that task teams must consider before deciding to use an ML algorithm for road safety analysis in their project. Again, this is not an exhaustive list. Some factors may be more relevant to some projects than others, while additional considerations may be required for certain projects. It is worth noting that many of these factors are also interrelated. For example, the feasibility of using ML for a project can be affected by time and budget constraints, the availability of data, and the anticipated resource intensiveness of the data preparation process. Table 16 provides a SWOT analysis of the use of ML in road safety analysis.

TABLE 16: **SWOT analysis of using ML in road safety analysis**

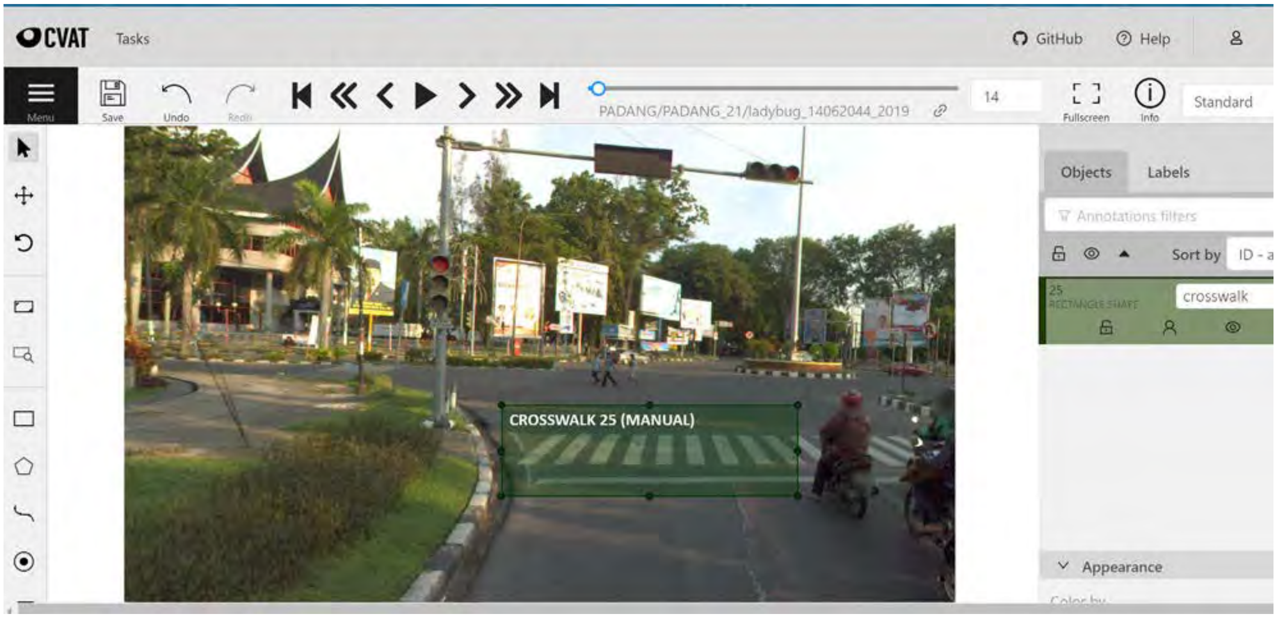
STRENGTHS	WEAKNESSES
• Offers tools and techniques to process big data that may be more precise compared to traditional methods. • Especially effective for feature learning, parameter optimization, and processing large amounts of big data. • ML algorithms tend to perform better than traditional statistical techniques in cases where high-dimensional and high-nonlinear data is involved. • As the technology develops, novel techniques create new opportunities to understand complex relationships between multiple, interrelated variables and predict outcomes with greater accuracy. • ML algorithms can be improved continuously as more data is generated or made available for training.	• Algorithms can be limited in their applicability; models may not perform well on data that is different from the training data's distribution. • Large amounts of data are needed to train the models and yield more accurate models, which may be difficult in data-scarce contexts. • Some ML algorithms (e.g., ANN) works like a black box, and can be hard to interpret, therefore an ML algorithm usually requires thorough validation and test processes before it can be deployed in the real environment and assist decision-making. • The technology still needs further development before it can be mainstreamed for use in road safety assessments.
OPPORTUNITIES	THREATS/CHALLENGES
• May eliminate the need for manual coding of road safety data in the future, making the process less labor-intensive and time consuming. • Possible to train datasets in one location or for one purpose and use them for another. • Provides a powerful method for complex crash risk modelling and other types of predictive analytics in road safety. • As the technology develops, a platform powered by ML could be used across geographies for road assessments. • As more and more data is generated and collected everyday, this could be potentially analyzed with ML algorithms to discover new patterns and insights.	• Requires specialist expertise, tools, and knowledge which may make its usefulness limited in some contexts, especially in developing countries. • May require additional investment in computer power and analytical software. • Complexity of ML algorithms can make them difficult to implement and analyze. • Ethical considerations, such as bias in ML systems. • As a data-driven approach, ML relies on high-quality data for training. Significant bias in the training data could lead to the failure of model training. Quality control of training data could be difficult, especially when annotating the data requires professional knowledge.

SOURCE: Original table for this publication.

Feasibility with project objectives and client requirements. Before deciding to use ML for any project, it must be ascertained if ML is suitable for the project. Some ML algorithms, such as neural networks, are not interpretable. They work like a black box. Clients may not have confidence in using them for significant decision-making, unless their predictions can be sufficiently validated.

Preparing data to train ML algorithms. ML is a data-driven approach. Therefore, as with any data-related project, it is important to plan the data collection and preparation process. To facilitate this process, make sure to have clearly defined the inputs and outputs of the model at the outset of the project. Section 2.1 provides guidance on how to select data sources, especially where big data may be involved. It is common that, during the training stage, an ML team may find the data is not enough to train a model with satisfactory performance. In this case, more data needs to be collected. In terms of data preparation, teams should be aware of the need to aggregate, clean and annotate data before it can be used for ML modelling. Annotation of data is especially necessary for supervised learning algorithms and entails manually identifying an object drawing a box or polygon around it and giving it a label such as “pothole” or “crosswalk” (figure 10).

FIGURE 10: **Labeling a crosswalk in Padang, Indonesia using the Computer Vision Annotation Tool (CVAT)**



SOURCE: World Bank Global Program for Resilient Housing.

Teams are advised to incorporate a quality control process to ensure data being used for any ML model, especially test data, is of good quality and truly valid and representative of the population or situation under study. For an ML-based project, steps include: (i) identifying data required for the model; (ii) data collection, cleaning, annotation; (iii) trial and error training; (iv) validation; (v) deployment. Task teams should estimate the duration of these tasks, considering their expected complexity and potential challenges (which can vary by context and availability of resources such as expertise and processing power). This will help them determine if ML is feasible for their project, how it compares to traditional methods and how incorporating ML can impact project timelines. It is worth noting that once deployed in the production environment, ML provides significant acceleration for the whole process, for example, DL-based image analysis can exponentially save the time for collecting data to be used in the road risk estimation.

A challenge for most ML algorithms is generalization, or how well a model can perform based on test data (also called unseen data). Models may not perform well on unseen data that is different from the training data's distribution. For example, a model that is trained on images collected on rural roads in an arid climate may not achieve the same level of performance on images in urban roads in another country. The transferability of the model depends on how similar the features in the images are. Therefore, before training ML algorithms, it is prudent to consider the diversity of the training data, especially in terms of where, how and when it was collected. It is worth noting that some researchers have found that artificial intelligence and ML algorithms can be easily and accurately applied to different types of urban networks within the same city.[52]

To determine if using ML fits a budget or can even deliver a cost-advantage, it is important to understand associated costs. Costs of using ML can arise from the hiring of experts to develop and program models, as well as from the data collection and preparation process (which includes cleaning

[52] Apostolos Ziakopoulos and George Yannis, "Using AI for Spatial Predictions of Driver Behavior" (presentation, ITF International Transport Forum Roundtable on Artificial Intelligence in Road Traffic Crash Prevention, 2021). https://www.nrso.ntua.gr/geyannis/conf/cp450-using-ai-for-spatial-predictions-of-driver-behavior/

and annotation). The cost of storing data (on local hardware or on the cloud) should also be accounted for, especially if the inputs involve big data. Depending on the model and quantity of data being input, and especially if a DL model is employed, you may also need to invest in additional computational resources (graphics processing unit-equipped local computers or nodes on the cloud). Front-end and back-end systems may also need to be established for automatic analysis services.

Deploying ML algorithms requires specialized expertise, often in the form of dedicated team members that are ML experts. TTLs can choose to hire experts and manage the process internally or acquire resources externally. An in-house, "do-it-yourself" approach ensures more control over every aspect of the process, which may be especially important where significant customization or trial and error may be required. However, this approach requires labor and time, and may be more costly in the long run. Using an external resource or tool, on the other hand, may be a faster option but can come at the expense of some visibility and control over the development of the model. It is important to consider these trade-offs to ensure the team is adequately resourced to use ML effectively in the project.

2.3 Big Data, Machine Learning and the Future of Road Safety Assessments

Artificial intelligence presents many exciting possibilities for automation and analysis in transport and infrastructure development. ML is increasingly used for road safety analysis. ML's inherent capability of managing uncertainties in data and models makes it extremely suitable for solving road safety related issues. Uncertainty is a defining element of crash risk modelling and, in fact, a source of complexity that has thus far limited the usefulness of traditional statistical models. Moreover, ML algorithms such as deep ANN can capture nonlinear patterns in data, making them the first choice for processing road safety big data. Table 17 provides a summary of possible applications of big data and ML in road safety analysis given the current state of the technologies.

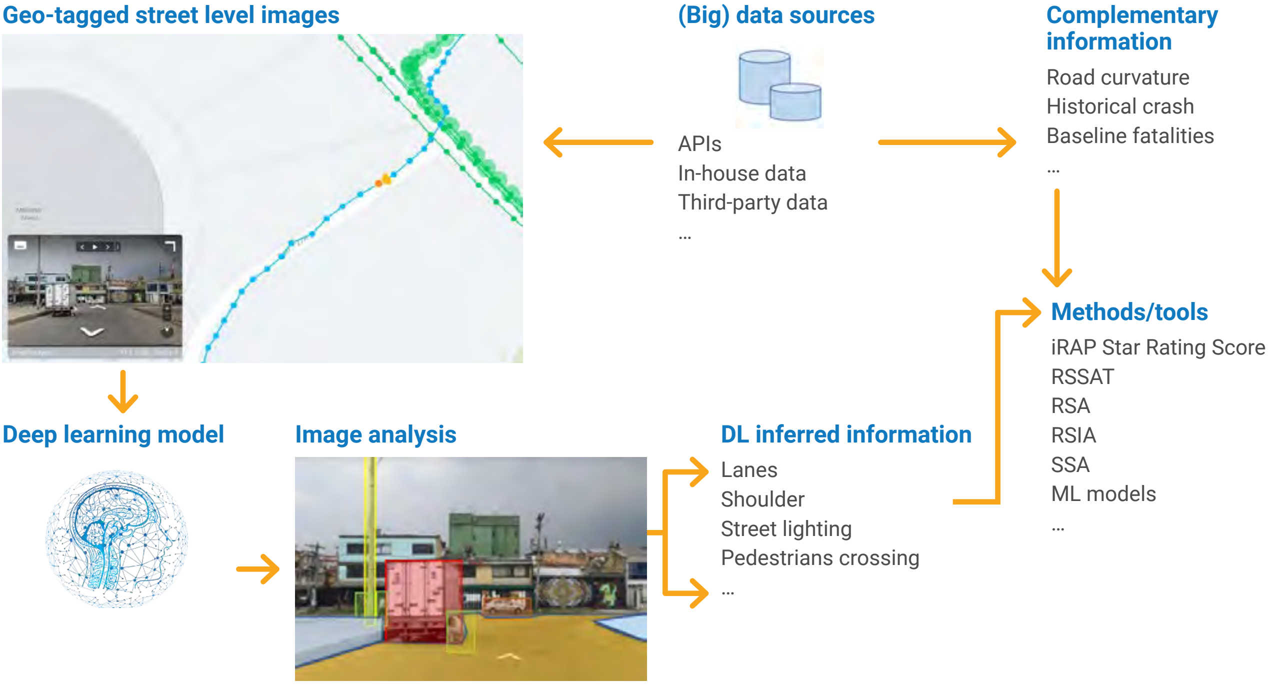
TABLE 17: **Potential applications of big data and ML in Methods I to VII**

POTENTIAL APPLICATIONS	HOW BIG DATA CAN HELP	HOW ML CAN HELP
Estimating Road Infrastructure Risk (Methods III, V, VI, and VII)	Video and photo images, APIs, satellite imagery and/or crowdsourced images	• Process images to evaluate road attributes • Identify road features that could cause crashes • Identify risk factors contributing to crash occurrence • Identify safety conditions in infrastructure
Traffic Flows (Methods IV to VII)	APIs, aerial imagery, open-source traffic data, road sensor data, wireless technology, street cameras, GPS data, mobile devices, real-time traffic data	• Process images to classify vehicles, identify congestion hotspots, vehicle detection, or speeds • Assess traffic flows • Develop risk maps • Map the safety performance and Star Rating • Traffic flows prediction
Crash Risk Assessment (Methods III to VII)	Meteorology data, geo-located crash data, video and photo images, APIs, open-source traffic data, road sensor data, historical crash data, crowdsourced crash data (e.g., Waze)	• Create crash prediction models • Develop risk maps • Analyze different conflict scenarios and high-risk behavior
Incident Reporting/Crash Data (Methods I, V and VI)	Video recording, crash data, photo images, crowdsourced data (Google Maps, Waze)	• Identify hotspots through clustering techniques
Analyzing Crash Severity (Methods III to VII)	Video and photo images, sensor data	• Process images to evaluate road attributes • Develop crash prediction models

SOURCE: Original table for this publication.

Combining big data and ML can provide an integrated framework for automatic road safety analysis and management. This framework, demonstrated in figure 11, employs platforms (such as Mapillary) to provide geo-tagged street level imagery for inputs to the DL model to infer useful information (e.g., road characteristics). The DL-inferred data is then combined with multi-source big datasets (e.g., region-specific historical crash data) for better analysis and management of road safety. For example, the combined information can readily be used as the input to Method I-VII for estimating the OPTRSR. Moreover, ML algorithms (e.g., ANN) have the potential to substitute traditional methods and tools (iRAP, RSSAT, etc.) for evaluating risks and safety indicators like OPTRSR.

FIGURE 11: Framework for automatic road safety analysis and management powered by ML



SOURCE: Original figure for this publication.

At present, much of the research and innovation in the use of ML for advanced road safety and risk modelling is being driven by universities and other research institutions. As other stakeholders, such as governments, developers of road safety tools and international organizations such as the World Bank look to apply ML in their projects, there is an opportunity to create dedicated tools that would harness big data and ML for road safety analysis. Such applications have the potential to reduce the risk of human error and allow road safety assessments to be mostly, if not fully, automated.

The following section presents practical examples of how big data and ML can assess urban road safety. It applies an integrated framework introduced in section 2.3 to explore the opportunities and limitations of new data sources and assess the ML models. To evaluate the robustness of the proposed framework, the Integrated Framework for Road Risk Prediction was applied in two cities of different sizes, regions, and data availability were chosen: Bogotá, Colombia, a rapidly urbanizing metropolis in Latin America, and Padang, Indonesia, a secondary city in East Asia. The study found that ML applied to street view imagery identified relevant road (and road user) characteristics to generate a model that predicts road risk with 72.5 percent accuracy in Bogotá. This framework was applied in Padang to test its replicability; preliminary results are encouraging for its potential to predict road safety for areas with limited crash data. The section concludes with a reflection and guidance for replicability.

PART 3

Case Studies: Applying Big Data and Machine Learning to Assess Road Safety

3.1 Objectives of the Case Studies

This section presents how the Integrated Framework for Road Risk Prediction can be applied in two different cities of interest: Bogotá, Colombia and Padang, Indonesia. The study examines how useful ML is in evaluating road safety and how easily the integrated framework can be replicated. All code is freely available for other teams to use and develop further.[53]

The objectives of the case studies are to:

1. Learn how well big data and ML can be used to identify road features, estimate road safety, categorize road segments based on their risk level, and identify high-risk segments.
2. Evaluate the utility of several big data sources that are freely available for road safety analysis in diverse geographic areas.[54]
3. Assess the replicability of the proposed approach.

Located on two different continents, the selected locations offer an opportunity to apply the framework on paved, urban roads in contrasting environments, particularly related to data availability and usability. For example, the government of Bogotá has made significant efforts to increase crash data collection and dissemination. The government offers an online portal with the location of each crash over the past year publicly available. In addition, there was high coverage for data derived from mobile phones, such as crowd-reported crashes. In contrast, information on the crash locations for Padang could not be found online, and methods for data collection are largely manual or paper based.[55] In addition, mobile application data was scarce for crowdsourced crash reports. As a result, Padang offers the opportunity to explore the utility of ML when data coverage is limited.

[53] The code for the Integrated Framework for Road Risk Prediction is open source and accessible on GitHub: https://github.com/datapartnership/IntegratedFrameworkForRoadSafety. However, some datasets require partnership with DDP to access.

[54] Freely available meaning at no cost; however, some data sources are not publicly available and require a license.

[55] World Bank, *Indonesia Public Expenditure Review 2020: Spending for Better Results* (Washington, DC: World Bank, 2020). https://openknowledge.worldbank.org/handle/10986/33954

BOGOTÁ AND PADANG: BACKGROUND AND CONTEXT

With a population of more than 7 million, the capital district of Bogotá is Colombia's largest city. As a critical economic hub with a growing population, Bogotá stands out as one of the most congested cities in the world.[56] The government has prioritized road safety and achieved significant gains over the past few decades, reducing the city's traffic fatality rate by more than 60 percent between 1996 and 2006 alone.[57] More recent interventions during the UN Decade for Action for Road Safety include establishing a National Road Safety Plan and a National Road Safety Agency (Agencia Nacional de Seguridad Vial) featuring a National Road Safety Observatory in collaboration with the World Bank.[58] In addition, in 2017, the city's government launched "Vision Zero," which aimed to implement a range of speed management strategies to eliminate pedestrian and driver fatalities. The program has delivered measurable results, such as a 27 percent reduction in fatalities across corridors where speed limits have been introduced, and further interventions are planned to sustain its impact.[59] Despite these initiatives and road safety improvements in Bogotá, challenges remain, and new policies would benefit from timely and affordable analytics on road safety.

Padang is the capital of the Indonesian province of Western Sumatra with a population of around 1 million. The government of Indonesia introduced various initiatives to address road safety during the UN Decade of Action for Road Safety. Established in 2011, the National Road Safety Master Plan achieved a 10 percent reduction in annual road fatalities between 2013 and 2016. However, data collection and management systems that rely on manual screening significantly challenge the country's progress in road performance and safety.[60] Initiatives such as the establishment of the Integrated Road Asset Management System and the World Bank's new Asia-Pacific Road Safety Observatory present a valuable opportunity for the country to improve its road safety data systems.[61] For this case study in Padang, crash data was scarce from alternative sources. Therefore, it offers the opportunity to explore the utility of the pre-trained ML models in a new region with limited data coverage.

[56] INRIX 2018 Global Traffic Scorecard. In 2018, drivers lost 272 hours in road congestion.

[57] ODI (Overseas Development Institute), "Bogotá," ODI: Think Change. Accessed October 12, 2021, from https://odi.org/en/about/features/bogot%C3%A1/

[58] World Bank, *Colombia - Programmatic Productive and Sustainable Cities Development Policy Loans* (Washington, DC: World Bank, 2020). http://documents.worldbank.org/curated/en/426591583968971309/Colombia-Programmatic-Productive-and-Sustainable-Cities-Development-Policy-Loans

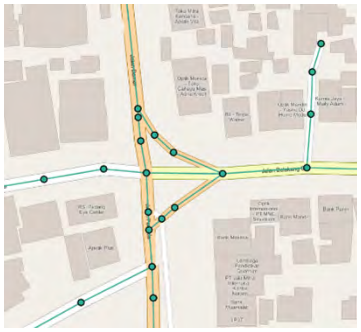
[59] Darío Hidalgo and Claudia Adriazola-Steil, "Bogotá's Vision Zero Road Safety Plan Is Saving Lives," TheCityFix, last modified September 26, 2019, https://thecityfix.com/blog/bogotas-vision-zero-road-safety-plan-saving-lives-dario-hidalgo-claudia-adriazola-steil/

[60] World Bank, *Indonesia Public Expenditure Review 2020: Spending for Better Results.*

[61] DT Global, "Indonesia: Establishment of Integrated Road Asset Management Systems," accessed October 4, 2021, https://dt-global.com/projects/irams-dc

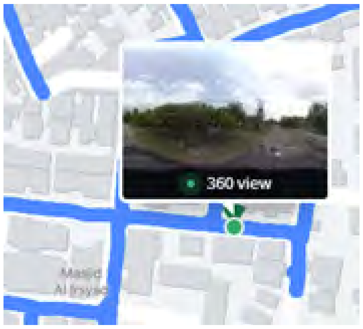
3.2 Methodology

The ML-based framework implemented in these case studies was developed to provide a quick screen to evaluate road safety. The framework ascertains road characteristics traditionally collected or annotated to provide a road safety prediction. ML models were developed specifically for this framework during these case studies, one to extract road characteristics from street view images and one to determine road risk based on the derived road characteristics. To do so, first, the models needed to be trained to extract road characteristics and determine the road risk based on crash data. Then the models could be applied to make predictions in new areas without crash data. Therefore, there were two phases in this framework, first the **training phase** to train the models (figure 12), and then the **deployment phase** to make new predictions with the models (figure 13). In each phase there were three steps, both of which began with data collection and preparation. OpenStreetMap (OSM), Waze, and Mapillary were used to develop this framework (additional examples of these datasets and related analysis can be found in Annex 3).



The **OSM** road network provided the foundation for analysis. It is freely available and scalable. OSM uses lines to represent roads and points to represent links among the roads. In OSM, the geometric road lines are split into road segments (called ways) that are connected by the points (called nodes). No modifications were made to the OSM geometry to maintain its synchronicity with other big datasets referencing OSM ways and nodes.

The **Waze** crash data consists of coordinates representing the location where users of the Waze application are when they see and report a crash.[62] The Waze crash points were joined to the nearest OSM road segment (within 20 meters). For each road segment, the crash frequency, or crash per meter, was calculated to normalize the frequency of crashes. Since OSM road segments vary in length and there could be multiple reports per crash, calculating the crash frequency provided crash trends. To identify road segments with more frequent crashes per meter, the crash frequency was split into high and low risk.



Mapillary was used to obtain street view images, which were primarily collected by the World Bank's Global Program for Resilient Housing. Since many images are captured along a street, and many images can be linked to a single road segment, the image closest to the centroid of the road segment was selected. The radius for this selection was within three meters of the centroid. This approach standardizes the image selection and classification: one image represents the scene of one road segment. For each OSM road segment, a street view image taken near the centroid of the segment was downloaded using Mapillary API v4.

SOURCE: Original examples for this publication based on data from OSM, Waze, and Mapillary. Copyright OpenStreetMap contributors, Microsoft, Esri Community Maps contributors. Basemap from Esri, HERE, Garmin, METI/NASA, USGS.

[62] Data provided by Waze App. Learn more at waze.com.

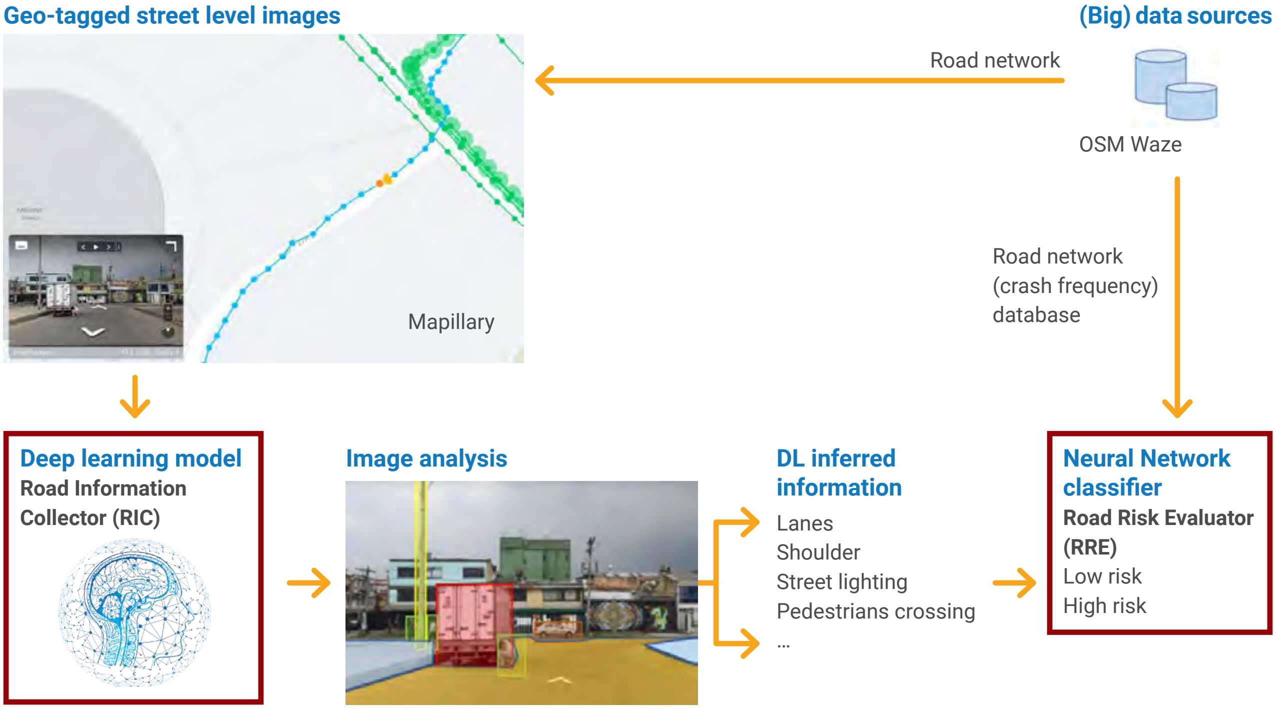
The Training Phase

The training phase consisted of two significant steps that were powered by ML to extract information from street view images and to make predictions on risk level based on extracted data. Each step had an ML model at its core that needed to be trained based on data. Therefore, there were three steps in the training phase.

Step 1. Select the region of interest and prepare data

A generalized polygon of the region of interest was used to collect data from OSM, Waze, and Mapillary. The road network database was prepared, and the street view images closest to the centroid of the road segment were downloaded as inputs for the models.

FIGURE 12: **Training phase for road safety segment analysis using ML**



SOURCE: Original figure for this publication.

Step 2. Develop ML model for identifying road characteristics

The first custom ML model developed for this case study was the Road Information Collector (RIC), shown in figure 12. It is a deep convolutional neural network, Mask R-CNN, which can classify and count objects detected in images.[63] The RIC model was trained with images from the updated Mapillary Vistas Dataset (initially released in 2017), which provides detailed characteristics for types of road markings and barriers, traffic lights and signs, and vulnerable road users such as pedestrians, motorcyclists, and bicyclists.[64] Other identifiable characteristics include flat terrain, which characterizes road gradient, and the presence of potholes, which could indicate paved, urban road quality. The RIC takes street view images as the input and can detect more than 100 classes of objects as the output (for a complete list of the features the RIC model detects, refer to Annex 4). The model can

[63] Kaiming He et al., “Mask R-CNN,” 2017 *IEEE International Conference on Computer Vision* (2017): 2980-2988.

[64] G. Neuhold et al., “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes,” *2017 IEEE International Conference on Computer Vision (ICCV)* (2017): 5000-5009, doi: 10.1109/ICCV.2017.534

detect and classify some road features better than others (for the precision score in detecting and classifying the objects, see Annex 5).

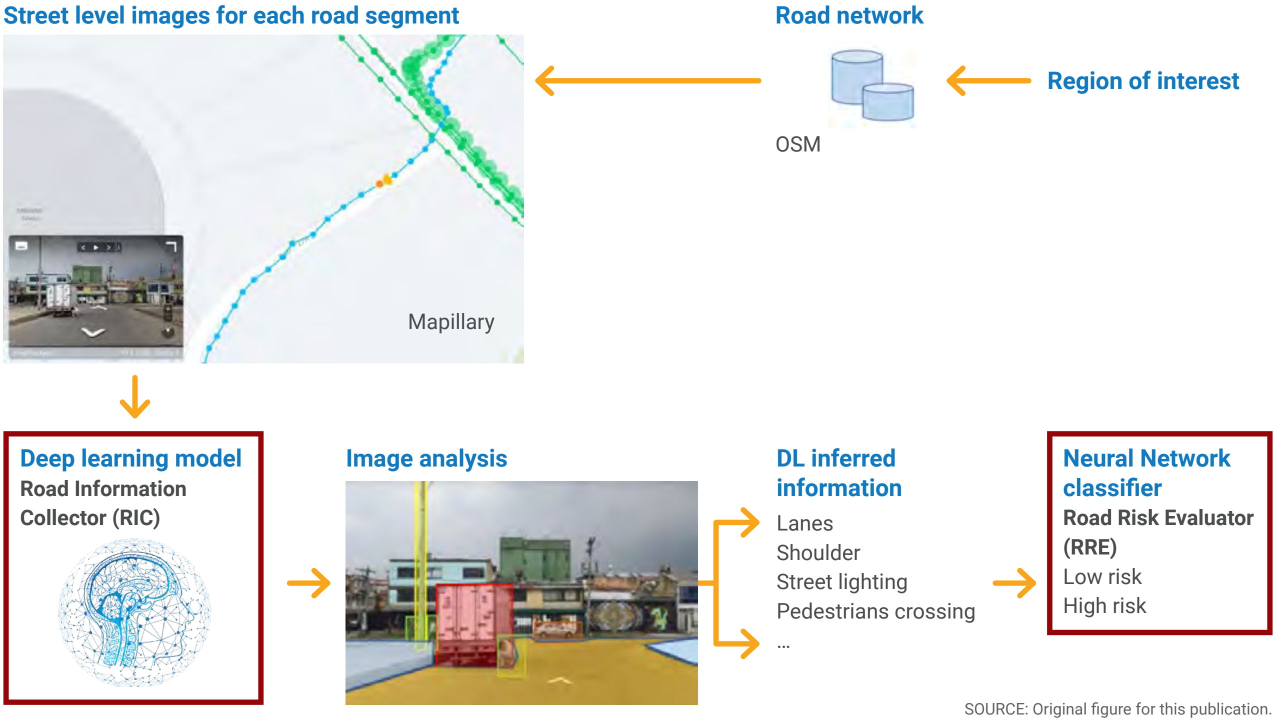
Step 3. Develop ML model for evaluating road risk

The second ML model developed was the Road Risk Evaluator (RRE). The RRE is a neural network classifier with two hidden layers; each has 50 neurons. The RRE was trained using paired data for each road segment, the road attributes from the RIC and the assigned road risk from the road network database. Similar work was conducted by a team using a neural network to predict the crash frequency of road segments.[65]

The Deployment Phase

Once the two ML models are trained, they can be added to an automated workflow in the deployment phase. This means the trained ML models can now predict the risk level for any road segment with the required input data – a street view image. Crash data is not required in the deployment phase.

FIGURE 13: **Deployment phase to predict road safety**



SOURCE: Original figure for this publication.

The deployment phase uses three steps to predict risk within an automated workflow (figure 13).

Step 1. Select the region of interest and download data

For the selected region of interest, the code will download the road network from OSM and calculate the centroid of each road segment. The code will then download from Mapillary API a street view image taken near the centroid of the road segment.

[65] Qiang Zeng et al., "Rule Extraction from an Optimized Neural Network for Traffic Crash Frequency Modeling," *Accident Analysis & Prevention* 97 (2016): 87-95.

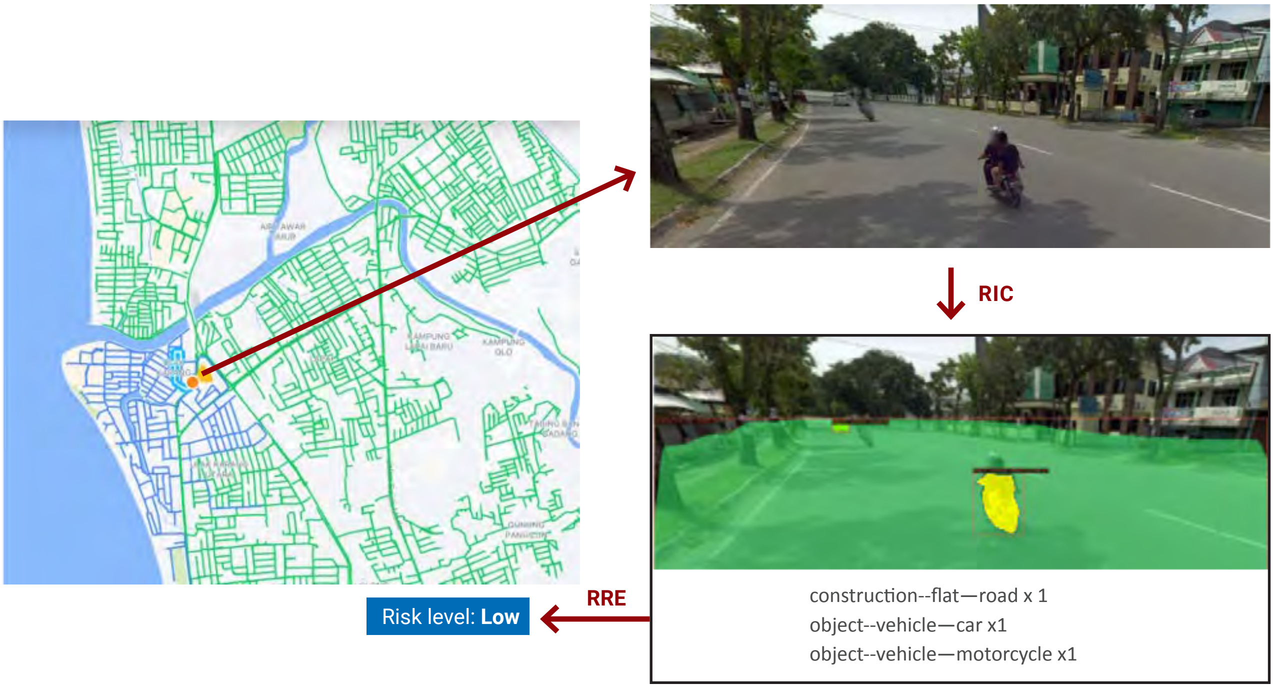
Step 2. Identify road characteristics

For each road segment, the downloaded image will be fed into the RIC to extract road characteristics. For each image, the RIC will output the numbers of detected objects for each class (refer to Annex 4 for classes). These numbers are put together to form a vector for each image.

Step 3. Evaluate road risk

Each vector produced by the RIC will be fed into the RRE to calculate the risk level: high or low. To illustrate the automated workflow of the deployment phase, figure 14 shows the risk prediction for a road segment. The RIC detected a flat road, car, and motorcycle; therefore, the RRE predicted the road segment as low risk. This framework requires no historical crash data to identify high- or low-risk roads.

FIGURE 14: **RIC and RRE applied to predict road segment risk**



SOURCE: Original figure for this publication, based on data from Mapillary and annotated with classifications from the model.

The two case studies presented illustrate the **training** and **deployment** phases.

The training phase was conducted in Bogotá, where data was collected to train the ML model RRE, while the RIC model was trained on the Mapillary Vista Dataset. Then the models were applied in the deployment phase to predict the risk level for each road segment in Bogotá, Colombia.

The second case study was in Padang, Indonesia. The RIC and RRE models trained in the previous case study were applied directly (i.e., without re-training) in a deployment phase to predict road risk in Padang. This demonstrates that, ideally, there is no need to re-run the training phase for future applications since the RIC and RRE are already trained.

3.3 Case Study 1: Bogotá, Colombia

The Training Phase

Step 1. Select the region of interest and prepare data

In Bogotá, a road network database was created to prepare training data for the ML models. First, a generalized polygon of the region was used to retrieve roads from OSM and six months of crash reports from Waze (July–December 2020). The crashes were joined to the nearest OSM road segment within 20 meters. The crash frequency, or crash per meter, was calculated and road segments were divided into high risk (crash frequency >0.5) and low risk (crash frequency <=0.5) in the road network database. This means a crash per meter of 1 represents one crash per meter in the six months of the Waze data collected. Street view imagery was downloaded using the Mapillary API to collect images close to the centroid of each road segment. Table 18 provides an overview of the data sources for this case study.

TABLE 18: Data used for case study in Bogotá, Colombia

DATA SOURCES	ATTRIBUTES	REMARKS
ROAD NETWORK		
OSM	**Road network** (road segment length)	Provided through an open license.
CRASHES		
Waze	**Road alerts** (crashes reported by users, coordinates)	Obtained through DDP.
ROAD CHARACTERISTICS		
Mapillary (images and tags)	**Street view image detections** (crosswalk, curb, guard rail, human, marking, pothole, sidewalk, sign, streetlight, traffic sign, utility pole)	Selection of image annotation tags used in study; more available through Mapillary Traffic Sign and Vistas. Multiple detections per image are possible.

SOURCE: Original table for this publication.

Step 2. Develop ML model for identifying road characteristics

The RIC was developed and trained to perform instance segmentation. It is a deep convolutional neural network that identified the classes, or objects in the image, and provided the count of these classifications. The model was trained using the Mapillary Vistas Dataset using a total of 124 classes (Annex 4).[66] The resulting output is a count of the classes identified by the bounding boxes, shown in figure 15, which is represented through a series of integers.

Training data: Mapillary Vistas Dataset (124 classes)

Input: Street view image near the centroid of a road segment

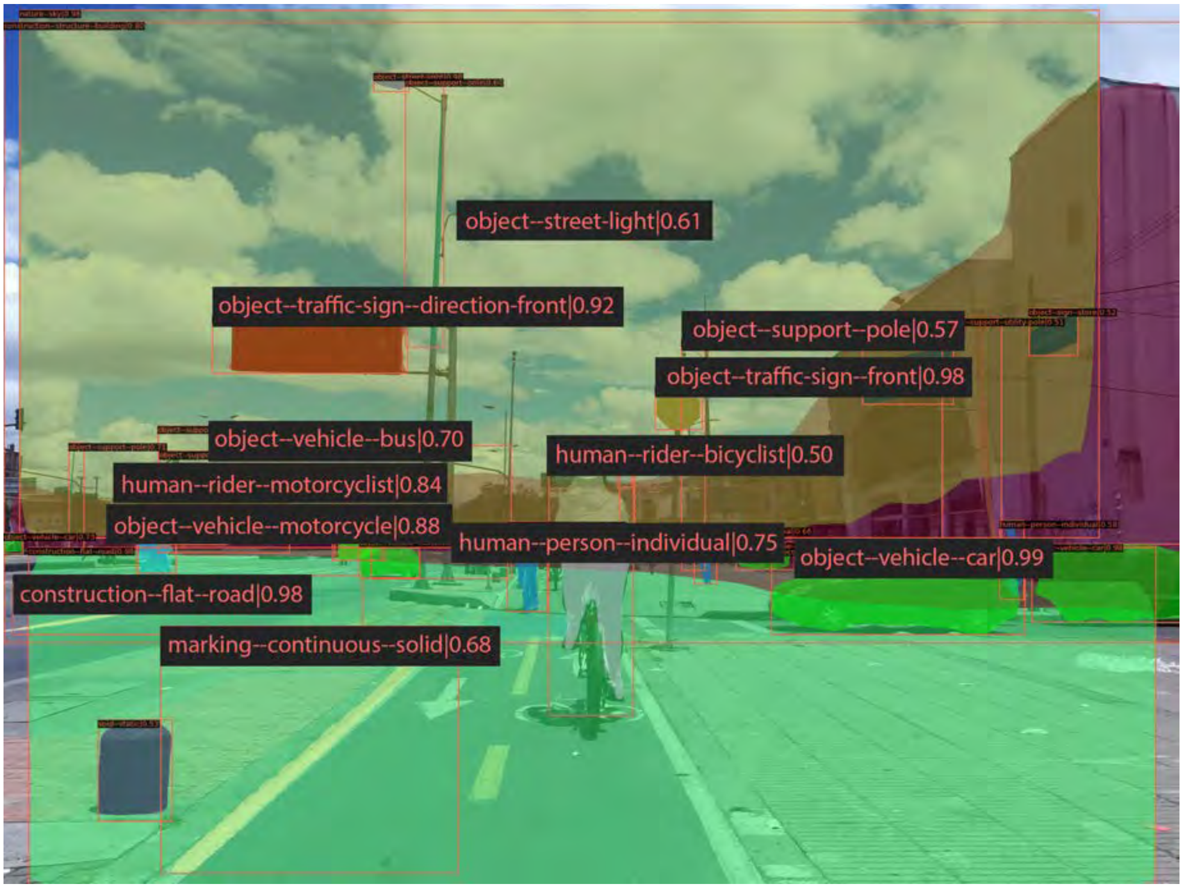
Output: A vector of integers (each element represents the count of detected objects that belong to a class)

Figure 15 depicts the RIC in action on an image from Bogotá. The bounding boxes surrounding each object in the image indicate classes the model identified. Confidence levels are provided next to the name of the object segmented by the bounding box. The closer the confidence level is to 1, the higher the confidence in the prediction. Looking at the center of the image, the bicyclist was identified with 0.5 confidence, and other vulnerable road users were recognized, such as a motorcyclist (0.84) and pedestrian (0.75). Vehicles were segmented with high confidence for the bus (0.7), motorcycle (0.88),

[66] G. Neuhold et al., "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes."

and car (0.99). The RIC segmented traffic signs, support and utility poles, flat road, and road markings as well.

FIGURE 15: **Image segmentation in Bogotá**


object--street-light|0.61
object--traffic-sign--direction-front|0.92
object--support--pole|0.57
object--traffic-sign--front|0.98
object--vehicle--bus|0.70
human--rider--bicyclist|0.50
human--rider--motorcyclist|0.84
object--vehicle--motorcycle|0.88
human--person--individual|0.75
object--vehicle--car|0.99
construction--flat--road|0.98
marking--continuous--solid|0.68

SOURCE: Original figure for this publication, based on data from Mapillary.

The sample image shows favorable results for image segmentation. The performance of the RIC model in terms of the average precision of the bounding box detection and classification for each class is provided in Annex 5. In the next step, road attribute data extracted through the RIC were inputs for the prediction model to link the road characteristics with the likelihood of a crash in the road networks examined.

Step 3. Develop the ML model RRE for evaluating road risk

To develop the RRE, six study areas in Bogotá, Colombia were selected to reduce computational load. These study areas were drawn to include a wide variety of neighborhoods (poor, rich) and placed throughout the city. They also contain high and low crash frequency road segments and comprehensive street view image coverage. Figure 16 shows the six study areas along with the crash risk from the road network database, high risk (crash frequency >0.5) and low risk (crash frequency <=0.5).

The low- and high-risk road segments in these areas were the training data for the model. Based on the segment risk derived from the road network database and the characteristics for each road segment derived from the RIC, the model was trained to evaluate a road segment as high or low risk.

Training data: The following input-output pairs obtained from road segments in six study areas in Bogotá, Colombia.

Input: A vector of integers, which is the output of RIC*

Output: 0 (low risk) or 1 (high risk)

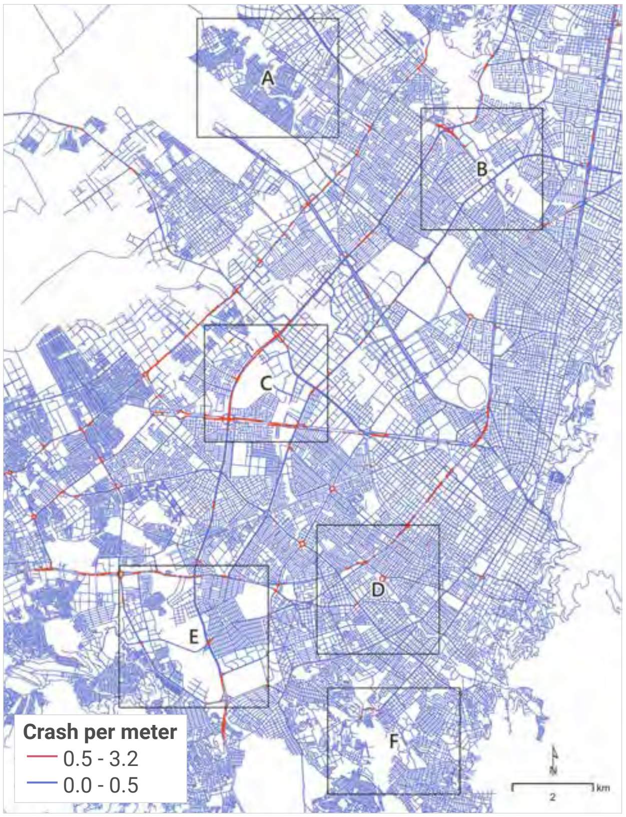
*Only 106 out of 124 classes are used as the input to RRE. A total of 18 classes irrelevant to road characteristics, such as sky, bird, etc., were removed from the vector before entering into the RRE.

In searching for an optimal architecture of the neural network, the number of layers and neurons were tested for the best performance. Testing showed that more layers or neurons do not significantly improve the performance on this dataset. The RRE was used to evaluate whether a road segment was low or high risk based on a street view image.

Overall performance of the ML

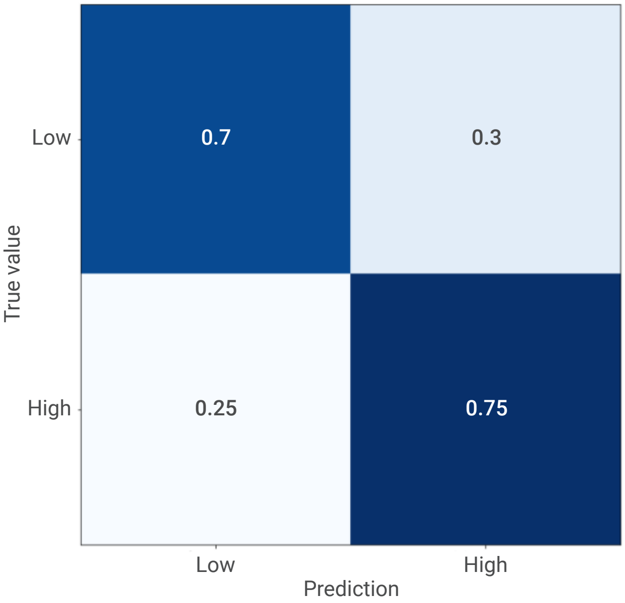
Predictions of low-risk road segments were correct 70 percent of the time, and predictions of high-risk road segments were correct 75 percent of the time (figure 17). The mean accuracy and F1-score were both 72.5 percent. The closer the accuracy and F1-score are to 100 percent, the better the performance of the model. In the case of this model, a random guess of a binary classification is 50 percent, which makes these results promising. These results suggest the model would perform well in similar contexts as Bogotá. If needed, there would be potential to fine-tune the model for increased accuracy and precision in other areas.

FIGURE 16: **Six study areas and crash frequency in Bogotá**



SOURCE: Original figure for this publication, based on data from OSM and data provided by the Waze App. Learn more at waze.com.

FIGURE 17: **Confusion matrix showing the accuracy of the RRE model**



SOURCE: Original figure for this publication.

TIPS FOR INTERPRETING ML PERFORMANCE

The performance of an ML model can be evaluated using accuracy, precision, recall, and the F1-score. These are derived by counting the correct predictions (true positives and true negatives) and incorrect predictions (false positives and false negatives).

accuracy = correct predictions / all predictions

precision = true positives / (true positives + false positives)

recall = true positives / (true positives + false negatives)

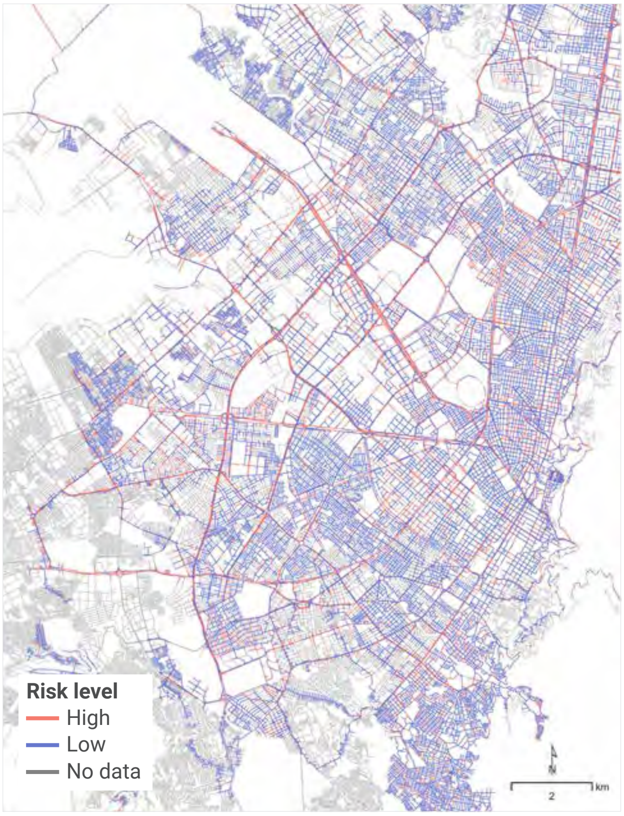
F1-score = 2*((precision * recall) / (precision + recall))

A confusion matrix shows how well the model performed in predicting road risk through a comparative chart of the true positives, true negatives, false positives, and false negatives.

Bogotá Results

Following the three-step workflow of the deployment phase described in section 3.2, road risk was predicted for the entire road network in Bogotá. In total, 98,488 images were processed to make the predictions shown in figure 18. Road segments without an image within 3 meters were not predicted. Overall, high crash frequency from Waze and high-risk predictions exhibited similarity along some segments, particularly on arterial roads; however, the model tended to moderately overpredict high risk.

FIGURE 18: **Road risk prediction in Bogotá**



SOURCE: Original figure for this publication, based on data from Mapillary, OSM and Waze.

3.4 Case Study 2: Padang, Indonesia

The Deployment Phase

The model that was built in Bogotá was applied in Padang. Similar to Bogotá, the road network was accessed through OSM, and street view images were downloaded using the Mapillary API. Waze crash data was joined to the OSM road network to compare with risk predictions. Padang had limited geospatial crash data to validate the model. Table 19 provides a description of the datasets.

TABLE 19: Data used for case study in Padang, Indonesia

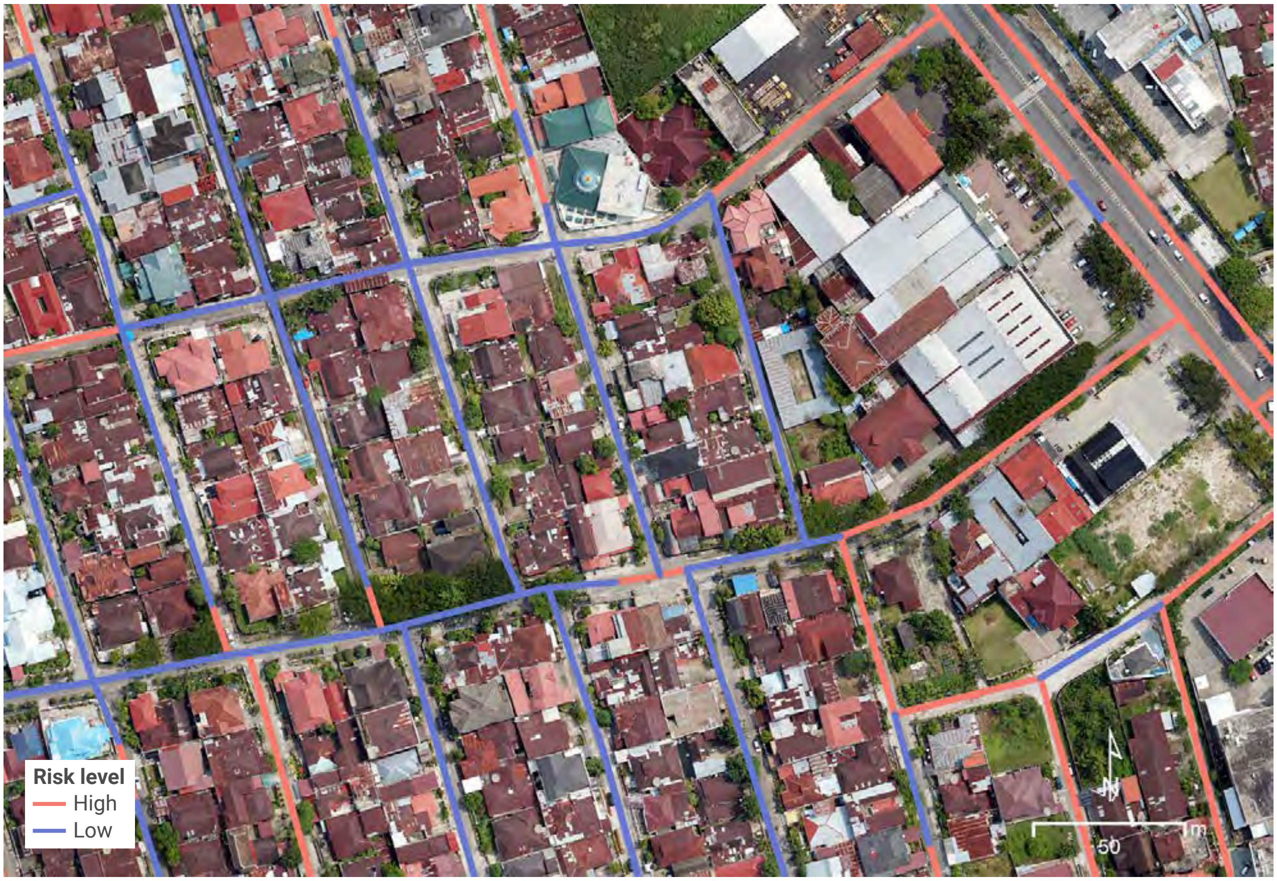
DATA SOURCES	ATTRIBUTES	REMARKS
ROAD NETWORK		
OSM	**Road network** (road segment length)	Provided through an open license.
CRASHES		
Waze	**Road alerts** (crashes reported by users, coordinates)	Obtained through DDP.
ROAD CHARACTERISTICS		
Mapillary (images and tags)	**Street view image detections** (crosswalk, curb, guard rail, human, marking, pothole, sidewalk, sign, streetlight, traffic sign, utility pole)	Selection of image annotation tags in study; more available through Mapillary Traffic Sign and Vistas. Multiple detections per image are possible.

SOURCE: Original table for this publication.

Padang Results

In Padang, preliminary results pointed to the framework's potential in scanning roads for safety. Figure 19 shows predictions where arterial road segments were predominately designated as high risk (red lines). Residential areas were interspersed with low- and high-risk road segments. Similar patterns of road segments predicted as high risk along arterial roads and a mix of low and high risk along residential and tertiary road segments were largely found.

FIGURE 19: **Road risk prediction in Padang**



SOURCE: Original figure for this publication, based on data from OSM and data provided by the Waze App. Learn more at waze.com. Drone imagery provided by the World Bank Global Program for Resilient Housing.

In general, where there were crashes reported by Waze, high-risk road segments were predicted. These preliminary results were encouraging; however, verifying the results was difficult because there was not sufficient data. While the deployment of the framework in Padang requires further validation with more data, ML-based approaches such as this are promising to offer initial road safety scans.

3.5 Findings

The Integrated Framework for Road Risk Prediction demonstrates the strength of ML to identify road segment safety with substantial accuracy (72.5 percent) in Bogotá. Preliminary results in Padang support replicating the framework with further validation in other areas. Using advanced ML techniques, the framework applied a streamlined approach that relied on road characteristics and crash frequency to determine crash risk in the training phase. Then the ML models applied in the deployment phase could predict road risk based on road characteristics without historical crash data.

The alternative data sources used to train the models were robust – thousands of annotations, high-resolution images, and crash data joined to extensive road networks – and of suitable quality for the models to provide a road safety scan. To identify road characteristics, the RIC was trained using the Mapillary Vistas Dataset, which has a breadth and depth of annotations from different contexts, providing geographic diversity. The RRE was trained using a pairing of the road characteristics and a road network database created from OSM road segments and Waze crash data. OSM road segments offered global scalability and were sufficient for a coarse assessment in these case studies. Waze data availability was dependent on the area (and the users of the app). Given the potential for duplicate crash reports, Waze data was not relied on for accurate crash data in Bogotá; instead, it was used to identify crash patterns of high- and low-risk road segments.

The framework is not suitable for detailed road assessments. However, it can be applied to screen roads for safety without historical crash data if the RIC model is enhanced with more training data and calibrated for the local street view context; the RRE model can be modified and enhanced with fine-grained training data. It is replicable in other areas with the following recommendations, which are applicable for developing other ML-based frameworks for road safety.

Incorporate training data to fine-tune the model for a specific location. Typically, ML models trained on data collected from one region do not work well for a new region. This is called domain shift: the testing data has a different distribution than the training data. In this case, including data collected from the new region in the training phase will usually help. It is important to evaluate the data and consider any influences the collection method may have on the potential to introduce bias into the project. For example, if local crash data is introduced to train the RRE, it would help validate and potentially improve the model's application in the location of interest. Both RIC and RRE can be continually trained with newly obtained data so that the knowledge learned from previous data can be carried on for new regions while the model is still applicable to the previous regions.

It is essential to ensure that models are based on sufficient, high-quality training data. In general, at least a few thousand annotations are recommended to identify objects from images with simple context, depending on the characteristics of the object. Whether the street view images are obtained through big data platforms such as Mapillary or collected by the team, street view imagery covering different geographical regions makes the trained object detection model, like the RIC, more robust. Since street level images capture the visual scene (road characteristics and road users) at a single point in time, it is important to consider these implications when using a snapshot of that time of day, day of week, and season. Relatedly, a road characteristic may be covered or occluded in a street view image; for instance, when a passing truck blocks a sign. Imagery collected at a frequent distance, such as every two meters, permits greater flexibility to analyze the road scene and predict risk using the RIC and RRE. OSM road networks require review for recency and accuracy, and possibly editing to ensure suitable quality and coverage in other areas. If high-quality, granular crash data shows a clear pattern of more risk classes, three classes could be predicted: for example, high, medium, and low risk.

Conclusion

Big data and ML offer promising opportunities to improve current road safety assessment procedures for sustainable development. Road safety assessments are often required for new transport and infrastructure developments to be approved or as part of their monitoring and evaluation once they are completed. However, conducting road safety assessment procedures can be expensive and time-consuming. Alternative data sources and ML can optimize this process by identifying patterns using complex predictive models. The Integrated Framework for Road Safety offers one approach using street view imagery that can be accessed through Mapillary or collected by the team to provide a road safety scan. With further training, this framework has the potential to provide detailed road safety assessments, mitigating the need for manual annotations (or years of historical crash data). In addition to the pilots and studies conducted by the researchers and representatives of road safety organizations interviewed for this note, there are many ML models contributing to road safety efforts, which typically outperform statistical models in predicting road safety.[67]

Integrate alternative data sources and ML into road safety assessments with care. Finding valid, representative data can be a significant challenge in evaluating risks and reducing crash fatalities and injuries through data-driven, evidence-based interventions. Teams can directly partner with private companies and data providers to retrieve alternative sources of data. And data sharing platforms, such as DDP, offer streamlined solutions. However, commercial data sources are not typically established to collect data for road safety analysis, and their data may be inadequate for road safety assessment methods and procedures. Data can be biased, incomplete, and challenging to synchronize with conventional analytical tools. The implications of collecting and analyzing big data using ML require thorough consideration. Data privacy and security are central concerns; data needs to be de-identified and anonymized and stored according to institutional guidelines.[68] Data and models need to be screened for biases that can affect their outcomes. For example, imbalanced access to smartphones or social media may amplify gender or community bias.[69] Teams can adhere to best practices and data policies and make their ML models and results transparent and openly shared. Resources such as “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle” and “The Ethics of Artificial Intelligence” may be helpful for teams implementing ML in their projects.[70]

The approach used for the case studies in this note can be extended to evaluate specific measures of road safety. For example, while the framework uses the crash frequency and may identify the number of relevant road users in a street view image, it does not thoroughly consider the number of (vulnerable) road users nor does it consider the probability of a crash causing fatalities or serious injuries. The approach could

[67] Philippe Silva, Michelle Andrade, and Sara Ferreira, “Machine Learning Applied to Road Safety Modeling: A Systematic Literature Review,” *Journal of Traffic and Transportation Engineering* 7, no. 6 (2020): 775-790, https://doi.org/10.1016/j.jtte.2020.07.004

[68] World Bank, *World Development Report 2021: Data for Better Lives* (Washington, DC: World Bank, 2021). doi:10.1596/978-1-4648-1600-0

[69] World Bank, *Use of AI Technology to Support Data Collection for Project Preparation and Implementation: A ‘Learning-by-doing’ Process* (Washington, DC: World Bank, 2021).

[70] Harini Suresh and John Guttag, “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle” in *Proceedings of Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO ‘21),* https://doi.org/10.1145/3465416.3483305; Nick Bostrom and Eliezer Yudkowsky, “The Ethics of Artificial Intelligence,” in *The Cambridge Handbook of Artificial Intelligence*, ed. Keith Frankish and William M. Ramsey (Cambridge: Cambridge University Press, 2014): 316-334.

also be extended using complementary data such as road geometry, traffic flow, traffic volume, traffic speed, weather, season, and other factors affecting visibility along the road or road surface conditions. The case studies illustrate the potential of big data and ML to reduce the manual inspection of roadways and provide road safety insight where otherwise the information is in short supply, thereby contributing to safer roads.

For big data to be fully leveraged for road safety analysis, governments, road safety advocates, and international development organizations will want to consider investing in platforms and tools that specialize in collecting and analyzing data for road safety. Ongoing efforts to establish regional road safety data observatories provide an opportunity to gather data providers and create a data marketplace specifically for road safety analysis, especially where alternative or traditional sources are scarce. Government regulations and initiatives to encourage private companies to share data could further integrate big data in international development projects, including road safety. It is essential for key stakeholders in road safety assessment to collaborate closely with pioneers of these technologies to realize their potential in road safety analysis.[71] Initiatives such as the Artificial Intelligence in Road Traffic Crash Prevention Roundtable hosted by the International Transport Forum (ITF) in early 2021 is an example of one such opportunity. Conversations with World Bank team leaders and transport specialists reveal that developing a tool to provide a single, easy-to-use solution to access and utilize big data for road safety analysis is in high demand. There is potential to automate some of the processing and analysis for which specialist expertise is currently required, and initiatives such as Ai-RAP and the World Bank Simplified Methodology suggest that practical, scalable solutions could be a reality soon.[72] As big data and ML become more accessible, and as their adoption accelerates worldwide, road safety practitioners, governments, road safety advocates, and international organizations can unlock their immense potential to improve the quality and efficiency of road safety assessments.

[71] Subasish Das and Greg P. Griffin, "Investigating the Role of Big Data in Transportation Safety," *Transportation Research Record* 2674, no. 6 (2020): 244–52, https://doi.org/10.1177/0361198120918565

[72] Monica Olyslagers (Safe Cities and Innovation Specialist, iRAP) and Satoshi Ogita (Senior Transport Specialist, World Bank), in discussion with the authors, April 2021.

ANNEX 1:

Most Relevant Big Data Types for Road Safety Analysis

DATA COLLECTION	POTENTIAL SOURCES	POTENTIAL APPLICATIONS	ADVANTAGES	LIMITATIONS
Street view imagery	• Apple Look Around • Google Street View • KartaView • Mapillary • Collected by team	Identify road attributes for road safety assessments.	• Provides objective evidence of conditions in the field. • Can be used in regions where government data is not available.	• Coverage is incomplete, particularly in rural and low-income areas. • Licensing restrictions for ML application.
Mobile applications and telematics	• Mobile application data • Telematic companies • Rideshare companies	Identify vehicle movement, traffic flows and road use by various types of users for crash risk. identification and road safety assessments.	• App data is usually low cost and current. • Telematic data could show risky driving behavior.	• Coverage is lighter in rural areas or cities where use of app is low. • Often requires data sharing agreements with private companies.
Crowdsourced	• Waze • Delivery drivers • OSM • Social media	Obtain crash data and information related to road use, such as types of road users and their relative density at a specific location. Can help to identify road risks.	• Can supplement government data, particularly if incidents are underreported or government provided road networks are unavailable.	• Requires app use in the region of interest. • Needs coordination and resources to collect reports from delivery drivers. • Data quality may be low. • Social desirability bias can occur, where users feel inclined to share specific types of information to reinforce a positive or negative perspective.
Government	• Government transport agencies • Road safety observatories	Most frequently used to obtain crash data, including statistics related to crash severity, crash frequency as well as fatalities and injuries statistics.	• Data often has many attributes or details that have been manually added. • Data often has been collected for many years in the same manner, allowing for temporal analysis.	• Data can be messy (human error). • Data often not shared.
Aerial and satellite imagery	• Earth observation agencies • Private companies	Identify road attributes for road safety assessments.	• Covers large geographic area.	• Requires balancing the cost with recency and granularity of imagery.
Meteorological sensors	• Meteorological agencies • Local universities • Private companies	Review weather conditions that may affect road safety, such as crashes.	• Infer driving conditions (i.e., if road surface conditions are not available in government crash data).	• There are varying levels of granularity.

SOURCE: Original table for this publication.

ANNEX 2:
Overview of Big Data Sources

Data sources accessible through DDP are indicated as free for World Bank task teams.

DATA	ACCESS	ATTRIBUTES	RESOLUTION AND FORMAT	COST	COMMENTS
STREET VIEW IMAGERY					
Apple Look Around	Early stages; contact company	Requires processing to derive physical features related to road safety, such as: crosswalks, speedbumps, painted lines, roads, road shoulders, sidewalks, streetlights, traffic signs and others specific to region of interest.	Image	N/A	Offers extremely limited geographic coverage.
Google Street view	Not accessible according to license		360 photos must be at least 4K (image)	N/A	Global coverage is fairly extensive.
KartaView	Open license		Depends on camera (image)	Free	Images are free, though image processing is required (see street view training data); global coverage is variable.
Mapillary	DDP		Depends on camera (image)	Free	Images are free, though image processing is required (see street view training data); global coverage is variable.
Collected by team	Requires permission and coordination with local government		Depends on camera (image or video)	High	Collection every two meters recommended for images. Images or video require processing; see street view training data.
STREET VIEW TRAINING DATA					
Mapillary Traffic Sign	DDP	**Traffic signs**	Resolution can be very high or very low. The model performs best on images with the same resolution level of the training dataset. (image)	Free	More than 300 traffic sign classes covering six continents.
Mapillary Vistas	DDP	**Physical features related to road** crosswalks, speedbumps, painted lines, roads, road shoulders, sidewalks, streetlights, traffic signs (others possible)		Free	Coverage spans six continents.
Annotation by team	Hire a team	**Physical features related to road, specific to region of interest** crosswalks, speedbumps, painted lines, roads, road shoulders, sidewalks, streetlights, traffic signs (others possible)		High	Consider collaborating with stakeholders in a region of interest to label images using a Computer Vision Annotation Tool (CVAT) or a labeling team with training. 2,000 labels per class is recommended for a simple classification.
World Bank's GRSF Road Risk Assessment software* * The software is included in this section as video training data is limited in World Bank countries. Contact Satoshi Ogita (World Bank), for access.	Open source	**Physical features related to road** road grade and curvature, pedestrian crossings, delineation, roadside severity, lane width, and number of lanes		Free	Video analysis produces a richer dataset. Piloted in Liberia and Mozambique.

DATA	ACCESS	ATTRIBUTES	RESOLUTION AND FORMAT	COST	COMMENTS
MOBILE APPLICATIONS AND TELEMATICS					
Grab	Contact company	Contact company	N/A	N/A	Coverage offered in Cambodia, Indonesia, Malaysia, Myanmar, Philippines, Singapore, Thailand, Vietnam.
HERE	Not accessible according to standard license	**Traffic** current and historical speeds, jams, crashes, road closures and road construction	Every minute (text, number)	N/A	Detailed road network coverage in more than 200 countries and comprehensive traffic speeds in more than 80 countries.
Mapbox Movement	DDP	**Movement activity index;** driving activity index available in select locations	Aggregated daily or monthly at 100 meter resolution (text, number)	Free	See Mapbox Movement data processing guidelines for recommendations and considerations when using this dataset.
Mapbox Traffic	DDP	**Traffic** (typical speed) each road segment, identified by a start and end node, has 2,016 typical speed predictions (7 days × 24 hours × 12 five-minute periods)	Typical speed per road segment in five-minute increments over a week (text, number)	Free	
Moovit	Contact company	**Urban transit** (public and on-demand)	Contact company	N/A	
Ola Cabs	DDP	**Travel time and potholes**	Contact DDP	Free	Coverage provided in India.
Orbital Insight	DDP	**Foot traffic** time of day, day of week, velocity (stationary, walking), dwell time	Each minute; 2019 to present (text, number)	Free	Foot traffic using mobile location data in region of interest, subject to data availability per country.
TomTom	Contact company	**Traffic** current and historical speeds, jams, crashes, road closures and road construction	Every minute per road segment (text, number)	Free to Medium	Global coverage is variable.
Uber Movement	Contact company	**Traffic** travel times between zones, average speed per segment and traffic density	Average travel time, average speeds per hour, time of day or quarter of year (text, number)	Free	Geographic coverage is limited to a selection of major cities. Currently no API. Was previously part of DDP.
Unacast	DDP	**Human movement**	Coordinates, horizontal accuracy, timestamp, time zone (text, number)	Free	Mobile Location Data Inventory for geographic coverage available through the DDP website.
Veraset	DDP	**Human movement**	Coordinates, horizontal accuracy, timestamp (text, number)	Free	
Waze	DDP	**Traffic** (alerts, jams, irregularities) major and minor crashes; severity of congestion or irregularities; current and typical speed on jammed segments; coordinates, road segment (start and end node), street name; road type; driving direction (NSEW); turn type; alerts (construction, road closure and weather)	Every minute; location provided as coordinates, road segment, street name (text, number)	Free	Includes weather alerts and major and minor crashes by application users; see Waze under Crowdsourced section.

DATA	ACCESS	ATTRIBUTES	RESOLUTION AND FORMAT	COST	COMMENTS
WhereIsMy Transport	DDP	**Informal transit network**	Determined in collaboration with team	Medium to High	Specializes in producing informal transit data according to General Transit Feed Specifications (GTFS). Supports team in collecting and processing data in exchange for the team covering in-field costs of data collection and facilitating engagement with local transport authorities.
CROWDSOURCED					
OSM	Open license	**Road segments** road type, length, and features	Centerline of road segments, referred to as ways and relations (text, number)	Free	May include additional road attributes: lanes, name, smoothness, surface, speed limit, and width, and other information such as overtaking permitted or lighting.
Twitter	DDP	**Road incidents** tweeted	User-dependent; can be associated with a place or location (text, number)	Free	
Waze	DDP	**Road incidents** reported using app	Every minute; location provided as coordinates, road segment, street name (text, number)	Free	
Delivery drivers	Coordinated by team	**Road incidents** reported using app	Depends on collection (text, number)	High	
GOVERNMENT					
Government or road safety observatory	Government contact or open data platform	**Incidents** (date, time, severity, type)	XY coordinate per incident (text, number)	Free to Low	Processing requires standard GIS software such as ArcGIS or QGIS (free). Storage is small, typically <1GB per urban area over multiple years.
		Road segments (type, width, speed limit)	Road segments (text, number	Low	
		Traffic lights (intersection type)	XY coordinate per traffic light (text, number)	Low	May include intersection type (pedestrian, bicyclist, for example).
SATELLITE AND AERIAL IMAGERY (AND OTHER REMOTE SENSING)					
Maxar Technologies	Contact company	**Elevation and roads**	Less than 1m (image)	High	Requires processing to derive road networks. Was previously part of DDP.
Orbital Insight	DDP	**Car and truck count; roads**	Car and truck count: high resolution, 2013 to present; roads: medium resolution, 2016 to present (image, number)	Free	Car and truck count derived from satellite imagery. Limited Geospatial Intelligence Platform credits to derive roads in region of interest; not for routable road networks; not suitable for narrow roads in urban areas or dirt or mountainous roads in rural areas.

DATA	ACCESS	ATTRIBUTES	RESOLUTION AND FORMAT	COST	COMMENTS
Security or traffic cameras	Collected by team or through external resource	**Traffic density and volume**	Depends on camera (image or video)	Medium to High	
Unmanned aerial vehicle (UAV)	Collected by team	**Elevation, roads, traffic density and volume**	Depends on camera (image or video)	Medium to High	Recent research suggests traffic density and volume are possible to calculate.
METEOROLOGICAL SENSORS					
OpenWeather	Contact company	**Weather** (weather type, temperature, wind speed and direction, cloud coverage; rain and snow volume by hour and per 3 hours)	40-year historical archive for any coordinates by the hour; or by city or 1 km, 5 km, 10 km or customized grid (text, number)	Low	Price is economical for the 40-year history of a single coordinate or city. Contact provider for details on pricing and to download many locations.
Tomorrow.io	DDP	**Weather** (weather type, temperature and humidity; wind speed, direction, gust; precipitation type, intensity; snow and ice accumulation; visibility, moon phase)	500m radius with precipitation recordings as low as 30 feet off the ground; time steps range from one day to one minute (text, number)	Free	

SOURCE: Original table for this publication.

ANNEX 3:

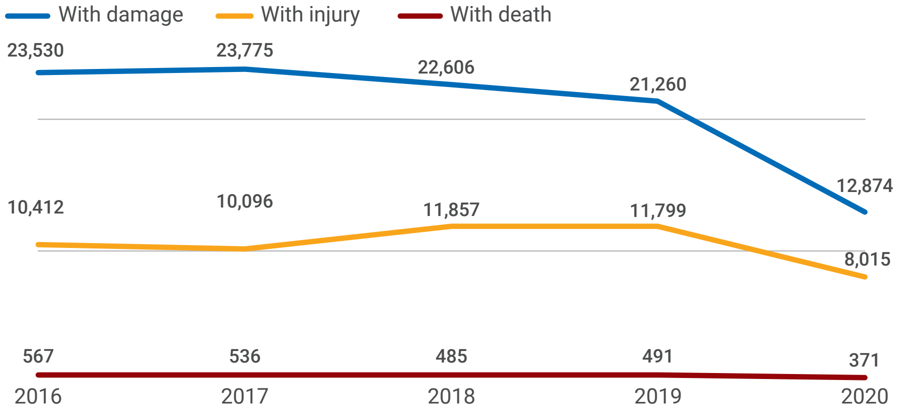
Hotspots and Heatmaps: Uncovering Data Patterns for Road Safety

Data visualizations are provided in the case study regions using alternative data sources, such as OSM, Mapbox, and Waze, as well as a select government dataset.

Bogotá, Colombia

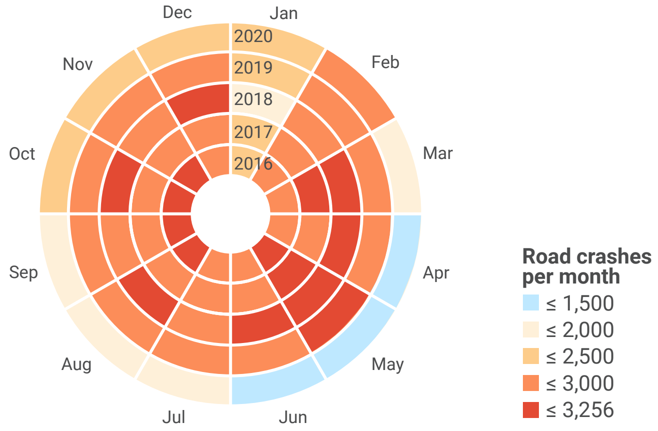
Temporal data visualizations show road safety patterns between years, seasons, months, weeks, days, and times of day. The Waze crash data used to train the ML model covered a period of six months, from July through December 2020. It was anticipated that the pandemic would affect the number of Waze crash reports, and potentially traffic patterns, as crashes reported by the government noticeably decreased compared to prior years (figure 3.1). The government dataset revealed fewer incidents starting in March 2020, suggesting that the number of crashes was affected by the pandemic, though it is worth noting that the speed limit was also reduced from 60km/h to 50 km/h in May 2020 (figure 3.2). With this in mind, the Waze data was used to identify road safety trends.

FIGURE 3.1: **Road crashes with damage, injury or death in Bogotá, 2016–2020**



SOURCE: Original figure for this publication, based on data from *Datos Abiertos Secretaría Distrital de Movilidad*.

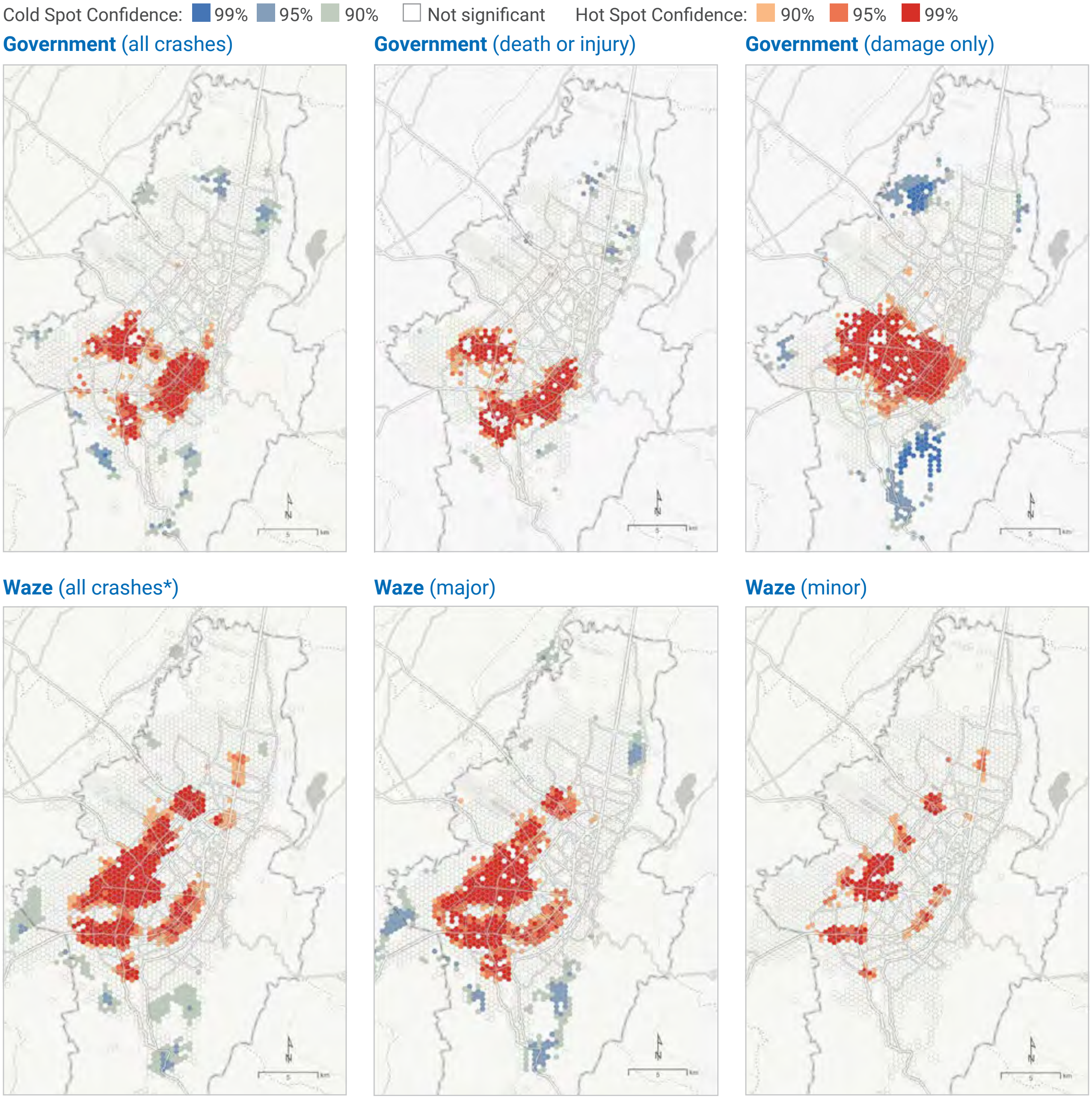
FIGURE 3.2: **Road crashes per month in Bogotá, 2016–2020**



SOURCE: Original figure for this publication, based on data from *Datos Abiertos Secretaría Distrital de Movilidad.*

Hotspot analysis groups crash locations to determine statistically significant clusters of crashes. Government and Waze datasets were analyzed during the same six-month window (figure 3.3). Between the two datasets, similar hotspots were found near Avenida Boyacá and Calle 6 along the highway in the south, Avenida Norte-Quito-Sur (NQS). Overall, Waze had more hotspots than the government dataset. Some minor road incidents captured by Waze may have gone unreported to the police. This trend can be seen in minor collisions clustering further north in the city. This cluster does not appear in the government data. Instead, clusters of government-reported crashes with only damage (no injury or fatality) appear in a central band. The approach to identify hotspots can vary, including the clustering method, size, shape, and search area of neighboring hotspots.

FIGURE 3.3: **Hotspot analysis of government and Waze crash data in Bogotá, July–December 2020**


Cold Spot Confidence: 🟦 99% 🟦 95% 🟩 90% □ Not significant Hot Spot Confidence: 🟧 90% 🟧 95% 🟥 99%
Government (all crashes)
Government (death or injury)
Government (damage only)
Waze (all crashes*)
Waze (major)
Waze (minor)
N
5
km

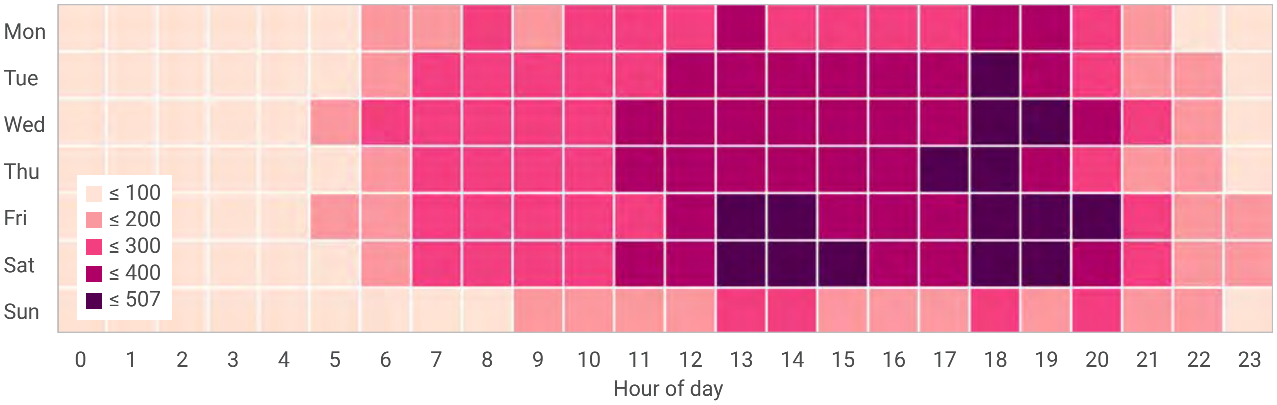
*Includes major and minor crashes, as well as those not categorized as either type.

SOURCE: Original figure for this publication, based on data from Datos Abiertos Secretaría Distrital de Movilidad and the Waze App. Learn more at waze.com. Basemap provided by Esri, HERE, Garmin, METI/NASA, USGS.

As with other alternative sources of data derived from mobile devices and apps, Waze crash reports are influenced by the location of the users, which affects where and when the crashes are reported. While Waze data notes major and minor incidents, the dataset will not include additional crash details typically obtained from an official source, such as type, severity, class, and reason. Even though users can validate reports (e.g., thumbs up) to provide a confidence and reliability rating and flag false reports, there is potential for duplication in Waze data. Deduplication was not conducted for this analysis because this study was interested in relative crash patterns.

Identifiable temporal patterns display when major crashes are aggregated by the day of the week and hour of the day (figure 3.4). In Bogotá, major crash reports increased between 6 and 7 p.m., having the most crashes during this window on Friday. Fewer incidents occurred on Sunday.

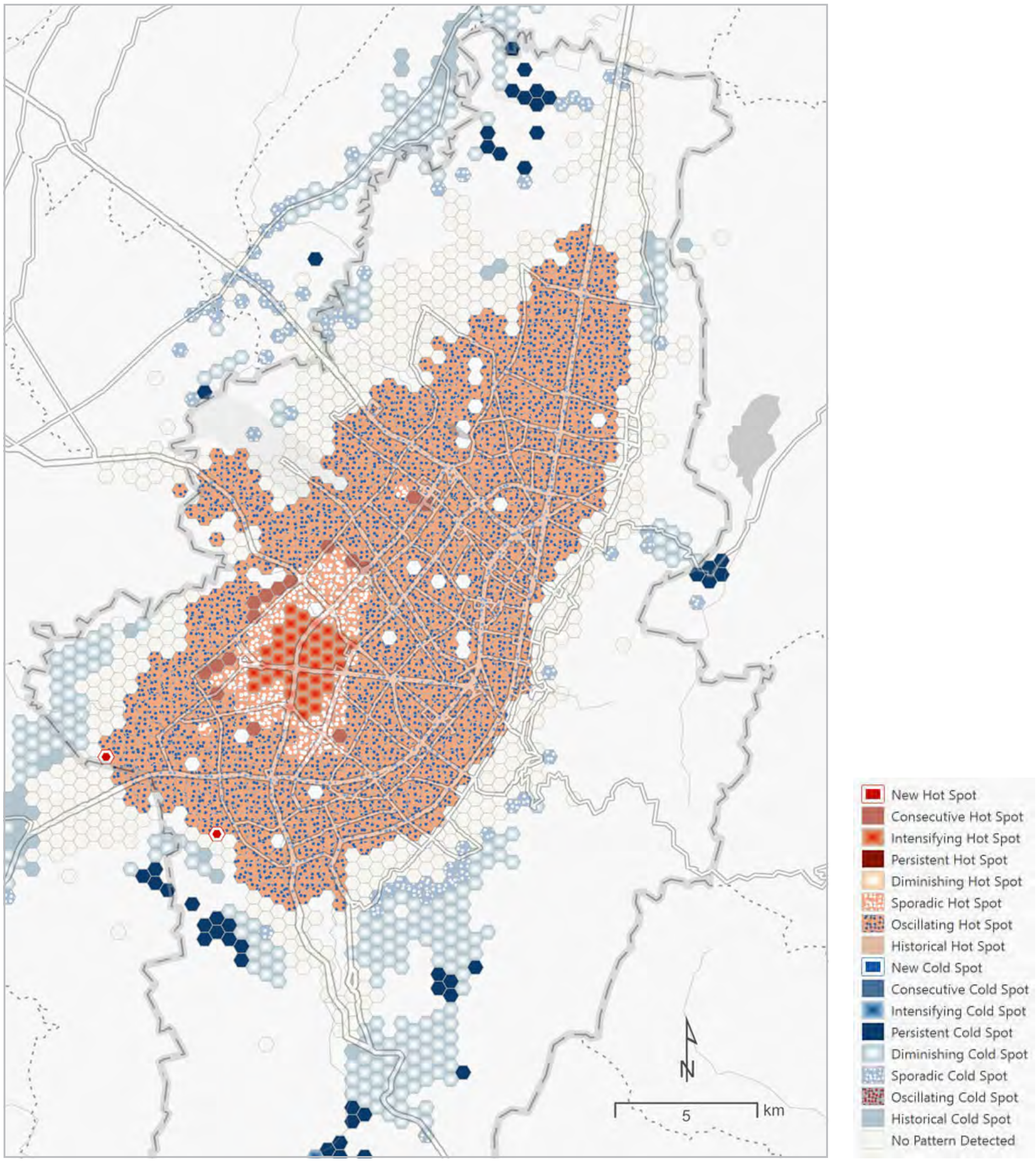
FIGURE 3.4: **Major crashes reported on Waze in Bogotá, July–December 2020**



SOURCE: Original figure for this publication, based on data provided by the Waze App. Learn more at waze.com.

Spatial and temporal analysis can be combined to identify areas for closer inspection that exhibit patterns over time. This is valuable given human movement or behavioral changes, including the effects of a pandemic, road construction, or updated speed limits, during the examined period. Emerging hotspot analysis reviews clusters of crashes that are consistent over time and ones that are intensifying or diminishing (figure 3.5).[73] In this example, each week was analyzed. Intensifying hotspot areas were statistically significant hotspots for 90 percent of the weeks analyzed with increasing intensity of hotspots, including the final week.

FIGURE 3.5: **Emerging hotspot analysis of Waze crashes in Bogotá, July–December 2020**

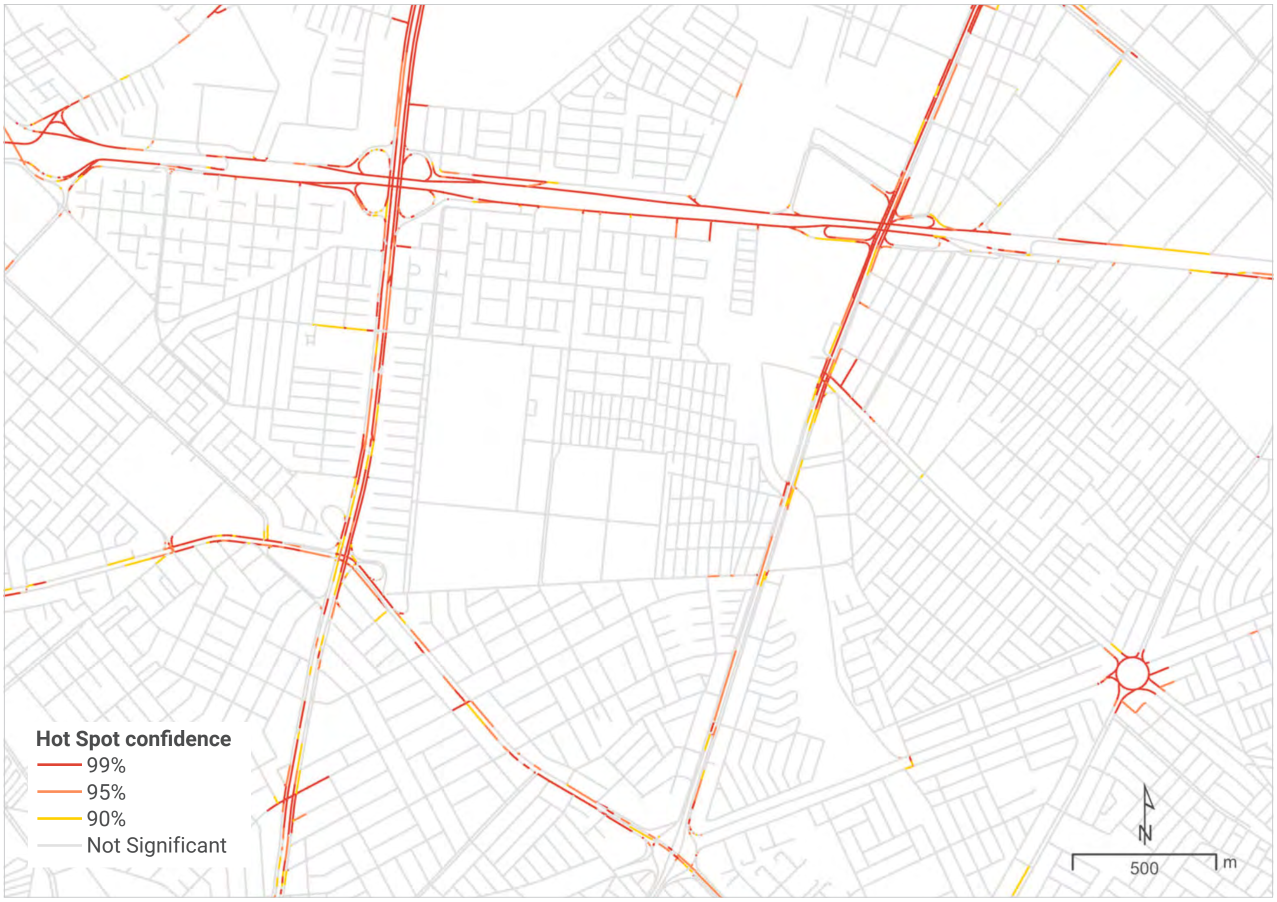


SOURCE: Original figure for this publication, based on data provided by the Waze App. Learn more at waze.com. Basemap provided by Esri, HERE, Garmin, METI/NASA, USGS.

[73] For a complete list of definitions, see "How Emerging Hot Spot Analysis Works": https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/learnmoreemerging.htm

If interventions or investments target a specific road, more geographically detailed information is required to make decisions. Hotspot analysis applied to road segments visualizes statistically significant crash frequencies along roads, as shown in figure 3.6.

FIGURE 3.6: **Hotspot analysis using Waze crash frequencies in Bogotá, July–December 2020**

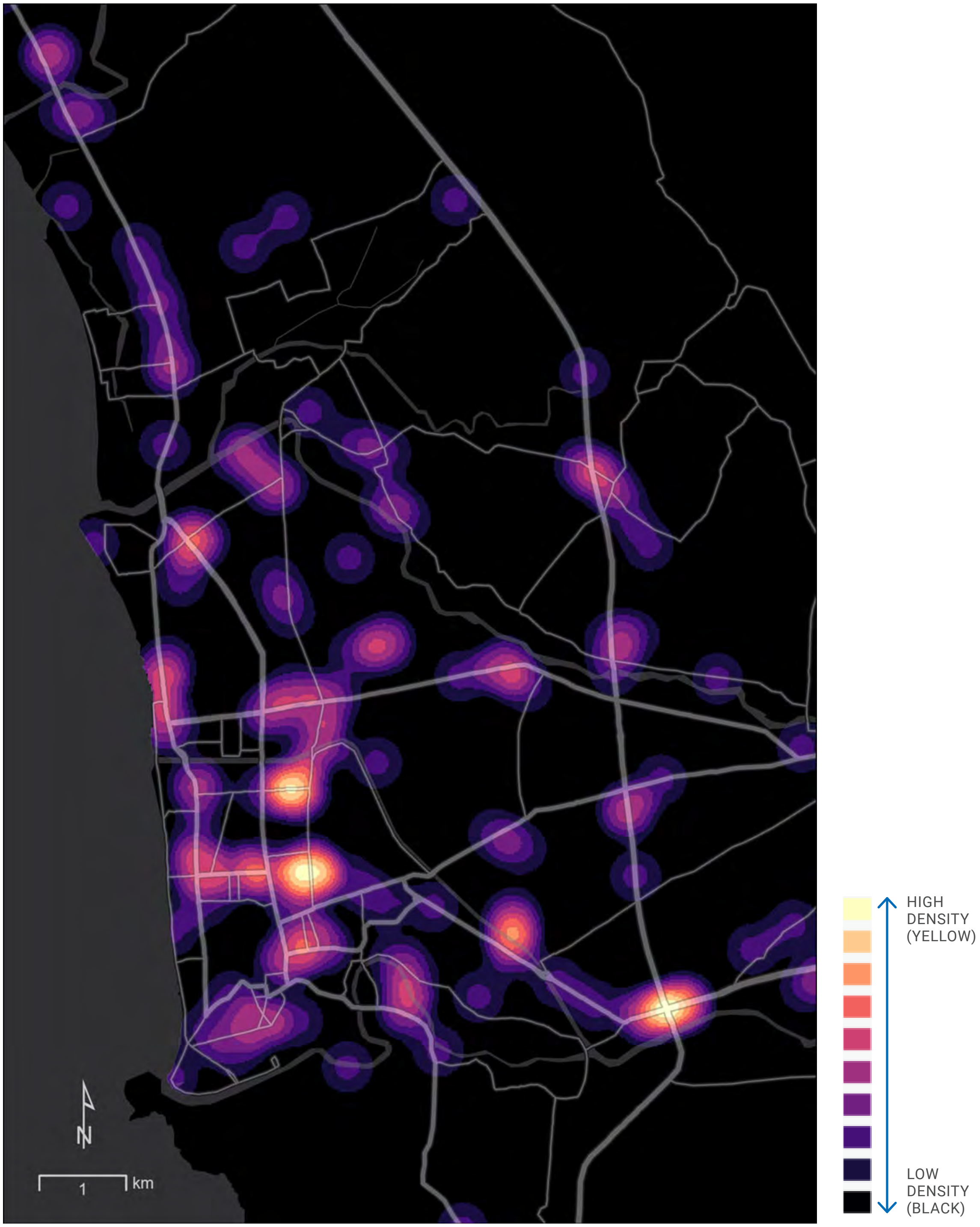


SOURCE: Original figure for this publication, based on data provided by OSM and the Waze App. Learn more at waze.com

Padang, Indonesia

Heatmaps visualize the density of crashes. While Waze data was sparse in Padang, some spatial patterns could be detected. A heatmap shows at least three distinct areas of high crash density that could be further examined during a site inspection (figure 3.7).

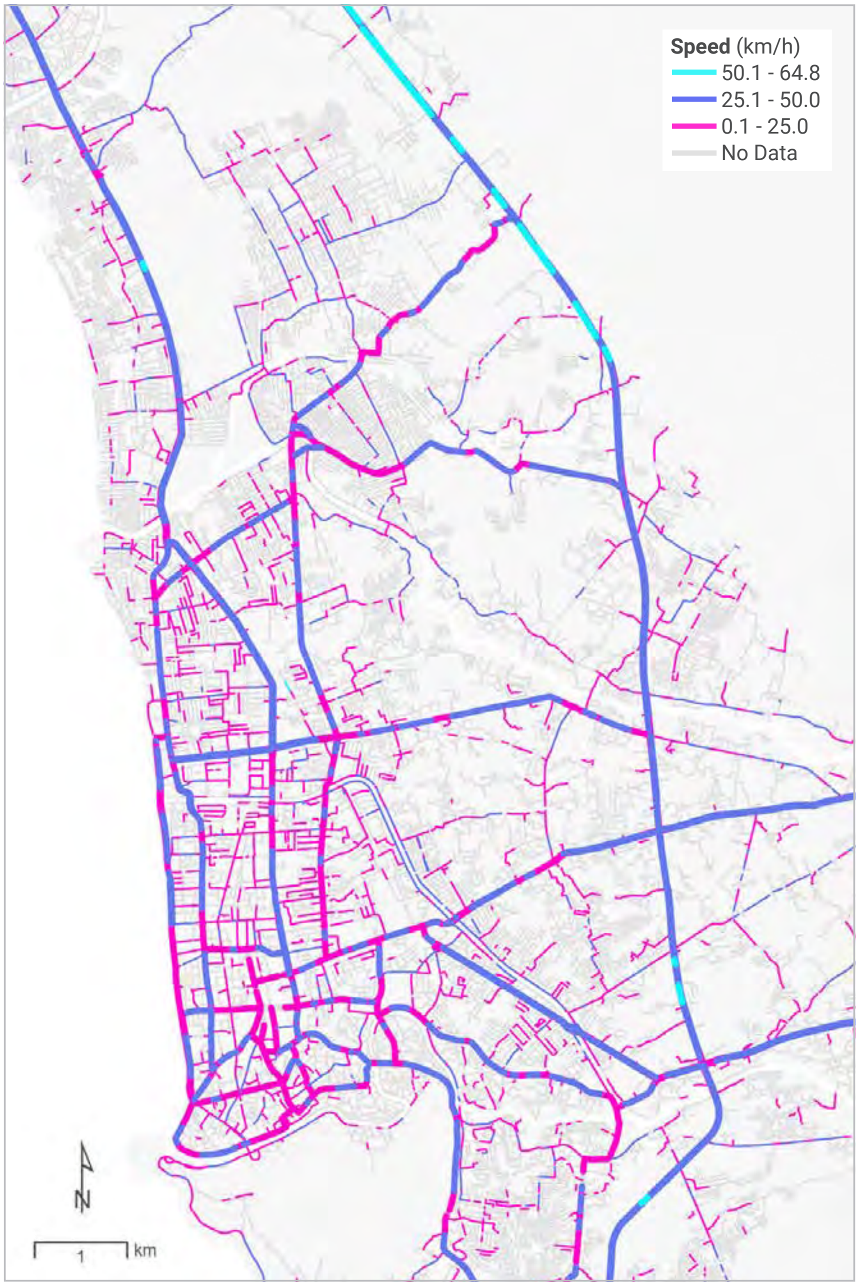
FIGURE 3.7: **Heatmap of crashes reported using the Waze app in Padang, April 2019–July 2021**



SOURCE: Original figure for this publication, based on data provided by the Waze App. Learn more at waze.com. Basemap provided by Esri, HERE, Garmin, METI/NASA, USGS.

Road safety assessments may require operating speeds of road segments. Mapbox collects this data from mobile devices and provides typical speeds per road segment in 5-minute increments. In Padang, Mapbox speeds were visualized for a Thursday from 5:00 p.m. to 6:00 p.m. (figure 3.8). Using the OSM road type to group and designate minor and major roads as a proxy for a low or high-speed limit (speed limits were sparsely noted in OSM), minor roads are visualized with thinner lines than major roads. The average speed typically slowed near intersections in pink (<25 km/h) when compared to major roads in purple (25-50 km/h). High-speed road segments exceeding 50 km/h are found heading north and south along Jalan By Pass. Identifying road segments with high speeds using Mapbox supports road safety assessments and the implementation of speed management or traffic calming measures.

FIGURE 3.8: Mapbox typical speeds in Padang on Thursday, 5:00 p.m. to 6:00 p.m.



SOURCE: Original figure for this publication, based on data provided by Mapbox. Basemap provided by Esri, HERE, Garmin, METI/NASA, USGS.

Annex 4: Classes Detected Using Mapillary Vistas Dataset in RIC Model and Input Classes for the RRE Model

All classes listed were detected using the Mapillary Vistas Dataset. Classes in bold were the input for the RRE Model.

animal--bird

animal--ground-animal

construction--barrier--ambiguous

construction--barrier--concrete-block

construction--barrier--curb

construction--barrier--fence

construction--barrier--guard-rail

construction--barrier--other-barrier

construction--barrier--road-median

construction--barrier--road-side

construction--barrier--separator

construction--barrier--temporary

construction--barrier--wall

construction--flat--bike-lane

construction--flat--crosswalk-plain

construction--flat--curb-cut

construction--flat--driveway

construction--flat--parking

construction--flat--parking-aisle

construction--flat--pedestrian-area

construction--flat--rail-track

construction--flat--road

construction--flat--road-shoulder

construction--flat--service-lane

construction--flat--sidewalk

construction--flat--traffic-island

construction--structure--bridge

construction--structure--building

construction--structure--garage

construction--structure--tunnel

human--person--individual

human--person--person-group

human--rider--bicyclist

human--rider--motorcyclist

human--rider--other-rider

marking--continuous--dashed

marking--continuous--solid

marking--continuous--zigzag

marking--discrete--ambiguous

marking--discrete--arrow--left

marking--discrete--arrow--other

marking--discrete--arrow--right

marking--discrete--arrow--split-left-or-straight

marking--discrete--arrow--split-right-or-straight

marking--discrete--arrow--straight

marking--discrete--crosswalk-zebra

marking--discrete--give-way-row

marking--discrete--give-way-single

marking--discrete--hatched--chevron

marking--discrete--hatched--diagonal

marking--discrete--other-marking

marking--discrete--stop-line

marking--discrete--symbol--bicycle

marking--discrete--symbol--other

marking--discrete--text

marking-only--continuous--dashed

marking-only--discrete--crosswalk-zebra

marking-only--discrete--other-marking

marking-only--discrete--text

nature--mountain

nature--sand

nature--sky

nature--snow

nature--terrain

nature--vegetation

nature--water

object--banner

object--bench

object--bike-rack

object--catch-basin

object--cctv-camera

object--fire-hydrant

object--junction-box

object--mailbox

object--manhole

object--parking-meter

object--phone-booth

object--pothole

object--sign--advertisement

object--sign--ambiguous

object--sign--back

object--sign--information

object--sign--other

object--sign--store

object--street-light

object--support--pole

object--support--pole-group

object--support--traffic-sign-frame

object--support--utility-pole

object--traffic-cone

object--traffic-light--general-single

object--traffic-light--pedestrians

object--traffic-light--general-upright

object--traffic-light--general-horizontal

object--traffic-light--cyclists

object--traffic-light--other

object--traffic-sign--ambiguous

object--traffic-sign--back

object--traffic-sign--direction-back

object--traffic-sign--direction-front

object--traffic-sign--front

object--traffic-sign--information-parking

object--traffic-sign--temporary-back

object--traffic-sign--temporary-front

object--trash-can

object--vehicle--bicycle

object--vehicle--boat

object--vehicle--bus

object--vehicle--car

object--vehicle--caravan

object--vehicle--motorcycle

object--vehicle--on-rails

object--vehicle--other-vehicle

object--vehicle--trailer

object--vehicle--truck

object--vehicle--vehicle-group

object--vehicle--wheeled-slow

object--water-valve

void--car-mount

void--dynamic

void--ego-vehicle

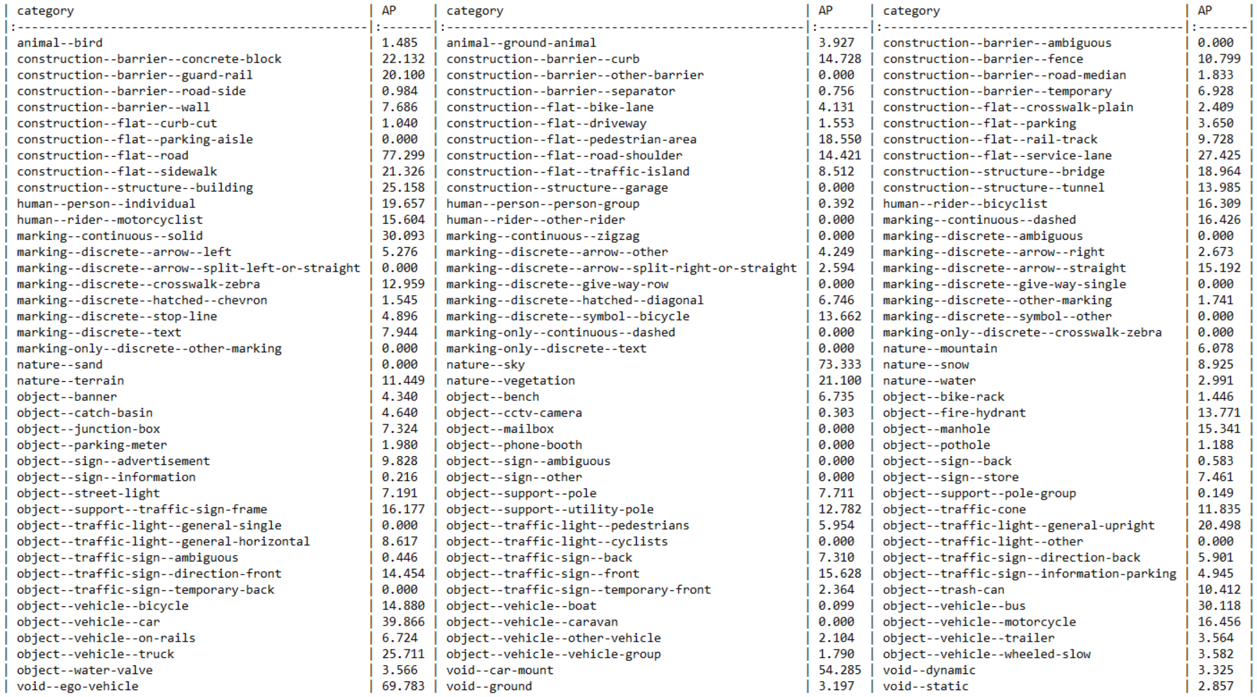
void--ground

void--static

void--unlabeled

Annex 5: Average Precision of the Bounding Box Detection and Classification

An Average Precision (AP) score closer to 100 indicates a better performance in correctly detecting and classifying an object. AP scores equal to zero mean that no data is available.



category	AP	category	AP	category	AP
animal--bird	1.485	animal--ground-animal	3.927	construction--barrier--ambiguous	0.000
construction--barrier--concrete-block	22.132	construction--barrier--curb	14.728	construction--barrier--fence	10.799
construction--barrier--guard-rail	20.100	construction--barrier--other-barrier	0.000	construction--barrier--road-median	1.833
construction--barrier--road-side	0.984	construction--barrier--separator	0.756	construction--barrier--temporary	6.928
construction--barrier--wall	7.686	construction--flat--bike-lane	4.131	construction--flat--crosswalk-plain	2.409
construction--flat--curb-cut	1.040	construction--flat--driveway	1.553	construction--flat--parking	3.650
construction--flat--parking-aisle	0.000	construction--flat--pedestrian-area	18.550	construction--flat--rail-track	9.728
construction--flat--road	77.299	construction--flat--road-shoulder	14.421	construction--flat--service-lane	27.425
construction--flat--sidewalk	21.326	construction--flat--traffic-island	8.512	construction--structure--bridge	18.964
construction--structure--building	25.158	construction--structure--garage	0.000	construction--structure--tunnel	13.985
human--person--individual	19.657	human--person--person-group	0.392	human--rider--bicyclist	16.309
human--rider--motorcyclist	15.604	human--rider--other-rider	0.000	marking--continuous--dashed	16.426
marking--continuous--solid	30.093	marking--continuous--zigzag	0.000	marking--discrete--ambiguous	0.000
marking--discrete--arrow--left	5.276	marking--discrete--arrow--other	4.249	marking--discrete--arrow--right	2.673
marking--discrete--arrow--split-left-or-straight	0.000	marking--discrete--arrow--split-right-or-straight	2.594	marking--discrete--arrow--straight	15.192
marking--discrete--crosswalk-zebra	12.959	marking--discrete--give-way-row	0.000	marking--discrete--give-way-single	0.000
marking--discrete--hatched--chevron	1.545	marking--discrete--hatched--diagonal	6.746	marking--discrete--other-marking	1.741
marking--discrete--stop-line	4.896	marking--discrete--symbol--bicycle	13.662	marking--discrete--symbol--other	0.000
marking--discrete--text	7.944	marking-only--continuous--dashed	0.000	marking-only--discrete--crosswalk-zebra	0.000
marking-only--discrete--other-marking	0.000	marking-only--discrete--text	0.000	nature--mountain	6.078
nature--sand	0.000	nature--sky	73.333	nature--snow	8.925
nature--terrain	11.449	nature--vegetation	21.100	nature--water	2.991
object--banner	4.340	object--bench	6.735	object--bike-rack	1.446
object--catch-basin	4.640	object--cctv-camera	0.303	object--fire-hydrant	13.771
object--junction-box	7.324	object--mailbox	0.000	object--manhole	15.341
object--parking-meter	1.980	object--phone-booth	0.000	object--pothole	1.188
object--sign--advertisement	9.828	object--sign--ambiguous	0.000	object--sign--back	0.583
object--sign--information	0.216	object--sign--other	0.000	object--sign--store	7.461
object--street-light	7.191	object--support--pole	7.711	object--support--pole-group	0.149
object--support--traffic-sign-frame	16.177	object--support--utility-pole	12.782	object--traffic-cone	11.835
object--traffic-light--general-single	0.000	object--traffic-light--pedestrians	5.954	object--traffic-light--general-upright	20.498
object--traffic-light--general-horizontal	8.617	object--traffic-light--cyclists	0.000	object--traffic-light--other	0.000
object--traffic-sign--ambiguous	0.446	object--traffic-sign--back	7.310	object--traffic-sign--direction-back	5.901
object--traffic-sign--direction-front	14.454	object--traffic-sign--front	15.628	object--traffic-sign--information-parking	4.945
object--traffic-sign--temporary-back	0.000	object--traffic-sign--temporary-front	2.364	object--trash-can	10.412
object--vehicle--bicycle	14.880	object--vehicle--boat	0.099	object--vehicle--bus	30.118
object--vehicle--car	39.866	object--vehicle--caravan	0.000	object--vehicle--motorcycle	16.456
object--vehicle--on-rails	6.724	object--vehicle--other-vehicle	2.104	object--vehicle--trailer	3.564
object--vehicle--truck	25.711	object--vehicle--vehicle-group	1.790	object--vehicle--wheeled-slow	3.582
object--water-valve	3.566	void--car-mount	54.285	void--dynamic	3.325
void--ego-vehicle	69.783	void--ground	3.197	void--static	2.857

Glossary of Terms

Big Data	Large data sets that require significant processing power and/or complex computational techniques to reveal patterns, trends, and correlations.
Development Data Partnership (DDP)	A partnership between international organizations and companies, created to facilitate the use of third-party data in research and international development.
Deep Learning (DL)	A branch of artificial intelligence that involves creating algorithms for deep artificial neural networks, inspired by the human brain, to learn complex patterns from high dimensional and large quantities of data.
Fatalities and Serious Injuries (FSI)	A metric of those killed or seriously injured in a traffic crash which is used to monitor traffic safety performance. Fatalities are defined as those who die within 30 days of the crash.
Intelligent Transport System (ITS)	The collection, analysis, and transmission of transportation, vehicle, and infrastructure data that informs users with real-time updates and improves future operations and predictions.
Internet of Things (IoT)	Devices that are connected to the internet to send and/or receive data.
Machine Learning (ML)	Method to systematically derive patterns, identify trends, and make conclusions from data with minimal human intervention.
Neural Network	A set of connected algorithms typically organized in three layers: input layer, hidden layer(s), and an output layer.
Overall Project Traffic and Road Safety Risk (OPTRSR)	The entire traffic and road safety risk of a project that evaluates the road infrastructure, vehicle operating speeds, road user behavior, vehicle standards, and post-crash trauma care.
Road Crash	The collision of a vehicle with another entity, such as a car, bicycle, stationary object, pedestrian, or animal, that causes injury or damage to one or more of the entities on a road or road-related area.
Road Safety	System to reduce risks to road users, preventing death or injury.
Road Safety Assessments	Systematic review of the current road or traffic scheme to identify hazardous areas.
Road Safety Audit (RSA)	Independent, systematic evaluation of the modification or addition to the road or traffic scheme to determine the crash potential and safety performance for all road users.
Road Safety Impact Assessment (RSIA)	The safety performance ranking of planned road construction or modification design schemes and their effect on the surrounding road network.
Road Safety Observatory (RSO)	A regional network of government representatives that facilitates the sharing and exchange of road safety data and expertise. The World Bank operates RSOs in Latin America (OISEVI), Africa (ARSO), and Asia-Pacific (APRSO).
Safe System	An approach to road safety that integrates principles for safer vehicles, safer roads, and safer users to eliminate death and serious injuries.
Supervised Learning	A machine learning task using labeled data to train the model with input-output pairs.
Unsupervised Learning	A machine learning technique that extracts patterns from unlabeled data. For example, grouping or clustering data with similar attributes.
Vulnerable Road Users	Individuals at a higher risk using the road because they do not have the protection of an enclosed vehicle, such as pedestrians, motorcyclists, bicyclists, and those on animals or animal drawn carts.

References

Australian BITRE (Bureau of Infrastructure and Transport Research Economics). "Australian Road Deaths Database (ARDD)." Australian BITRE. Updated May 13, 2021. https://data.gov.au/data/dataset/australian-road-deaths-database

Bedoya Arguelles, Guadalupe, Svetoslava Petkova Milusheva, Arianna Legovini, and Sarah Elizabeth Williams. "Smart and Safe Kenya Transport (SMARTTRANS)." Washington, DC: World Bank, 2019. https://documents1.worldbank.org/curated/en/723411574361015073/pdf/Smart-and-Safe-Kenya-Transport-SMARTTRANS.pdf

Bliss, Tony and Jeanne Breen. "Meeting the Management Challenges of the Decade of Action for Road Safety." *IATSS Res*. 35 (2012): 48–55. https://doi.org/10.1016/j.iatssr.2011.12.001

Bostrom, Nick and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 316-334. Cambridge: Cambridge University Press, 2014.

Das, Subasish and Greg P. Griffin. "Investigating the Role of Big Data in Transportation Safety." *Transportation Research Record* 2674, no. 6 (2020): 244–52. https://doi.org/10.1177/0361198120918565

Diop, Makhtar. "All Road Deaths Are Preventable. We Can Make It Happen." World Bank. Accessed May 14, 2021. https://blogs.worldbank.org/transport/all-road-deaths-are-preventable-we-can-make-it-happen

DT Global. "Indonesia: Establishment of Integrated Road Asset Management Systems." Accessed October 4, 2021. https://dt-global.com/projects/irams-dc

Google. "Google Maps, Google Earth, and Street View." Accessed May 14, 2021. https://about.google/brand-resource-center/products-and-services/geo-guidelines/

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask R-CNN." *2017 IEEE International Conference on Computer Vision* (2017): 2980-2988.

Hidalgo, Darío and Claudia Adriazola-Steil. "Bogota's Vision Zero Road Safety Plan Is Saving Lives." TheCityFix. Last modified September 26, 2019. https://thecityfix.com/blog/bogotas-vision-zero-road-safety-plan-saving-lives-dario-hidalgo-claudia-adriazola-steil/

Institute for Transportation and Development Policy. "Pune, India Wins 2020 Sustainable Transport Award." Last modified June 27, 2019. https://www.itdp.org/2019/06/27/pune-india-wins-2020-sustainable-transport-award/

International Transport Forum. "Best Practice for Urban Road Safety: Case Studies." *International Transport Forum Policy Papers,* no. 76 (2020).

International Transport Forum. *Zero Road Deaths and Serious Injuries: Leading a Paradigm Shift to a Safe System*. Paris: OECD Publishing, 2016. https://doi.org/10.1787/9789282108055-en

iRAP (International Road Assessment Programme). *iRAP Star Rating and Investment Plan Implementation Support Guide*. London: iRAP, March 2017.

Krambeck, Holly, Magreth Kakoko, and Mireille Raad. *Using Computer Vision to Automatically Detect Road Features for Road Safety Audits and Assessments: Inception Report*. Washington, DC: World Bank, 2019.

Lovón-Melgarejo, Jesús, Alonso Tenorio-Trigoso, Manuel Castillo-Cara, and Daniel Miranda. "Identification of Risk Zones for Road Safety through Unsupervised Learning Algorithms." In *16th LACCEI International Multi-Conference for Engineering, Education, and Technology: Innovation in Education and Inclusion, Lima, Peru, July 2018.* http://www.laccei.org/LACCEI2018-Lima/full_papers/FP413.pdf

Milusheva, Sveta, Robert Marty, Guadalupe Bedoya, Sarah Williams, Elizabeth Resor, and Arianna Legovini. "Applying Machine Learning and Geolocation Techniques to Social Media Data (Twitter) to Develop a Resource for Urban Planning." *PLoS ONE* 16, 2 (2021): https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0244317

Neilson, Alex, Indratmo, Ben Daniel, Stevanus Tjandra. "Systematic Review of the Literature on Big Data in the Transportation Domain: Concepts and Applications." *Big Data Res.* 17 (2019): 35-44. https://doi.org/10.1016/j.bdr.2019.03.001

Neuhold, G., T. Ollmann, S. R. Bulò, and P. Kontschieder. "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes." *2017 IEEE International Conference on Computer Vision (ICCV)* (2017): 5000-5009. doi: 10.1109/ICCV. 2017.534

ODI (Overseas Development Institute). "Bogotá." ODI: Think Change. Accessed October 12, 2021. https://odi.org/en/about/features/bogot%C3%A1/

ODPH (Open Data Philippines). "Open Data Philippines." ODPH. Accessed June 3, 2021. https://data.gov.ph/

OECD (Organisation for Economic Co-operation and Development) /ITF (International Transport Forum). *Big Data and Transport: Understanding and Assessing Options*. Paris: OECD/ITF, 2015. https://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf

Omdena. "Rating Road Safety Through Machine Learning to Prevent Road Accidents." Accessed May 28, 2021. https://omdena.com/projects/ai-road-safety/

Ospina-Mateus, Holman, Leonardo Augusto Quintana Jiménez, Francisco José López-Valdés, Natalie Morales-Londoño, and Katherinne Salas-Navarro. "Using Data-Mining Techniques for the Prediction of the Severity of Road Crashes in Cartagena, Colombia." In *Applied Computer Sciences in Engineering*. Edited by J. Figueroa-García, M. Duarte-González, S. Jaramillo-Isaza, A. Orjuela-Cañon, Y. Díaz-Gutierrez, 309-20. Cham: Springer, 2019. https://doi.org/10.1007/978-3-030-31019-6_27

Silva, Philippe Barbosa, Michelle Andrade, and Sara Ferreira. "Machine Learning Applied to Road Safety Modeling: A Systematic Literature Review." *Journal of Traffic and Transportation Engineering* 7, no. 6 (2020): 775-790. https://doi.org/10.1016/j.jtte.2020.07.004

Suresh, Harini and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." In *Proceedings of Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), Association for Computing Machinery, New York, October 2021.* https://doi.org/10.1145/3465416.3483305

US NHTSA (United States National Highway Traffic Safety Administration). "Data." US NHTSA. Accessed May 28, 2021. https://www.nhtsa.gov/data

WHO (World Health Organization). *Global Status Report on Road Safety 2018*. Geneva: WHO, 2018.

World Bank. "Better Data for Safer Roads: The Powerful Mission of Road Safety Observatories." Last modified November 5, 2020. https://www.worldbank.org/en/news/video/2020/11/05/better-data-for-safer-roads-the-powerful-mission-of-road-safety-observatories

World Bank. *Colombia - Programmatic Productive and Sustainable Cities Development Policy Loans*. Washington, DC: World Bank, 2020. http://documents.worldbank.org/curated/en/426591583968971309/Colombia-Programmatic-Productive-and-Sustainable-Cities-Development-Policy-Loans

World Bank. *GRSF DRIVER Completion Report*. Washington, DC: World Bank, 2019. https://documents1.worldbank.org/curated/en/245151560919065747/pdf/Data-for-Road-Incident-Visualization-Evaluation-and-Reporting-Lowing-the-Barriers-to-Evidence-Based-Road-Safety-Management-in-Resource-Constrained-Countries.pdf

World Bank. *Environmental and Social Framework for IPF Operations, ESS4: Community Health and Safety*. Washington, DC: World Bank, 2018.

World Bank. *Good Practice Note: Road Safety*. Washington, DC: World Bank, 2019. https://pubdocs.worldbank.org/en/648681570135612401/Good-Practice-Note-Road-Safety.pdf

World Bank. *Guide for Road Safety Opportunities and Challenges: Low and Middle Income Country Profiles*. Washington, DC: 2020. https://openknowledge.worldbank.org/handle/10986/33363

World Bank. *Indonesia Public Expenditure Review 2020: Spending for Better Results*. Washington, DC: World Bank, 2020. https://openknowledge.worldbank.org/handle/10986/33954

World Bank. *Innovative Road Safety Risk Assessment Tool with Automated Image Analysis Technology*. Washington, DC: World Bank, 2019.

Word Bank. *Making Roads Safer*. Washington, DC: World Bank, 2014.

World Bank. *Mobile Metropolises: Urban Transport Matters: An IEG Evaluation of the World Bank Group's Support for Urban Transport*. Washington, DC: World Bank, 2017.

World Bank. "Open Traffic Data to Revolutionize Transport." Last modified December 19, 2016. https://www.worldbank.org/en/news/feature/2016/12/19/open-traffic-data-to-revolutionize-transport

World Bank. *Open Traffic: Easing Urban Congestion*. Washington, DC: World Bank, n.d. https://olc.worldbank.org/system/files/WBG_BD_CS_OpenTraffic_1.pdf

World Bank. *Road Safety Indicators for Project Monitoring*. Washington, DC: World Bank, 2021.

World Bank. *The High Toll of Traffic Injuries: Unacceptable and Preventable*. Washington, DC: World Bank, 2017.

World Bank. *Use of AI Technology to Support Data Collection for Project Preparation and Implementation: A 'Learning-by-doing' Process*. Washington, DC: World Bank, 2021.

World Bank. *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank, 2021. doi:10.1596/978-1-4648-1600-0

Zeng, Qiang, Helai Huang, Xin Pei, S.C. Wong, and Mingyun Gao. “Rule Extraction from an Optimized Neural Network for Traffic Crash Frequency Modeling.” *Accident Analysis & Prevention* 97 (2016): 87-95. doi: 10.1016/j.aap.2016.08.017

Zhang, Min, Yang Liu, Shaohua Luo, Siyan Gao. “Research on Baidu Street View Road Crack Information Extraction Based on Deep Learning Method.” *Journal of Physics: Conference Series* no. 1616 (2020). https://iopscience.iop.org/article/10.1088/1742-6596/1616/1/012086/pdf

Ziakopoulos, Apostolos and George Yannis. “Using AI for Spatial Predictions of Driver Behavior.” (ITF) International Transport Forum Roundtable on Artificial Intelligence in Road Traffic Crash Prevention. Presentation, February 2021.
https://www.nrso.ntua.gr/geyannis/conf/cp450-using-ai-for-spatial-predictions-of-driver-behavior/

This guidance note offers a practical introduction to integrating big data and machine learning in road safety evaluations. It outlines data requirements for several road safety assessments, provides a convenient overview of relevant big data sources, and explains machine learning fundamentals for the application of these advanced technologies, specifically for road safety. The note proposes an Integrated Framework for Road Safety, which takes the reader step-by-step through a machine learning workflow to evaluate road risk, using case studies in Bogotá, Colombia and Padang, Indonesia.

The Integrated Framework for Road Safety uses machine learning to identify road characteristics from street view images and predict road segment risk based on those identifiable characteristics. As a result, road segment risk was predicted with 72.5 percent accuracy in Bogotá. While the preliminary results in Padang were encouraging, additional data is required to verify the performance in a new context. However, the workflow illustrated through these case studies shows potential for replicability. All code for the Integrated Framework for Road Safety is free and publicly available for repurposing and refining to local context through a link provided in the note.

The framework exemplifies current capabilities to reduce the reliance on manual image annotations and highlights the potential to conduct a road safety scan without years of historical crash data. The increasing availability of big data and the growing use of machine learning models for road safety point to rapidly evolving technological solutions that have immense capacity to improve the quality and efficiency of road safety assessments in developing countries.

GRSF
Global Road Safety Facility

UKaid
from the British people