AI 894: Design & Implementation of AI Systems

Automated Narrative Creation of Collisions Involving a Pedestrian, Bicycle, or Micromobility Device

Project 7
Daniel Zielinski

Master of Artificial Intelligence (AI)
FA 2024

Instructor: Wang Chengfei, Ph.D.

Blank Page

General Guidelines

1. To complete all the homework assignments for this course please use this template document.
2. Each assignment has to be submitting by Sunday 11:59 PM EST.
3. Each figure should be followed by a brief description about the figure.
4. The figures can be hand drawn and scanned in some circumstances, but the hand drawn figure should be clear and legible to obtain full credits. Unclear hand drawn figures will receive partial credits. For constructing figures and diagrams it is advised to use tools.
5. Figures and tables should have appropriate captions. For documenting and referencing styles please follow the APA or MLA writing style.
6. Please make sure that you provide a reference section.
7. Any material text or figure taken from books, journals or Internet should be referenced. If you have a sentence or a figure that does not belong (authorship) to you, they need to be clearly referenced. If you fail to do so your report will be considered as a case for plagiarism. It is your duty to make sure that your report is free from any activity related to plagiarism. In case you are suspected of attempting plagiarism then you will be responsible for the cause. The penalty for plagiarism will be a "0" awarded to your report. So it is good to keep simple, always have the principle to acknowledge people for their contributions.

Please go through the following instructions before submitting the report

Academic Integrity

Academic integrity — scholarship free of fraud and deception — is an important educational objective of Penn State. Academic dishonesty can lead to a failing grade or referral to the Office of Student Conduct.

Academic dishonesty includes, but is not limited to:

- cheating
- plagiarism
- fait brication of information or citations
- facil ating acts of academic dishonesty by others
- unauthorized prior possession of examinations
- submitting the work of another person or work previously used without informing the instructor and securing written approval
- tampering with the academic work of other students

How Academic Integrity Violations Are Handled

In cases where academic integrity is questioned, procedure requires an instructor to notify a student of suspected dishonesty before filing a charge and recommended sanction with the college. Procedures allow a student to accept or contest a charge. If a student chooses to contest a charge, the case will then be managed by the respective college or campus Academic Integrity Committee. If a disciplinary sanction also is recommended, the case will be referred to the Office of Student Conduct.

All Penn State colleges abide by this Penn State policy, but review procedures may vary by college when academic dishonesty is suspected. Information about Penn State's academic integrity policy and college review procedures is included in the information that students receive upon enrolling in a course.

Additionally, Penn State students are expected to act with civility and personal integrity; respect other students' dignity, rights, and property; and help create and maintain an environment in which all can succeed through the fruits of their own efforts. An environment of academic integrity is requisite to respect for oneself and others, and a civil community.

For More Information on Academic Integrity at Penn State

Please see the Academic Integrity Chart for specific college contact information or visit one of the following URLs:

- Penn State Senate Policy on Academic Integrity
- iStudy for Success! — learn about plagiarism, copyright, and academic integrity through an educational module
- Turnitin a web-based plagiarism detection and prevention system

PROJECT PROPOSAL

Purpose:

This project intends to create a system to summarize and categorize automobile-versus-pedestrian and bicycle collisions using raw text of police dispatches to augment detailed state-reported crash reports. Only about half of ped/bike collisions are estimated to be reported to the state and this project aims to gauge and qualify if this under-reporting contains more value to traffic engineers and decision makers than state-reported crash data alone. The summarized and categorized data will be summarized into a safety ranking score for intersections and compared to a recently completed traffic engineering project in Cleveland, Ohio. This comparison will conclude if recommended data sourcing processes has a blind spot and if dispatch-derived collision data has the potential to redirect funding decisions for pedestrian safety improvement plans (PSIP) to other areas.

1. Project title

Automated Narrative Creation of Collisions Involving a Pedestrian, Bicycle, or Micromobility Device

2. Keywords: discipline / application

1. **Disciplines**: Deep Learning, NLP Classification, Generative AI, Applied Statistics
2. **Application domains**: Transportation, Traffic Engineering, Law Enforcement

3. Background and Challenges

- Bike Cleveland

"*[Bike Cleveland is a] 501(c)(3) advocacy non-profit for people on bikes in the Greater Cleveland area. Representing over 1000 due paying members and more than 32 local businesses, we make sure that any time the conversation turns to transportation — that people on bikes are being considered alongside people in cars. We work to improve policy, infrastructure, and legislation to help make our roads places that serve people and communities, not just traffic.*" (Bike Cleveland, 2015)

This group maintains a website here: https://www.bikecleveland.org/

Part of Bike Cleveland's advocacy efforts have involved collecting and disseminating information about pedestrian and bicycle collisions in the Cleveland area. This effort began as an automated Twitter Bot that posted the location of 911 police dispatches involving a pedestrian or bicycle. After a substantial number of police dispatches were collected, Bike Cleveland compared the tweeted dispatch reports to state-mandated OH-1 reports and discovered less than half of dispatch records, which clearly described a bicycle or pedestrian collision, had a corresponding OH-1 record. This prompted a discussion among group members and an effort has been led by the group to advocate for better reporting by city officials.

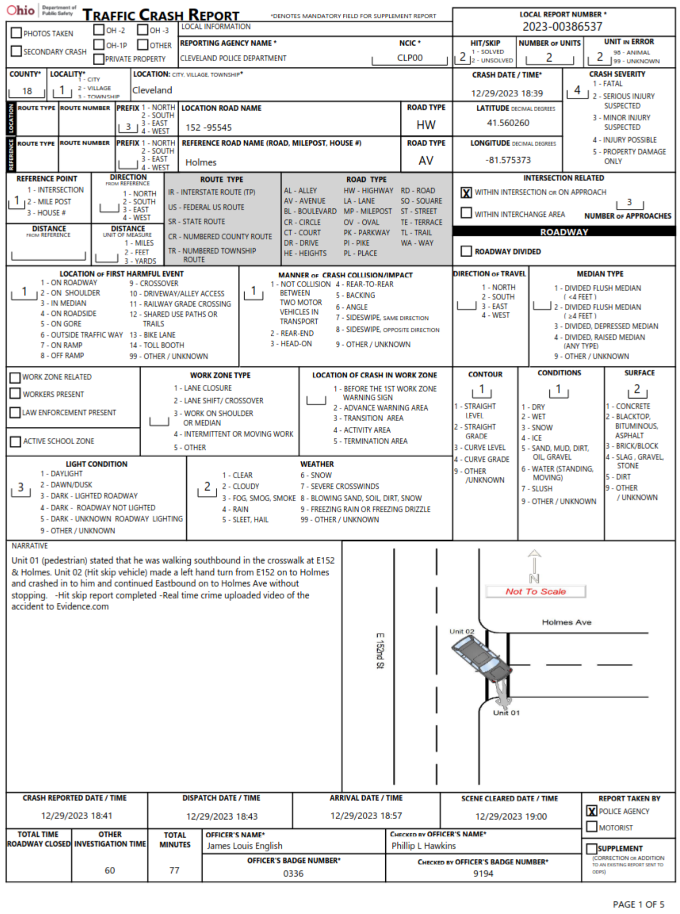
- Vision Zero

Is a type of government program which seeks to eliminate serious injury and death caused by traffic violence. The City of Cleveland has adopted a Vision Zero Action Plan resolution which aims to eliminate death and serious injury on Cleveland roads by 2032. This initiative is outlined on a website here: https://www.visionzerocle.org/pages/action-plan (Cleveland VZTF, 2022)

- Data Sources

The data for this project is sourced from Bike Cleveland and their ongoing effort to collect, publish, and advocate for street designs that are safe for cyclists and pedestrians. Feature data will consist of two inputs:

OH-1 crash data is sourced from the Ohio Crash System Tools: https://www.transportation.ohio.gov/programs/highway+safety/highway-safety-resources/08-crash-trends-resources and is primarily stored on PDF files. All collisions are available, but this project will focus only on those involving a pedestrian, bicylcist, or person using a micromobility device. These PDFs are generally consistent and use the layout described in this procedure manual: https://dam.assets.ohio.gov/image/upload/publicsafety.ohio.gov/HSY7010.pdf. Basic information, such as location and date/time, is stored in a structured file, referred to as GCAT. Some information, such as collision narratives, is not extracted into the GCAT structured data and will require using OCR to strip text out of a certain area within each report to be used as a feature input. It may be possible to extract information from the diagram drawings, but this may be unnecessary if the text inputs produce acceptable results. A challenge with these PDFs is that in 2023 Cleveland Police started using a computer system to create OH-1 reports. In 2022 and prior most reports are hand-written and will present a challenge to extract text. Page 1 of example OH-1:



Ohio | Department of Public Safety **TRAFFIC CRASH REPORT** *DENOTES MANDATORY FIELD FOR SUPPLEMENT REPORT

LOCAL REPORT NUMBER * 2023-00386537

□ PHOTOS TAKEN □ OH -2 □ OH -3
□ SECONDARY CRASH □ OH-1P □ OTHER
□ PRIVATE PROPERTY

LOCAL INFORMATION

REPORTING AGENCY NAME * CLEVELAND POLICE DEPARTMENT
NCIC * CLP00
HIT/SKIP 2 (1 - SOLVED, 2 - UNSOLVED)
NUMBER OF UNITS 2
UNIT IN ERROR 2 (98 - ANIMAL, 99 - UNKNOWN)

COUNTY* 18
LOCALITY* 1 (1 - CITY, 2 - VILLAGE, 3 - TOWNSHIP)
LOCATION: CITY, VILLAGE, TOWNSHIP* Cleveland
CRASH DATE / TIME* 12/29/2023 18:39
CRASH SEVERITY 4 (1 - FATAL, 2 - SERIOUS INJURY SUSPECTED, 3 - MINOR INJURY SUSPECTED, 4 - INJURY POSSIBLE, 5 - PROPERTY DAMAGE ONLY)

LOCATION
ROUTE TYPE **ROUTE NUMBER** **PREFIX** 3 (1 - NORTH, 2 - SOUTH, 3 - EAST, 4 - WEST)
LOCATION ROAD NAME 152 -95545
ROAD TYPE HW
LATITUDE DECIMAL DEGREES 41.560260

REFERENCE
ROUTE TYPE **ROUTE NUMBER** **PREFIX** (1 - NORTH, 2 - SOUTH, 3 - EAST, 4 - WEST)
REFERENCE ROAD NAME (ROAD, MILEPOST, HOUSE #) Holmes
ROAD TYPE AV
LONGITUDE DECIMAL DEGREES -81.575373

REFERENCE POINT 1 (1 - INTERSECTION, 2 - MILE POST, 3 - HOUSE #)
DIRECTION FROM REFERENCE (1 - NORTH, 2 - SOUTH, 3 - EAST, 4 - WEST)
DISTANCE FROM REFERENCE
DISTANCE UNIT OF MEASURE (1 - MILES, 2 - FEET, 3 - YARDS)

ROUTE TYPE
IR - INTERSTATE ROUTE (TP)
US - FEDERAL US ROUTE
SR - STATE ROUTE
CR - NUMBERED COUNTY ROUTE
TR - NUMBERED TOWNSHIP ROUTE

ROAD TYPE
AL - ALLEY, AV - AVENUE, BL - BOULEVARD, CR - CIRCLE, CT - COURT, DR - DRIVE, HE - HEIGHTS
HW - HIGHWAY, LA - LANE, MP - MILEPOST, OV - OVAL, PK - PARKWAY, PI - PIKE, PL - PLACE
RD - ROAD, SO - SOUARE, ST - STREET, TE - TERRACE, TL - TRAIL, WA - WAY

INTERSECTION RELATED
☒ WITHIN INTERSECTION OR ON APPROACH
□ WITHIN INTERCHANGE AREA
NUMBER OF APPROACHES 3

ROADWAY
□ **ROADWAY DIVIDED**

LOCATION OF FIRST HARMFUL EVENT 1
1 - ON ROADWAY, 2 - ON SHOULDER, 3 - IN MEDIAN, 4 - ON ROADSIDE, 5 - ON GORE, 6 - OUTSIDE TRAFFIC WAY, 7 - ON RAMP, 8 - OFF RAMP, 9 - CROSSOVER, 10 - DRIVEWAY/ALLEY ACCESS, 11 - RAILWAY GRADE CROSSING, 12 - SHARED USE PATHS OR TRAILS, 13 - BIKE LANE, 14 - TOLL BOOTH, 99 - OTHER / UNKNOWN

MANNER OF CRASH COLLISION/IMPACT 1
1 - NOT COLLISION BETWEEN TWO MOTOR VEHICLES IN TRANSPORT, 2 - REAR-END, 3 - HEAD-ON, 4 - REAR-TO-REAR, 5 - BACKING, 6 - ANGLE, 7 - SIDESWIPE, SAME DIRECTION, 8 - SIDESWIPE, OPPOSITE DIRECTION, 9 - OTHER / UNKNOWN

DIRECTION OF TRAVEL
1 - NORTH, 2 - SOUTH, 3 - EAST, 4 - WEST

MEDIAN TYPE
1 - DIVIDED FLUSH MEDIAN (<4 FEET), 2 - DIVIDED FLUSH MEDIAN (≥4 FEET), 3 - DIVIDED, DEPRESSED MEDIAN, 4 - DIVIDED, RAISED MEDIAN (ANY TYPE), 9 - OTHER / UNKNOWN

□ WORK ZONE RELATED
□ WORKERS PRESENT
□ LAW ENFORCEMENT PRESENT
□ ACTIVE SCHOOL ZONE

WORK ZONE TYPE
1 - LANE CLOSURE, 2 - LANE SHIFT/ CROSSOVER, 3 - WORK ON SHOULDER OR MEDIAN, 4 - INTERMITTENT OR MOVING WORK, 5 - OTHER

LOCATION OF CRASH IN WORK ZONE
1 - BEFORE THE 1ST WORK ZONE WARNING SIGN, 2 - ADVANCE WARNING AREA, 3 - TRANSITION AREA, 4 - ACTIVITY AREA, 5 - TERMINATION AREA

CONTOUR 1
1 - STRAIGHT LEVEL, 2 - STRAIGHT GRADE, 3 - CURVE LEVEL, 4 - CURVE GRADE, 9 - OTHER /UNKNOWN

CONDITIONS 1
1 - DRY, 2 - WET, 3 - SNOW, 4 - ICE, 5 - SAND, MUD, DIRT, OIL, GRAVEL, 6 - WATER (STANDING, MOVING), 7 - SLUSH, 9 - OTHER / UNKNOWN

SURFACE 2
1 - CONCRETE, 2 - BLACKTOP, BITUMINOUS, ASPHALT, 3 - BRICK/BLOCK, 4 - SLAG , GRAVEL, STONE, 5 - DIRT, 9 - OTHER / UNKNOWN

LIGHT CONDITION 3
1 - DAYLIGHT, 2 - DAWN/DUSK, 3 - DARK - LIGHTED ROADWAY, 4 - DARK - ROADWAY NOT LIGHTED, 5 - DARK - UNKNOWN ROADWAY LIGHTING, 9 - OTHER / UNKNOWN

WEATHER 2
1 - CLEAR, 2 - CLOUDY, 3 - FOG, SMOG, SMOKE, 4 - RAIN, 5 - SLEET, HAIL, 6 - SNOW, 7 - SEVERE CROSSWINDS, 8 - BLOWING SAND, SOIL, DIRT, SNOW, 9 - FREEZING RAIN OR FREEZING DRIZZLE, 99 - OTHER / UNKNOWN

NARRATIVE

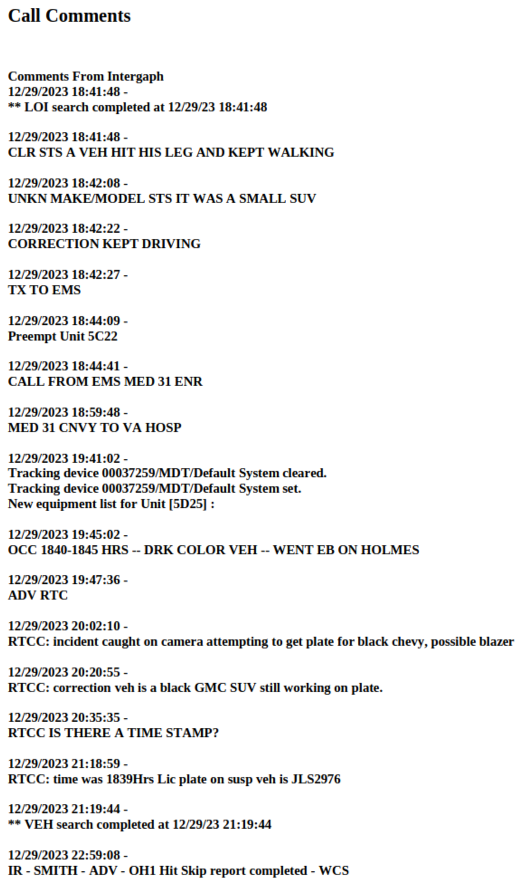
Unit 01 (pedestrian) stated that he was walking southbound in the crosswalk at E152 & Holmes. Unit 02 (Hit skip vehicle) made a left hand turn from E152 on to Holmes and crashed in to him and continued Eastbound on to Holmes Ave without stopping. -Hit skip report completed -Real time crime uploaded video of the accident to Evidence.com

CRASH REPORTED DATE / TIME 12/29/2023 18:41	**DISPATCH DATE / TIME** 12/29/2023 18:43	**ARRIVAL DATE / TIME** 12/29/2023 18:57	**SCENE CLEARED DATE / TIME** 12/29/2023 19:00	**REPORT TAKEN BY** ☒ POLICE AGENCY □ MOTORIST

TOTAL TIME ROADWAY CLOSED	**OTHER INVESTIGATION TIME** 60	**TOTAL MINUTES** 77	**OFFICER'S NAME*** James Louis English **OFFICER'S BADGE NUMBER*** 0336	**CHECKED BY OFFICER'S NAME*** Phillip L Hawkins **CHECKED BY OFFICER'S BADGE NUMBER*** 9194	□ **SUPPLEMENT** (CORRECTION OR ADDITION TO AN EXISTING REPORT SENT TO ODPS)

PAGE 1 OF 5

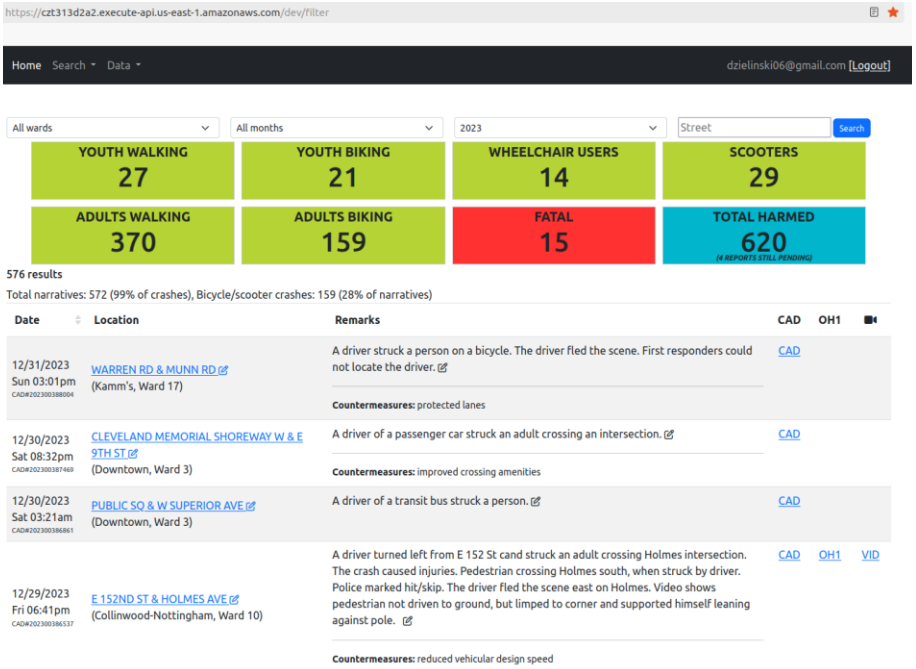
Police dispatch records have been gathered through a campaign of bot-driven public records requests to the City of Cleveland's Public Records Request Center for dispatches categorized as "pedestrian-struck." https://clevelandoh.govqa.us/WEBAPP/_rs/supporthome.aspx . These files stored as text-only PDF files so it should be straight-forward to extract text using OCR from these documents. All iterations of information are included in each file reflecting iterations of information as a dispatcher receives it. So, information later in this text is more accurate than early information. Also, dispatchers use a lot of acronyms to reduce the keystrokes needed to receive and send information quickly. Preprocessing of the text from "CLR STS" to "Caller states" will likely be required to be able to fully leverage LLM tokens. This is an important point and anticipated to be a particular challenge for Llama, which is intended for use to supplement the narrative creation. Example of police dispatch notes:



Call Comments

```
Comments From Intergaph
12/29/2023 18:41:48 -
** LOI search completed at 12/29/23 18:41:48

12/29/2023 18:41:48 -
CLR STS A VEH HIT HIS LEG AND KEPT WALKING

12/29/2023 18:42:08 -
UNKN MAKE/MODEL STS IT WAS A SMALL SUV

12/29/2023 18:42:22 -
CORRECTION KEPT DRIVING

12/29/2023 18:42:27 -
TX TO EMS

12/29/2023 18:44:09 -
Preempt Unit 5C22

12/29/2023 18:44:41 -
CALL FROM EMS MED 31 ENR

12/29/2023 18:59:48 -
MED 31 CNVY TO VA HOSP

12/29/2023 19:41:02 -
Tracking device 00037259/MDT/Default System cleared.
Tracking device 00037259/MDT/Default System set.
New equipment list for Unit [5D25] :

12/29/2023 19:45:02 -
OCC 1840-1845 HRS -- DRK COLOR VEH -- WENT EB ON HOLMES

12/29/2023 19:47:36 -
ADV RTC

12/29/2023 20:02:10 -
RTCC: incident caught on camera attempting to get plate for black chevy, possible blazer

12/29/2023 20:20:55 -
RTCC: correction veh is a black GMC SUV still working on plate.

12/29/2023 20:35:35 -
RTCC IS THERE A TIME STAMP?

12/29/2023 21:18:59 -
RTCC: time was 1839Hrs Lic plate on susp veh is JLS2976

12/29/2023 21:19:44 -
** VEH search completed at 12/29/23 21:19:44

12/29/2023 22:59:08 -
IR - SMITH - ADV - OH1 Hit Skip report completed - WCS
```

Narratives/Labels are sourced from a data store hosted by Bike Cleveland and published through an AWS web portal: https://czt313d2a2.execute-api.us-east-1.amazonaws.com/dev/filter
which features a graphical interface listing collision summaries and offers basic filtering function. An example is shown here:


https://czt313d2a2.execute-api.us-east-1.amazonaws.com/dev/filter
Home
Search
Data
dzielinski06@gmail.com [Logout]
All wards
All months
2023
Street
Search

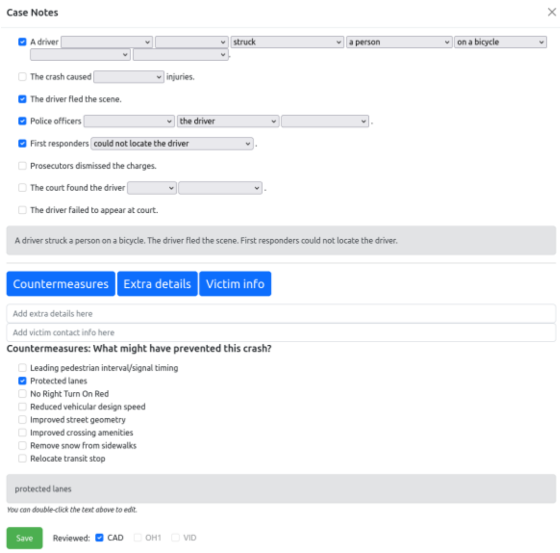
YOUTH WALKING	YOUTH BIKING	WHEELCHAIR USERS	SCOOTERS
27	21	14	29
ADULTS WALKING	**ADULTS BIKING**	**FATAL**	**TOTAL HARMED**
370	159	15	620 (4 REPORTS STILL PENDING)

576 results

Total narratives: 572 (99% of crashes), Bicycle/scooter crashes: 159 (28% of narratives)

Date	Location	Remarks	CAD	OH1	📹
12/31/2023 Sun 03:01pm CAD#202300388004	WARREN RD & MUNN RD (Kamm's, Ward 17)	A driver struck a person on a bicycle. The driver fled the scene. First responders could not locate the driver. **Countermeasures:** protected lanes	CAD		
12/30/2023 Sat 08:32pm CAD#202300387469	CLEVELAND MEMORIAL SHOREWAY W & E 9TH ST (Downtown, Ward 3)	A driver of a passenger car struck an adult crossing an intersection. **Countermeasures:** improved crossing amenities	CAD		
12/30/2023 Sat 03:21am CAD#202300386861	PUBLIC SQ & W SUPERIOR AVE (Downtown, Ward 3)	A driver of a transit bus struck a person.	CAD		
12/29/2023 Fri 06:41pm CAD#202300386537	E 152ND ST & HOLMES AVE (Collinwood-Nottingham, Ward 10)	A driver turned left from E 152 St cand struck an adult crossing Holmes intersection. The crash caused injuries. Pedestrian crossing Holmes south, when struck by driver. Police marked hit/skip. The driver fled the scene east on Holmes. Video shows pedestrian not driven to ground, but limped to corner and supported himself leaning against pole. **Countermeasures:** reduced vehicular design speed	CAD	OH1	VID

These narratives were entered by volunteers who have read the dispatch and OH-1 reports and generated "Remarks" to summarize the collision and are extracted into a csv file. An anticipated challenge with this data is that when this project started narrative creation was a free-text entry but has recently evolved to consist of entries into a drop-down form that produces a sentence narrative, like this:



☑ A driver [ ] [ ] [struck] [a person] [on a bicycle] [ ] [ ].

☐ The crash caused [ ] injuries.

☑ The driver fled the scene.

☑ Police officers [ ] [the driver] [ ].

☑ First responders [could not locate the driver].

☐ Prosecutors dismissed the charges.

☐ The court found the driver [ ] [ ].

☐ The driver failed to appear at court.

A driver struck a person on a bicycle. The driver fled the scene. First responders could not locate the driver.

Add extra details here

Add victim contact info here

Countermeasures: What might have prevented this crash?

☐ Leading pedestrian interval/signal timing
☑ Protected lanes
☐ No Right Turn On Red
☐ Reduced vehicular design speed
☐ Improved street geometry
☐ Improved crossing amenities
☐ Remove snow from sidewalks
☐ Relocate transit stop

protected lanes

You can double-click the text above to edit.

Reviewed: ☑ CAD ☐ OH1 ☐ VID

This helps with speed and ease of data entry and label consistency for recently created labels, but older narratives may contain unneeded information. This may hinder a model's ability to fit required narrative creation.

- AI Ethical Challenge

The production of this data relies on communication between humans and data entry by humans, so it’s certain that intrinsic biases held by the humans behind these reports will be imperceptibly embedded in a model using them. This is especially true for dispatch reports as these reports are momentary impressions communicated over the phone and keyed into a computer system as quickly as possible. The audio of the 911 calls are not the focus of this project but some are available in this dataset. Comparison of eyewitness descriptions in these audio files against detailed investigative findings reveals puzzling inaccuracies in dispatch reports. The project's main goal, however, is to test the operationalization of data and details otherwise lost. Details which are more likely to be inaccurate carry less weight in furthering this project’s goal and core elements, such as the size and type of vehicle and if a person was walking or on a bicycle, will comprise the most focused of the project. Entries from volunteers for collision remarks may also introduce bias, but with a considerable number of the narratives entered with drop down selections, this should be limited. The design and aim of this project have been selected in such a way as to limit the concern of AI-perpetuated biases.

Also, the info in these reports is sensitive and sometimes documents the most traumatic, or last, moments of a persons life. Delicate and careful consideration will be made when referring to serious injury and fatal events.

4. Project Objectives

Existing research in understanding underreporting of pedestrian and bicycle collisions has focused on quantifying the extent and nature of collision underreporting. This project accepts those findings and aims to develop a method to fill the underreporting gap and conclude if additional information of underreported collisions would produce a different action in a recent transportation project.

1) Reconcile Cleveland Police 911 dispatches of pedestrian and bicycle collisions and OH-1 reports and create an algorithm to generate narratives summarizing information for each collision. This algorithm will be a two-step combination of predictive ML and Gen AI. The predictive ML layer will focus on the classification of vehicle types, maneuvering, and pedestrian characteristics and will be an input to a Gen AI algorithm which will summarize those inputs into a plain-spoken sentence. The predictive ML layer will predict:

Driver vehicle type:
* Unknown
* Car
* Large Passenger Vehicle
* Very Large Vehicle
* Motorcycle or Micro-mobility Device

Maneuvered:
* Unknown
* Straight
* A Left Turn
* A Right Turn
* Backward
* Vehicle as Weapon {maybe}

Pedestrian Type:
* Unknown
* An Adult
* A Child
* Multiple People

Pedestrian vehicle type:
* Unknown
* On Foot
* A Bicycle
* A Micro-mobility Device
* A Wheelchair or Stroller

In the Area Of:
* Unknown
* An Intersection
* Mid-block
* A Bike Lane
* A Driveway

2) The narrative generation algorithm will be deployed into the volunteer narrative-generation process and be self-updating using pickle.

3) Collect and summarize collision narratives for planning areas, such as intersections, and describe typical pedestrian and bicycle collision characteristics in that area. Summary should conform to guidance in Section 2.3.4. Analyze Historical/Observed Crash Data in ODOT Safety Analysis Guidance (ODOT, 2022).

4) In 2021, the City of Cleveland contracted for improvements to pedestrian infrastructure at 42 intersections along Principal Arterial roadways as part of a public works project entitled *CUY - Cleveland PSIP (*Burgess & Niple, 2021). Thie project was pursued as part of a Pedestrian Safety Improvement (PSIP) focus of an ODOT funding program, The Highway Safety Improvement Project (HSIP). The project cost $1,468,664 and Cleveland received $830,000 from ODOT to fund the project (ODOT, 2023). This objective intends to summarize pedestrian safety at all intersections along principal arterials in Cleveland using only OH-1 reports, and then summarize again using OH-1 reports plus police dispatch reports. A conclusion of consistency will be made of intersection prioritizations between the two datasets and intersections selected for the *Cleveland PSIP* project will receive additional consideration.

The scoring method practiced by traffic engineers is expected to be somewhat different from OH-1 only data because of professional judgment. More difference is expected when using OH-1 plus dispatch data.

Because of this, a rank comparison is anticipated to be the most appropriate statistical test. First, a validation test will be used to verify if the OH-1 only list identifies at least 33 of the 42 intersections treated with the Cleveland PSIP project in the top 42 positions. Then a comparison of the rankings of all intersections along major arterials will be made between the OH-1 only ranking and the OH-1 + dispatch ranking using three statistics: Spearman, Kendall tau, and Hoeffding. Tentatively, ranking coefficients above 0.8 will be used as a threshold to test if the two data sources produce materially similar rankings. There is an expectation that if underreporting is influenced by equity issues the Hoeffding coefficient will be higher than Spearman and Kendall.

5. Feasibility Studies

4.1 Scientific Survey

- Identify keywords that describe your project

NLP categorization
NLP summarization
Generative AI
Car Accident Underreporting
Applied Statistics

- Find at least 6 scientific articles (papers and journals) that handled similar problem to your project.

Doggett is the closest recent analog to this project and seemed closest to operationalizing multiple data sources into the underreporting gap:

Doggett, S., Ragland, D. R., & Felschundneff, G. (2018). Evaluating Research on Data Linkage to Assess Underreporting of Pedestrian and Bicyclist Injury in Police Crash Data. *UC Berkeley: Safe Transportation Research & Education Center*. https://escholarship.org/uc/item/0jq5h6f5

Hauer and Scott have explored the theme of underreporting but because these papers are dated they are cited here to establish a timeline for how long this problem has been knowns and to highlight the intractability of underreporting.

Hauer, E., & Hakkert, A. S. (1988). Extent and Some Implications of Incomplete Accident Reporting. Transportation Research Record, 1185. http://onlinepubs.trb.org/Onlinepubs/trr/1988/1185/1185-001.pdf

Scott, Robert, & Carroll, Philip. (1971). Acquisition of Information on Exposure and on Non-Fatal Crashes - Accident Data Inaccuracies Vol. II. *Highway Safety Research Institute, The University of Michigan Ann Arbor, Michigan*. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/1054/15848.0001.001.pdf

Abay, Ahmed, and Khadka are broad explorations of collision underreporting trends and characteristics:

Abay, Kibrom A. (2015). Investigating the nature and impact of reporting bias in road crash data. *Transportation Research Part A: Policy and Practice*, *71*, 31–45. https://doi.org/10.1016/j.tra.2014.11.002

Ahmed, A., Sadullah, A. F. M., & Yahya, A. S. (2019). Errors in accident data, its types, causes and methods of rectification-analysis of the literature. *Accident Analysis & Prevention*, *130*, 3–21. https://doi.org/10.1016/j.aap.2017.07.018

Khadka, Anish, Parkin, John, Pilkington, Paul, Joshi, Sunil Kumar, & Mytton, Julie. (2022). Completeness of police reporting of traffic crashes in Nepal: Evaluation using a community crash recording system. *Traffic Injury Prevention*, *23*(2), 79–84. https://doi.org/10.1080/15389588.2021.2012766

Nauman explores an action plan for implementing Vision Zero safety improvements by improving organizational networks within large cities. This is an interesting idea for operationalizing solutions but is too far out of scope to include these ideas in the current project.

Naumann, Rebecca B., Heiny, Stephen, Evenson, Kelly R., LaJeunesse, Seth, Cooper, Jill F., Doggett, Sarah, & Marshall, Stephen W. (2019). Organizational networks in road safety: Case studies of U.S. Vision Zero cities. *Traffic Injury Prevention*, *20*(4), 378–385. https://doi.org/10.1080/15389588.2019.1587752

- Explain how solutions/approaches in these scientific articles differ from your contributions and findings.

The current project accepts the findings of past work, but this work has functioned to quantify and characterize underreporting. The intention of this project is to operationalize alternative data and conclude if it would make a difference to include it into a traffic planning process. This project also differs from other work in the focus on police dispatch reports instead of hospital data.

4.2 Technical Survey

- Identify and discuss existing business products (start up, software, ..) close to your project.

AASHTO Safety Software is SAAS licensed product used to screen intersections and sections on collision metrics and aligns with objective 3 of this project: https://www.aashtoware.org/products/safety/safety-overview/

- Find and analyze related projects with code if possible (Github, pre-trained models, … for examples)

This project is intended to augment an existing documentation project at Bike Cleveland posted here: https://czt313d2a2.execute-api.us-east-1.amazonaws.com/dev/filter

Objective 1 of this project will create a two-step AI pipeline of two distinct ML models. The first will be a an NLP categorization model identifying unit characteristics, and the second will be an on-device LLM and producing a GenAI summary narrative:

A related example of an on-device application of the Llama LLM is 'easy-llama' here: https://github.com/ddh0/easy-llama

Another possibly related example makes use of images in the pipeline, and while it's not anticipated be part of the first iteration of this project, a related example doing this is here:

https://github.com/Rajeevsinghsisodiya/Integrating-LLaMA-Language-Model-with-FLUX-Pipeline-for-Enhanced-Image-and-Text-Generation

- Explain how your project and your solution differ from existing business products and technical projects available online?

This project will differ from the AASHTO's SAAS Safety Software because it will include data that is historically underreported.

6. Methodology (AI System)

- Is the dataset already used in other projects?

No. This is a novel dataset sourced through a years-long campaign of public records requests performed by Bike Cleveland.

- Briefly explain the data analytics workflow and pipeline to solve your problem.

  Text in police dispatch PDFs and OH-1 narratives (when available) will be converted into string variables and used as input into a predictive ML model to categorize unit types. Those outputs will be used as prompts to the Llama LLM, along with the original text, with goal of summarizing original text.

- Clearly list the task that will be completed in the following weeks.

  Text will be stripped from PDFs and a training label dataset will be assembled using classic regex and manipulation functions.

- How your methodology proves/disapproves your project objectives.

  Each objective builds data to objective 4, which then narrows the proof to a classic statistical test. Right now, success parameter values are naive but offer a framework to simply findings into a yes or no.

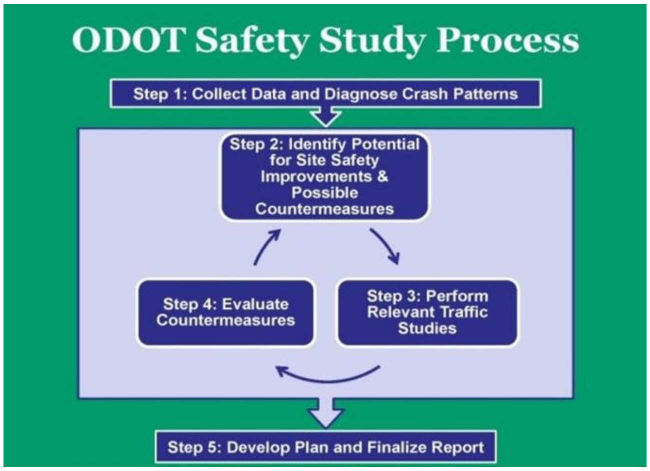
7. Deliverables

- Minimum viable prototype
- Final report
- Reproducible AI Workflow
  - Installation instructions
  - Deployment instructions
  - Commented code
  - Datasets
  - Saved trained models.
- Slides (Presentation)
- Video (Demonstration)

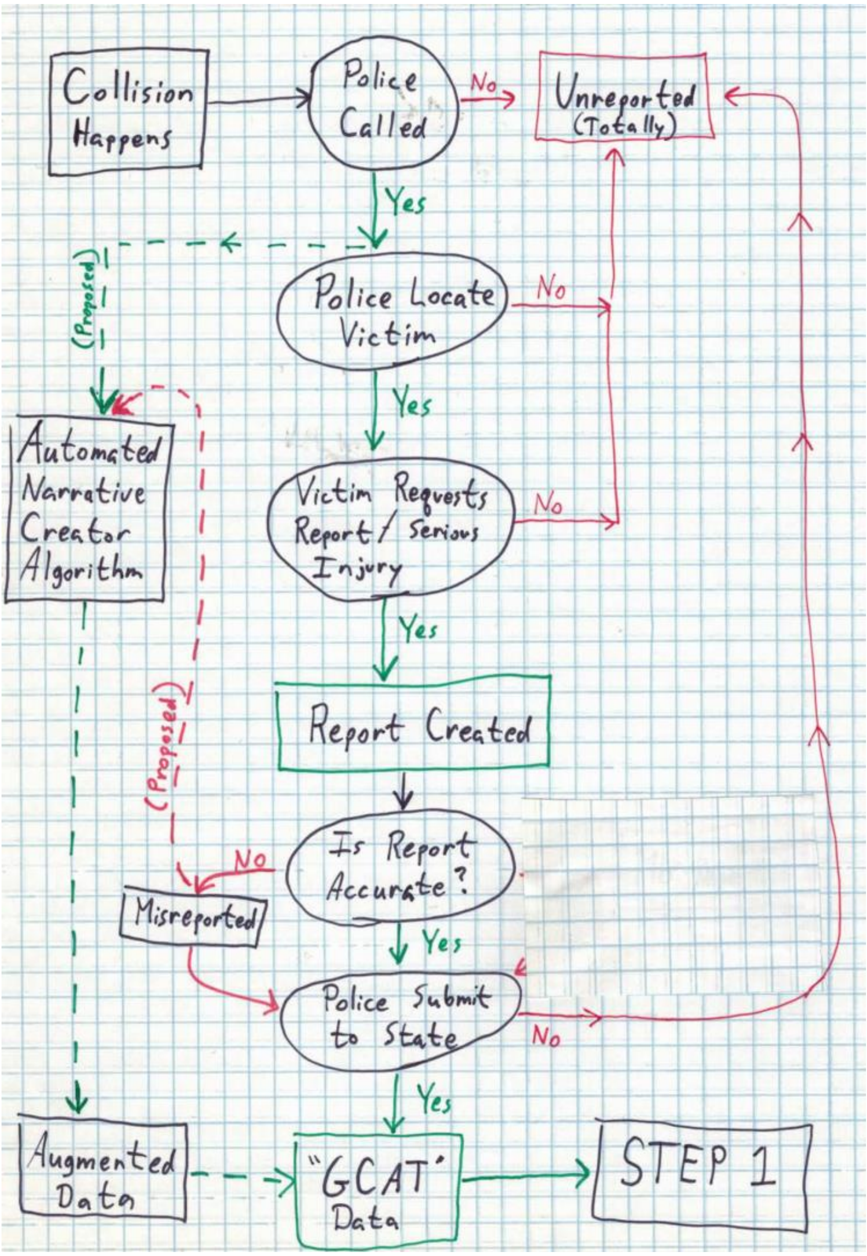
8. Importance and Impacts

The recommended first step in a data-driven traffic safety study is to collect data (ODOT, 2022) and the only recommended data is confirmed crashes submitted to the state on OH-1 reports. The only way to make an OH-1 record of a collision is by making a report to a police officer. While reports may be made to an officer through any interview, this process most often starts through the 911 phone system and the police dispatch. However, 'pedestrian-struck' dispatches don't always result in a police officer creating a report. A notable percentage of time first responders will go to a location only to find all the relevant parties have left the scene. The result is a discrepancy between evident collisions and OH-1 reports. Past research has identified underreporting of pedestrian and bicycle collisions as a severe issue with estimates ranging "from 44 to 75 percent for pedestrian crashes, and from 7 to 46 percent for bicycle crashes" (Doggett, et al, 2018). Preliminary analysis of the current dataset partially confirms these findings with only 49% of pedestrian-struck dispatches having an associated official OH-1 collision report in 2023.

Once created, police submit OH-1 reports to the state for inclusion in the Ohio Crash System Tools. Use of state-provided Crash System Tools, such as *GIS Crash Analysis Tool (GCAT), Crash Analysis Module* (*CAM) Tool,* and *OSTATS Public Crash Dashboard* are the recommended data sources and crash-analysis tools from the Ohio State Highway Patrol (OSHP), despite the disclaimer "only confirmed crashes are in the Crash System" (OSHP, 2024).


ODOT Safety Study Process
Step 1: Collect Data and Diagnose Crash Patterns
Step 2: Identify Potential for Site Safety Improvements & Possible Countermeasures
Step 3: Perform Relevant Traffic Studies
Step 4: Evaluate Countermeasures
Step 5: Develop Plan and Finalize Report

(OSHP, 2024)

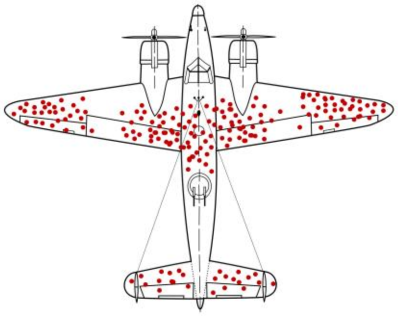


(Simplified current and proposed information pipeline of collision information prior to "Step 1")

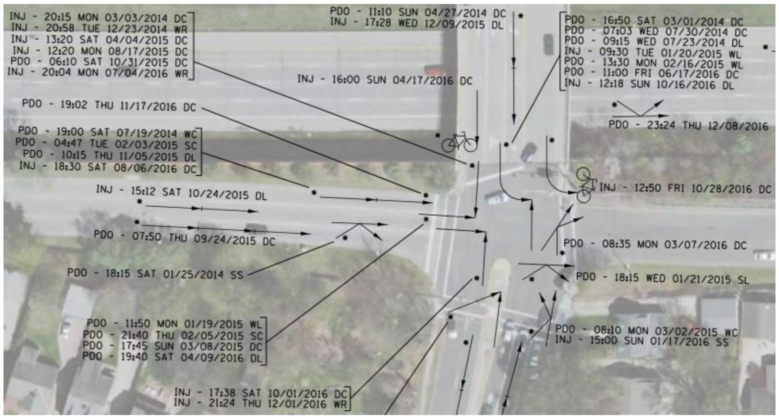
Survivorship Bias

The classic study by Abraham Wald '*Estimating Plane Vulnerability Based on Damage of Survivors*' demonstrated how survivorship bias can trick an audience into believing an unsafe assumption that data with specificity is also comprehensive (Wald, 1980). By counting the location and quantity of bullet holes in airplanes returning from combat, Wald points out observations were only being made of planes that were not shot down. The bias is almost comically obvious when one infers that the most harmful bullet holes are likely unobserved because the plane crashed. A similar bias is likely occurring in collision data but for low injury incidents.

City of Cleveland pedestrian or cyclist safety initiatives adhere to the OHSP recommendation and only use state-provided GCAT data in the *Vision Zero* and *Speed Table Program* (Cleveland, VZTF 2022; Cleveland, 2023). Because only OH-1data is used, this begs the question: "*are all crashes officially reported*?" The existence of under-reporting in collision data extends back more than 50 years with Scott and Carrol stating in 1971, "Highway accident data are seriously biased due to under-reporting" (Scott, 1971). It's curious then to see diagrams from modern safety studies commissioned by ODOT (Crawford, 2018) that do not mention, or seem to consider, underreporting in their analysis. Indeed, diagrams in these studies have an eerie similarity to diagrams of airplanes studied in Wald's analysis:



Survivorship Bias Illustration of WWII Bomber Plane (Moll, 2022)


INJ - 20:15 MON 03/03/2014 DC
INJ - 20:58 TUE 12/23/2014 WR
INJ - 13:20 SAT 04/04/2015 DC
INJ - 12:20 MON 08/17/2015 DC
PDO - 06:10 SAT 10/31/2015 DC
INJ - 20:04 MON 07/04/2016 WR
PDO - 11:10 SUN 04/27/2014 DC
INJ - 17:28 WED 12/09/2015 DL
PDO - 16:50 SAT 03/01/2014 DC
PDO - 07:03 WED 07/30/2014 DC
PDO - 09:15 WED 07/23/2014 DL
INJ - 09:30 TUE 01/20/2015 WL
PDO - 13:30 MON 02/16/2015 WL
PDO - 11:00 FRI 06/17/2016 DC
INJ - 12:18 SUN 10/16/2016 DL
INJ - 16:00 SUN 04/17/2016 DC
PDO - 19:02 THU 11/17/2016 DC
PDO - 23:24 THU 12/08/2016
PDO - 19:00 SAT 07/19/2014 WC
PDO - 04:47 TUE 02/03/2015 SC
PDO - 10:15 THU 11/05/2015 DL
INJ - 18:30 SAT 08/06/2016 DC
INJ - 12:50 FRI 10/28/2016 DC
INJ - 15:12 SAT 10/24/2015 DL
PDO - 07:50 THU 09/24/2015 DC
PDO - 08:35 MON 03/07/2016 DC
PDO - 18:15 SAT 01/25/2014 SS
PDO - 18:15 WED 01/21/2015 SL
PDO - 11:50 MON 01/19/2015 WL
PDO - 21:40 THU 02/05/2015 SC
PDO - 17:45 SUN 03/08/2015 DC
PDO - 19:40 SAT 04/09/2016 DL
PDO - 08:10 MON 03/02/2015 WC
INJ - 15:00 SUN 01/17/2016 SS
INJ - 17:38 SAT 10/01/2016 DC
INJ - 21:24 THU 12/01/2016 WR

HSIP Intersection Safety Study (Crawford, 2018)

Studies of collision underreporting postulate that underreporting has a negative relationship along two axes: income, and injury severity (Ahmed, 2019; Abay, 2015; Khadka, 2022; Scott, 1971). Therefore, as underreporting increases so does the probability of incorrect forecasting, erroneous attribution of crash contributing factors, and mis-prioritization of funding and safety initiatives (Ahmed, 2019). Said otherwise, state-recommended engineering methods applied by the City of Cleveland may be inadvertently discounting pedestrian and bicycle safety because records of these collisions do not 'survive' the data collection process.

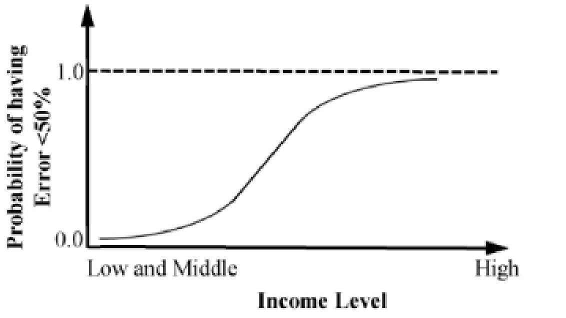


Fig. 4. The logistic relationship between error in reporting non-fatal accidents and income level of a country.

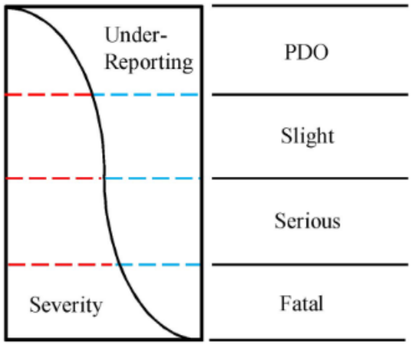


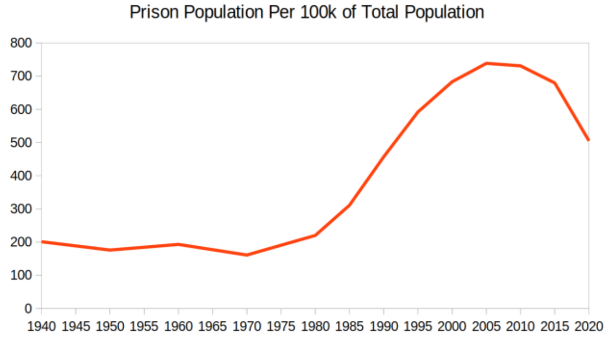
Fig. 5. Severity versus Under-Reporting Diagram.

(Ahmed, 2019)

Overcriminalization as Contributor of Selection Bias

While not the primary focus of this project, it is illustrative to explore how over-criminalization might interact with this collision data and discuss some cases observed in the current data.

In the early 1980s, Ronald Reagan proclaimed victims of crimes as "forgotten persons" (Reagan, 1981) and enacting a series of *tough-on-crime* legal and budgetary reforms resulting in a dramatic increase of incarceration from the mid-1980s through the mid-2000s.



(World Prison Brief, 2020)

This shift in the American legal system is referred to as "overcriminalization" and has been studied extensively by the legal scholar community. It may be summarized as the existence of too many laws applied too broadly (Smith, 2012). There is near universal agreement, among groups ranging from the American Civil Liberties Union (ACLU) and The Heritage Foundation, that overcriminalization has broad negative societal impacts (Dillon, 2012). One complicating factor of overcriminalization is a paradox that the criminal perpetrators of today were often victims of unrelated crimes in the past (Green, 2019). This victim-to-criminal dynamic develops into a practically, or morally, unworkable relationship between law enforcement and victims when victims are most vulnerable. Because OH-1 collision reporting relies on information provided to police, such as name and date of birth, collision parties may be distrustful of the police and choose to avoid providing information or interacting with the police in any way.

An example illustrating this is that of a 59-year-old woman riding a bicycle struck by a driver causing serious injury. The woman was unable to leave the scene due to her injuries. Responding officers discovered she had an outstanding warrant from an unrelated drug charge. Even though she was not at fault police promptly arrested her. An even more serious example of underreporting is that of a 65-year-old man who had a lengthy criminal record, once described by a judge as "a menace" after inflicting serious injury on a young girl while fleeing police in a stolen vehicle. He served six years in prison but after his release was reported as fatally wounded by the coroner as a pedestrian in another traffic collision. This may be a reporting error by the coroner but, for reasons unclear, the OH-1 regarding this fatal collision was never submitted to the state. (As of this writing, the author has submitted public record requests to multiple agencies with no reply.) It's unclear how much underreporting is driven by individuals and how much by reporting agencies. The state requires OH-1 to be submitted within five days of creation of a report (ORC, 5502), but examples of late reporting, non-reporting, and misreporting are observed in OH-1 data are common, even for fatalities. To quote Hauer "the problem of dealing with a rubbery yardstick is further complicated by the fact that not all that is reportable is reported" (Hauer, 1988).

After police and thoughtful consideration are rejected as a response to collisions, the systematic recourse offered through Crash System Tools and Vision Zero is potentially replaced with unsystematic and problematic behavior. There are troubling examples in this data of dispatches reading something like, "Juvenile struck by vehicle. Shots fired. EMS standing by for police to clear scene. All subjects gone on arrival." Even in these cases, no OH-1 report is created and the root cause of the seminal incident is lost in the data. Whatever the root cause of underreporting, situations of perpetrator-to-victim cycles shows how deeply entrenched traffic violence can be on American roads and the potential for spiraling severity of underreporting in over-criminalized areas.

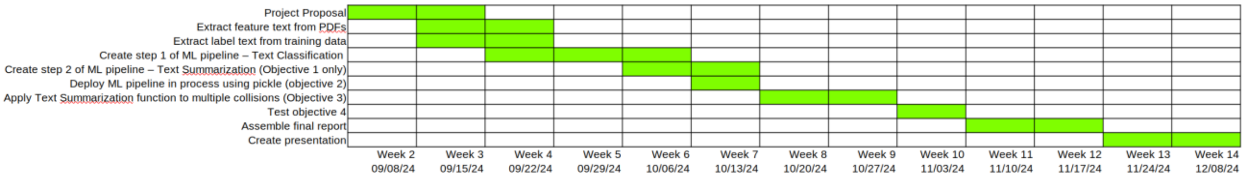
Data Solutions

It may be assumed that planners and decision makers do not want a community to be victimized repeatedly in the same manner because some evidence exists for consideration of economically disadvantaged areas. The data solution to this problem can be inferred from Appendix C of Cleveland's Vision Zero Action Plan, which describes the summarization of collision data. This process provides for an adjustment factor to counteract the effects in economically disadvantaged areas (Cleveland VZTC, 2022).

While it is possible this is sufficient to counterbalance underreporting, a blanket factor is not without problems because it has the possibility to still allow blind spots of safety issues. For example, if two areas have the same economic adjustment factor but one area has greater underreporting then valuable infrastructure money may be applied in the area with less actual need. If planning initiatives rely mainly on comprehensive officer-generated reports, well-intentioned conclusions may fail to address traffic safety issues in over-criminalized areas.

Revisiting Ronald Reagan's concern that "lack of concern for victims compounds failure [of our system]." The current project intends to create a solution to fill a 50-year gap in collision reporting. There have been gains made in adding structure to official reporting, like the OH-1, but it's evident further gains are needed. It's important for a community to be honest about traffic violence and reduce the actual and perceived risk of walking and cycling. Even after adjusting relative safety risk for underreporting, the health benefits of cycling outweigh the public health impacts of a collision injuries, but the perceived individual risk is disproportionally high and suppresses individual cycling activity (Götschi, et al., 2016).

9. Project Schedule



	Week 2 09/08/24	Week 3 09/15/24	Week 4 09/22/24	Week 5 09/29/24	Week 6 10/06/24	Week 7 10/13/24	Week 8 10/20/24	Week 9 10/27/24	Week 10 11/03/24	Week 11 11/10/24	Week 12 11/17/24	Week 13 11/24/24	Week 14 12/08/24
Project Proposal	🟩	🟩											
Extract feature text from PDFs		🟩	🟩										
Extract label text from training data		🟩	🟩										
Create step 1 of ML pipeline – Text Classification			🟩	🟩	🟩								
Create step 2 of ML pipeline – Text Summarization (Objective 1 only)					🟩	🟩							
Deploy ML pipeline in process using pickle (objective 2)						🟩							
Apply Text Summarization function to multiple collisions (Objective 3)							🟩	🟩					
Test objective 4									🟩				
Assemble final report										🟩	🟩		
Create presentation												🟩	🟩

10. Required Technologies

This project will be developed in Python using

For PDF processing and OCR:

- pdf2image
- Tensorflow.keras.preprocessing.image
- pytesseract
- PIL

For general data manipulation and preprocessing:

- pandas
- numpy
- matplotlib
- re

For general and enhanced computing:

- concurrent
- collections
- multiprocessing
- operator
- datasets (ThreadPoolExecuter)

For neural network creation and NLP processing:

- PyTorch
- Transformers (Llama)
- nltk

11. References

Abay, Kibrom A. (2015). Investigating the nature and impact of reporting bias in road crash data. *Transportation Research Part A: Policy and Practice*, *71*, 31–45. https://doi.org/10.1016/j.tra.2014.11.002

Ahmed, A., Sadullah, A. F. M., & Yahya, A. S. (2019). Errors in accident data, its types, causes and methods of rectification-analysis of the literature. *Accident Analysis & Prevention*, *130*, 3–21. https://doi.org/10.1016/j.aap.2017.07.018

Burgess & Niple. (2021). *CUY - Cleveland PSIP Contract Proposal - 210192 PID 113330*. Ohio Department of Transportation. https://contracts.dot.state.oh.us/common/searchAPI.do;jsessionid=c064FkjMF37QSwuiCcPlo8Azf8fS7ZAjgJAJnvbW.dotidpxep02?fetchCurrent=false&PID_NUM=113330&cabinetId=1002

Cleveland, City of. (2023). *2022 Speed Table Pilot Program Evaluation and Key Recommendations*. https://www.clevelandohio.gov/sites/clevelandohio/files/traffic/SpeedTablePilotEvaluationReport0523.pdf

Cleveland, Vision Zero Task Force of. (2022). *Vision Zero Action Plan*. https://www.visionzerocle.org/pages/action-plan

Crawford, Murphy &. Tilly. (2018). *CUY-90-9.09 SAFETY STUDY*. https://ftp.dot.state.oh.us/pub/Districts/D12/Production/Consultant_Programmatic/2023_January/116831/CUY-90-9.09%20Safety%20Study%202018.06.30.pdf

Doggett, S., Ragland, D. R., & Felschundneff, G. (2018). Evaluating Research on Data Linkage to Assess Underreporting of Pedestrian and Bicyclist Injury in Police Crash Data. *UC Berkeley: Safe Transportation Research & Education Center*. https://escholarship.org/uc/item/0jq5h6f5

Götschi, T., Garrard, J., & Giles-Corti, B. (2016). Cycling as a Part of Daily Life: A Review of Health Perspectives. *Transport Reviews*, *36*(1), 45–71. https://doi.org/10.1080/01441647.2015.1057877

Hauer, E., & Hakkert, A. S. (1988). Extent and Some Implications of Incomplete Accident Reporting. *Transportation Research Record*, *1185*. http://onlinepubs.trb.org/Onlinepubs/trr/1988/1185/1185-001.pdf

Khadka, Anish, Parkin, John, Pilkington, Paul, Joshi, Sunil Kumar, & Mytton, Julie. (2022). Completeness of police reporting of traffic crashes in Nepal: Evaluation using a community crash recording system. *Traffic Injury Prevention*, *23*(2), 79–84. https://doi.org/10.1080/15389588.2021.2012766

Moll, Cameron. (2022). *Abraham Wald and the airplane diagram with red bullet holes – here's the origin story*. https://www.cameronmoll.com/journal/abraham-wald-red-bullet-holes-origin-story

Naumann, Rebecca B., Heiny, Stephen, Evenson, Kelly R., LaJeunesse, Seth, Cooper, Jill F., Doggett, Sarah, & Marshall, Stephen W. (2019). Organizational networks in road safety: Case studies of U.S. Vision Zero cities. *Traffic Injury Prevention*, *20*(4), 378–385. https://doi.org/10.1080/15389588.2019.1587752

ODOT. (2022). *Safety Analysis Guidelines*. https://www.transportation.ohio.gov/programs/highway+safety/highway-safety-manual-guidance/safety-analysis-guidelines-cf/search

ODOT. (2023). *Ohio Highway Safety Improvement Program 2023 Annual Report*. https://highways.dot.gov/sites/fhwa.dot.gov/files/2024-04/HISP%28Ohio%29%202023%20Report.pdf

OSHP Statistical Analysis Unit. (2024). *Ohio Statistics and Analytics for Traffic Safety User Guide*. Ohio State Highway Patrol. https://dam.assets.ohio.gov/image/upload/statepatrol.ohio.gov/links/OSTATS_PublicUserGuide_2024.pdf

ORC. §5502.11 (2011). *Written Report of Motor Vehicle Accident*. https://codes.ohio.gov/ohio-revised-code/section-5502.11

Scott, Robert, & Carroll, Philip. (1971). Acquisition of Information on Exposure and on Non-Fatal Crashes - Accident Data Inaccuracies Vol. II. *Highway Safety Research Institute, The University of Michigan Ann Arbor, Michigan*. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/1054/15848.0001.001.pdf

Wald, Abraham. (1980). Reprint of Estimating Plane Vulnerability Based on Damage of Survivors. *Center for Naval Analysis*. https://lru.praxis.dk/Lru/microsites/virksomhederiundervisningen/novo_materiale/9-A%20method%20of%20estimating%20plane%20vulnerability%20....pdf