# ISIS End of Summer Showcase

## Project: Data Pipeline and Feature Selection Module

Hunter Baxter

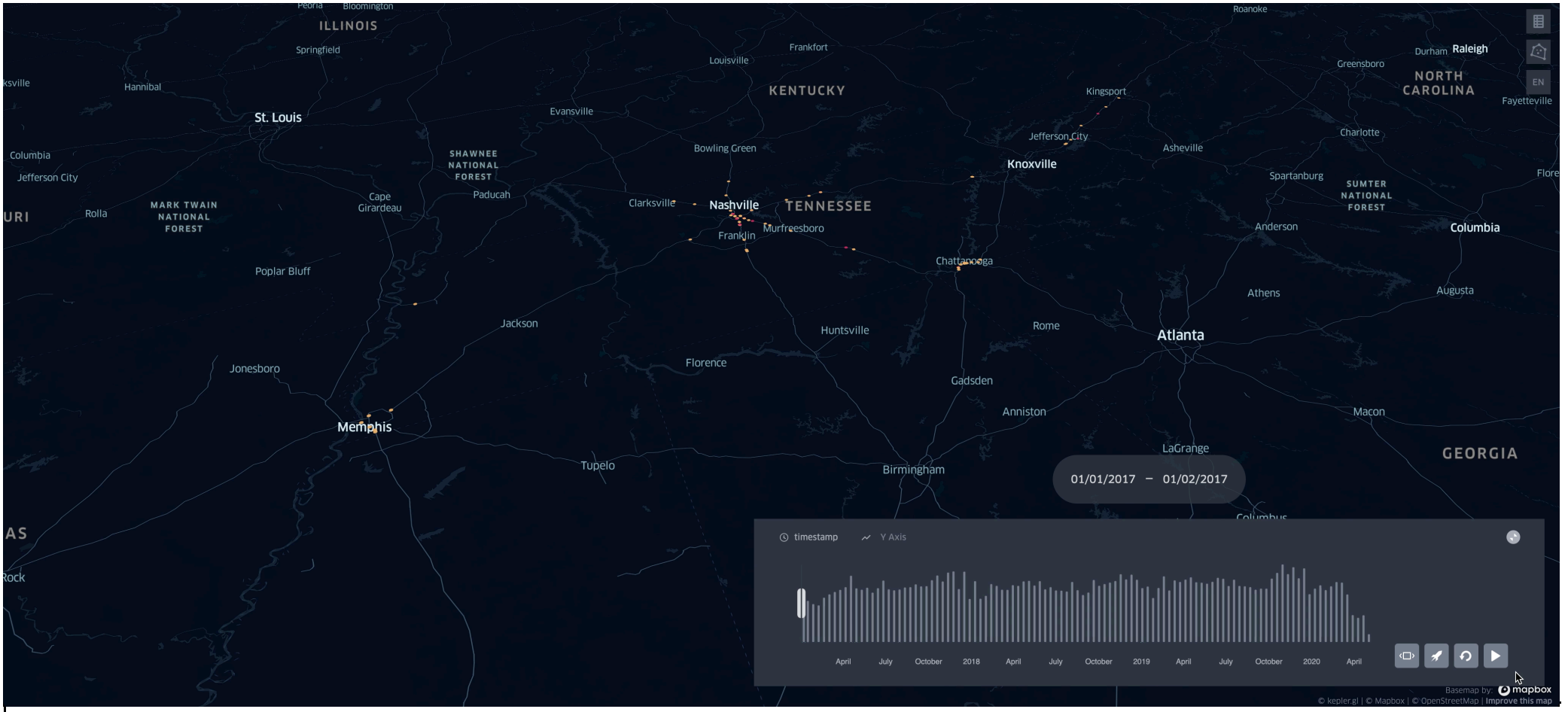
Mentors:

Abhishek Dubey

Sayyed Mohsen Vazirizade

# Introduction

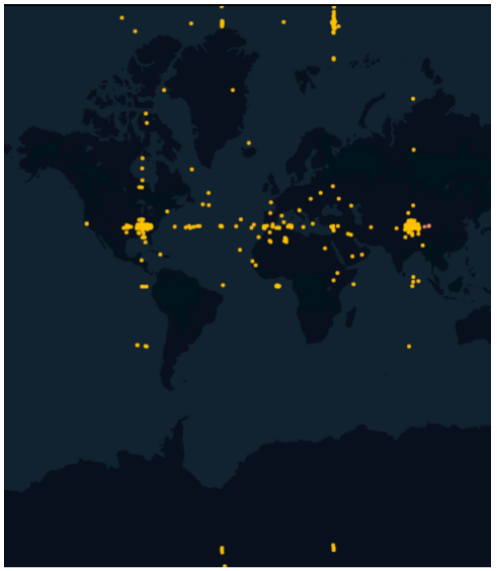
## Emergency Response Management




The big challenge for TDOT and other emergency responders is that they need to manage incidents spread across a large area with limited resources. The faster an incident is cleared, the less congestion and the better the recovery. To reach this goal, we collect various datasets (features) for long periods of time over a large area. The size of this data poses challenges.

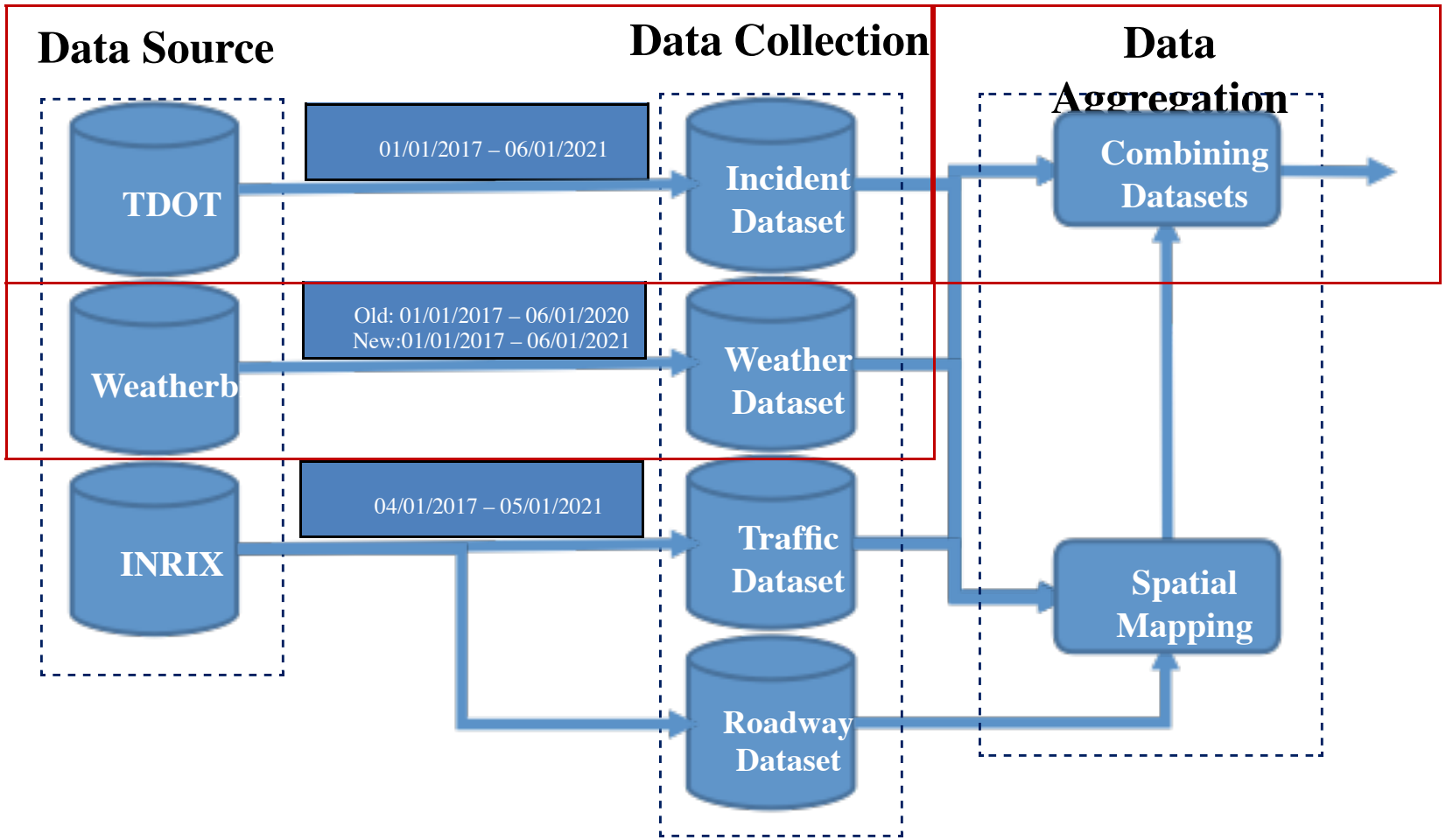
# Data

- Four data sets
  - Tennessee Department of Transportation automobile crash data
  - INRIX traffic and road segment data
  - Weatherbit weather data

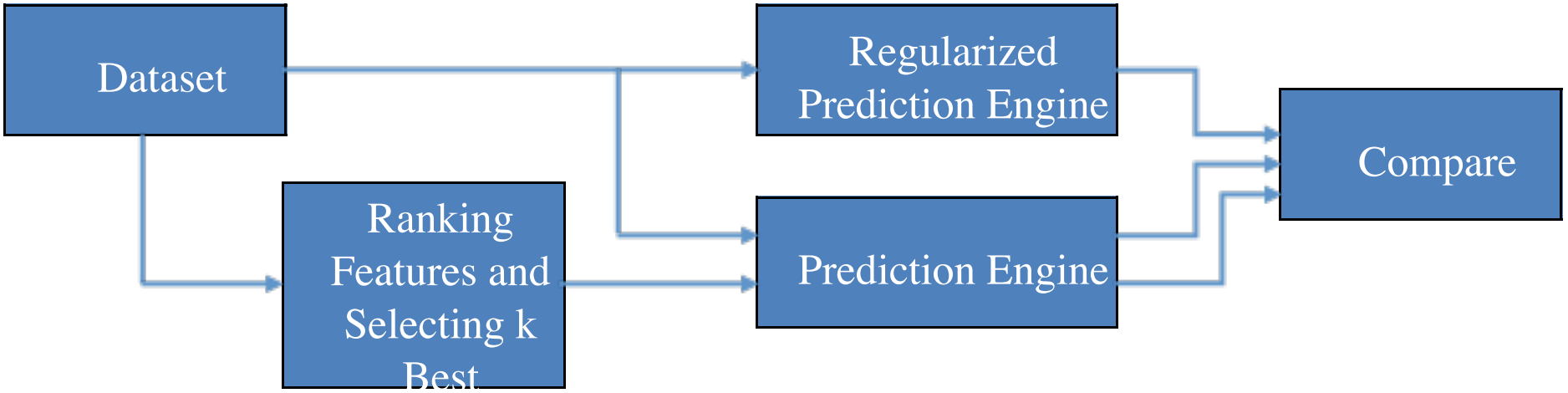


| Year | Erroneous % | Both Lat and Long Null % | Just Lat Null % | Just Long Null % | Inverse Lat and Long Count | Total Count | Both Lat and Long Null Count | Just Lat Null Count | Just Long Null Count | Out of TN Count | Zero Count | Just Inverse Lat Count | Just Inverse Long Count | Point In TN Count | Flipped Lat and Long Order Count | Flipped Lat and Long Order Inverse Lat and Long Count | Flipped Lat and Long Order Just Inverse Lat Count | Flipped Lat and Long Order Just Inverse Long Count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2002 | 99.95 | 99.92 | 0.00 | 0.01 | 0 | 115705 | 115615 | 2 | 13 | 13 | 2 | 0 | 41 | 19 | 0 | 0 | 0 | 0 |
| 2003 | 99.98 | 99.93 | 0.01 | 0.03 | 0 | 142263 | 142165 | 10 | 36 | 16 | 6 | 0 | 24 | 6 | 0 | 0 | 0 | 0 |
| 2004 | 100.00 | 99.99 | 0.00 | 0.00 | 0 | 159587 | 159569 | 4 | 4 | 7 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 |
| 2005 | 99.94 | 99.92 | 0.00 | 0.01 | 0 | 159811 | 159676 | 4 | 14 | 15 | 0 | 0 | 57 | 0 | 0 | 0 | 45 | 0 |
| 2006 | 99.64 | 99.61 | 0.01 | 0.01 | 0 | 151819 | 151230 | 11 | 16 | 23 | 0 | 0 | 411 | 22 | 0 | 0 | 106 | 0 |
| 2007 | 98.17 | 98.02 | 0.01 | 0.00 | 0 | 151061 | 148072 | 8 | 7 | 209 | 0 | 0 | 179 | 2444 | 27 | 0 | 113 | 2 |
| 2008 | 96.78 | 96.64 | 0.01 | 0.01 | 1 | 137668 | 133041 | 9 | 7 | 184 | 0 | 1 | 2604 | 1626 | 19 | 0 | 176 | 0 |
| 2009 | 96.00 | 95.66 | 0.01 | 0.01 | 0 | 139788 | 133722 | 14 | 11 | 114 | 332 | 1 | 4655 | 733 | 0 | 3 | 203 | 0 |
| 2010 | 92.11 | 76.38 | 0.01 | 0.01 | 0 | 157589 | 120362 | 8 | 14 | 177 | 24602 | 0 | 2898 | 9428 | 0 | 19 | 80 | 1 |
| 2011 | 83.40 | 26.93 | 0.00 | 0.01 | 0 | 173134 | 46631 | 4 | 9 | 125 | 97627 | 0 | 341 | 28376 | 0 | 10 | 11 | 0 |
| 2012 | 46.48 | 17.20 | 0.00 | 0.00 | 0 | 182689 | 31423 | 2 | 5 | 138 | 53354 | 2 | 66 | 97667 | 0 | 23 | 6 | 3 |
| 2013 | 39.37 | 14.98 | 0.00 | 0.00 | 0 | 182358 | 27325 | 1 | 2 | 506 | 43963 | 0 | 0 | 110547 | 0 | 13 | 0 | 1 |
| 2014 | 23.41 | 6.16 | 0.00 | 0.00 | 0 | 187791 | 11567 | 3 | 1 | 3450 | 28946 | 0 | 3 | 143820 | 0 | 1 | 0 | 0 |
| 2015 | 2.67 | 0.00 | 0.00 | 0.00 | 0 | 207500 | 6 | 0 | 0 | 0 | 5529 | 0 | 812 | 201153 | 0 | 0 | 0 | 0 |
| 2016 | 0.05 | 0.00 | 0.00 | 0.00 | 0 | 206546 | 9 | 0 | 0 | 0 | 102 | 0 | 2966 | 203469 | 0 | 0 | 0 | 0 |
| 2017 | 6.00 | 6.00 | 0.00 | 0.00 | 0 | 209899 | 12597 | 0 | 0 | 0 | 3 | 0 | 886 | 196413 | 0 | 0 | 0 | 0 |
| 2018 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 207853 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 207853 | 0 | 0 | 0 | 0 |
| 2019 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 214122 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 214122 | 0 | 0 | 0 | 0 |
| 2020 | 0.01 | 0.01 | 0.00 | 0.00 | 0 | 180685 | 14 | 1 | 0 | 0 | 0 | 0 | 0 | 180670 | 0 | 0 | 0 | 0 |
| 2021 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 74153 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 74151 | 0 | 0 | 0 | 0 |

# Data Pipeline



# Feature Selection




$$Importance(X_i) = Relevance(X_i, Y) - \sum_{s \in S} Redundancy(X_i, X_s)$$

# Internship Commentary

- What went well
  - Developed practical solutions for an industry partner
  - Learned tools for deployable solutions to Big Data problems
- Challenges
  - Eating a healthy lunch daily
- Lessons Learned
  - It is better to work slower and build a better solution than to work faster and need to debug or refactor later

# Conclusion

- There is intent to publish methodology and results of feature selection module in IEEE Big Data 2021 Conference
- I will be continuing research in the Fall 2021 semester for research credit hours