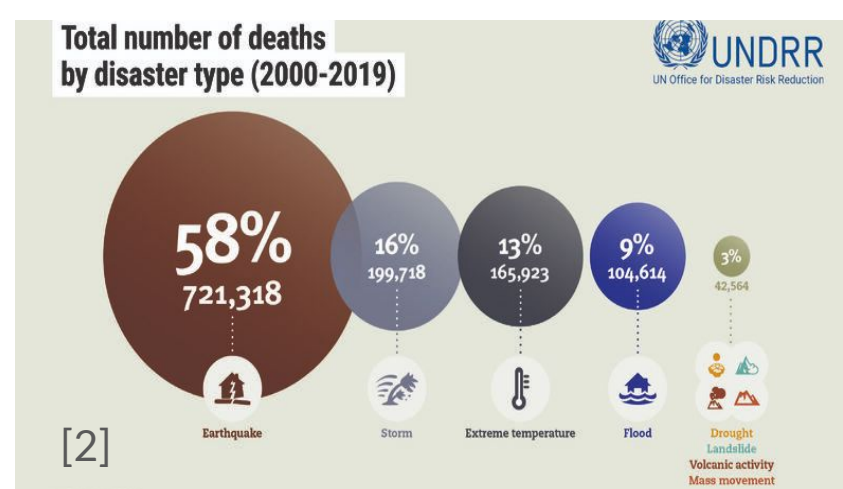


Introduction

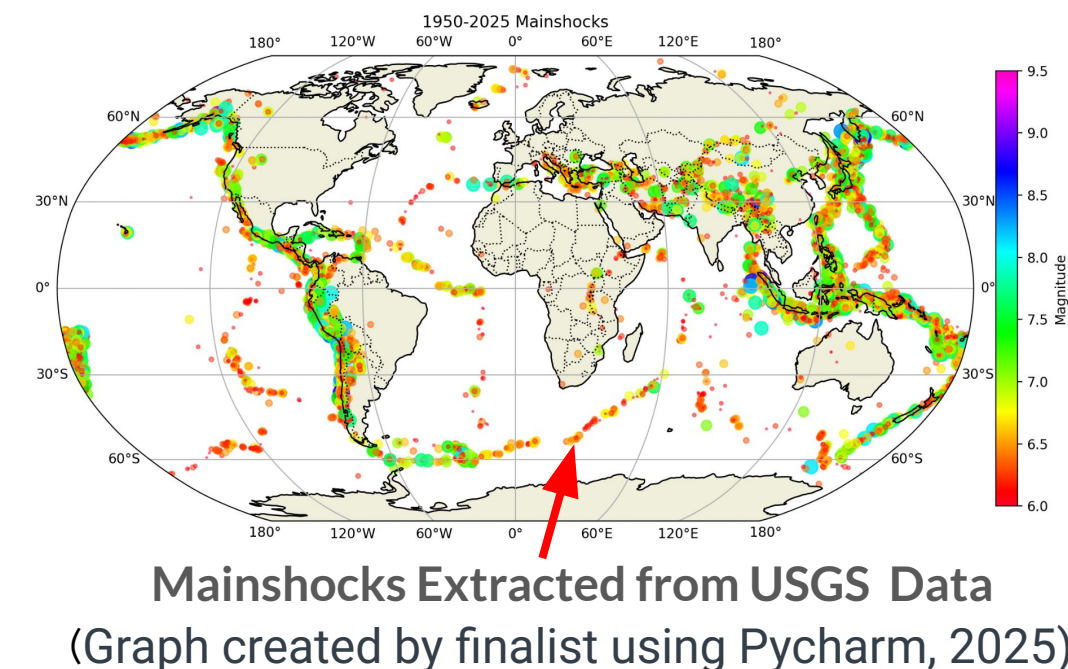
Earthquakes may just seem to be shocks that make buildings shake, but they can be deadly. From 2000 to 2019, only 8% of natural disasters were earthquakes, but they were 58% of total fatalities [1]! I live only a few miles away from a major active fault, the San Andreas Fault. Living so close to an active fault made me wonder what happens after a major earthquake. Major mainshocks are hard to predict, but aftershocks might be. After a major earthquake, aftershocks can cause additional damage while people are still recovering. It is critical if we can accurately forecast aftershocks.



[1]<https://disasterphilanthropy.org/cdp-resource/earthquakes/>
[2]https://www.linkedin.com/posts/undrr_earthquakes-are-the-most-deadly-type-of-disaster-activity-6851034509161451520-YHdt/
[3]<https://www.reuters.com/pictures/pictures-powerful-earthquake-strikes-southeast-asia-2025-03-28/>

USGS Earthquake Query

- Experiment Setup:**
 - Download 75 years of earthquake data from the USGS
 - Organize and classify data
 - Plot mainshocks over a world map
- Analysis of Results:**
Recorded earthquakes happen along the boundary of tectonic plates, matching my expectation.



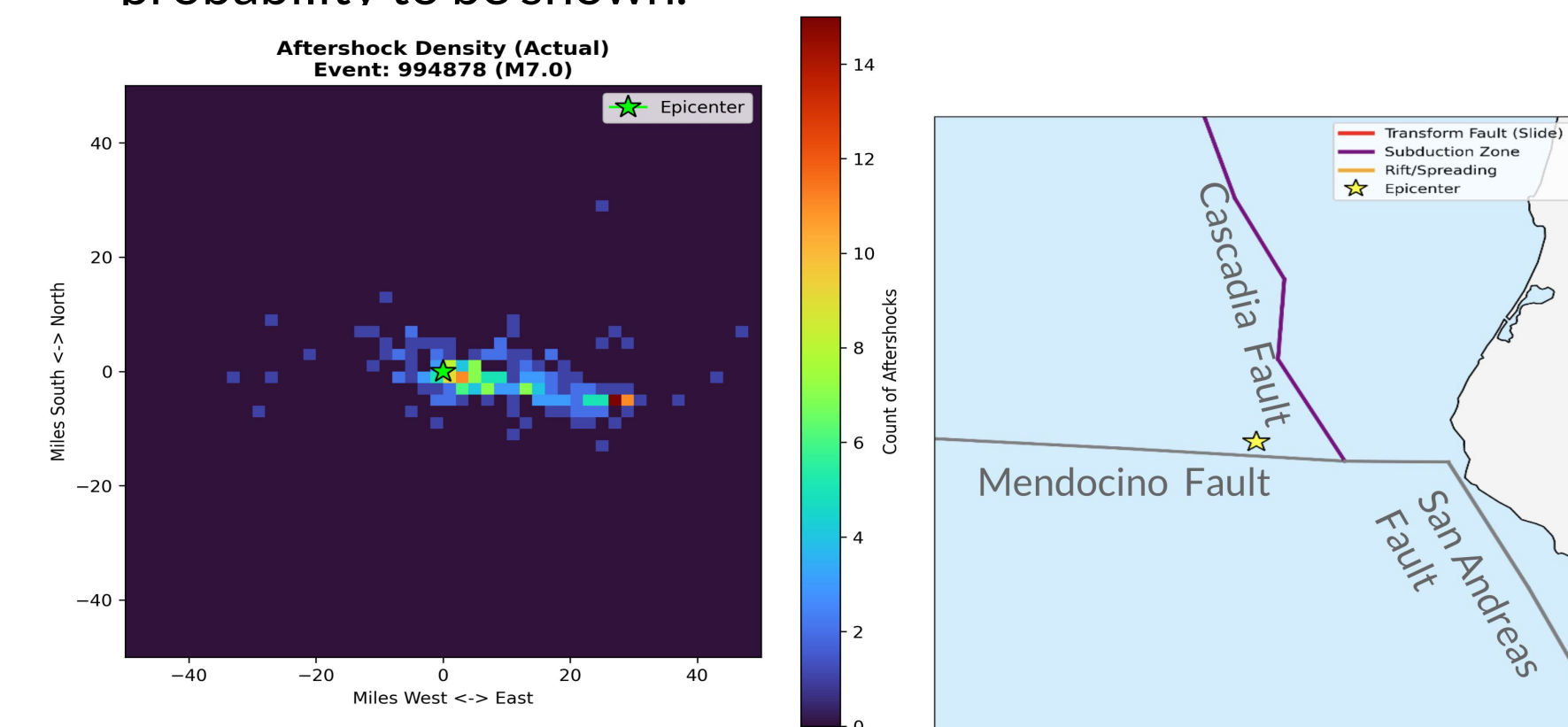
Mainshocks Extracted from USGS Data
(Graph created by finalist using Pycharm, 2025)

Spatial Distribution

"Which areas are at the highest risk?"

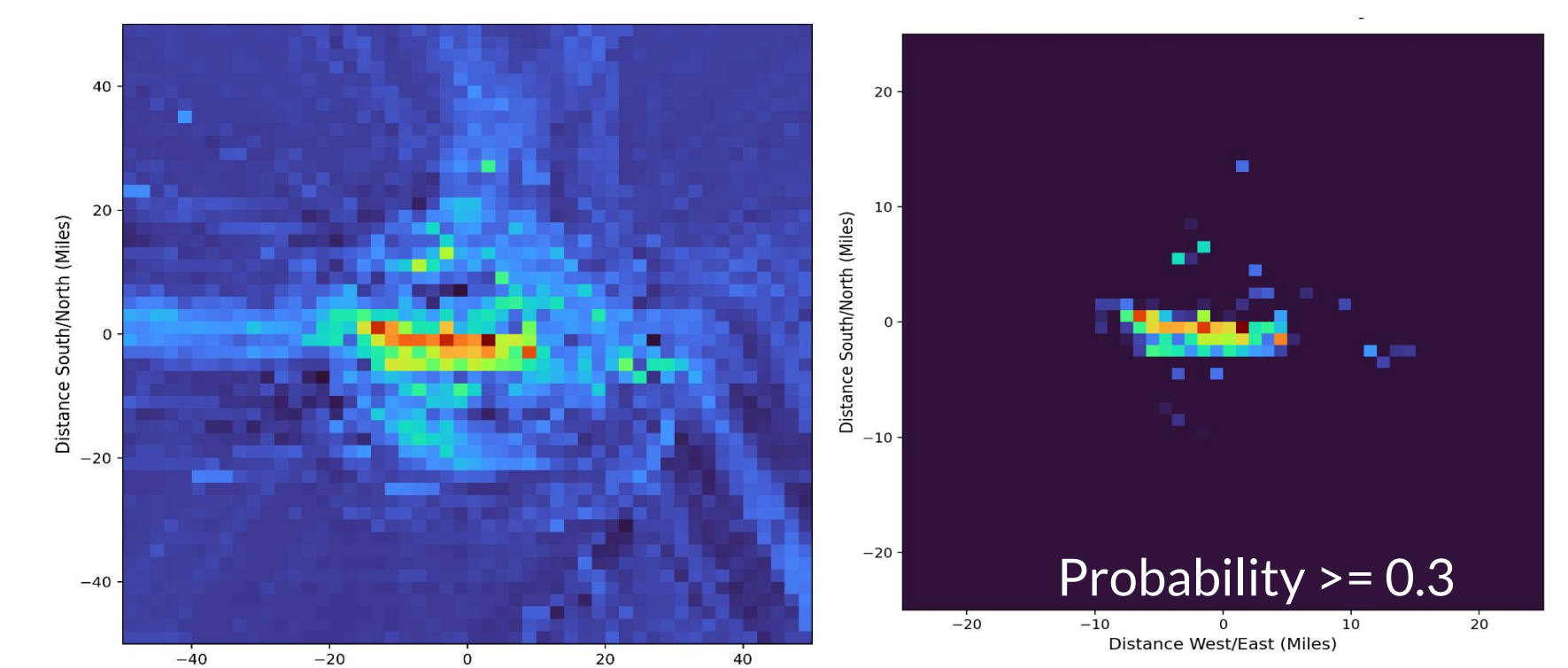
Forecast Likely Locations of Aftershocks

- New Feature:** Distance to nearest major fault
- Label:** Generate a 100 x 100 mile box around the epicenter of the earthquake and split it into a 50x50 grid, meaning each grid cell represents an area of 2x2 square miles.
- Case Study:** Dec 5, 2024, M7.0 Cape Mendocino Earthquake
I compared the forecasted probability heatmap with the recorded historical data. Findings are shown below.
 - The fault awareness (b) is clearly visible in actual historical data (a) as well as my forecasted data (c). The new feature added took an effect!*
 - Locations along the fault and close to the epicenter have a higher chance of aftershocks than other areas, similar to what I expected.
 - The forecasted aftershock location pattern is very close to the actual one.* Only showing the high-probability locations helps the forecasted result seem more similar to the actual locations.
 - There are some artificial effects can be seen in the heatmap for all probabilities. These effects can be removed by limiting the range of probability to be shown.



(a) Actual Aftershocks

(b) Fault Map



(c) Forecasted Aftershock Heatmap

(Graphs and images created by finalist using Pycharm, 2025)

Aftershock Oracle

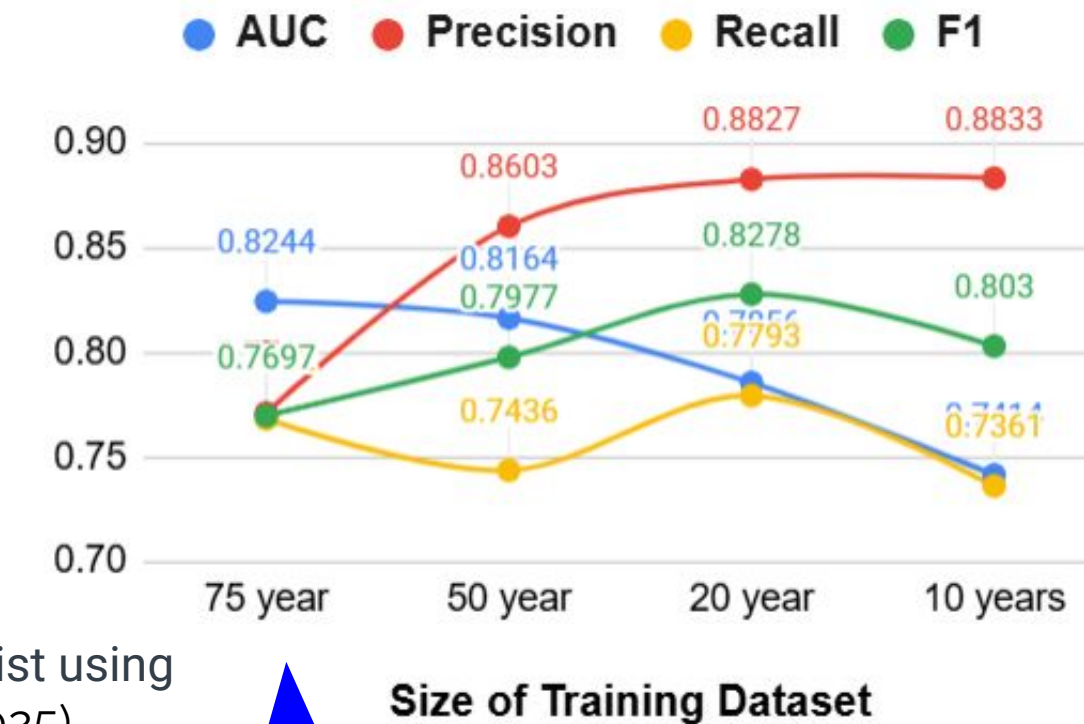
Machine Learning-Based Aftershock Risk Forecasting for Earthquake Safety

(Logo created by finalist using ChatGPT, November 28, 2025)



Q: Does larger dataset perform better?

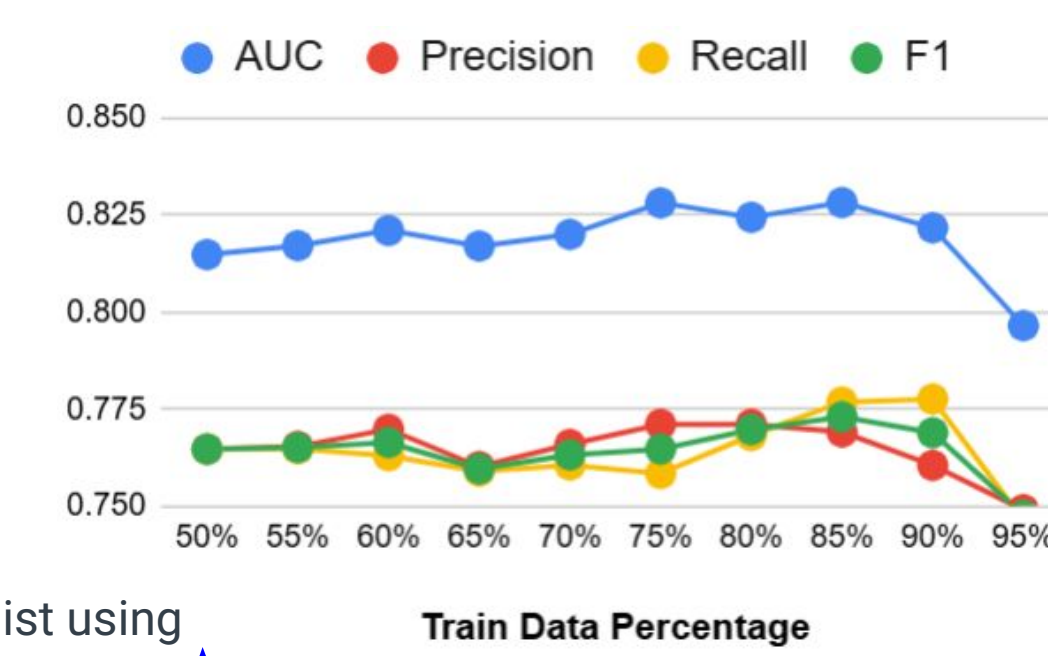
- Experiment Setup:**
Train 4 datasets of different length, 10/20/50/75 years
- Analysis of Results:** *More examples => better ranking (higher AUC).*



(Graph created by finalist using Google Sheets, 2025)

Q: What is the ideal split ratio?

- Experiment Setup:**
Split training data size from 50% to 95% in 5% steps
- Analysis of Results:** *85% is the best ratio overall*



(Graph created by finalist using Google Sheets, 2025)

Machine Learning Model Questions: How to select parameters to help model perform?

Aftershock Forecasting Questions: What do people care about after earthquakes occur?

XGBoost Models

Mainshock File

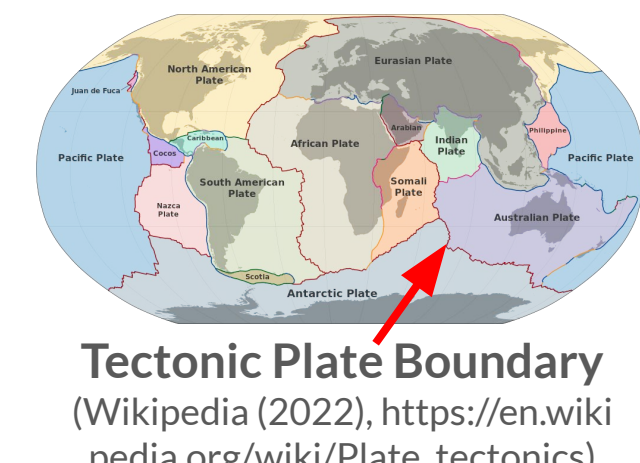
- Located mainshocks
 - Mag > 6
 - Max within 30 days

Aftershock Files

- Identify the aftershocks
 - Mag > 2.5
 - Within 30 days
 - 50 miles from mainshock

Features

- Magnitude
- Latitude
- Longitude
- Depth
- Distance to Nearest Fault



Tectonic Plate Boundary
(Wikipedia (2022), https://en.wikipedia.org/wiki/Plate_tectonics)

Labels

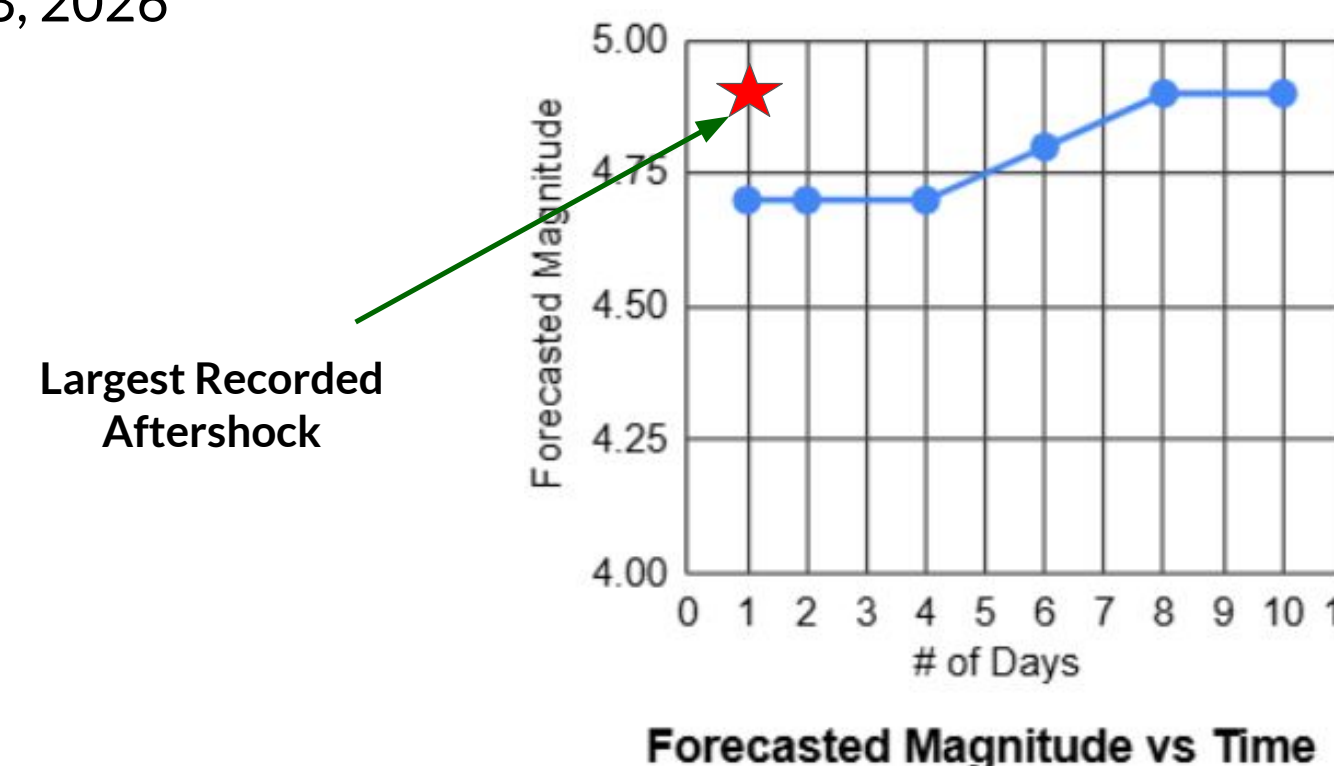
- Magnitude of Aftershocks
- Number of Aftershocks
- Location of Aftershocks

Magnitude Forecasting

"How intense will aftershocks be?"

Forecast Largest Magnitude Over Whole Area

- Label:** the magnitude of the maximum aftershock on the Nth day.
- Case Study:** One M6.2 earthquake occurred 128 km south of Vilyuchinsk, Russia, on Jan 13, 2026

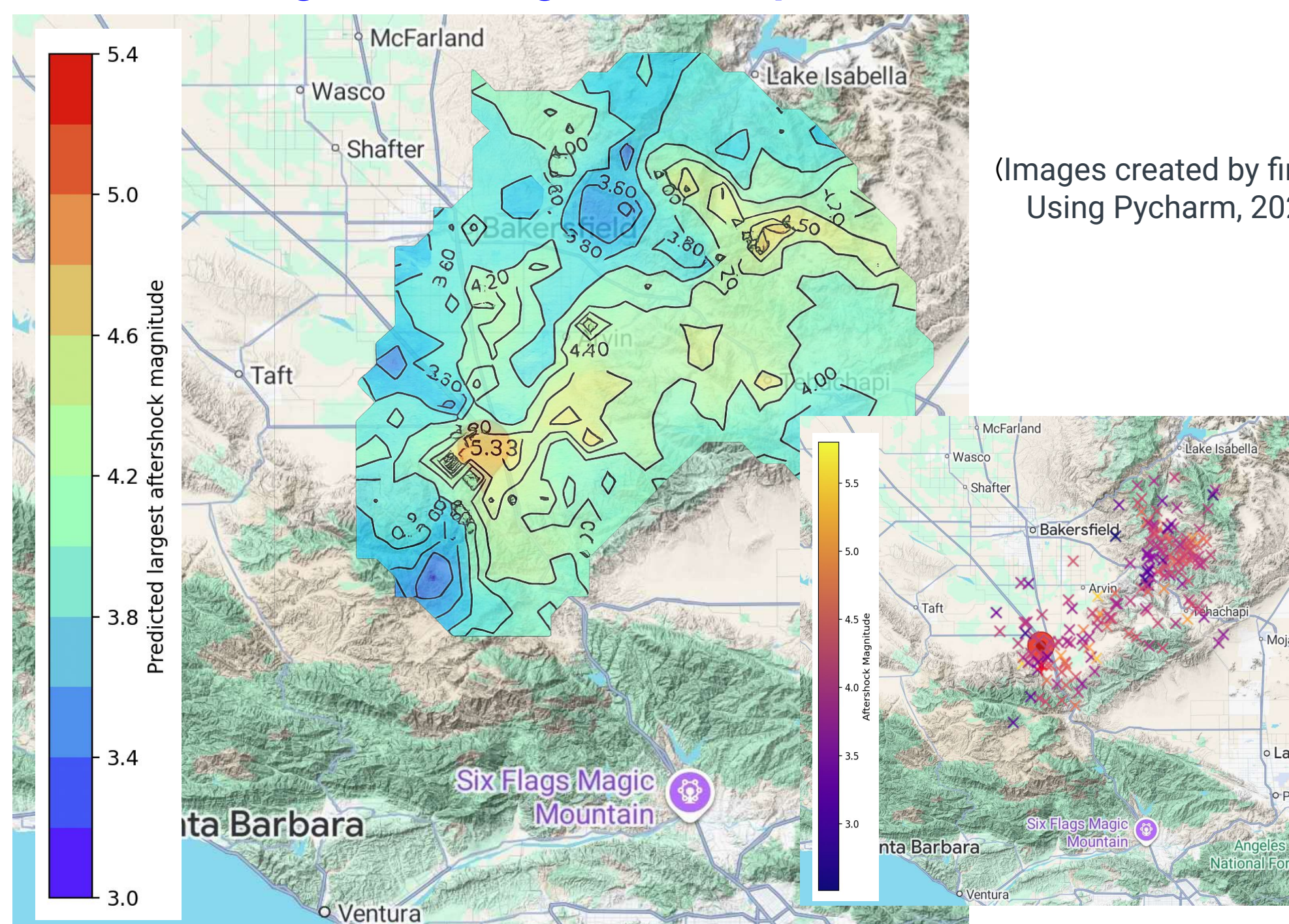


(Graphs created by finalist using Pycharm, 2026)

Forecast Largest Magnitude Distribution

- New Feature:** Distance to nearest major fault
- Label:** The magnitude of the maximum aftershock on the Nth day calculated separately in all 50x50 grids.
- Case Study:** A prominent earthquake in California History, on July 21, 1952, the M7.5 earthquake occurred 23 miles south of Bakersfield, California.

Both magnitude range and hotspots match!



(a) Contour diagram of magnitude distribution

(b) Historical records of aftershocks within 30 days

(Images created by finalist Using Pycharm, 2026)

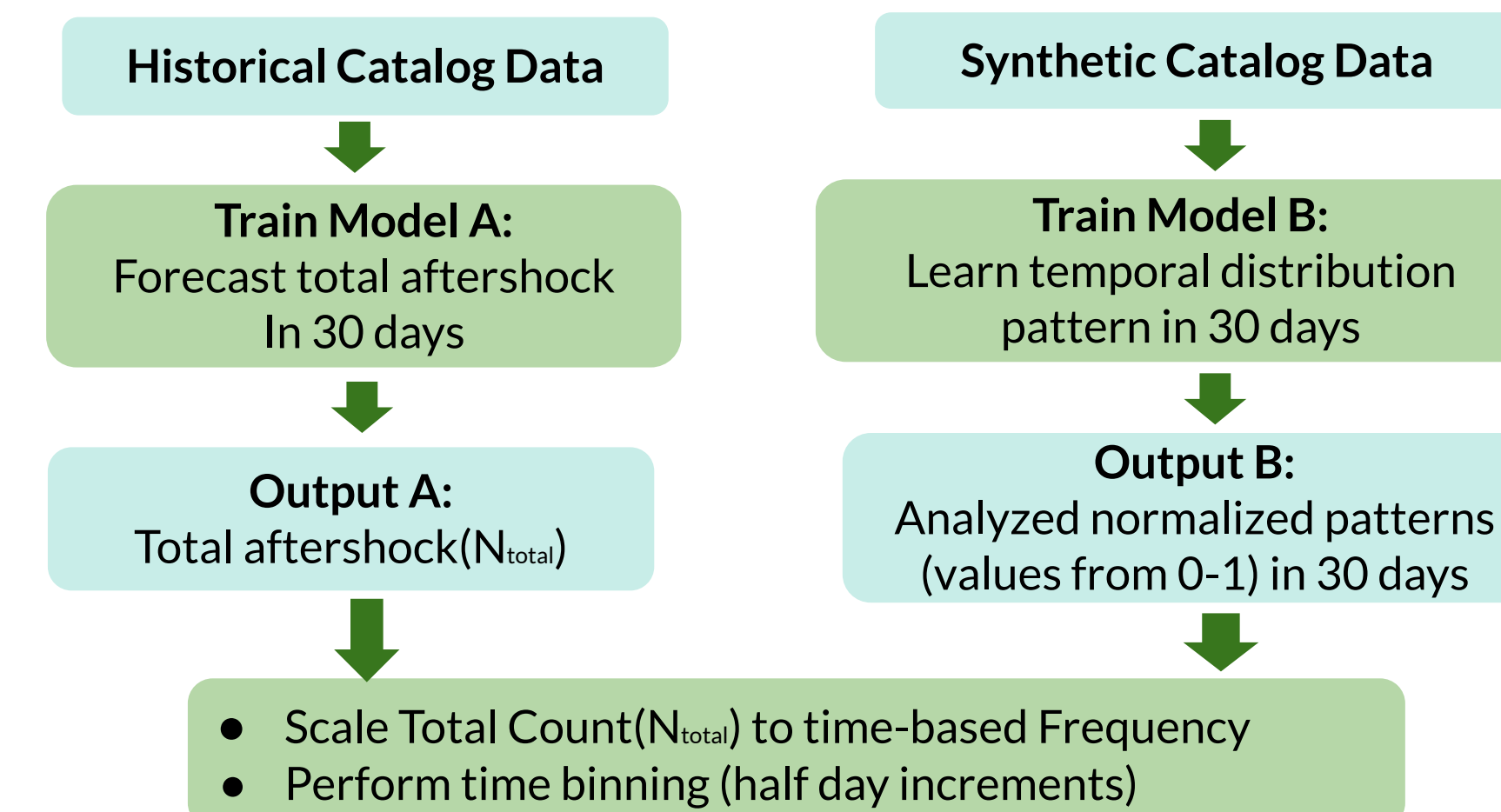
Temporal Frequency

"How often will aftershocks occur?"

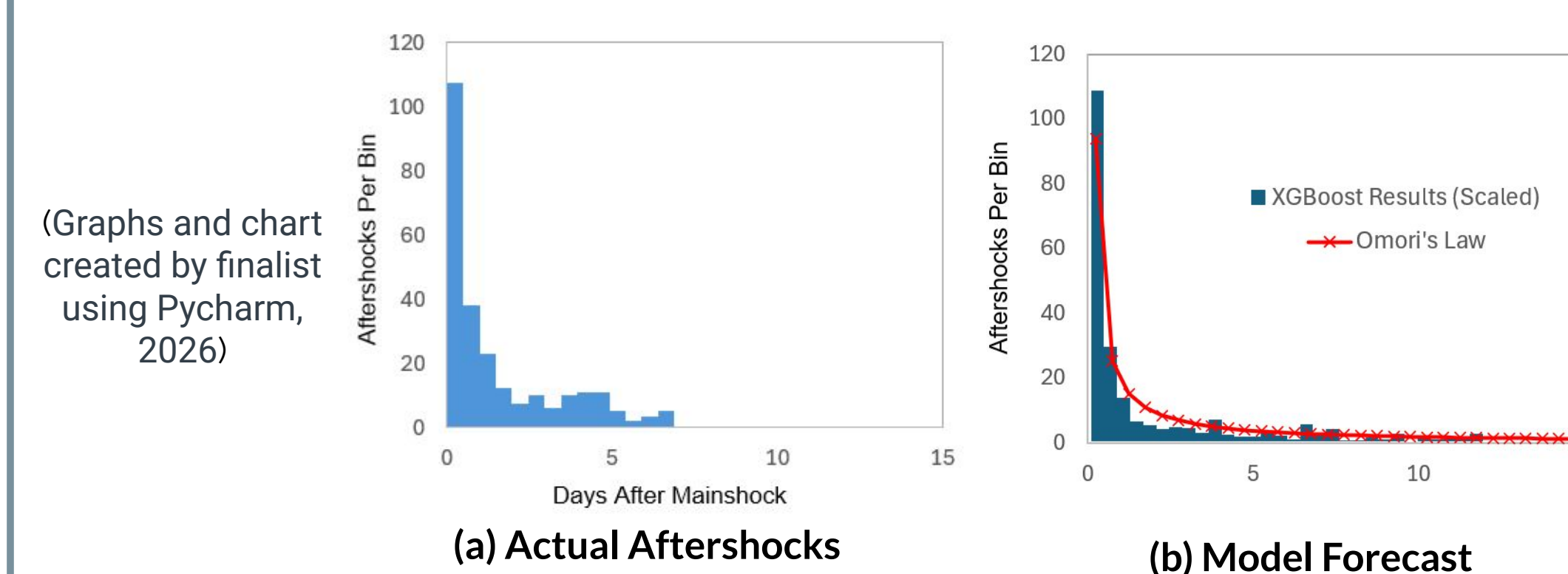
Synthetic Catalog: Human-generated earthquake data!

Questions	Answers
Purpose	Allows users to simulate longer years of data and minimize human-errors or inaccurate data in real historical catalogs, such as USGS. I used a synthetic catalog to find foundational empirical laws for prediction.
Creation	Computer generates it by algorithms (Statistical Models) or real-world simulations (Physics-Based Models).
Which Laws?	Omori's Law (frequency over time) and the Gutenberg-Richter Law (magnitude distribution).
Size of my Synthetic Catalog?	2000 mainshocks, each with up to 300 aftershocks.

Forecast Total Number of Aftershocks Over Time



- Case Study:** 2024 Cape Mendocino Quake
The time-based aftershock frequency forecasting curve (b) follows Omori's law and matches the aftershock frequency over time from the actual historical catalog (a) very well.



(a) Actual Aftershocks

(b) Model Forecast

Q: Do recent years have better results?

- Experiment Setup:**
 - Training data are from 4 different year periods (see below)
 - Test data from 2020-2025, without overlapping training data
- Analysis of Results:**
Model accuracy is not time-dependent to training data (Table created by finalist using Google Sheets, 2026)

Training Years	AUC	Precision	Recall	F1
1950-1969	0.7349	0.9189	0.5941	0.7456
1970-1989	0.7284	0.8886	0.5934	0.7116
1990-2009	0.7767	0.8970	0.6782	0.7724
2010-2019	0.7627	0.8763	0.7232	0.7924

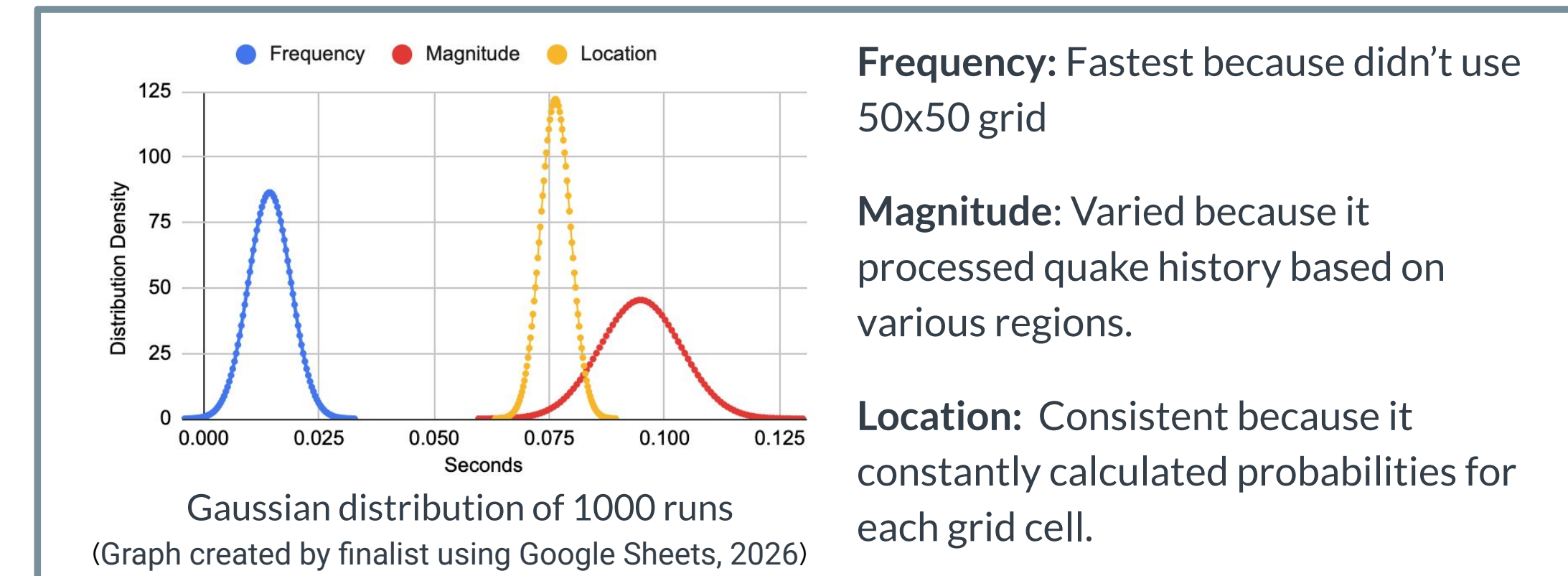
Machine Learning Model Selection

Model	What it is	Pros	Cons
Logistic Regression	Logistic classifier	Simple, fast, easy to explain	Misses complex nonlinear patterns
Random Forest	Many trees (bagging; training independently)	Handles nonlinear data well, robust	Can be slower/bigger; probs less reliable
XGBoost	Boosted (boosting: training one at a time)	Often best accuracy on tabular data	Needs tuning; can overfit if careless

(a) Comparison of Models (Graph and Table created by finalist using Google Sheets, 2025)

Forecasting Model Runtimes

"How fast will my model run?"



Frequency: Fastest because didn't use 50x50 grid

Magnitude: Varied because it processed quake history based on various regions.

Location: Consistent because it constantly calculated probabilities for each grid cell.

Forecasting Methods Background Research

Physics-Based:

- Uses physical principles like stress change and fault friction
- For long-term forecasting.

Statistical (Current Research method):

- Analyzes past seismicity patterns
- For quick, testable short-term predictions

Foundational Seismological Laws

- Omori's Law: States that aftershock rates decay over time.
- Gutenberg-Richter Law: Defines the magnitude vs frequency relationship in a region.

Major Model Families

- ETAS: Provides time and space based prospects for later events.
- Machine Learning:** Automates forecasting by learning from data rather than manual rules. *My model is this category!*
- Hybrid (ETAS + ML): Uses ETAS as a structural base while ML identifies complex residual patterns.

Conclusion + Future Work

In this project, I proved XGBoost was the best for training aftershock forecasts. With the optimized parameters, I was able to achieve 85% AUC. I applied the trained model to generate the timing pattern of the largest aftershock and multiple heatmaps. I also conducted multiple case studies based on real world earthquakes to show the strong correlation between real-world data and forecasted probability, time, magnitude, and location. By changing complex seismic data into accessible heatmaps and forecasts, my project provides a resource for emergency responders to prioritize high-risk zones and save lives during the 30-day window following a major earthquake.

To improve my project in future work, I will consider the following things:

- Quantify uncertainty to my project to show confidence ratings
- Analyze which features affect the results the most
- Incorporate ETAS into my XGBoost model
- Add the actual forecasting backend for users to input data and view results

I hope my work can help others and be applied in real-life scenarios. This project really helped me understand that technology can make a difference in forecasting natural disasters.