

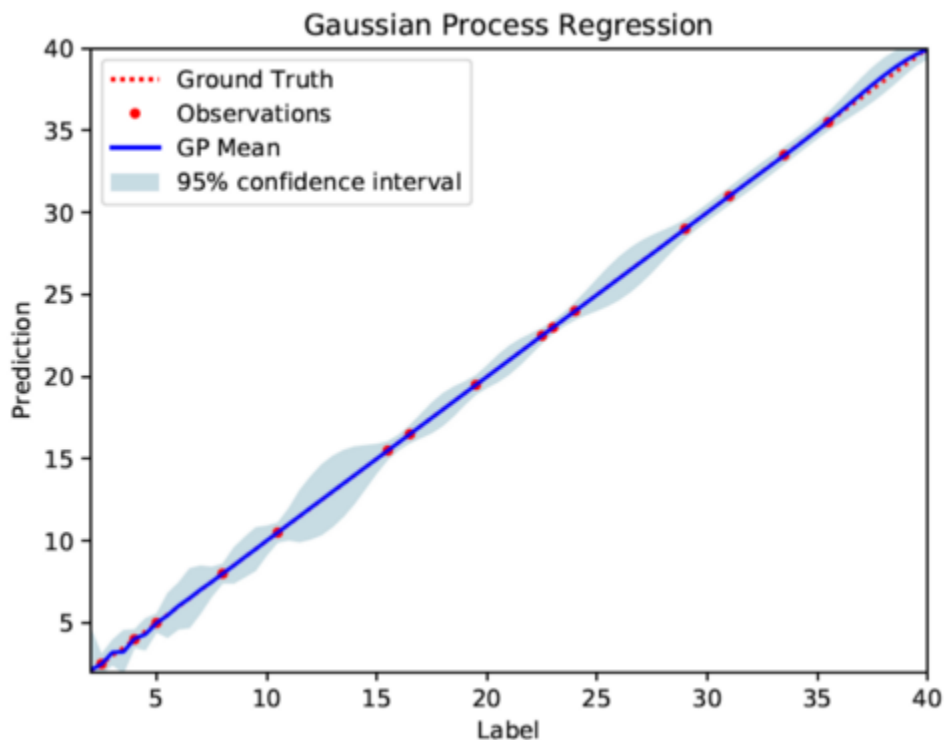
Predicting air quality after 2020.

Our team wanted to find out if air quality had improved significantly over the past 2 to 4 years. The current state of health that required many people to stop interacting in close groups small and large is reported in the news as showing many improvements anecdotally from 2019 to today.

Our team focused on datasets from the EPA air quality portal that reported the raw monitoring sensor data on a daily basis. The EPA data comes with an initial air quality index result for each pollutant monitored and the formula is published on the support site. [<https://www.epa.gov/outdoor-air-quality-data>]

With a very short time window, our team decided to focus on a regression model that could provide meaningful results to potentially noisy data. To get the benefit of some inline data normalization, we chose the [GPR] Gaussian process regression model that is available from scikit-learn. [site with <https://towardsdatascience.com/quick-start-to-gaussian-process-regression-36d838810319>]

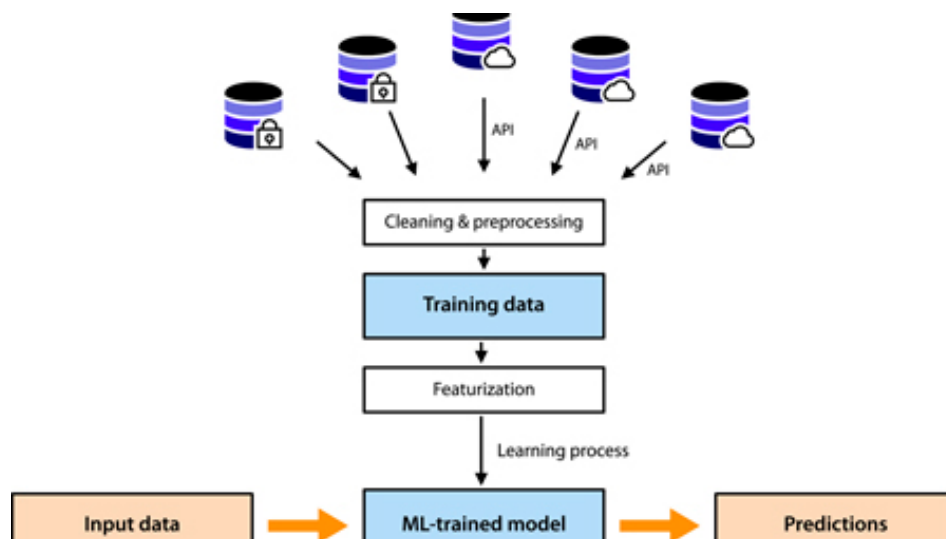
By using GPR, the team had confidence that using the raw data reading as a single feature to input, would show measurable patterns that could predict trends. The prediction results can be plotted to fit ground truth observations to a line that can show a slope and a range of confidence. [image from <https://towardsdatascience.com/quick-start-to-gaussian-process-regression-36d838810319>]



Our process approach to machine learning can be seen in the figure below:

- APIs and data saved from APIs were used
- Cleaning and processing data required the raw data collected be filtered and collated into different ways of looking at the data
 - City datasets was created to look at NO2, SO2, Ozone, CO and PM10 over 3 years, 10 years and a set going back to 2000
 - Each City dataset was filtered to look at one pollutant over 3 years, 10 years and back to 2000
 - The following models were tested to see what would give the best results
 - SVM model with various feature selection
 - Random Forest Classification with various kernels
 - KNN with various kernels
 - Logistic Regression with different feature selections and different kernels
 - Gaussian Process Regression – using the scikit Learn model set-up for noise in data

Results of testing showed that the GPR was the best model to use to get a result for prediction with the least amount of ground truth data.



Air Data: Air Quality Data Collected at Outdoor Monitors Across the US. (2021, August 12).

US EPA. <https://www.epa.gov/outdoor-air-quality-data>

Sit, H. (2020, June 12). *Quick Start to Gaussian Process Regression - Towards Data Science*.

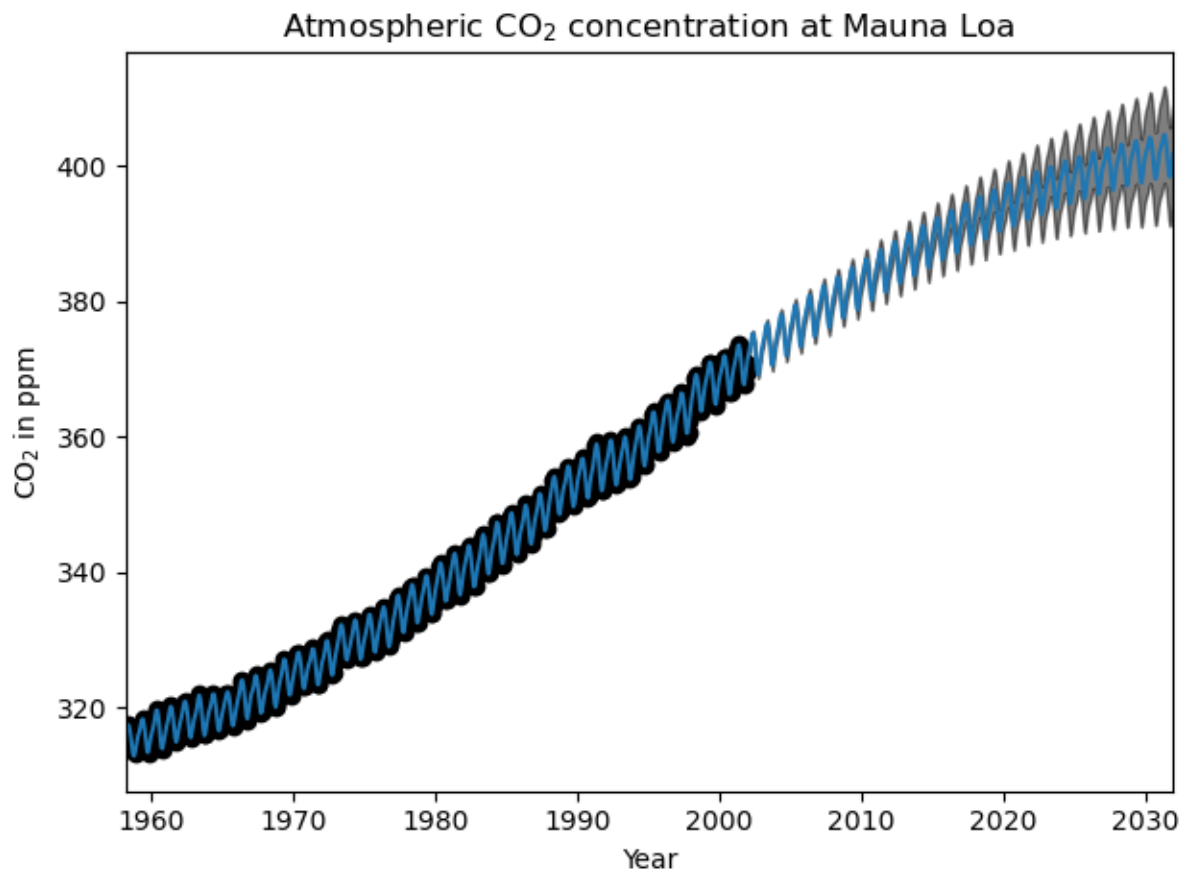
Medium. [https://towardsdatascience.com/quick-start-to-gaussian-process-regression-](https://towardsdatascience.com/quick-start-to-gaussian-process-regression-36d838810319)

36d838810319

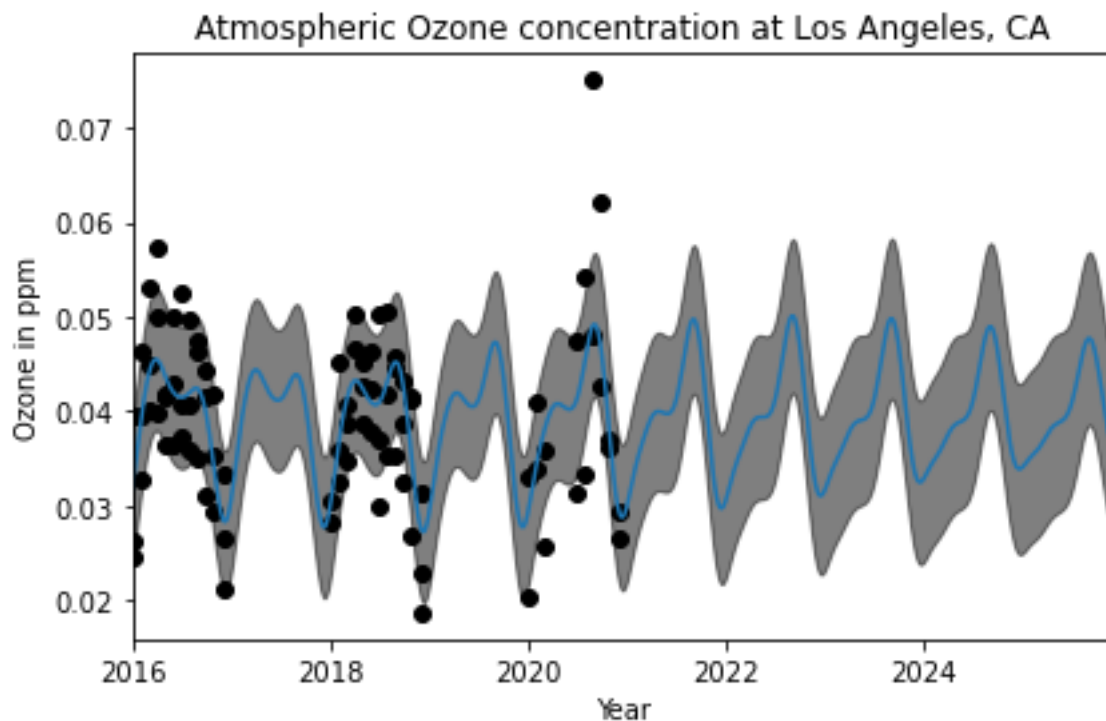
Our results


The Log-Marginal-Likelihood is similar to a confidence interval and is part of the prediction results that come from using the Gaussian Process Regression Model. It is a great tool for pilot studies and can help a team identify promising data in areas that they might not have considered.

A similar test was done using CO₂ data over 60 years for Mauna Loa with GPR. The results show a clear rise in CO₂ however the data had to be graphed over 60 years to catch the trend in a 5 year period and identify the overall results as going up or down.



The initial testing over 5 years with one-year gaps in between showed this a similar yearly pattern as the Mauna Loa dataset. After getting these results we realized we needed to look at data that covered more time to get a better sense of what the overall trends in air quality might be. It is noted that the United States has good air quality compared to other places in the world. To add to the data set we already had, we chose to only consider data that came from the EPA to have consistence in the quality of the raw that we were using, even though the earliest data was coming from Kaggle epa datasets. Our team ran data sets for Ozone, NO2 and PM10 for Los Angeles area, New York City area, Cheyenne area, Raleigh area and Seattle area. The Ozone results were the best and showed identifiable trends using the data we had readily available.





Air Quality

Final Group Project

The Questions

- Has Air Quality gotten better gotten better during the Pandemic?
- If air quality has gotten better, will the level of measurable pollutants in the air go back up when the restrictions are lifted?
- Could a regression or machine learning model show statistical significance and a clear prediction>

What is the data being measured?

- The EPA gives access to the data from their monitoring stations
- The data could be considered 'ground truth' and is reliable and considered the best quality as long as instruments are calibrated and in good repair
- Our team used the sensor data and a computed air quality score as features to look at and measure

The Data Domain

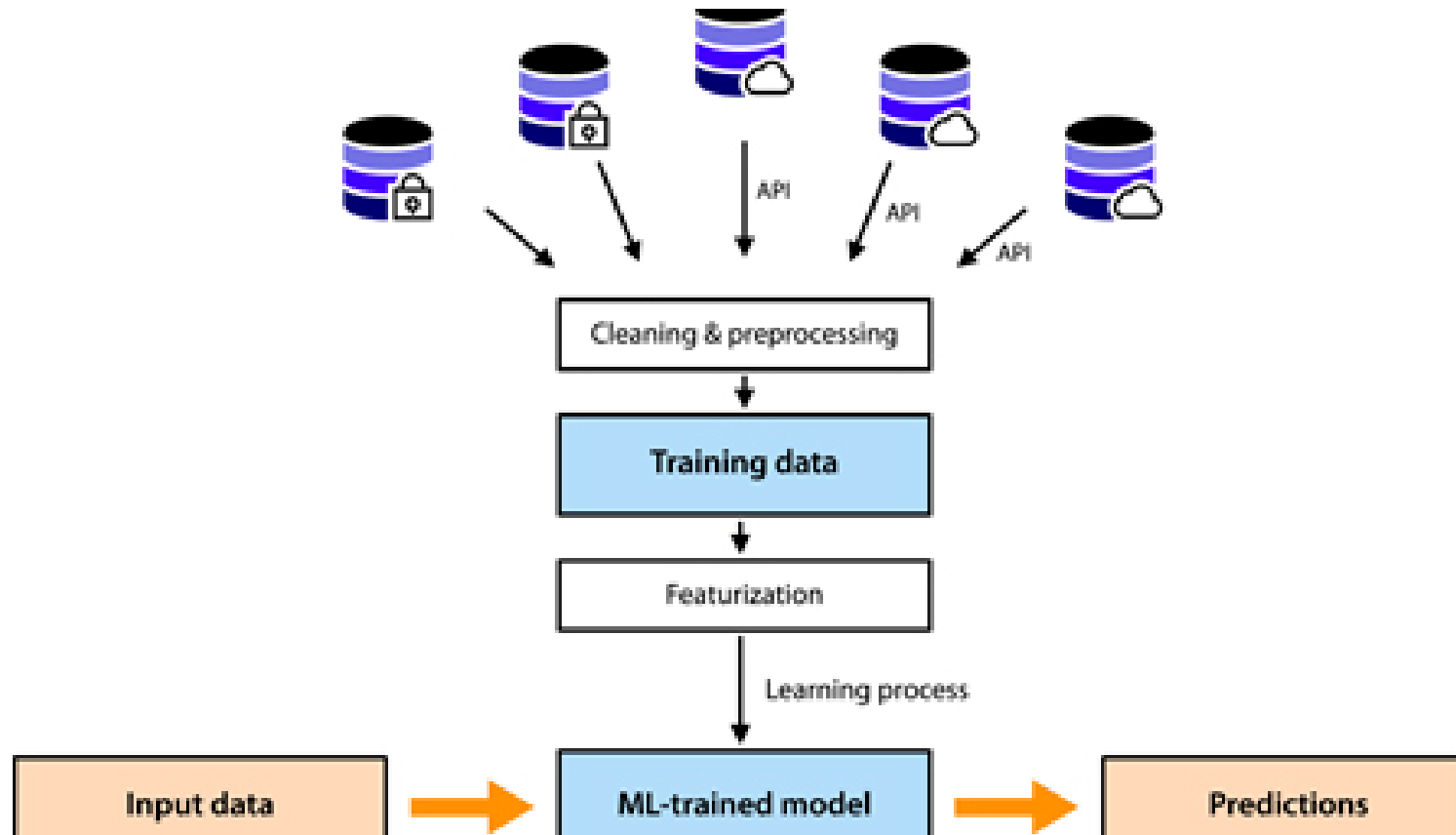
Municipal Areas Studied

- Cheyenne, Wyoming
- El Paso, Texas
- Los Angeles, California
- New York City, New York
- Raleigh, North Carolina
- Seattle, Washington

Pollutants affecting Air Quality

- Ozone - ppm
- NO₂ - ppb
- SO₂ - ppb
- CO - ppm
- PM₁₀ ug/m³

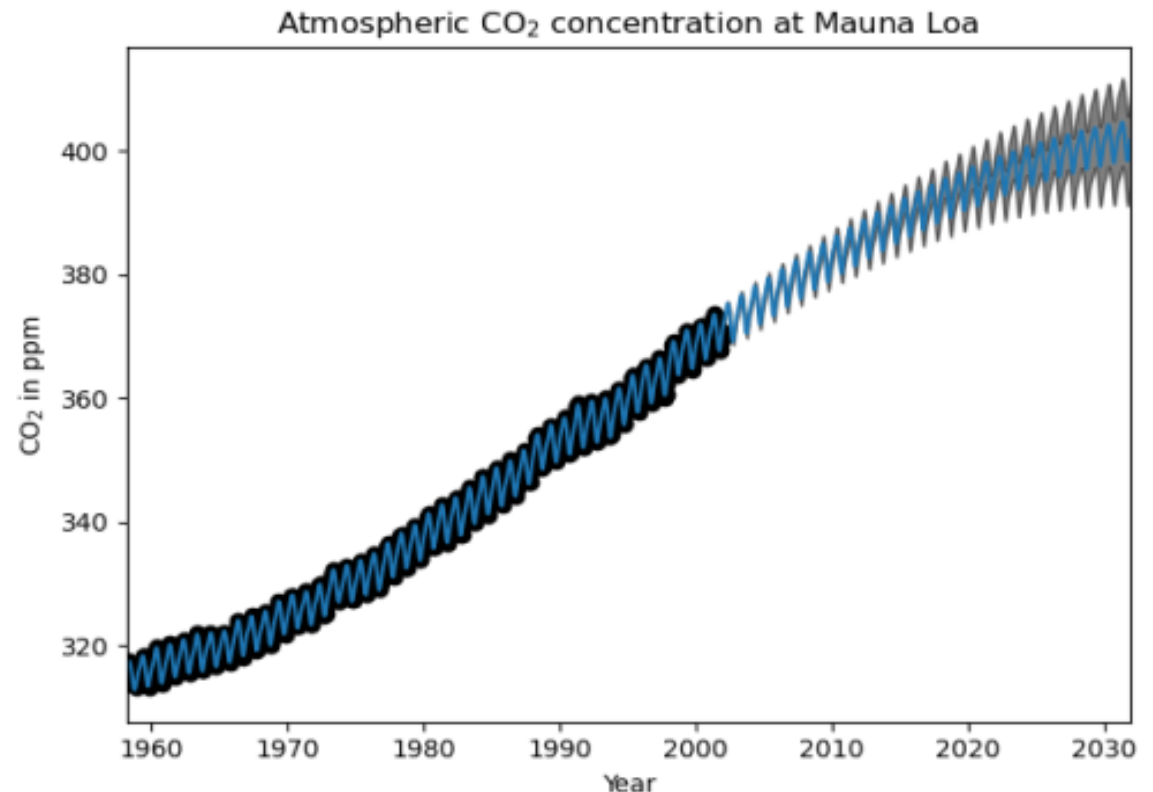
Our Approach to Machine Learning



The result we wanted:

Initial Results

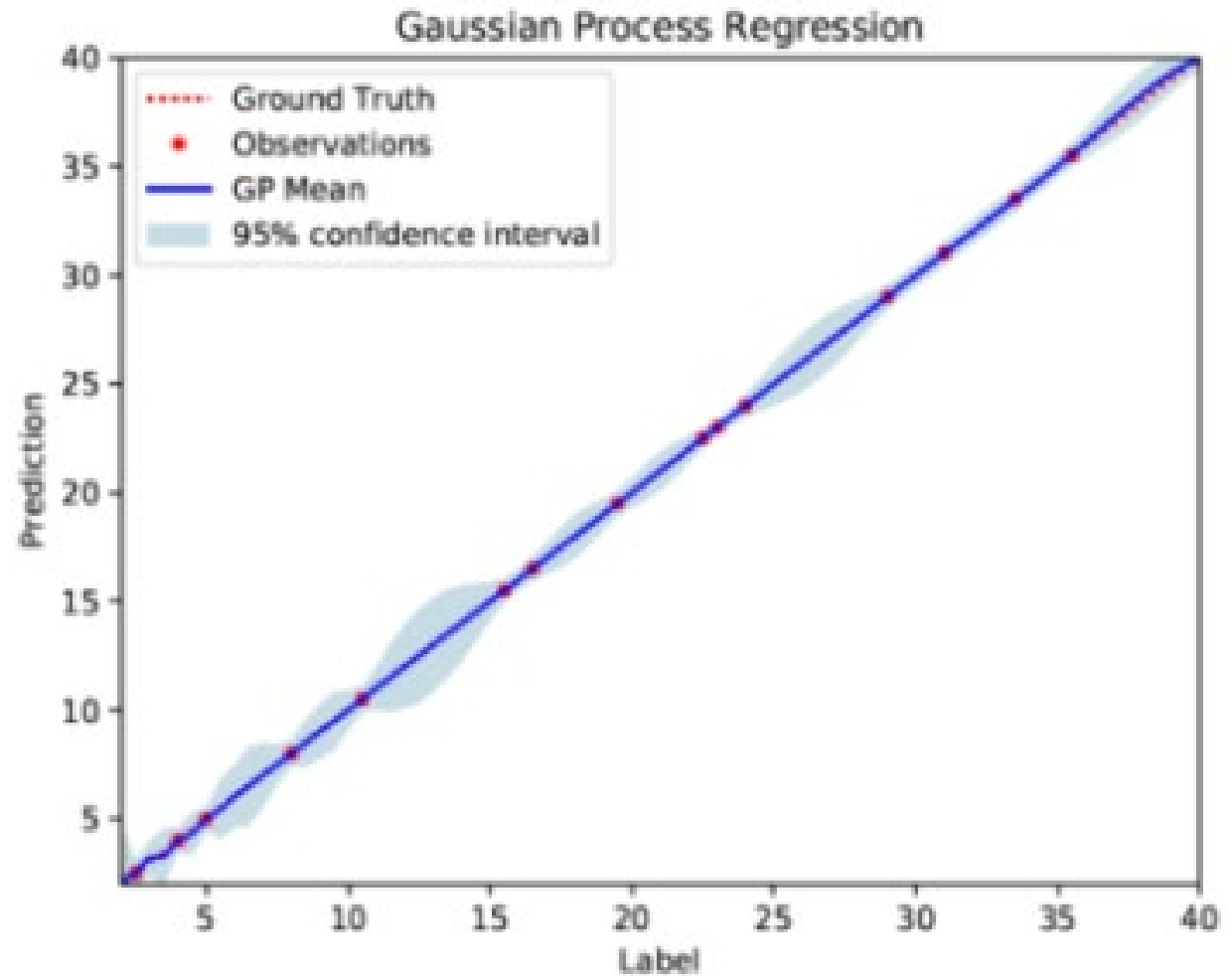
- During initial discovery and testing machine learning models that showed more promise were regression tests that could fit the data to a continuous line.
- Other models were checked - to verify this assumption
- It became clear that the Gaussian Process Regression model used to show atmospheric concentration at Maun Loa was a model that would show significant results with a small feature set
- The Mauna Loa results also showed that the dataset had to be augmented by added more years to get a noticeable trend over time



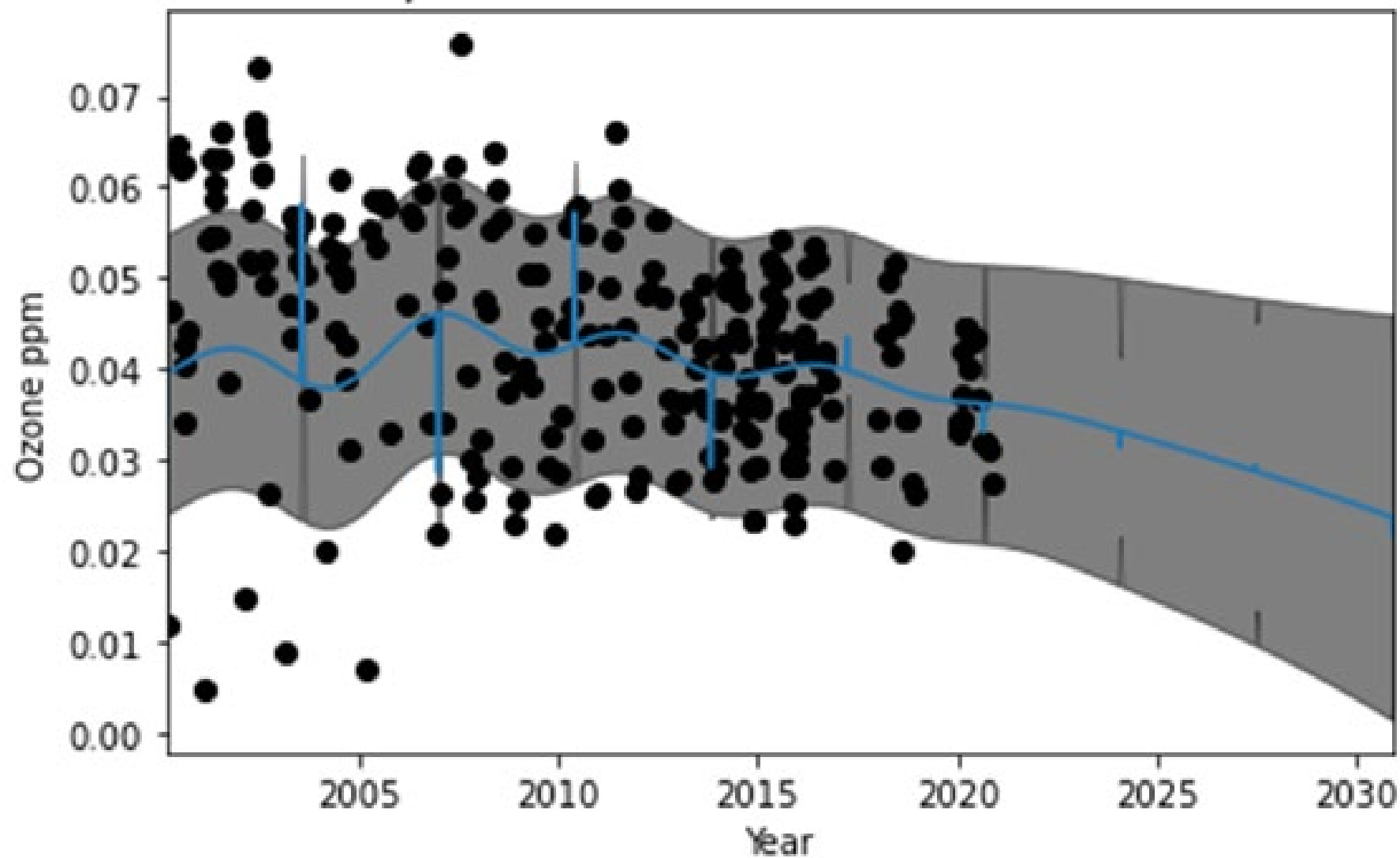
Our Best Results

We added more historic EPA data and were able to get the beginnings of a prediction. We still had a lot of noise in the data but using the GPR allowed the team to get results that would show whether it was worth considering a specific change in pollutants in the air. The GPR also showed what specific areas were worth studying more closely.

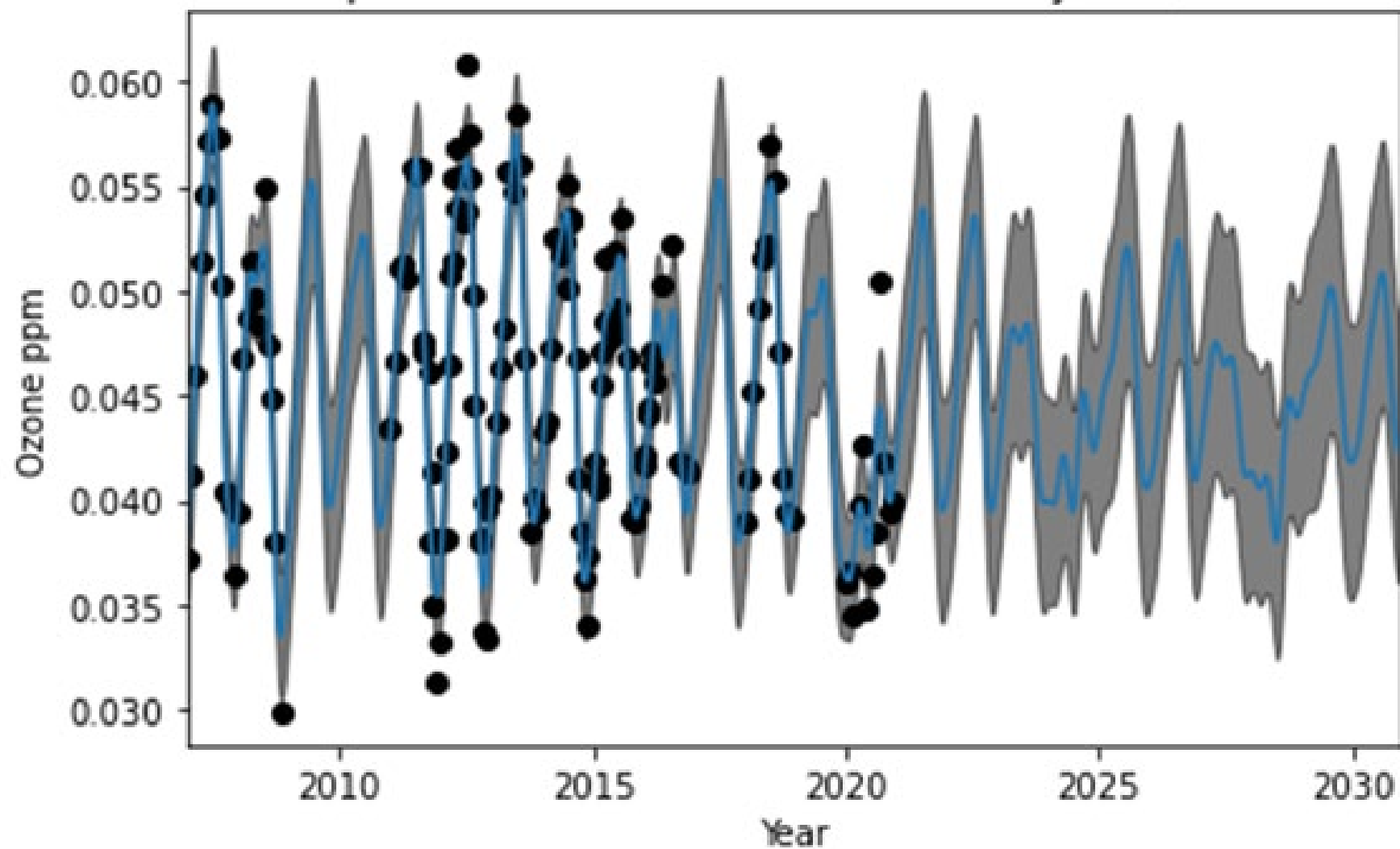
This image shows what the ideal data would look like.



Atmospheric Ozone concentration at RTP Area NC



Atmospheric Ozone concentration at Cheyenne, WY area



Atmospheric Ozone concentration at Seattle, WA area

