HSPH Capstone Project
Modeling Air Pollution

Members: Casey, Chris, Justin, and Keyan

TF: David Sondak

Our Partners

National Studies on Air Pollution and Health (NSAPH)

Prof. Francesca Dominici, Dr. Christine Choirat, Ben Sabath

Harvard T.H. Chan School of Public Health

IACS Institute for Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

HARVARD SCHOOL OF PUBLIC HEALTH

What’s the Problem?

Problem Statement

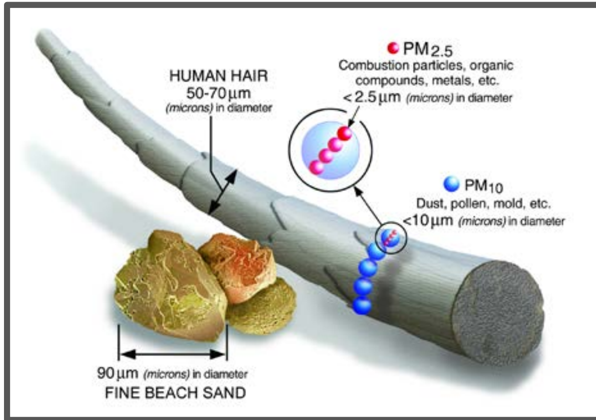
Motivation

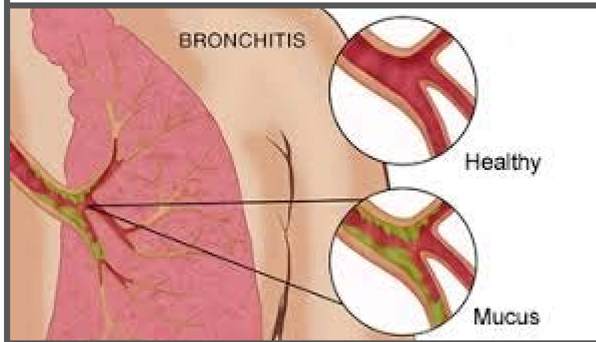
$PM_{2.5}$ Is the Problem

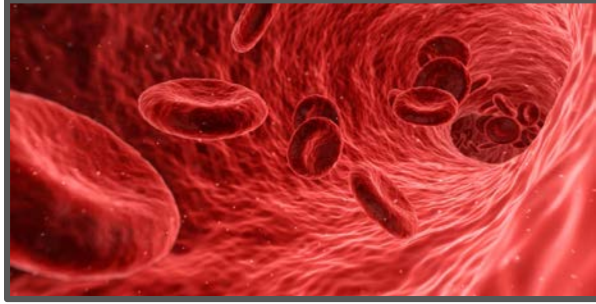
- Fine Particulate Matter < 2.5 μm in diameter

- Natural and artificial sources

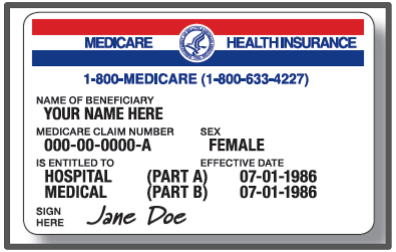
- Associated with a host of adverse **short-** and **long-term** health effects: asthma, bronchitis, lung cancer, ...





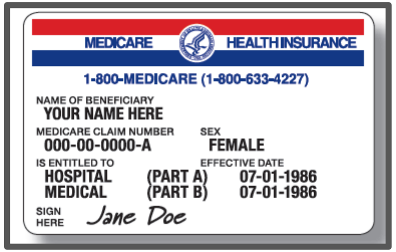


HSPH’s Goal



HARVARD
SCHOOL OF PUBLIC HEALTH

HSPH's Goal



VE RI TAS
HARVARD
SCHOOL OF PUBLIC HEALTH

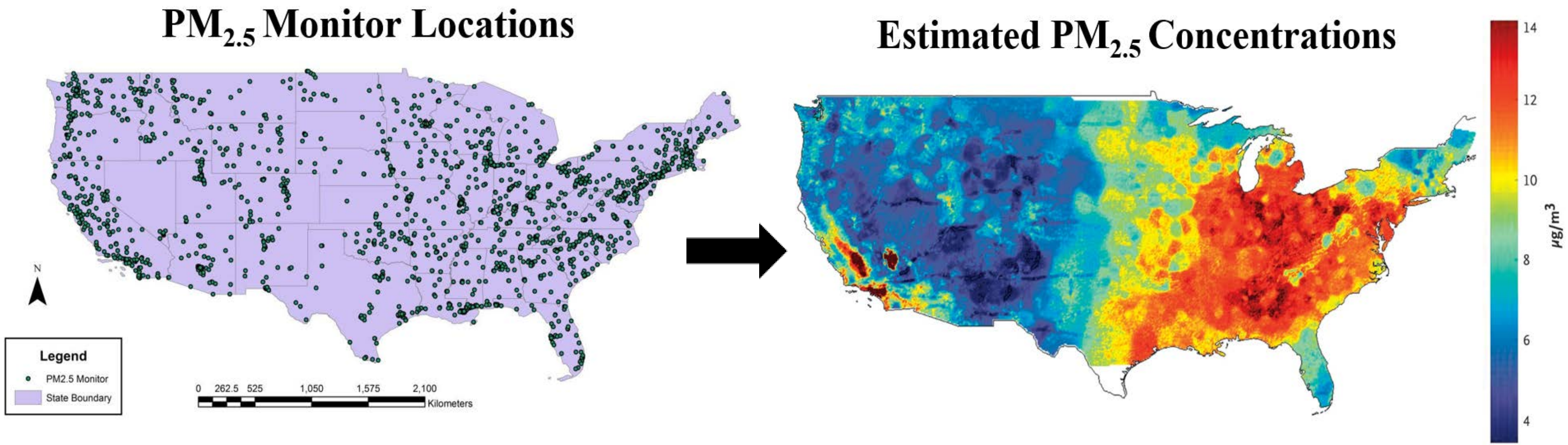
Motivation - Causal Inference

There are many areas in the US for which the $PM_{2.5}$ concentrations are **not** known.

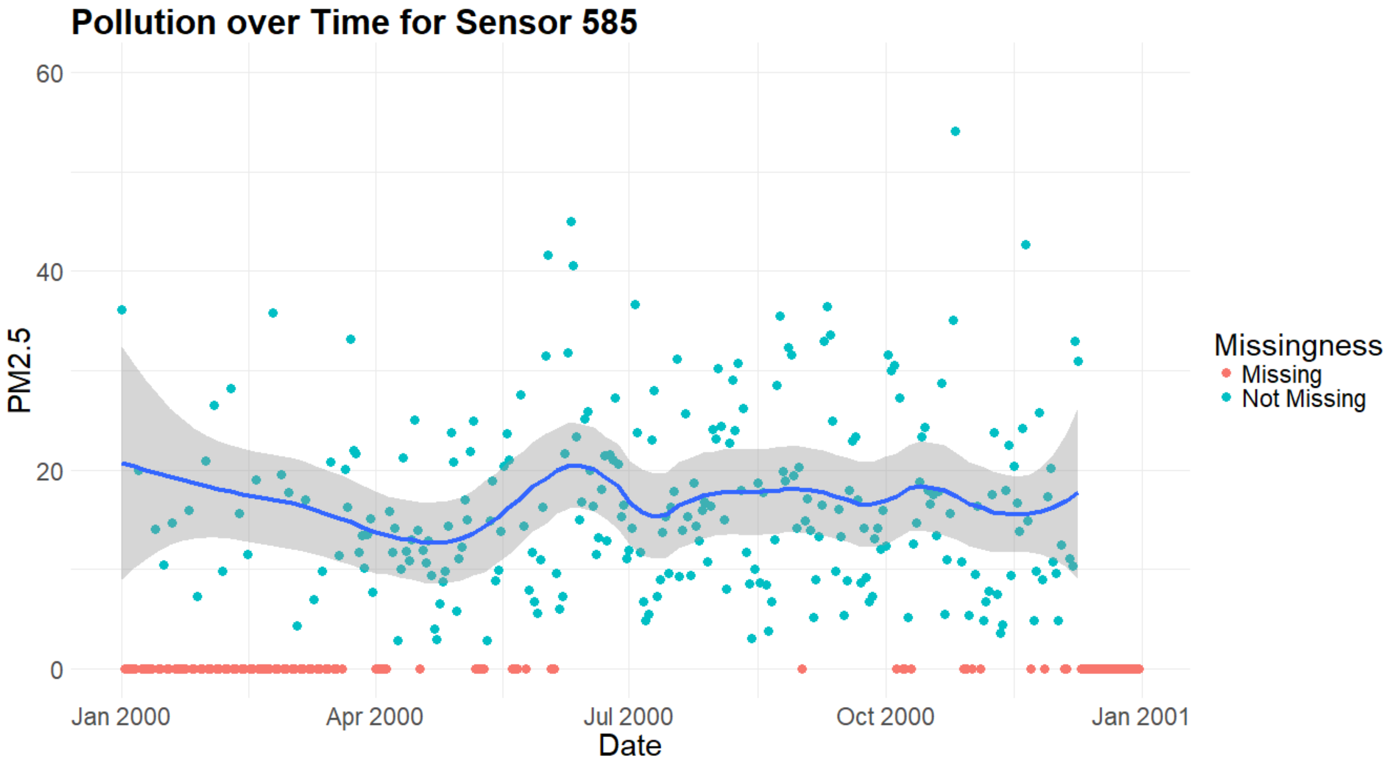
We cannot estimate the overall effect that pollution has on health if we do not know what the pollution is in those areas.

Thus, our goal was to fit a model that accurately interpolates pollution throughout the entire US on a daily basis.

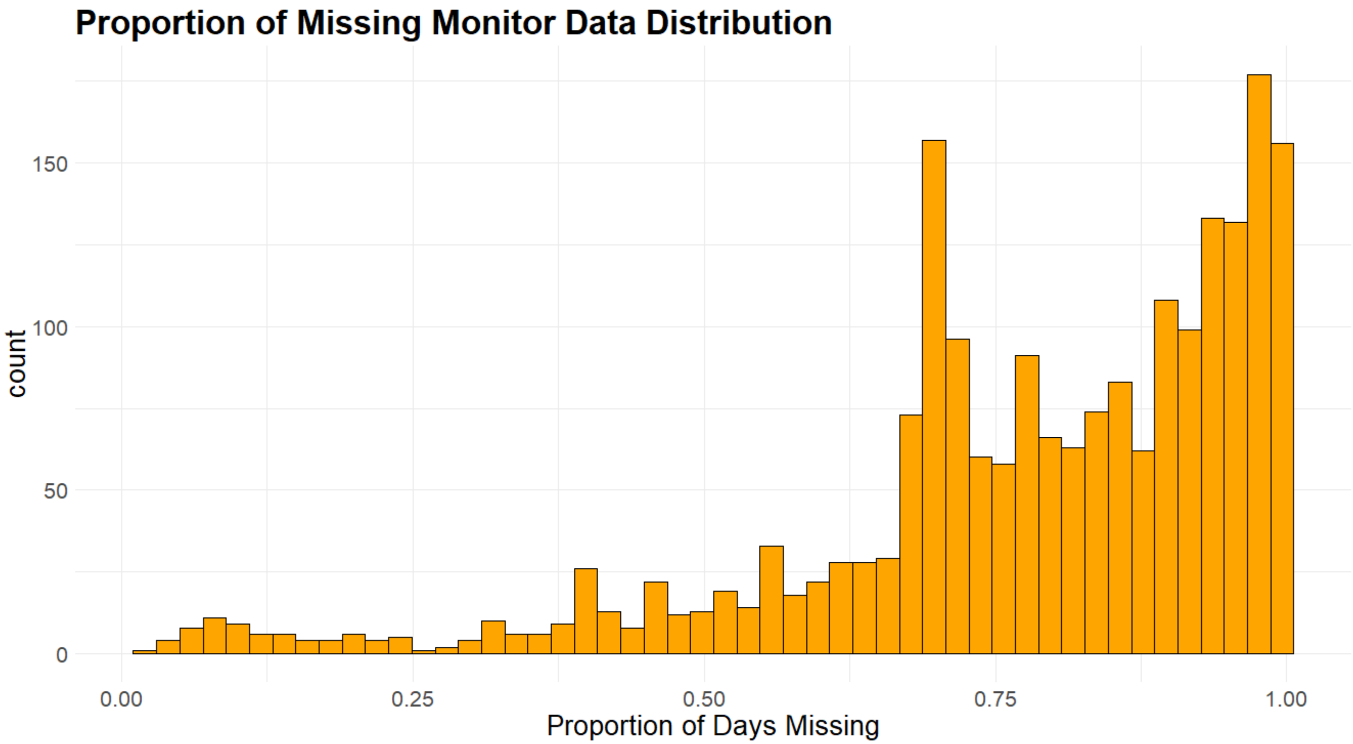
Spatial Interpolation

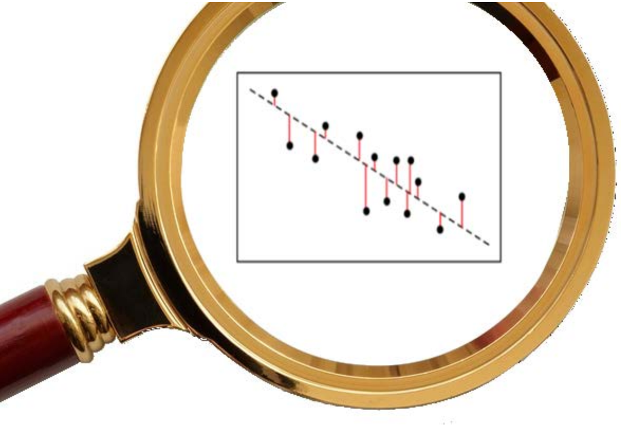


Temporal Interpolation



Temporal Interpolation





Data, Preprocessing, & EDA

Exploratory Data Analysis

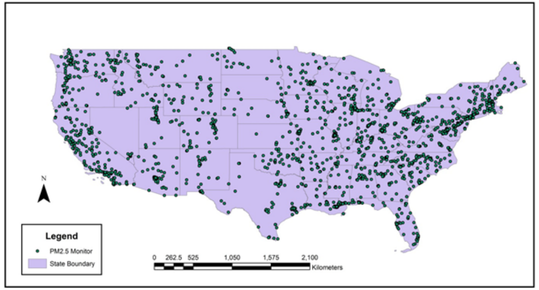
Feature Correlations

Missing Data

Given Data

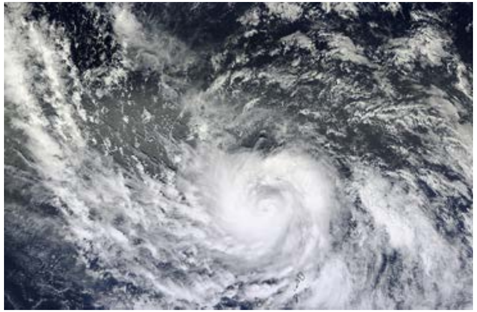
16yrs
2156 Sites

Sensor Data
(Response)



- 13M $PM_{2.5}$ Sensor-Days
- ~75% sensor-days missing
    - Defunded sites (costly)

Satellite Data
(Predictors)



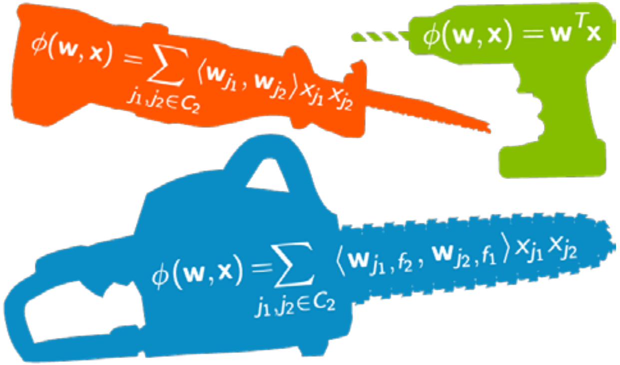
- 115 measurements/day/site
- Also severe missingness
    - Cloud cover
    - Snow reflection

Added Data

Census Data

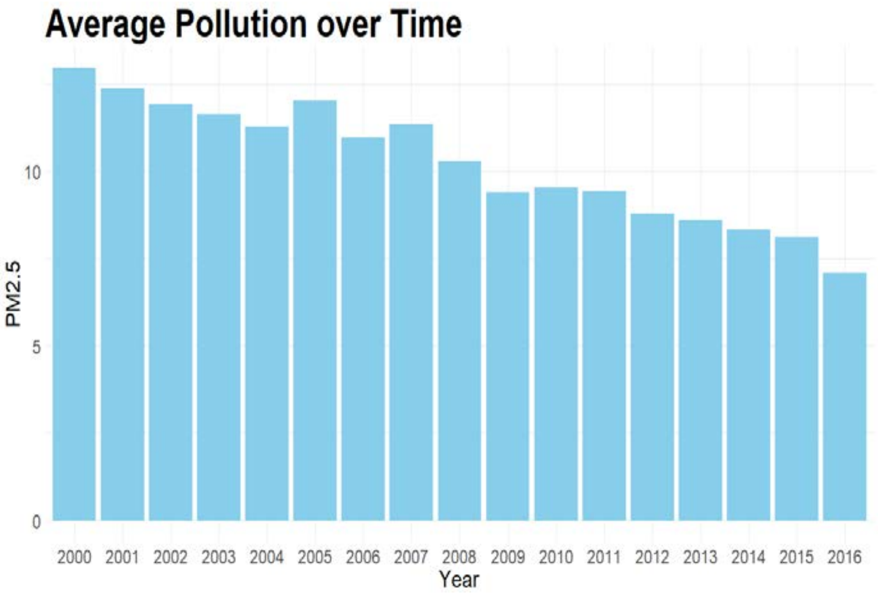
- Population Density
- Income distribution
- Education

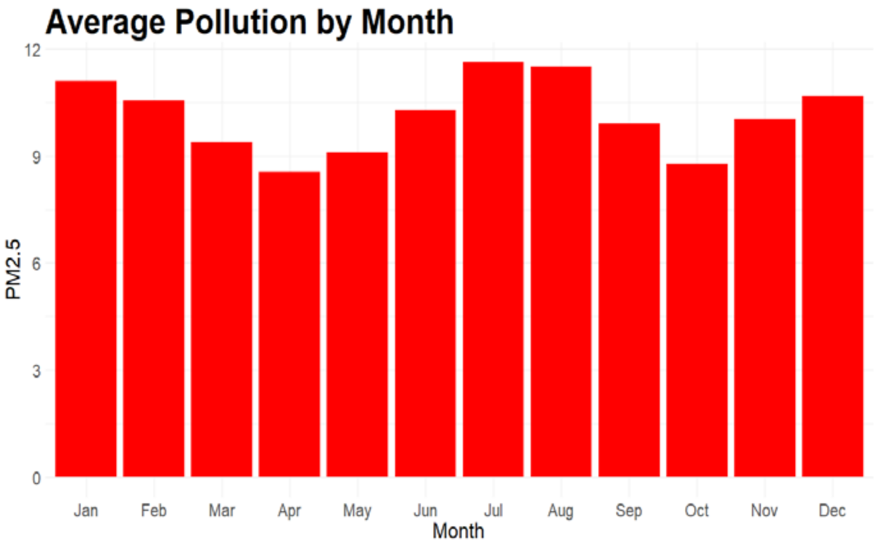
Engineered Features



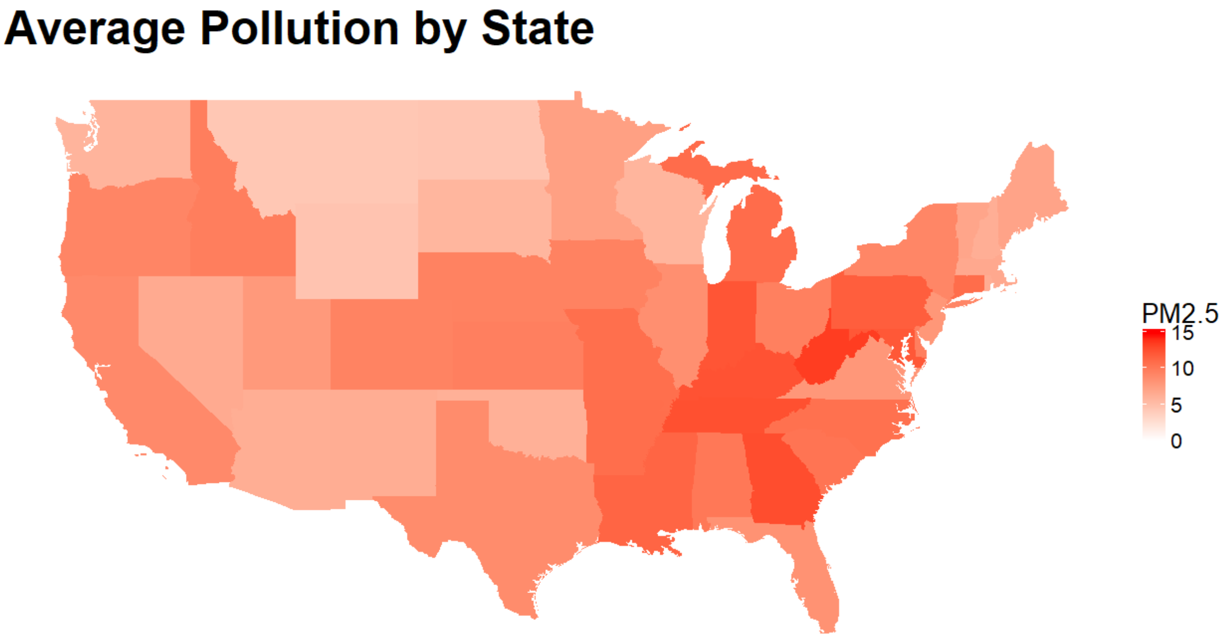
- Nearby $PM_{2.5}$ lead
- Periodic time domain features

EDA - Pollution Over Time

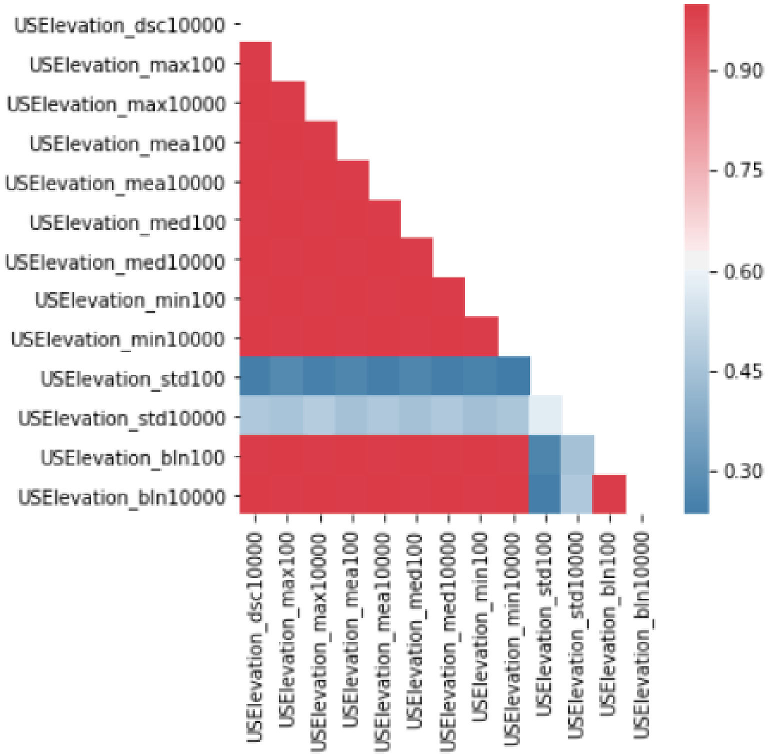


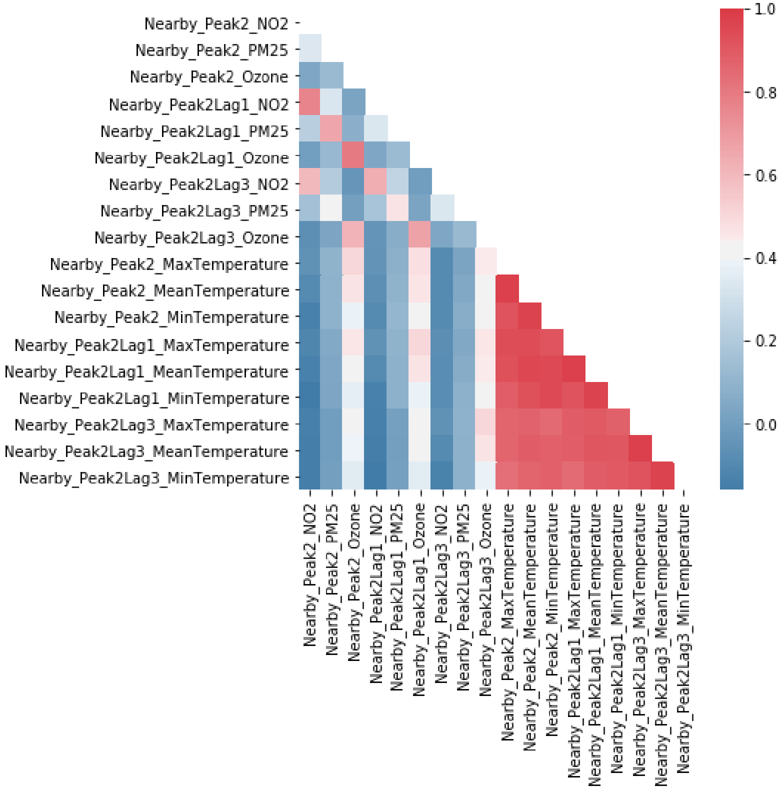


EDA - Pollution by Location



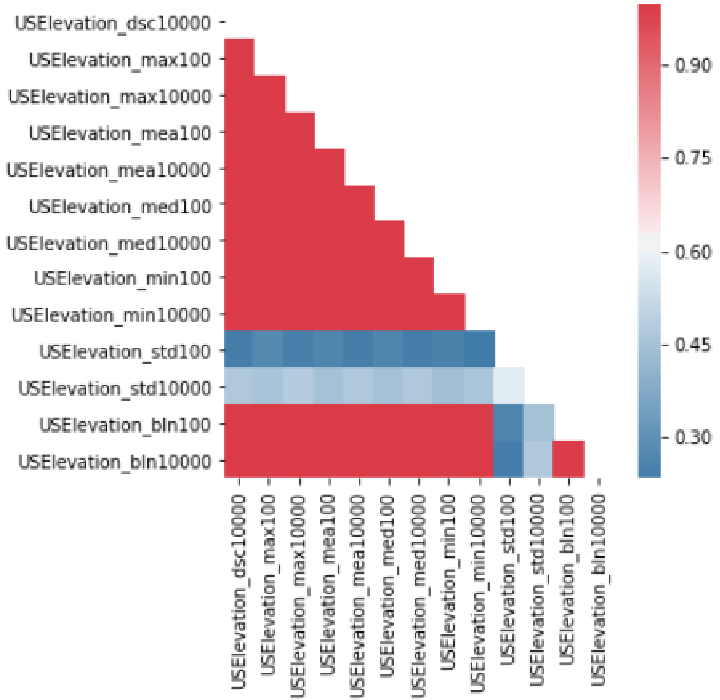
EDA - Pairwise Correlations





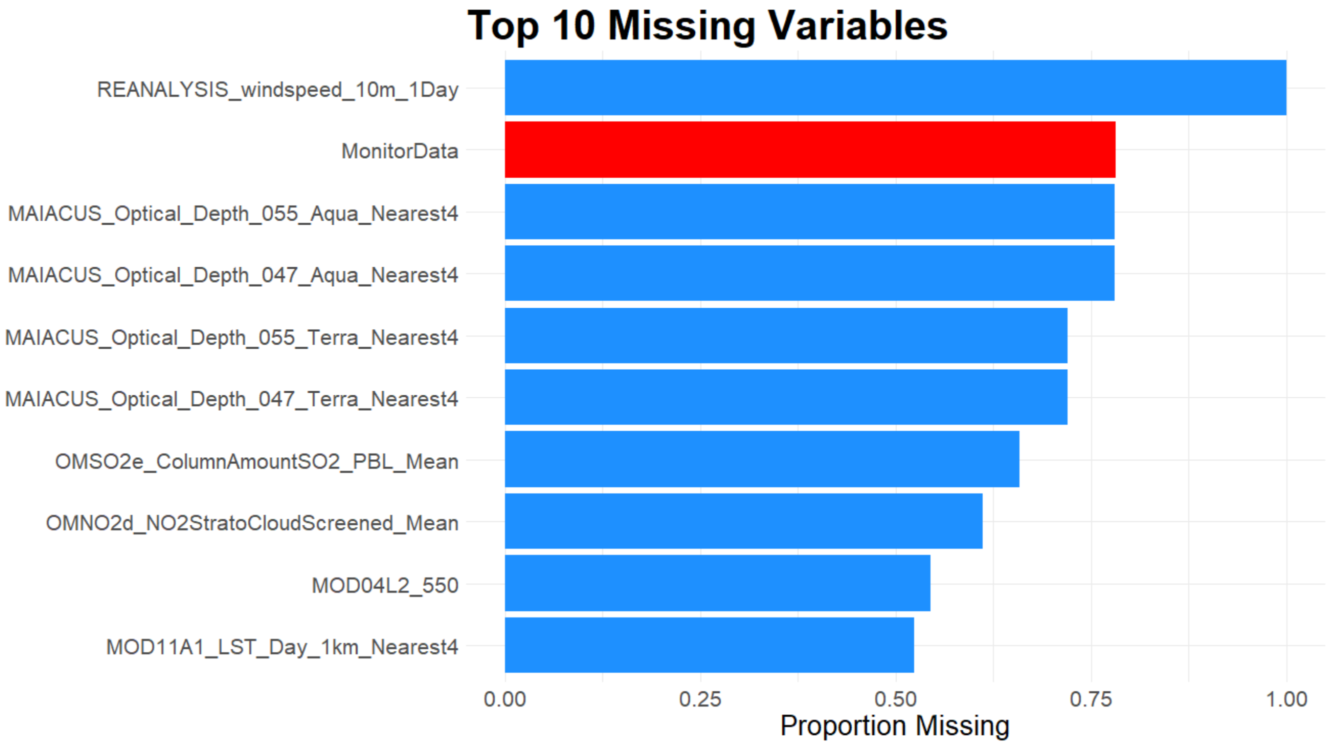
Dimensionality Reduction

- For each pair of variables with greater than 0.9 correlation, drop one based on:
    - Correlation with response
    - Amount of missingness



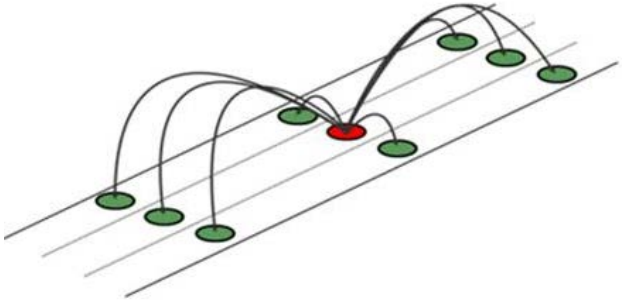
- Also removed variables that had non-significant partial correlations with $PM_{2.5}$ after controlling for nearby $PM_{2.5}$

Missing Data



Correlations with $PM_{2.5}$

Variable	Correlation
<chr>	<dbl>
Nearby_Peak2_PM25	0.8571
Nearby_Peak2Lag1_PM25	0.5917
MAIACUS_Optical_Depth_047_Terra_Nearest4	0.4155
MAIACUS_Optical_Depth_055_Terra_Nearest4	0.4062
Nearby_Peak2Lag3_PM25	0.3775
Nearby_Peak2_NO2	0.3409
MAIACUS_Optical_Depth_047_Aqua_Nearest4	0.3316
Nearby_Peak2Lag1_NO2	0.3252
MAIACUS_Optical_Depth_055_Aqua_Nearest4	0.3250
REANALYSIS_hpbl_DailyMean	-0.2924



Imputing Missing Data

MissForest Algorithm

Modifications and Additions

Evaluation

MissForest Algorithm

Data and text mining Advance Access publication October 28, 2011

MissForest—non-parametric missing value imputation for mixed-type data

Daniel J. Stekhoven[1,2,3,*] and Peter Bühlmann[1,3]

[1]Seminar for Statistics, Department of Mathematics, ETH Zurich, [2]Life Science Zurich PhD Program on Systems Biology of Complex Diseases and [3]Competence Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Modern data acquisition based on high-throughput technology is often facing the problem of missing data. Algorithms commonly used in the analysis of such large-scale data often depend on a complete set. Missing value imputation offers a solution to this problem. However, the majority of available imputation methods are restricted to one type of variable only: continuous or categorical. For mixed-type data, the different types are usually handled separately.

development of new and enhanced measurement techniques in these fields provides data analysts with challenges prompted not only by high-dimensional multivariate data where the number of variables may greatly exceed the number of observations, but also by mixed data types where continuous and categorical variables are present. In our context, categorical variables can arise as any kind ranging from technical settings in a mass spectrometer to a diagnostic expert opinion on a disease state. Additionally, such datasets often contain

MissForest Algorithm

Algorithm 1 Impute missing values with RF.

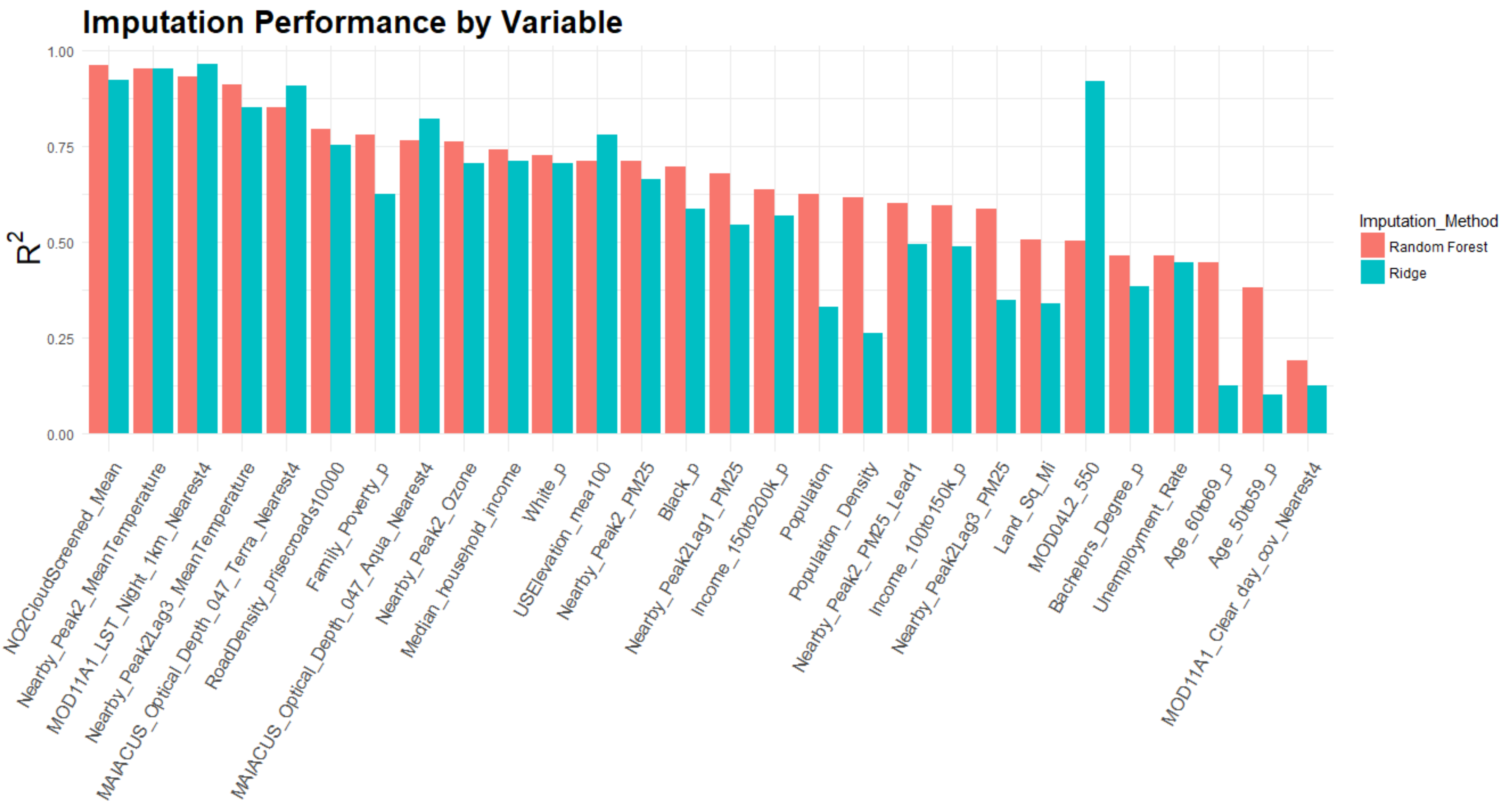
Require: $\mathbf{X}$ an $n \times p$ matrix, stopping criterion γ

Make initial guess for missing values;
$\mathbf{k} \leftarrow$ vector of sorted indices of columns in $\mathbf{X}$ w.r.t. increasing amount of missing values;
while not γ **do**
  $\mathbf{X}^{imp}_{old} \leftarrow$ store previously imputed matrix;
  for s in $\mathbf{k}$ **do**
    Fit a random forest: $\mathbf{y}^{(s)}_{obs} \sim \mathbf{x}^{(s)}_{obs}$;
    Predict $\mathbf{y}^{(s)}_{mis}$ using $\mathbf{x}^{(s)}_{mis}$;
    $\mathbf{X}^{imp}_{new} \leftarrow$ update imputed matrix, using predicted $\mathbf{y}^{(s)}_{mis}$;
  end for
  update γ.
end while
return the imputed matrix $\mathbf{X}^{imp}$

1. Perform **mean imputation**
2. For each variable with missing values:
   a. Fit **random forest** using non-missing values as response and all other variables as predictors
   b. Use fitted random forest to impute missing values
3. Repeat step 2 until:
   a. Convergence stopping criterion reached OR
   b. Max number of iterations reached

Modifications and Additions

- Allow for use of ridge regression (or any scikit-learn model)
    - Significantly improves algorithm runtime
- Imputation evaluation scheme
    - Perform imputation for non-missing values in holdout set and compute R^2
- Improvements for making imputations on new datasets
    - MissForest R package does not allow for imputation on new datasets without model re-fitting
    - Current Python implementation does not correctly follow the MissForest algorithm


Imputation Performance by Variable
R^2
1.00
0.75
0.50
0.25
0.00
Imputation_Method
Random Forest
Ridge
NO2CloudScreened_Mean
Nearby_Peak2_MeanTemperature
MOD11A1_LST_Night_1km_Nearest4
Nearby_Peak2Lag3_MeanTemperature
MAIACUS_Optical_Depth_047_Terra_Nearest4
RoadDensity_prisecroads10000
Family_Poverty_p
MAIACUS_Optical_Depth_047_Aqua_Nearest4
Nearby_Peak2_Ozone
Median_household_income
White_p
USElevation_mea100
Nearby_Peak2_PM25
Black_p
Nearby_Peak2Lag1_PM25
Income_150to200k_p
Population
Population_Density
Nearby_Peak2_PM25_Lead1
Income_100to150k_p
Nearby_Peak2Lag3_PM25
Land_Sq_Mi
MOD04L2_550
Bachelors_Degree_p
Unemployment_Rate
Age_60to69_p
Age_50to59_p
MOD11A1_Clear_day_cov_Nearest4

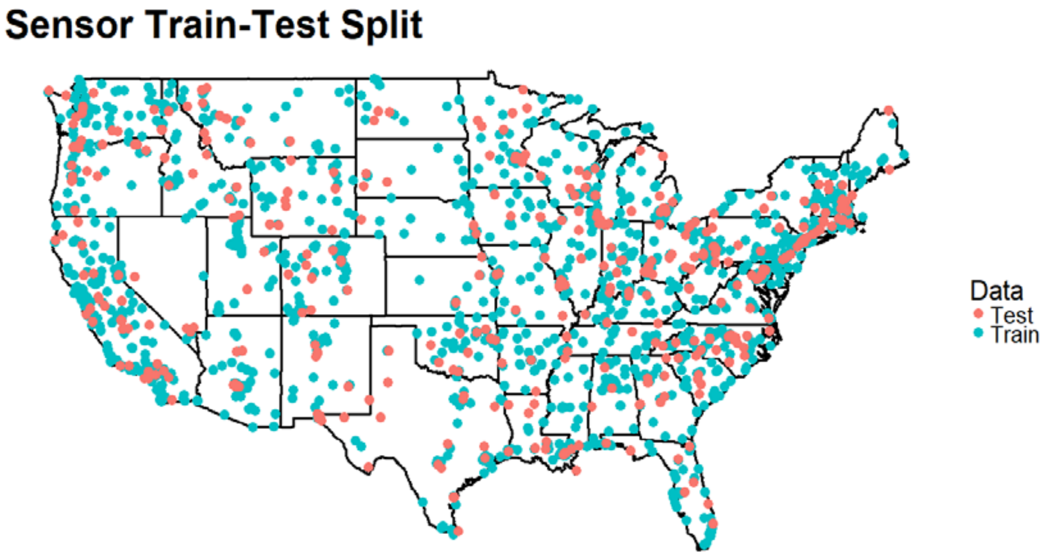
Modeling

- Setup
- Baseline Model
- Scikit-Learn Models
- CNN
- Evaluation

Modeling Setup

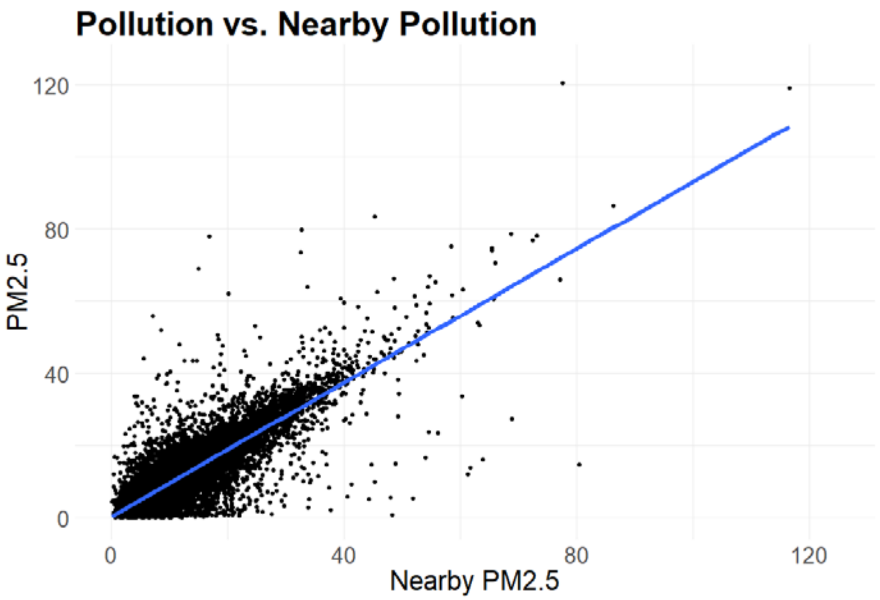
- Trained models on a random 80% sample of sensors
  - Reserved the remaining sensors for testing

- Tuned all hyper-parameters using K-fold cross-validation.
  - Optimizing for R^2



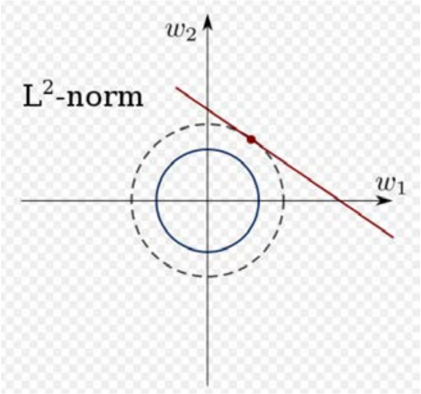
Baseline Model - Simple Linear Regression

Variable <chr>	Correlation <dbl>
Nearby_Peak2_PM25	0.8571
Nearby_Peak2Lag1_PM25	0.5917
MAIACUS_Optical_Depth_047_Terra_Nearest4	0.4155
MAIACUS_Optical_Depth_055_Terra_Nearest4	0.4062
Nearby_Peak2Lag3_PM25	0.3775
Nearby_Peak2_NO2	0.3409
MAIACUS_Optical_Depth_047_Aqua_Nearest4	0.3316
Nearby_Peak2Lag1_NO2	0.3252
MAIACUS_Optical_Depth_055_Aqua_Nearest4	0.3250
REANALYSIS_hpbl_DailyMean	-0.2924

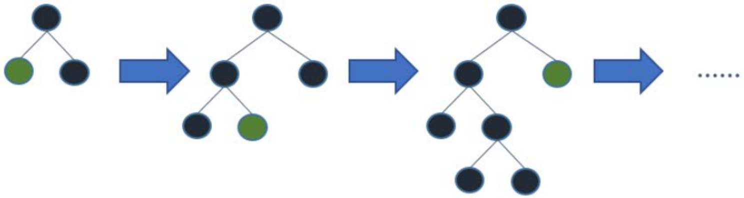


Modeling Methods Tested

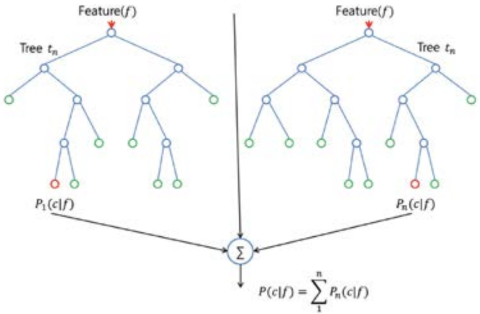
Ridge



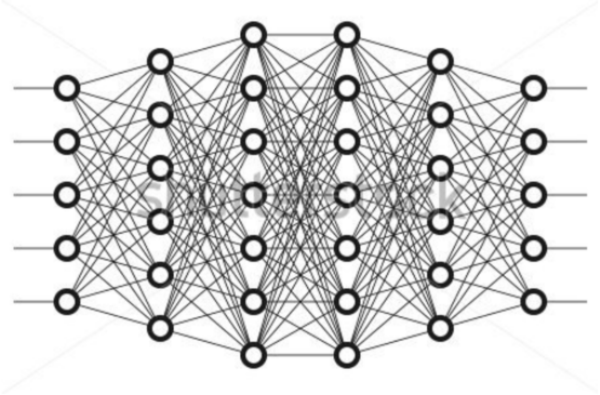
XGBoost



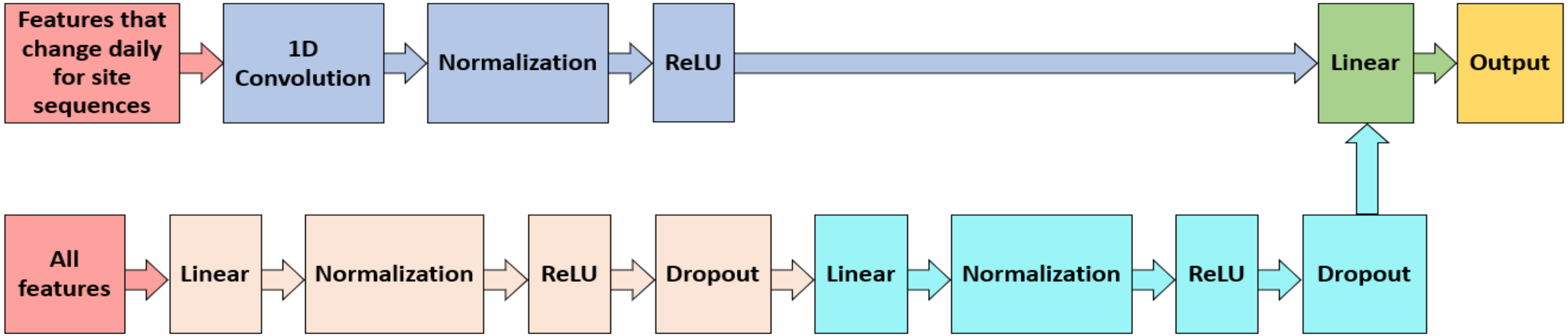
Random Forest



CNN



CNN Architecture

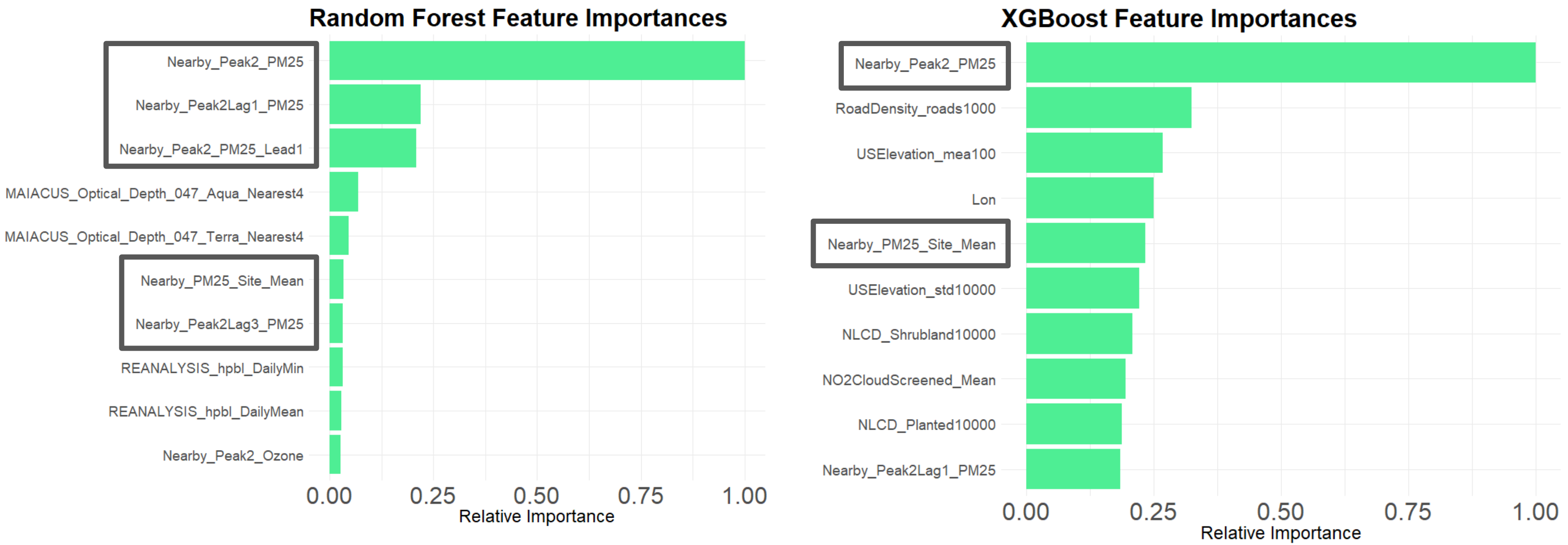


- 1D convolutional layer:
  - Inputs: Features that change on a daily basis within a site sequence
  - Kernel width of size 3: Relationships between features from previous day, current day, and following day accounted for when predicting pollution on current day
- All features inputted to 2-layer fully connected component
- Hidden units resulting from 1D convolution and 2-layer component concatenated

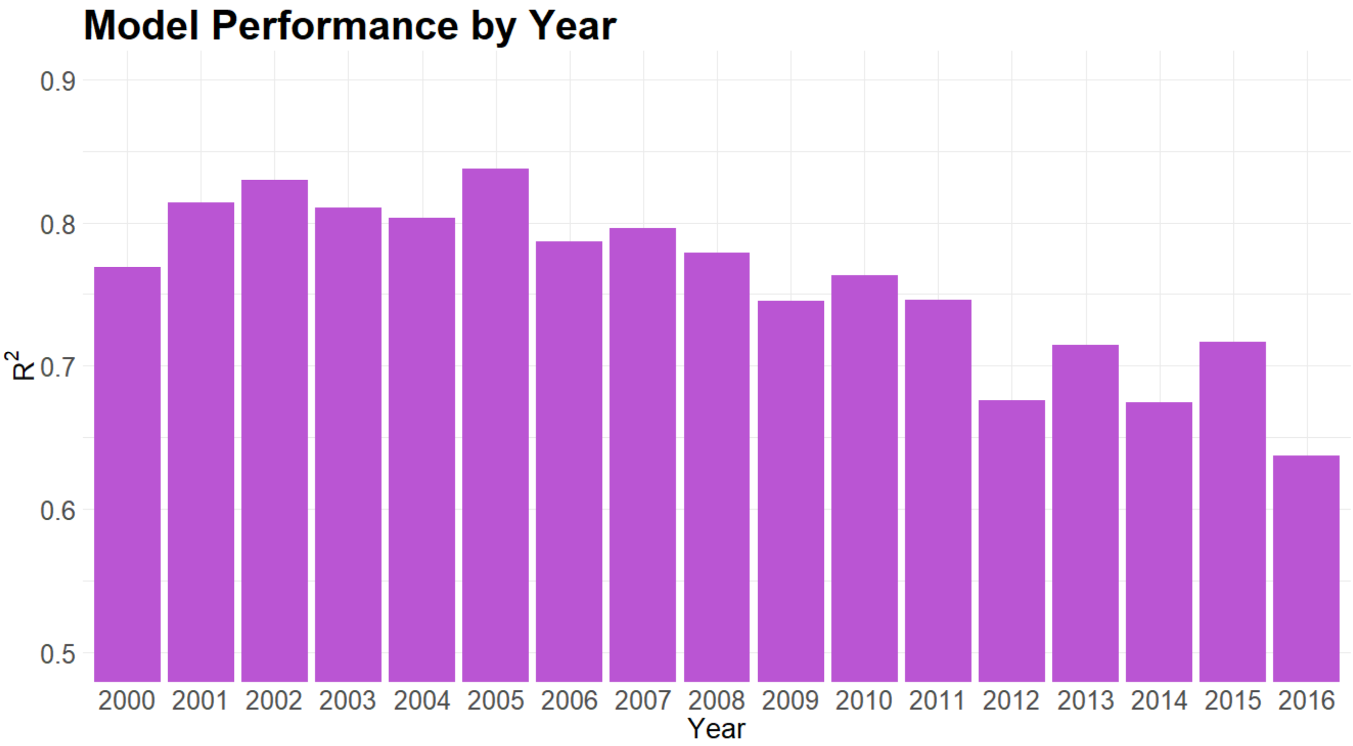
Model Test R^2 Results

	OLS	Ridge	RF	XGBoost	CNN	Ensemble
R^2	0.712	0.733	0.780	0.776	0.775	**0.784**

Feature Importances

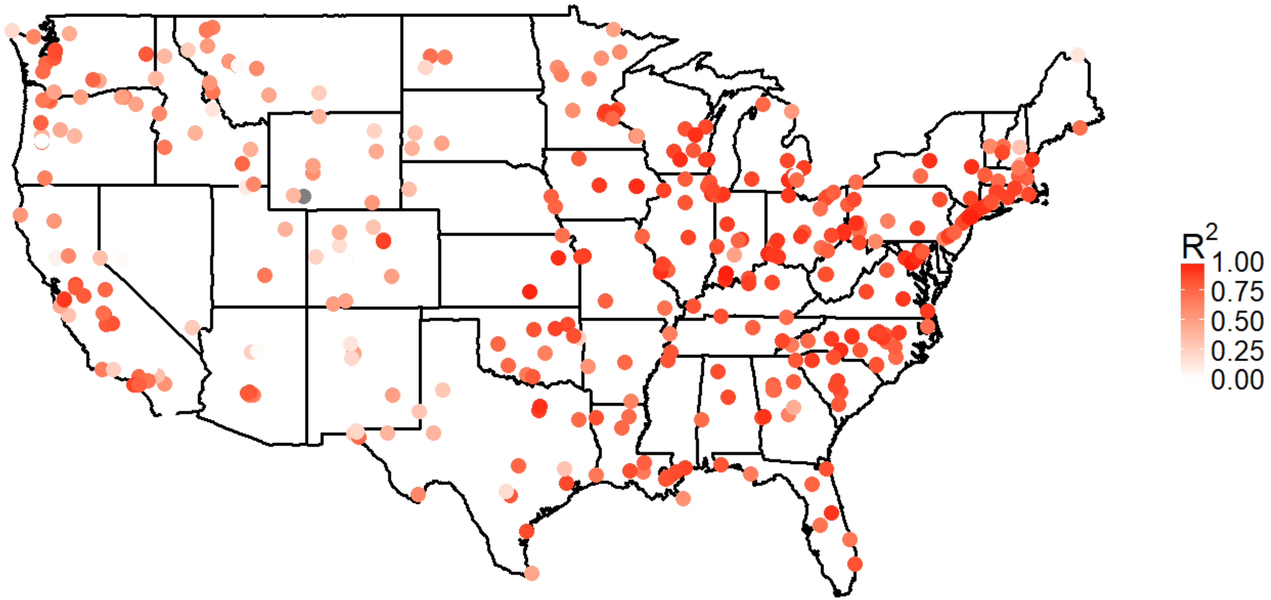


Model Diagnostics



Model Diagnostics

Model Performance by Location



Looking Ahead

Takeaways

Future Work

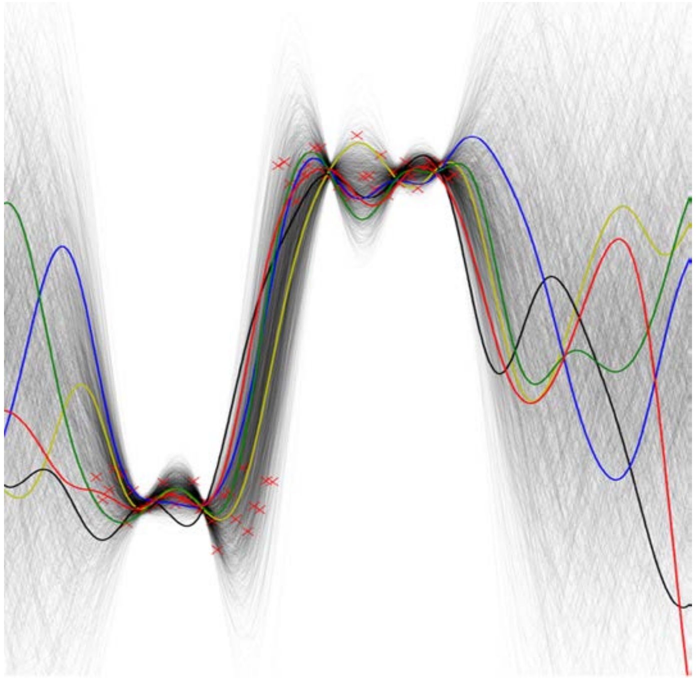
Takeaways

- Because of the disproportionate importance of nearby $PM_{2.5}$, we believe that it is absolutely essential for more pollution monitors to be installed, especially in regions where there are few.

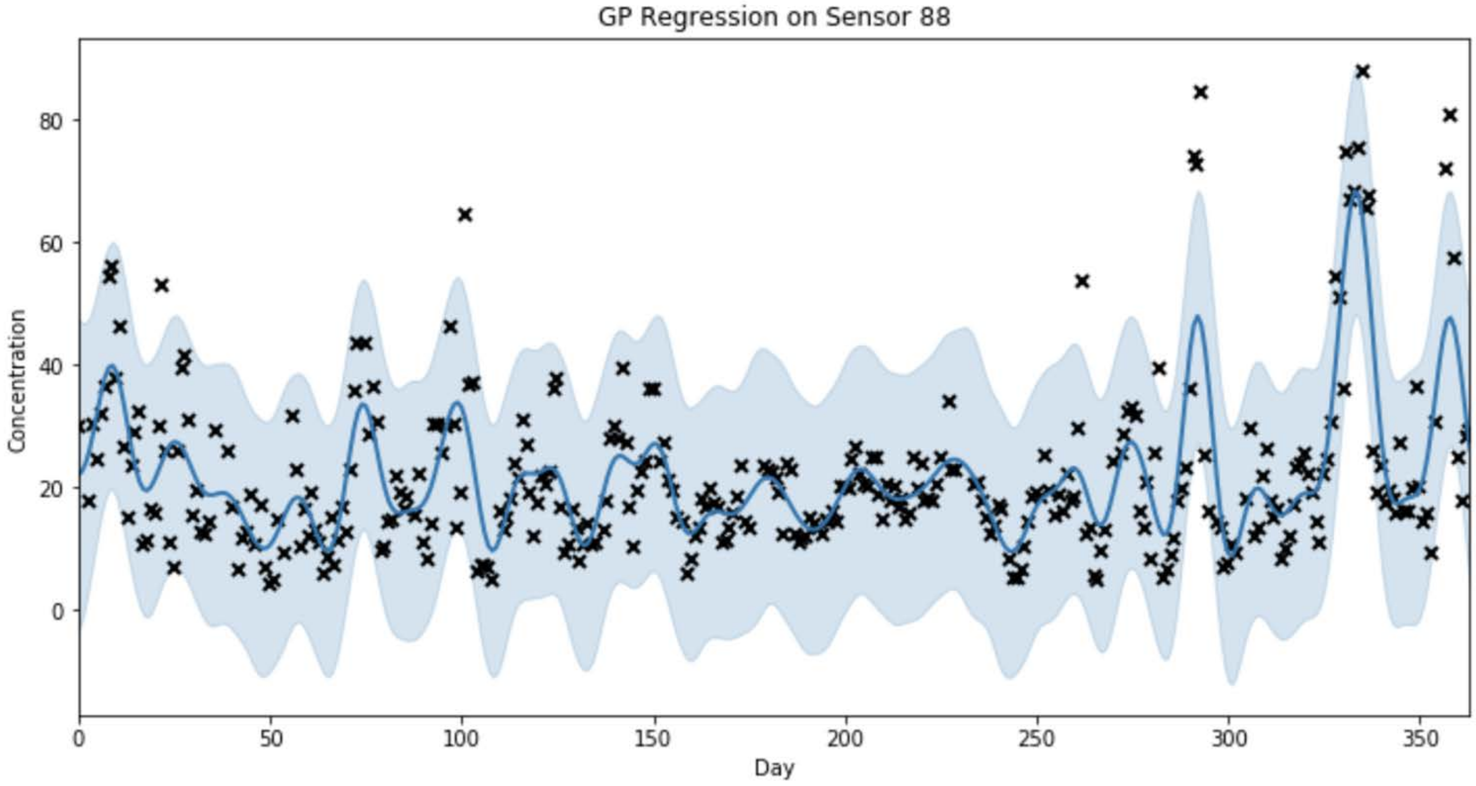
- We also have evidence to suggest that our imputation procedure provides high quality imputations, and since the procedure is quite easy to implement, we recommend that HSPH use it in the future.

Future Work

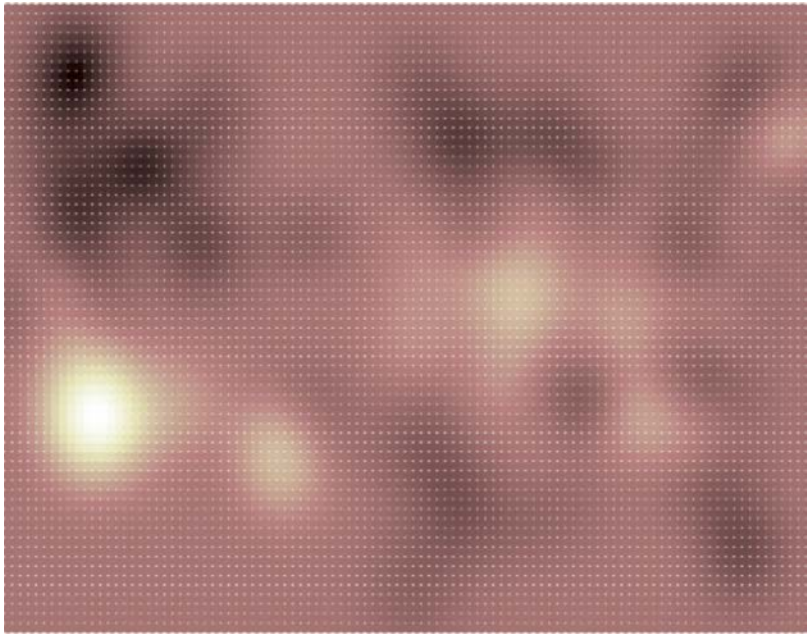
- LSTM, meta-learner
- Uncertainty Quantification
    - More insight into model performance in areas that are far away from sensors
    - Allow for more accurate variance estimates of any associated causal effects
    - Prioritize placement of new sensors



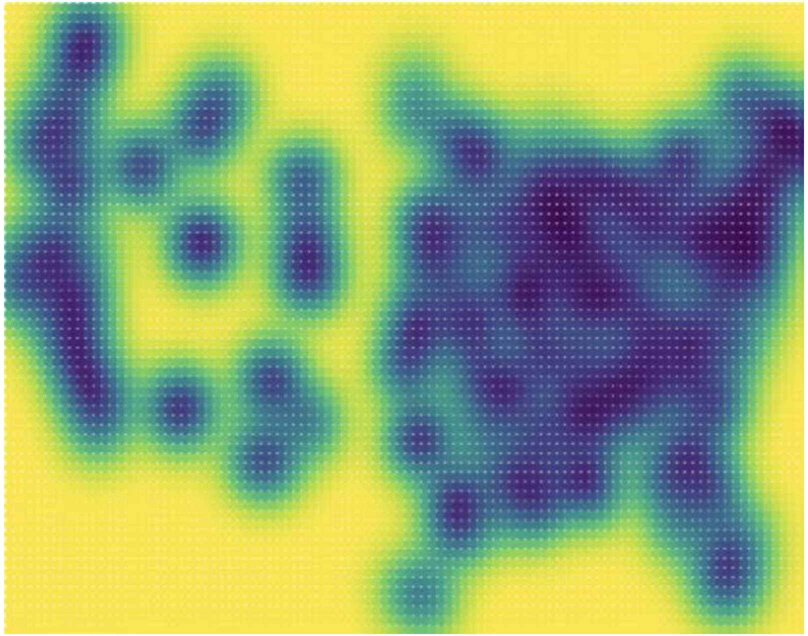
Gaussian Process Demo: Time

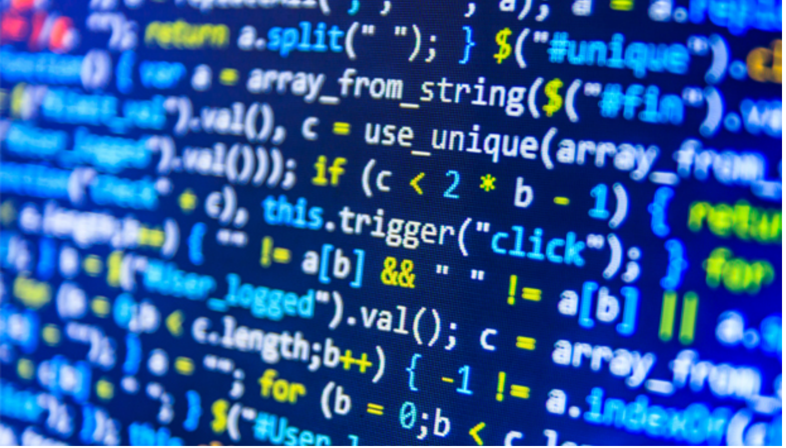


Gaussian Process Demo: Spatial (Mean Est.)



Gaussian Process Demo: Spatial (Variance)





Software Stuff

Extensible

Pipeline

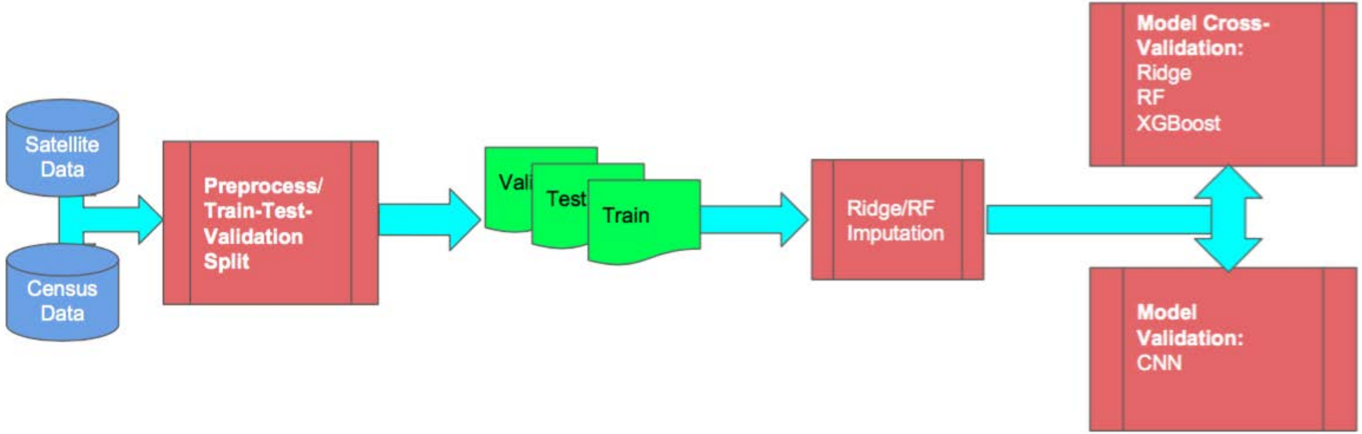
Architecture

Software Stuff

1. Efficient imputation software for overcoming missingness

2. Easy to use machine learning pipeline

3. Extensible package for HSPH to use and develop going forward




Thank you!