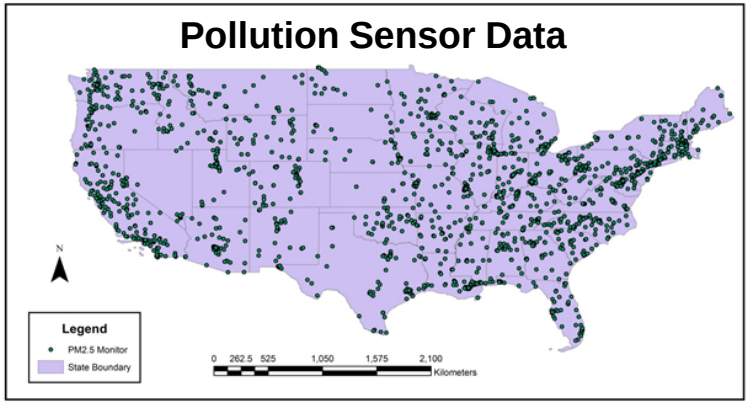
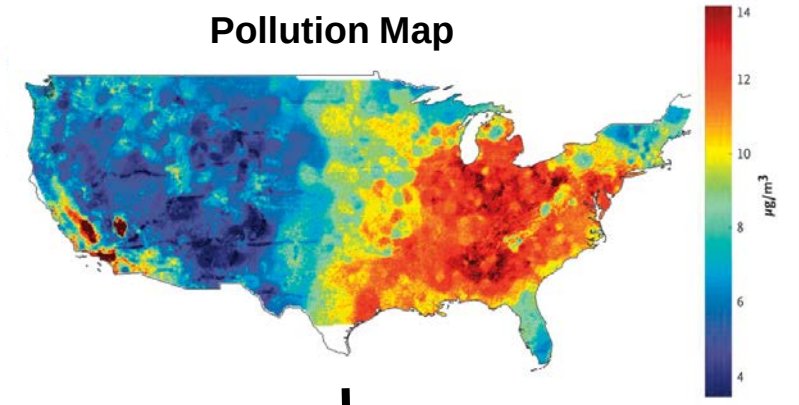
HSPH Capstone Project
Modeling Pollution

Members: Casey, Chris, Justin, and Keyan

TF: David Sondak

Big Picture - Problem & Motivation



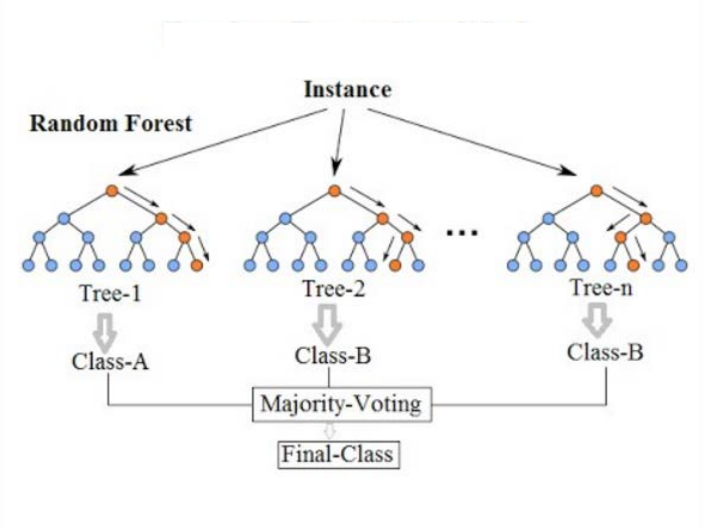


Causal Relationships

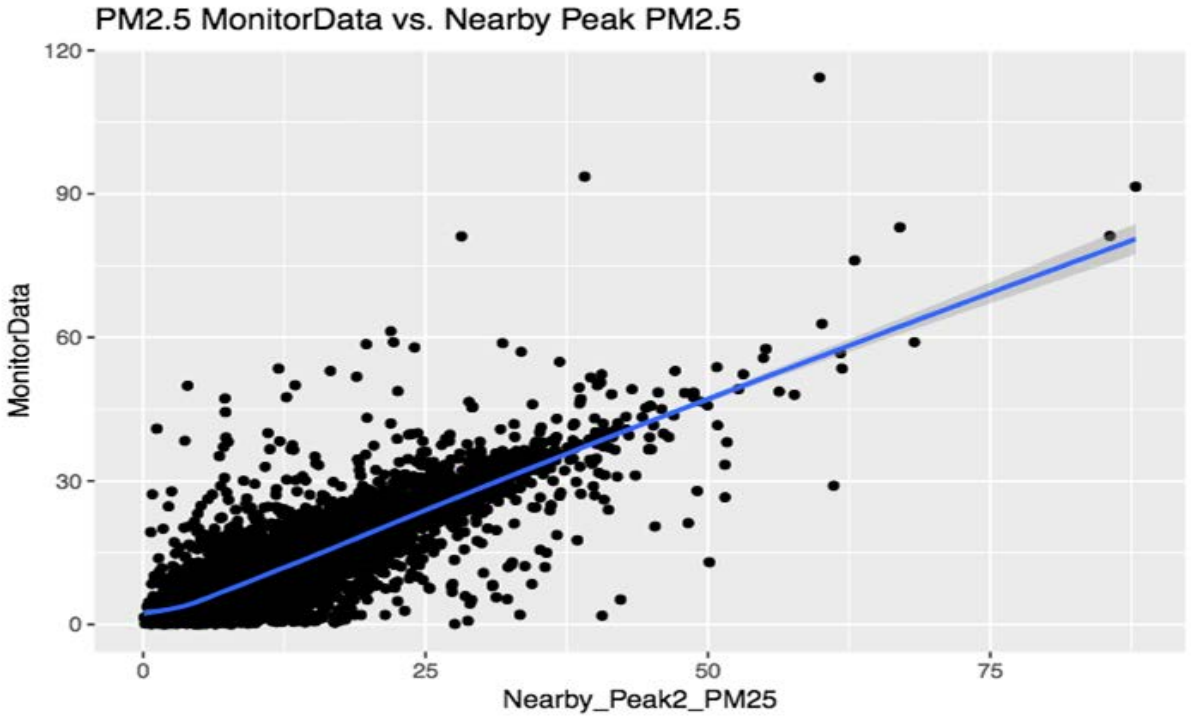
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Big Picture - Missing Data

- There is a lot of missing data in both the predictors and the response
  - We have implemented the iterative “missForest” algorithm to impute missing values



Big Picture - Nearby Pollution



Full-Scale Results

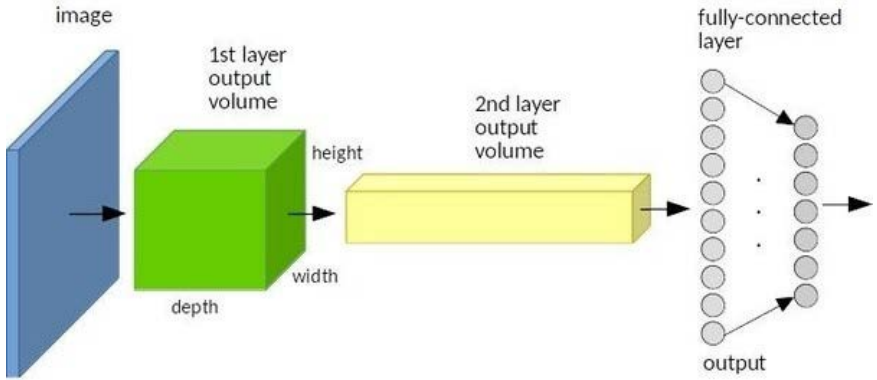
Model	Subset (1%)	Full Data
Ridge	0.76	0.75
Random Forest	0.78	0.77

Current Status - Software Package

- All HPC currently done on Odyssey
- HSPH team works on Harvard/MIT Research Computing Environment (RCE)
- Work with HSPH team to determine best options for turnover and continued use as semester ends

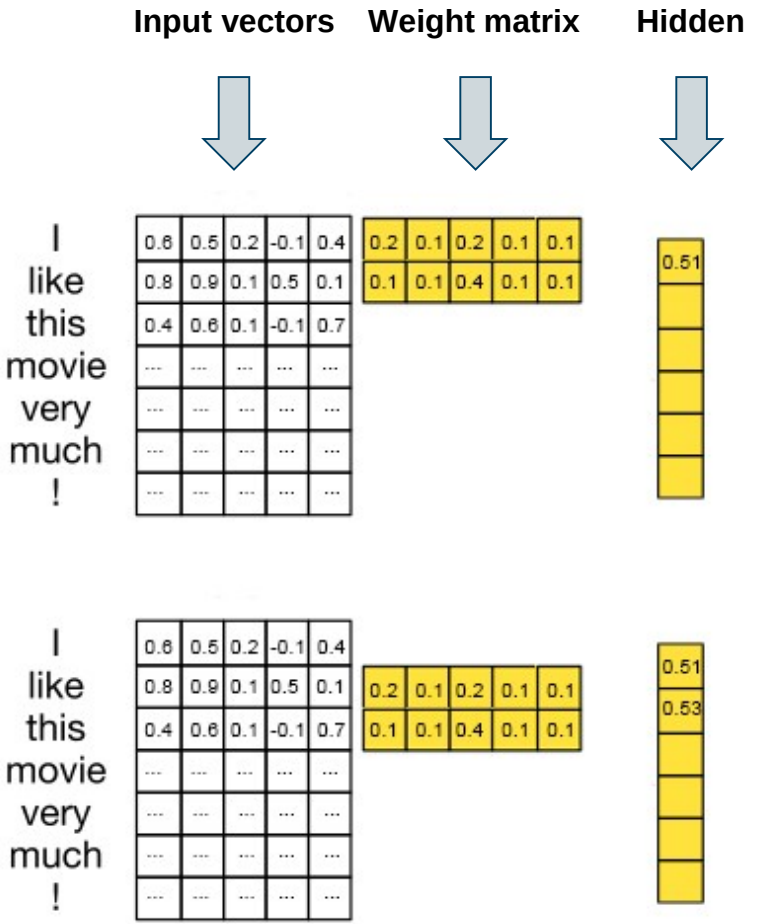
Current Status - Modelling

- Run and validate results of CNN on fully imputed dataset
- Implement Gaussian processes for interpolation and compare results to CNN and other methods



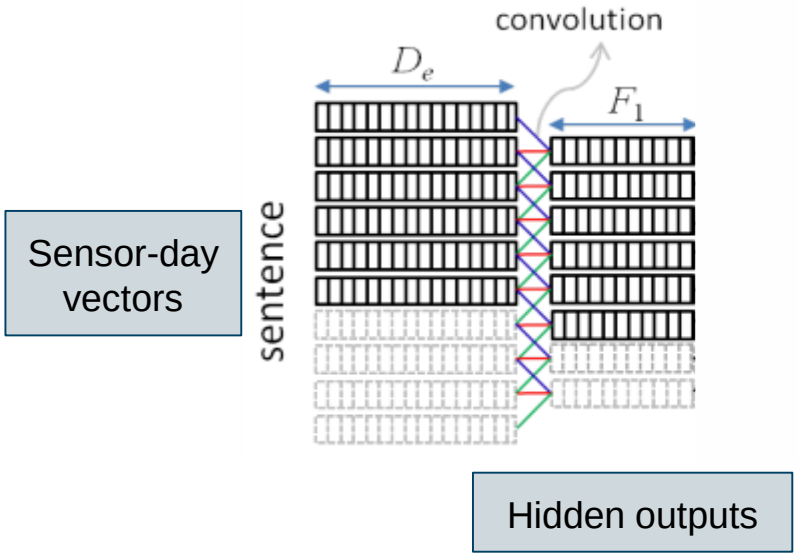
CNN Architecture

- CNN makes sense in this context
  - Days close together are likely to be related in ways that are relevant to pollution
- Usage of CNN in this context analogous to how CNNs used for NLP
  - 1D convolutions



CNN Architecture

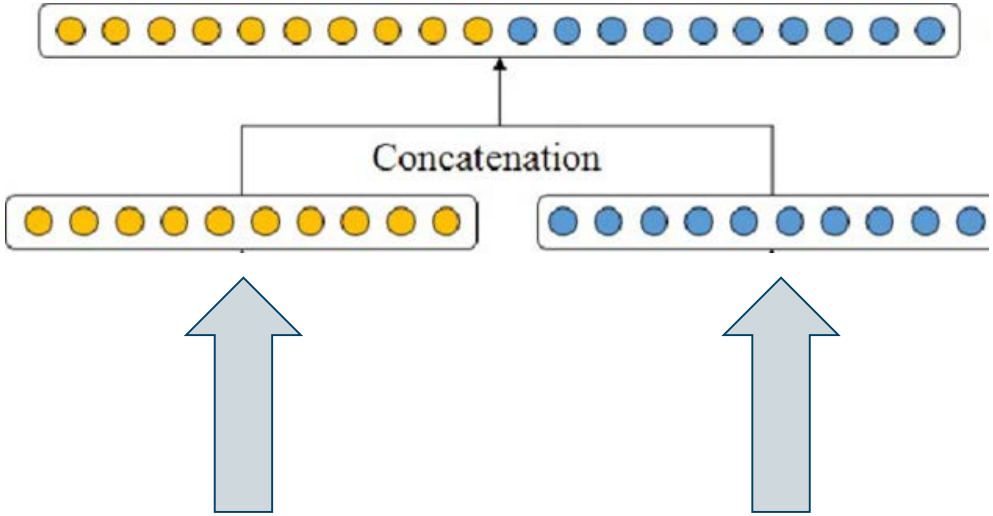
- Use kernel width of size 3 so that features from previous day, current day, day after are used for predicting pollution for current day
  - Will tune kernel width



CNN Architecture

- Inputs to convolutional layer:
  - Features that change on a daily basis within a sensor sequence

- Features that are constant throughout a sensor sequence:
  - Concatenated with hidden outputs from convolutions

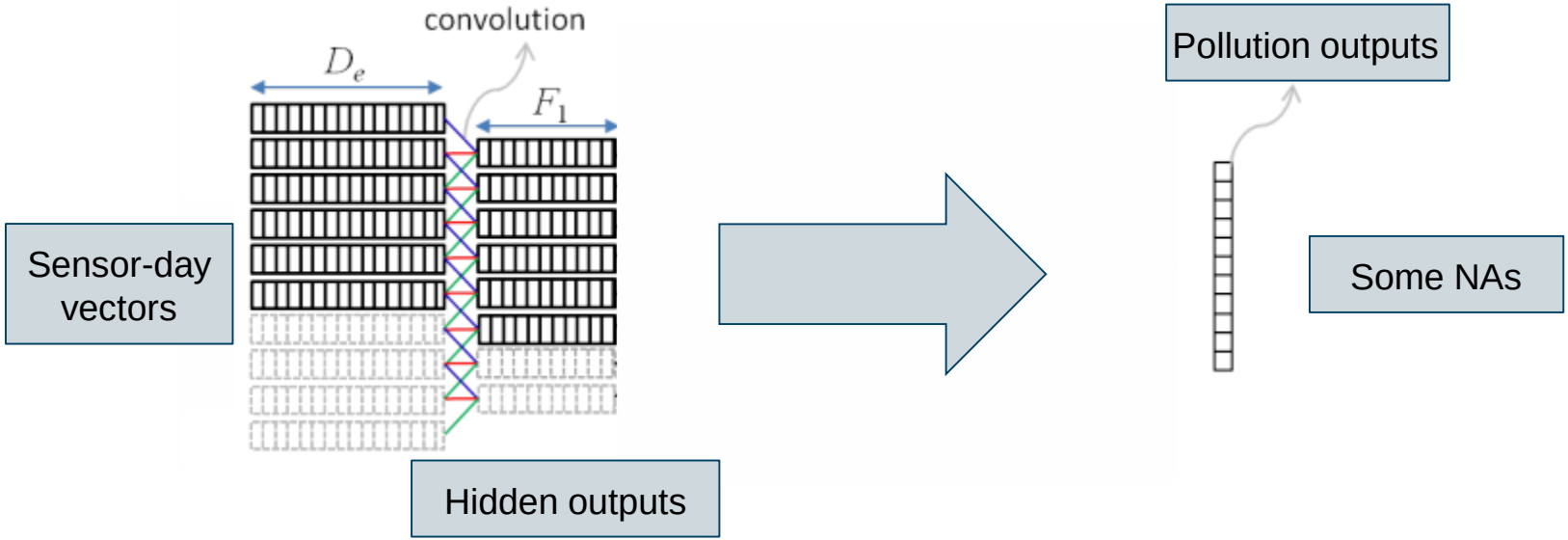


Hidden output from convolution applied to window of days

Features constant over a sensor sequence

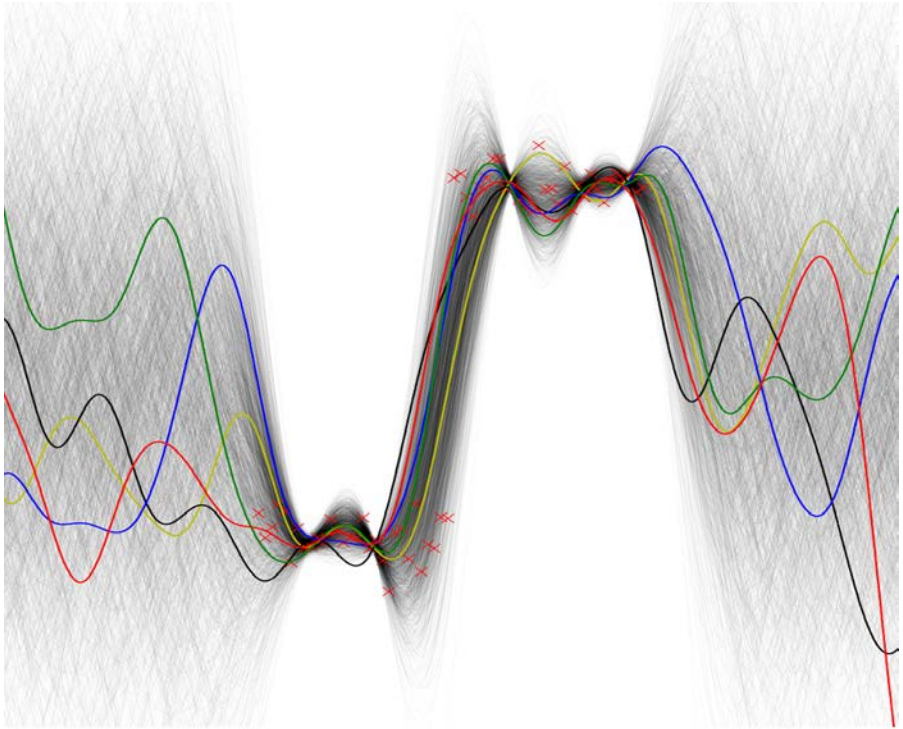
CNN Architecture

- Many days with missing pollution response within each sensor sequence
  - Have to extract hidden outputs for which there exists a pollution response

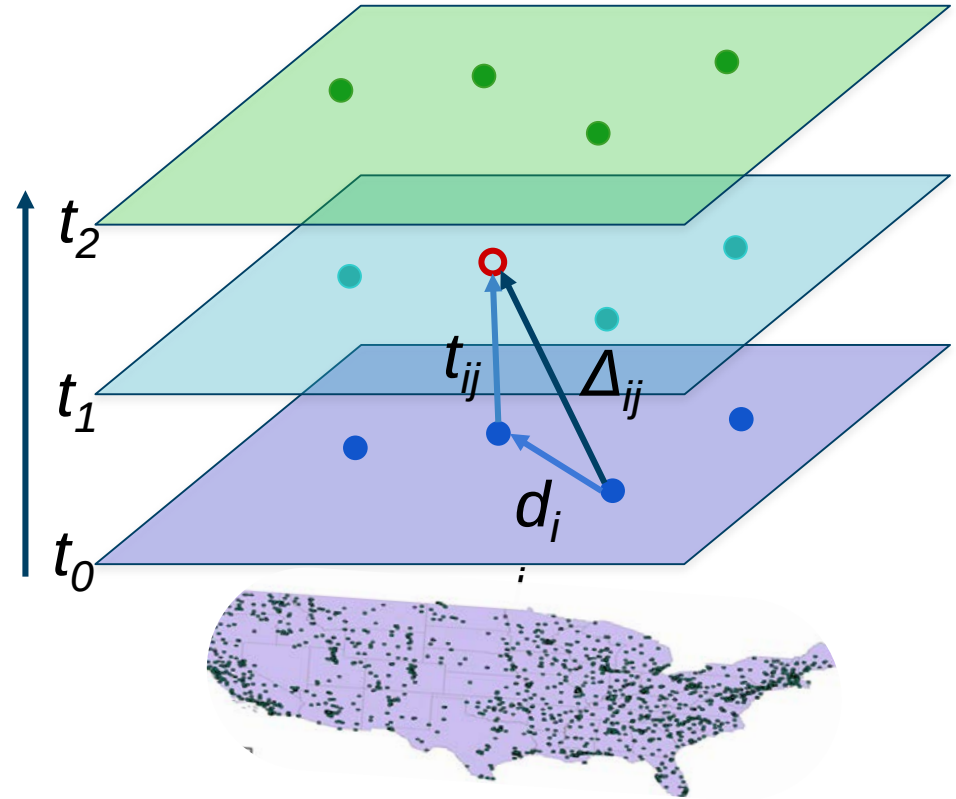


Using Nearby Terms - Gaussian Processes

- Use strong spatial & temporal correlations
- New strategy for predicting in sensor-less areas
- Inherent error estimates

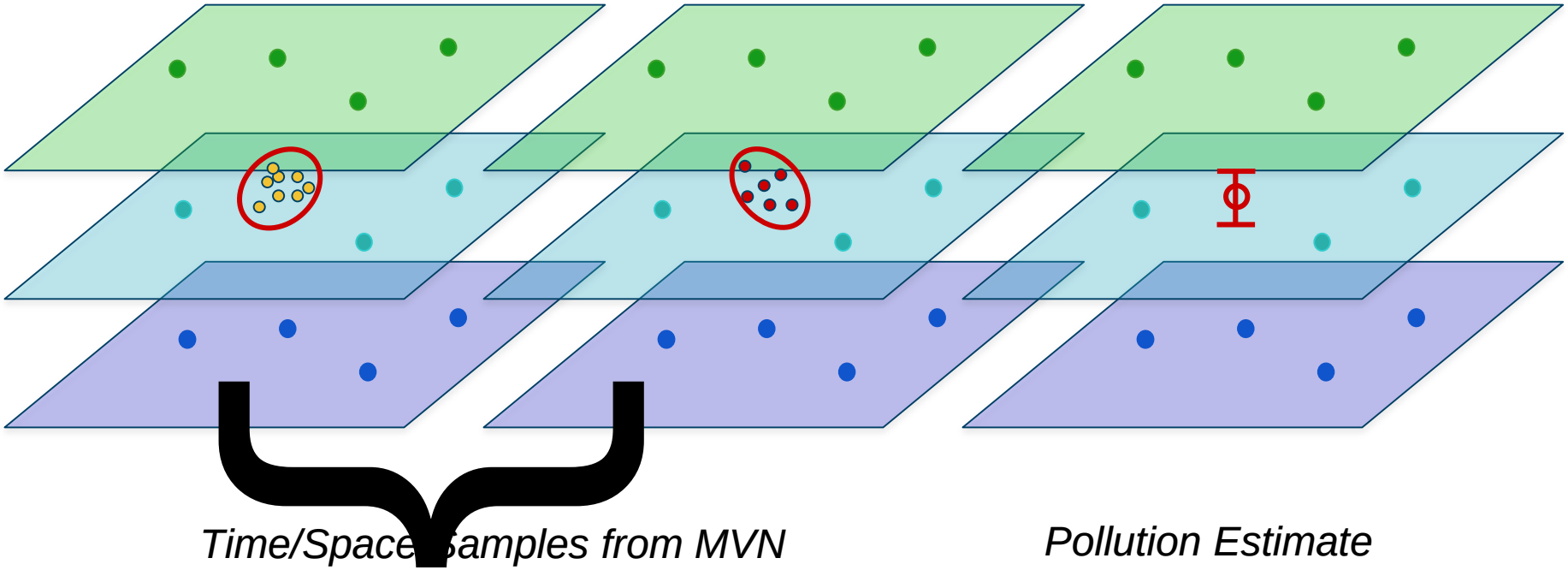


Correlations in Time & Space

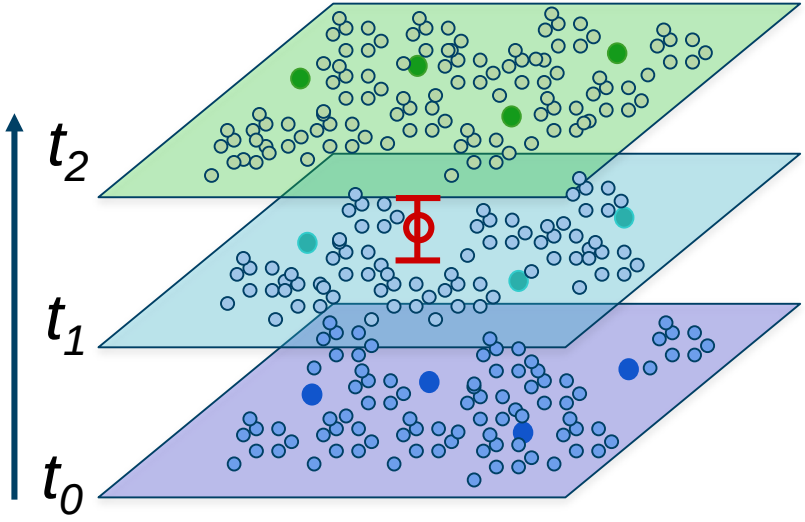


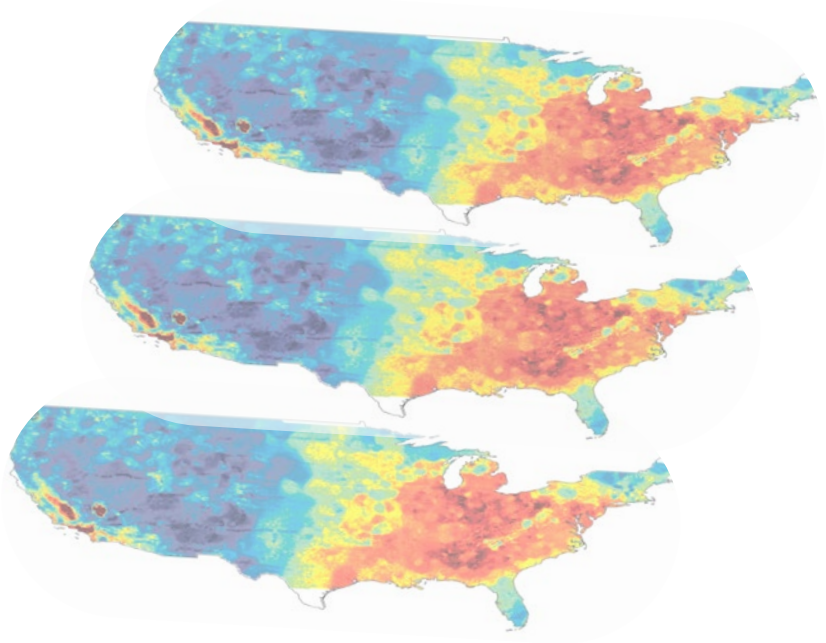
$$\sum_{ij} = \sigma^2 e^{-\frac{\Delta_{ij}^2}{2l^2}}$$

Sampling Provides Pollution Estimates



Sampling Also Estimates Uncertainty





But, Requires Herculean Matrix Inversion