

# EPA Tools and Resources Training Webinar: Virtual Beach

Mike Cyterski

*US EPA Office of Research and Development*

**June 3, 2021**

# Presentation Outline

1. Software Overview
2. Demonstration of Data Analysis Using Virtual Beach version 3
3. Introduction to Web-Based Virtual Beach

# What Problem Does Virtual Beach Address?

- Water quality may be costly, time consuming to measure
- Public health decisions often need to be timely
- 2000 BEACH Act amendment to the Clean Water Act:
  - EPA studies pathogens/indicators and issues criteria
  - “Coastal” states have 3 years to adopt these standards
  - EPA provides grants for monitoring and assessment of rec waters

# What is Virtual Beach?

- [Virtual Beach](#) (VB) is a decision support tool for the development of statistical models of water quality at site-specific locations
- [Version 3](#) – desktop software
- Valuation methods: linear regression and gradient boosting – a decision-tree based machine learning technique
- The user provides environmental features, such as:
  - Rainfall
  - Water temperature
  - Turbidity
  - Wave height
  - Number of beachgoers, dogs, birds
  - Nearby tributary discharge
- Water quality estimates used for site-based management decisions

# History of Virtual Beach

- Developed closely with stakeholders
  - State beach coordinators and managers from Great Lakes states like WI, MI, and OH
- Version 3 developed with USGS Water Science Center, Middleton, WI
- USGS Water Science Center in Columbus, OH
  - Model beach water quality and microcystin concentrations at public water intake sites
- State and local shellfish managers in SC, NC, GA, FL
- Great Lakes states - issuing swimming advisories
- Synergy with ShellBase and EPA Office of Water's new Sanitary Survey App
- Training workshops around the Great Lakes and Atlantic Coast
  - Participants from all over North America, Europe, even Guam!

# Demonstration of Virtual Beach version 3

## Example Dataset: Alaska

# New Version Being Developed: Web-Based Virtual Beach

- Provides powerful statistical techniques (machine learning pipelines)
- Simple user interface
- Browser-based
- Private user accounts
- Study sites, data files, and models saved online
- **Emphasis on cross-validation** for assessment of expected predictive performance
- Evaluation methods

Linear regression, gradient boosting, support vectors (radial basis functions)

Douglas Patton, Environmental Economist – ORISE Fellow

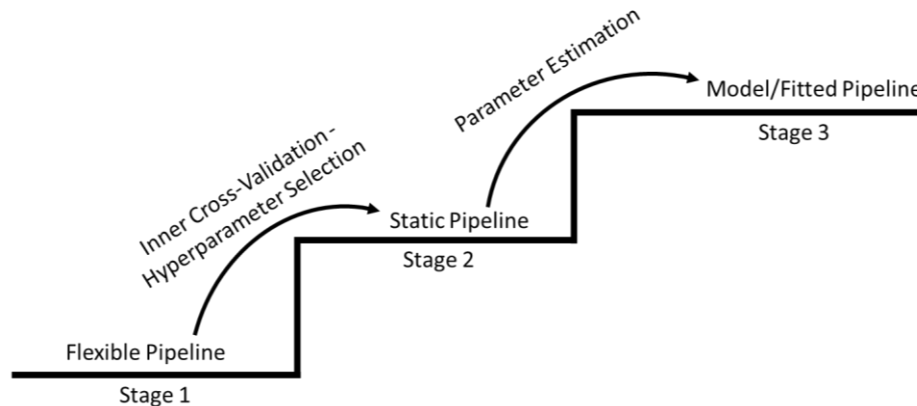
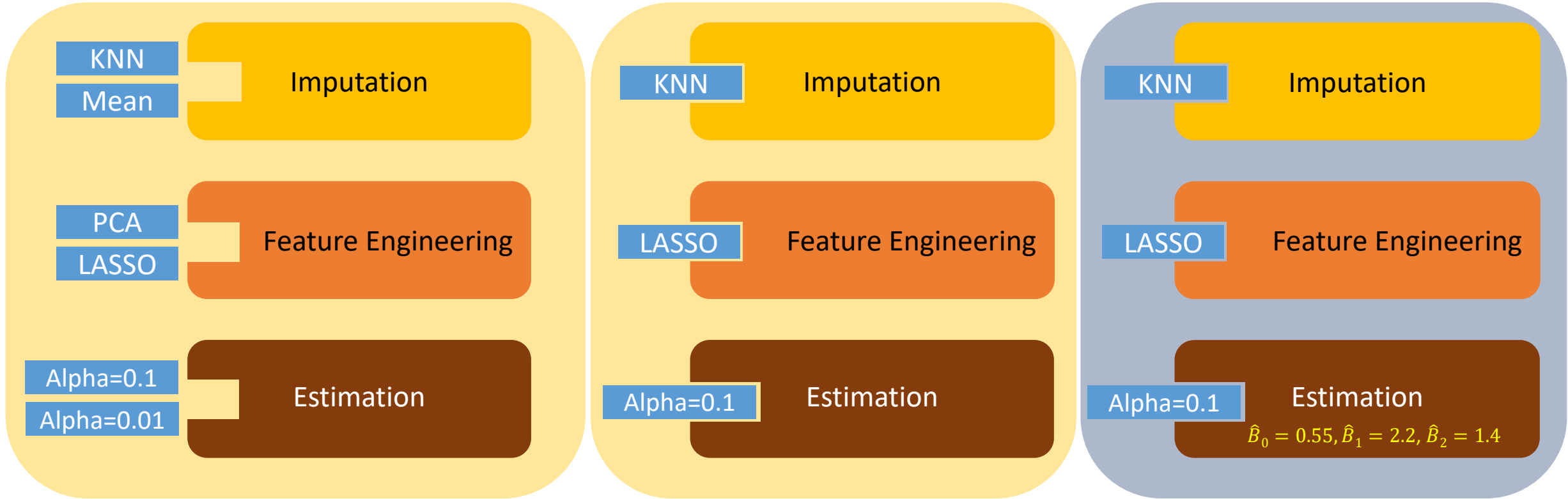
The ORISE Fellowship Program at the U.S. EPA, Office of Research and Development, Athens, GA, is administered by the Oak Ridge Institute for Science and Education through Interagency Agreement No. DW8992298301 between the U.S. Department of Energy and the U.S. Environmental Protection Agency

# Anatomy and Life Cycle of a Pipeline

Flexible

Static

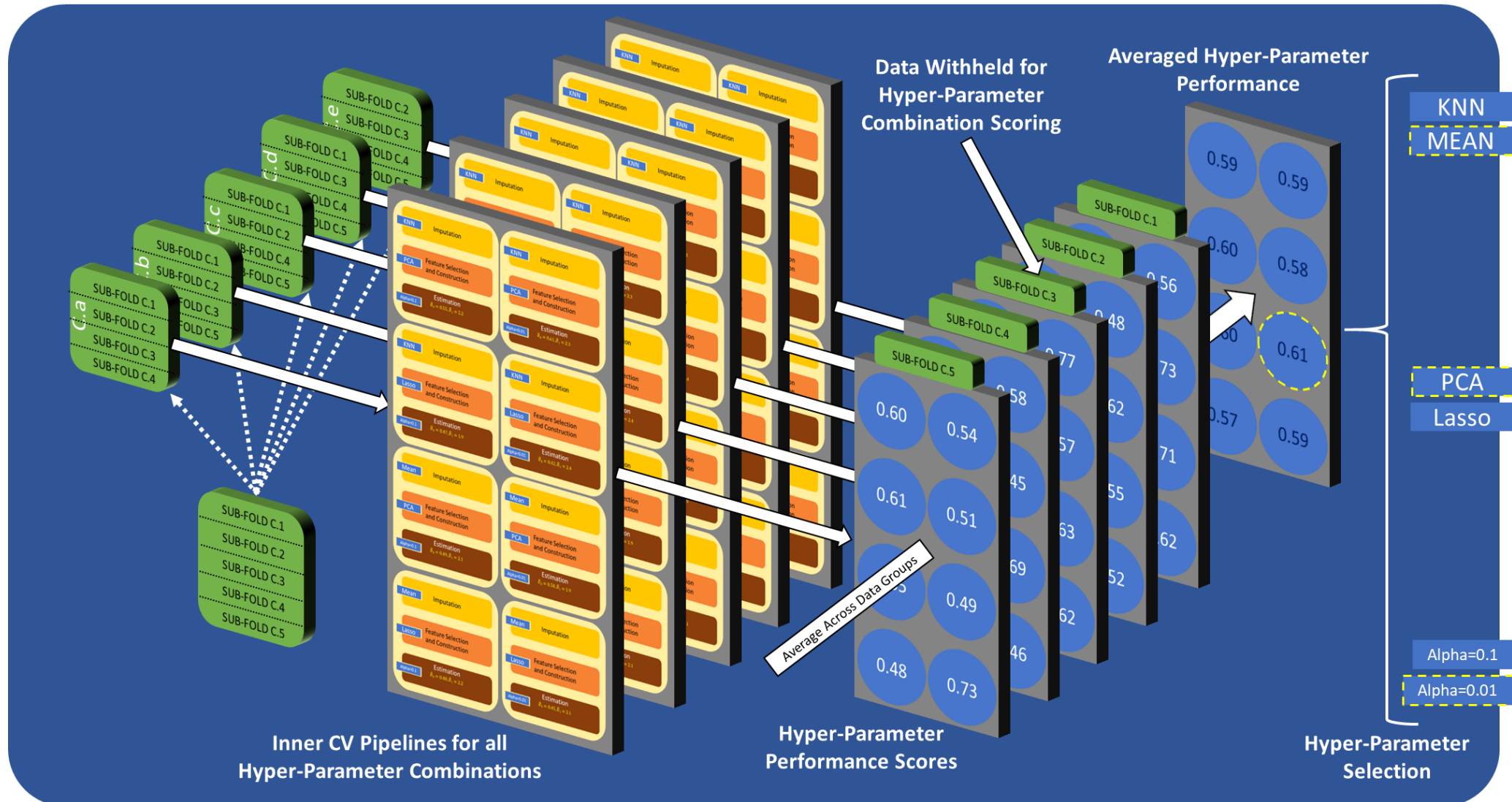
Model (Fitted)





# Web-Based Virtual Beach

## Inner Cross-Validation for Hyperparameter Selection



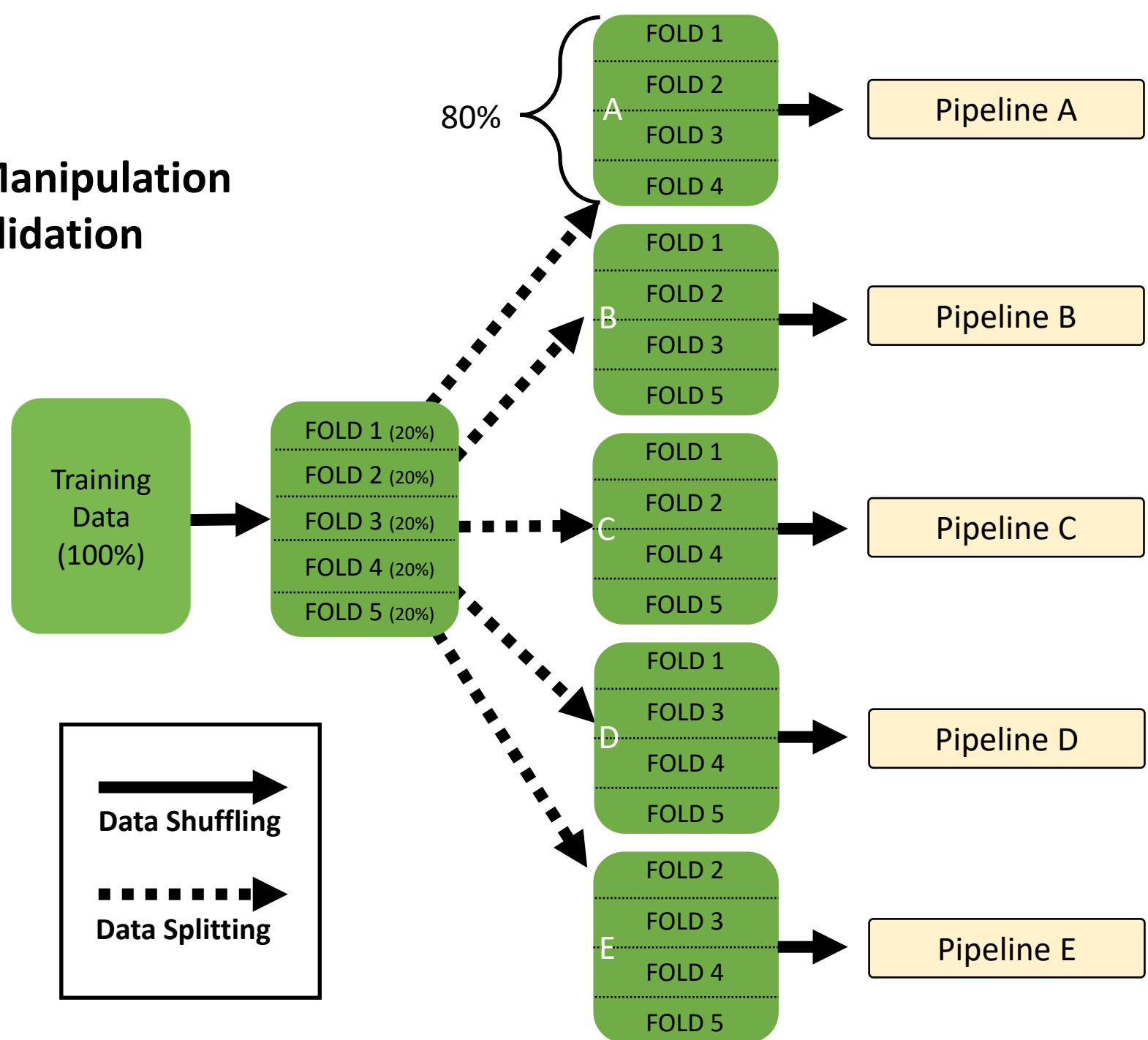
# Web-Based Virtual Beach

Two tasks within a VB project

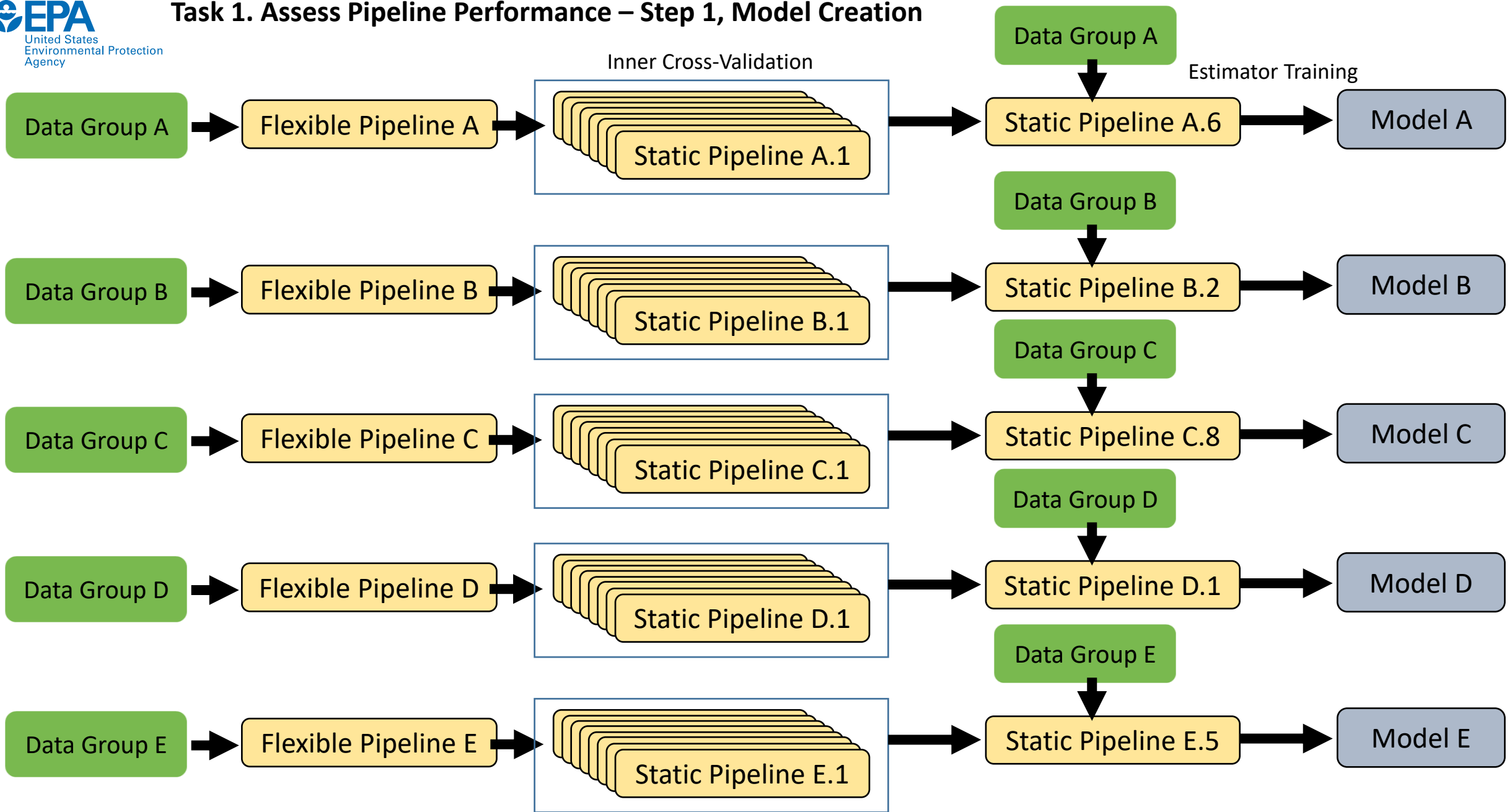
1. Assess *Pipeline* Performance (via Cross Validation)
2. Create a Final *Model* for Prediction

# Training Data Manipulation

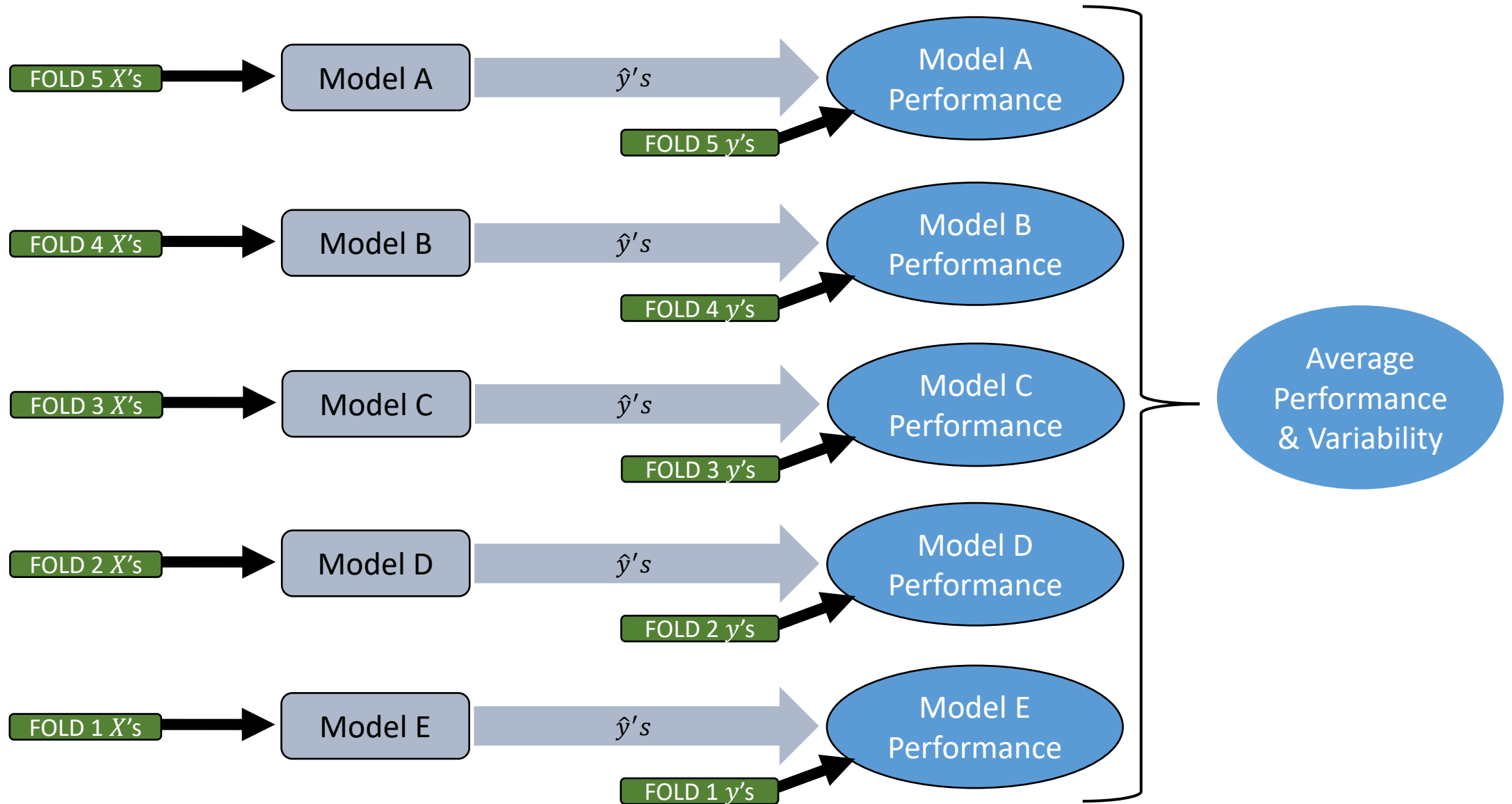
## 5-Fold Cross Validation



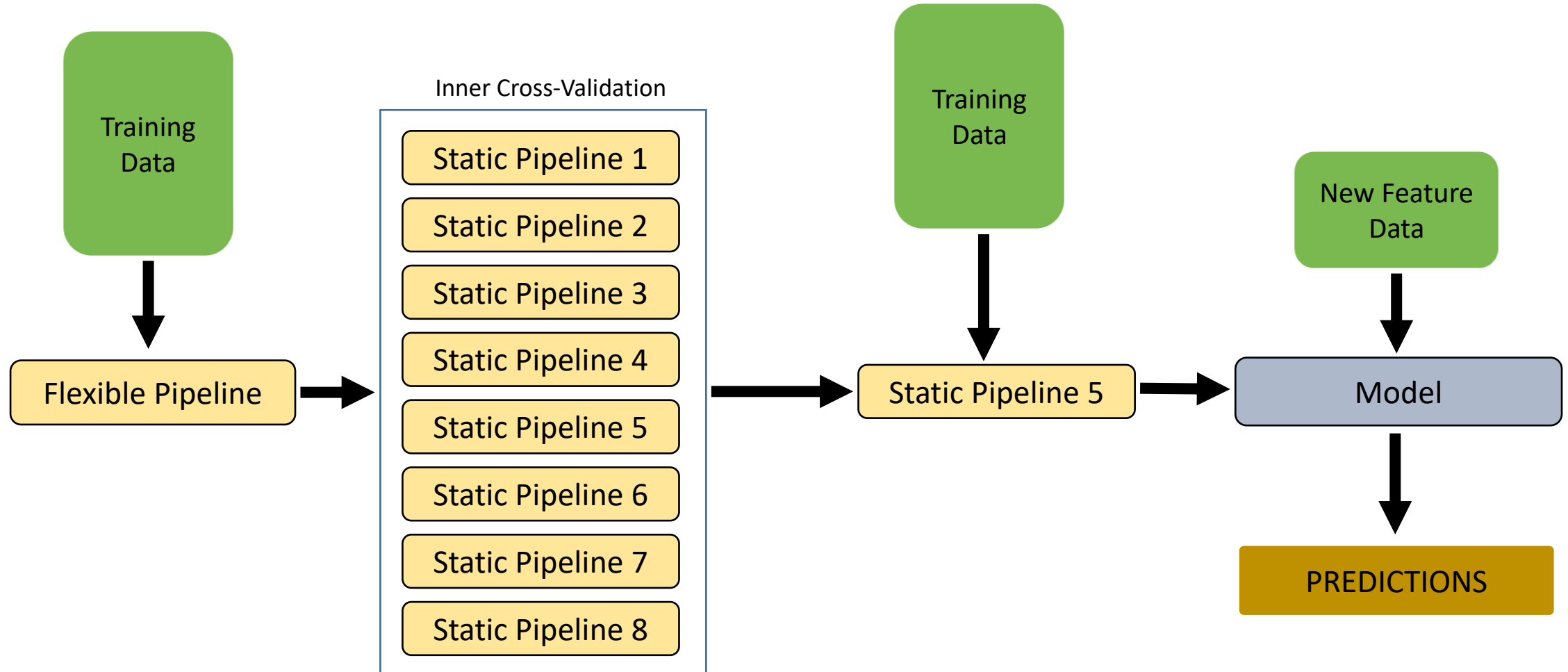
# Task 1. Assess Pipeline Performance – Step 1, Model Creation



## Task 1. Assess Pipeline Performance – Step 2, Calculate Average Performance



## Task 2: Develop a Model for Prediction



# Web-Based Virtual Beach Screenshots

## Location Mapping

Location Name: test location

Description: test location

Start Latitude: 29.65225560712191

Start Longitude: -94.09790039062501

End Latitude: 29.5232805008286

End Longitude: -94.46594238281251

Water Latitude: 29.403747057881503

Water Longitude: -94.21743383357212

Buttons: Fit map to beach, Save Location, Cancel, Clear

## Data Processing

test project this is a VB test project

Location: testlocation

Training Data

Dataset name: VB\_Data.csv

Dataset description:

Start row: End row: Selected rows: Total rows:

Buttons: Add range to selected rows, Select all rows, Clear selected rows

row	ID	Response	x1	x2	x3	x4	x5	x6	x7	x8	x9
0	49234	0.68	0.107117881	0.010985825	0.091048341	0.095417387	0.035454453	0.091338698	0.438714702	0.237704525	0.228544433
1	49235	1.98	0.63350822	0.004305831	-0.490986951	1.102006681	0.454200213	-0.783678663	1.223303746	0.664611271	0.63058086
2	49236	2.01	0.409501161	-0.007498787	0.069102227	0.618009793	0.2462478	-0.325194669	0.576934984	0.550451012	0.43186140
3	49237	3.54	0.152286488	0.026781817	-0.150537516	0.541950228	0.136512109	-0.243491605	0.434945197	0.245502777	0.56189749
4	49238	3.74	0.362999731	0.094217024	0.064188077	0.232064504	0.11992262	0.040068285	0.3481715	0.299046428	0.24972552

Items per page: 5 1 - 5 of 172

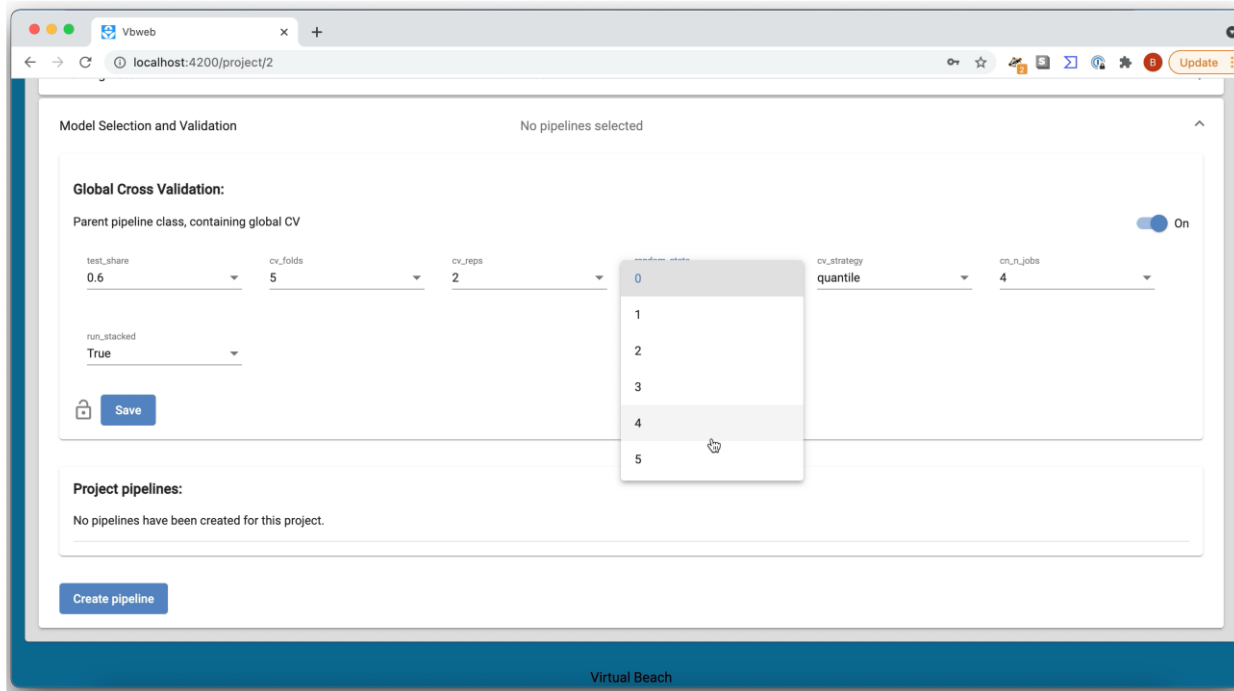
Response variable: Feature variables:

Define components to generate A/O values?  Define regulatory value?  Something about co-linearity?  Something about training split?

Buttons: Save Dataset, Cancel

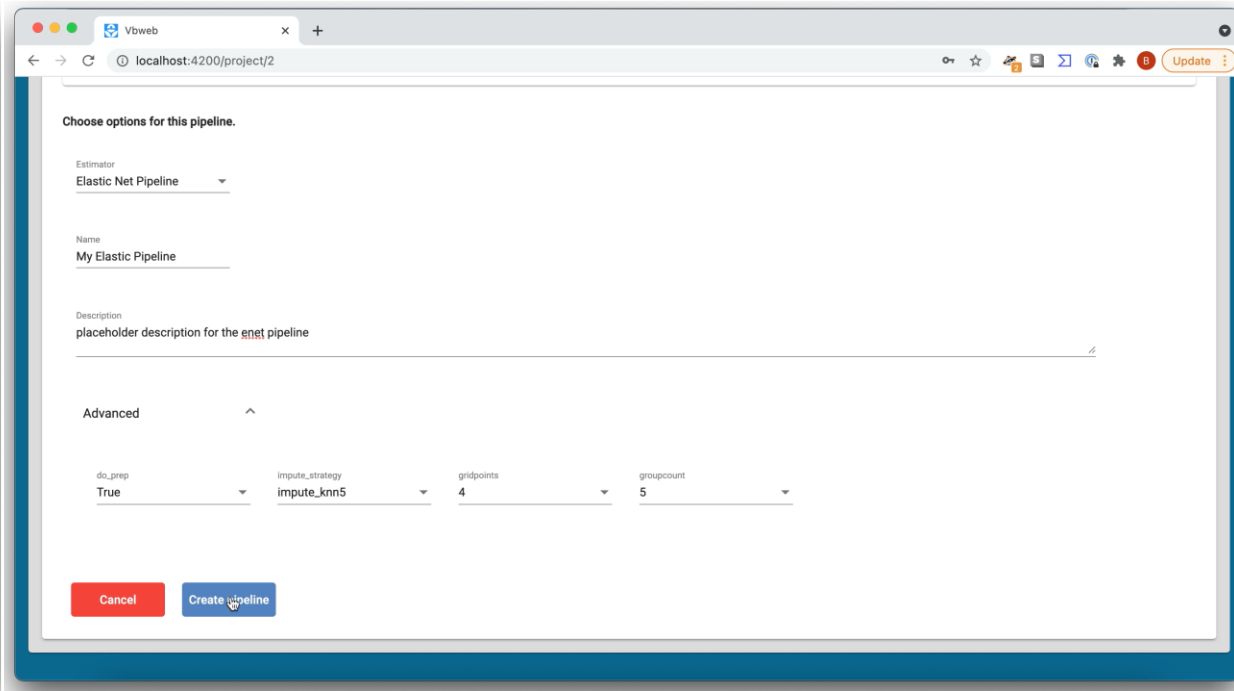
# Web-Based Virtual Beach Screenshots

## Cross Validation Setup



The screenshot shows the 'Model Selection and Validation' section of the Vbweb application. The page title is 'No pipelines selected'. Under the 'Global Cross Validation' section, there is a toggle switch for 'Parent pipeline class, containing global CV' which is currently turned 'On'. Below this, several configuration options are visible: 'test\_share' is set to 0.6, 'cv\_folds' is 5, 'cv\_reps' is 2, 'cv\_strategy' is 'quantile', and 'cn\_n\_jobs' is 4. A dropdown menu for 'cv\_strategy' is open, showing options from 0 to 5. The 'run\_stacked' option is set to 'True'. A 'Save' button is located at the bottom left of this section. Below the global settings, the 'Project pipelines' section indicates 'No pipelines have been created for this project.' and includes a 'Create pipeline' button at the bottom left. The browser address bar shows 'localhost:4200/project/2'.

## Pipeline Creation

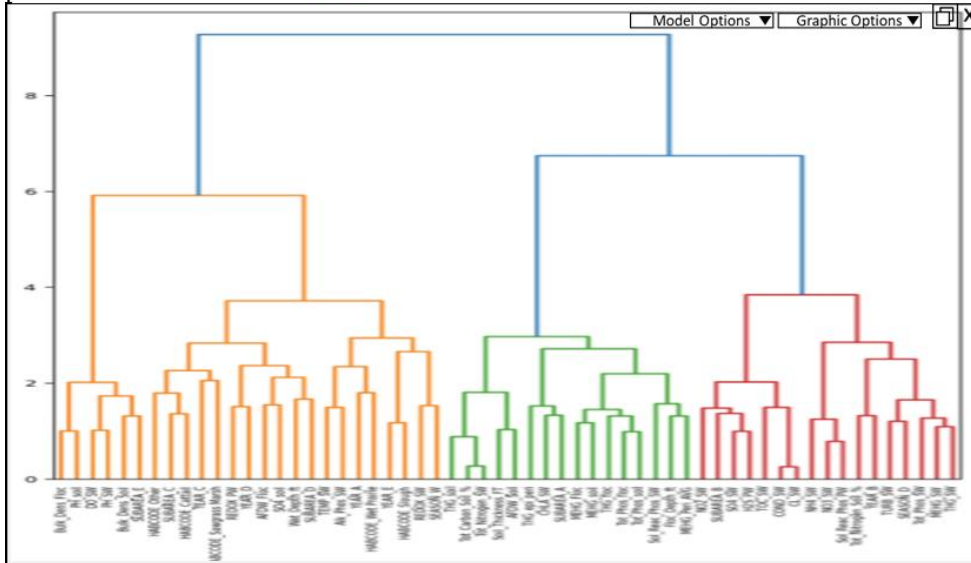
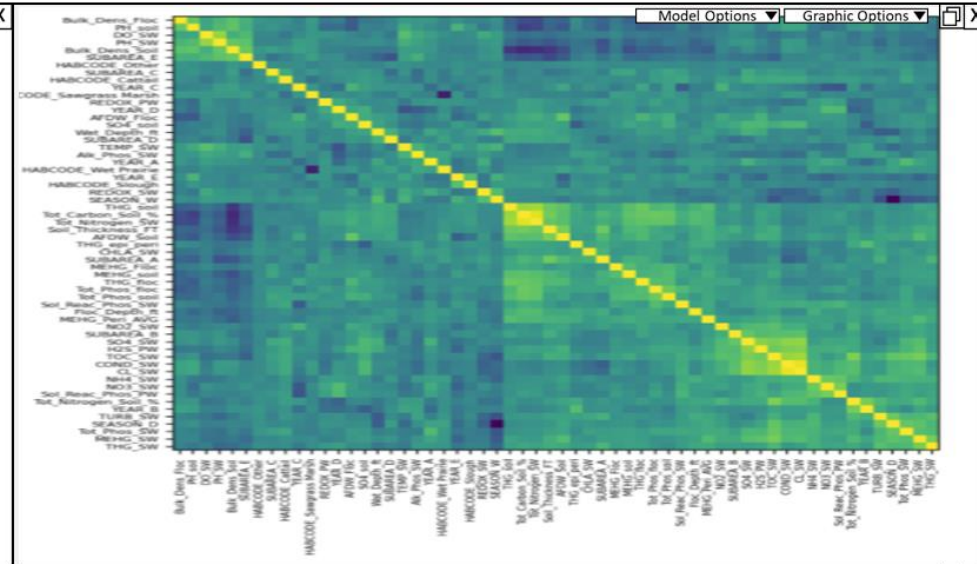
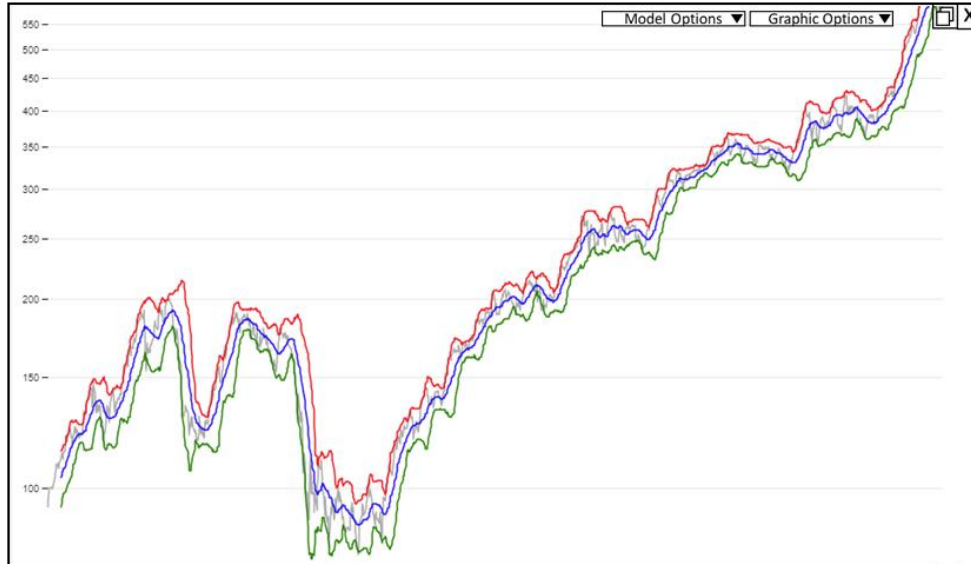


The screenshot shows the 'Choose options for this pipeline.' section of the Vbweb application. The 'Estimator' is set to 'Elastic Net Pipeline'. The 'Name' field contains 'My Elastic Pipeline'. The 'Description' field has a placeholder text: 'placeholder description for the enet pipeline'. Under the 'Advanced' section, 'do\_prep' is set to 'True', 'impute\_strategy' is 'impute\_knn5', 'gridpoints' is 4, and 'groupcount' is 5. At the bottom, there are 'Cancel' and 'Create pipeline' buttons. The browser address bar shows 'localhost:4200/project/2'.



# Web-Based Virtual Beach

## Training Dashboard



GeneID	LogFC	AvgExpr	Log2Sigma	sym	pval
11429	0.356246902	8.6398525485	-0.4551606329	Aco2	0.05010577
11666	0.3823529975	6.2514159019	-0.3875423468	Abcd1	0.05718461
11416	-0.5572913073	4.8391942468	0.0062846546	Slc33a1	0.05948051
11430	-0.6991014035	4.8267083892	0.5310049989	Acx1	0.0595969
11565	-0.987764648	5.2024262532	1.2303036019	Adssl1	0.06453492
11637	0.3865299927	8.0591071214	-0.0154754416	Ak2	0.06557879
11764	0.3189188265	6.3237657526	-0.8295617077	Ap1b1	0.06564359
11566	-0.415611363	7.559530734	0.1279183933	Adss	0.06660231
11744	0.3476253748	5.9243166713	-0.5841165376	Anxa11	0.0697055
11568	-3.3588422432	-0.3287561535	1.0046262521	Aebp1	0.0714946

# Take Home Messages

Virtual Beach used to create site-specific statistical models for water quality indicators

Current desktop version available at:

<https://www.epa.gov/ceam/virtual-beach-vb>

Web-based version under development

Training materials and support available

## **Mike Cyterski, PhD**

Research Ecologist and Data Scientist  
Center for Environmental Measurement and Modeling  
USEPA Office of Research and Development  
[cyterski.mike@epa.gov](mailto:cyterski.mike@epa.gov)  
706-355-8142

## **Web-VB Software Engineers and Development Support:**

Doug Patton - ORISE, EPA/ORD, Athens, GA  
Deron Smith – EPA/ORD, Athens, GA  
Jason Dunken - ORISE, EPA/ORD, Athens, GA  
Brandon Main - ORISE, EPA/ORD, Athens, GA  
Kurt Wolfe – EPA/ORD, Athens, GA  
Rajbir Parmar – EPA/ORD, Athens, GA

**Visit EPA's VB webpage: <https://www.epa.gov/ceam/virtual-beach-vb>**