The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency

An Analysis of Overfitting In Modern QSAR Models

Nathaniel Charest*, Gabriel Sinclair*, Christian Ramsland*,

Todd Martin**, Antony Williams**

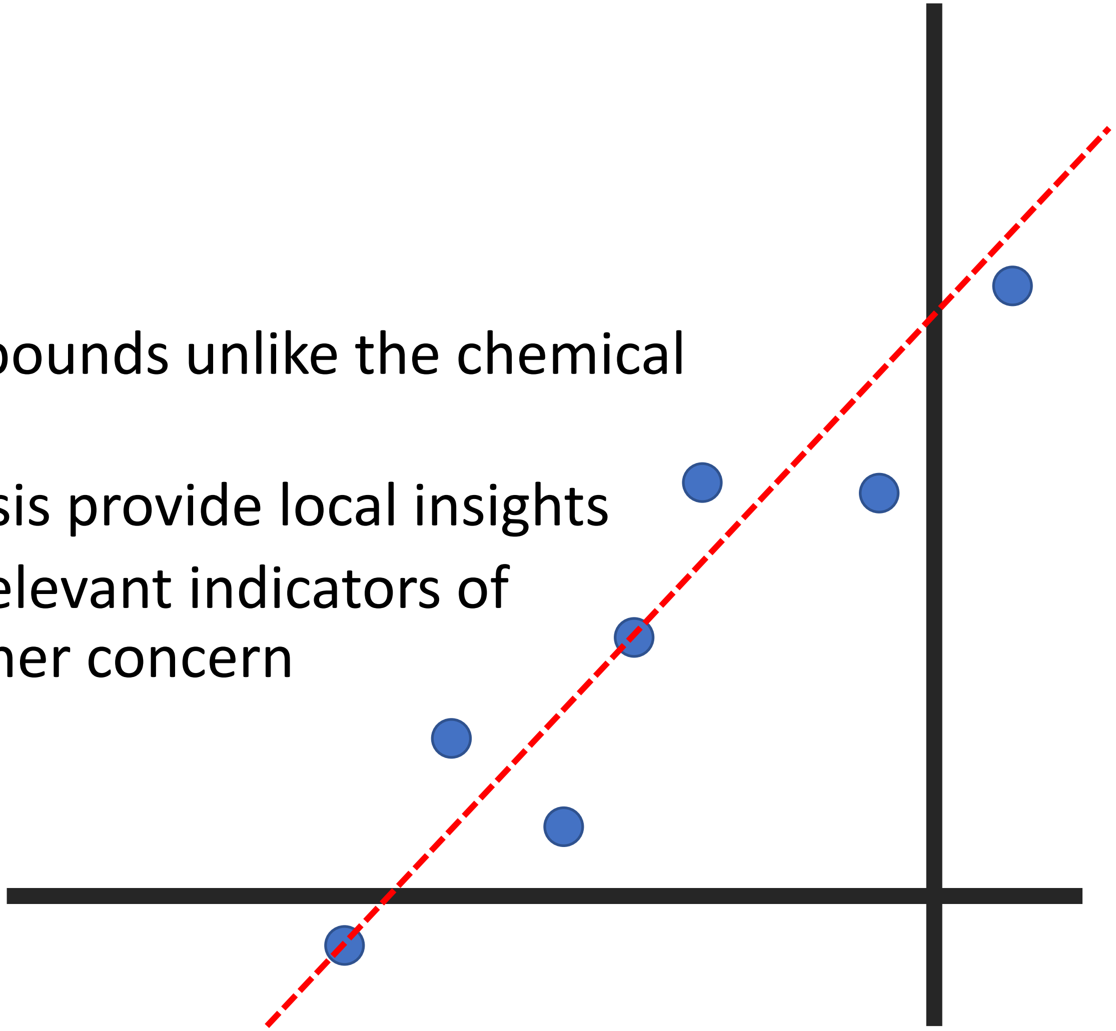
*Oak Ridge Associated Universities, National Student Services Contractor

**US Environmental Protection Agency, Center for Computation Toxicology & Exposure

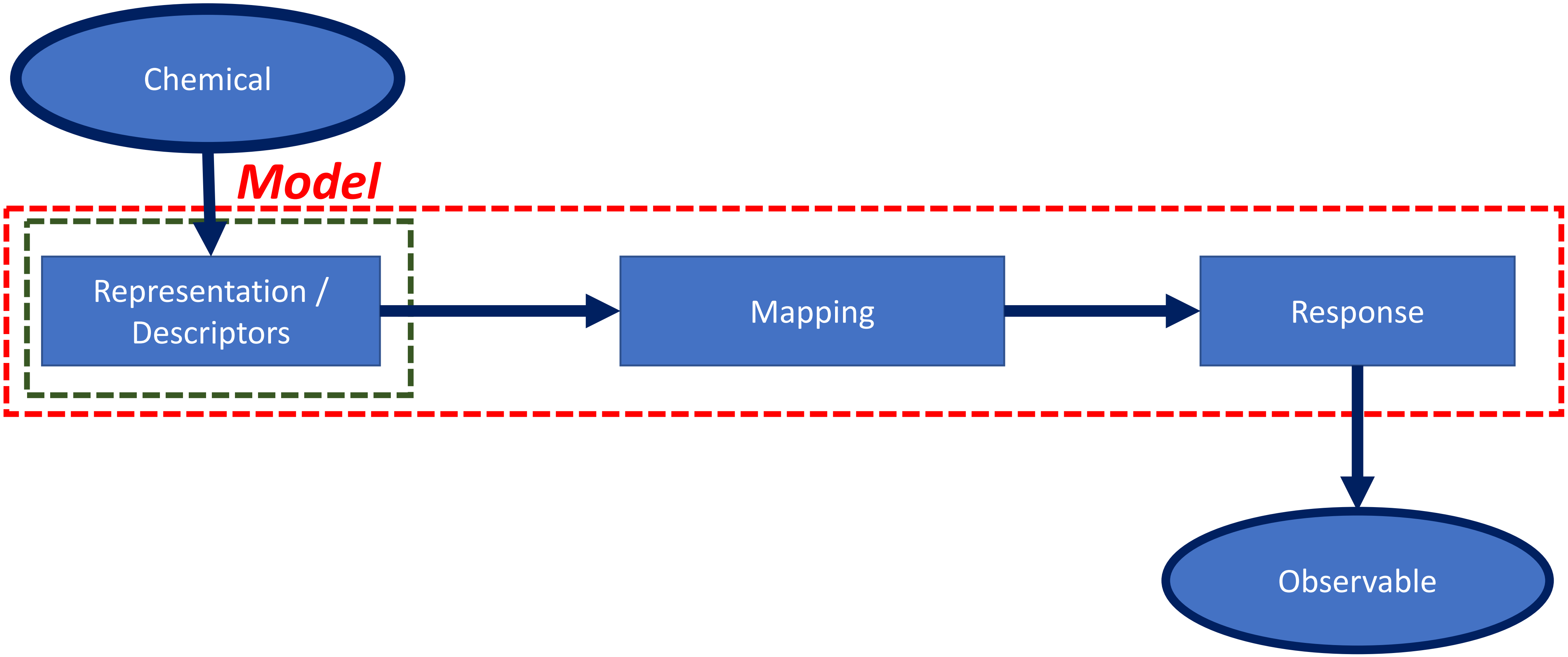
ACS Fall Chicago 2022

QSA/PR Model

- 31(000) Flavors…let’s go with vanilla
- Interpretability For Regulation
- Global vs. Local
  - Global models theoretically can flag compounds unlike the chemical space of training data
  - Techniques like GenRA or analogue analysis provide local insights
  - Regulators seek abstractions of globally relevant indicators of toxicity, environmental persistence, or other concern



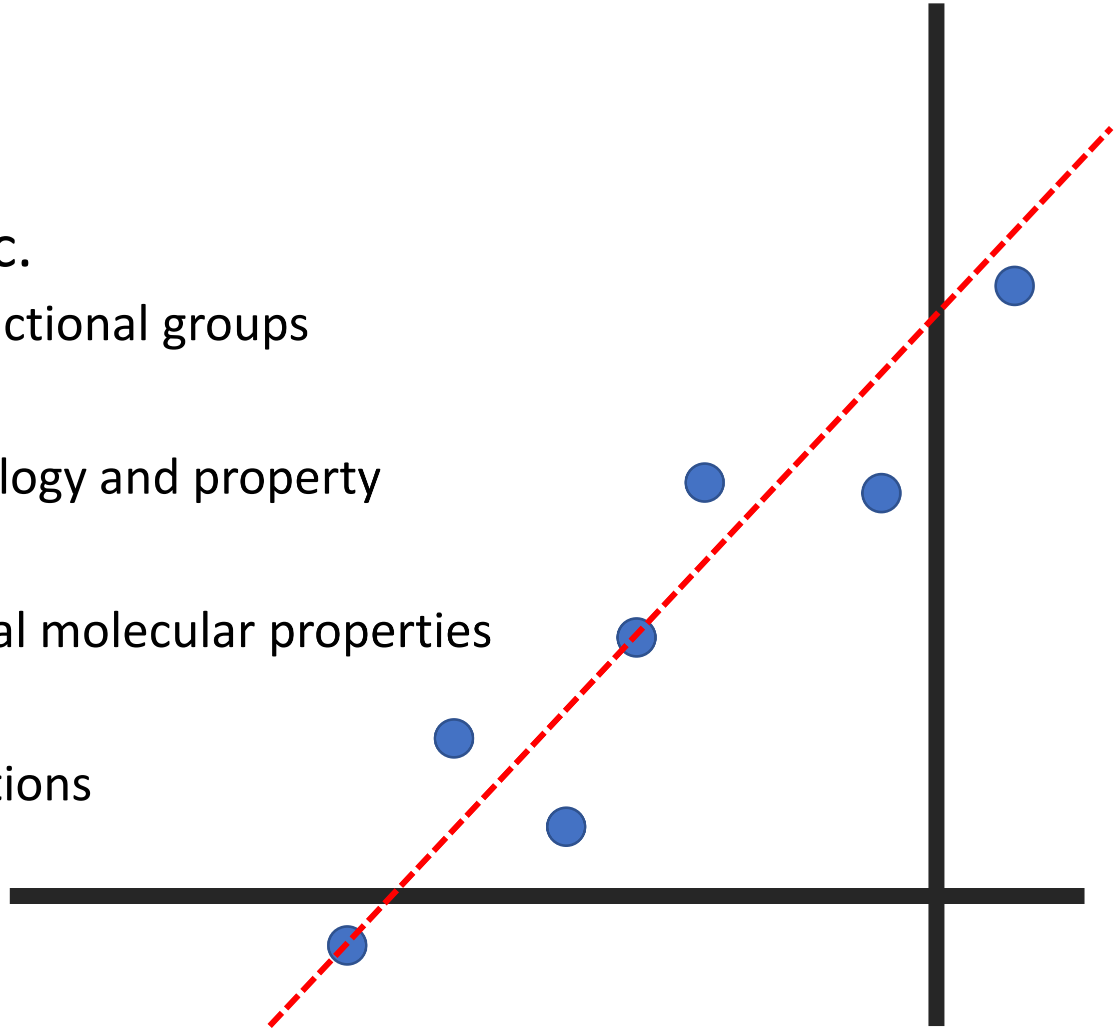
QSA/PR Model



Chemical
Model
Representation / Descriptors
Mapping
Response
Observable

QSA/PR Model

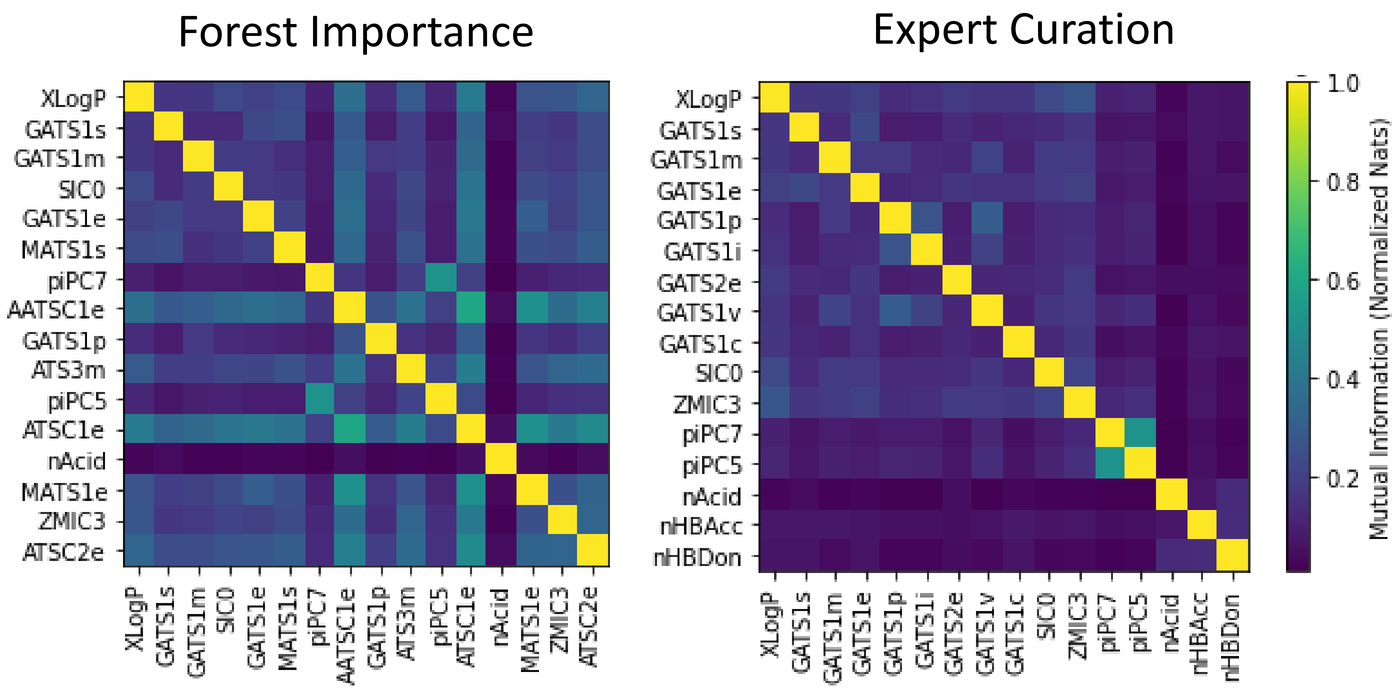
- Representation Matters
- “Descriptors”
  - Structure counts, fingerprints, SMILES, etc.
    - Embeds chemistry as glyphs representing functional groups
  - Physiochemical indices
    - Embeds chemistry as reals representing topology and property
  - Constitutional
    - Embeds chemistry as reals representing global molecular properties
  - Semi-empirical model predictions
    - Embeds chemistry as low-level model predictions



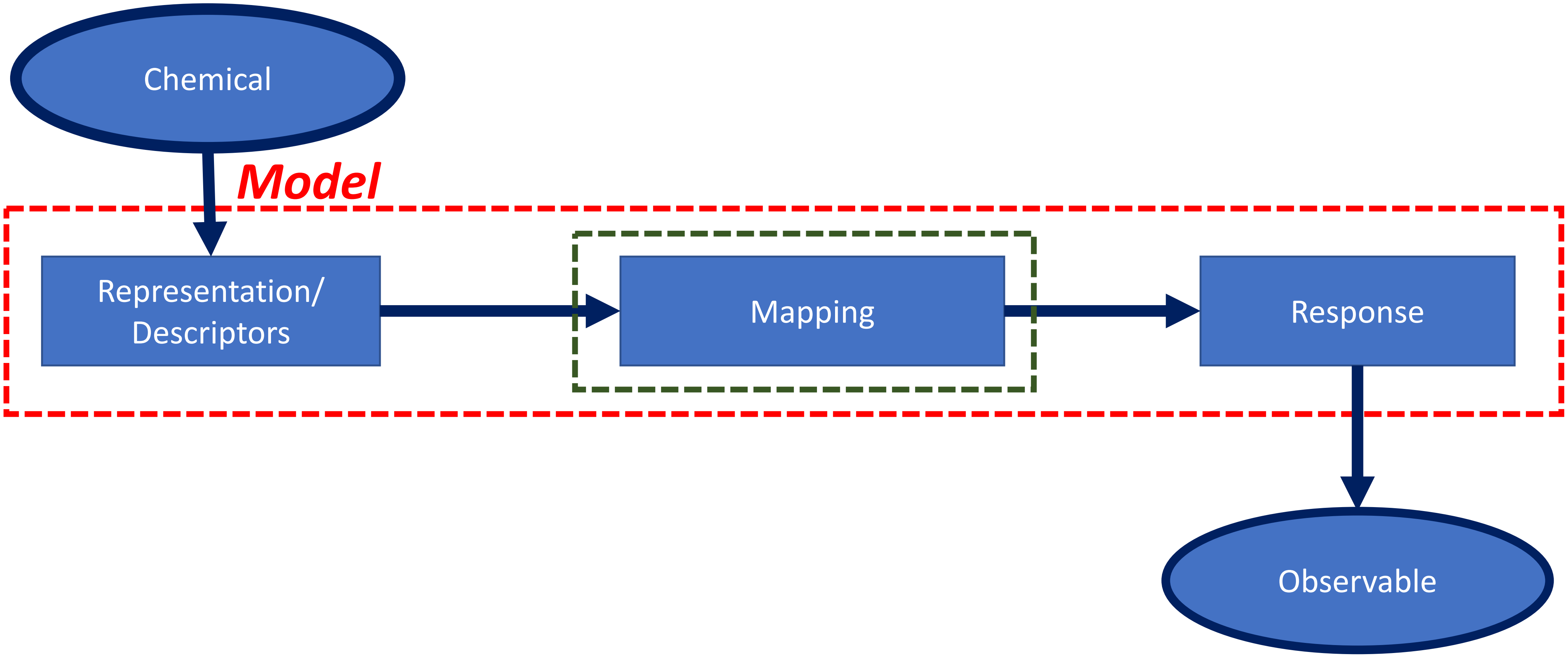
Automated Descriptor Selection

- Algorithmic selection can overrepresent informatically entangled facets of structure
- Depending on the structure of the dataset, this can “over-localize” the mechanisms described by the model

Mutual Information of Descriptors

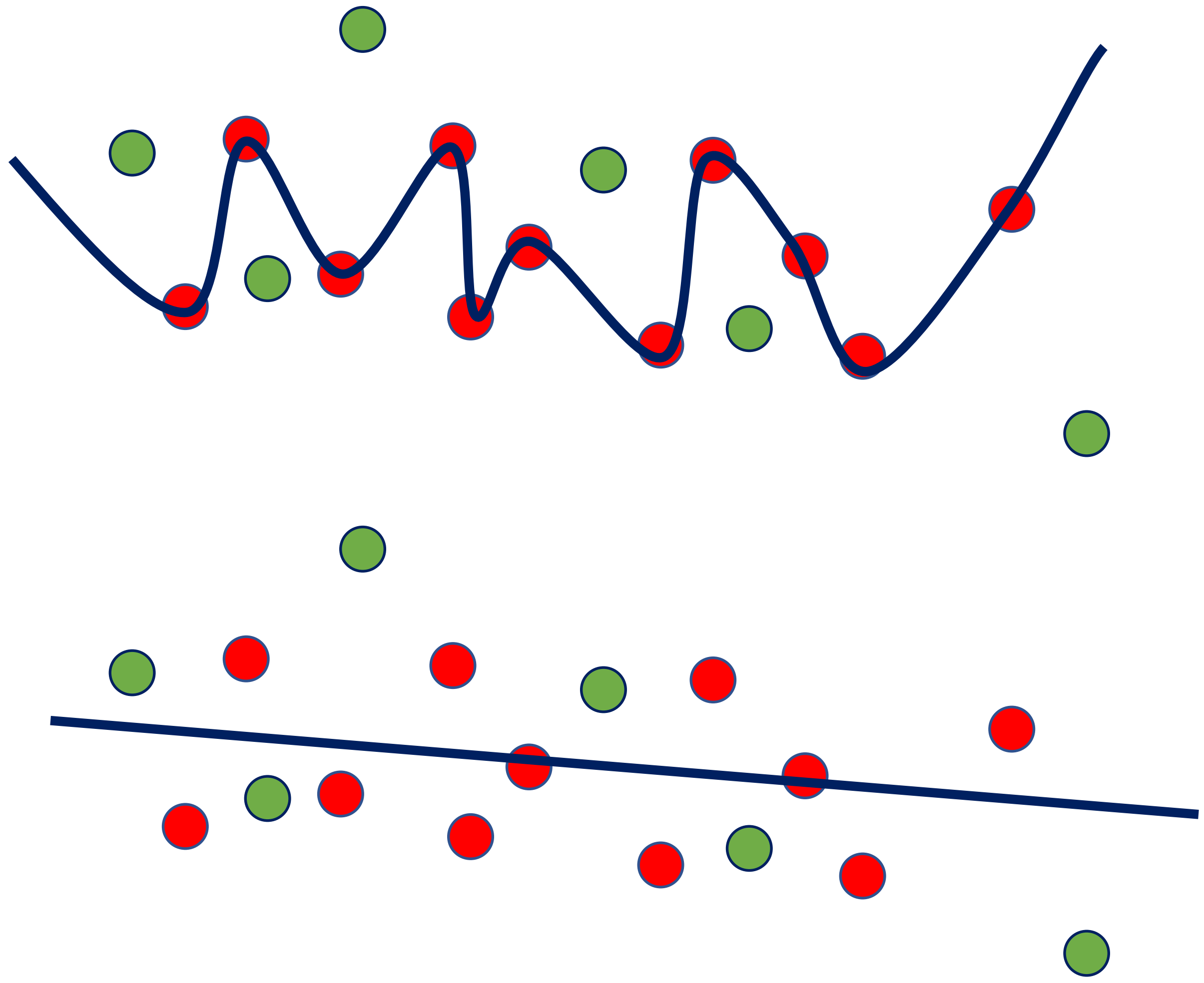


QSA/PR Model


Chemical
Model
Representation/
Descriptors
Mapping
Response
Observable

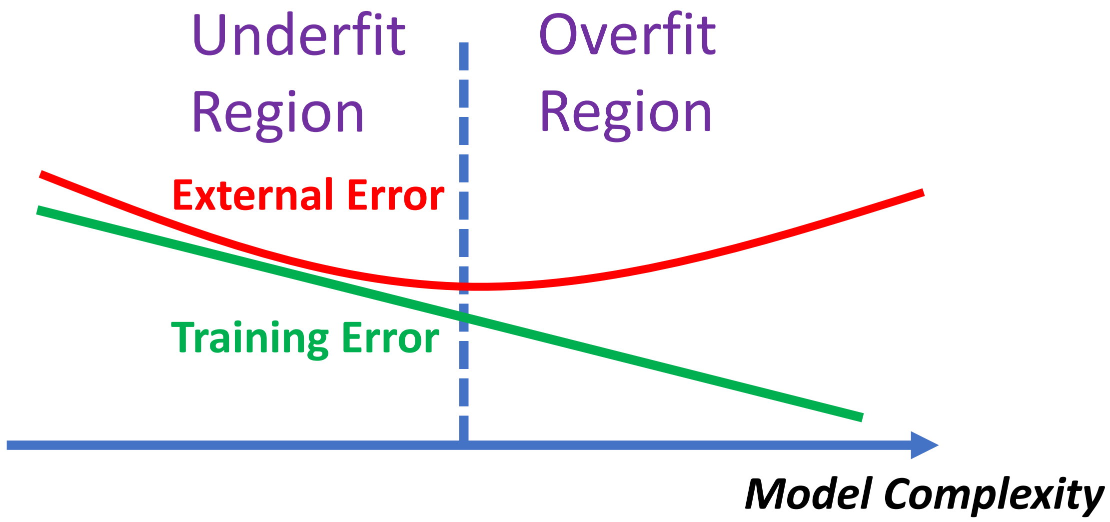
The Traditional Case of Overfitting

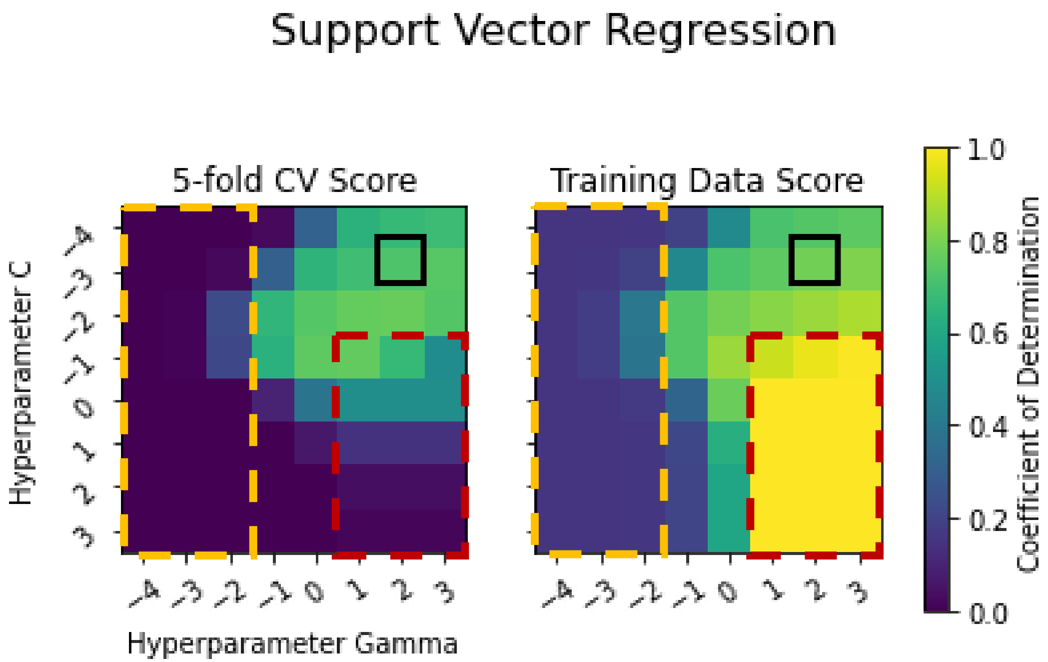
- Mappings can overfit because they *do not* necessarily abstract underlying principles that govern the chemistry or physics
- An 'overfit' model has mapped each training point directly to its response, memorizing the noise and local patterns of the data



Model Complexity & Fit

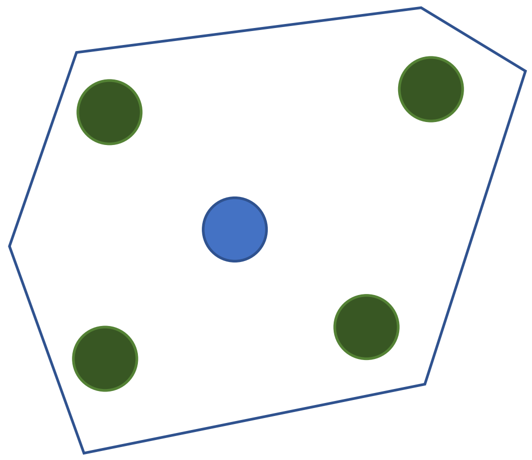
- Fitting is a function of model complexity – the more information a model can contain, the more capacity it has to memorize
- With more limited capacity, it learns the data more efficiently
- Efficiency means finding *useful, high-level abstractions within the data*





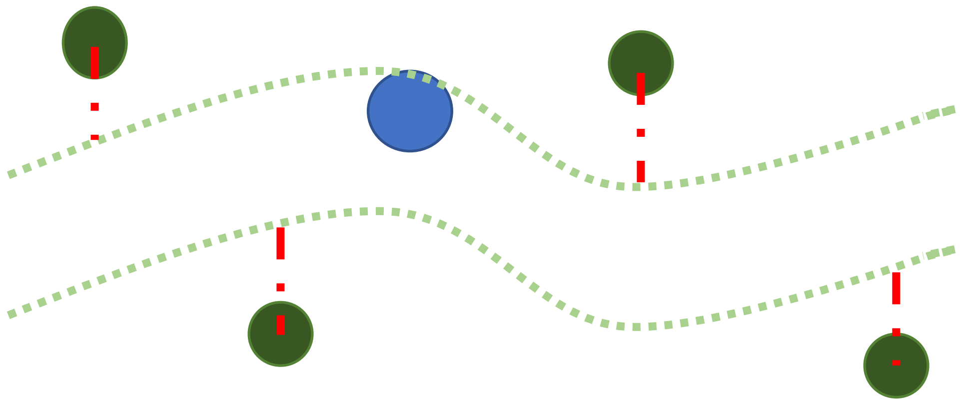
Common Types of Regressor

- "Neighborhood" models
  - K-Nearest Neighbors
  - Decision Trees
  - Random Forests



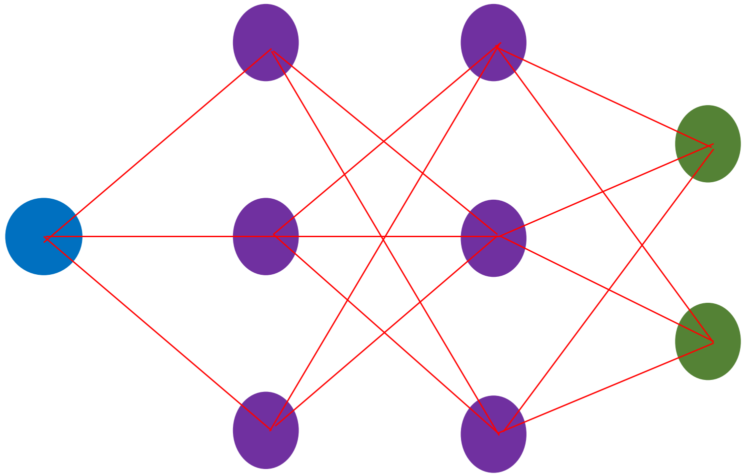
Common Types of Regressors

- "Neighborhood" models
  - K-Nearest Neighbors
  - Decision Trees
  - Random Forests
- "Geometric" models
  - Kernel machines
  - Parametric regression



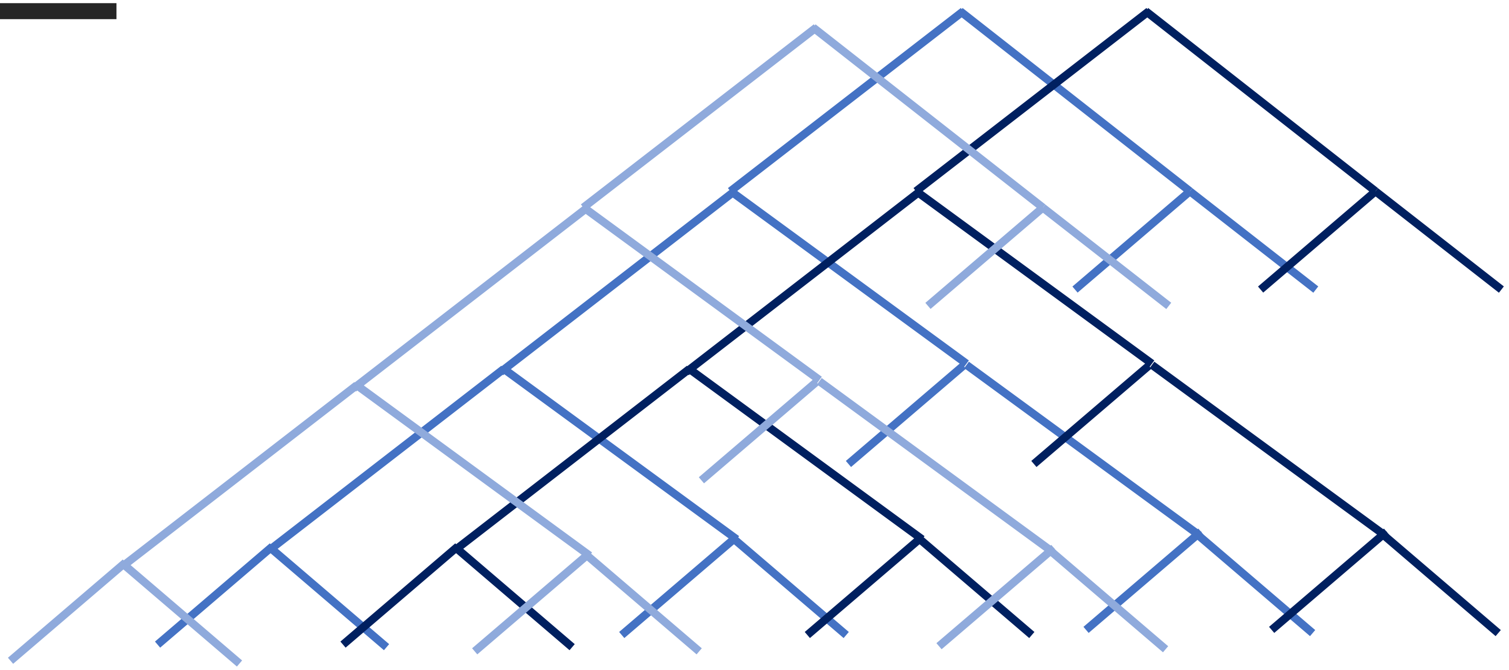
Common Types of Regressors

- "Neighborhood" models
  - K-Nearest Neighbors
  - Decision Trees
  - Random Forests
- "Geometric" models
  - Kernel machines
  - Parametric regression
- "Representation" models
  - Deep neural networks



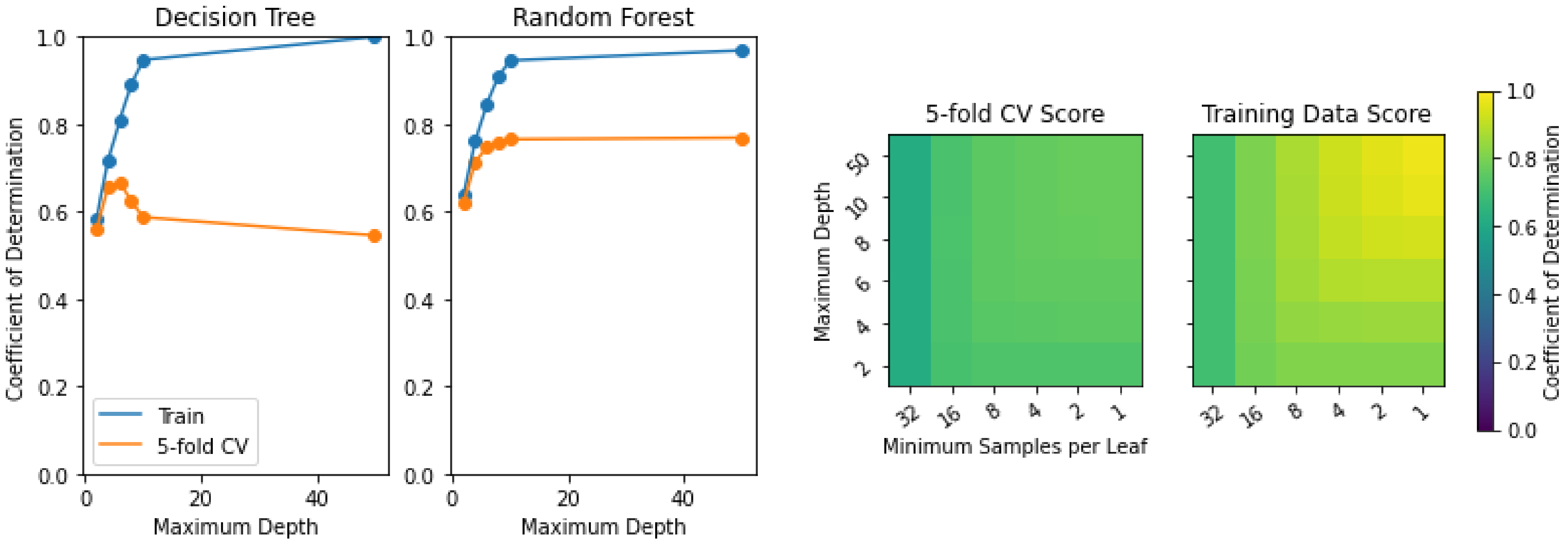
Common Types of Regressors

- "Neighborhood" models
  - K-Nearest Neighbors
  - Decision Trees
  - **Random Forests**
- "Geometric" models
  - Kernel machines
  - Parametric regression
- "Representation" models
  - Deep neural networks



Random Forests

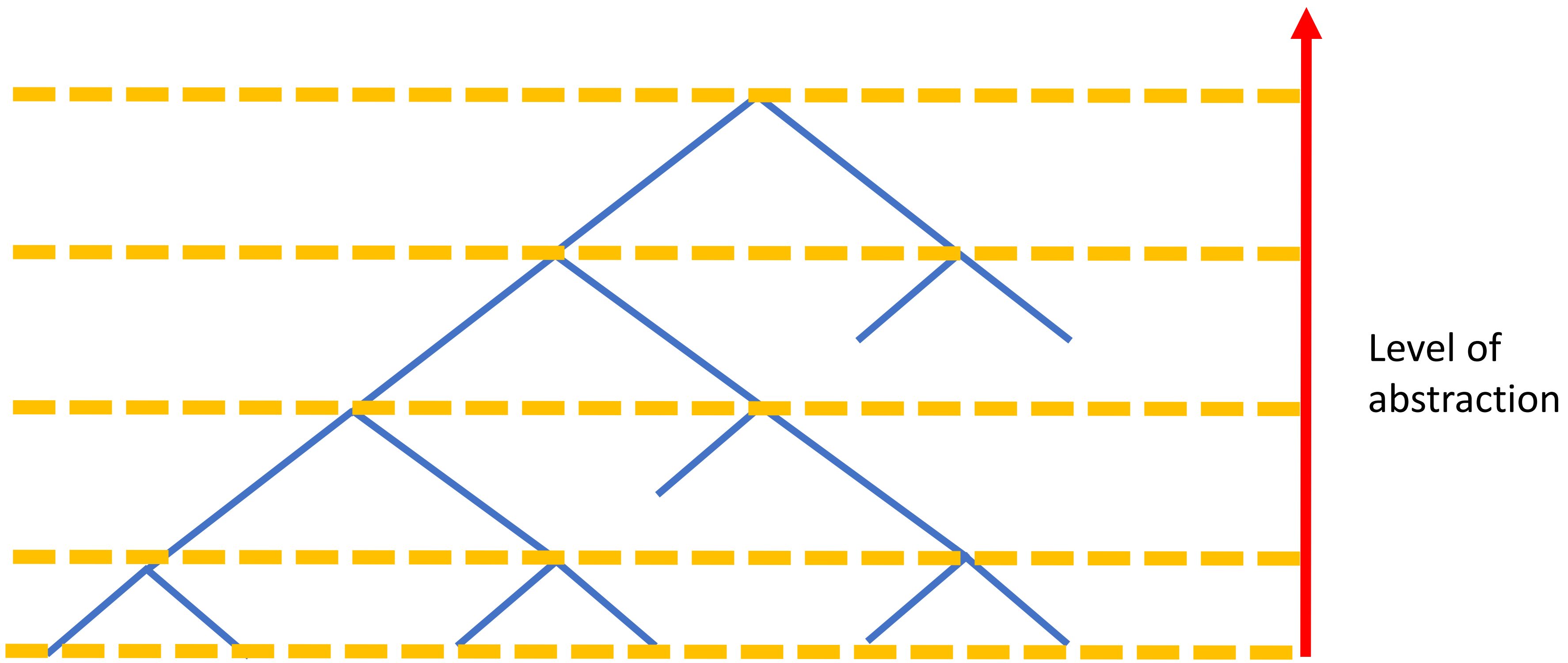
Random forests are an ensemble of a neighborhood model



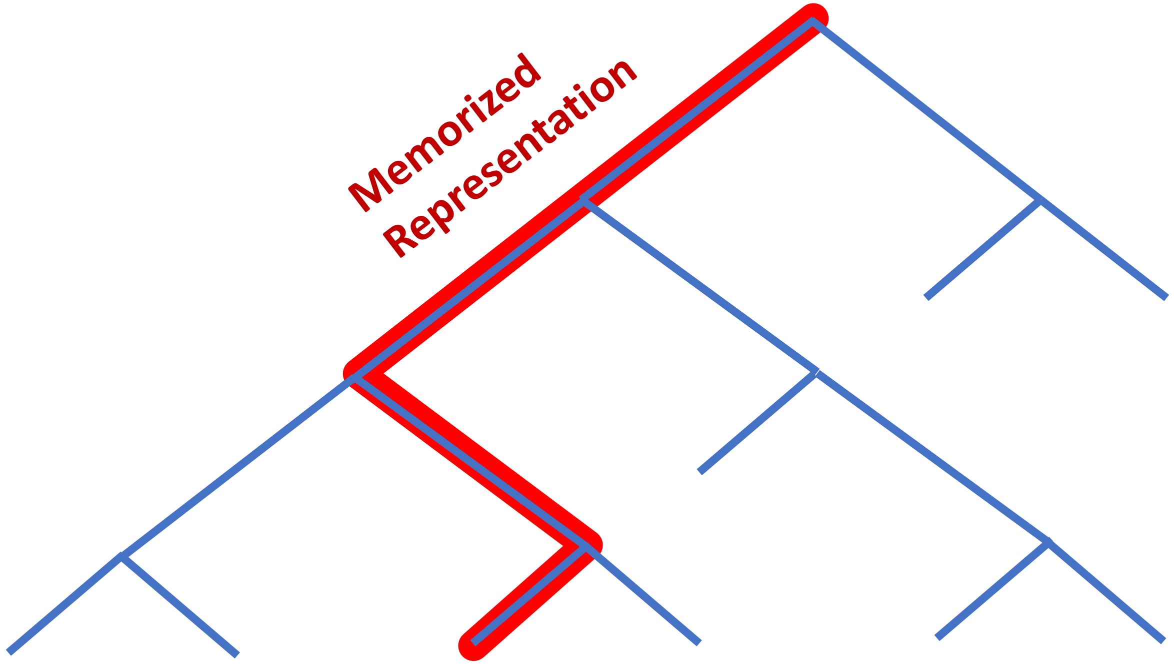
Immunity?

- In one sense
  - Breiman random forests are like k-Nearest Neighbor model in that they explicitly store a representation of the data they are trained on
    - Breiman forests grow trees without pruning, which often results in a data point getting its own leaf
    - This is an *explicit* representation of the data

Breiman Tree



Breiman Tree

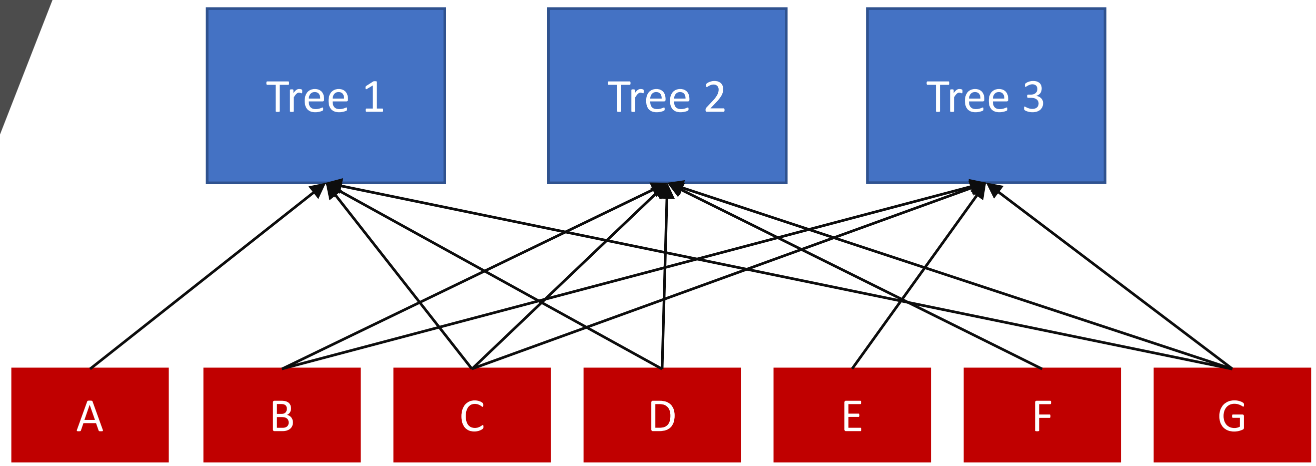


Immunity?

- Breiman forests bootstrap with replacement for each tree so that a given tree does not see the entire training set (1/e ≈ 63%)
- Do they “overfit”? Not really, because it memorizes its exposed training set by construction
- The “partially blind” ensemble effect of the bootstraps causes all these memorizations to wash out, so the memorization is “blurred”

“Partially blind” ensemble

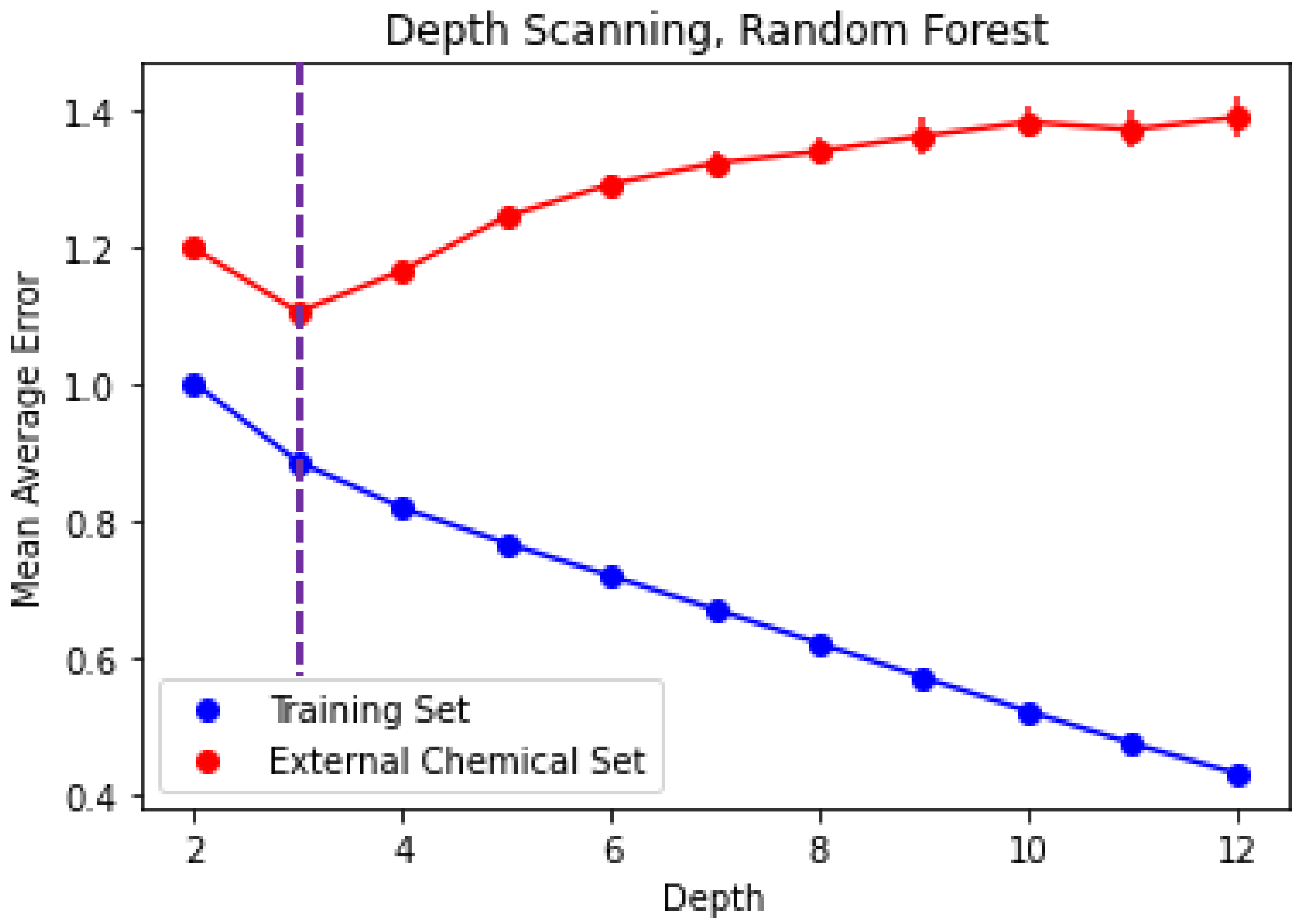
- Bootstrapping partially blinds the model



- The partially blind trees “wash out” their predictions, resulting in a more generalized model
- But the model contains a memorized form of the data *so the proportional representation of the training set matters a lot!*

Limitations

- There is a limit to the overfitting resistance of the random forest the is relevant to “global” modeling
- The high-level abstractions of the shallow trees perform better than the local chemistries of the training domain
- Careful selection of chemical representation can fix this, but short of that it may be savvy to use a more conservative model for highly general chemistries



Conclusions

- Demands for transparency, generality and clarity limit regulatory ability to rely on statistical summaries in model validation
- Idiosyncrasies of public data sets increase concern around overfitting or over-localization
- Due to EPA interest in exotic chemistries (carbon-fluoro bonds, metallics, etc.) we are integrating analysis to combat over-localization to produce more robust theoretical underpinnings for policy decisions