Supervised Learning via Ensemble Tensor Completion

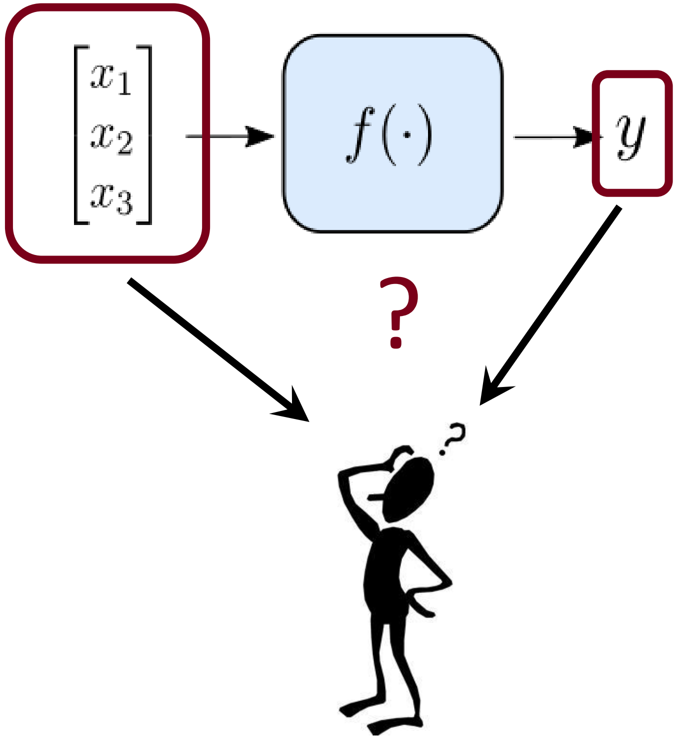
N. Kargas
University of Minnesota

N.D. Sidiropoulos
University of Virginia

2020 Asilomar Conference on Signals, Systems, and Computers.

The Supervised Learning Problem

- General nonlinear function identification
  - `Supervised' - from input-output data
  - Function approximation problem
  - Identifiability, performance, complexity...
- Previous work:
  - Canonical System Identification (CSID)
  - Based on tensor principal components
  - Advantages: `universal', intuitive, interpretable, backed by theory
  - Suitable only for discrete input...



Categorical (classification)
Real-valued (prediction, regression)

Roadmap

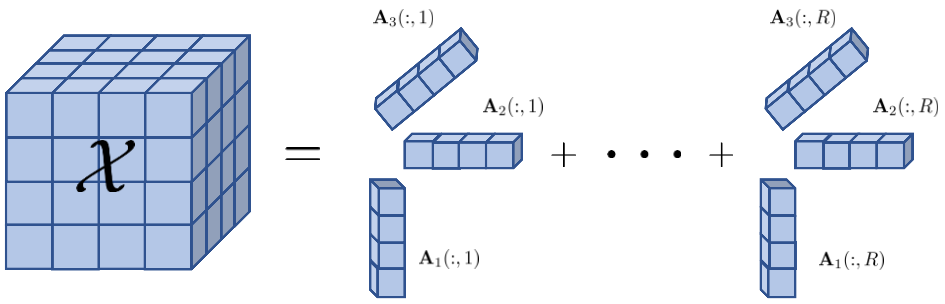
- Tensor Decomposition
- Canonical System Identification (CSID)
- Proposed Method: Ensemble Tensor Completion
- Experiments
- Conclusion

Canonical Polyadic Decomposition (CPD)

- An N-way tensor (multi-way array) admits a decomposition of rank *R* it can be decomposed as a sum of *R* rank-1 tensors

$$\mathcal{X} = \sum_{r=1}^{R} \mathbf{A}_1(:,r) \circ \mathbf{A}_2(:,r) \circ \cdots \circ \mathbf{A}_N(:,r)$$

- Tensor rank is smallest *R* for which such decomposition exists → *Canonical*



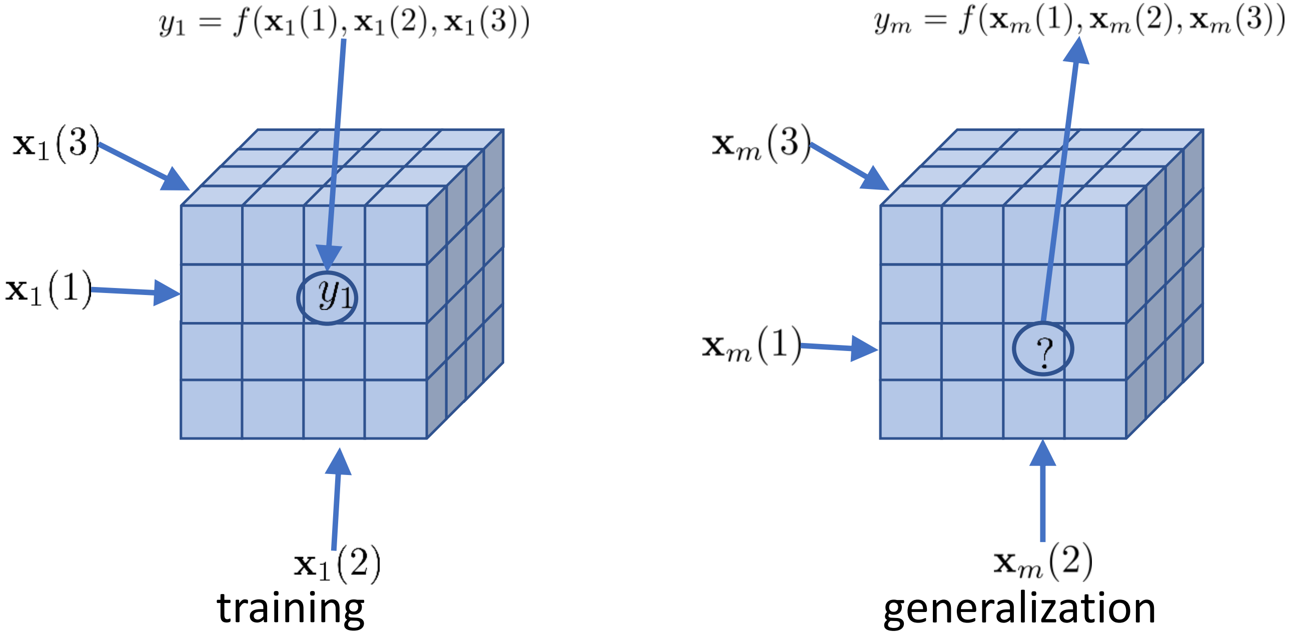
- Element-wise: $\mathcal{X}(i_1,\ldots,i_N) = \sum_{r=1}^{R} \prod_{n=1}^{N} \mathbf{A}_n(i_n,r)$
- Matrix unfolding: $\mathcal{X}^{(n)} = (\mathbf{A}_N \odot \cdots \odot \mathbf{A}_{n+1} \odot \mathbf{A}_{n-1} \cdots \odot \cdots \mathbf{A}_1)\mathbf{A}_n^T$
- Vector: $\mathrm{vec}(\mathcal{X}) = (\mathbf{A}_N \odot \cdots \odot \mathbf{A}_1)\mathbf{1}$
- <u>Property:</u> Unique under mild conditions!

Canonical System Identification (CSID)

- Single high-order tensor for learning a general nonlinear system

- Each input vector $[\mathbf{x}_m(1), \mathbf{x}_m(2), \mathbf{x}_m(3)]^T$ is viewed as a cell multi-index and the cell content is the estimated response of the system

$$\mathbf{x}_1(1) = 2$$
$$\mathbf{x}_1(2) = 3$$
$$\mathbf{x}_1(3) = 1$$
$$y_1 = f(2, 3, 1)$$

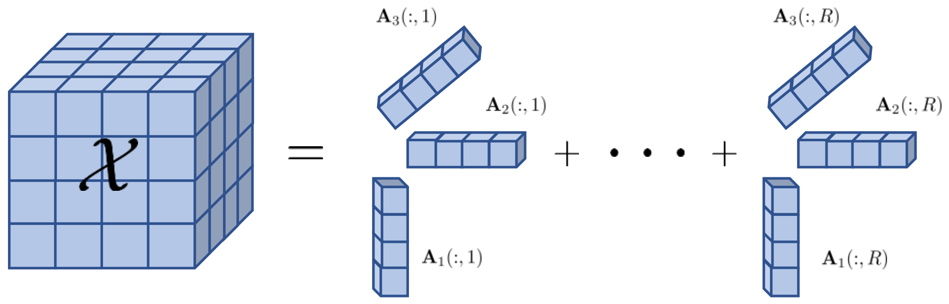


Kargas, N., and Sidiropoulos, N. D. “Nonlinear System Identification via Tensor Completion”, AAAI 2020, NYC, NY.

Canonical System Identification (CSID)

- Assuming a low-rank CPD model, the problem of finding the rank-R approximation which best fits the data is formulated as:

$$\min_{\{\mathbf{A}_n\}_{n=1}^N} \frac{1}{M}\sum_{m=1}^{M}\left(y_m - f\left(\mathbf{x}_m; \{\mathbf{A}_n\}_{n=1}^N\right)\right)^2 + \sum_{n=1}^{N}\rho\|\mathbf{A}_n\|_F^2 + \sum_{n=1}^{N}\mu\|\mathbf{T}_n\mathbf{A}_n\|_F^2,$$



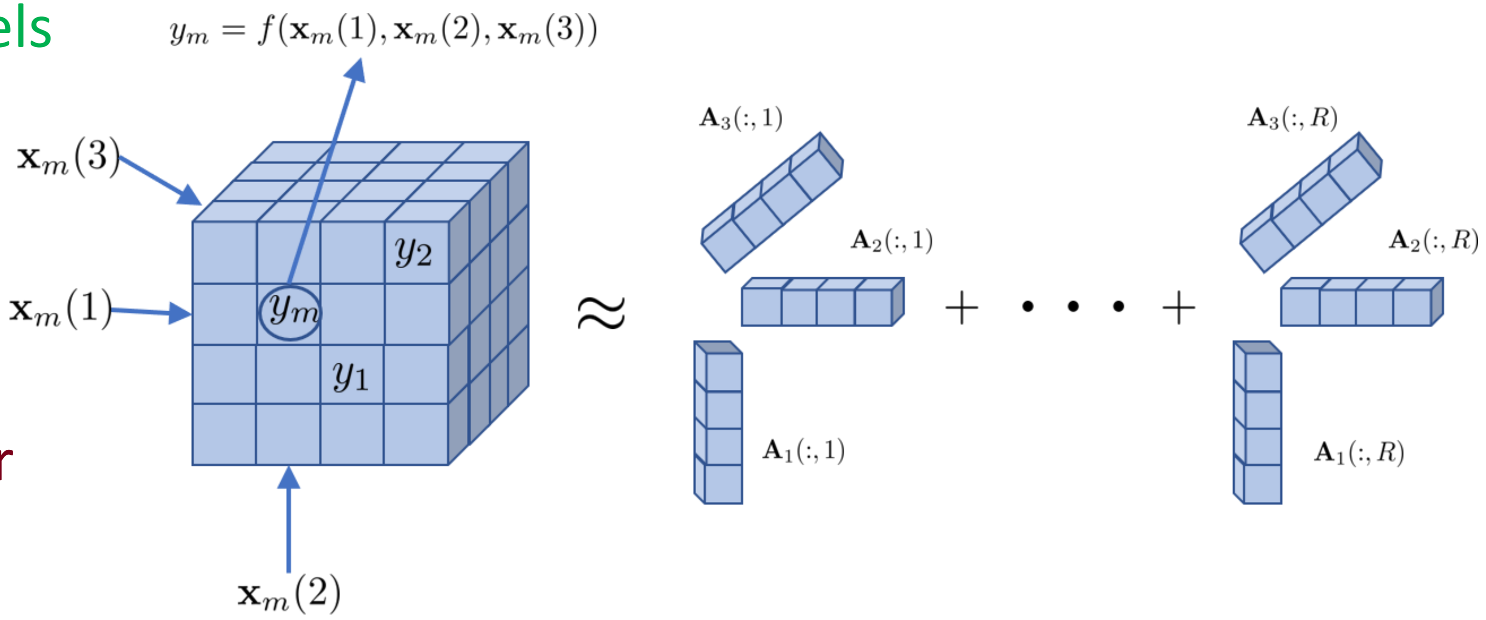
where $f\left(\mathbf{x}_m; \{\mathbf{A}_n\}_{n=1}^N\right) = \sum_{r=1}^{R}\prod_{n=1}^{N}\mathbf{A}_n(\mathbf{x}_m(n), r)$.

Kargas, N., and Sidiropoulos, N. D. “Nonlinear System Identification via Tensor Completion”, AAAI 2020, NYC, NY.

Canonical System Identification (CSID)

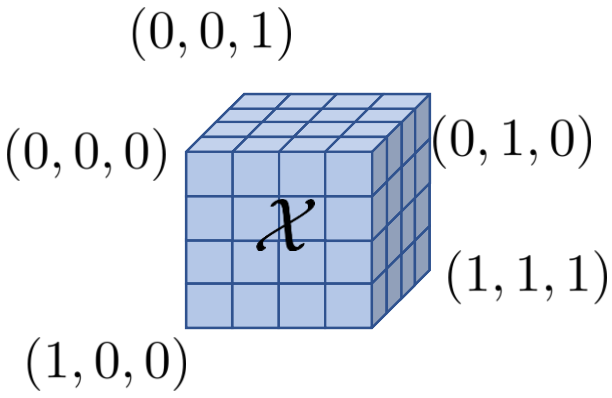
- Advantages: `universal', models any nonlinearity.
- If low rank, we can learn the true mapping!

- Drawback: naturally suited for discrete input data



Kargas, N., and Sidiropoulos, N. D. “Nonlinear System Identification via Tensor Completion”, AAAI 2020, NYC, NY.

Ensemble Learning



- Coarse discretization
  + Complexity
  - Resolution
  - Poor performance

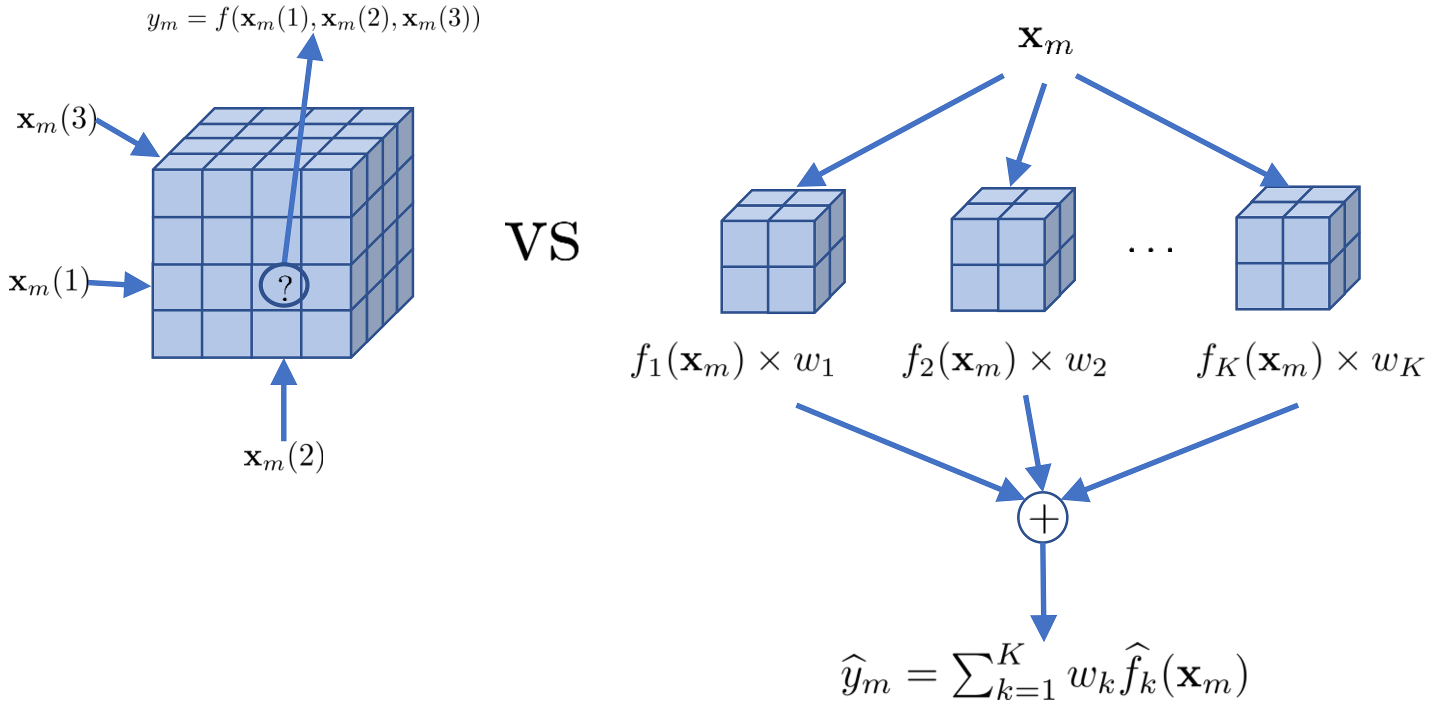
- Fine discretization
  + Resolution
  - Complexity
  - No generalization

- Ensemble learning
  - Combination of multiple ``weak'' models outperforms a single model
  - Examples: Bagging, boosting

Ensemble Tensor Completion

- Bagging
  - Create different training sets by sampling
  - Parallel training
  - Average the results

- Boosting
  - Sequential training
  - Models are fit on the prediction errors



Bagging

- Step 1: Create different training datasets by sampling with replacement
- Step 2: Select a discretization method for each dataset
  - Intervals have identical widths
  - Intervals have same number of points
  - K-means
- Step 3: Solve K independent problems using Stochastic Gradient Descent (SGD):

$$\min_{\{\mathbf{A}_n\}_{n=1}^N} \frac{1}{M}\sum_{m=1}^{M}\left(y_m - f\left(\mathbf{x}_m; \{\mathbf{A}_n\}_{n=1}^N\right)\right)^2 + \sum_{n=1}^{N}\rho\|\mathbf{A}_n\|_F^2 + \sum_{n=1}^{N}\mu\|\mathbf{T}_n\mathbf{A}_n\|_F^2,$$

- Combine the results: $f_{\mathrm{CSID-Bag}}(\mathbf{x}) = \sum_{k=1}^{K} w_k f_k\left(\mathbf{x}; \{\mathbf{A}_n^k\}_{n=1}^N\right), \quad w_k = \frac{1/\mathrm{Err_k}}{\sum_{k=1}^K 1/\mathrm{Err_k}}$

Boosting (Forward State-wise Additive Modeling)

- Models are fit on the prediction errors
- Step 1: At iteration k, choose between 3 discretization methods
- Step 2: Solve:

$$\min_{\{\mathbf{A}_n^k\}_{n=1}^N} \frac{1}{M}\sum_{m=1}^{M}\left(y_m - \widehat{y}_m^{k-1} - f_k\left(\mathbf{x}_m; \{\mathbf{A}_n^k\}_{n=1}^N\right)\right)^2 + \sum_{n=1}^{N}\rho\|\mathbf{A}_n^k\|_F^2 + \sum_{n=1}^{N}\mu\|\mathbf{T}_n\mathbf{A}_n^k\|_F^2,$$

where $\widehat{y}_m^{k-1} = \sum_{k'=1}^{k-1} f_k\left(\mathbf{x}_m; \{\mathbf{A}_n^{k'}\}_{n=1}^N\right).$

- At each iteration, a new model is added to the expansion
- Predict the output of new data points as

$$f_{\mathrm{CSID-Boost}}(\mathbf{x}) = \sum_{k=1}^{K} f_k\left(\mathbf{x}; \{\mathbf{A}_n^k\}_{n=1}^N\right).$$

Experiments

- Compare the ensemble models against a single CSID model
- Regression task using 4 UCI repository datasets
- We combine K=10 CSID models to build the ensemble models
- We fix the alphabet size to be I=20 and discretize all continuous inputs

- 85% training, 15% test (5-fold cross-validation for parameter selection)
- Evaluate the performance using RMSE
- All the methods are trained using Adam with a learning rate 1e-2 for a maximum of 50 epochs

Results

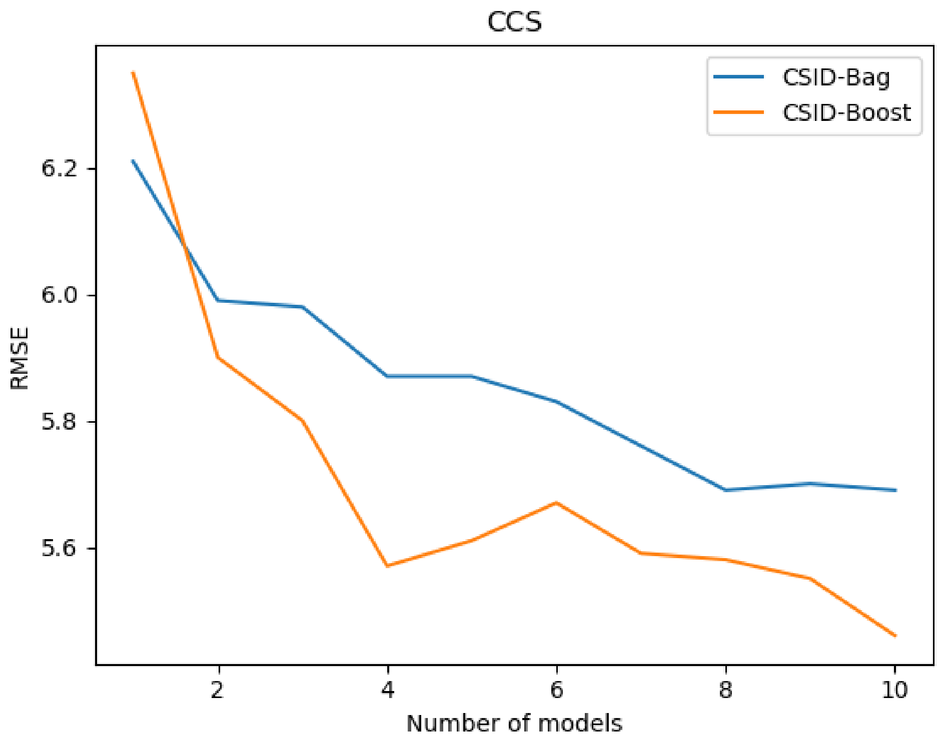
Table 1: Dataset Information.

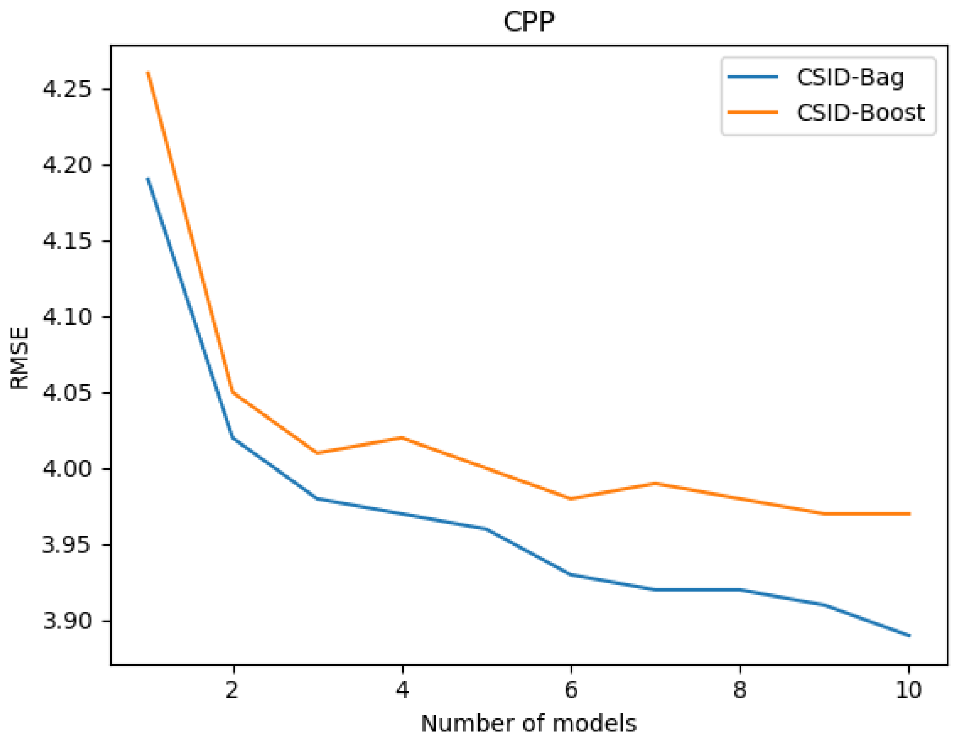
Dataset	N	M
QSAR AQUATIC TOXICITY (QSAR)	8	546
CONCRETE COMPRESSIVE STRENGTH (CCS)	9	1030
CYCLE POWER PLANT (CPP)	4	9568
PHYSICOCHEMICAL PROPERTIES (PP)	9	45730

Table 2: Comparison of RMSE performance of different models on multi-output regression.

Dataset	CSID	CSID-Bag (10)	CSID-Boost (10)
QSAR	1.51	1.37	1.49
CCS	6.25	5.69	5.46
CPP	4.22	3.89	3.97
PP	4.29	3.95	3.98

RMSE Performance vs Number of Models





Take-home points

- Concluding remarks:
  - Tensor based method for supervised learning
  - Ensemble learning can enhance the prediction accuracy of the CSID model
  - Counter the performance degradation resulting from the discretization step

- Coming up:
  - So far, non-parametric; what if we know something about $f(\boldsymbol{x})$? – in review
  - Other tensor models

THANK YOU!