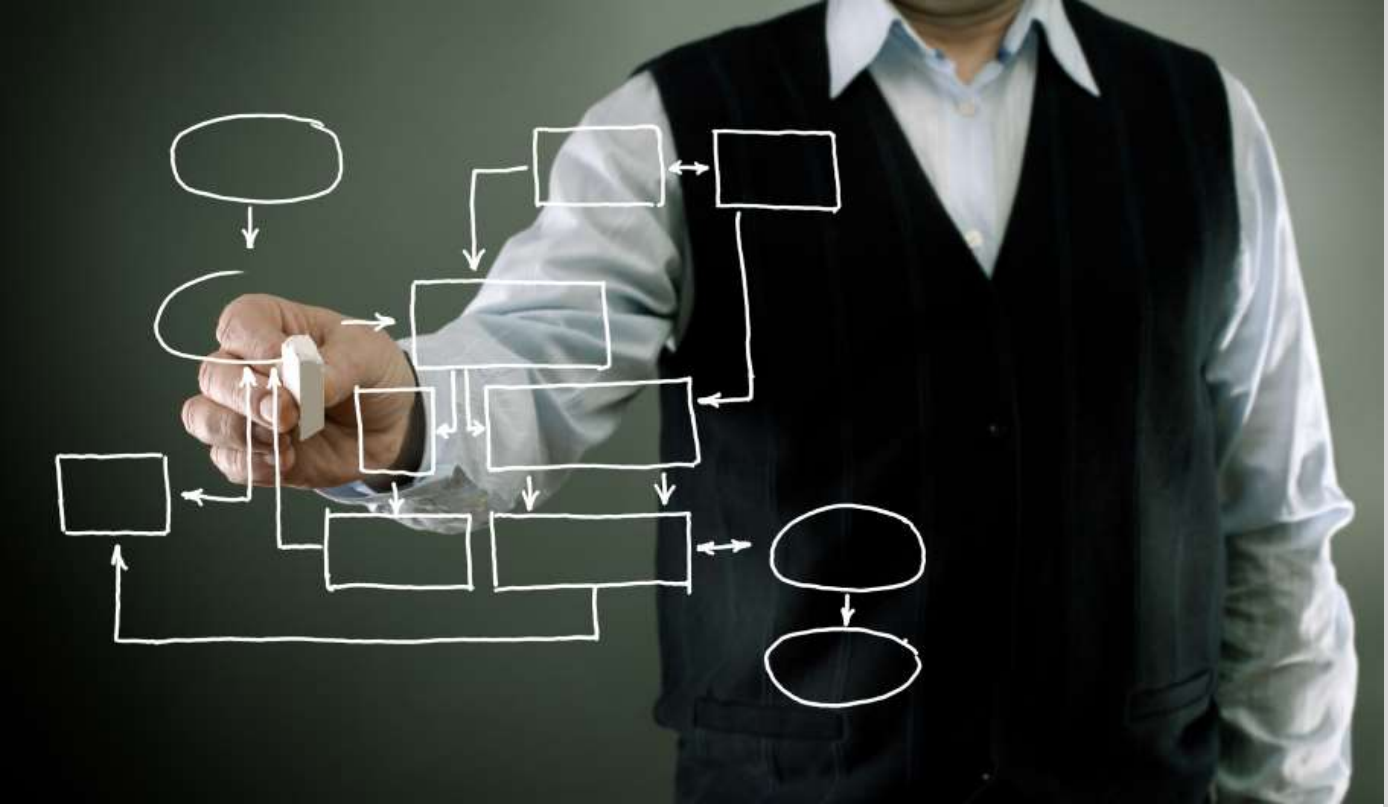


Data Aggregation for Reducing Training Data in Symbolic Regression

Lukas Kammerer, Gabriel Kronberger and Michael Kommenda

Josef Ressel Centre for Symbolic Regression (JRZ)
University of Applied Sciences Upper Austria, Hagenberg (HEAL)

HEAL
HEURISTIC AND EVOLUTIONARY ALGORITHMS LABORATORY

SymReg
JOSEF RESSEL CENTER FOR SYMBOLIC REGRESSION

OBERÖSTERREICH

HeuristicLab

EuroCAST 2019

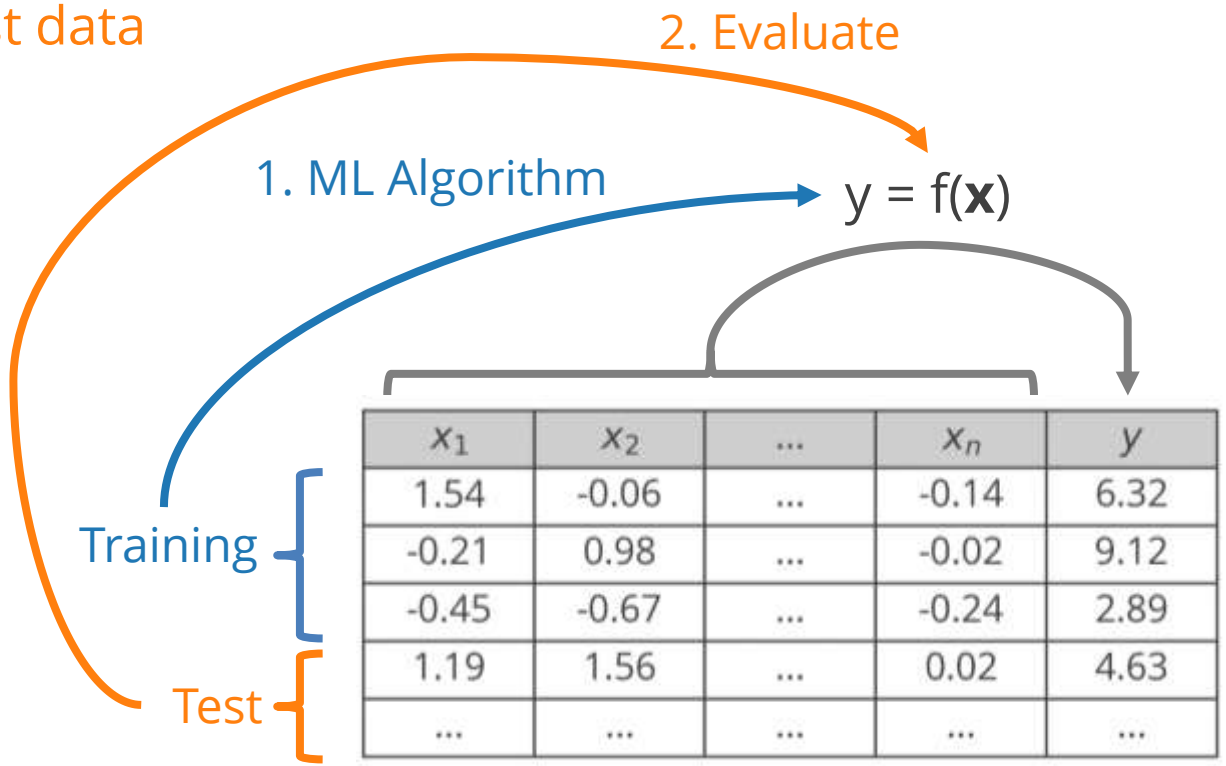
Contact:
Lukas Kammerer
Heuristic and Evolutionary Algorithms Lab (HEAL)
Softwarepark 11
4232 Hagenberg, Austria

E-mail:
lukas.kammerer@fh-hagenberg.at

Web:
http://heal.heuristiclab.com

Supervised Machine Learning

1. Find model with minimal error on training data
2. Check that model works with unseen test data

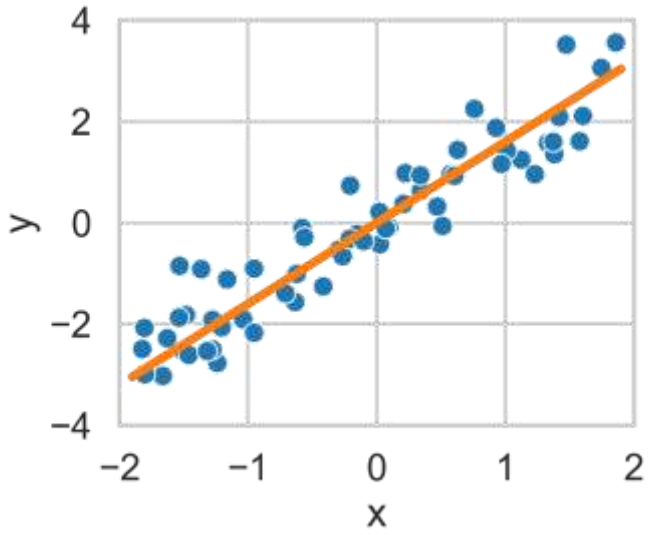


x_1	x_2	...	x_n	y
1.54	-0.06	...	-0.14	6.32
-0.21	0.98	...	-0.02	9.12
-0.45	-0.67	...	-0.24	2.89
1.19	1.56	...	0.02	4.63
...	...	...	...	...

Regression Algorithms

Linear Regression

- Find linear relations in data



Symbolic Regression

- Find mathematical formulas of any structure
- Search with genetic programming with strict offspring selection[1]
- Optimize constants with gradient descent

Random Forest Regression

- Build many decision trees with differently sampled training data

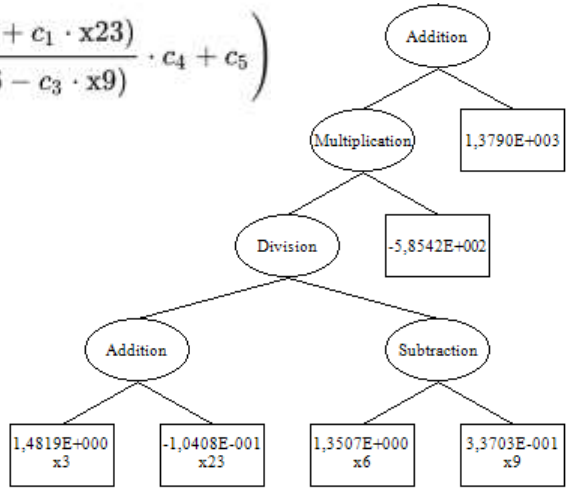
1) Affenzeller, M., Wagner, S., Winkler, S., & Beham, A. (2009). *Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications*. Chapman and Hall/CRC.

Regression Algorithms

- Linear Regression
  - Find linear relations in data

- **Symbolic Regression**
  - Find mathematical formulas of any structure
  - Search with genetic programming with strict offspring selection[1]
  - Optimize constants with gradient descent

$$e = \left(\frac{(c_0 \cdot \mathrm{x3} + c_1 \cdot \mathrm{x23})}{(c_2 \cdot \mathrm{x6} - c_3 \cdot \mathrm{x9})} \cdot c_4 + c_5 \right)$$



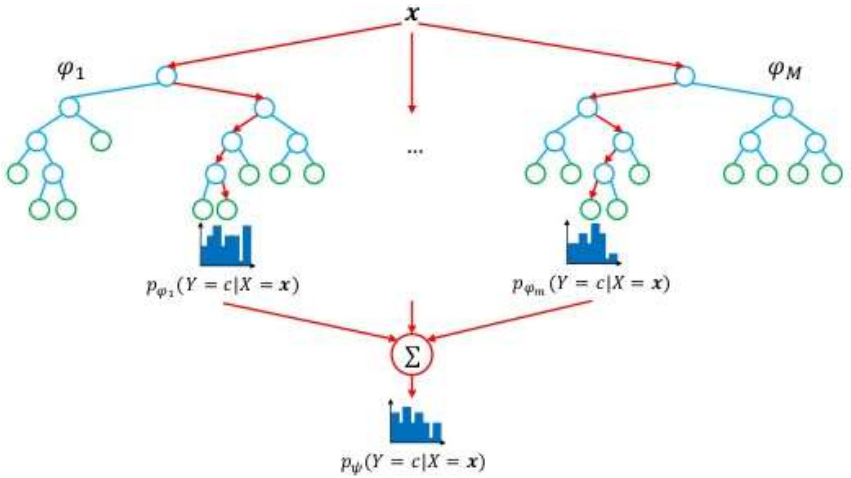
- Random Forest Regression
  - Build many decision trees with differently sampled training data

1) Affenzeller, M., Wagner, S., Winkler, S., & Beham, A. (2009). *Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications*. Chapman and Hall/CRC.

Regression Algorithms

- Linear Regression
  - Find linear relations in data
- Symbolic Regression
  - Find mathematical formulas of any structure
  - Search with genetic programming with strict offspring selection[1]
  - Optimize constants with gradient descent
- **Random Forest Regression**
  - Build many decision trees with differently sampled training data

Image: Ilan Reinstein, KDnuggets™, www.kdnuggets.com/2017/10/random-forests-explained.html

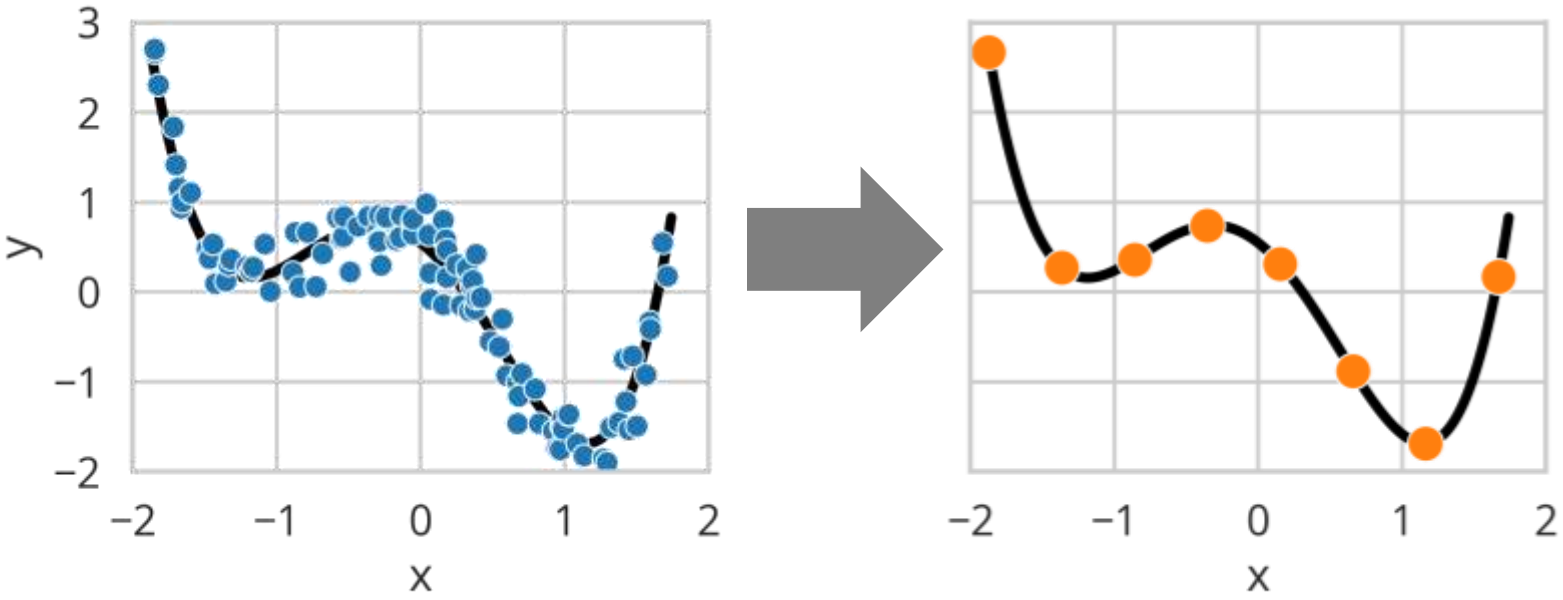


1) Affenzeller, M., Wagner, S., Winkler, S., & Beham, A. (2009). *Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications*. Chapman and Hall/CRC.

Make Big Data small again!

GP with Big Data is impractical → Problem: Evaluation

Can we reduce training data with without loss in accuracy?

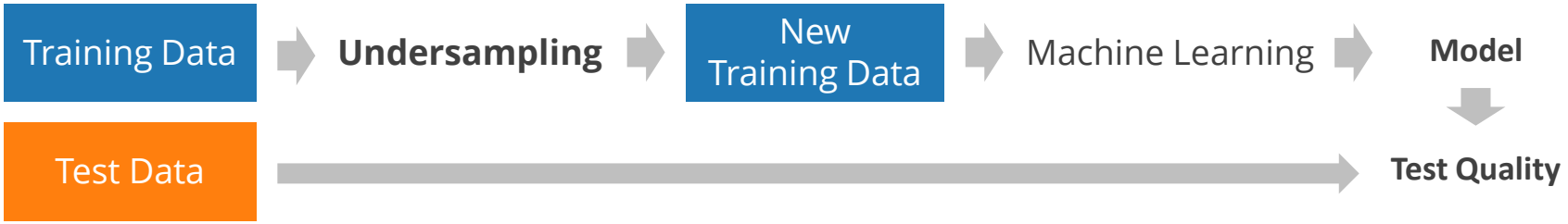


Idea

Kugler et al.[1] suggest data clustering

→ Cancel out noise

→ Remove inconsistencies



Used undersampling methods:

- Random sampling
- K-means clustering
- Data binning

1) Kugler, C., Hochrein, T., Dietl, K., Heidemeyer, P., Bastian, M.: *Softsensoren in der Kunststoffverarbeitung: Qualitätssicherung für die Compoundierung und Extrusion*. SKZ - Forschung und Entwicklung. Shaker Verlag GmbH (2015)

K-Means Clustering

Finds *K* centroids, so that the total (Euclidean) distance of all points to their closest centroid becomes minimal.

- → Stochastic
- → Scaling of variables
- → Mini-Batch K-Means[1]

  Much faster with similar accuracy[1]

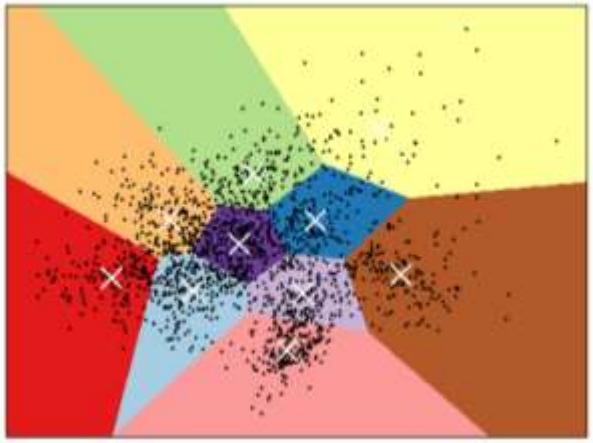


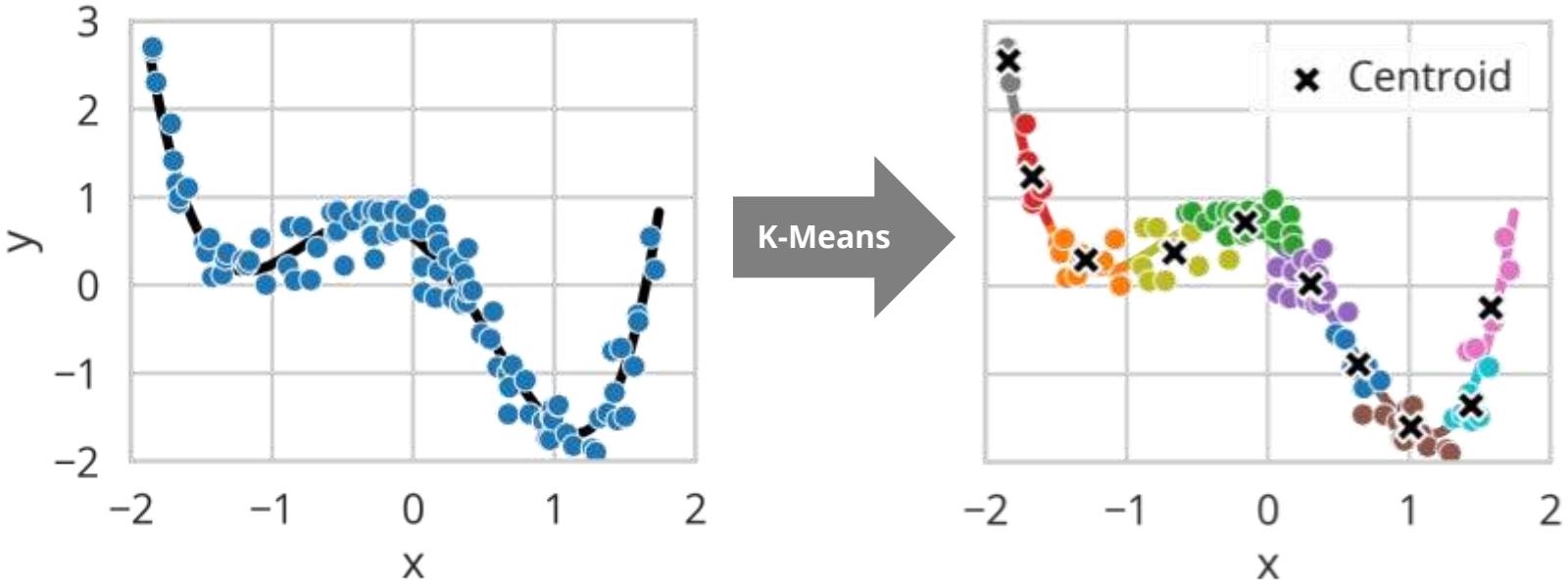
Image: Scikit-learn: *Machine Learning in Python*, Pedregosa et al., JMLR 12 pp. 2825-2830, 2011

1) Sculley, D. (2010, April). *Web-Scale K-Means Clustering*. In Proceedings of the 19th international conference on World wide web (pp. 1177-1178). ACM.

K-Means Clustering

Result of K-Means are *K* centroids

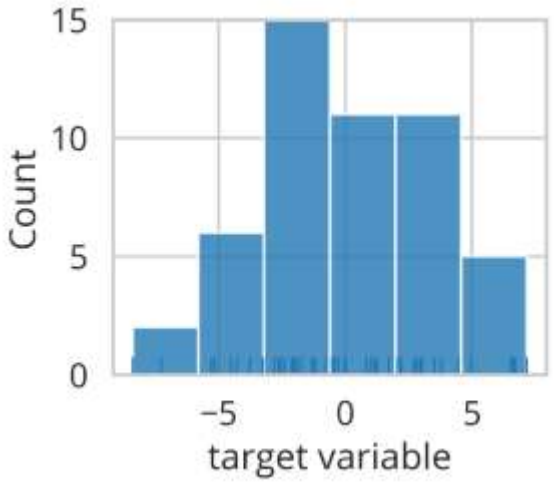
→ Use these centroids as training data!



Data Binning

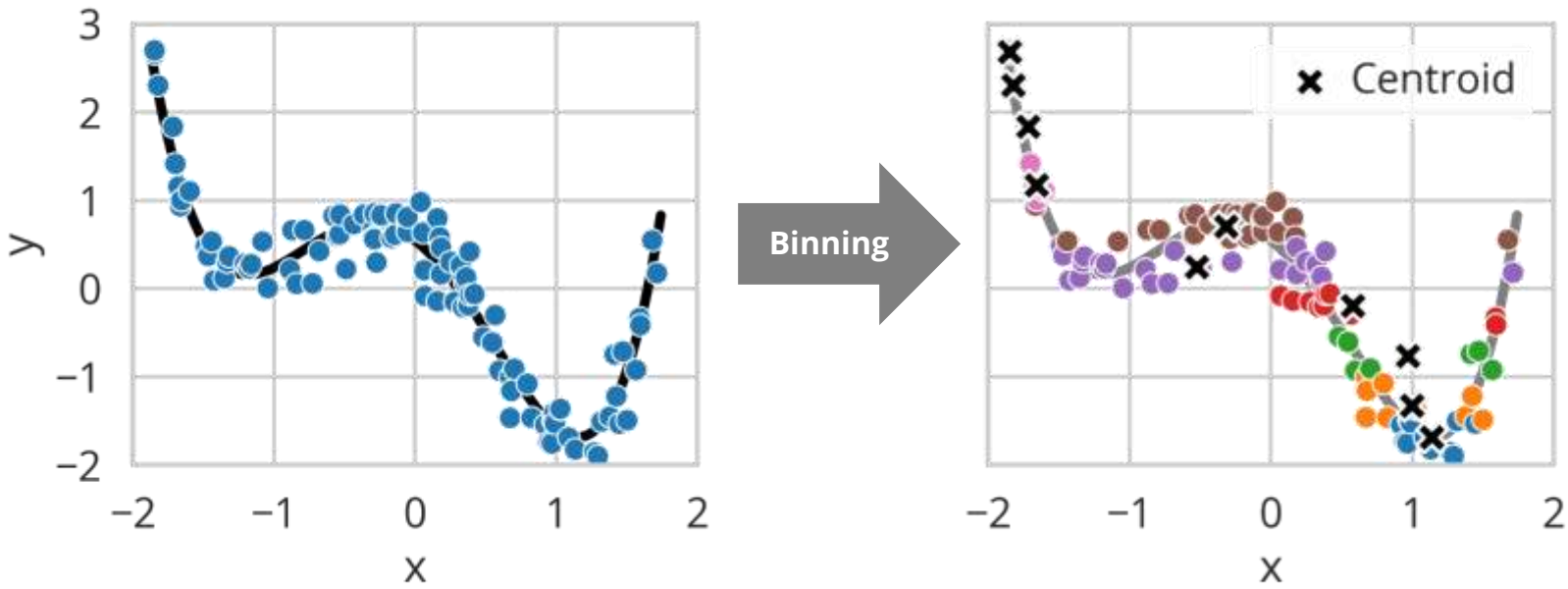
1. Group the data along the target variable into bins of fixed width (similar to a histogram).
2. Use the average of each variable per bin as representative.
3. Use these representatives as training data.

→ Deterministic
→ Imposes n:1 mapping of features to target variable
→ Should cancel out inconsistencies in process states

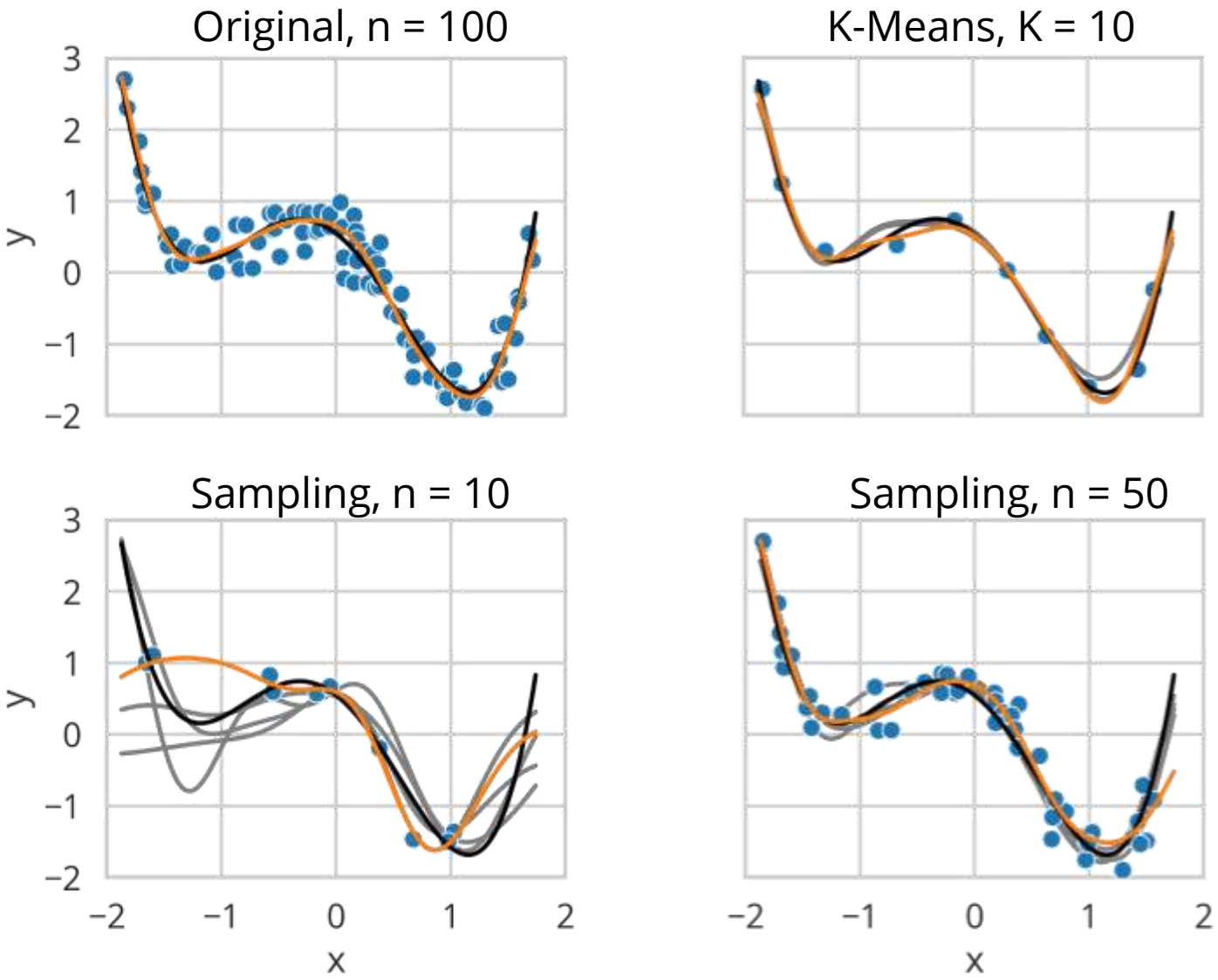


Data Binning

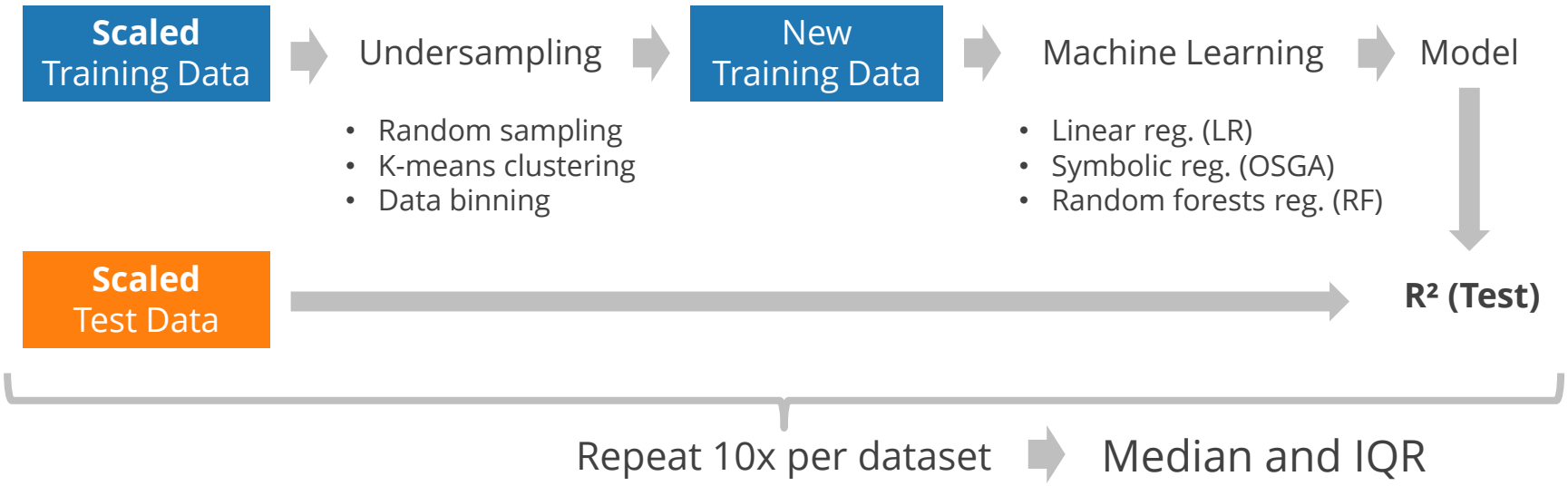
1. Group the data along the target variable into bins of fixed width (similar to a histogram).
2. Use the average of each variable per bin as representative.
3. Use these representatives as training data.



Intuition


Original, n = 100
K-Means, K = 10
Sampling, n = 10
Sampling, n = 50
y
x
3
2
1
0
−1
−2
−2
−1
0
1
2

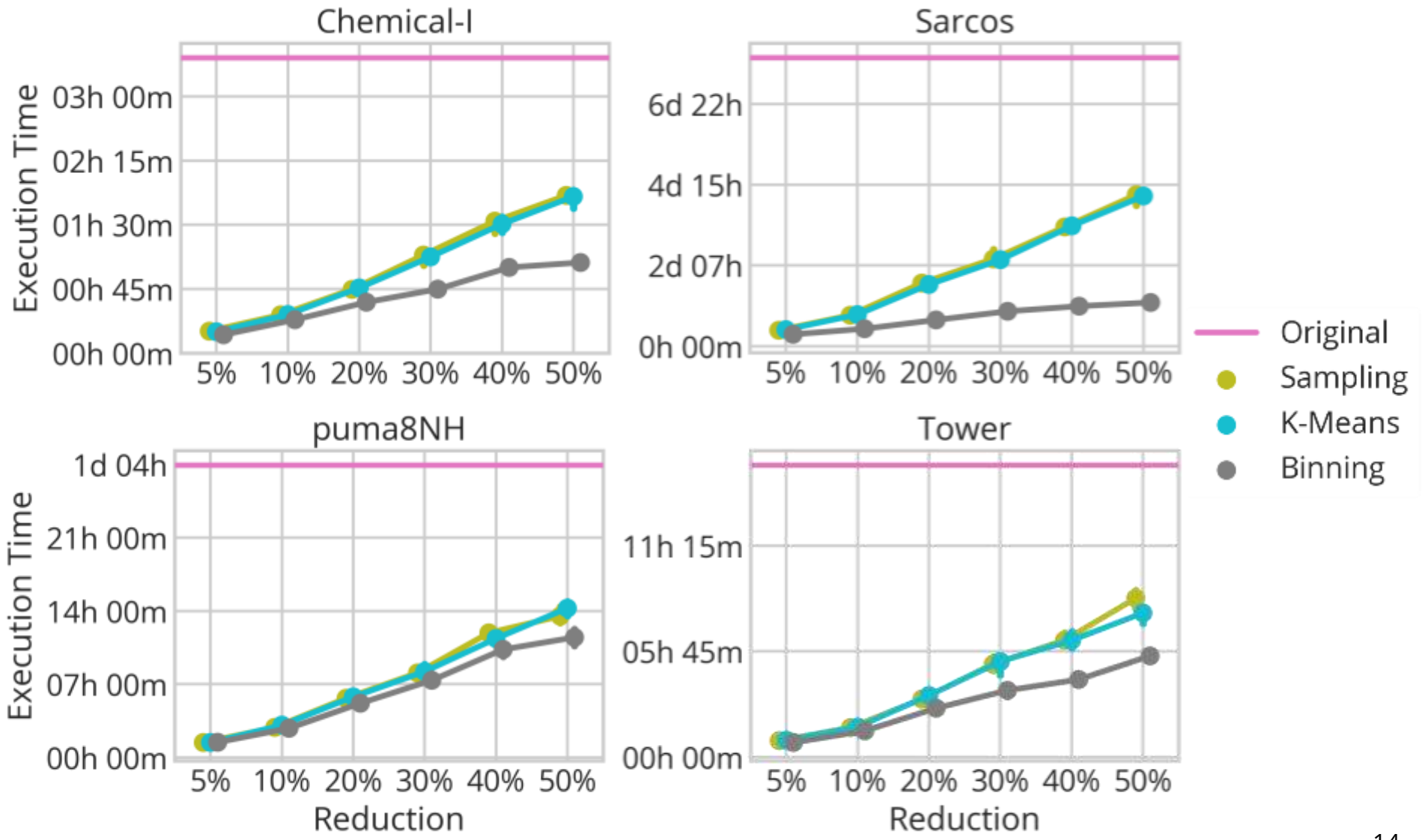
Experiment Setup



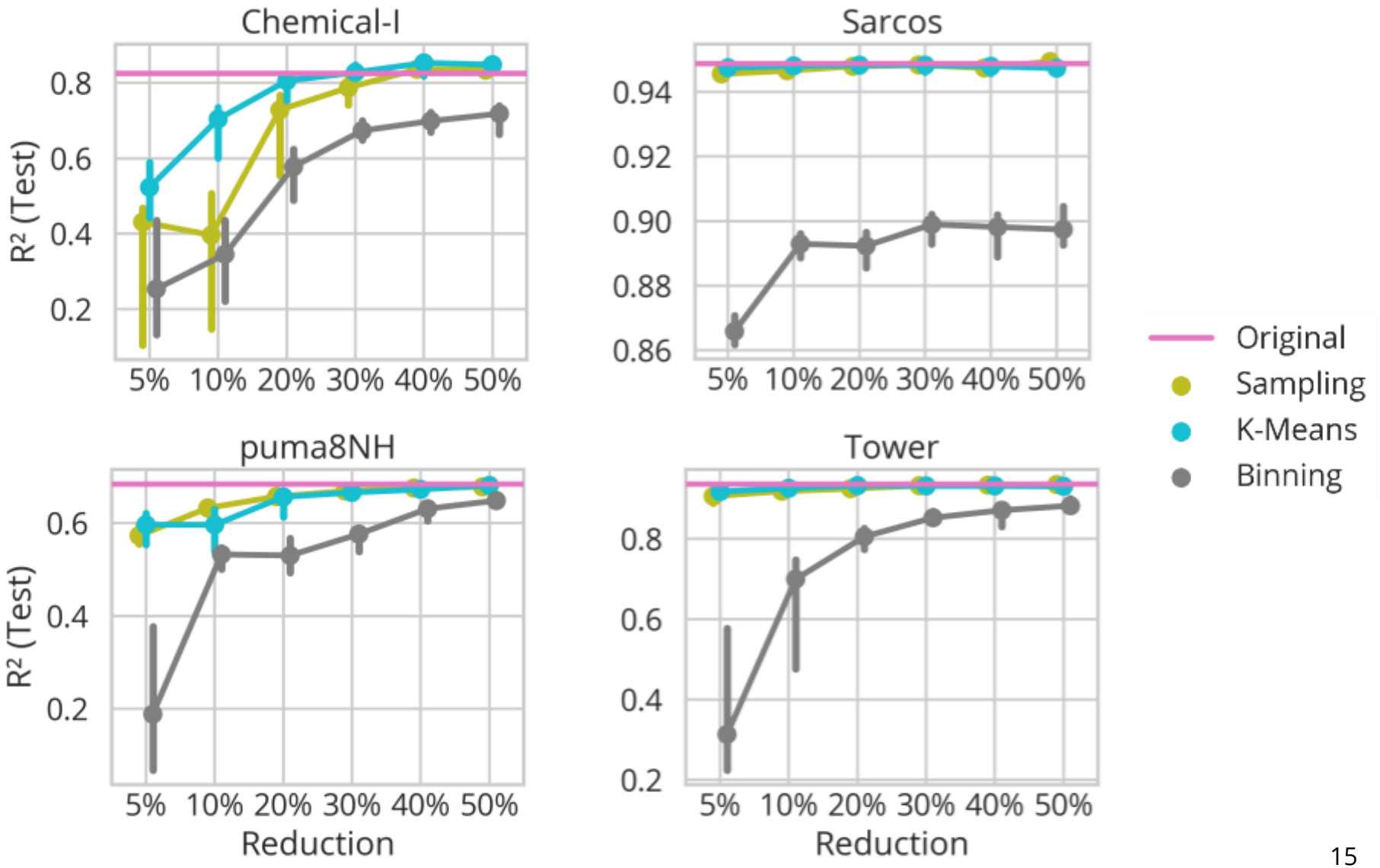
Used datasets:

	training rows	features
Chemical-I	711	57
Sarcos	~44 500	21
puma8NH	6144	8
Tower	3136	25

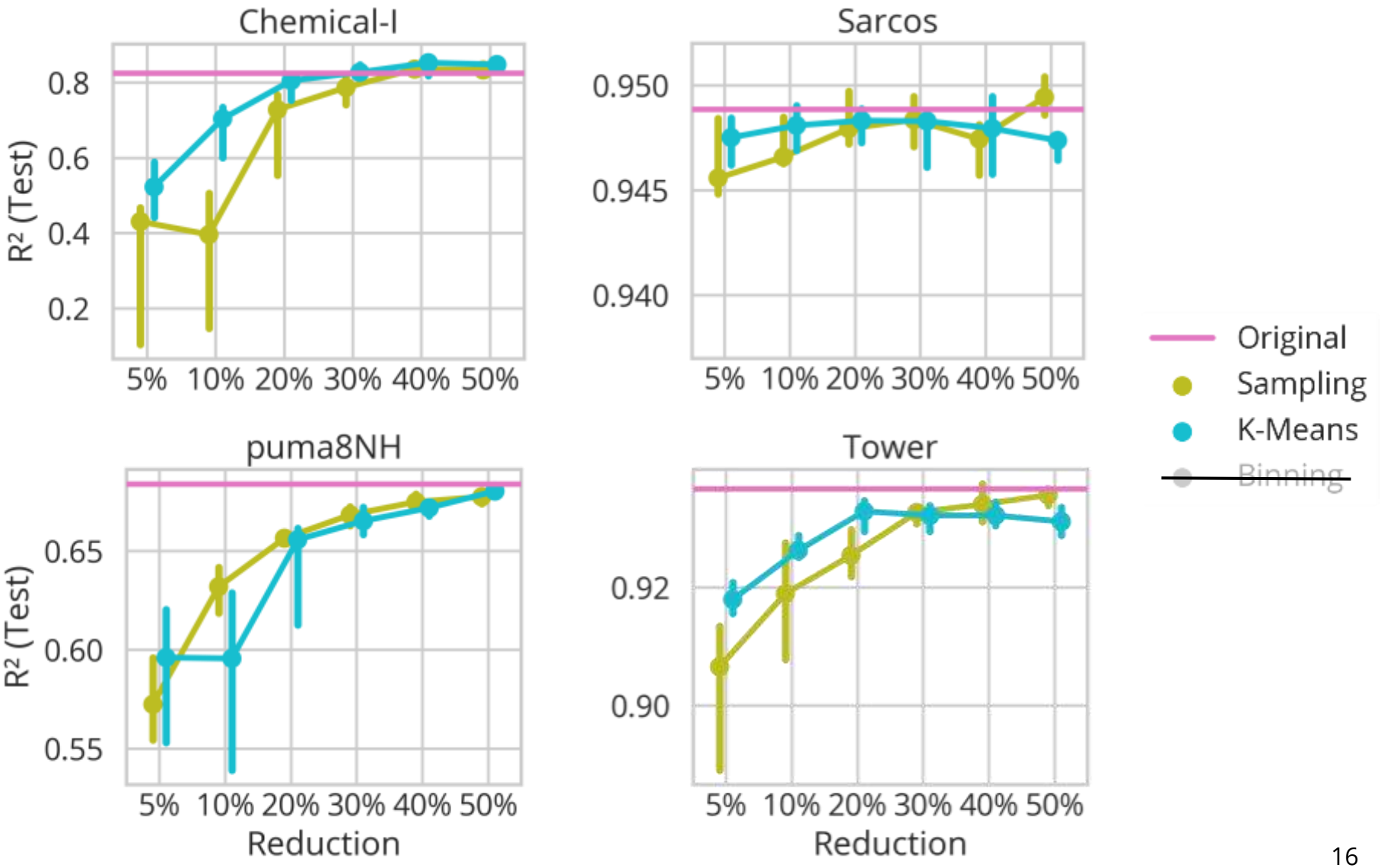
OSGA Speedup



Loss in Quality with OSGA

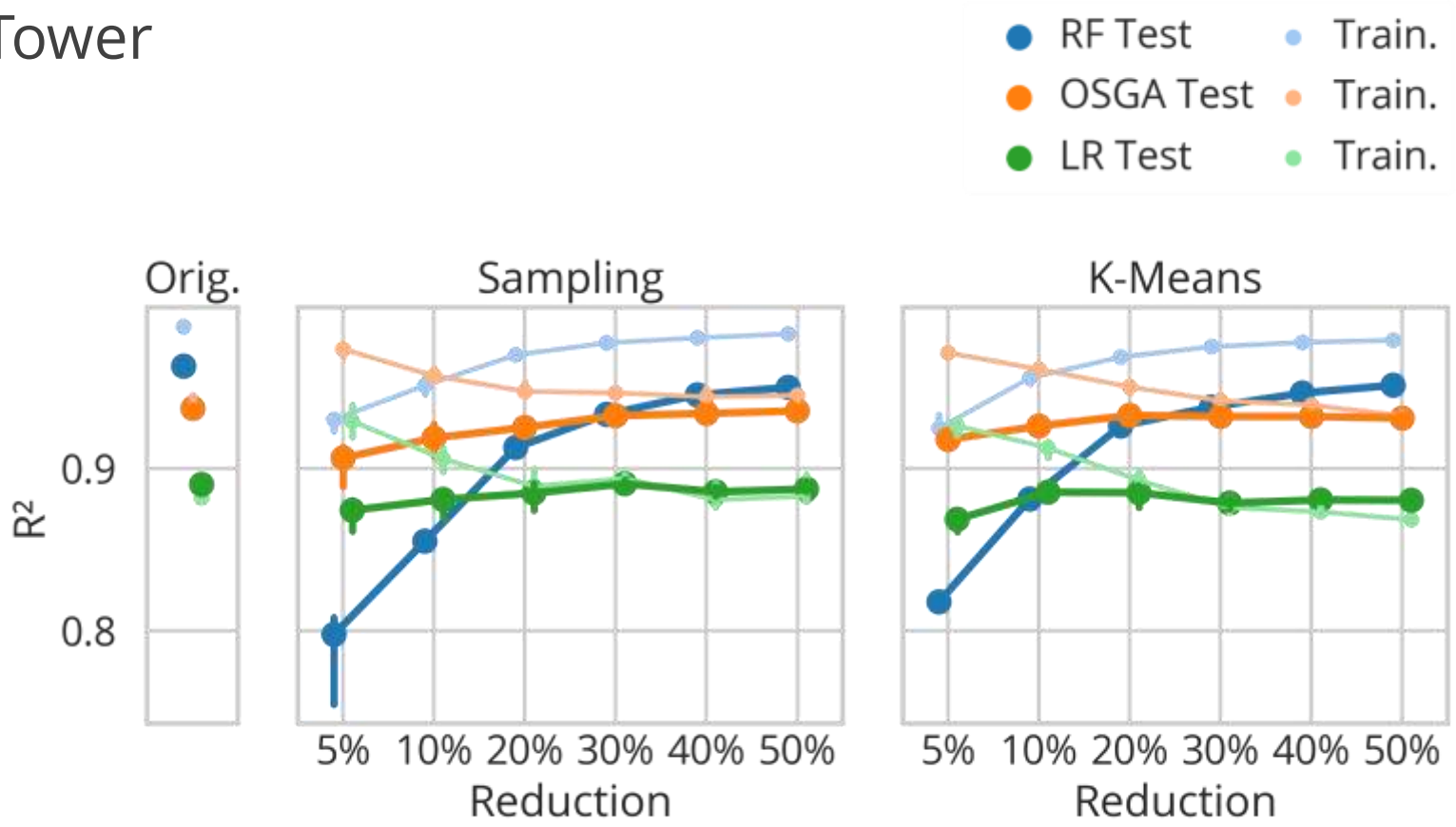


Loss in Quality with OSGA



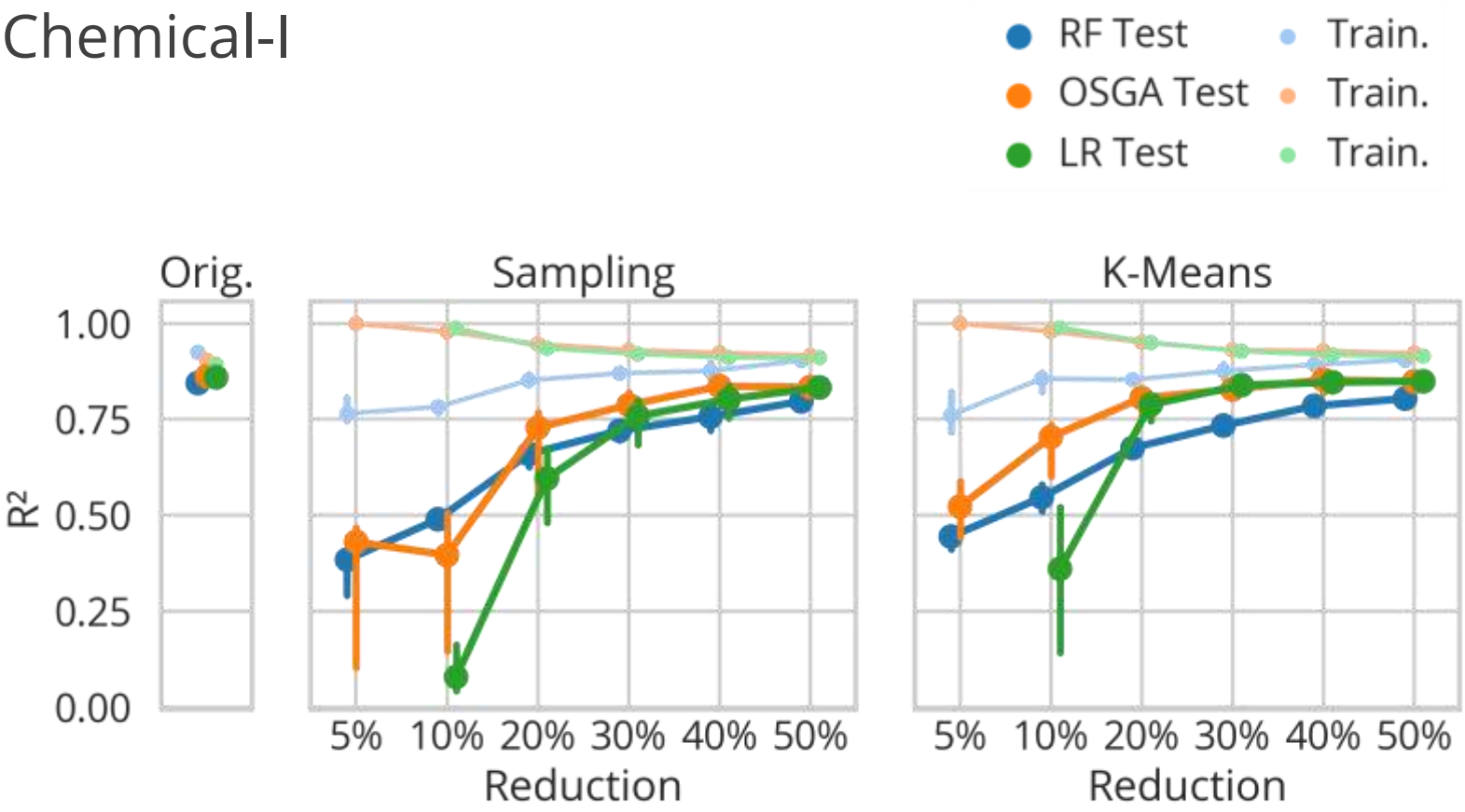
Algorithm Comparison

Tower



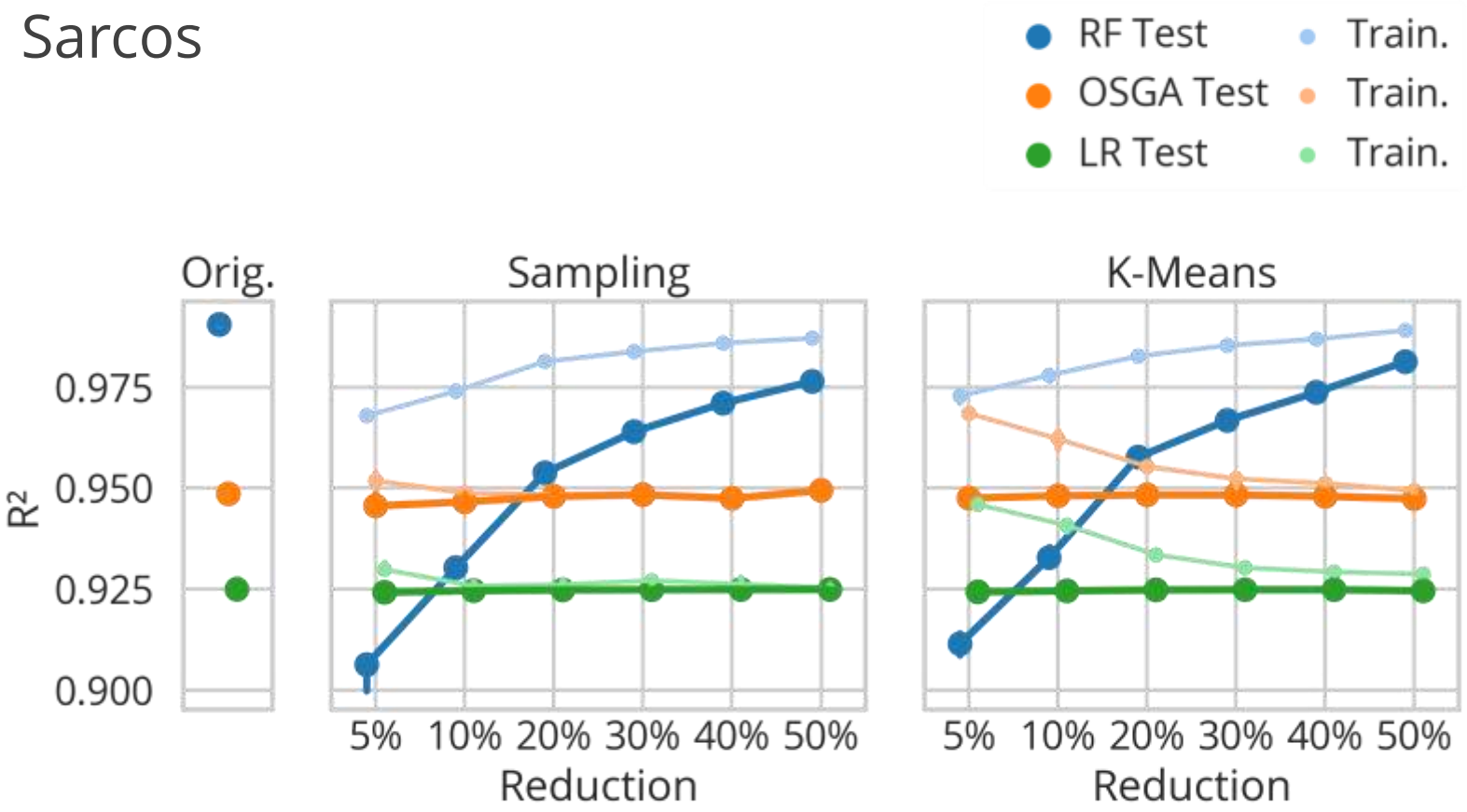
Algorithm Comparison

Chemical-I



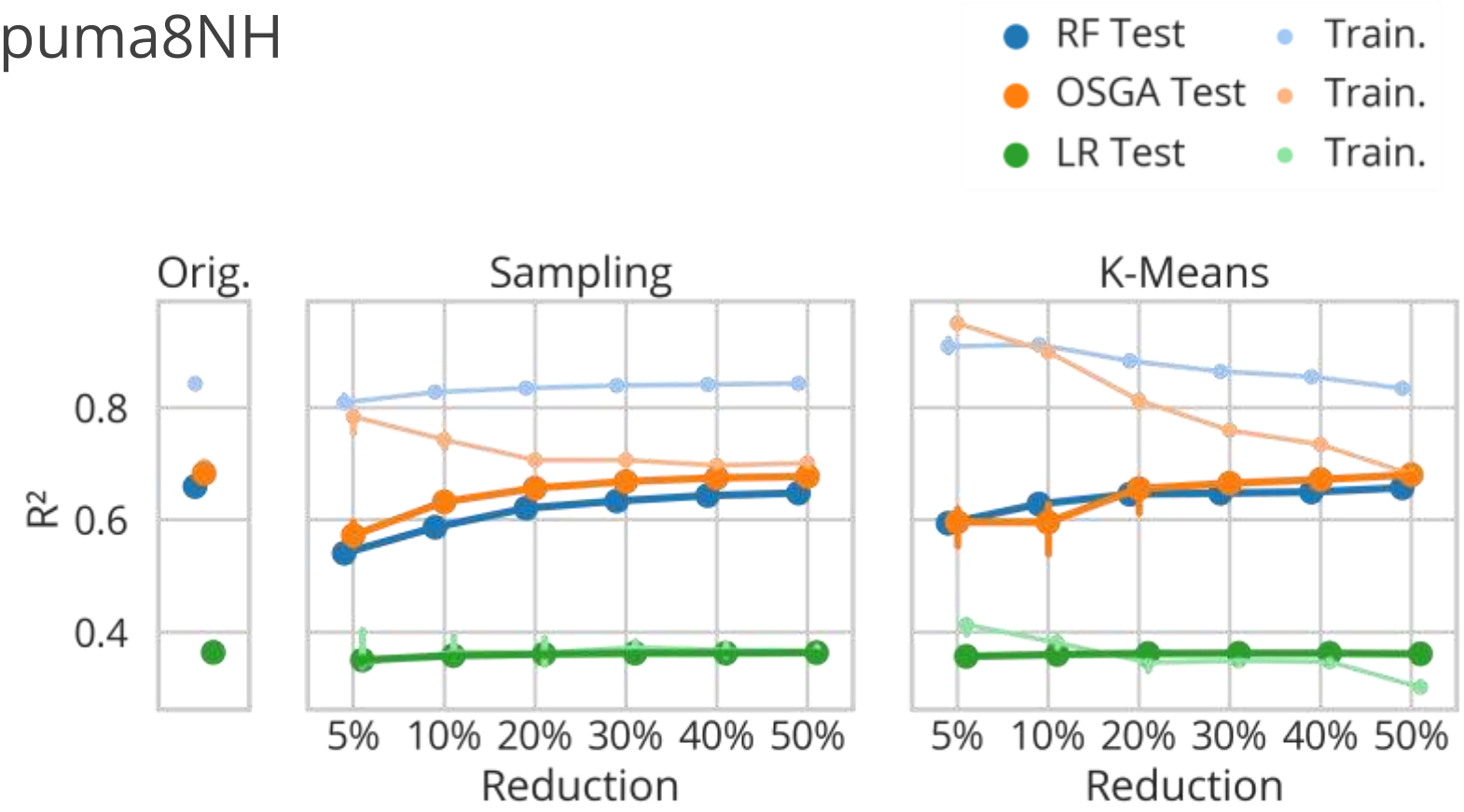
Algorithm Comparison

Sarcos



Algorithm Comparison

puma8NH



Conclusion

- Results of k-means and random sampling comparable
  - Implications on…
    - hyperparameter optimization?
    - statistical tests?
    - etc.?
- Binning not usable
  - Unusable loss in accuracy in all cases

- Little loss of accuracy with >30% undersampling rate
- Speedup linear to undersampling rate

Discussion

Data Aggregation for Reducing Training Data in Symbolic Regression

Lukas Kammerer, Gabriel Kronberger and Michael Kommenda

Josef Ressel Centre for Symbolic Regression (JRZ)
University of Applied Sciences Upper Austria, Hagenberg (HEAL)

HEAL
HEURISTIC AND EVOLUTIONARY ALGORITHMS LABORATORY

SymReg
JOSEF RESSEL CENTER FOR SYMBOLIC REGRESSION

OBERÖSTERREICH

HeuristicLab

EuroCAST 2019

Contact:
Lukas Kammerer
Heuristic and Evolutionary Algorithms Lab (HEAL)
Softwarepark 11
4232 Hagenberg, Austria

E-mail:
lukas.kammerer@fh-hagenberg.at
Web:
http://heal.heuristiclab.com

Algorithm Settings

Offspring Selection Genetic Programming

Max. Tree Size	Depth: 30 Length:50
Grammar	+ - × ÷ exp log square sqrt cbrt sin cos tan
Population Size	300
Mutation Rate	20 %
Selector	GenderSpecific (Random & Proportional)

Random Forest

R	30 %
M	50 %
Number of Trees	50

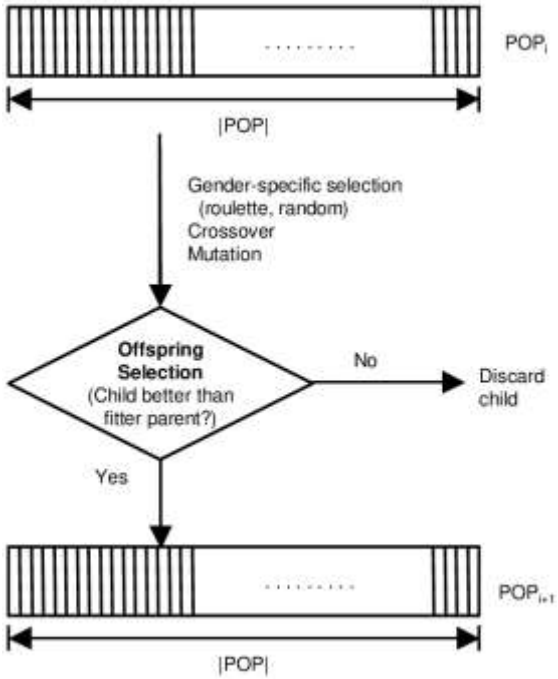


Image: Affenzeller, M., Wagner, S., & Winkler, S. (2007, July). *Aspects of adaptation in natural and artificial evolution*. In Proceedings of the 9th annual conference companion on Genetic and evolutionary computation (pp. 2595-2602). ACM