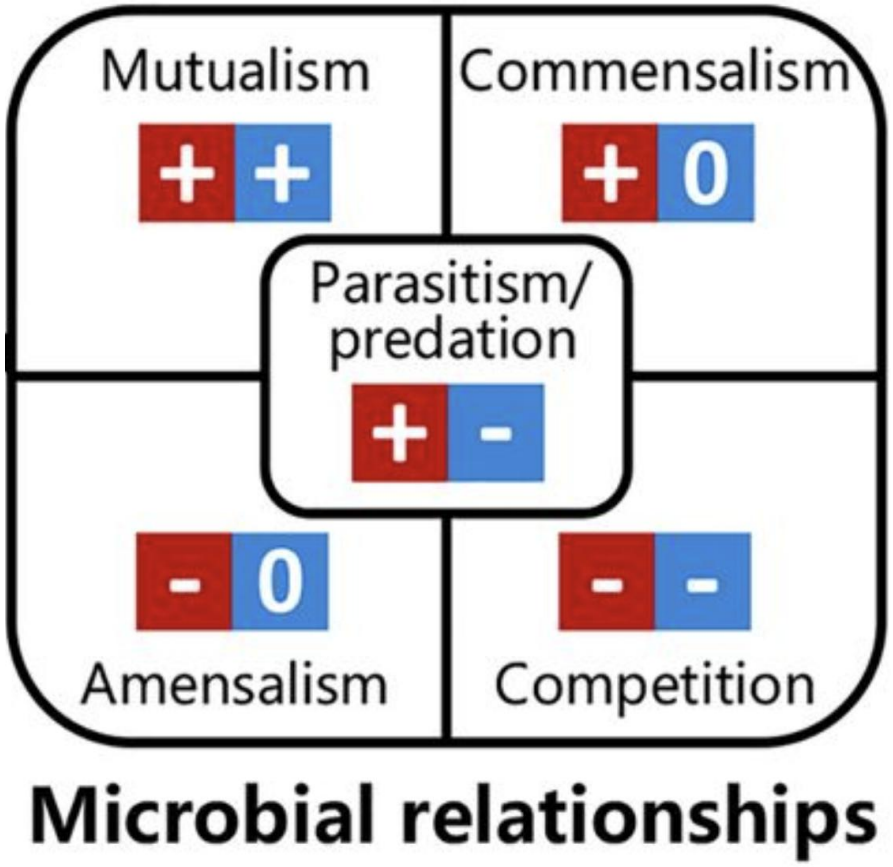
# Cross-Validation for training and testing co-occurrence network inference algorithms


Daniel Agyapong
da2343@nau.edu
PhD student

Dr. Toby Hocking
Toby.Hocking@nau.edu
Machine Learning Director

NAU
NORTHERN ARIZONA UNIVERSITY
School of Informatics, Computing, and Cyber Systems

# INTRODUCTION




**Microbial relationships**

Source:
https://journals.asm.org/doi/10.1128/mSystems.00124-19

- Most state-of-the-art methods focus on inferring **positive** and **negative** associations between bacteria.
- Reconstructing microbial networks to represent these interactions would help to understand the complex behaviors in microbial communities.

# Real Microbiome Abundance Data

| Data | Citation | Samples | Taxa |
|---|---|---|---|
| amgut1 | https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226 | 289 | 127 |
| amgut2 | | 296 | 138 |
| hmp216S | https://ibdmdb.org/tunnel/public/summary.html | 47 | 45 |
| hmp2prot | | 47 | 43 |
| enterotype | https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217 | 280 | 553 |
| esophagus | | 3 | 58 |
| crohns | https://www.mcgill.ca/statisticalgenetics/software | 100 | 5 |
| Baxter_CRC | http://www.raeslab.org/companion/ocean-interactome.html | 490 | 117 |
| glne007 | | 490 | 338 |
| iOraldat | https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03911-w | 86 | 63 |

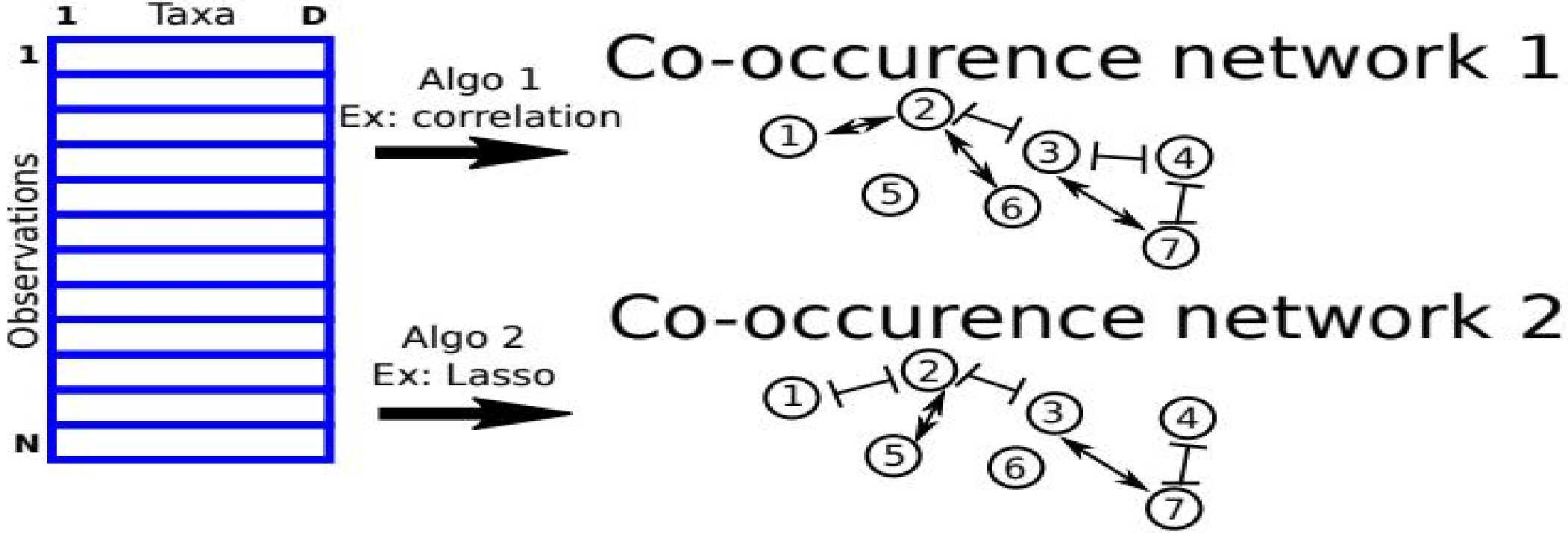
Each data set is a matrix of counts, for example:

**Taxa**

**Samples**

| | | |
|---|---|---|
| 0 | 15 | 761 |
| 4 | 0 | 98 |
| 53 | 74 | 0 |
| 0 | 32 | 0 |
| 11 | 0 | 0 |
| 0 | 24 | 65 |

# Different algorithms infer different co-occurence networks




Which is a more accurate interpretation for these data?

**Associations**:
Positive ↔
Negative ⊣

# Research Questions

For some particular real data sets:

- How can we automatically learn hyper-parameters? (let the data tell us the “best” threshold, rather than choosing arbitrarily)
- Which of the available microbial network analysis algorithms is most accurate and gives least test error ?
- How many samples are needed for Cross Validation to be useful.

# What are some of the Existing Algorithms?

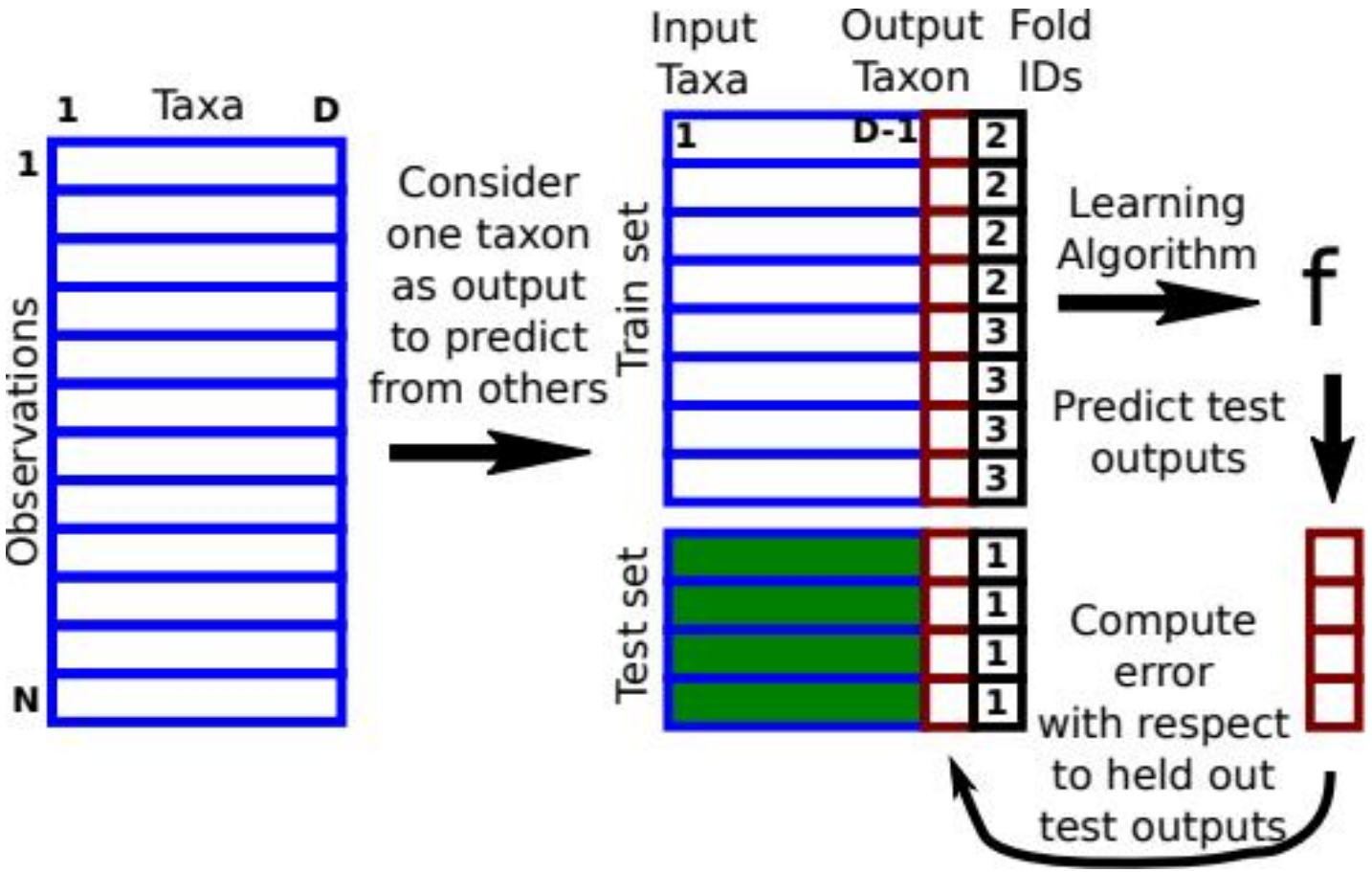
There are a lot of existing algorithms, each with various hyper-parameters which determine the sparsity (number of edges) in network.

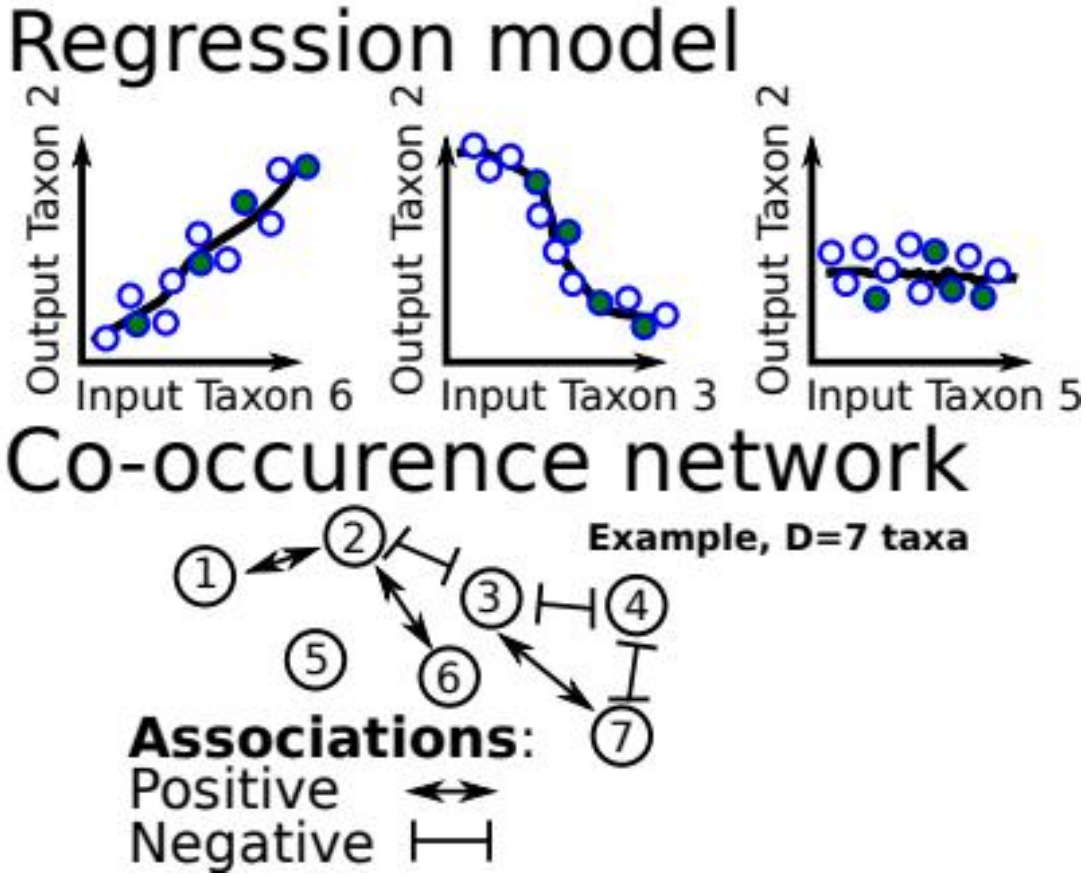
| Pearson/Spearman Correlation | Least Absolute Shrinkage and Selection Operator (LASSO) | Gaussian Graphical Model (GGM) |
| --- | --- | --- |
| • Threshold on correlation constant | • Degree of L1 regularization | • Inverse Covariance (Precision) Matrix |

# Existing Evaluation Types

| Category of Evaluation Type | External Data | Sub-Sample Analysis |
| --- | --- | --- |
| Methods | SparCC, REBACCA, SPIEC-EASI, gCoda, COZINE, HARMONIES, mLDM | CClasso |
| Issues | • Lack of generalizability<br>• Lack of ground truth<br>• Biases in external data | • Sensitivity to the choice of subsampling parameters<br>• Limited scope |

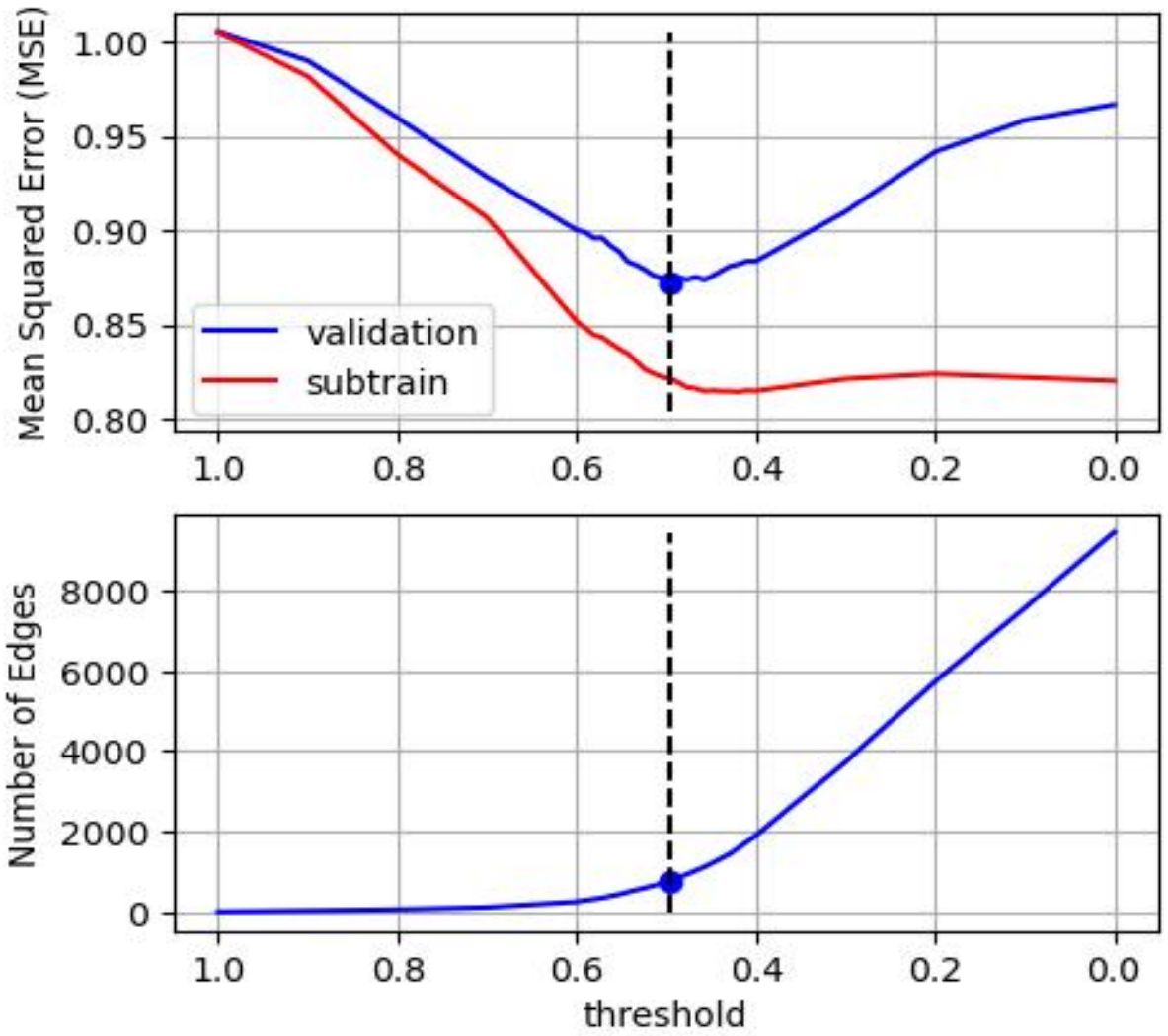
# Proposal: Cross-validation for training and testing co-occurrence network inference algorithms




**Repeat for each output taxon and Fold ID**

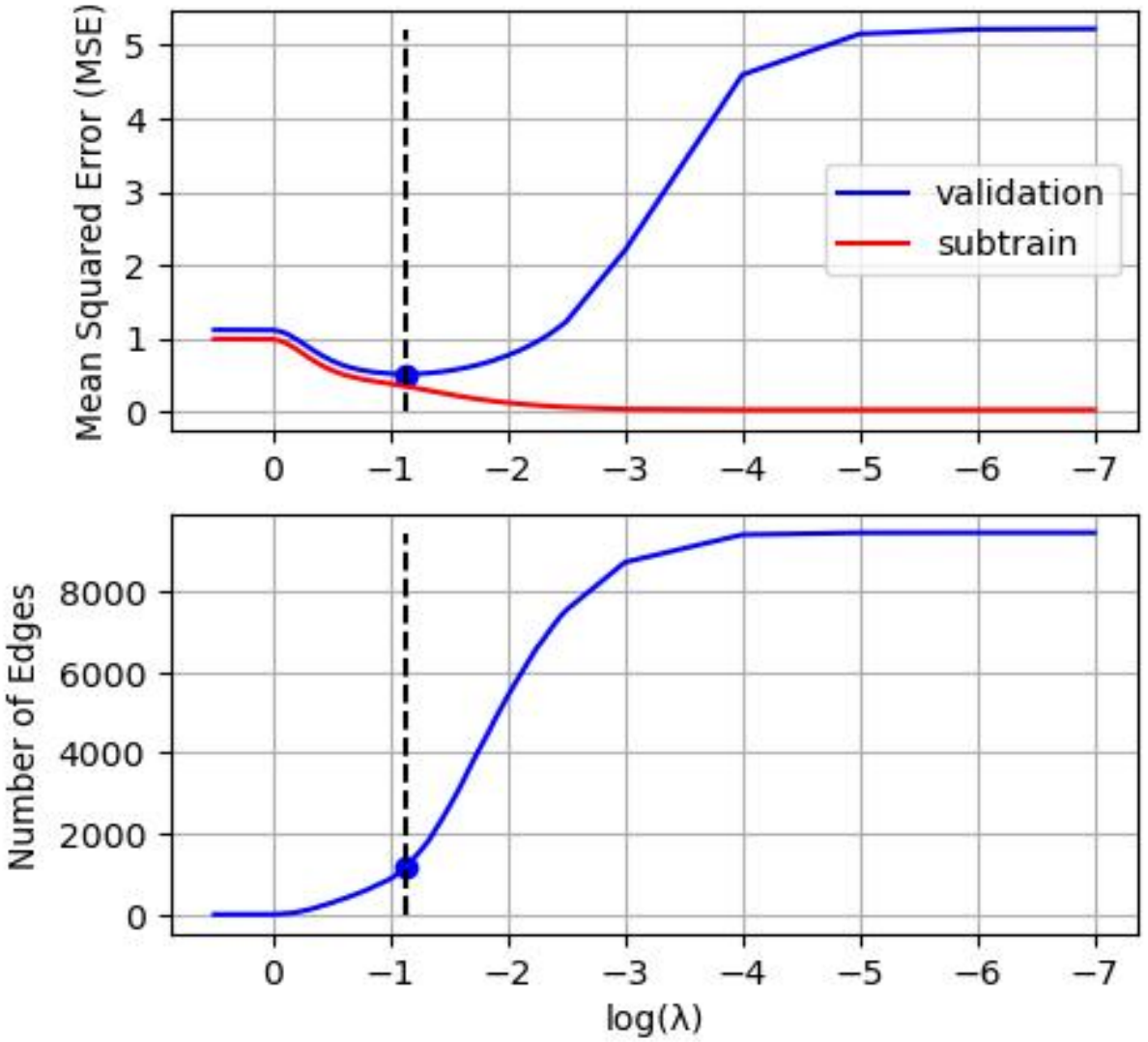
# Results: Training the Pearson correlation threshold using cross-validation




- Subtrain error decreases as the model complexity increases whilst the validation error shows a U shape.

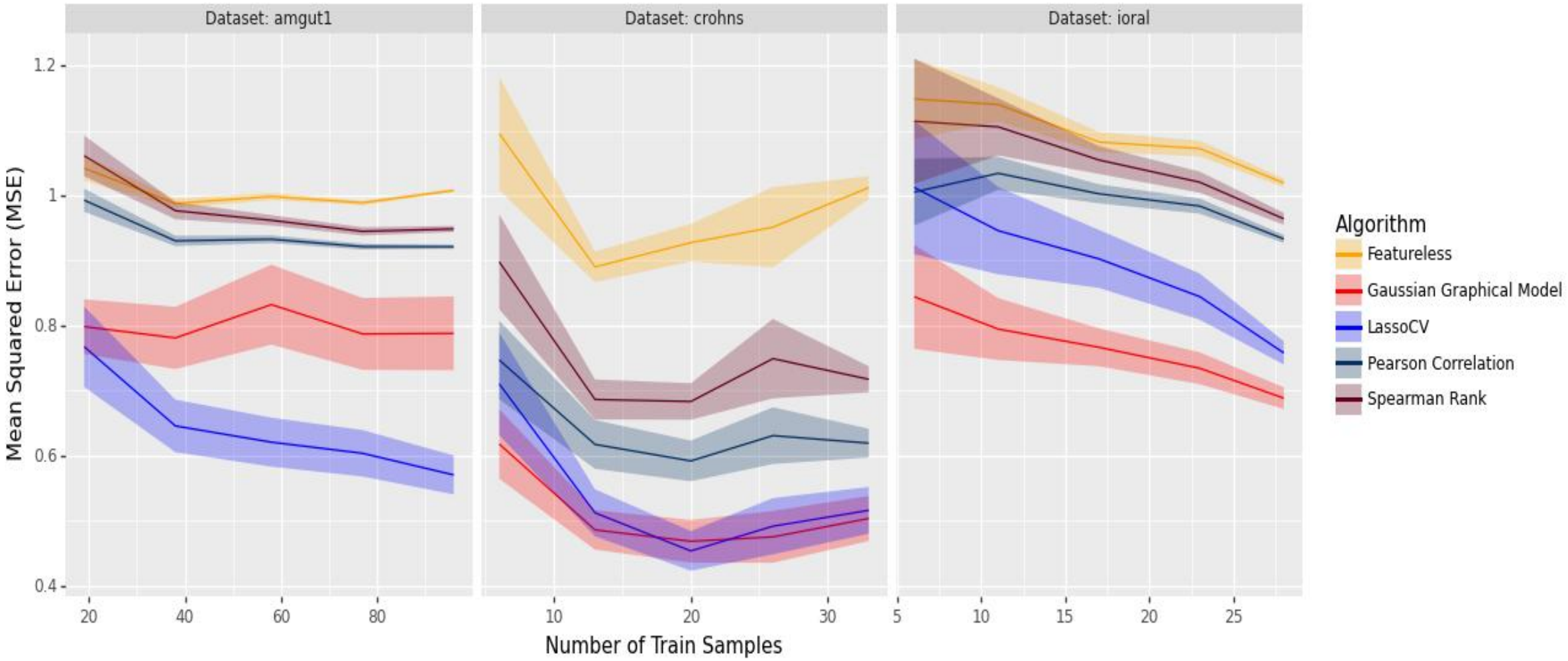
- We select the threshold which gives the minimum validation error, in this example r2=0.5 (Best number of edges = 785).

# Results: Training the Lasso algorithm with cross-validation

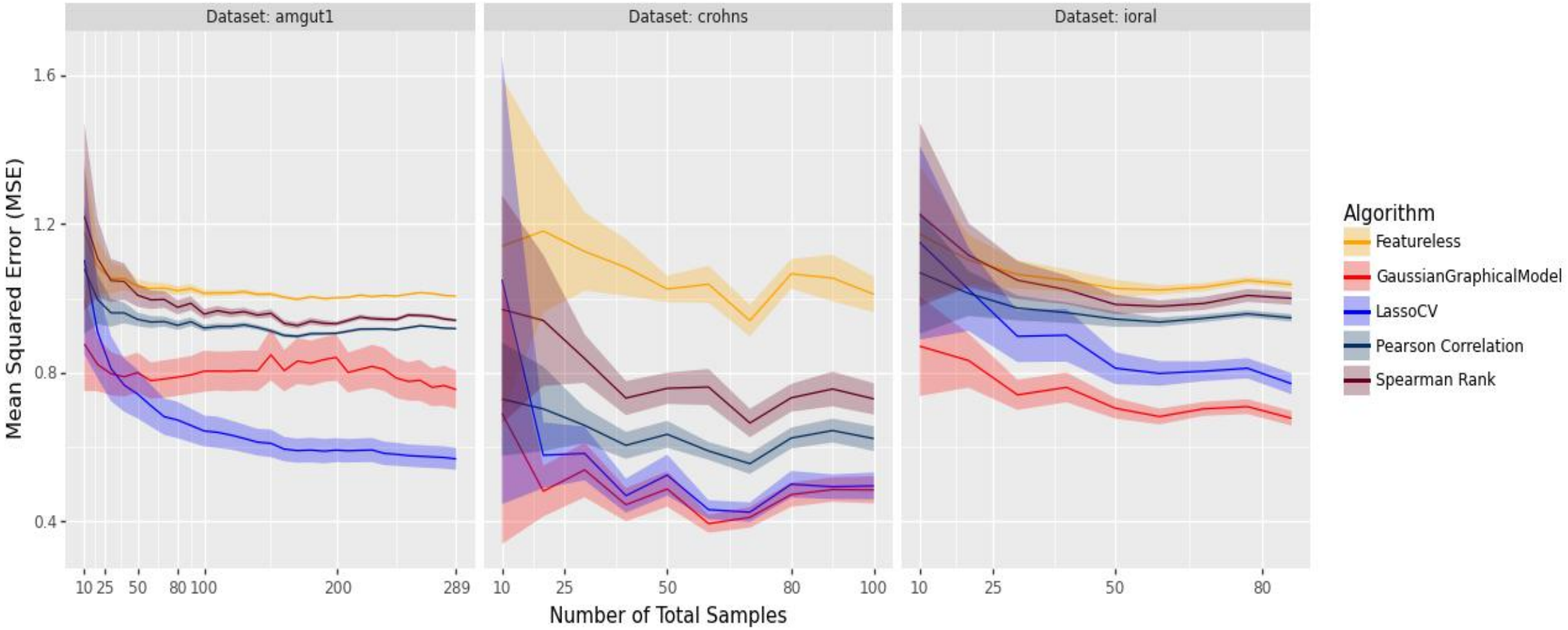



- Train set split into subtrain set (used to learn regression coefficients) and validation set (used to learn model complexity, degree of L1 regularization, which controls sparsity / number of edges in co-occurence network).
- The lamda value (degree of L1 regularization) which has the minimimum the validation error corresponds to 1176 edges.

# Results: Algorithms can be compared using test error



# Results: How many samples are needed for CV to be useful



# REFERENCES


- https://www.liebertpub.com/doi/10.1089/cmb.2021.0406
- https://smnh.tau.ac.il/en/interactions-among-living-organisms/
- https://scikit-learn.org/stable/modules/cross_validation.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7768662/
- https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004226
- https://www.sciencedaily.com/releases/2018/05/180515092931.htm



Contact: da2343@nau.edu


Reproducibility: https://github.com/EngineerDanny/CS685-Microbe-Network-Research

# THANK YOU

# ANY QUESTIONS?