Cross-validation: what, how and which?

Pradeep Reddy Raamana

Statistics [from cross-validation] are like bikinis👙.
What they reveal is suggestive, but what they conceal is vital!

Baycrest

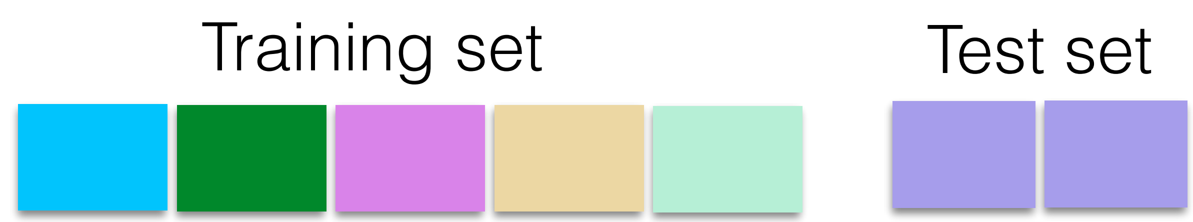
Reference: Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2016). **Assessing and tuning brain decoders: cross-validation, caveats, and guidelines.** NeuroImage.

ONTARIO BRAIN INSTITUTE

Goals for Today

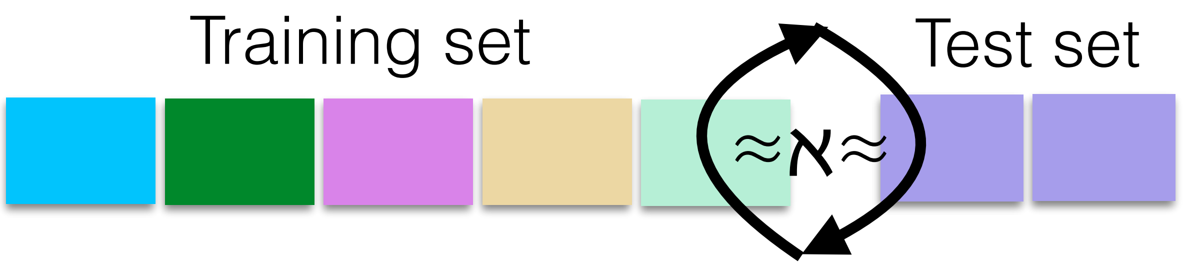
Goals for Today

- What is cross-validation?



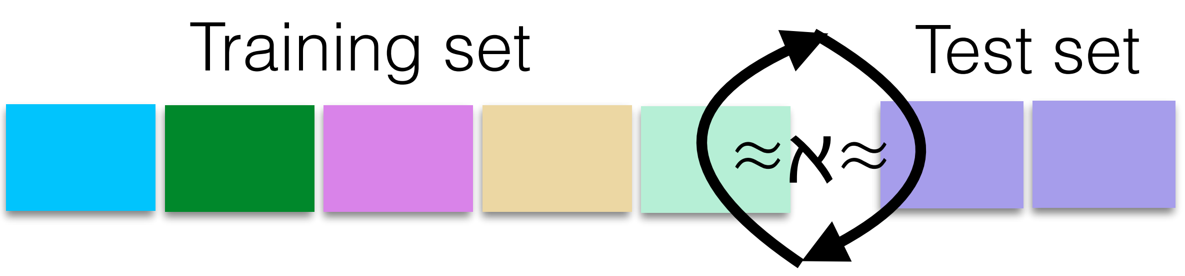
Goals for Today

- What is cross-validation?
- How to perform it?

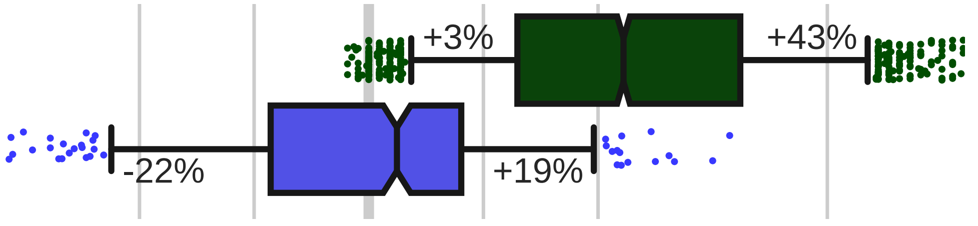


Goals for Today

- What is cross-validation?

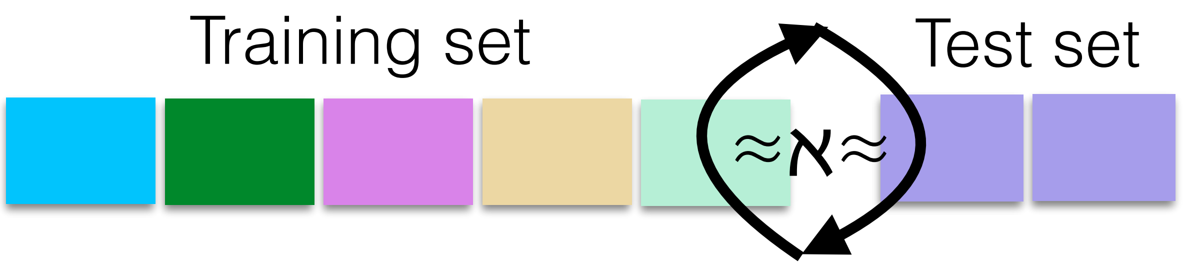


- How to perform it?
- What are the effects of different CV choices?

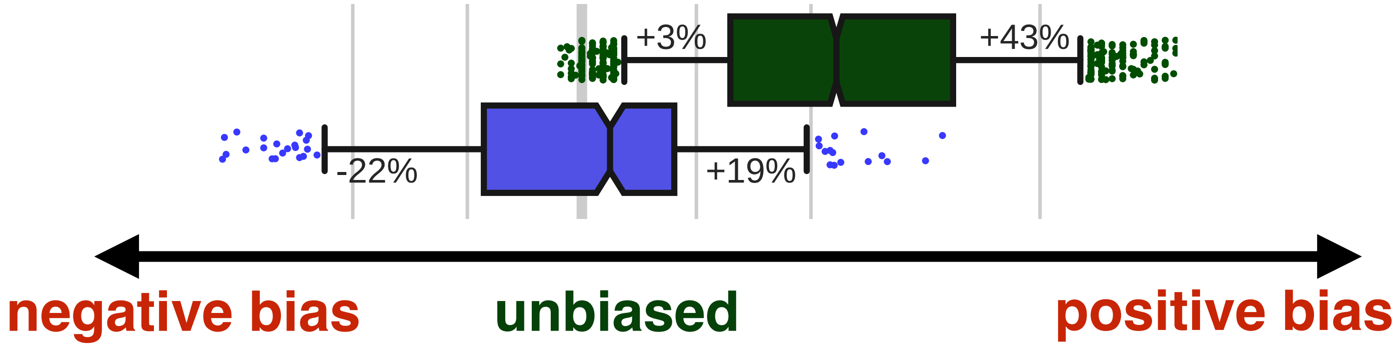


Goals for Today

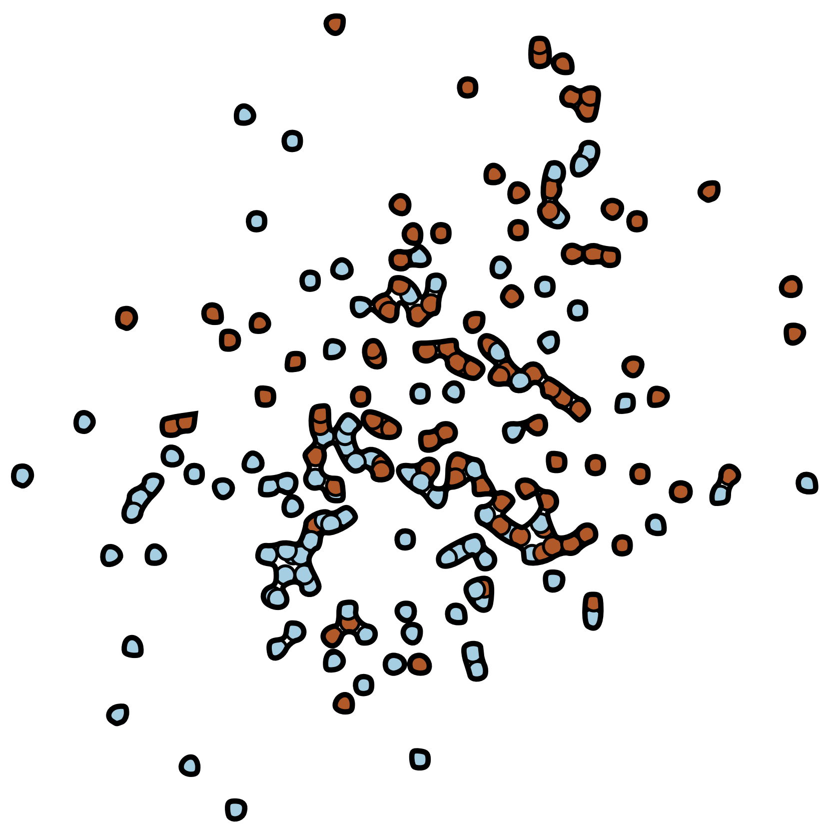
- What is cross-validation?



- How to perform it?
- What are the effects of different CV choices?

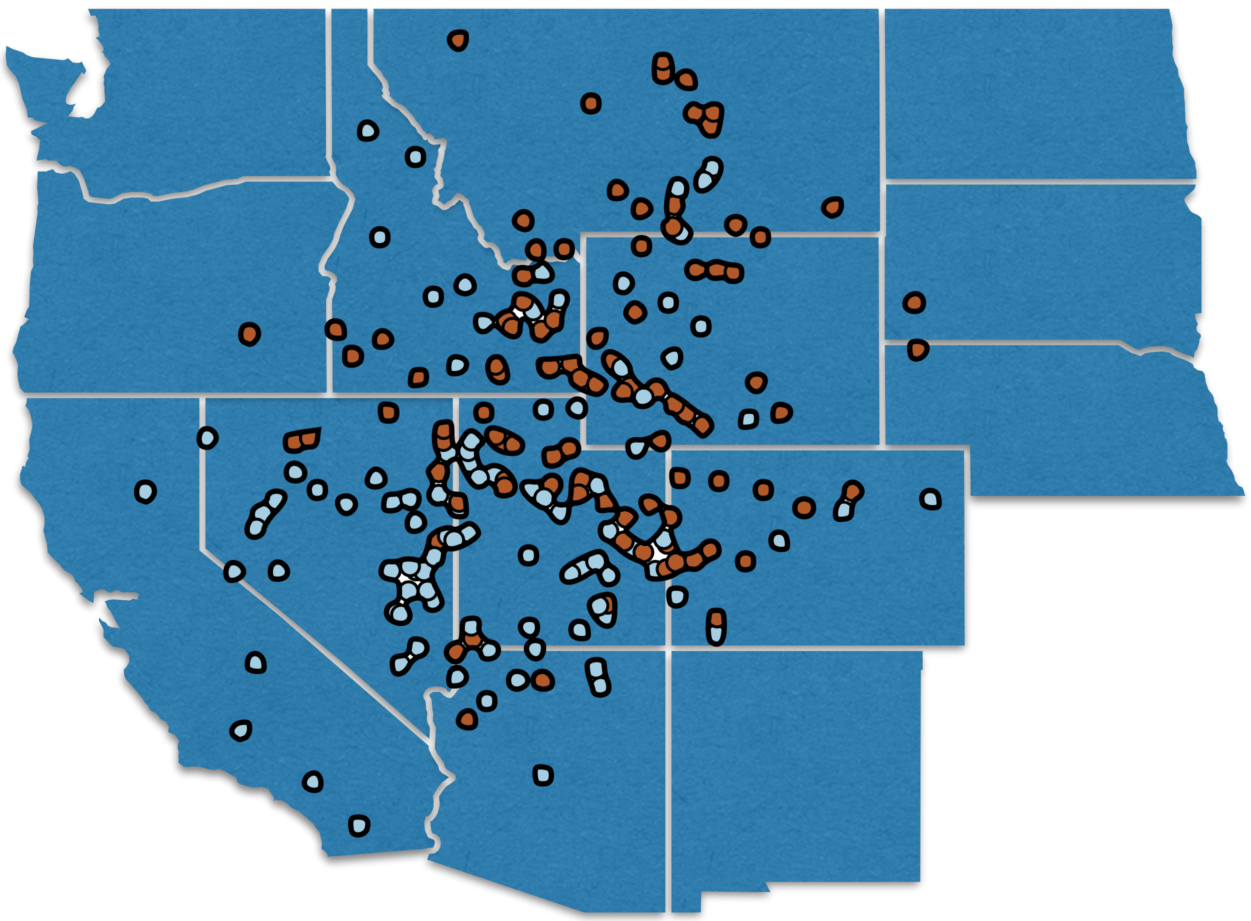


What is generalizability?



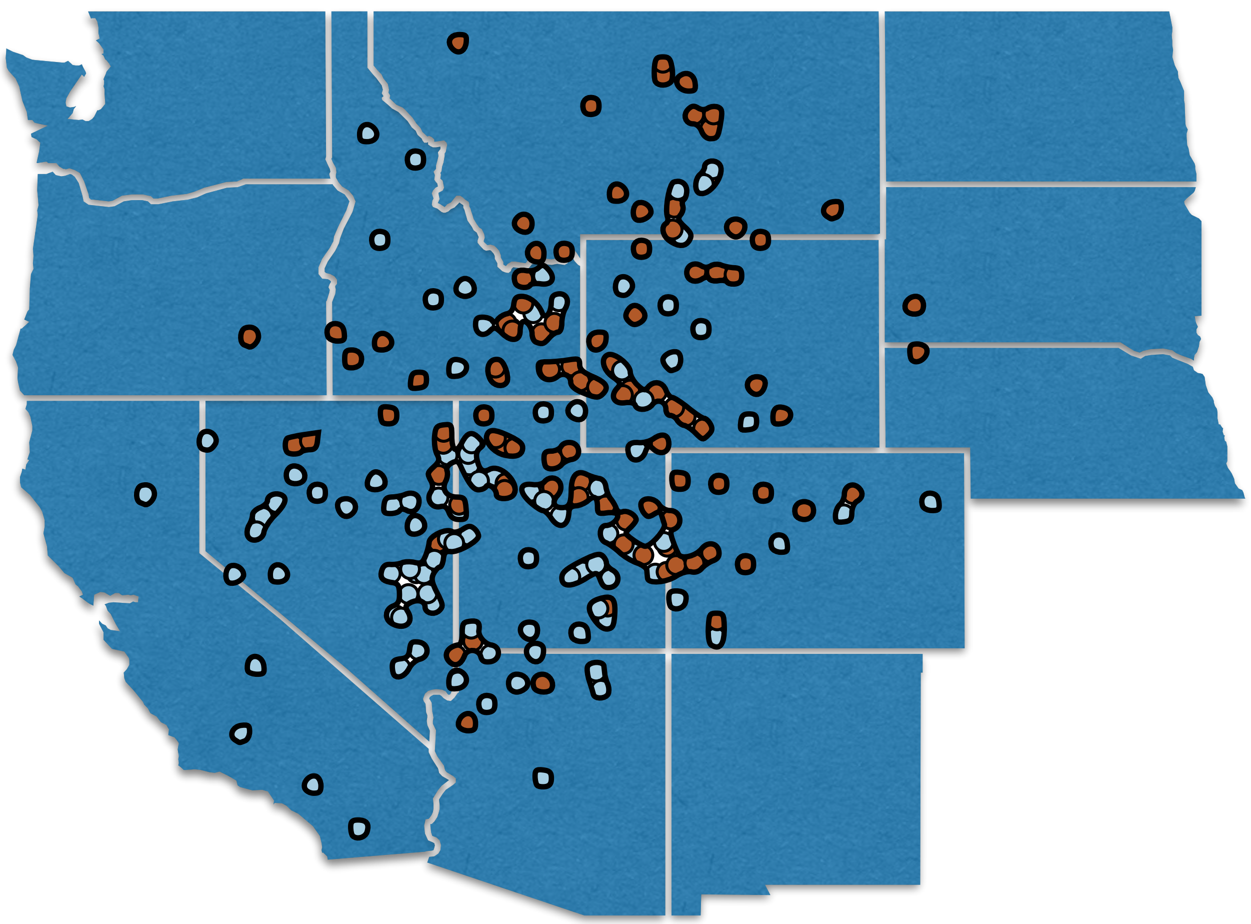
available
data (sample)

What is generalizability?



available
data (sample)

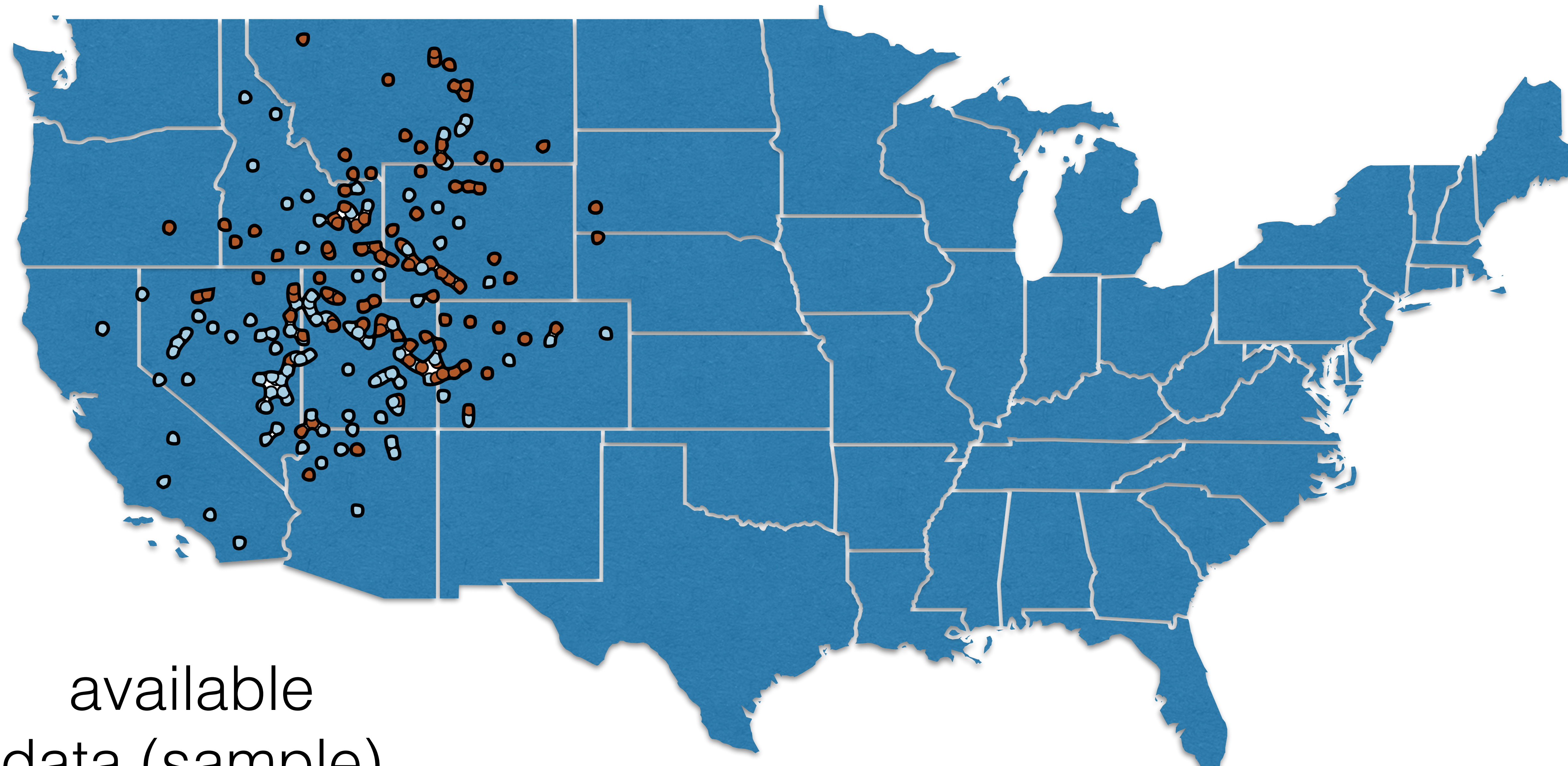
What is generalizability?



available data (sample)

desired: accuracy on **unseen** data (population)

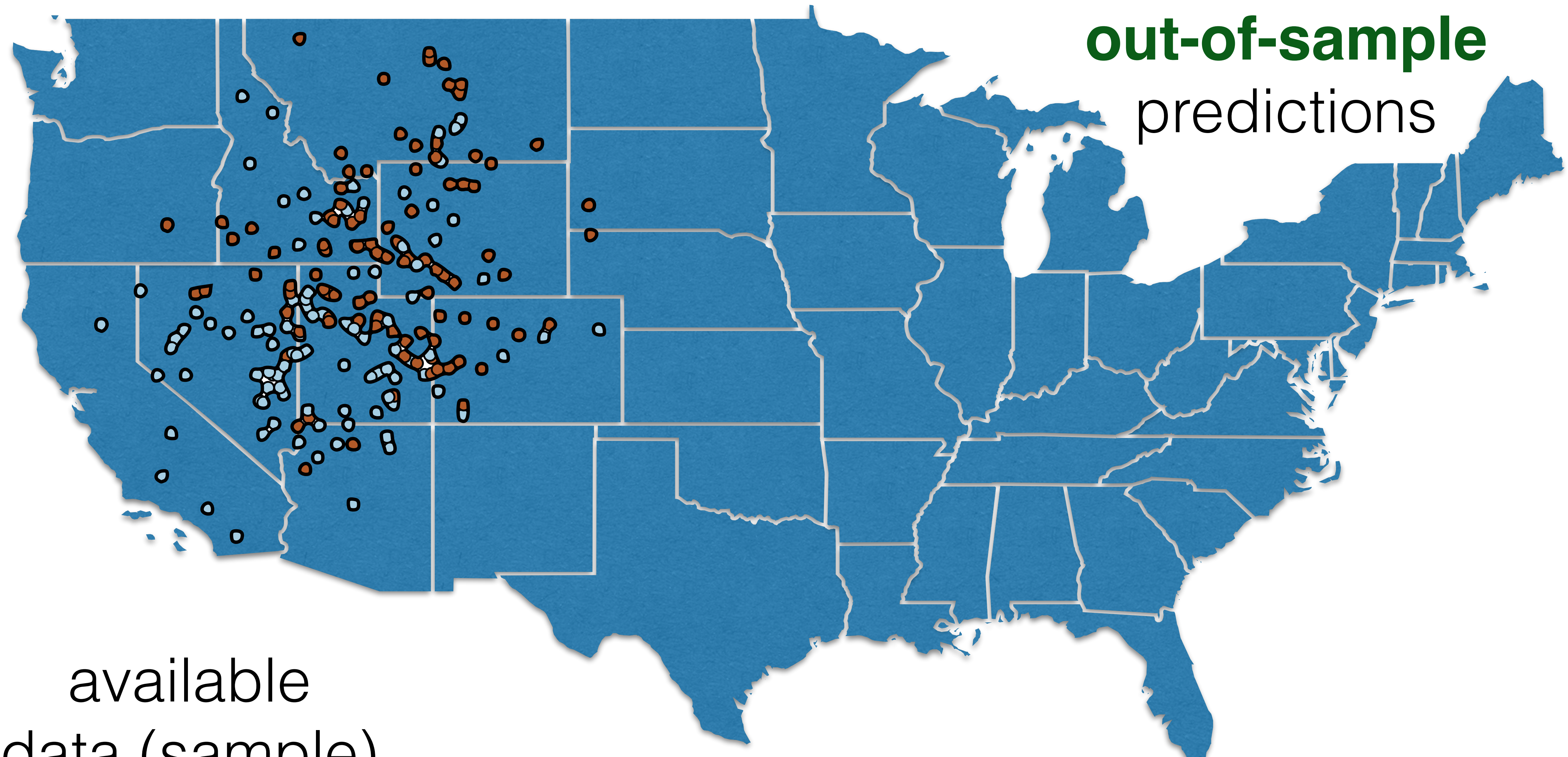
What is generalizability?



available
data (sample)

desired: accuracy on
unseen data (population)

What is generalizability?

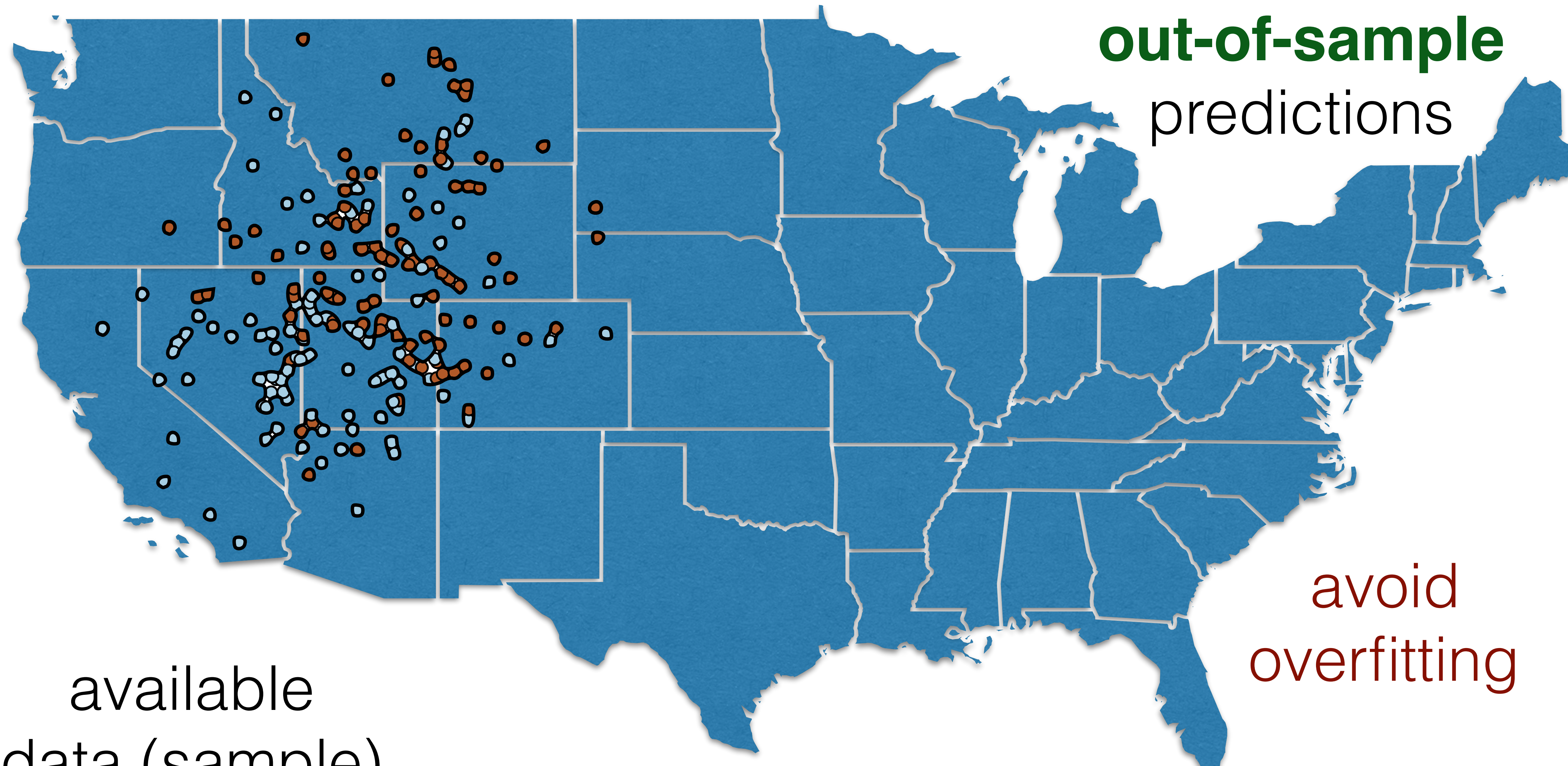


out-of-sample predictions

available data (sample)

desired: accuracy on **unseen** data (population)

What is generalizability?



out-of-sample predictions

avoid overfitting

available data (sample)

desired: accuracy on **unseen** data (population)

Why cross-validate?

Why cross-validate?

bigger training set

↓

better **learning**

Why cross-validate?

Training set

Test set

bigger training set

↓

better **learning**

bigger test set

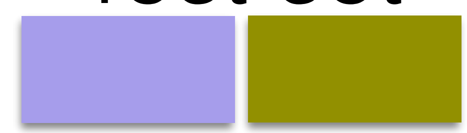
↓

better **validation**

Why cross-validate?

Training set

Test set



bigger training set

↓

better **learning**

bigger test set

↓

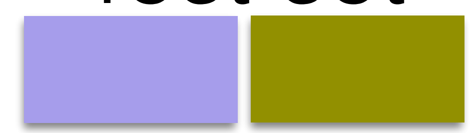
better **validation**

Key: training and test sets are **disjoint.**

Why cross-validate?

Training set

Test set



bigger training set

↓

better **learning**

bigger test set

↓

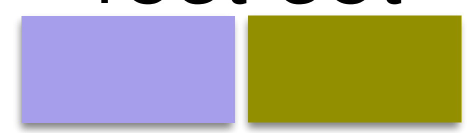
better **validation**

Key: training and test sets are **disjoint.**
And the dataset or sample size is fixed.

Why cross-validate?

Training set

Test set



bigger training set

↓

better **learning**

bigger test set

↓

better **validation**

Key: training and test sets are **disjoint.**
And the dataset or sample size is fixed.
They grow at the expense of each other!

Why cross-validate?

bigger training set → better **learning**

bigger test set → better **validation**

Key: training and test sets are **disjoint.**
And the dataset or sample size is fixed.
They grow at the expense of each other!

Why cross-validate?

Training set

Test set

bigger training set → better **learning**

bigger test set → better **validation**

Key: training and test sets are **disjoint.**
And the dataset or sample size is fixed.
They grow at the expense of each other!

⇒ **cross**-validate to maximize both

Use cases

Use cases

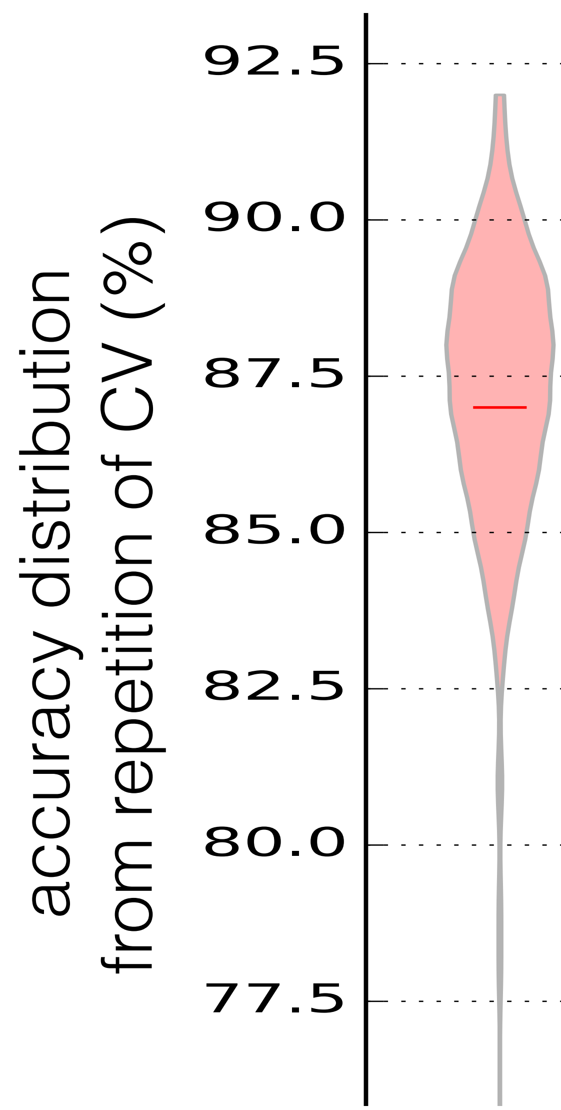
- “When setting aside data for parameter estimation and validation of results can not be afforded, cross-validation (CV) is typically used”

Use cases

- "When setting aside data for parameter estimation and validation of results can not be afforded, cross-validation (CV) is typically used"
- Use cases:

Use cases

- "When setting aside data for parameter estimation and validation of results can not be afforded, cross-validation (CV) is typically used"
- Use cases:
    - to estimate generalizability (test accuracy)



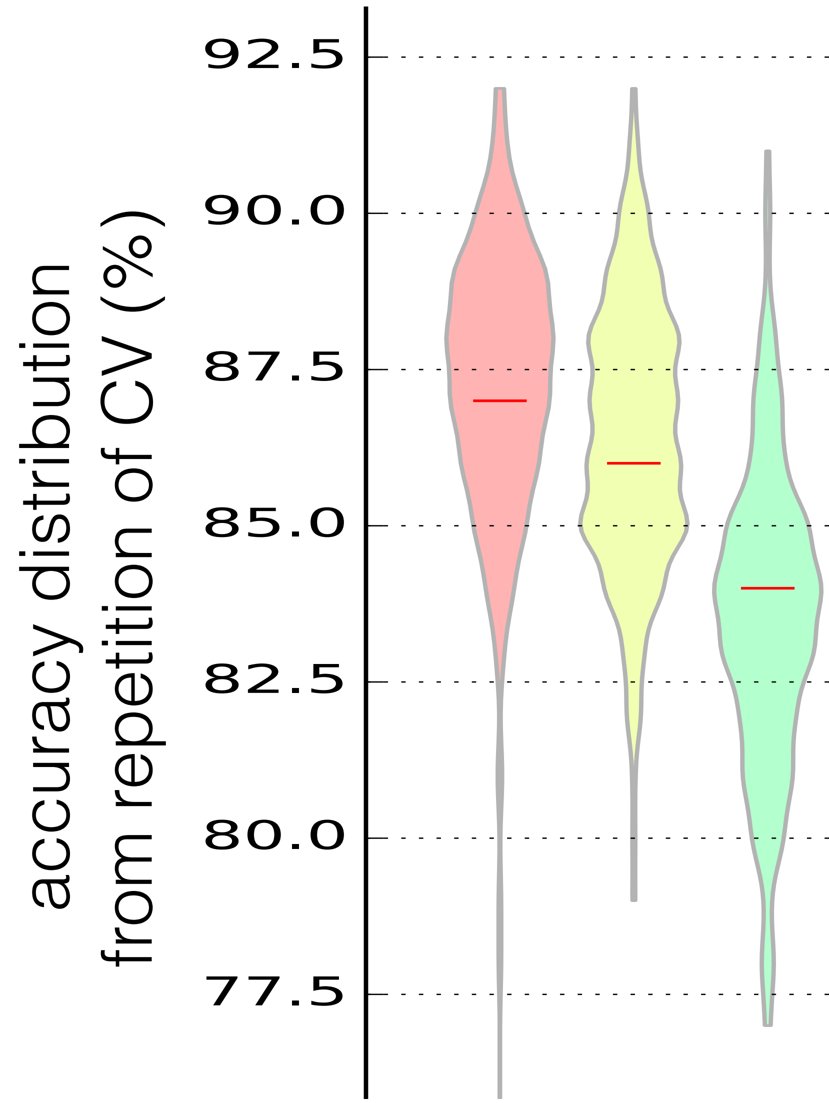
Use cases

- "When setting aside data for parameter estimation and validation of results can not be afforded, cross-validation (CV) is typically used"

- Use cases:

  - to estimate generalizability (test accuracy)

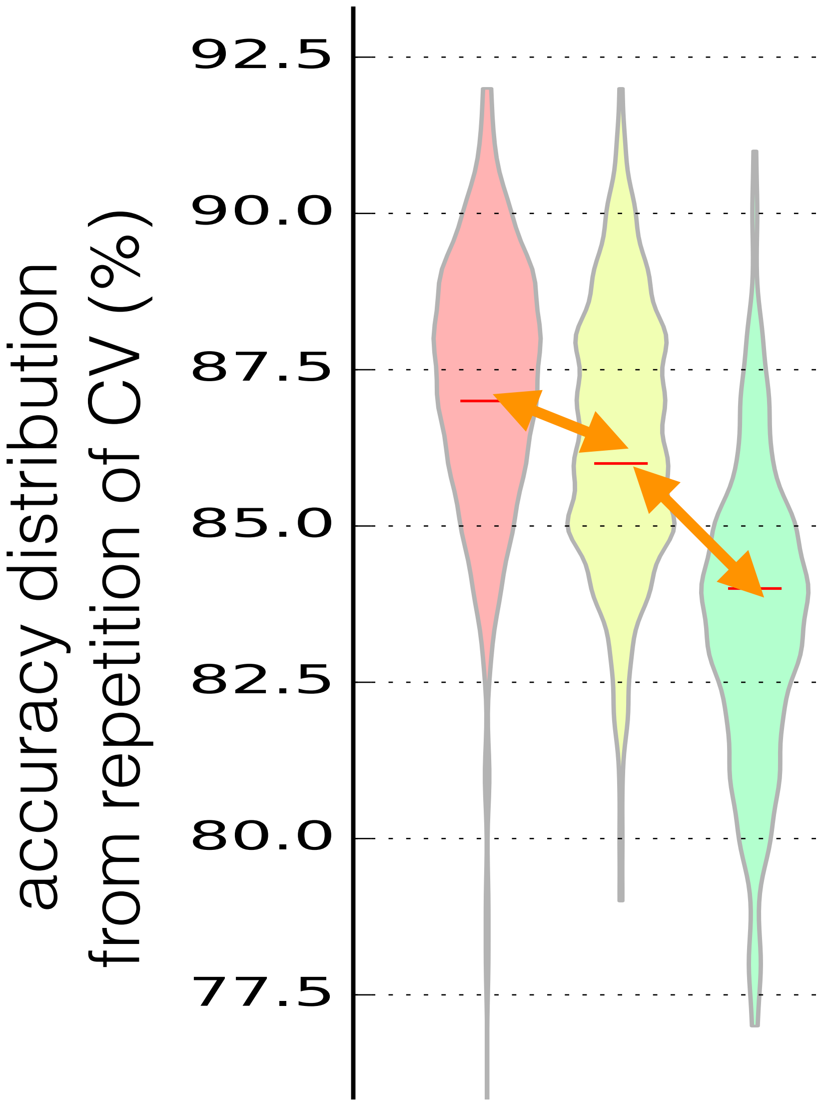
  - to pick optimal parameters (model selection)



Use cases

- "When setting aside data for parameter estimation and validation of results can not be afforded, cross-validation (CV) is typically used"

- Use cases:
    - to estimate generalizability (test accuracy)
    - to pick optimal parameters (model selection)
    - to compare performance (model comparison).



Types of CV

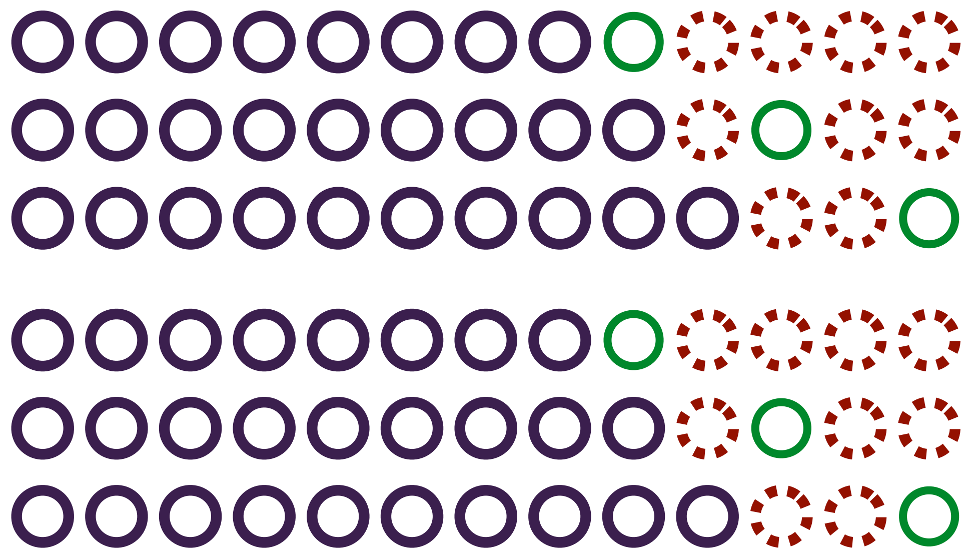
Types of CV

1. **How you split** the dataset into train/test

Types of CV

1. **How you split** the dataset into train/test

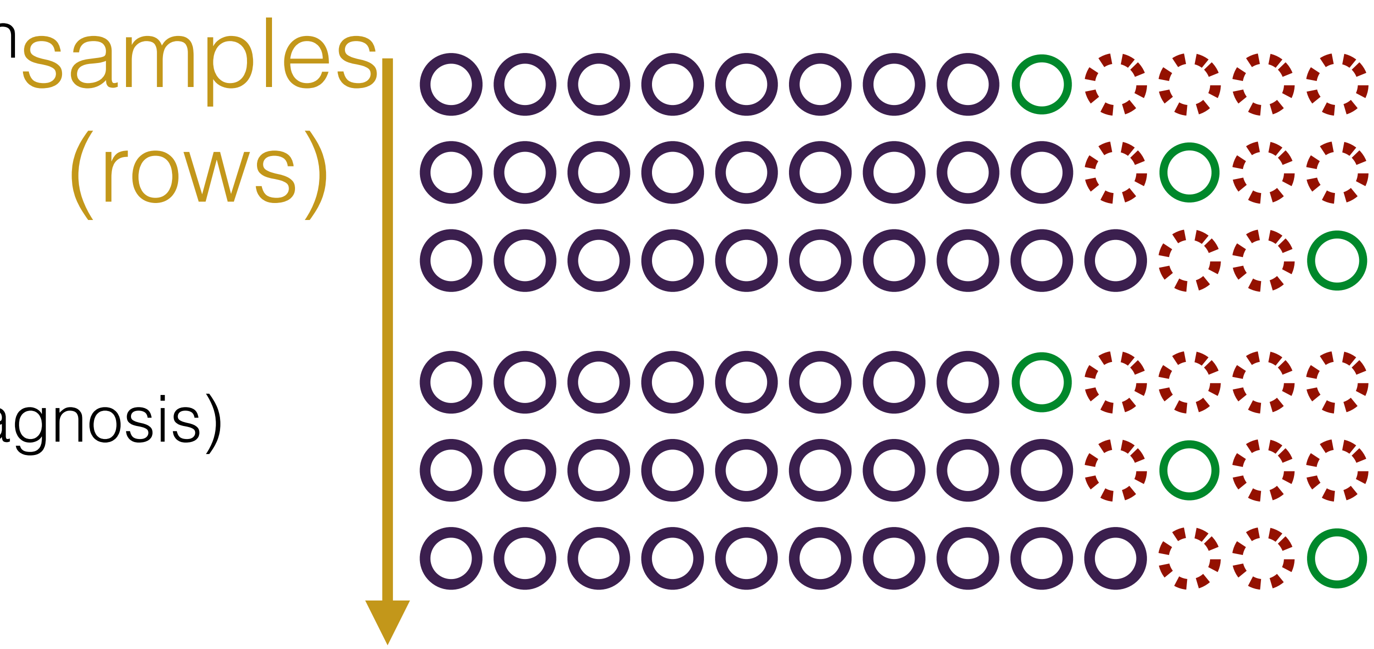
   - maximal independence between training and test sets is desired.



Types of CV

1. **How you split** the dataset into train/test

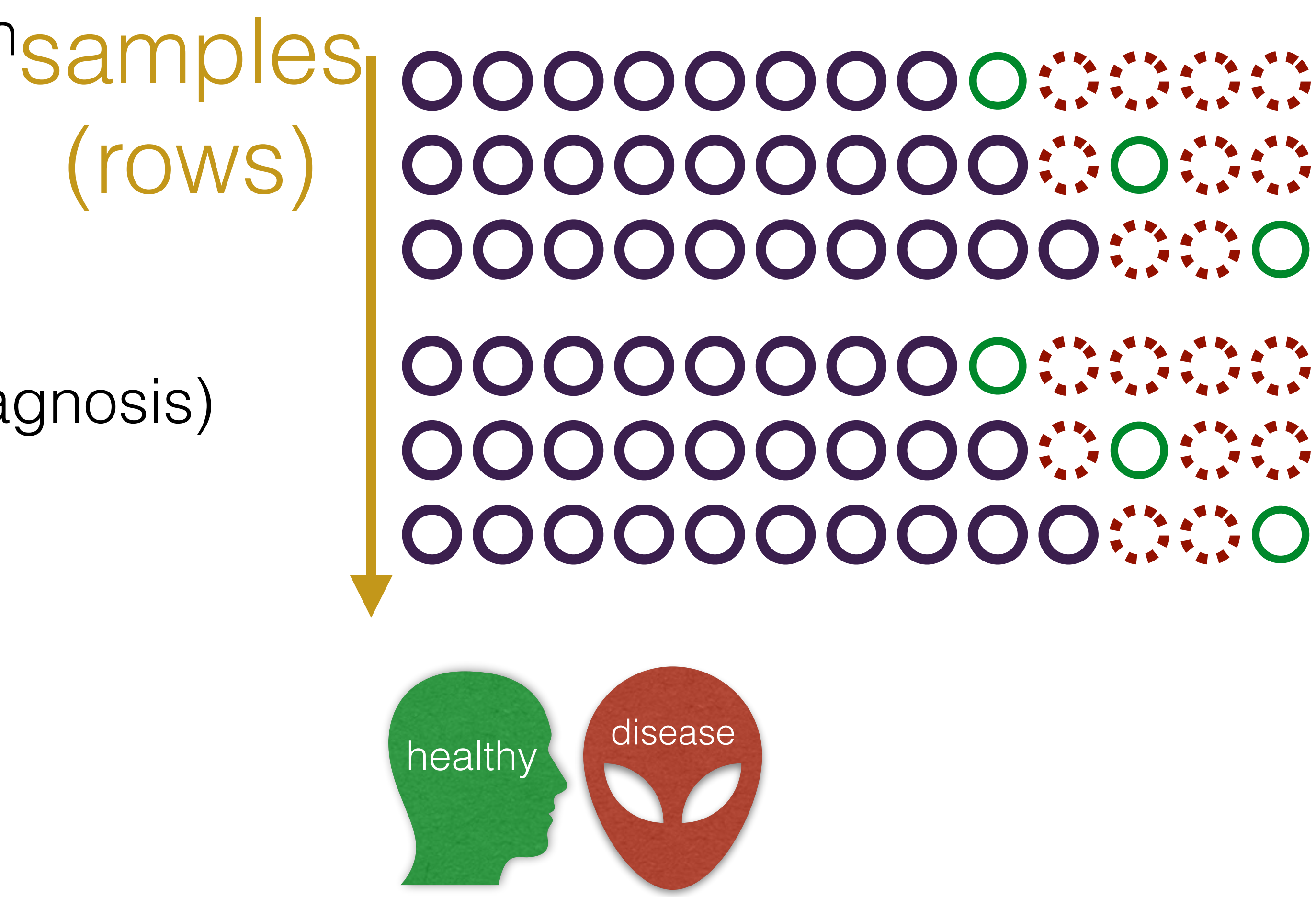
- maximal independence between training and test sets is desired.
- This split could be
  - over samples (e.g. indiv. diagnosis)



Types of CV

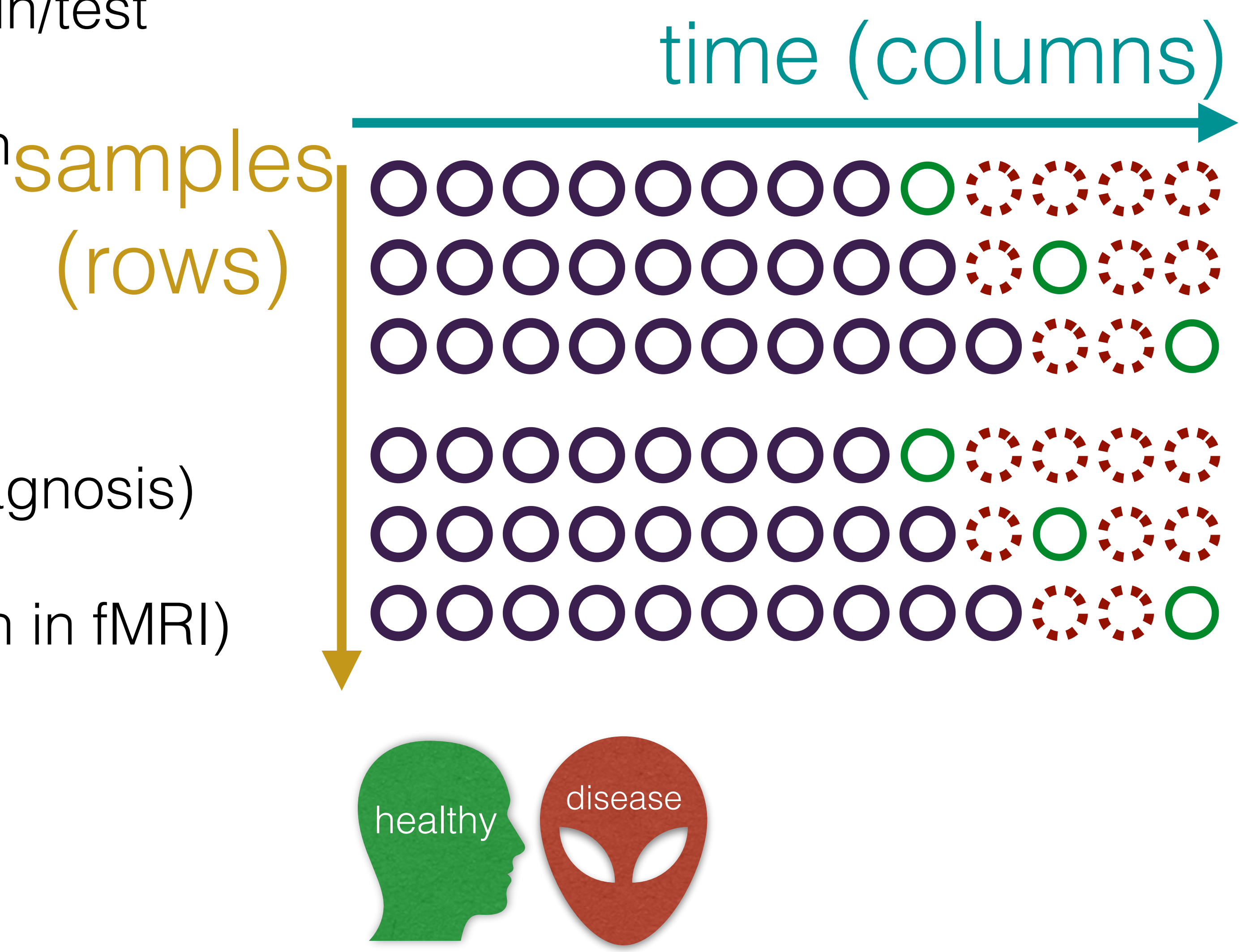
1. **How you split** the dataset into train/test

- maximal independence between training and test sets is desired.
- This split could be
  - over samples (e.g. indiv. diagnosis)



Types of CV

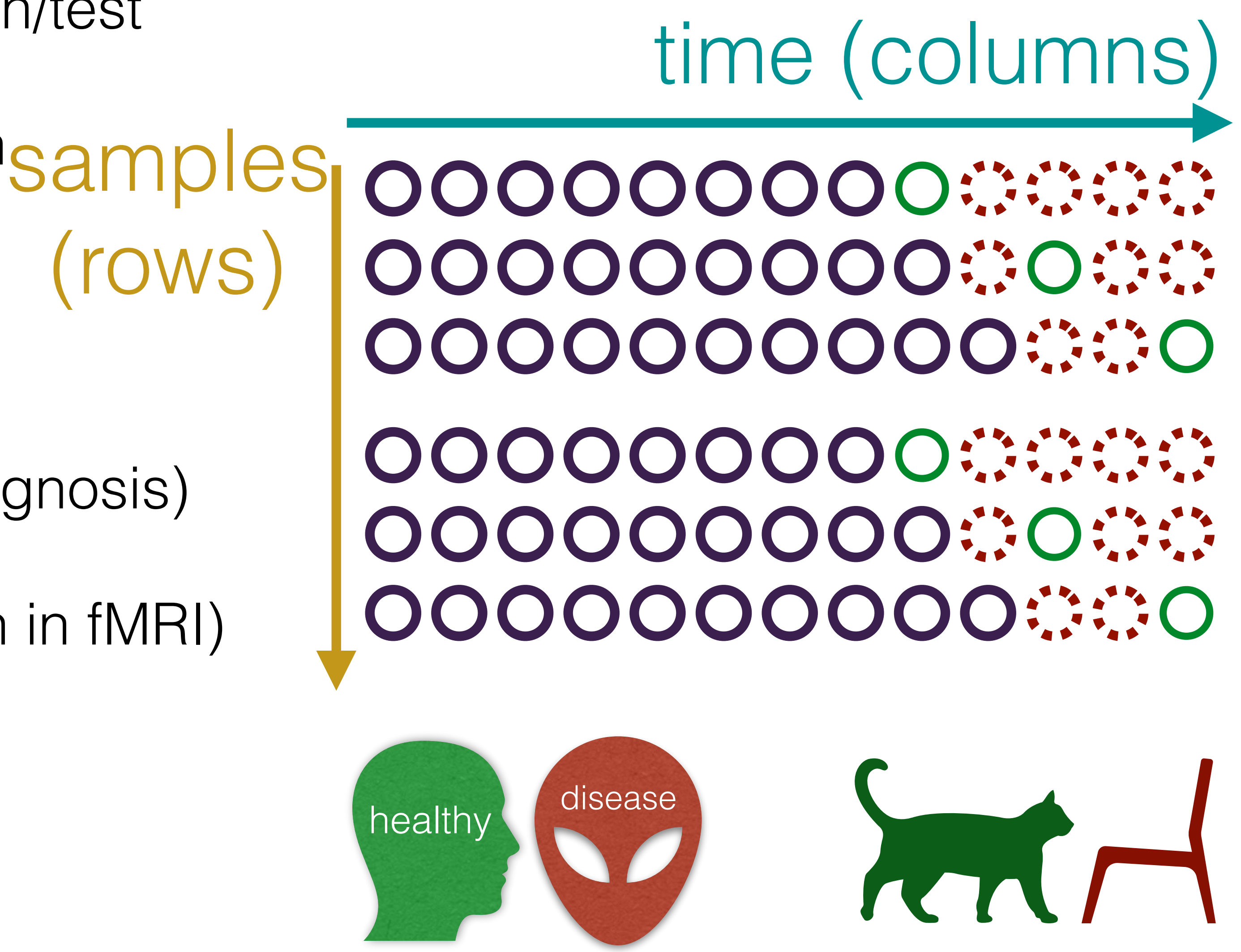
1. **How you split** the dataset into train/test
   - maximal independence between training and test sets is desired.
   - This split could be
     - over samples (e.g. indiv. diagnosis)
     - over time (for task prediction in fMRI)



Types of CV

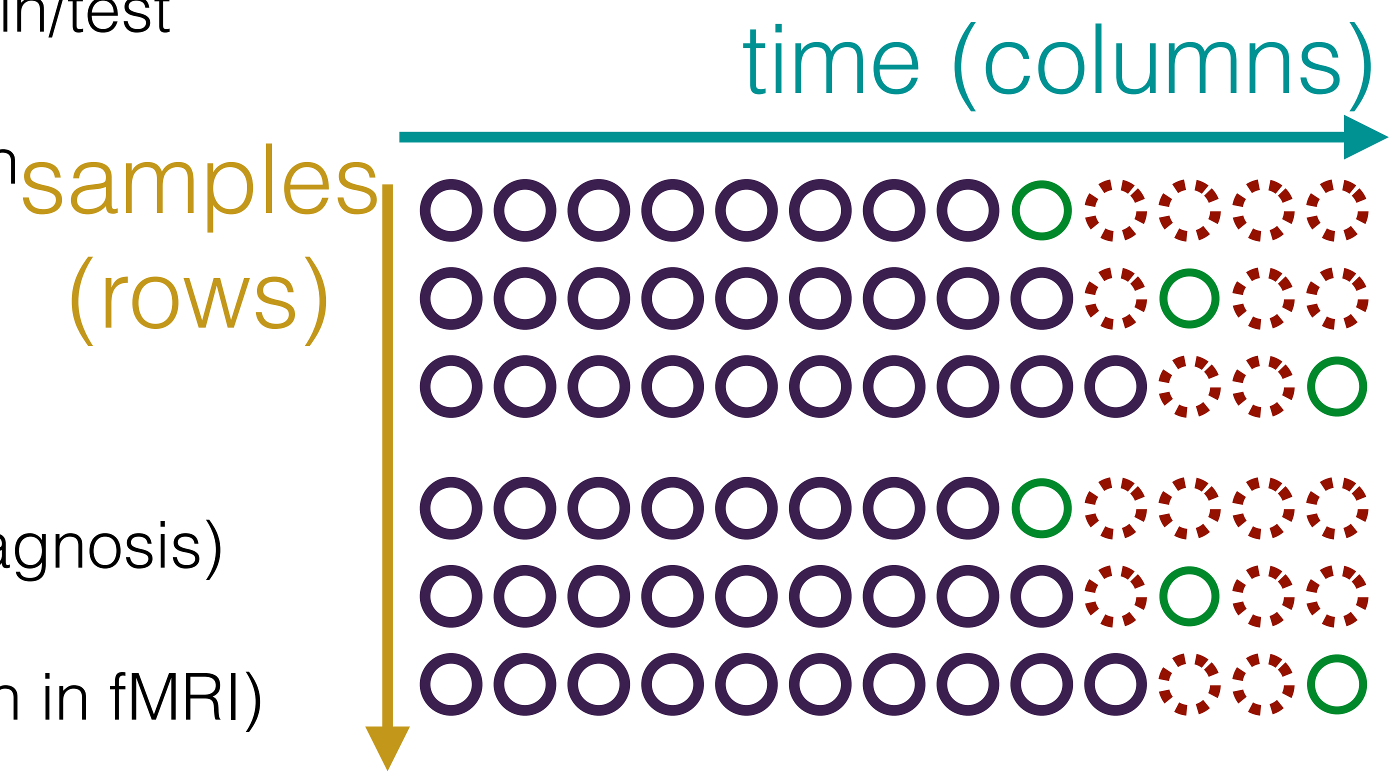
1. **How you split** the dataset into train/test

- maximal independence between training and test sets is desired.
- This split could be
  - over samples (e.g. indiv. diagnosis)
  - over time (for task prediction in fMRI)

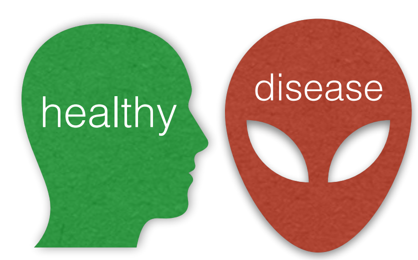


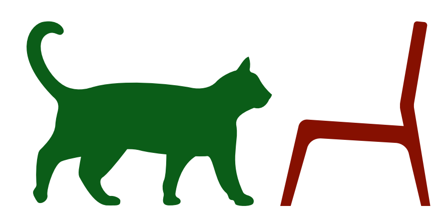
Types of CV

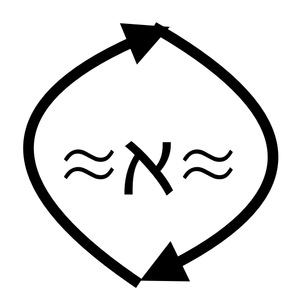
1. **How you split** the dataset into train/test
    - maximal independence between training and test sets is desired.
    - This split could be
        - over samples (e.g. indiv. diagnosis)
        - over time (for task prediction in fMRI)



2. **How often you repeat randomized splits?**
    - to expose classifier to full variability
    - As many as times as you can e.g. 100







Many other variations!

- k-fold, k = 2, 3, 5, 10, 20
- hold-out,
  train % = 50, 63.2, 75, 80, 90
- **stratified**
  - across train/test
  - across classes
- **inverted**:
  very small training, large testing
- leave one sample / pair / tuple
  condition / task / block out

1. 2-fold cross-validation (kf2)
2. 3-fold cross-validation (kf3)
3. 5-fold cross-validation (kf5)
4. 10-fold cross-validation (kf10)
5. 2 times repeated 5-fold (2xkf5)
6. 2 times repeated 10-fold (2xkf10)
7. 5, 10, and 20 times repeated bootstrap (5xboot, 1
8. 80/20 hold-out (80/20) — a training set of size
   data, and test set of 20%, with similar proportion
9. resubstitution (resub), training and testing in the
10. inverted 5-fold (invkf5): learning on a single fold,
11. 20/20 hold out (20/20) — training and test sets c
12. 5 times repeated 20/20 hold out (5x20/20)
13. 20/10 holdout (20/10)
14. 10/10 hold out (10/10)
15. 5 times repeated 10/10 hold out (5x10/10)

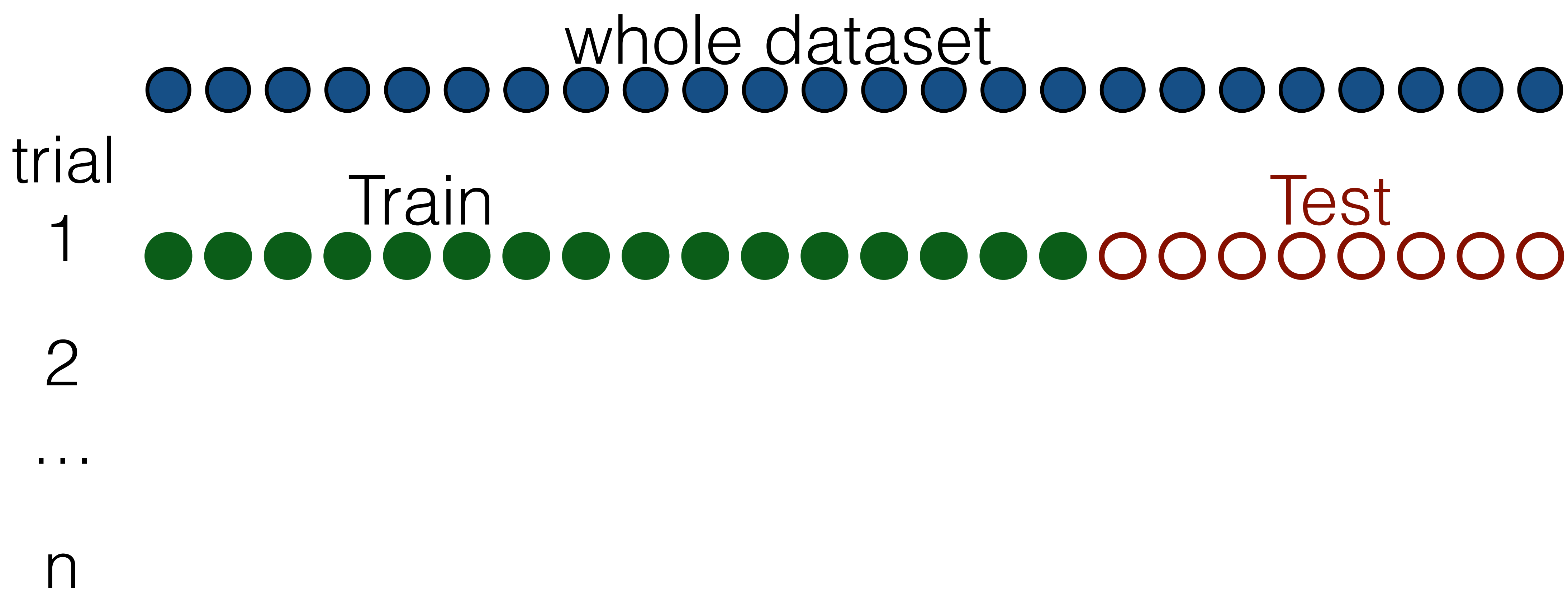
Wainer, J., & Cawley, G. (2017). Empirical Evaluation of Resampling Procedures for Optimising SVM Hyperparameters. Journal of Machine Learning Research, 18(15), 1–35.

Hold-out CV

Set aside a fixed percentage (e.g. 30%) for testing

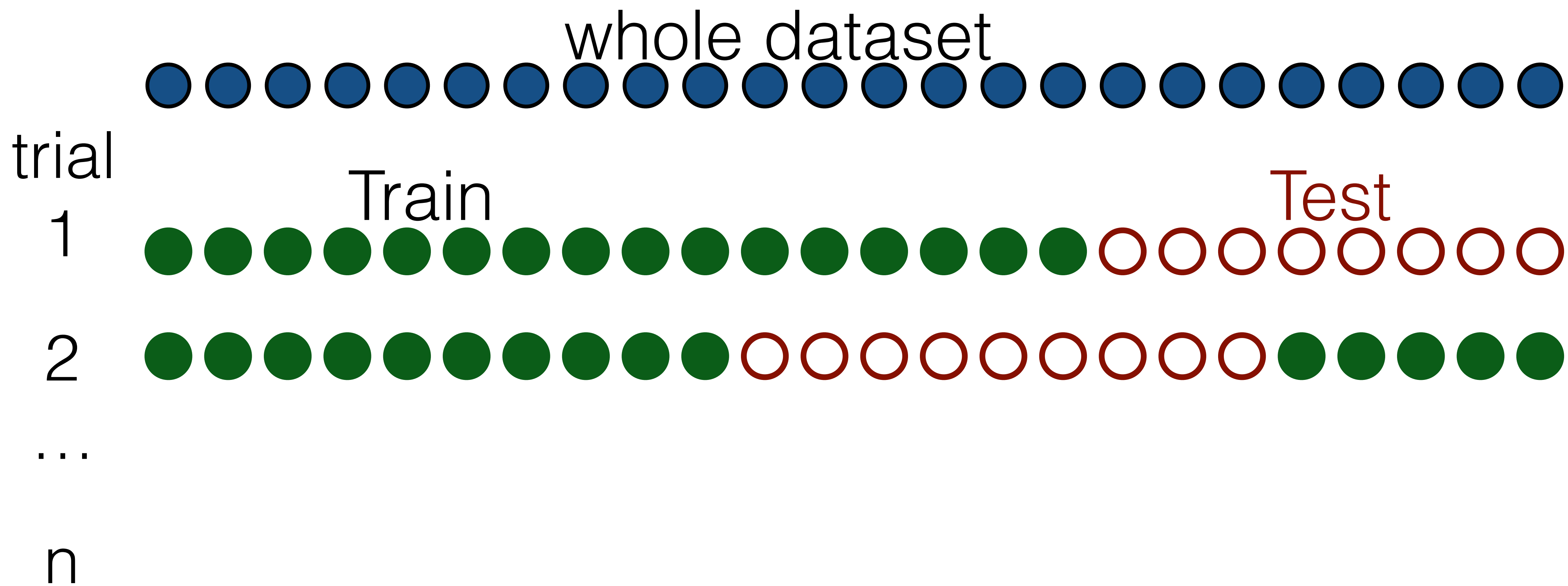
Hold-out CV

Set aside a fixed percentage (e.g. 30%) for testing



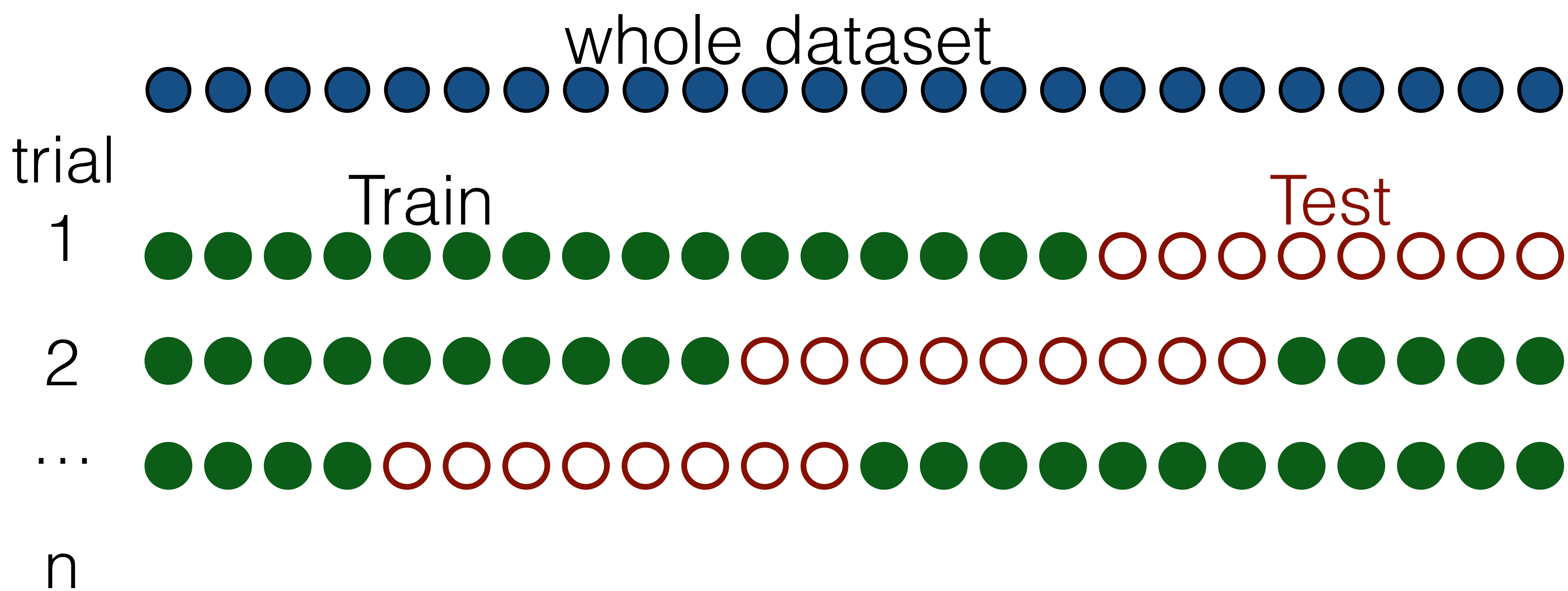
Hold-out CV

Set aside a fixed percentage (e.g. 30%) for testing



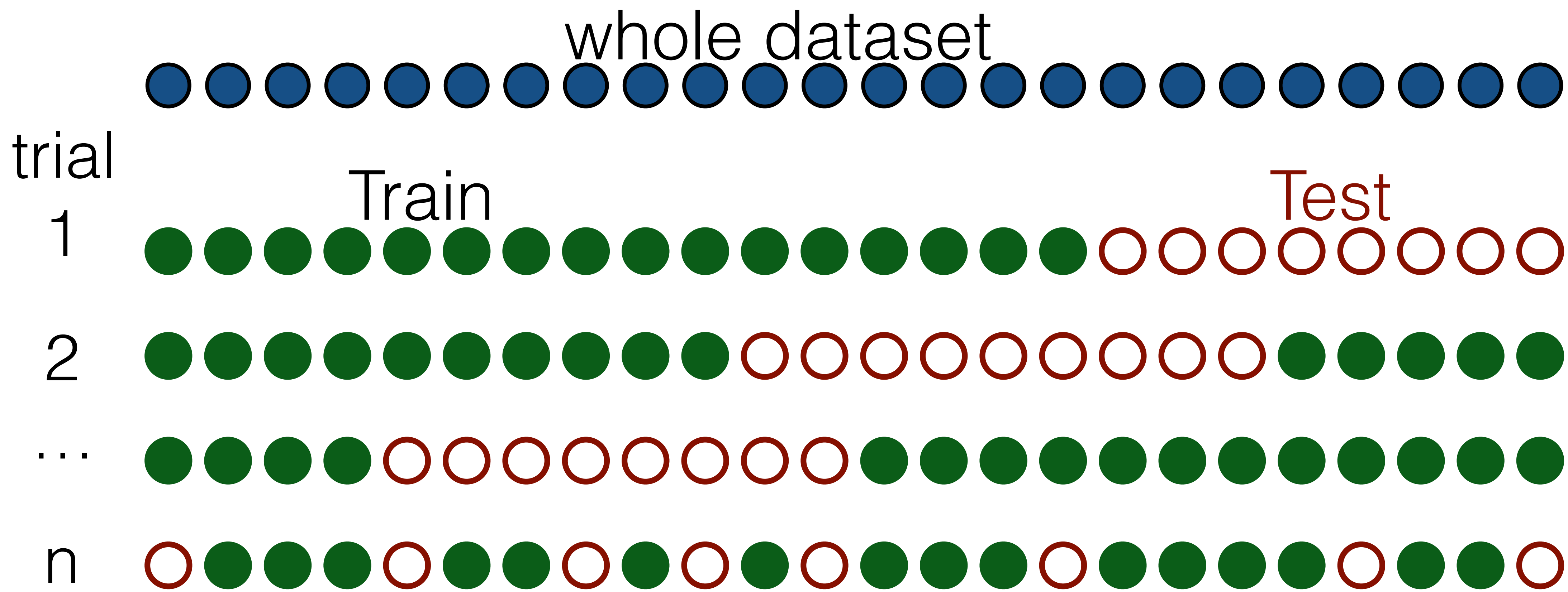
Hold-out CV

Set aside a fixed percentage (e.g. 30%) for testing



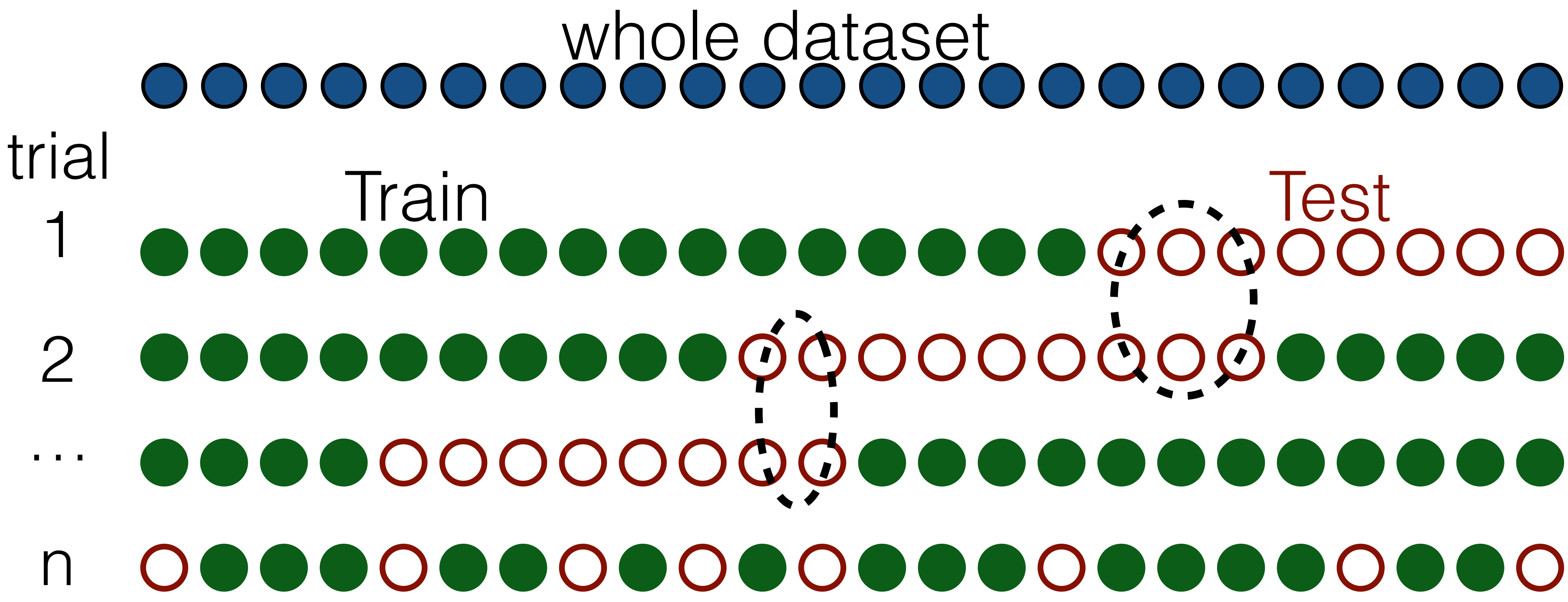
Hold-out CV

Set aside a fixed percentage (e.g. 30%) for testing



Hold-out CV

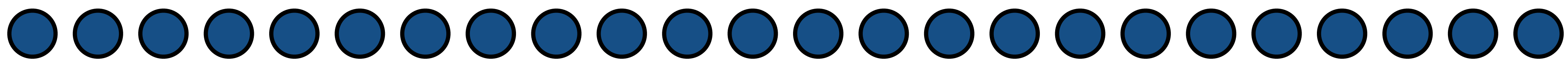
Set aside a fixed percentage (e.g. 30%) for testing



Note: there could be **overlap** among the test sets!

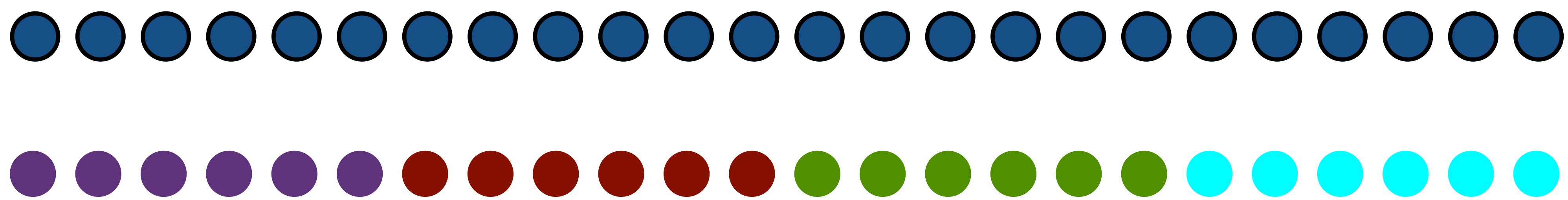
i.e. test sets in different iterations could have common samples

K-fold CV



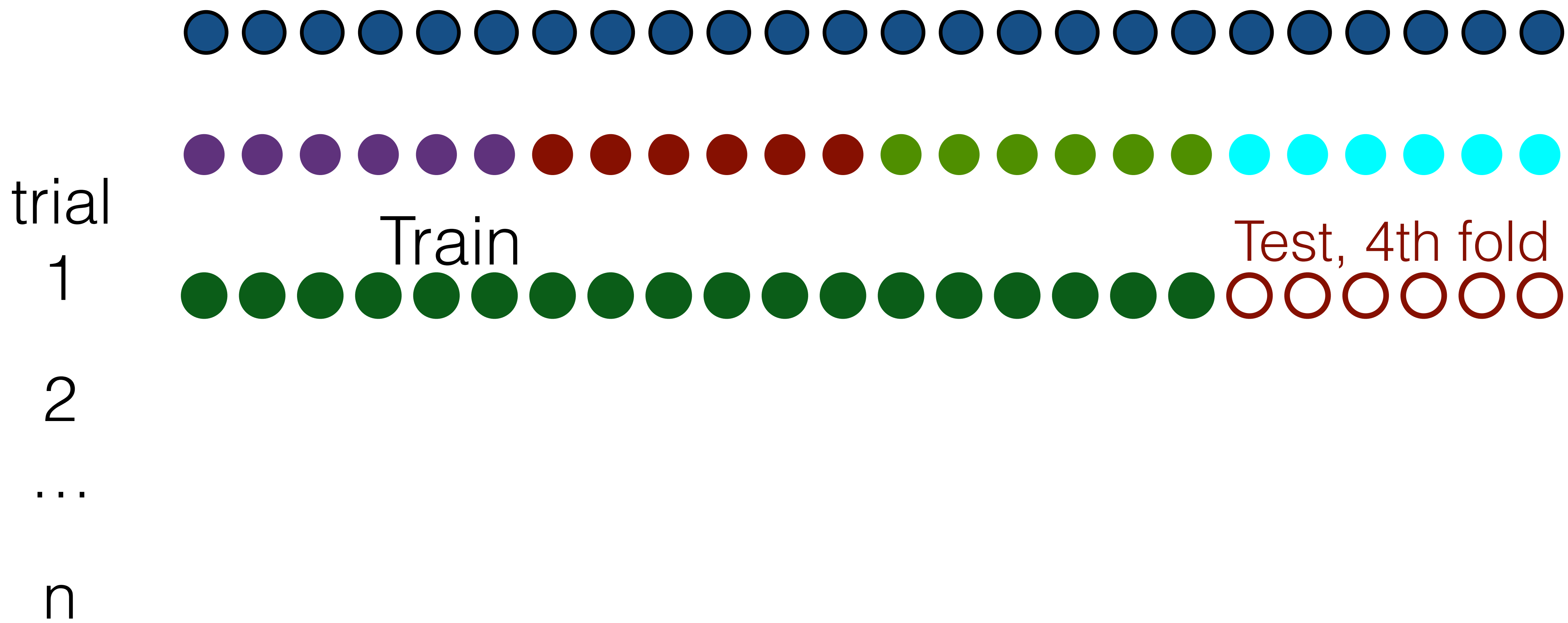
Note: different folds won’t be contiguous.

K-fold CV



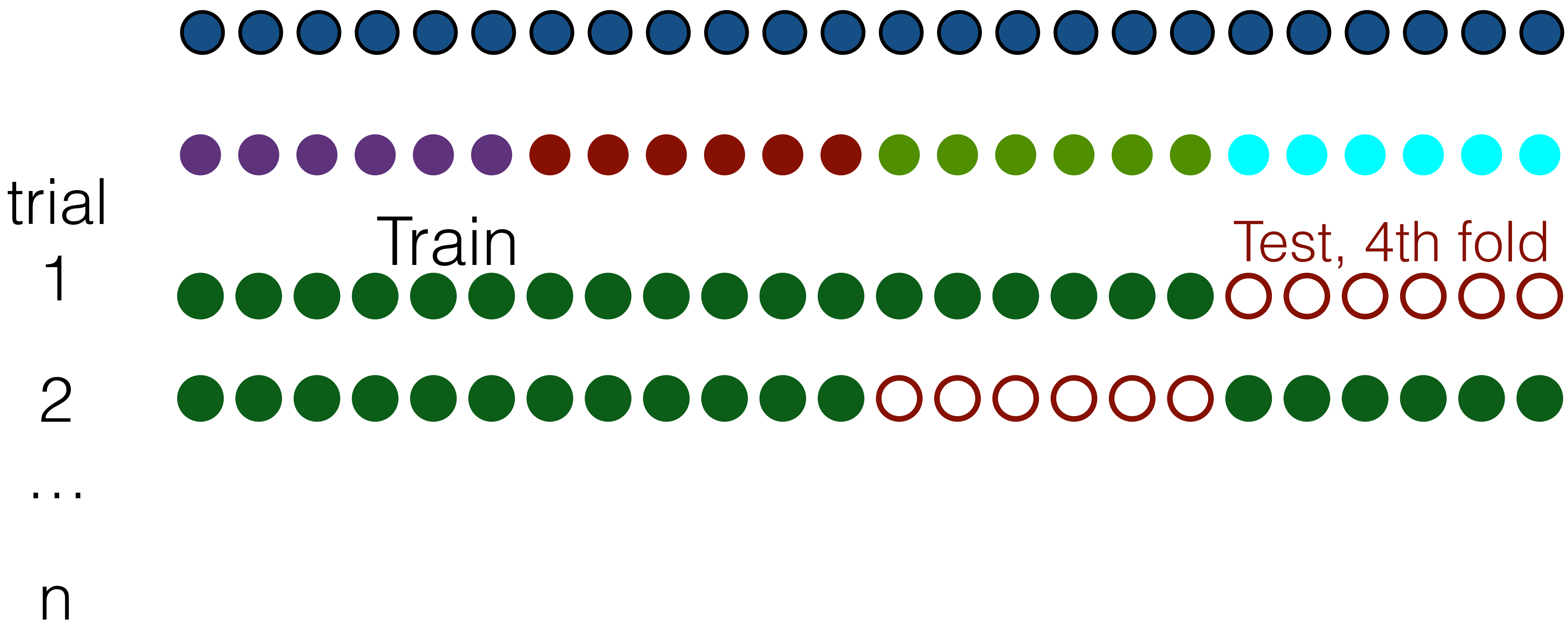
Note: different folds won't be contiguous.

K-fold CV



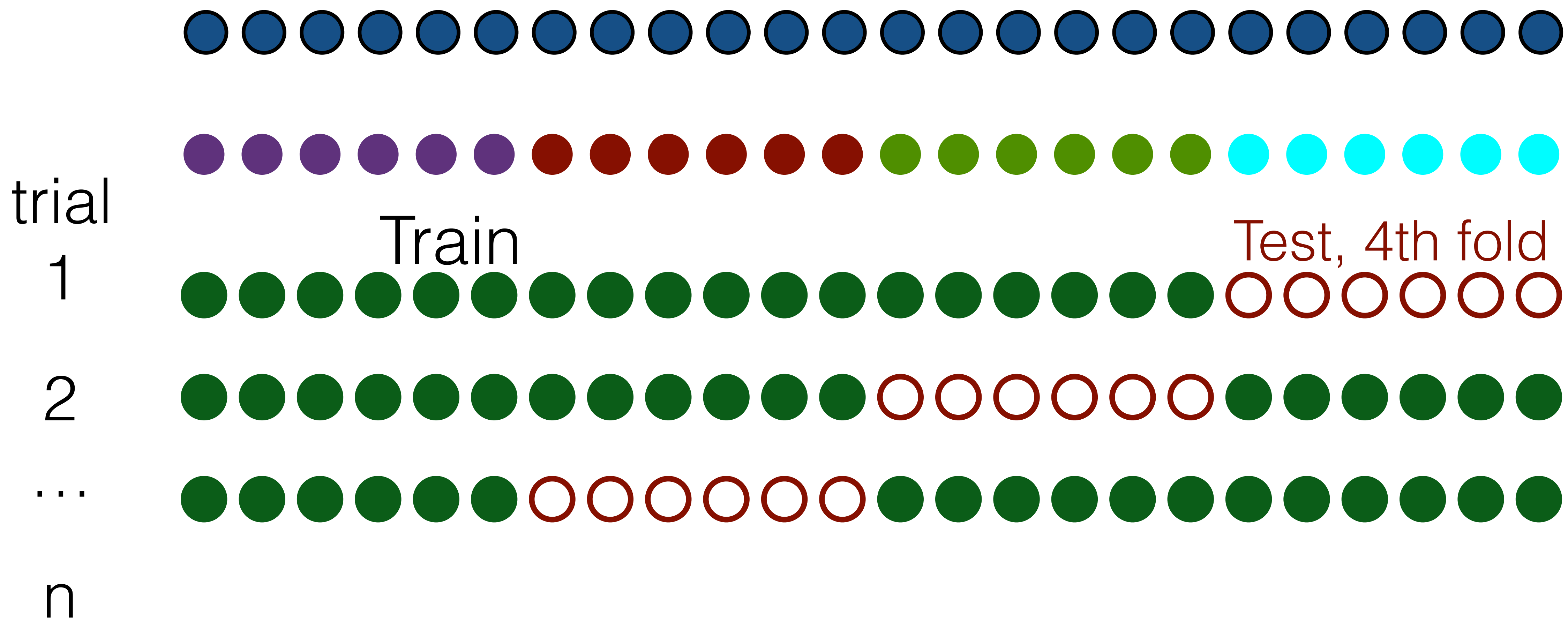
Note: different folds won't be contiguous.

K-fold CV



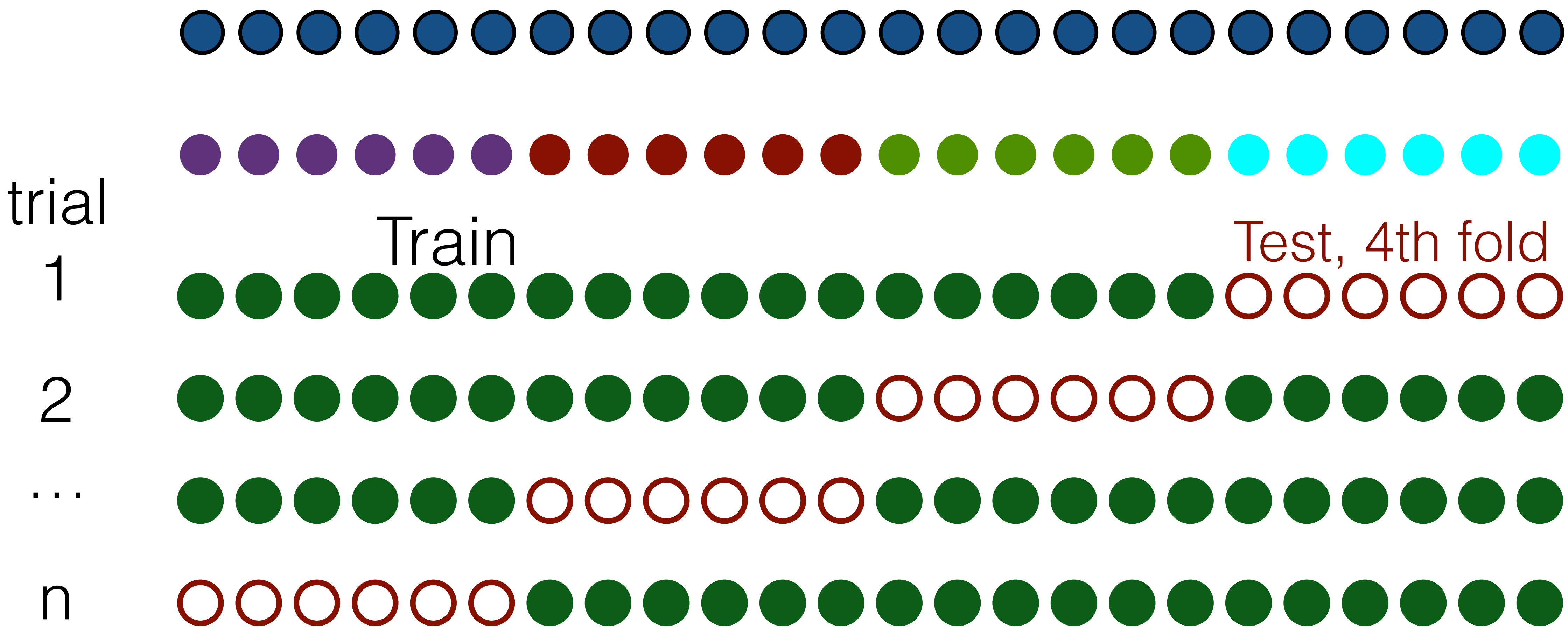
Note: different folds won't be contiguous.

K-fold CV



Note: different folds won't be contiguous.

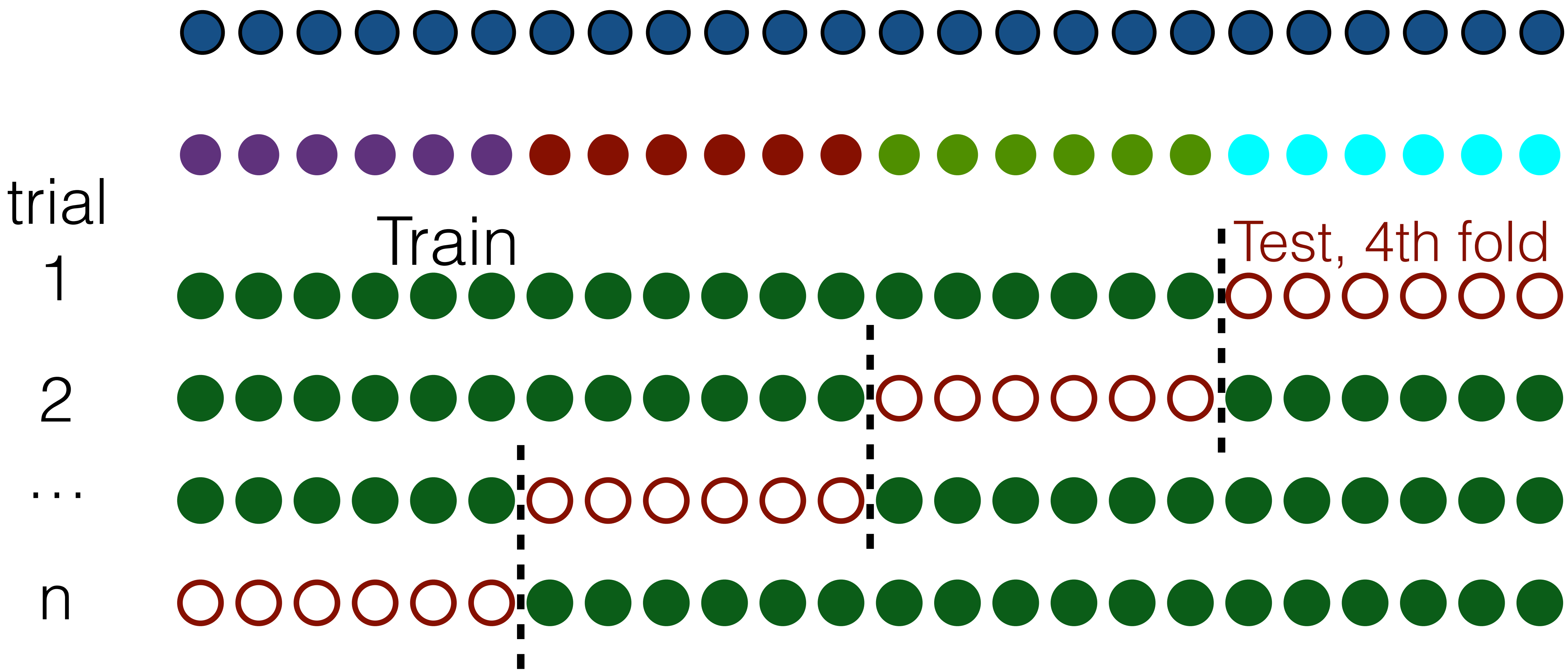
K-fold CV



Note: different folds won't be contiguous.

K-fold CV

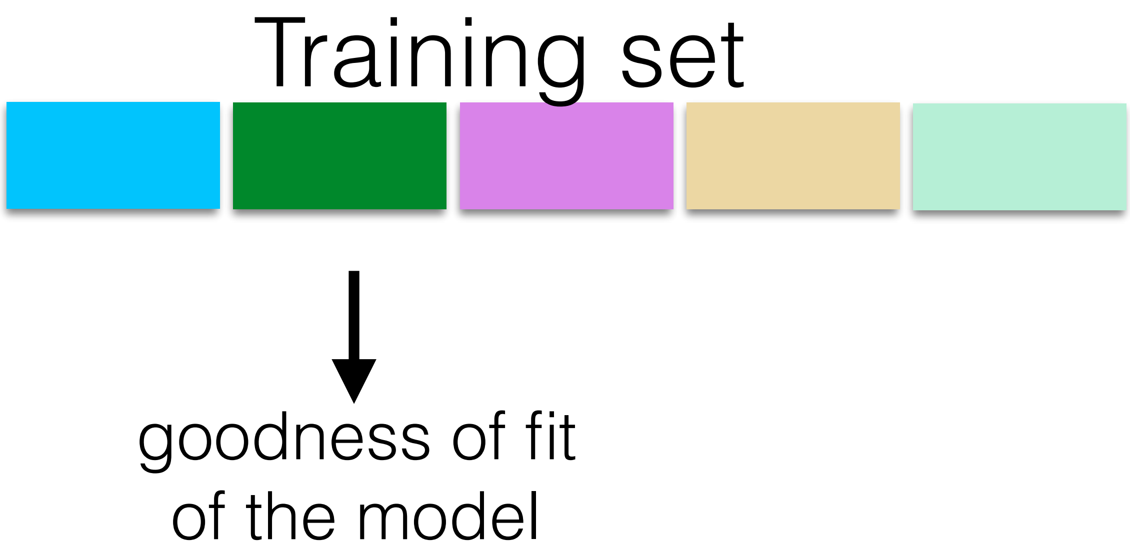
Test sets in different trials are indeed mutually disjoint



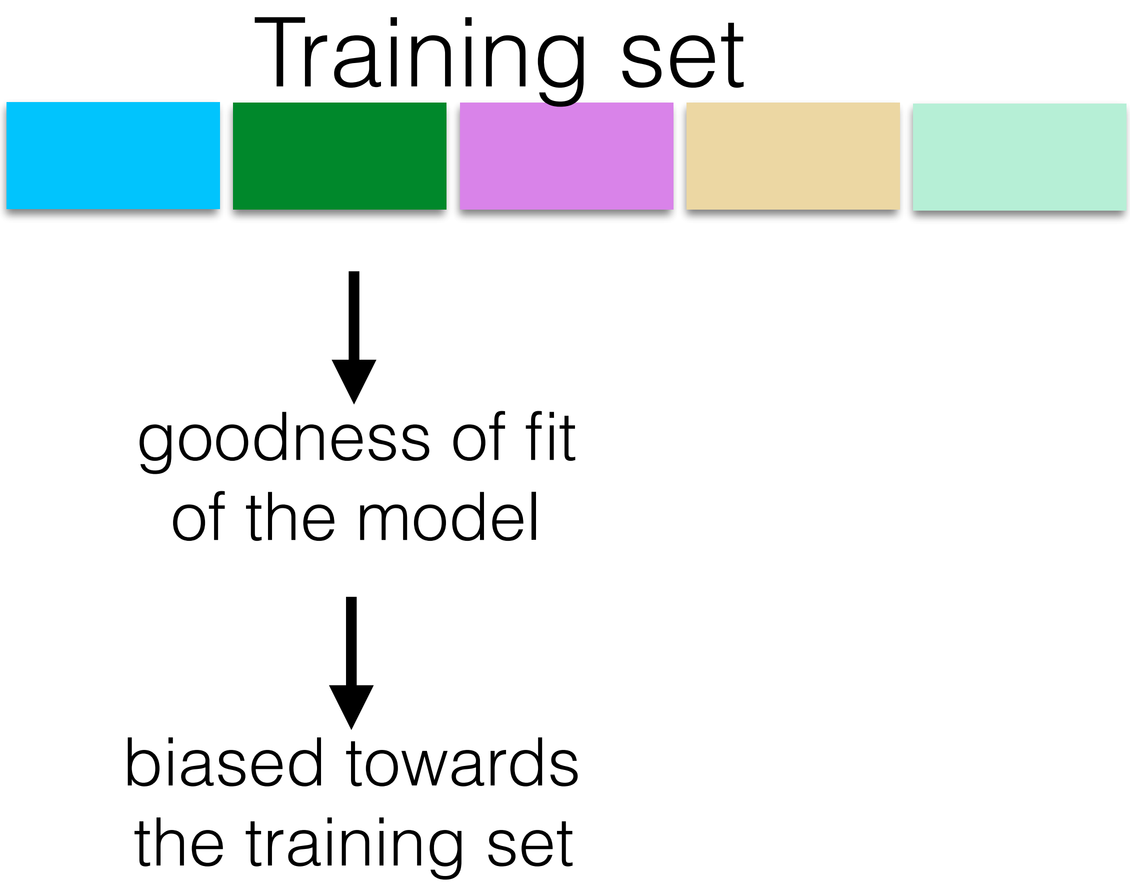
Note: different folds won't be contiguous.

Validation set

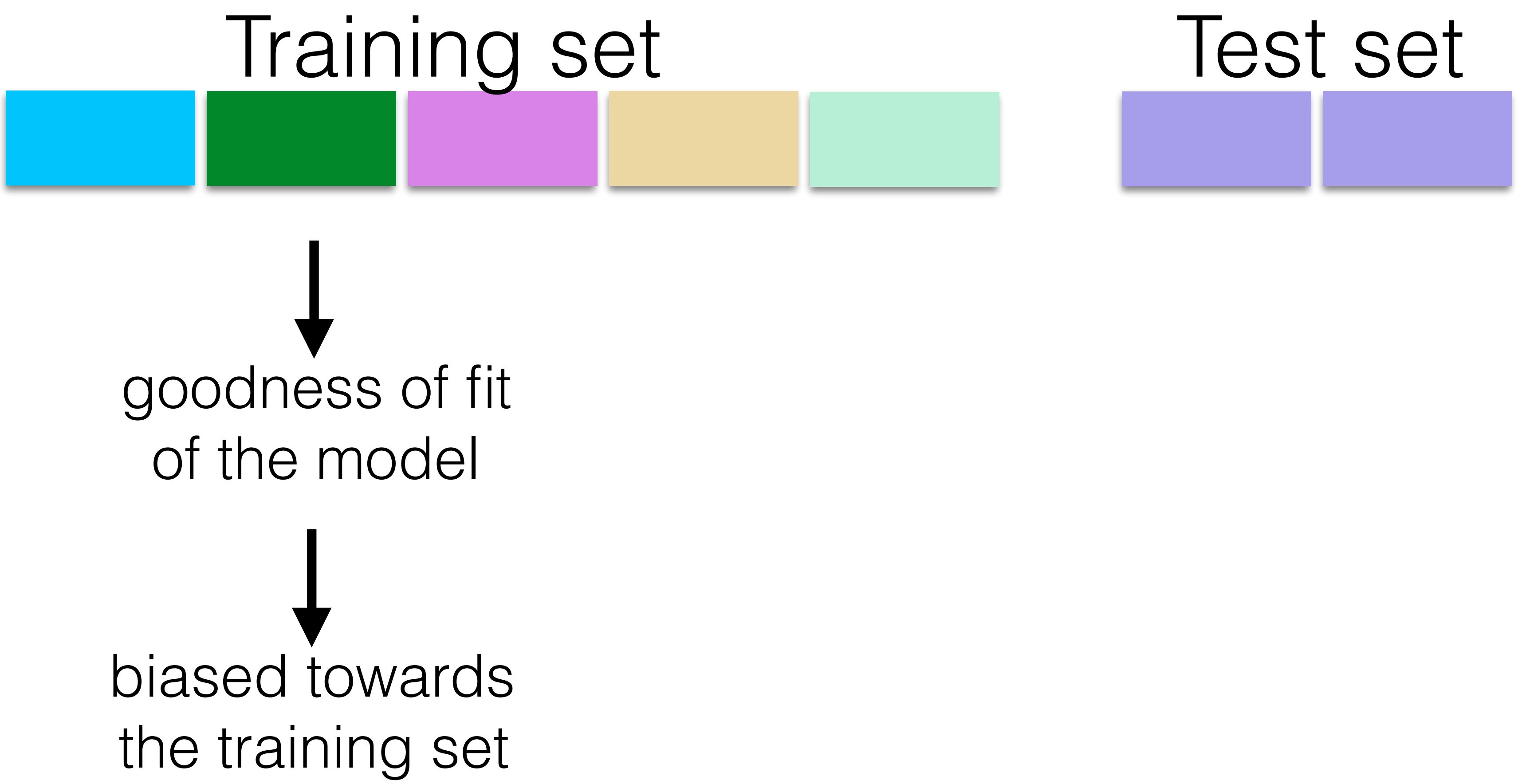
Validation set



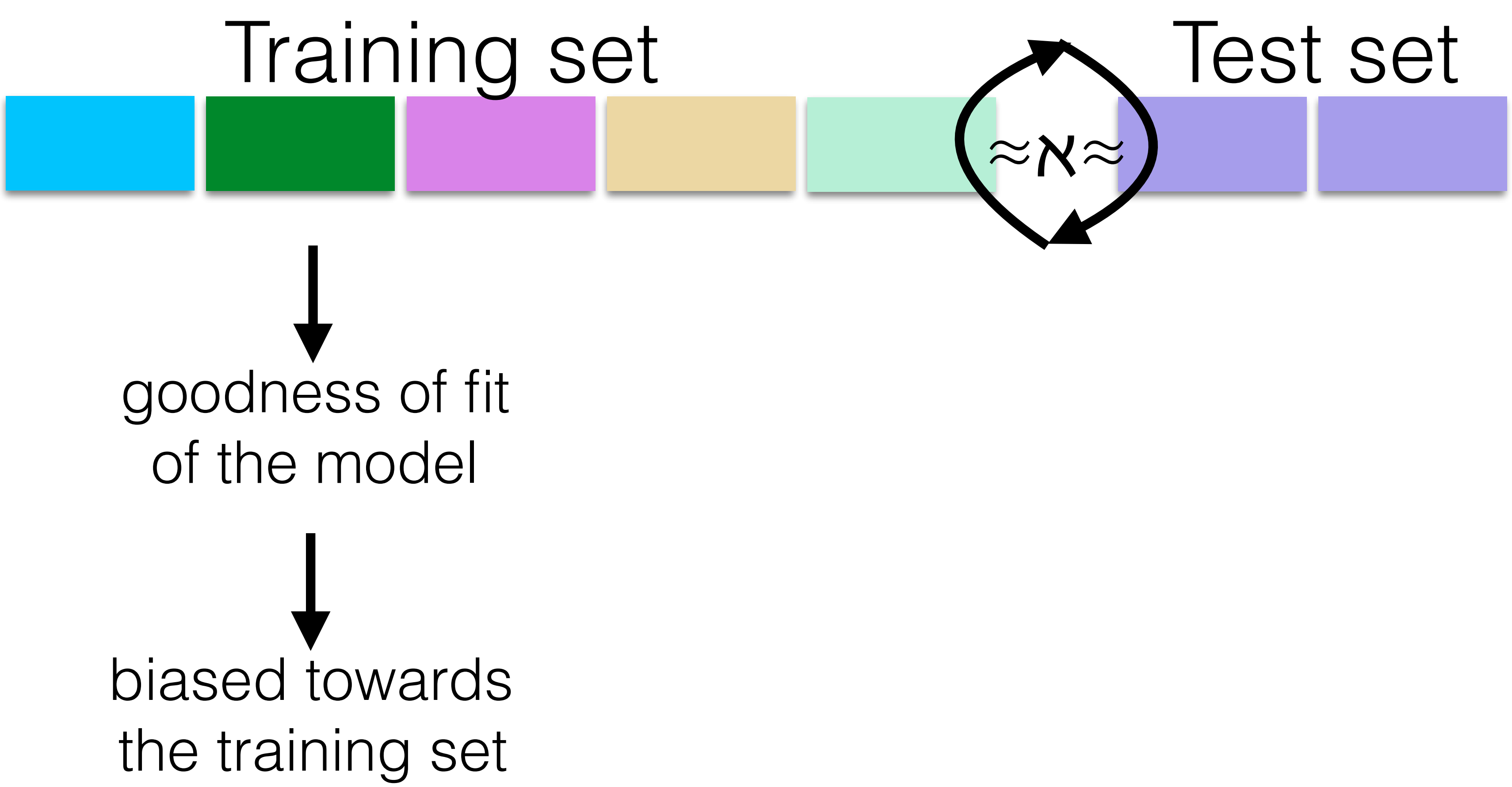
Validation set



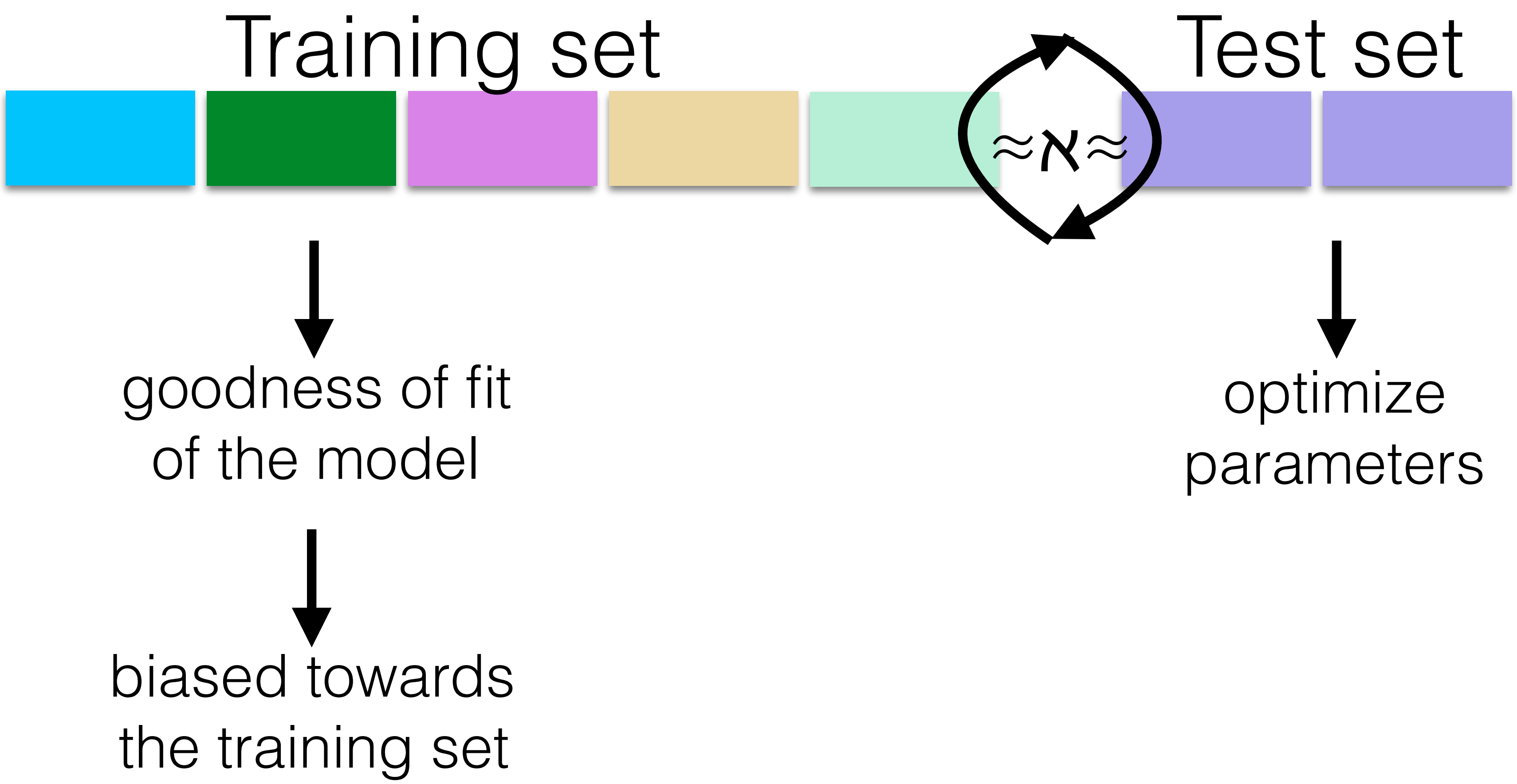
Validation set



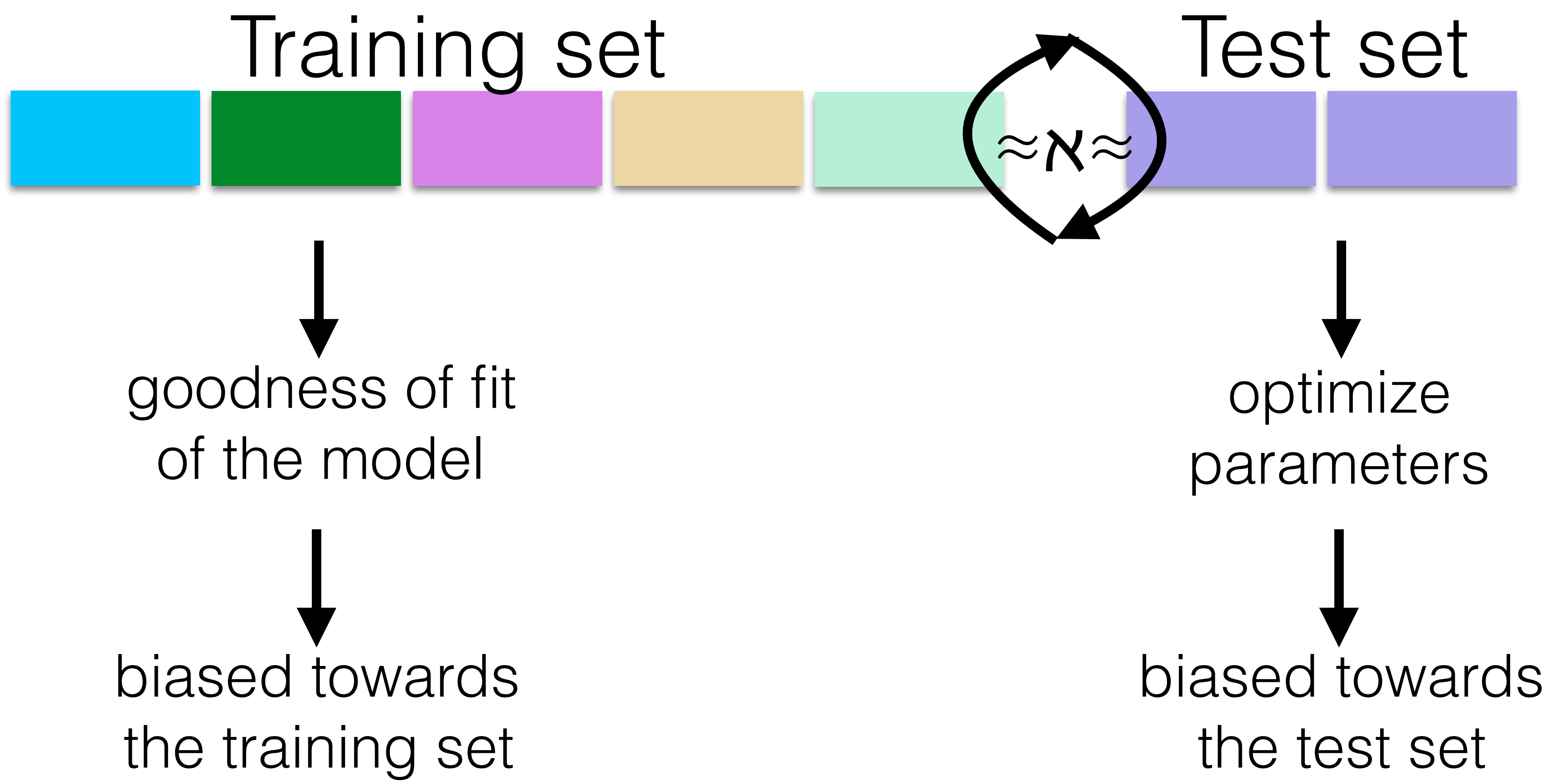
Validation set



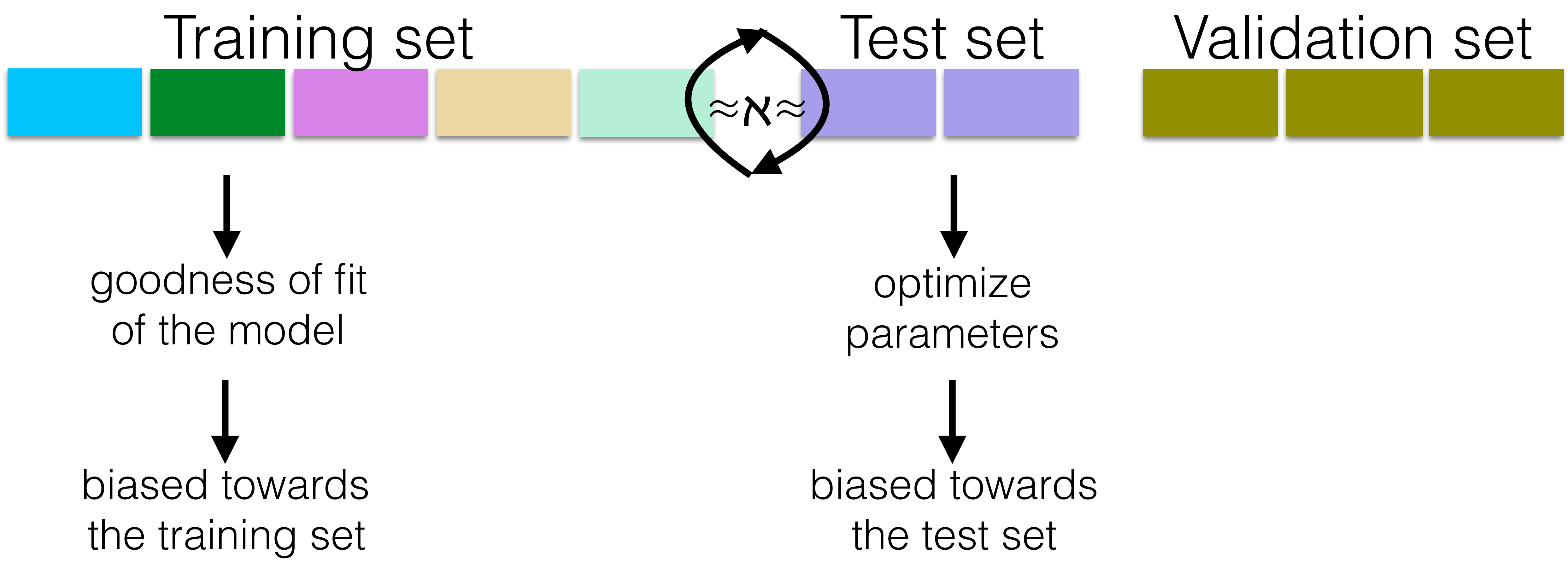
Validation set



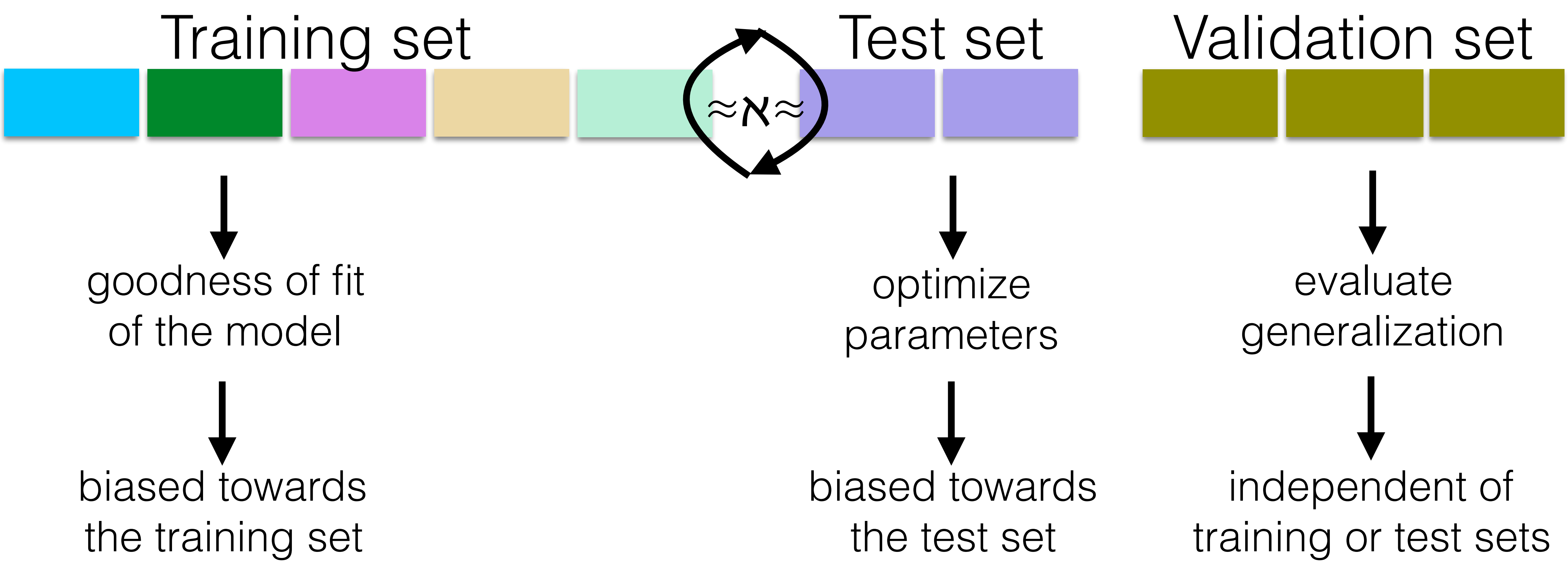
Validation set



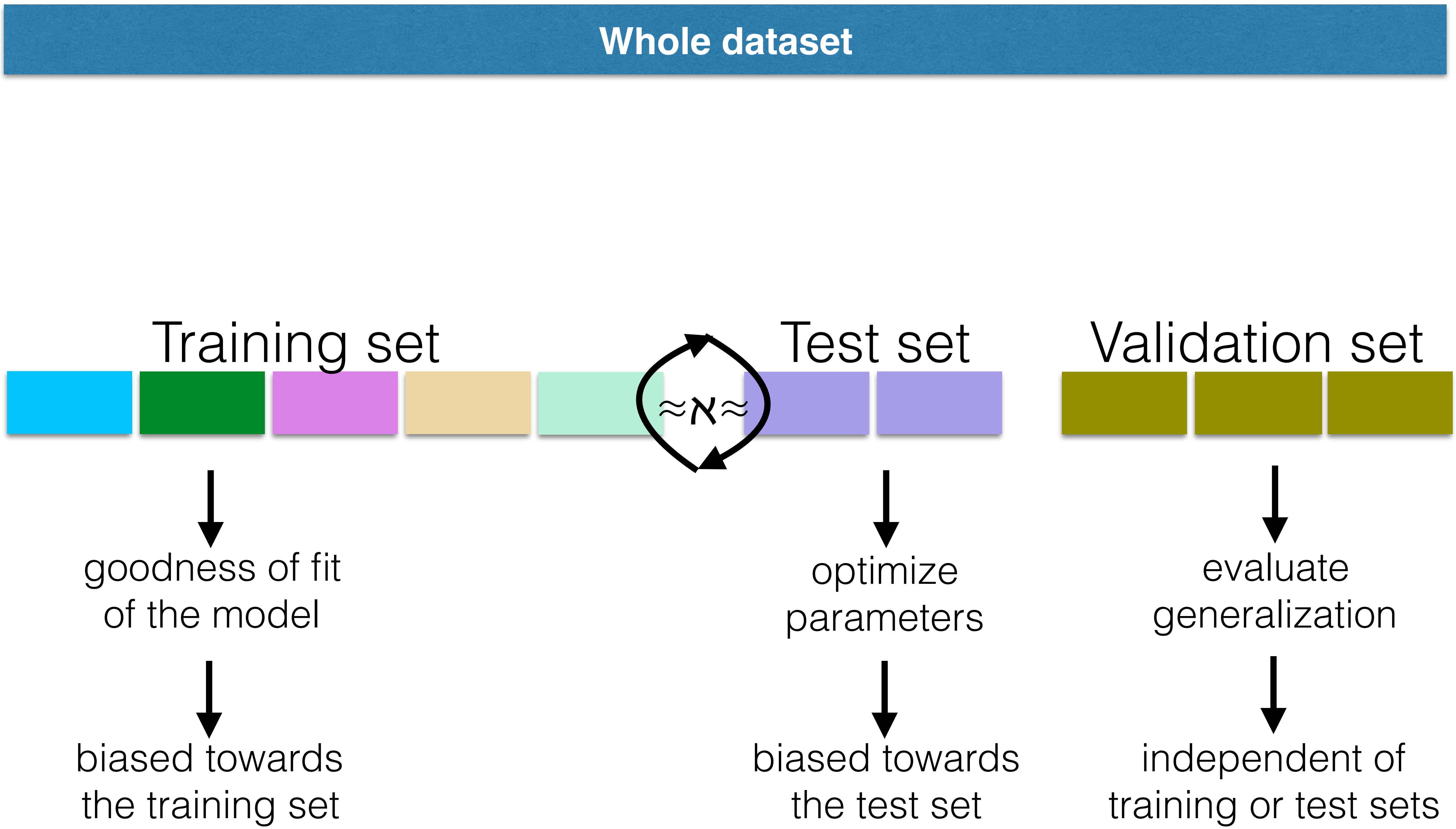
Validation set



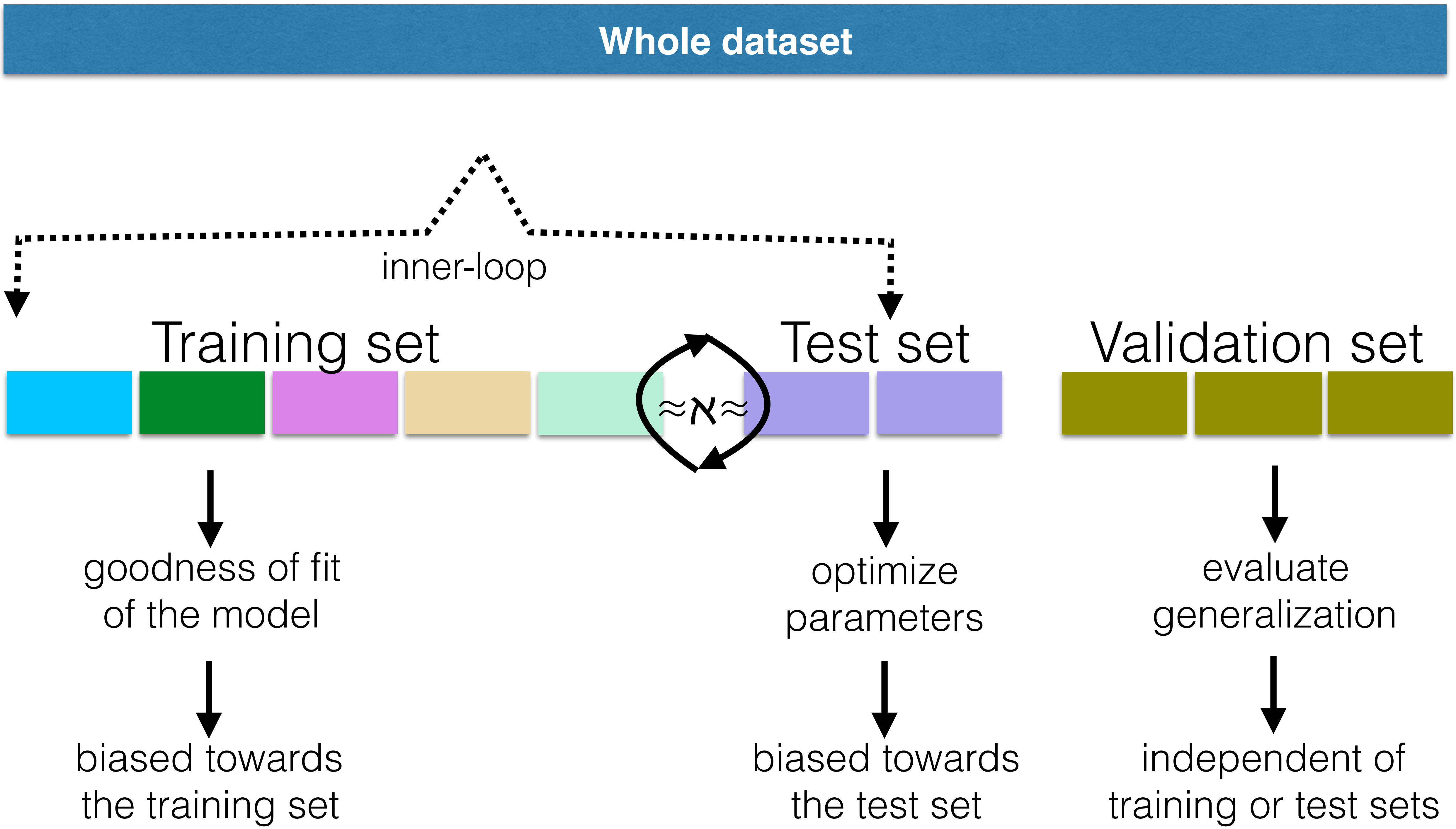
Validation set



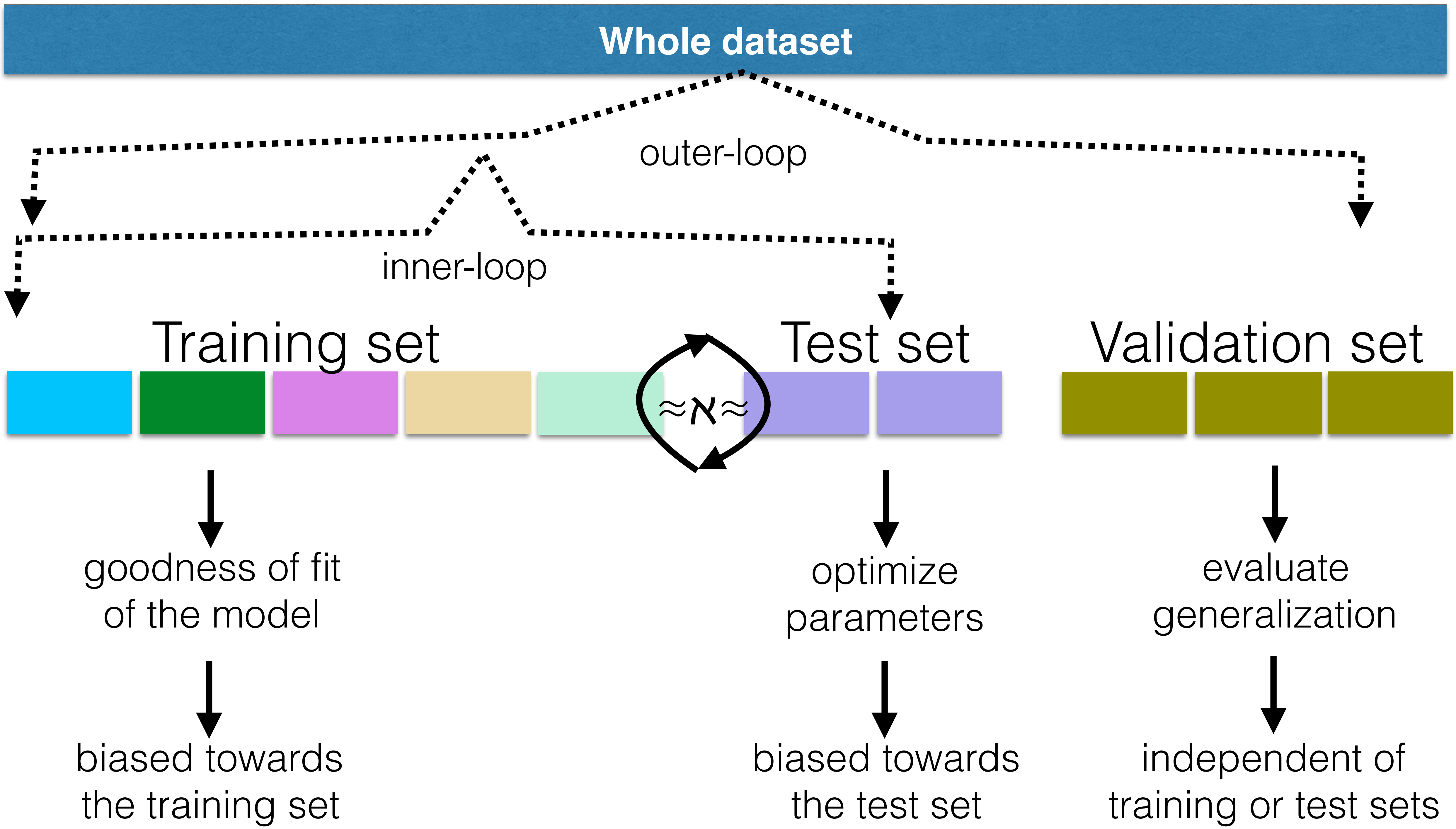
Validation set



Validation set



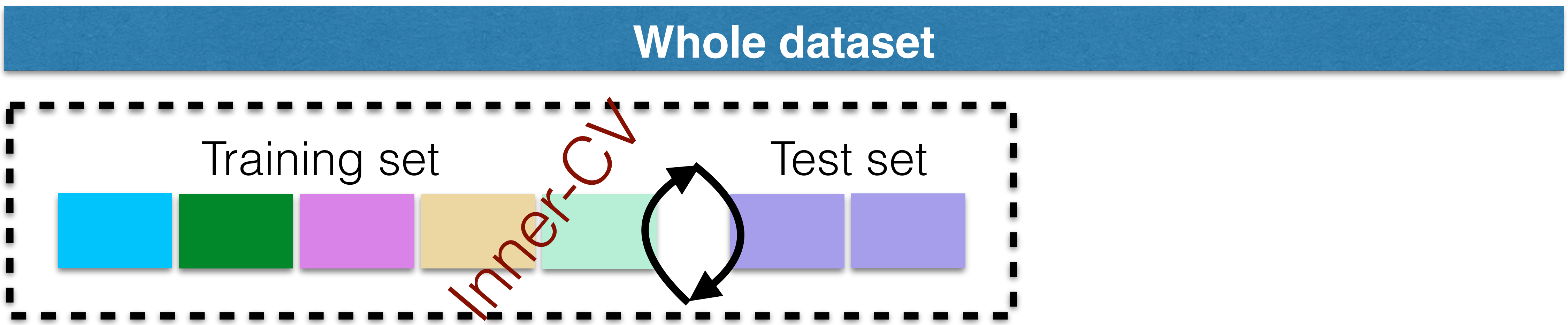
Validation set



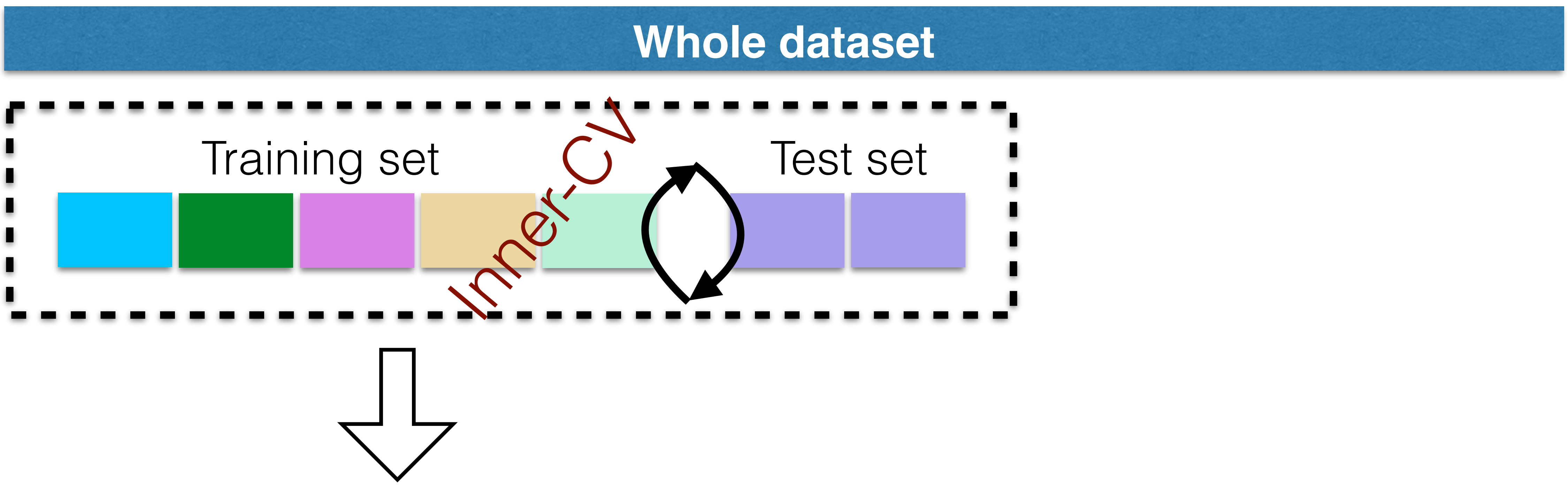
Measuring bias in CV measurements

Whole dataset

Measuring bias in CV measurements

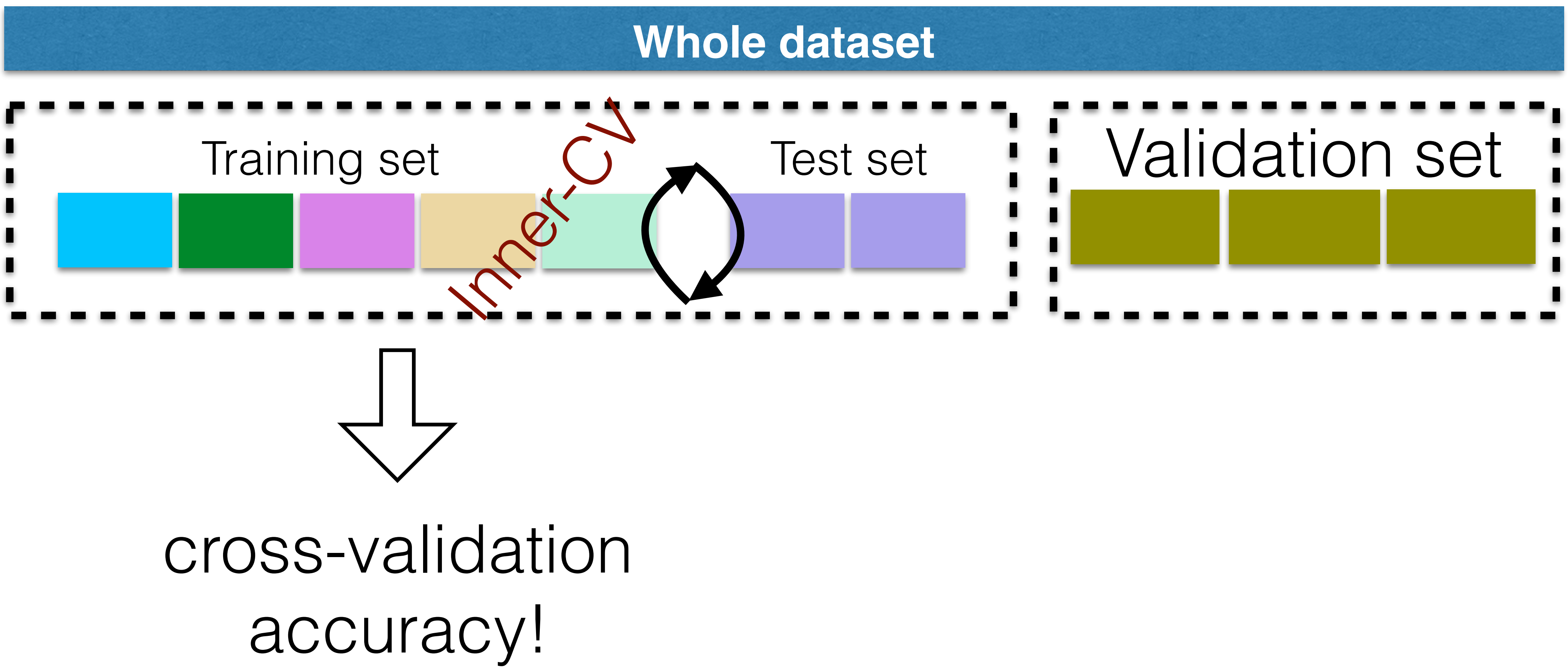


Measuring bias in CV measurements

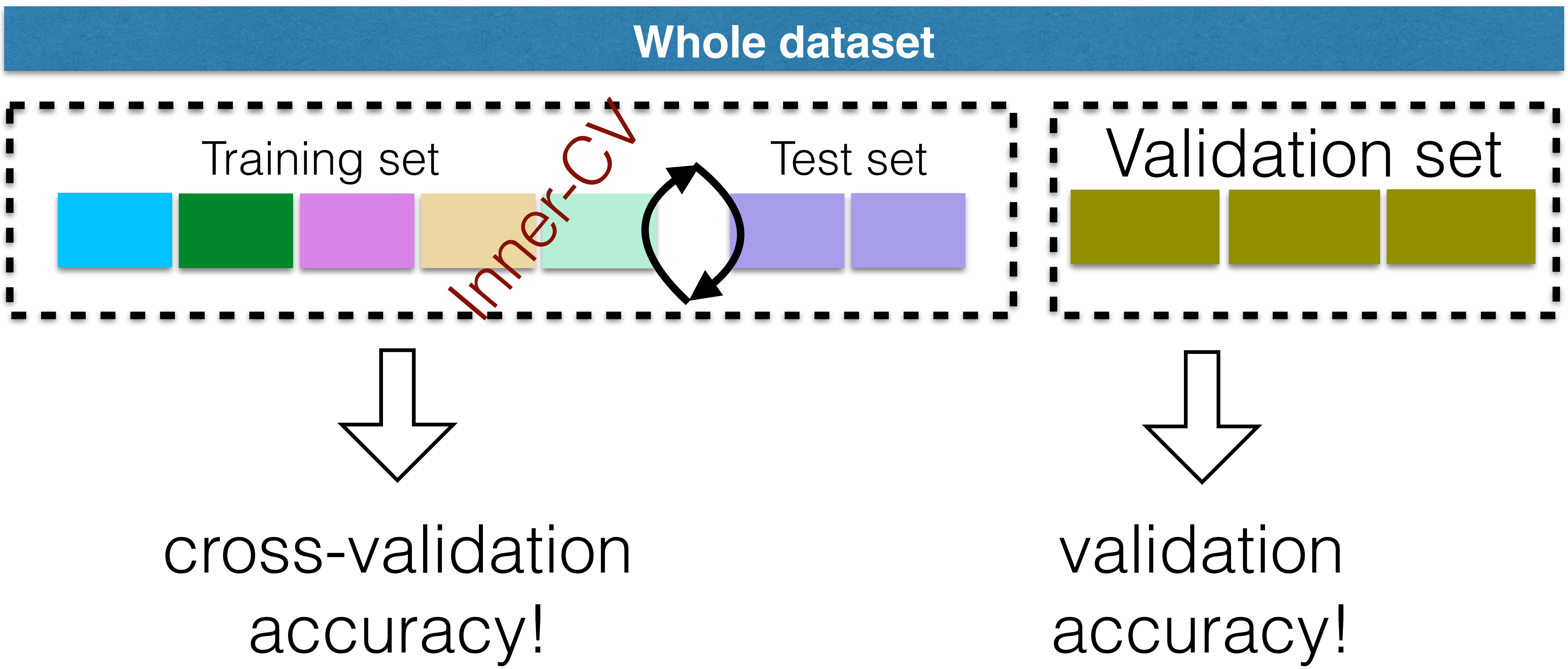


cross-validation accuracy!

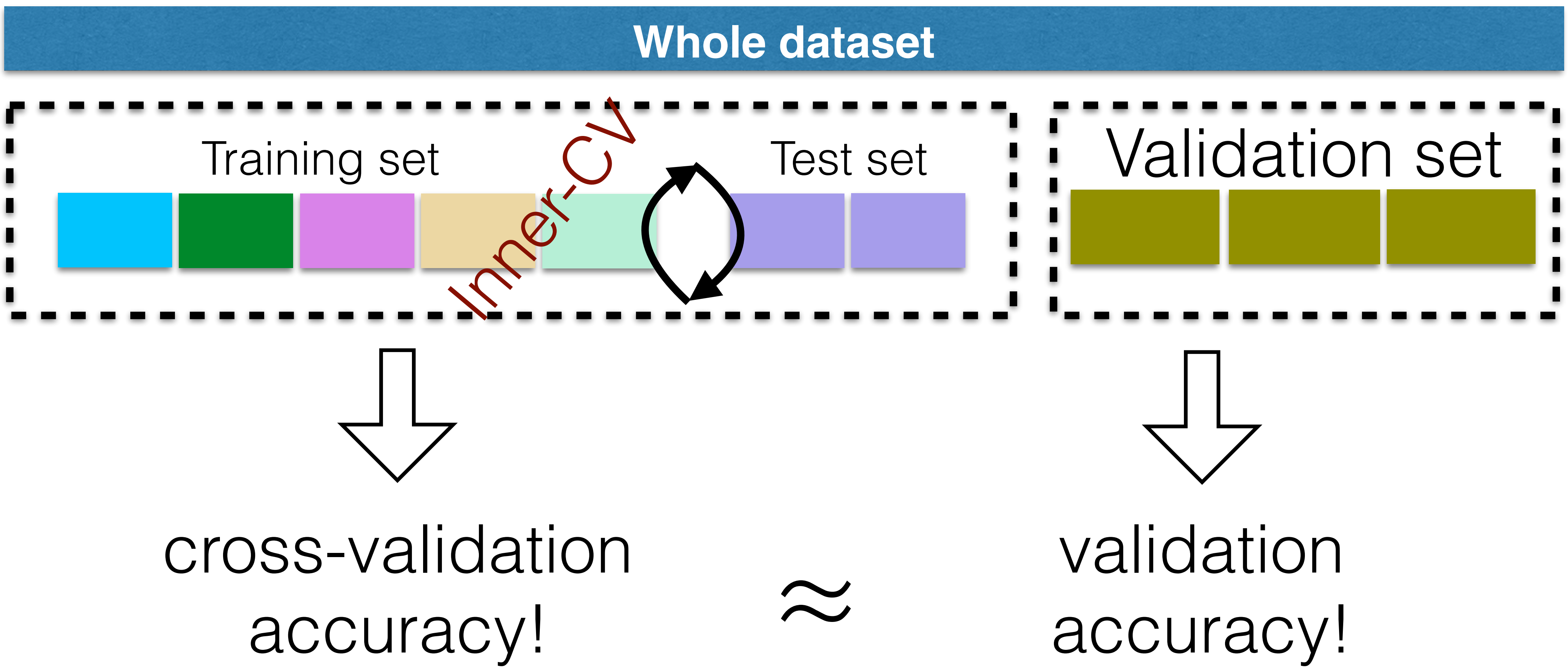
Measuring bias in CV measurements



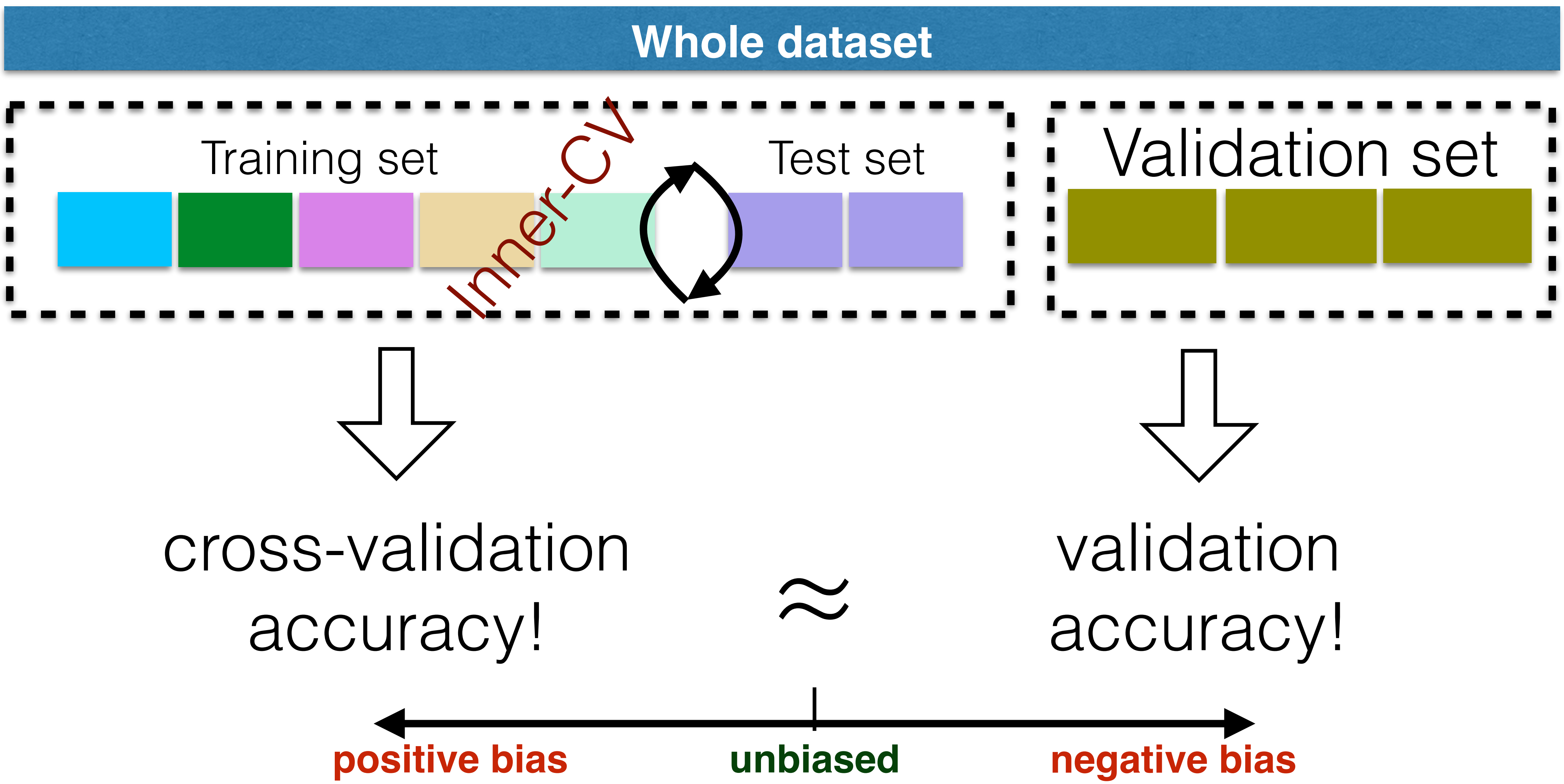
Measuring bias in CV measurements



Measuring bias in CV measurements



Measuring bias in CV measurements



Intra-subject datasets: Haxby

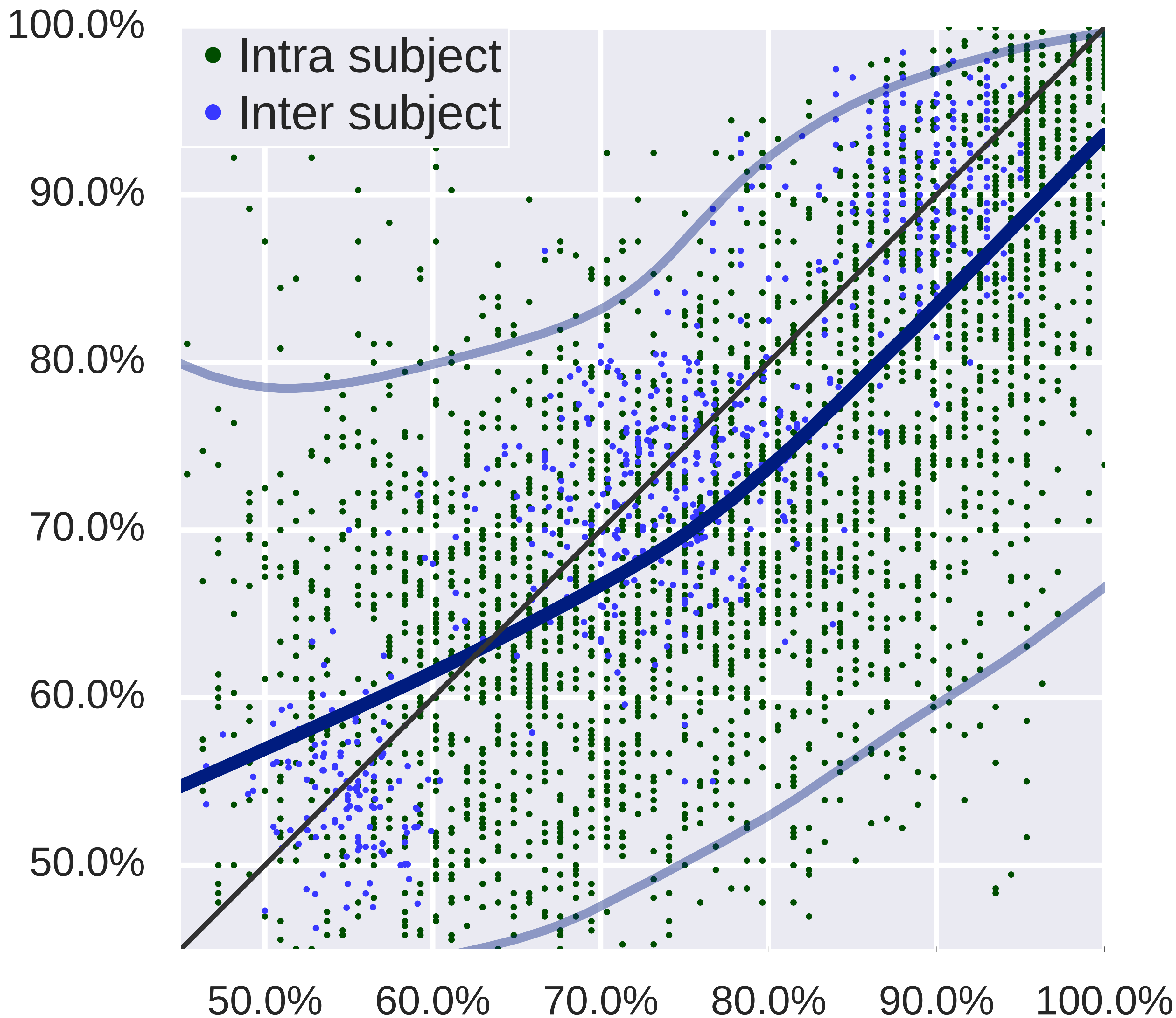
Task	# samples	#blocks	mean accuracy of SVM *l2*	mean accuracy of SVM *l1*
bottle / scramble	209	12 secs	75%	86%
cat / bottle			62%	69%
cat / chair			69%	80%
cat / face			65%	72%
cat / house			86%	95%
cat / scramble			83%	92%
chair / scramble			77%	91%
chair / shoe			63%	70%
face / house			88%	96%
face / scissors			72%	83%
scissors / scramble			73%	87%
scissors / shoe			60%	64%
shoe / bottle			62%	69%
shoe / cat			72%	85%
shoe / scramble			78%	88%

Inter-subject fMRI datasets

Dataset	Description	# samples	# blocks (sess./subj.)	Task	mean accuracy SVM ℓ_2	mean accuracy SVM ℓ_1
Duncan [9]	fMRI, across subjects	196	49 subj.	consonant / scramble	92%	88%
				consonant / word	92%	89%
				objects / consonant	90%	88%
				objects / scramble	91%	88%
				objects / words	74%	71%
				words / scramble	91%	89%
Wager [53]	fMRI across subjects	390	34 subj.	negative cue / neutral cue	55%	55%
				negative rating / neutral rating	54%	53%
				negative stim / neutral stim	77%	73%
Cohen (ds009)	fMRI across subjects	80	24 subj.	successful / unsuccessful stop	67%	63%
Moran [34]	fMRI across subjects	138	36 subj.	false picture / false belief	72%	71%
Henson [19]	fMRI across subjects	286	16 subj.	famous / scramble	77%	74%
				famous / unfamiliar	54%	55%
				scramble / unfamiliar	73%	70%
Knops [23]	fMRI, across subjects	114	19 subj.	right field / left field	79%	73%

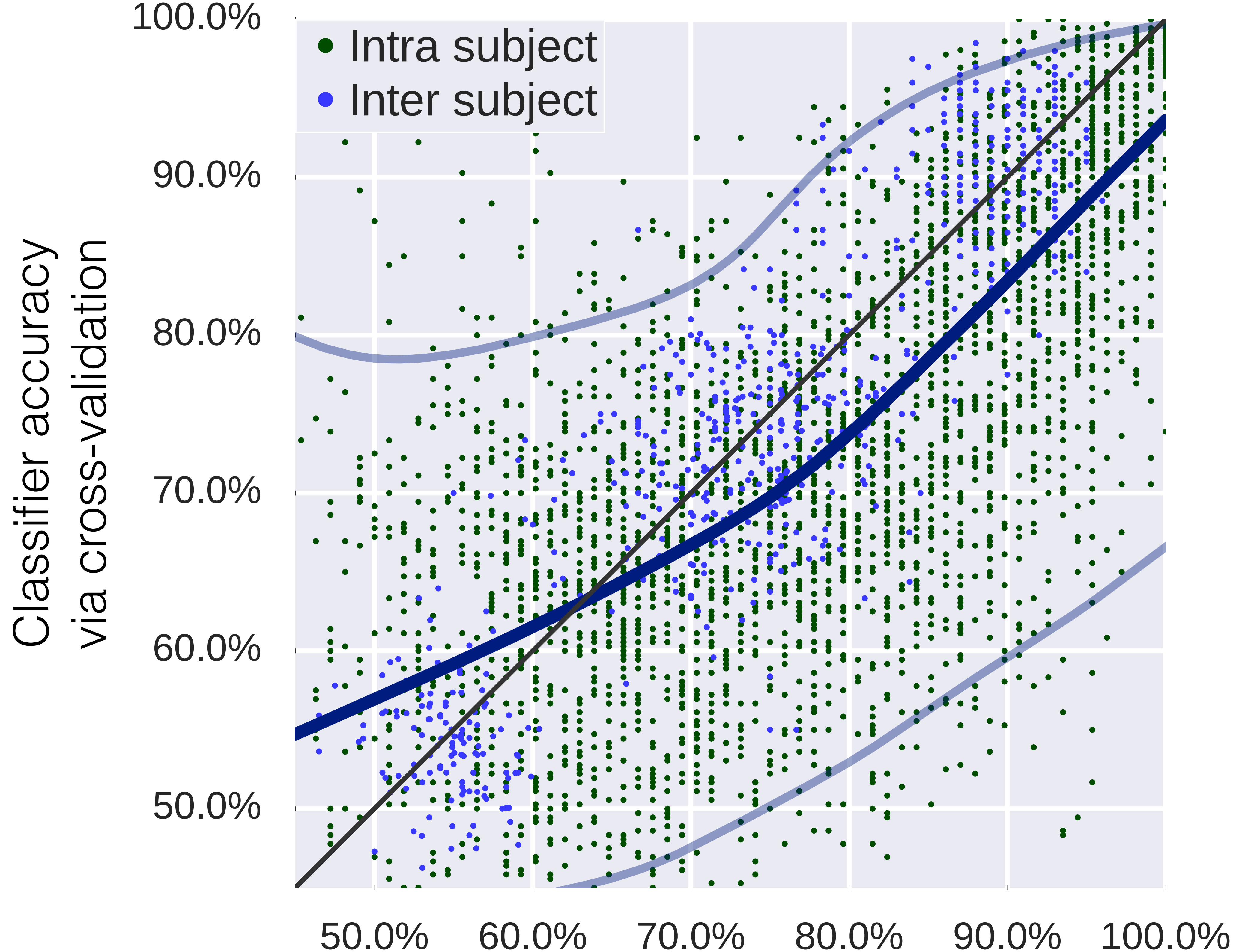
Reference: Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2016). **Assessing and tuning brain decoders: cross-validation, caveats, and guidelines.** NeuroImage.

Results: hold-out (10 trials)

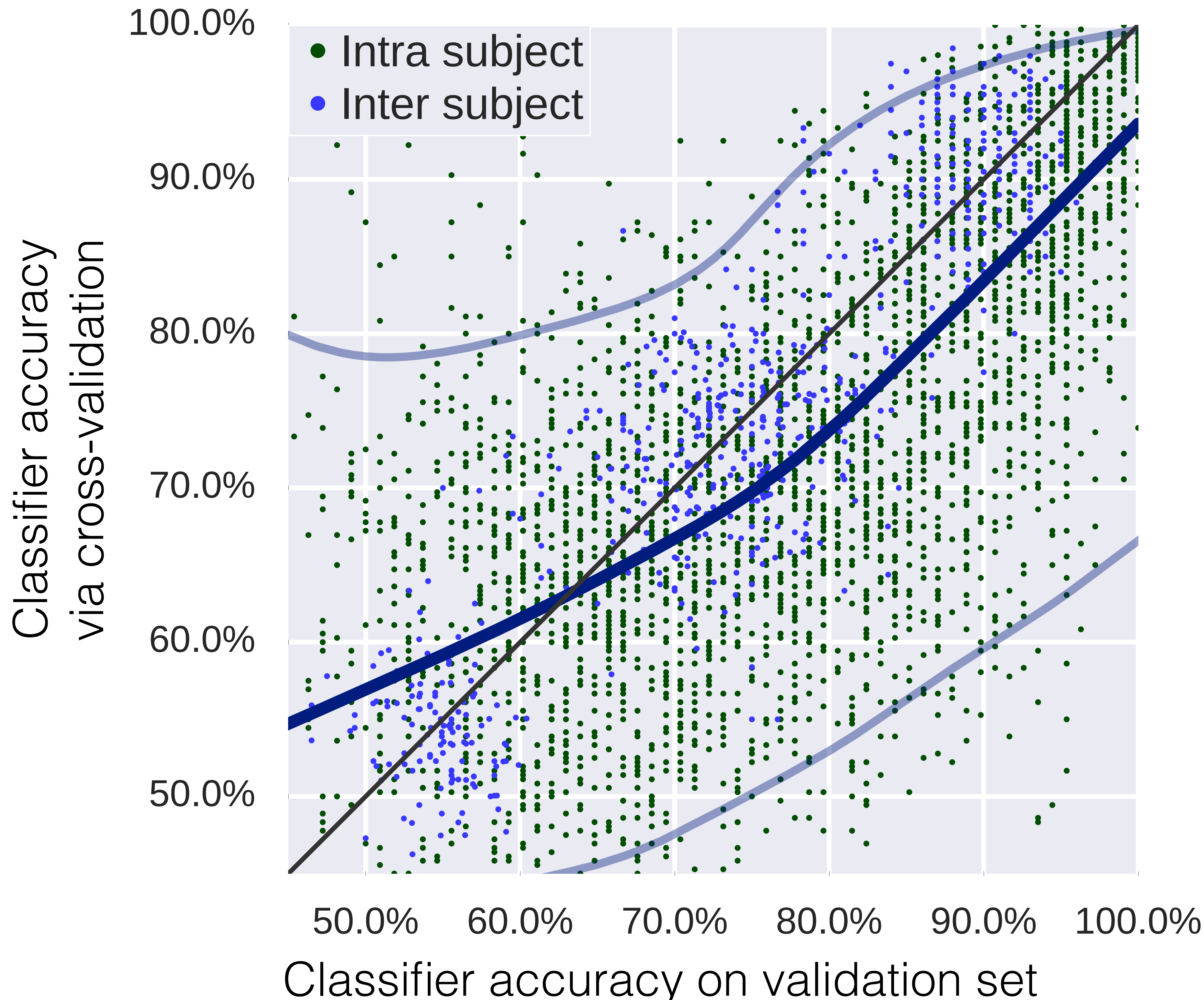


P. Raamana

Results: hold-out (10 trials)

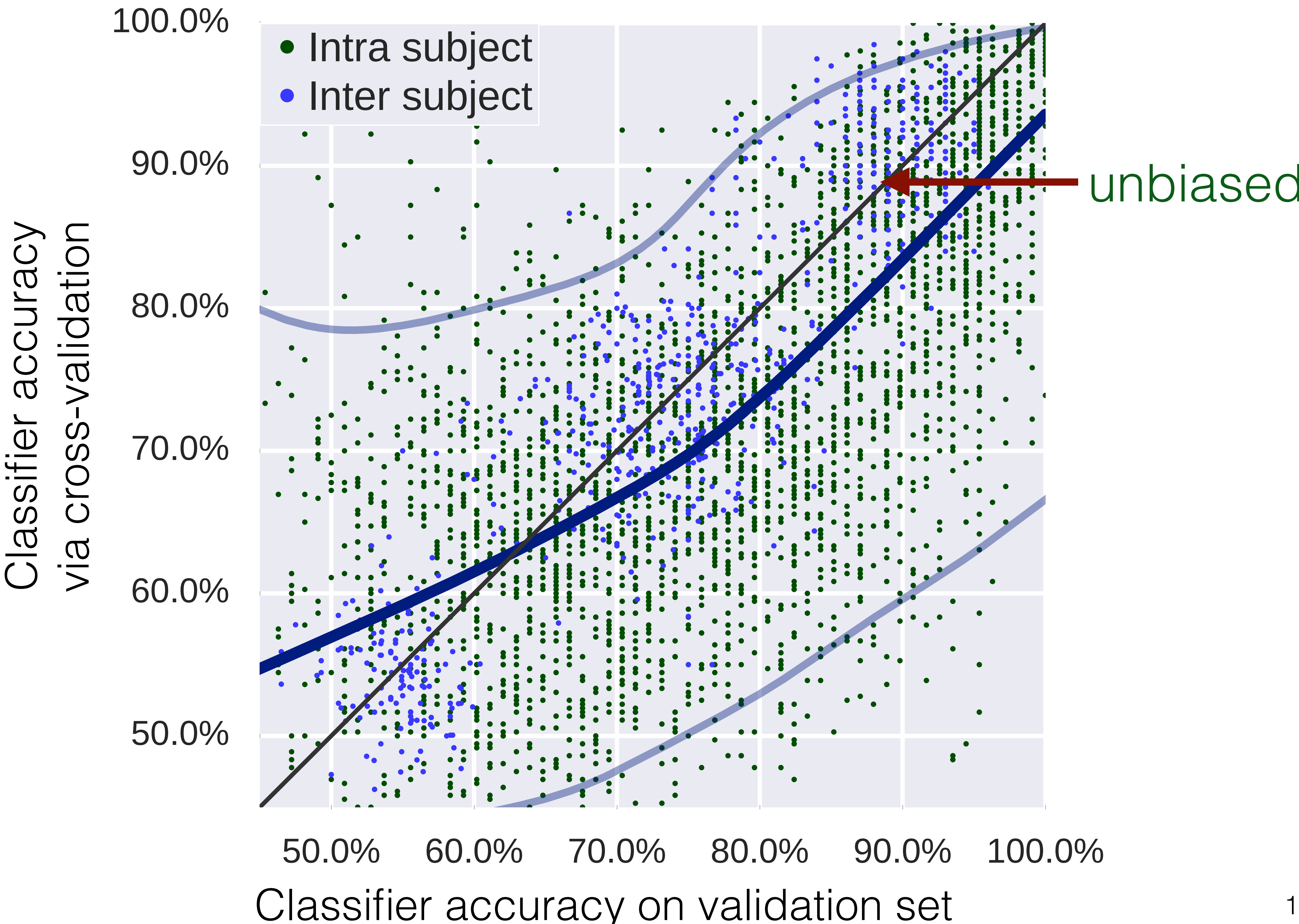


Results: hold-out (10 trials)



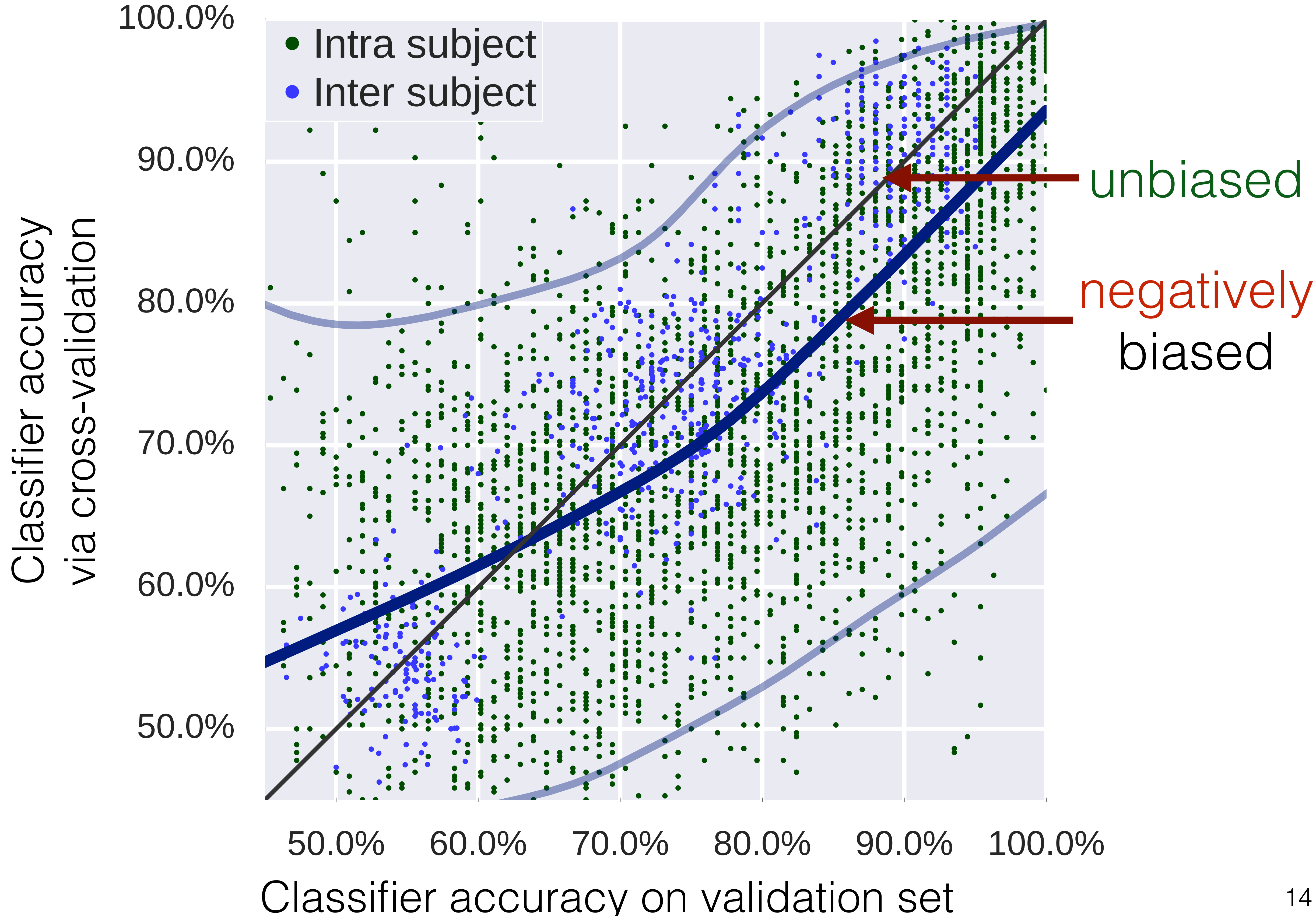
P. Raamana

Results: hold-out (10 trials)



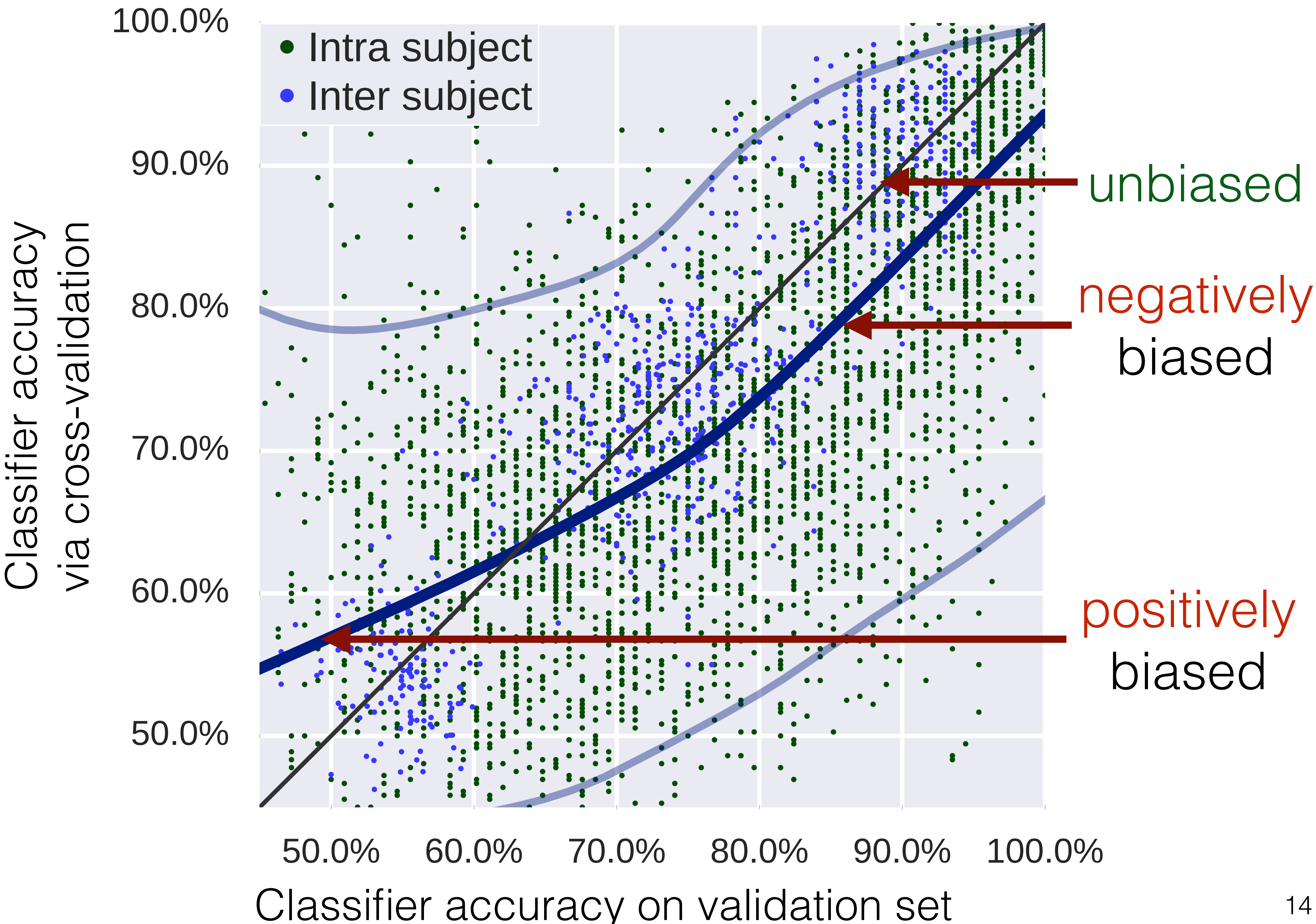
P. Raamana

Results: hold-out (10 trials)



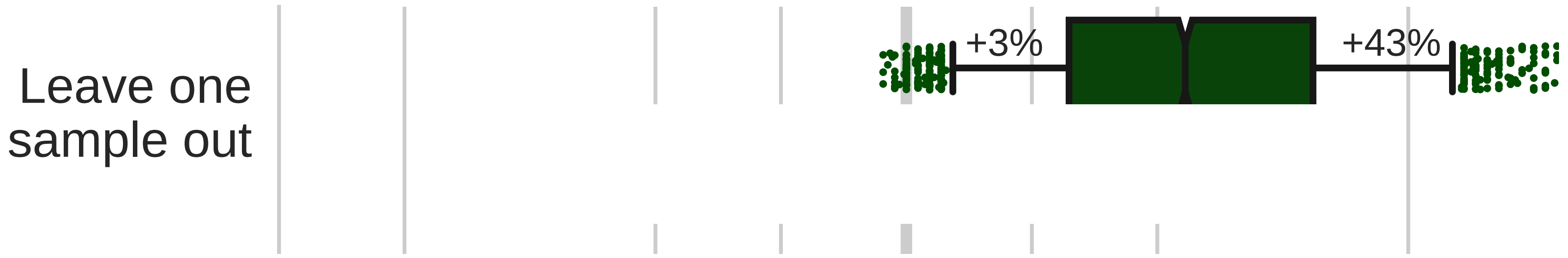
P. Raamana

Results: hold-out (10 trials)

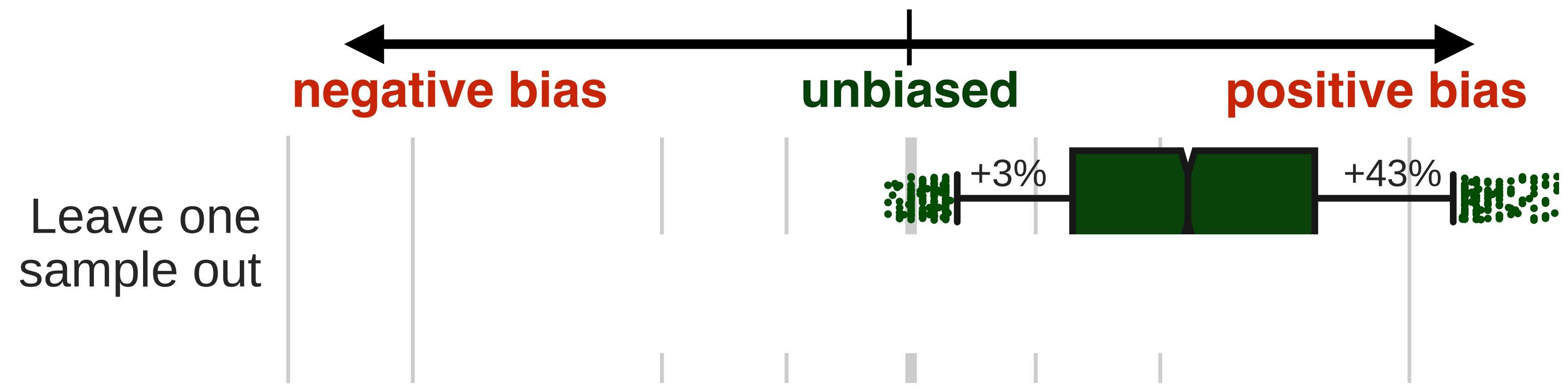


P. Raamana

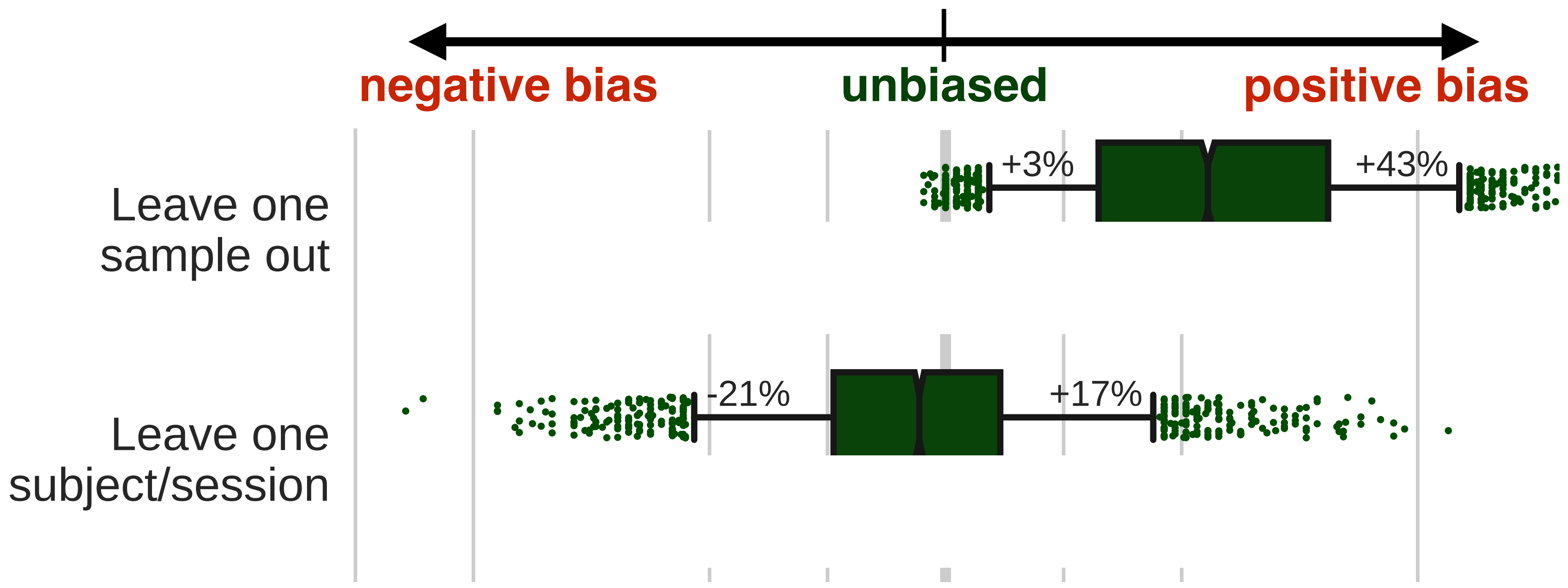
CV vs. Validation: real data



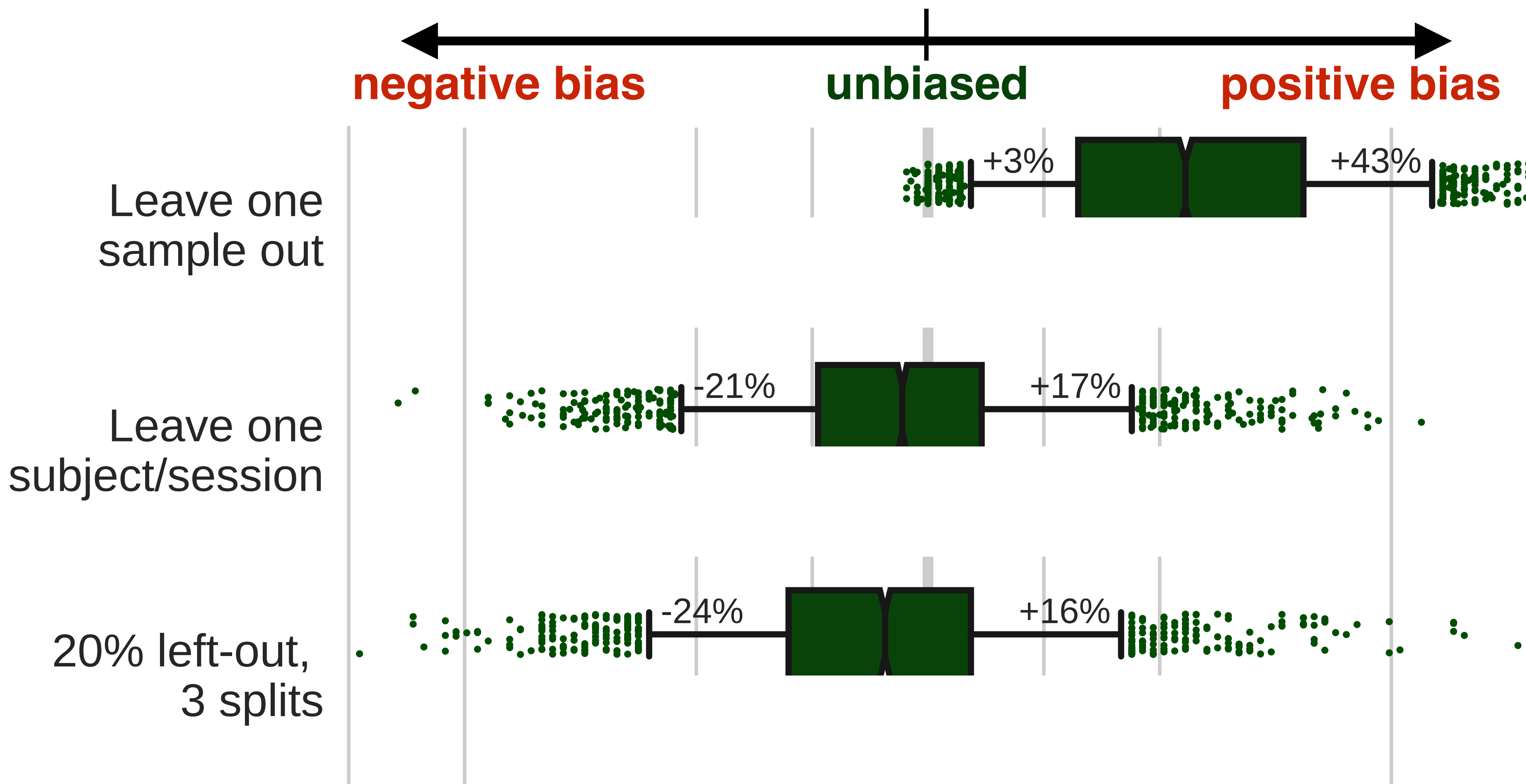
CV vs. Validation: real data



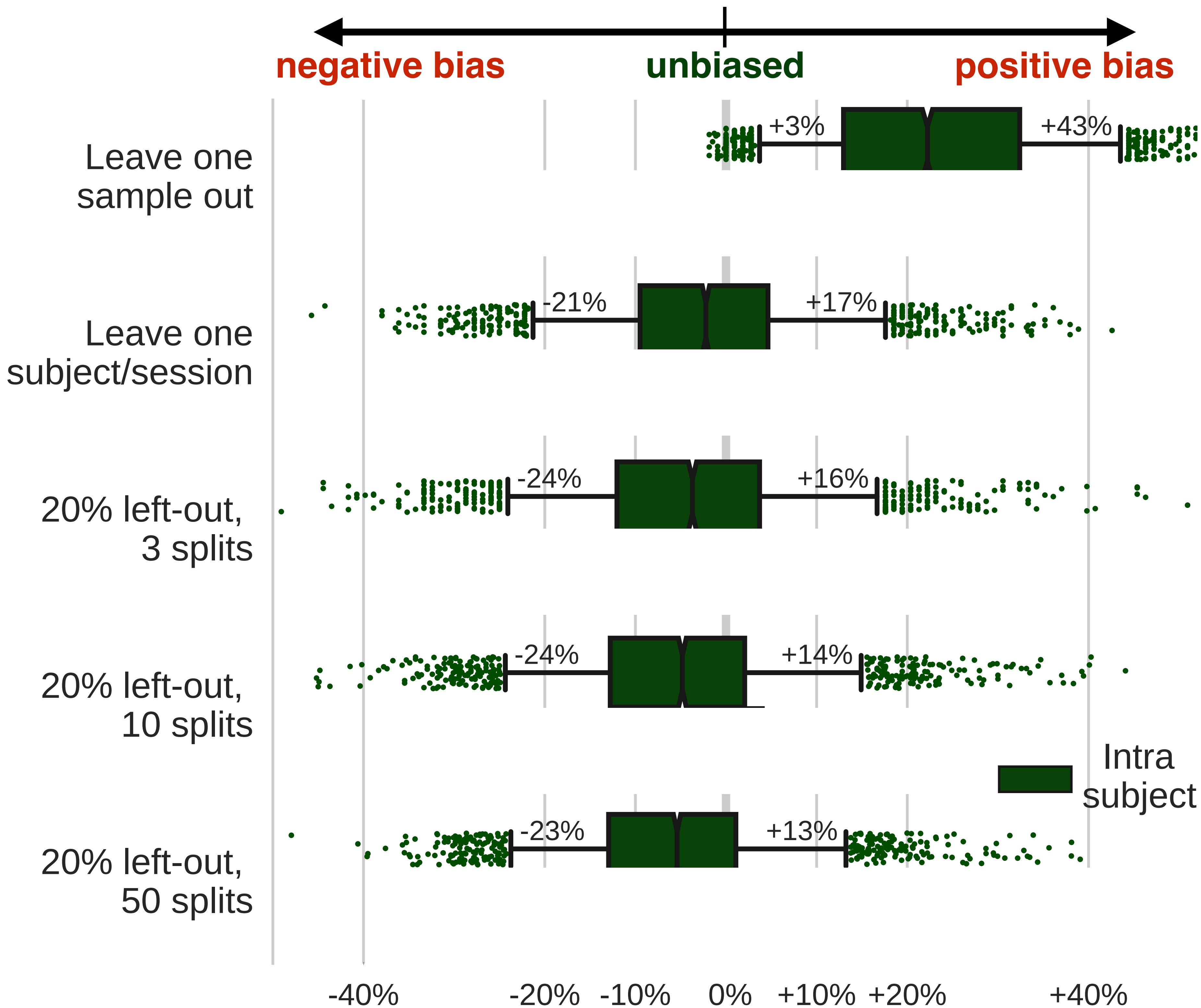
CV vs. Validation: real data



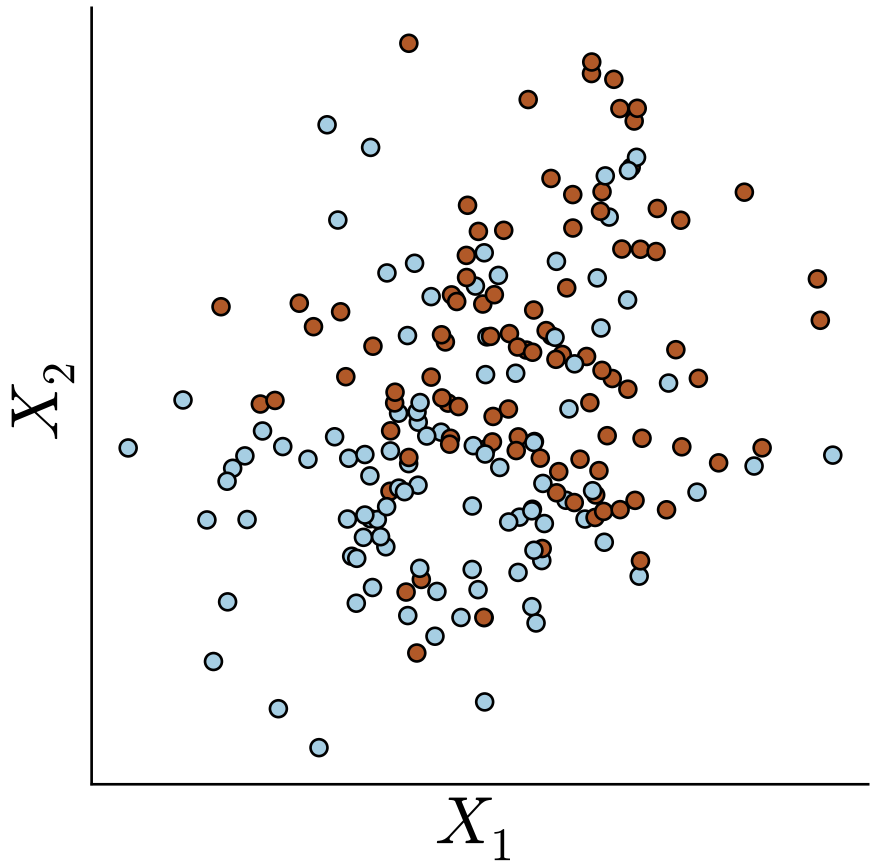
CV vs. Validation: real data



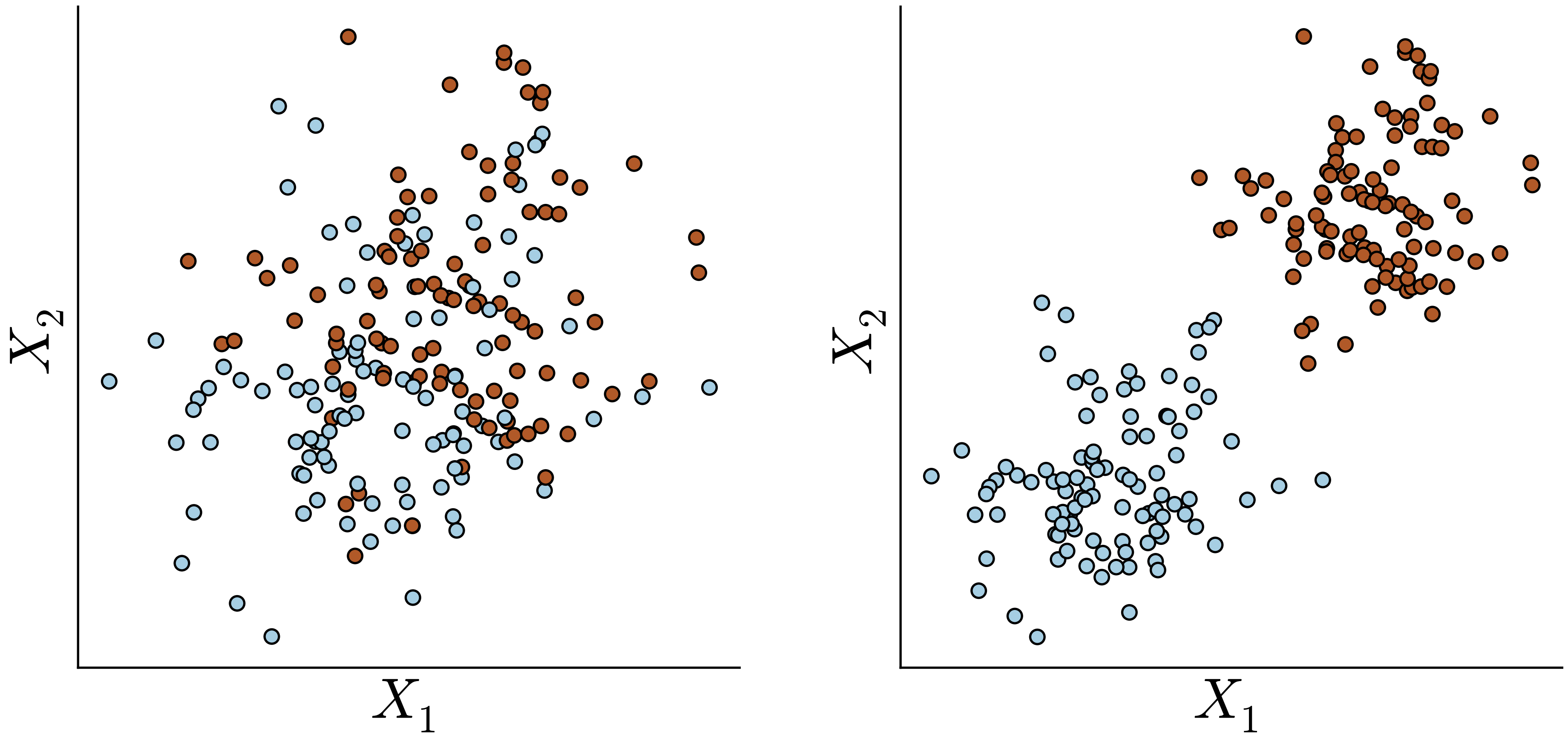
CV vs. Validation: real data



Simulations: known ground truth

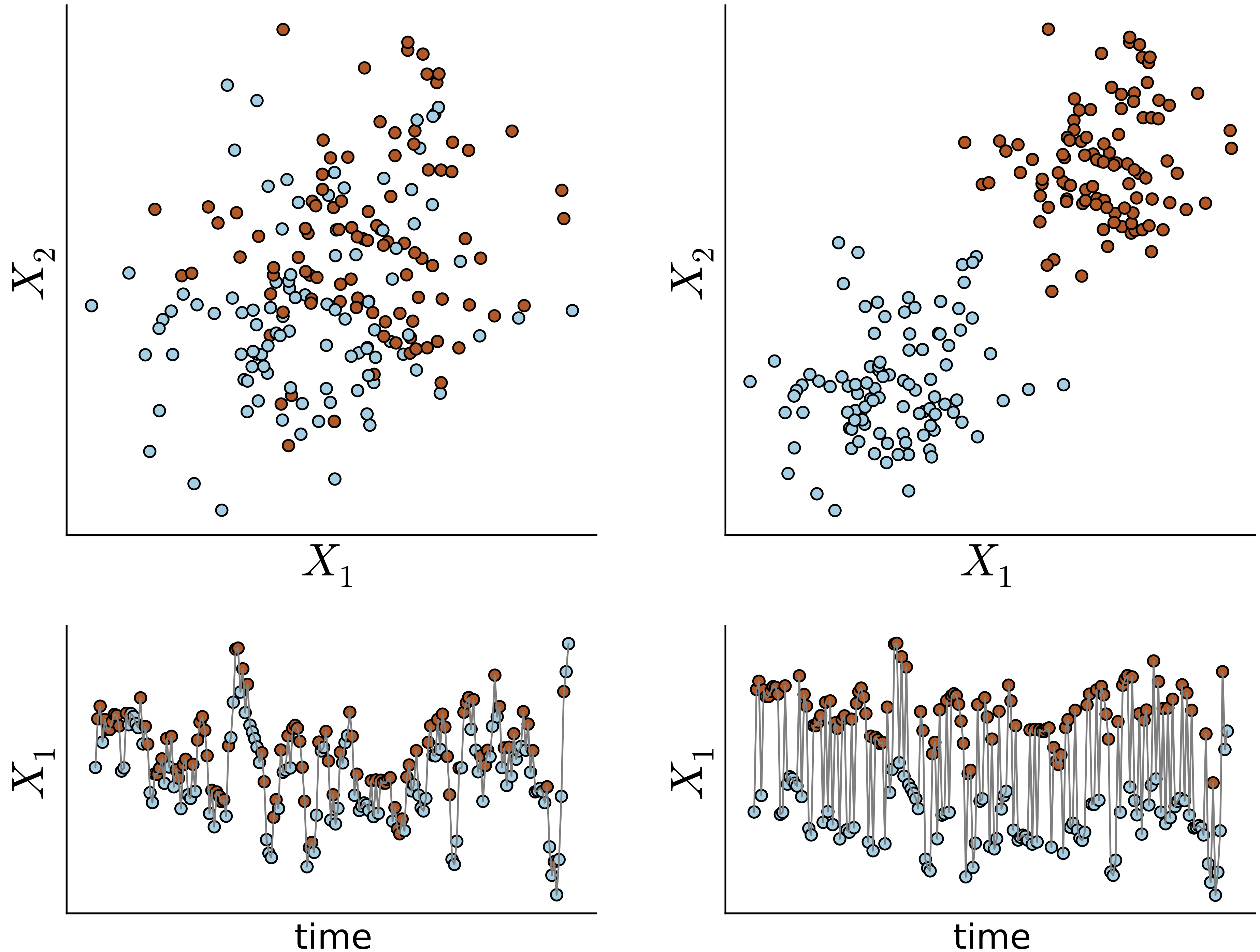


Simulations: known ground truth

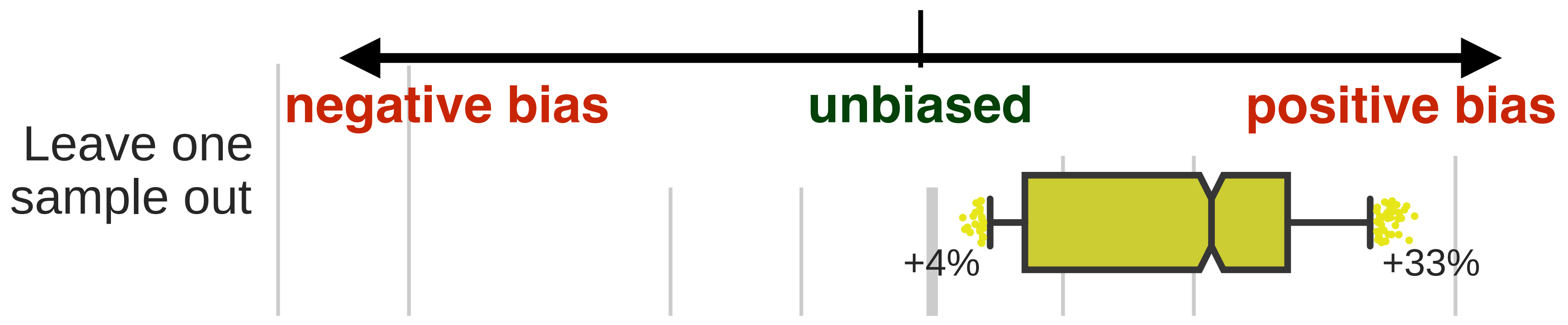


P. Raamana

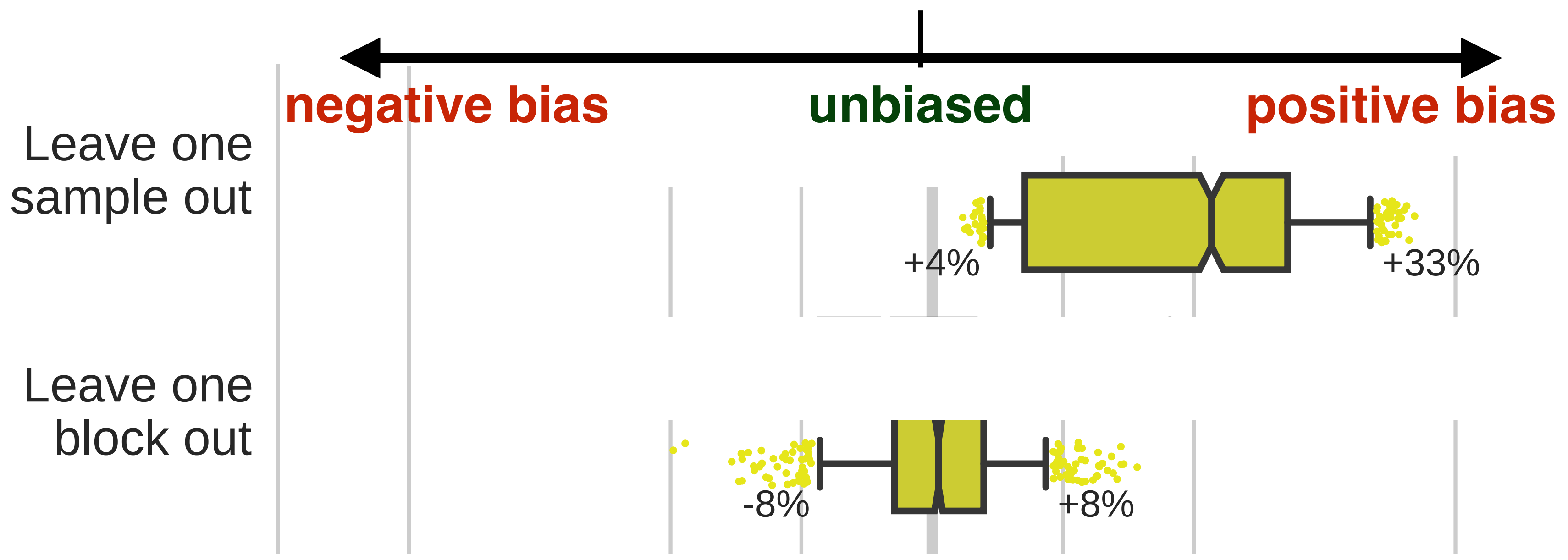
Simulations: known ground truth



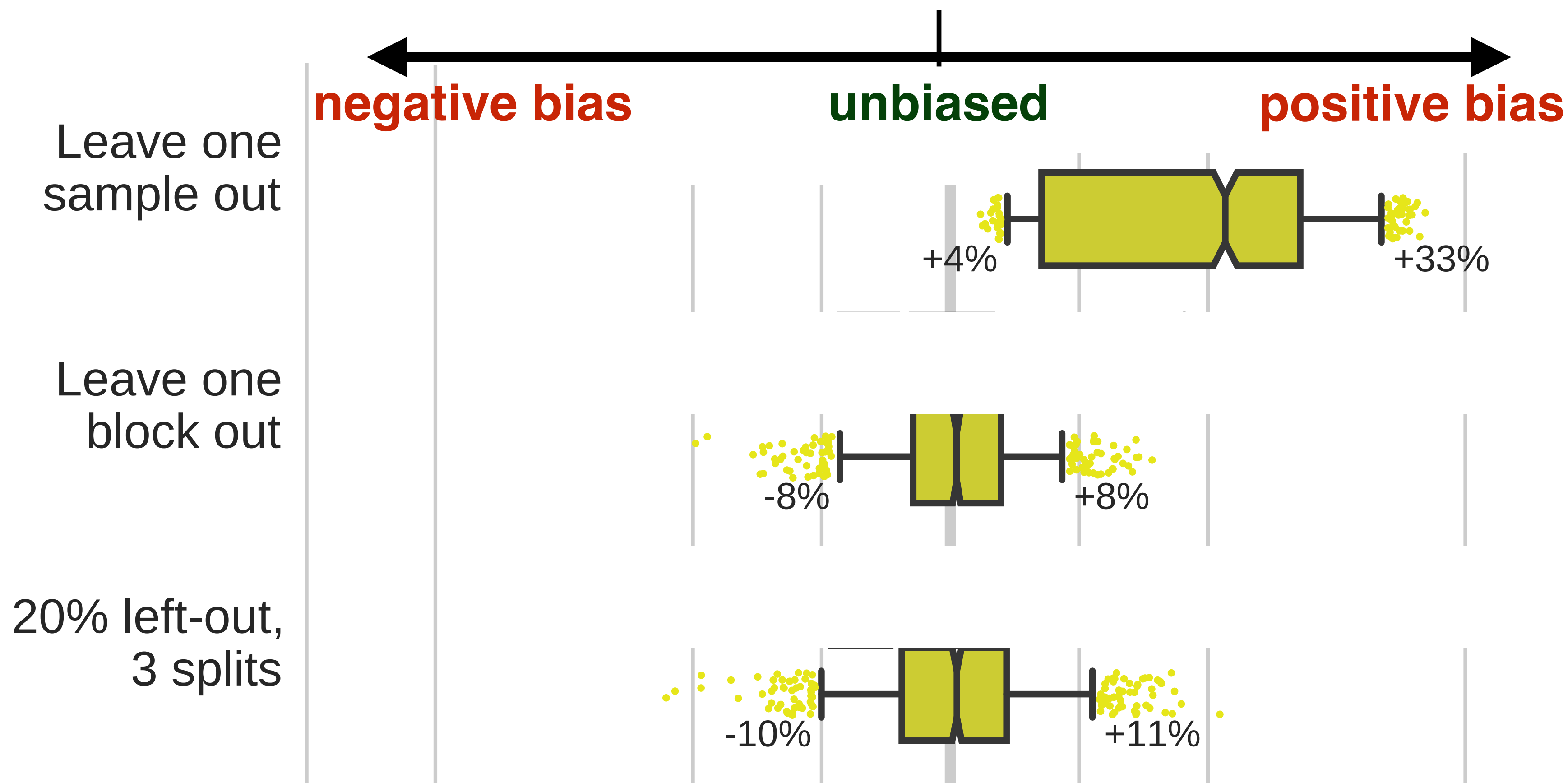
CV vs. Validation



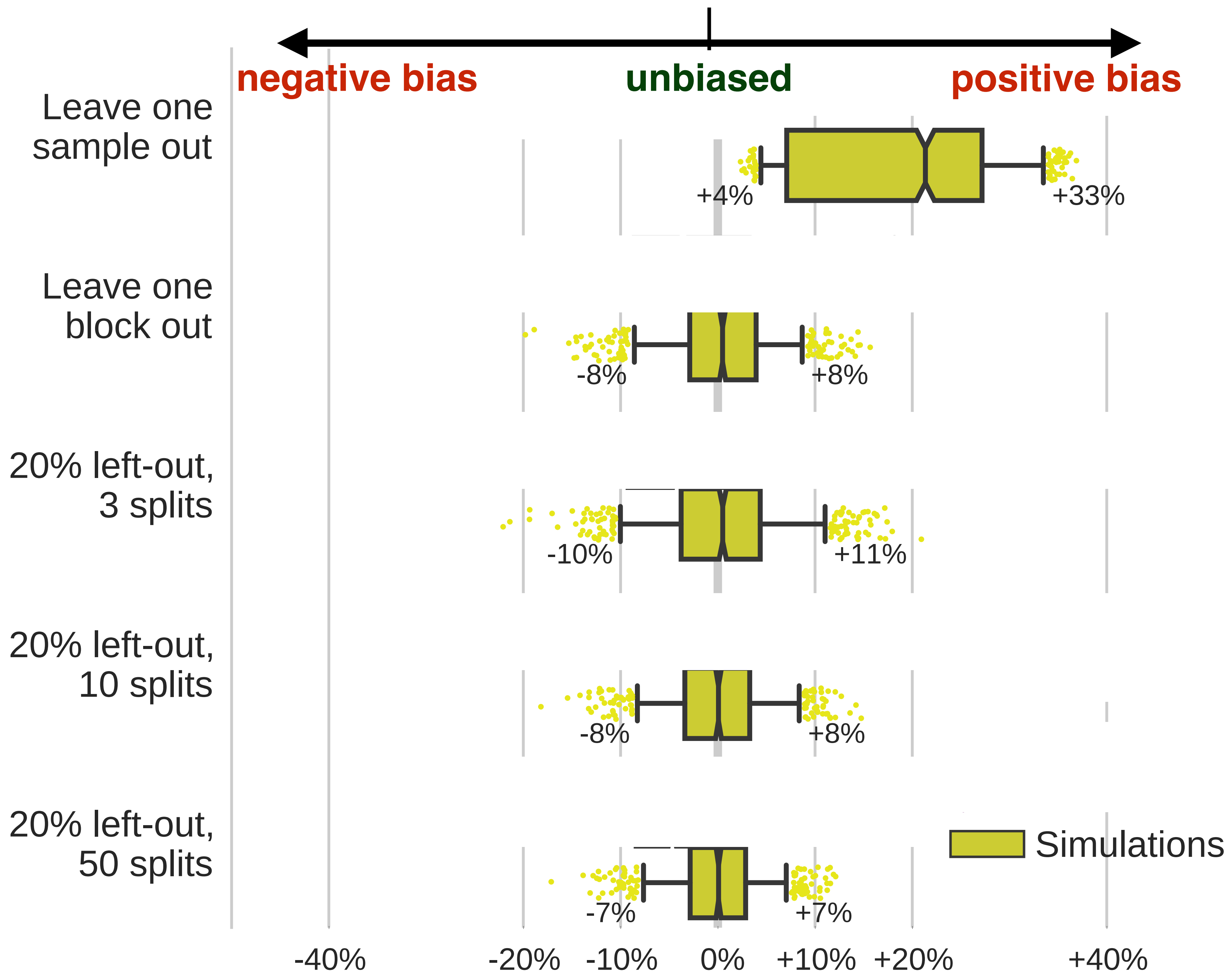
CV vs. Validation



CV vs. Validation



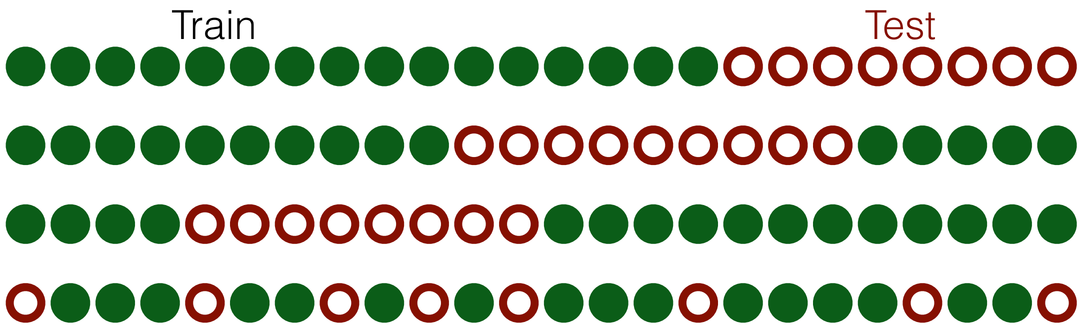
CV vs. Validation



Aggregation across folds

- It's not enough to properly split each fold, and accurately evaluate classifier performance!
- Not all measures across folds are *commensurate*!
    - e.g. decision scores from SVM (reference plane and zero are different!)
    - hence they can not be pooled across folds to construct an ROC!
    - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!

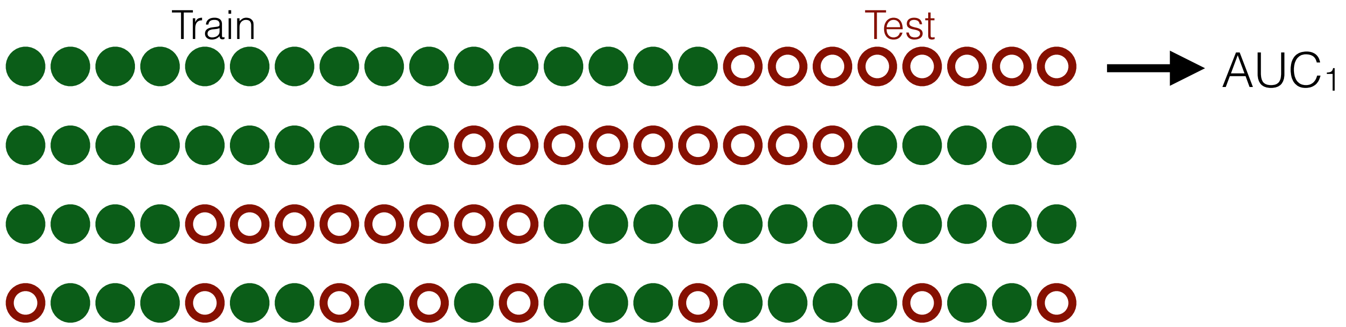
Aggregation across folds



- It's not enough to properly split each fold, and accurately evaluate classifier performance!
- Not all measures across folds are *commensurate*!
  - e.g. decision scores from SVM (reference plane and zero are different!)
  - hence they can not be pooled across folds to construct an ROC!
  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!

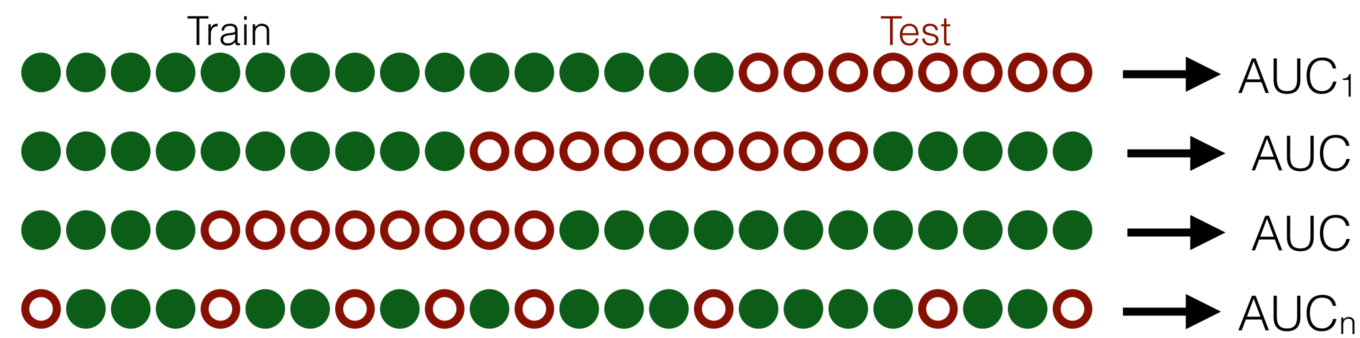
Aggregation across folds

- It's not enough to properly split each fold, and accurately evaluate classifier performance!
- Not all measures across folds are *commensurate*!
    - e.g. decision scores from SVM (reference plane and zero are different!)
    - hence they can not be pooled across folds to construct an ROC!
    - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!



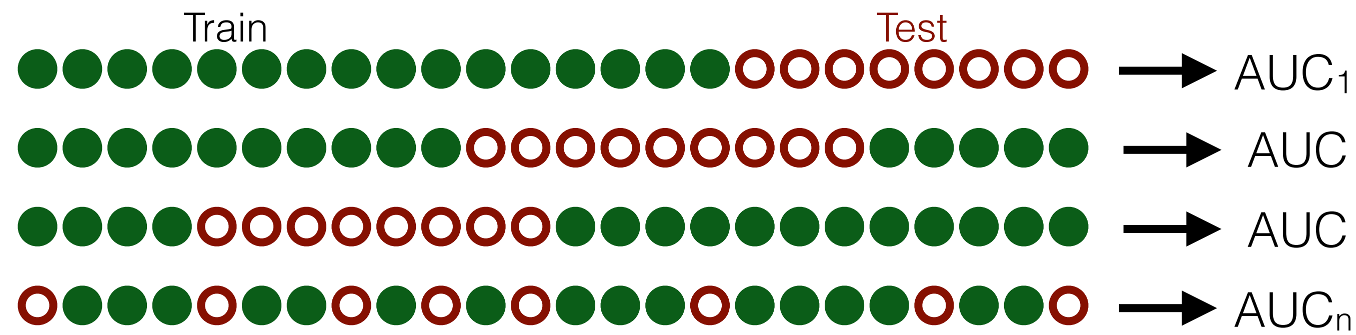
Aggregation across folds

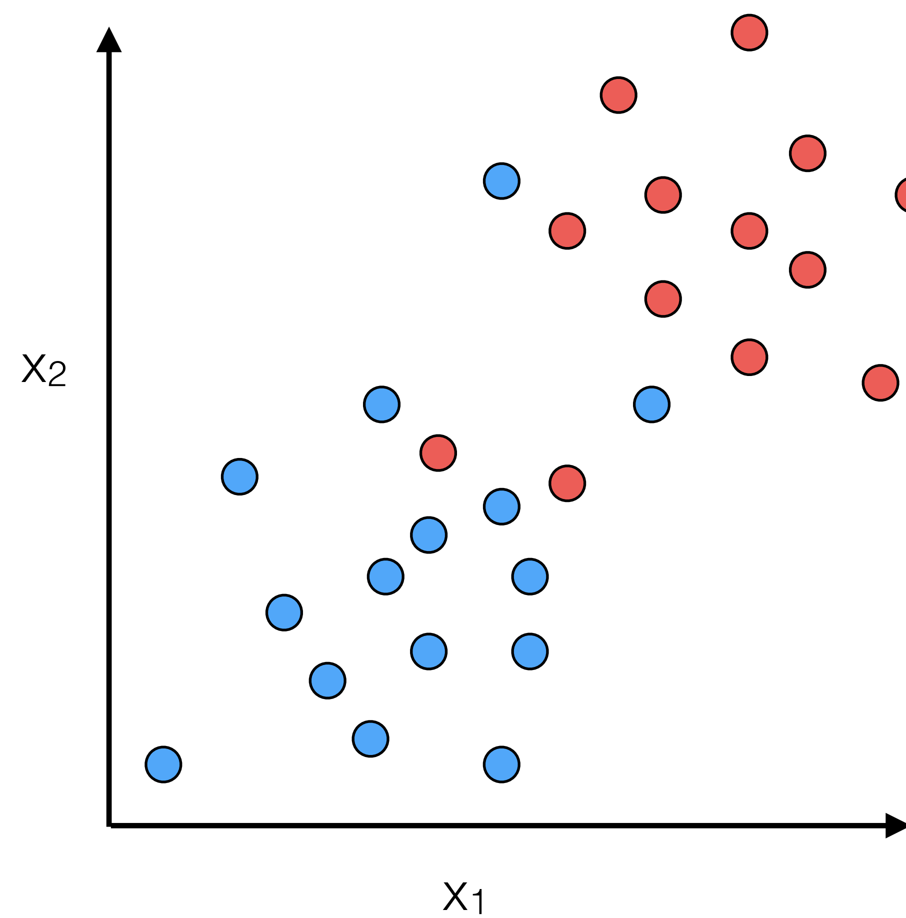
- It's not enough to properly split each fold, and accurately evaluate classifier performance!
- Not all measures across folds are *commensurate*!
    - e.g. decision scores from SVM (reference plane and zero are different!)
    - hence they can not be pooled across folds to construct an ROC!
    - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!



Aggregation across folds

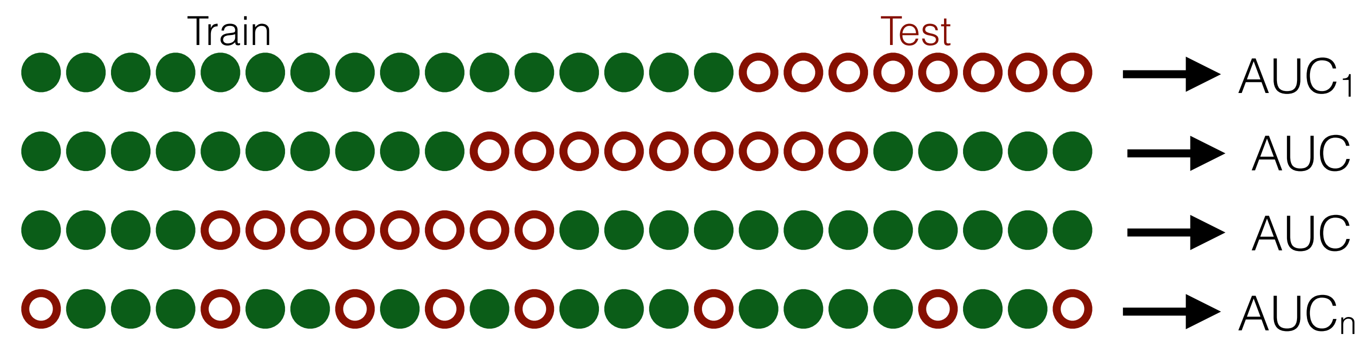
- It's not enough to properly split each fold, and accurately evaluate classifier performance!
- Not all measures across folds are *commensurate*!
  - e.g. decision scores from SVM (reference plane and zero are different!)
  - hence they can not be pooled across folds to construct an ROC!
  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!

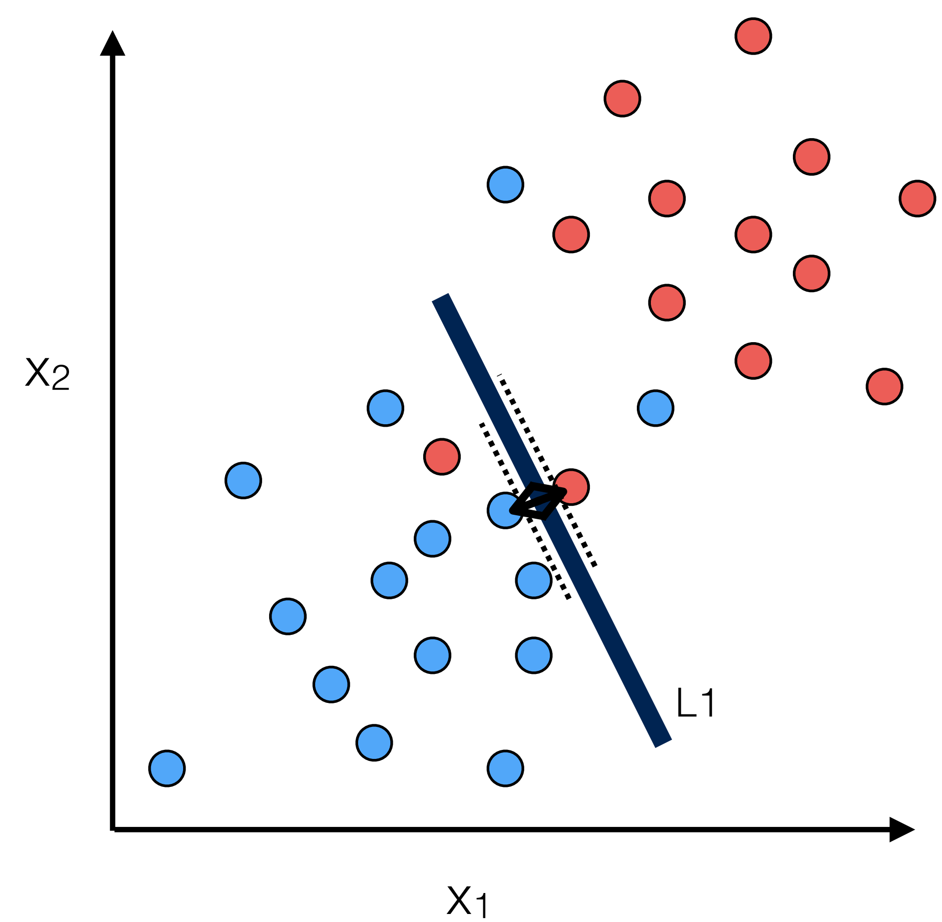




Aggregation across folds

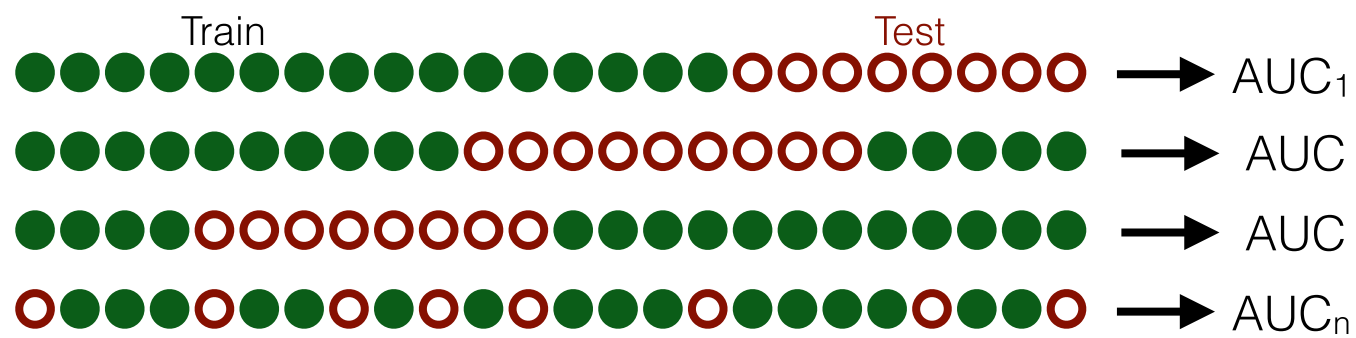
- It's not enough to properly split each fold, and accurately evaluate classifier performance!
- Not all measures across folds are *commensurate*!
  - e.g. decision scores from SVM (reference plane and zero are different!)
  - hence they can not be pooled across folds to construct an ROC!
  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!

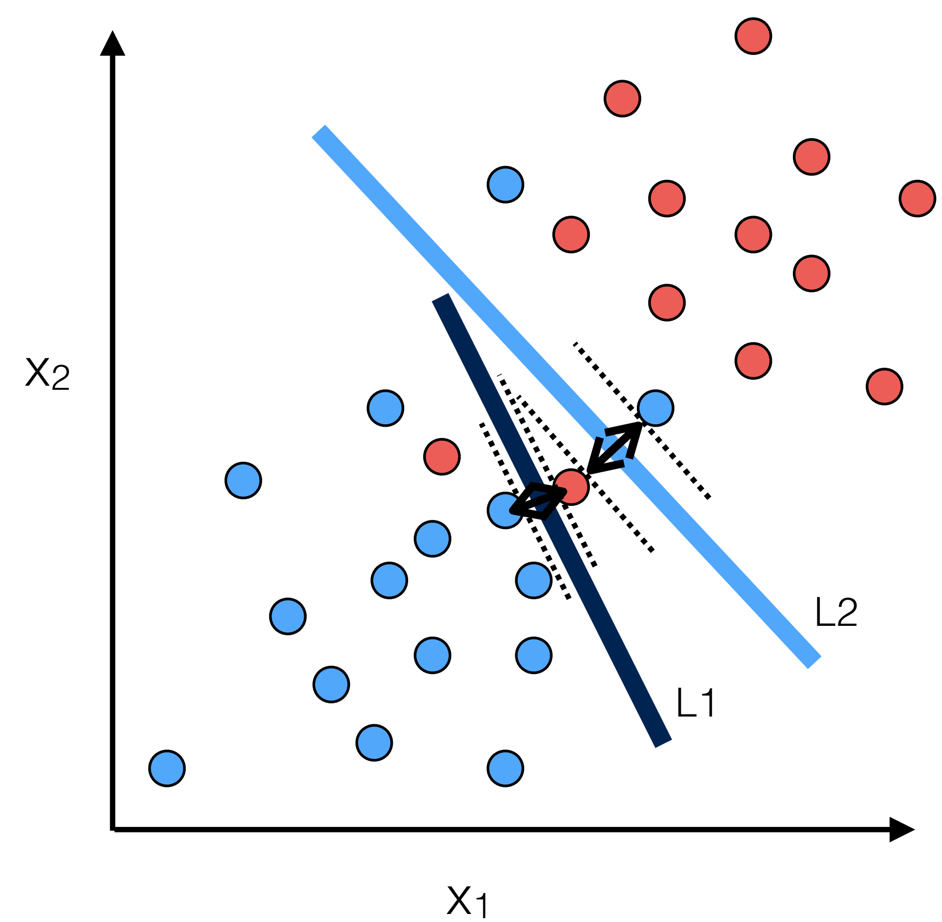




Aggregation across folds

- It's not enough to properly split each fold, and accurately evaluate classifier performance!
- Not all measures across folds are *commensurate*!
  - e.g. decision scores from SVM (reference plane and zero are different!)
  - hence they can not be pooled across folds to construct an ROC!
  - Instead, make ROC per fold and compute AUC per fold, and then average AUC across folds!





Conclusions

- Avoid leave-one-out cross-validation
  - esp. when correlations are present in your data
  - produces optimistic estimates with high variance
- Use repeated-holdout (10-50% for testing)
  - respecting sample/dependency structure
  - maximizing independence between train & test sets

In God we trust, but all others must cross-validate!

- Results could vary drastically with a different CV scheme
- CV results have variance (>10%)
- Document CV scheme in detail:
    - type of split
    - number of repetitions
    - Full distribution of estimates
- Proper splitting is not enough, proper pooling is needed too.

References

- Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2016). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. NeuroImage. http://doi.org/10.1016/j.neuroimage.2016.10.038
- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40–79.
- Forman, G. (2010). Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. ACM SIGKDD Explorations Newsletter.

Acknowledgements

Baycrest

ONTARIO NEURODEGENERATIVE DISEASE RESEARCH INITIATIVE

ONTARIO BRAIN INSTITUTE

CAN-BIND

SFU

UNIVERSITY OF TORONTO

Follow @raamana_