GENERAL ASSEMBLY

MODEL VALIDATION (TRAIN/TEST SPLIT AND CROSS VALIDATION)

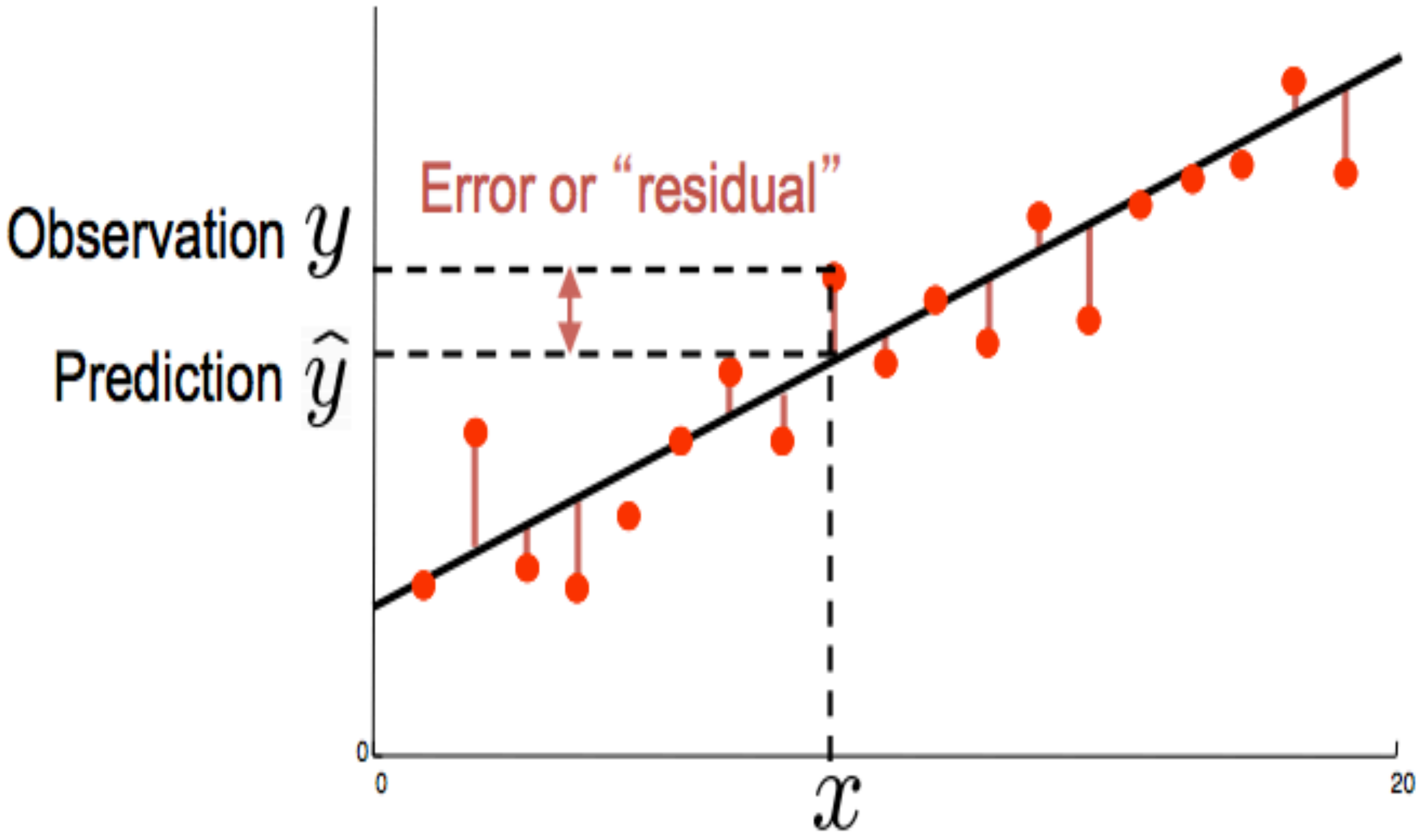
Joseph Nelson, Data Science Immersive

AGENDA

- Review: Modeling
- Training, Validating, Testing
- Cross Validation
- Three-way Train/Test Split
- Coding Implementation

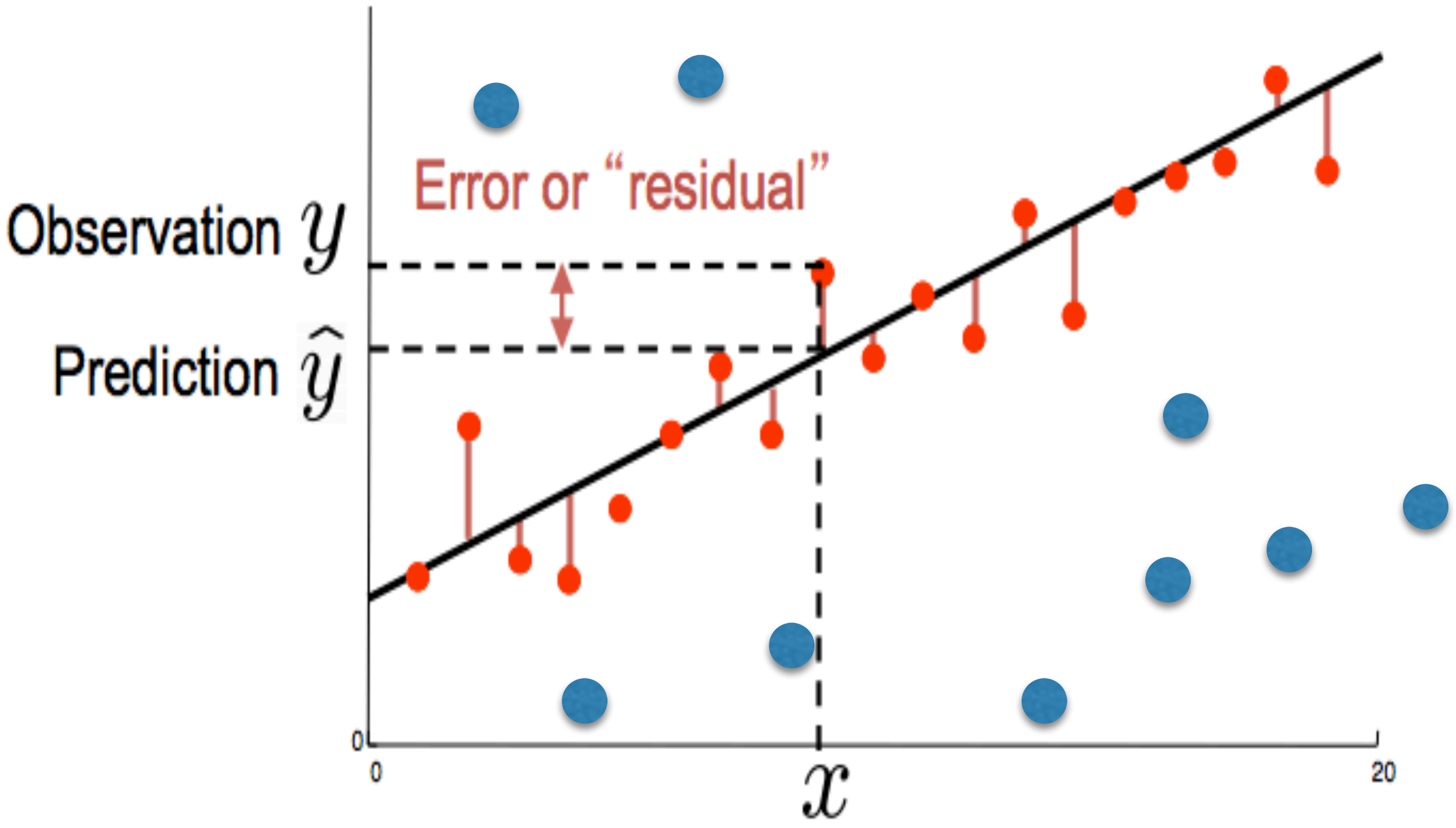
MODELING REVIEW

- Imagine we have EVERY point possible in the universe
- How would we model our data?



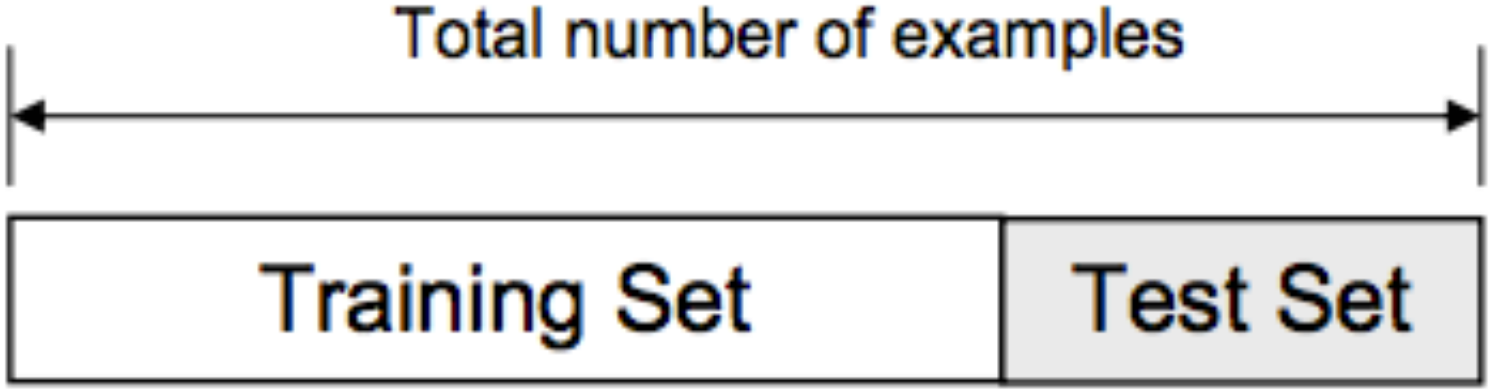
MODELING REVIEW

- Imagine we DO NOT have every point possible in the universe
- How would we model our data?
- Any possible solutions?



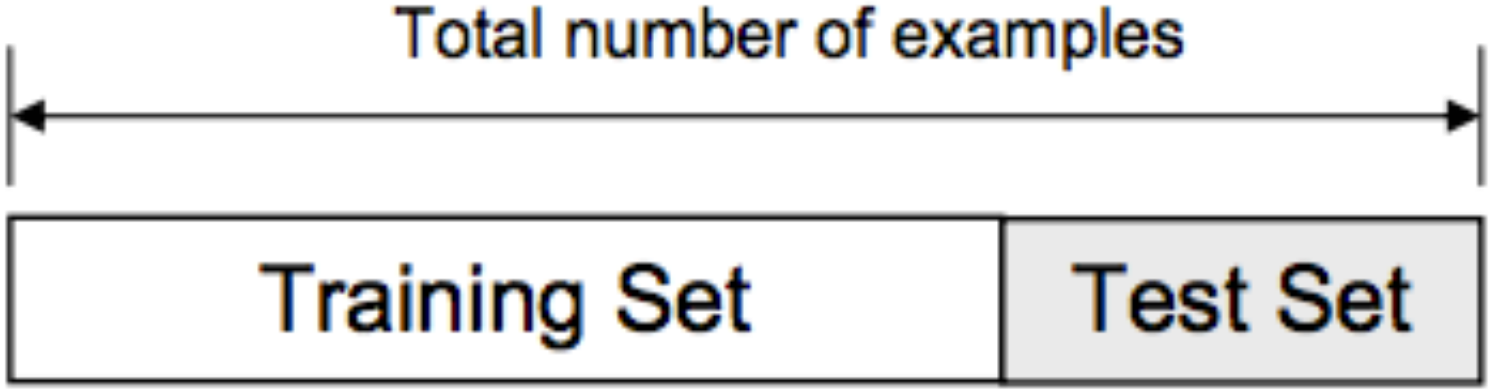
SPLITTING OUR DATA: TRAINING SET, TESTING SET

- THE HOLDOUT METHOD: Train/Test Split
- **Training Set**: Used to train the classifier
- **Testing Set:** Used to estimate the error rate of the trained classifier
- **Advantages?**
- **Disadvantages?**



SPLITTING OUR DATA: TRAINING SET, TESTING SET

- THE HOLDOUT METHOD: Train/Test Split
- **Training Set**: Used to train the classifier
- **Testing Set:** Used to estimate the error rate of the trained classifier
- **Advantages?** Fast! Simple! Computationally inexpensive!
- **Disadvantages?** Eliminating data! Imperfect splits!



THERE MUST BE ANOTHER WAY!

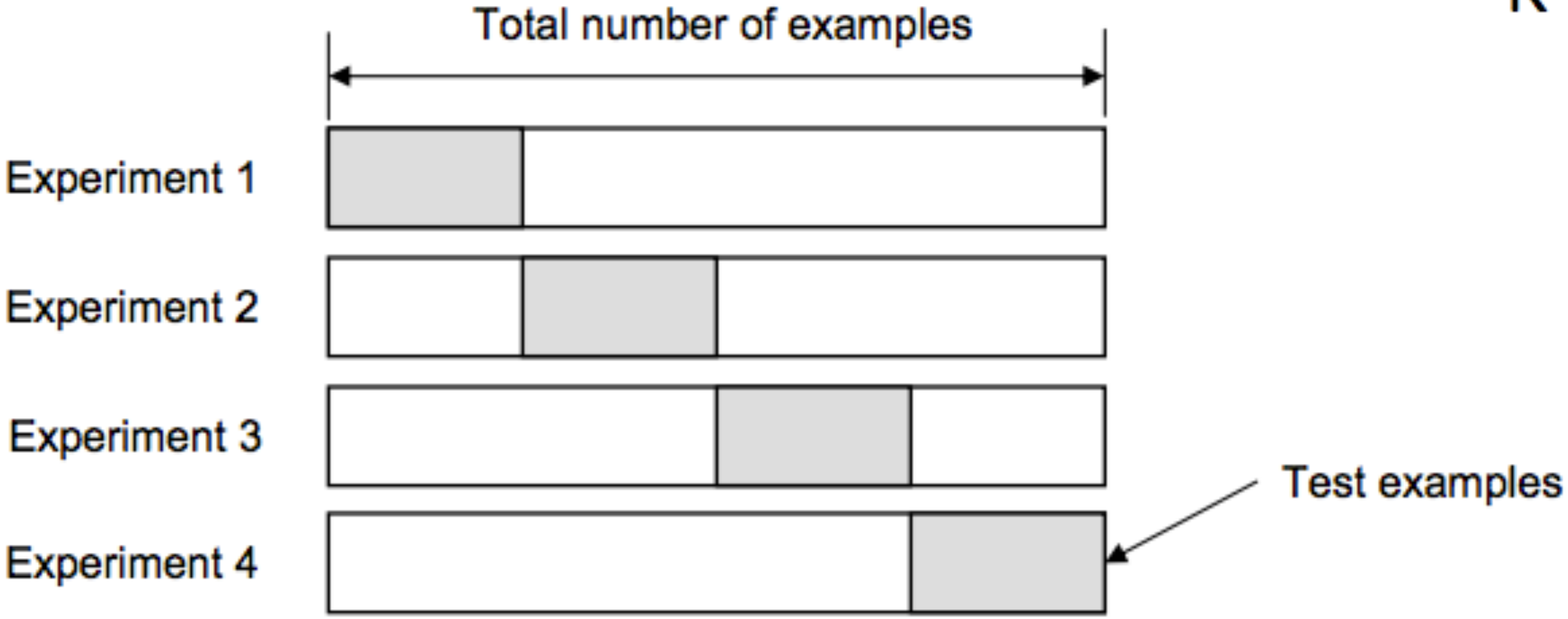
- **How can we use the maximum amount of our data points while still ensuring model integrity?**

- Toss out answers – your answers are valuable parts of being an inquisitive data scientist wanting to test your assumptions

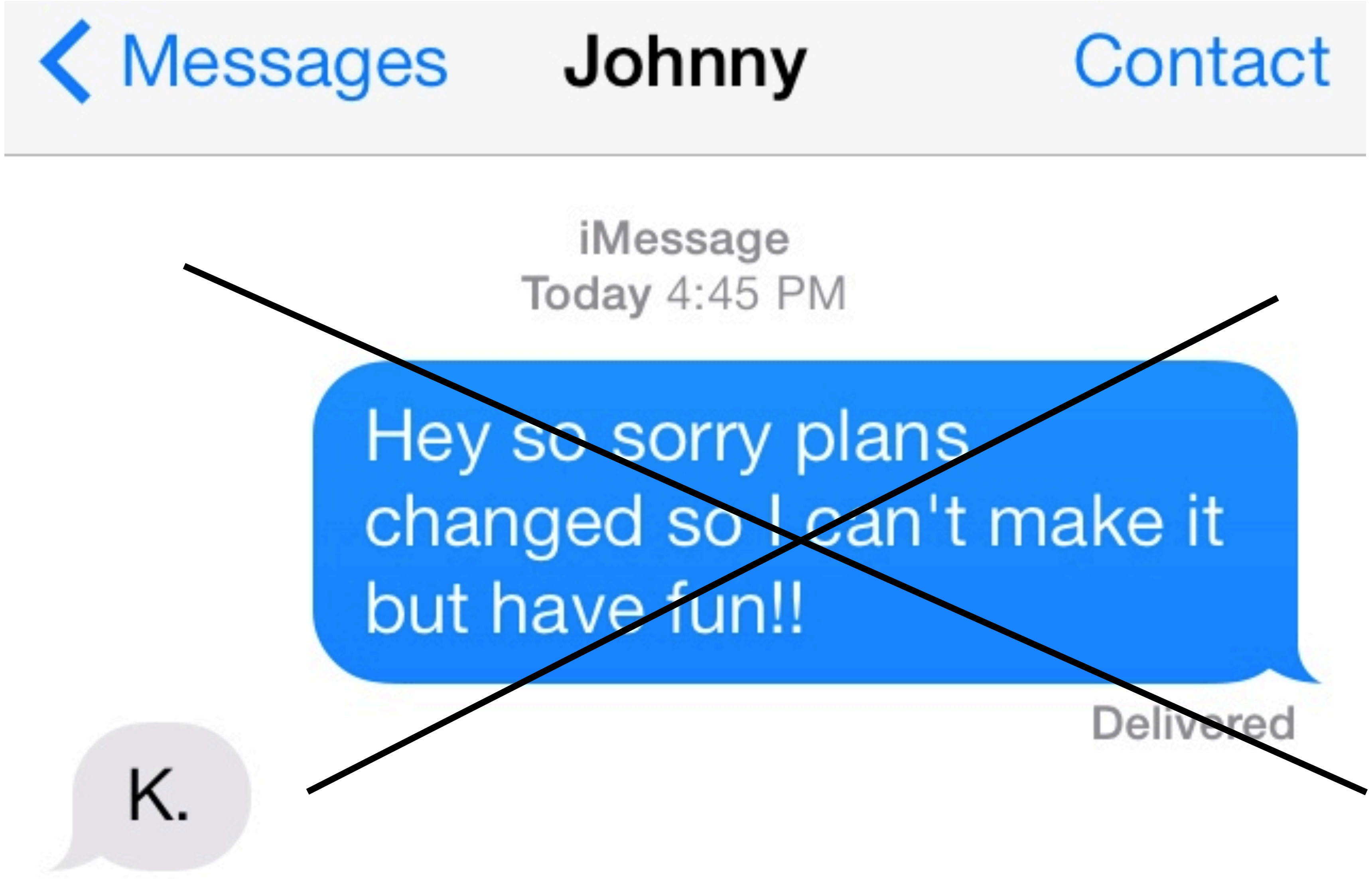
K-FOLDS CROSS VALIDATION

- Split our data into a number of different pieces (folds)
- Train using k-1 folds for training and a different fold for testing
- Average our model against EACH of those iterations
- Choose our model and TEST it against the final fold

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$



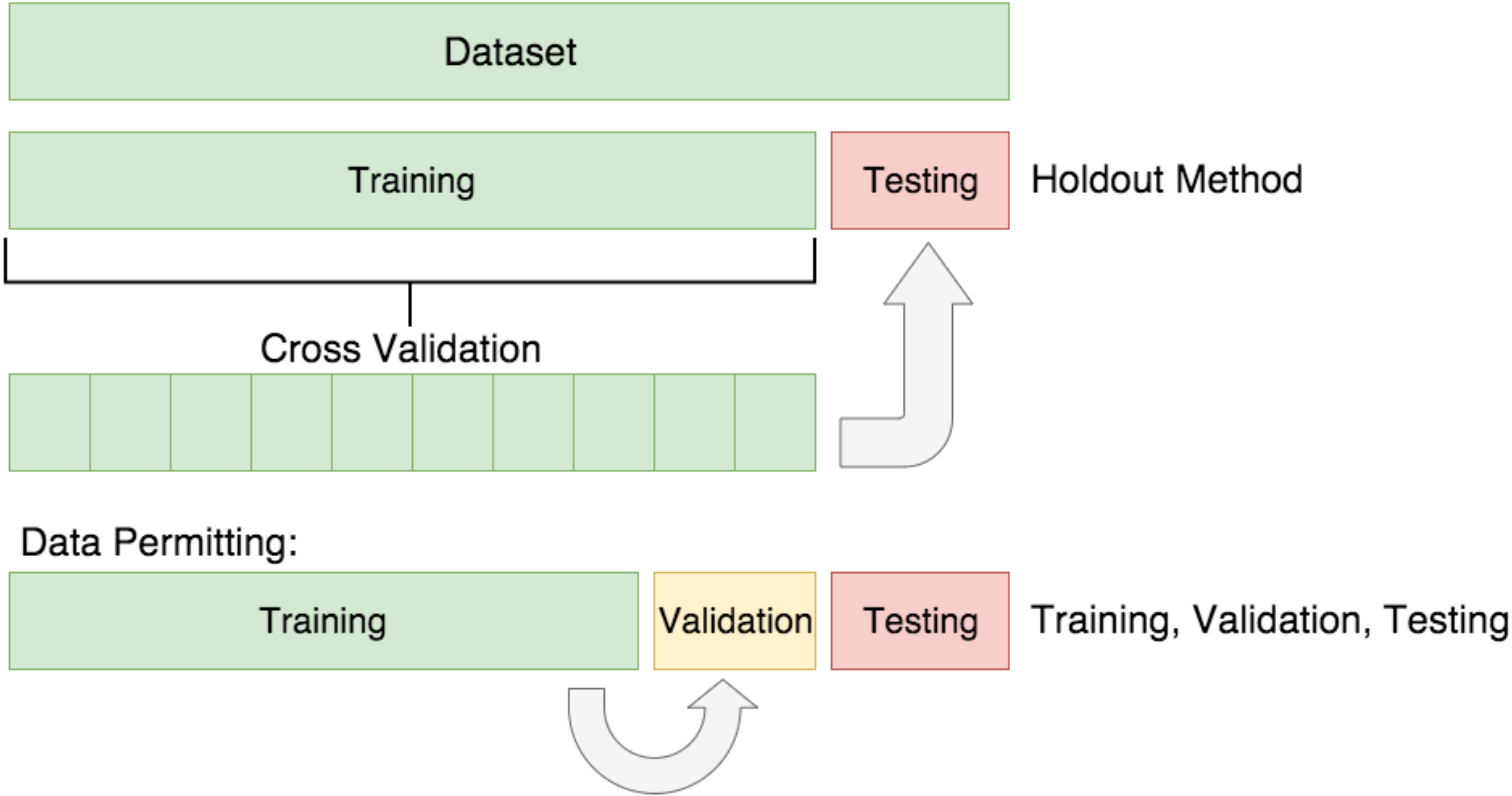
HOW MANY FOLDS ARE IN A K-FOLDS MODEL?



~~Hey so sorry plans changed so I can't make it but have fun!!~~

K.

PROCEDURE



Dataset

Training

Testing

Holdout Method

Cross Validation

Data Permitting:

Training

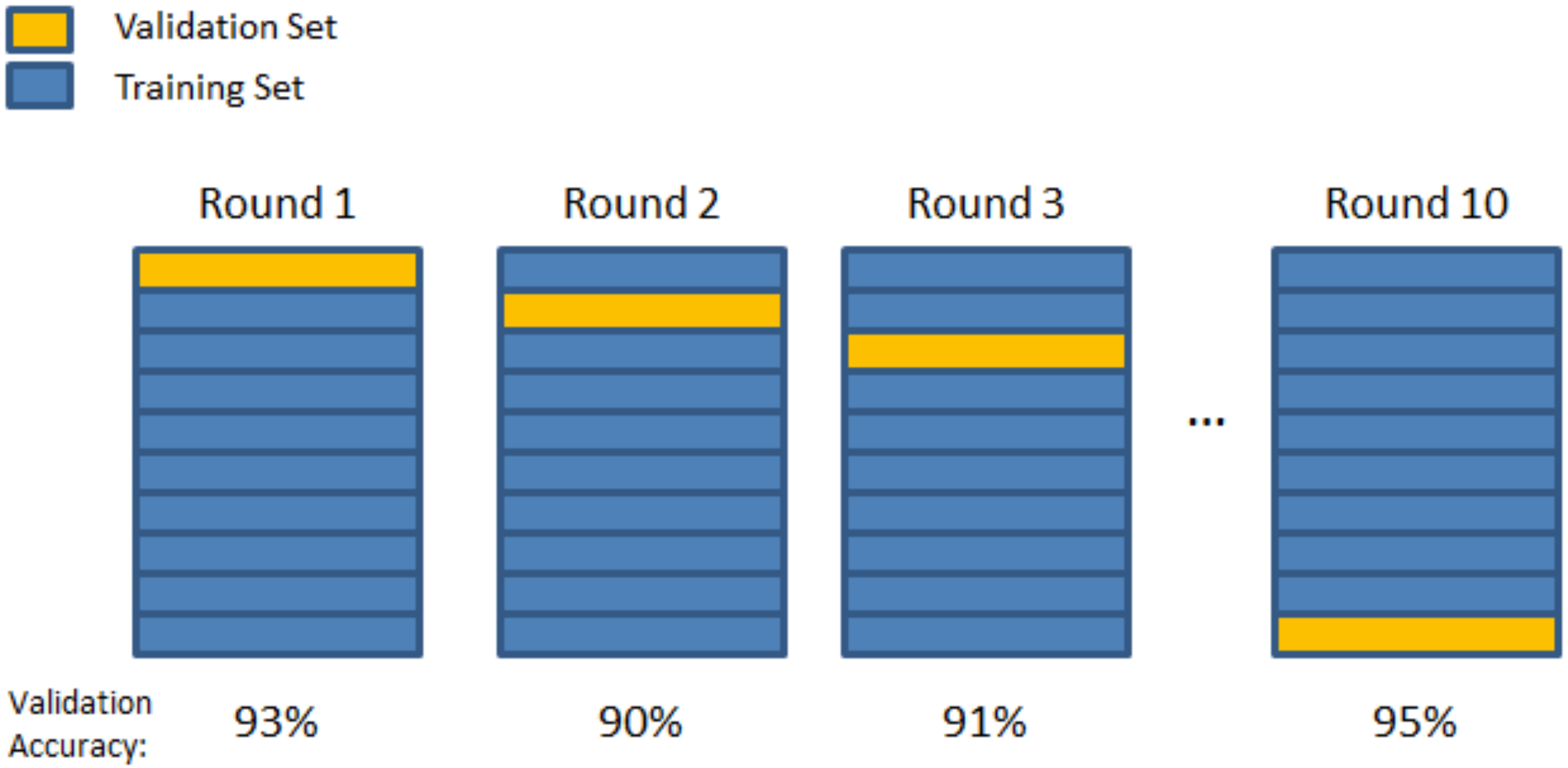
Validation

Testing

Training, Validation, Testing

K-FOLDS CROSS VALIDATION

- K=10
- Round 1: Check 9 training sets against one validation set. . . Round 2. . .

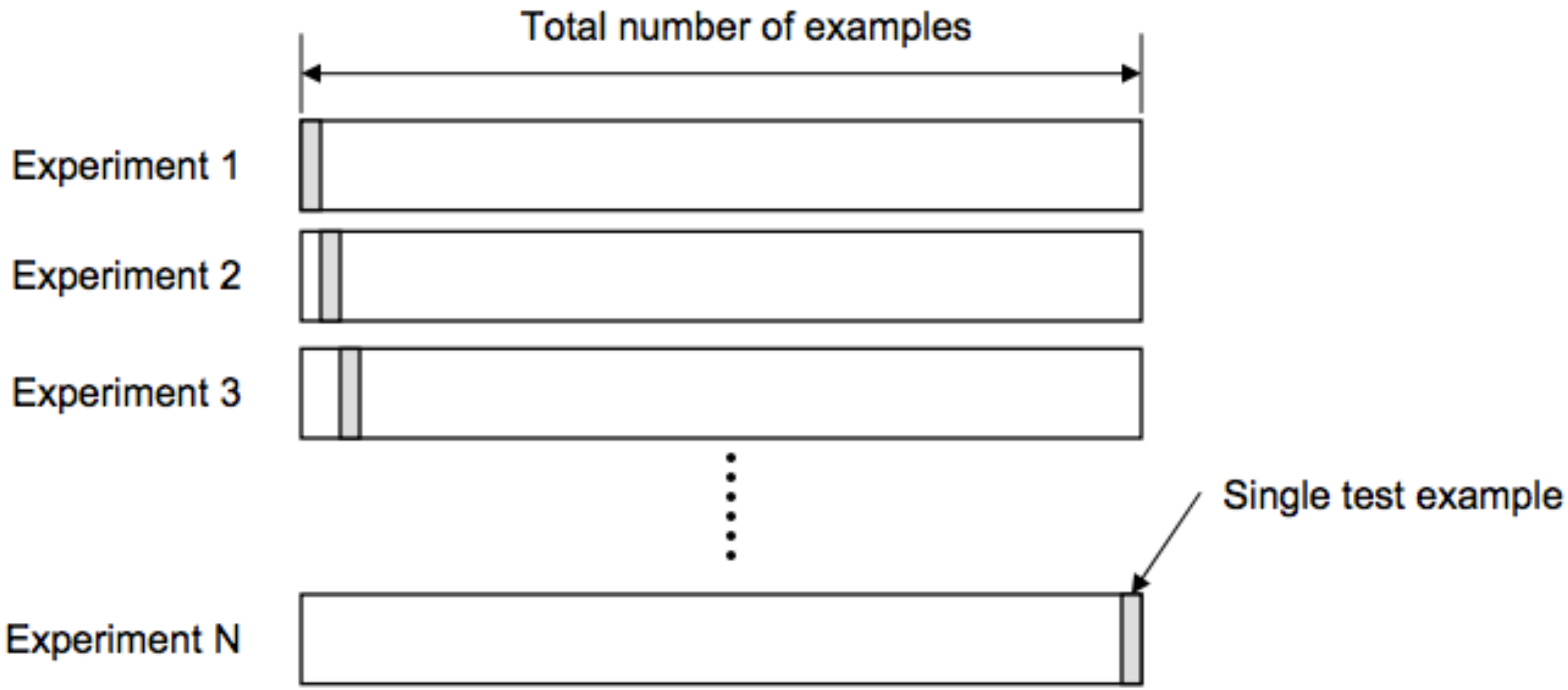


Final Accuracy = Average(Round 1, Round 2, ...)

LEAVE ONE OUT CROSS VALIDATION (LOOCV)

- K-folds is taken to the logical extreme: K = N
- For a dataset of N examples, perform N experiments
- Average our model against EACH of those iterations
- Choose our model and TEST it against the final fold

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$



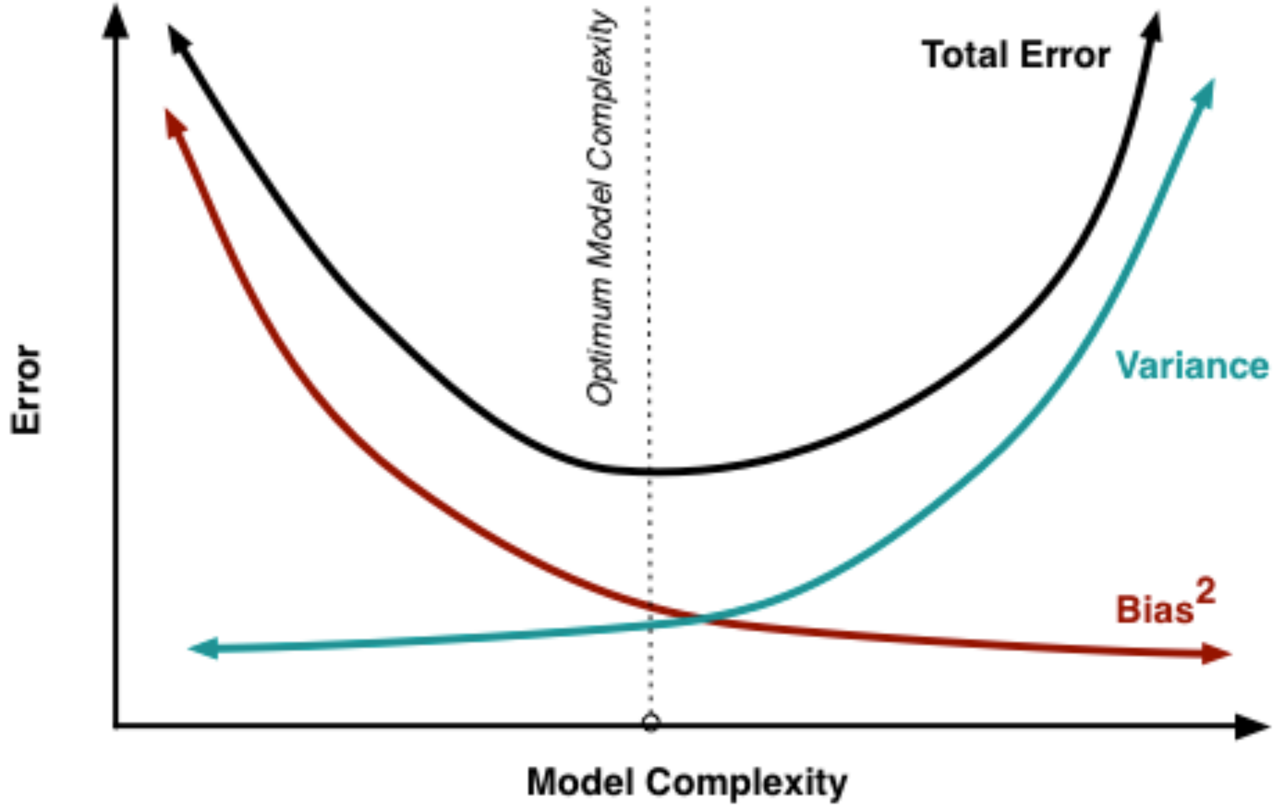
HOW MANY FOLDS SHOULD WE CHOOSE?

- A high number of folds results in what?

- A low number of folds results in what?

HOW MANY FOLDS SHOULD WE CHOOSE?

- **With a large number of folds:**
- Error due to bias is low
- Variance is quite high
- Computationally expensive
- **With a low number of folds:**
- Error due to variance is low
- The error due to bias will be large
- Computationally cheaper
- **Thus…**
- For large datasets, k=3 typically ok
- Sparse datasets, LOOCV



THREE WAY DATA SPLITS

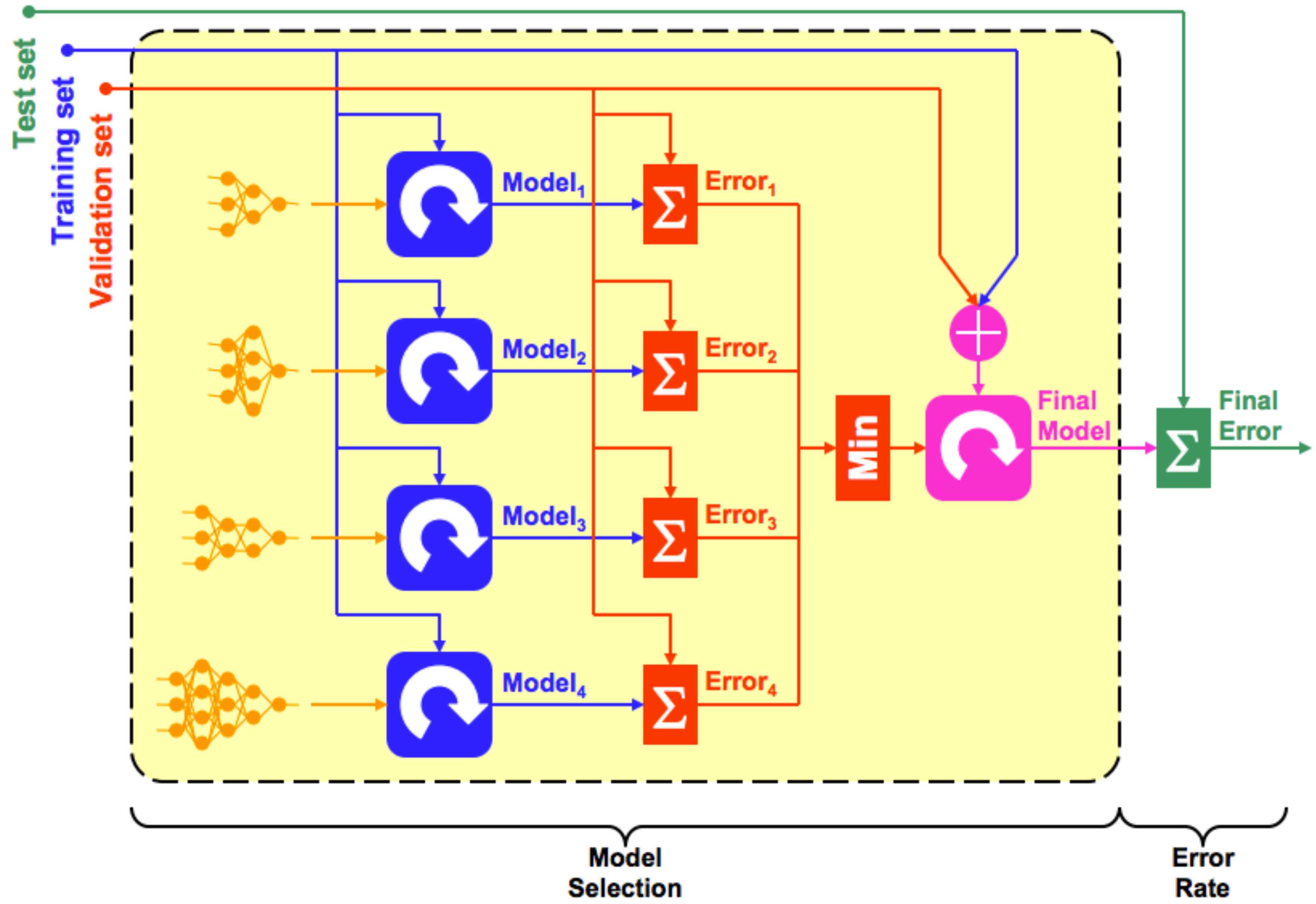
- If model selection and true error estimates are to be computed simultaneously, three disjoint data sets are best.
- **Training set:** a set of example used for learning – what parameters of the classifier
- **Validation set**: a set of examples used to tune the parameters of the classifier
- **Testing set:** a set of examples used ONLY to assess the performance of the fully-trained classifier

- **Validation and testing must be separate data sets.** Once you have the final model set, you cannot do any additional tuning after testing.

PROCEDURE

- 1. Divide data into training, validation, testing sets
- 2. Select architecture (model type) and training parameters (k)
- 3. Train the model using the training set
- 4. Evaluate the model using the training set
- 5. Repeat 2-4 selecting different architectures (models) and tuning parameters
- 6. Select the best model
- 7. Assess the model with the final testing set

PROCEDURE


Test set
Training set
Validation set
Model$_1$
Model$_2$
Model$_3$
Model$_4$
Error$_1$
Error$_2$
Error$_3$
Error$_4$
Min
Final Model
Final Error
Model Selection
Error Rate

PARTING QUESTIONS

- The demo covers a basic test/train split as well as k-fold cross-validation Check: Is 2-fold cross-validation the same as a 50:50 test/train split?

- Will two different 50:50 (or x:y) splits produce the same model score?

ADDITIONAL RESOURCES

https://www.youtube.com/watch?v=_2ij6eaaSl0&t=2m34s

http://www.win-vector.com/blog/2015/01/random-testtrain-split-is-not-always-enough/

https://en.wikipedia.org/wiki/Cross-validation_(statistics)