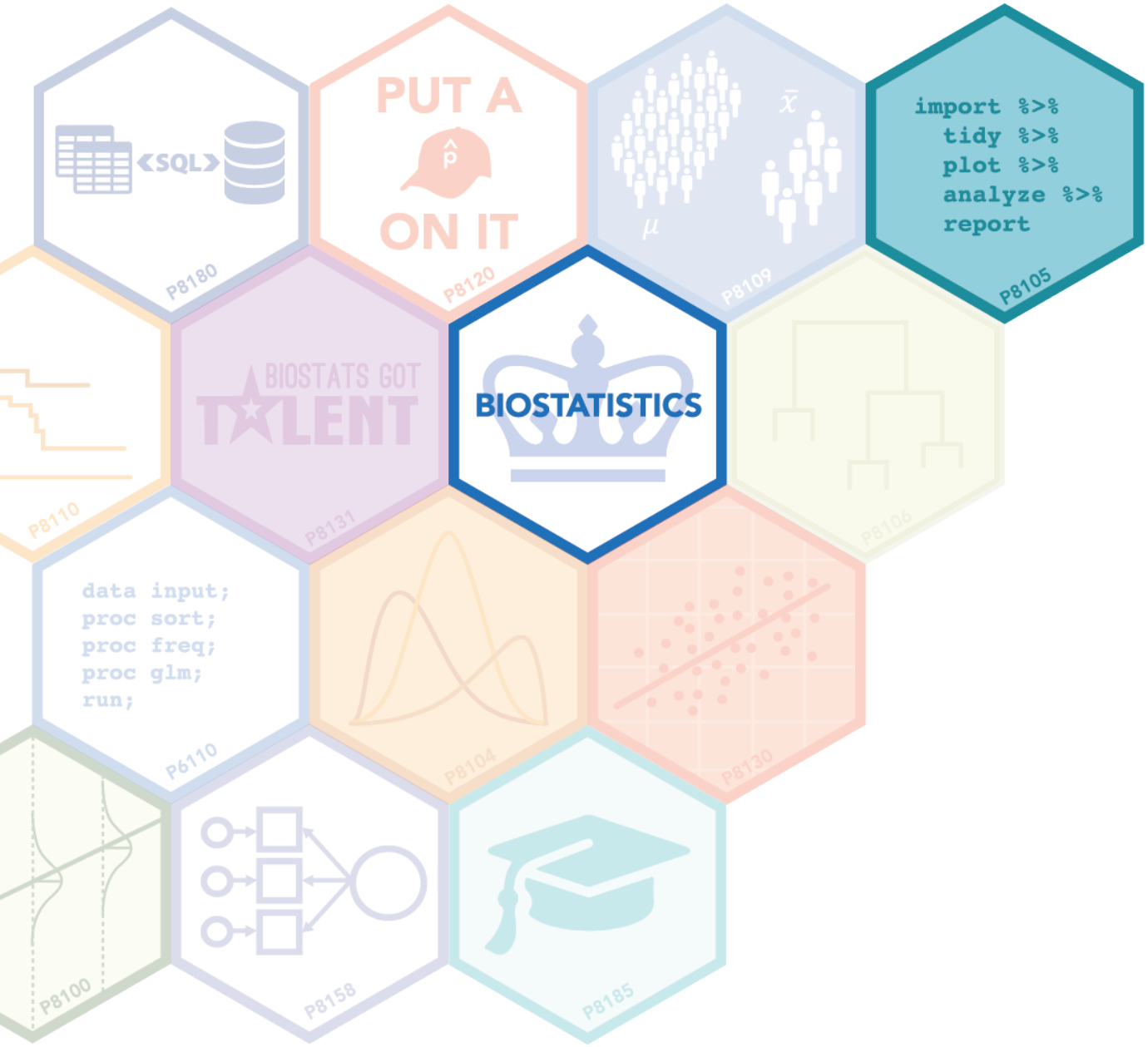


CROSS VALIDATION

Jeff Goldsmith, PhD
Department of Biostatistics

Model selection

- When you have lots of possible variables, you have you choose which ones will go in your model
- In the best case, you have a clear hypothesis you want to test in the context of known confounders
- (Always keep in mind that no model is “true”)

Model selection is hard

- Lots of times you're not in the best case, but still have to do something
- This isn't an easy thing to do

- For nested models, you have tests
  - You have to be worried about multiple comparisons and "fishing"

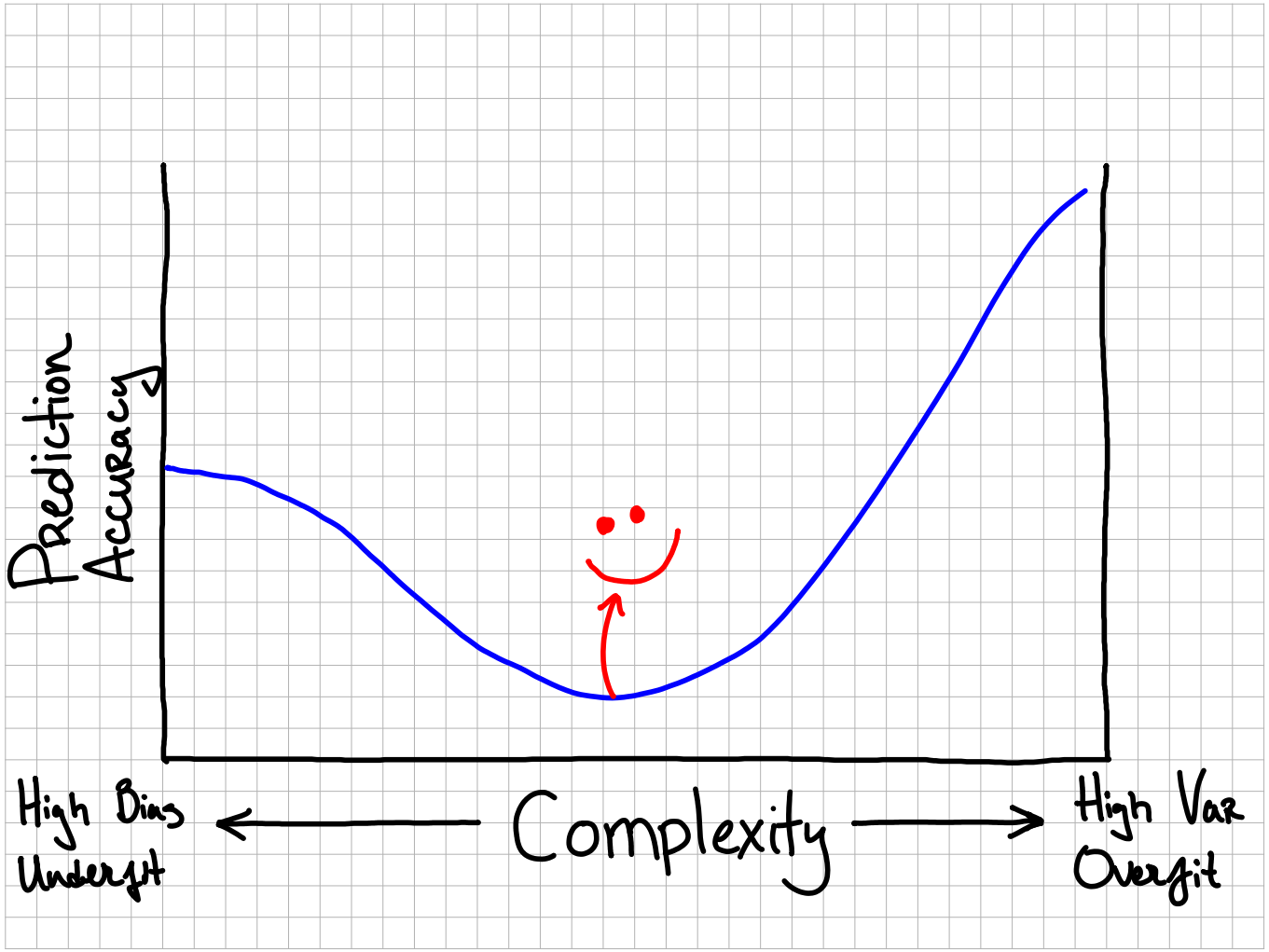
- For non-nested models, you don't have tests
  - AIC / BIC / etc are traditional tools
  - Balance goodness of fit with "complexity"

Questioning fit

- These are basically the same question:
  - Is my model not complex enough? Too complex?
  - Am I underfitting? Overfitting?
  - Do I have high bias? High variance?

- Another way to think of this is out-of-sample goodness of fit:
  - Will my model generalize to future datasets?

Flexibility vs fit



Prediction accuracy

- Ideally, you could
  - Build your model given a dataset
  - Go out and get new data
  - Confirm that your model “works” for the new data
- That doesn’t really happen
- So maybe just act like it does?

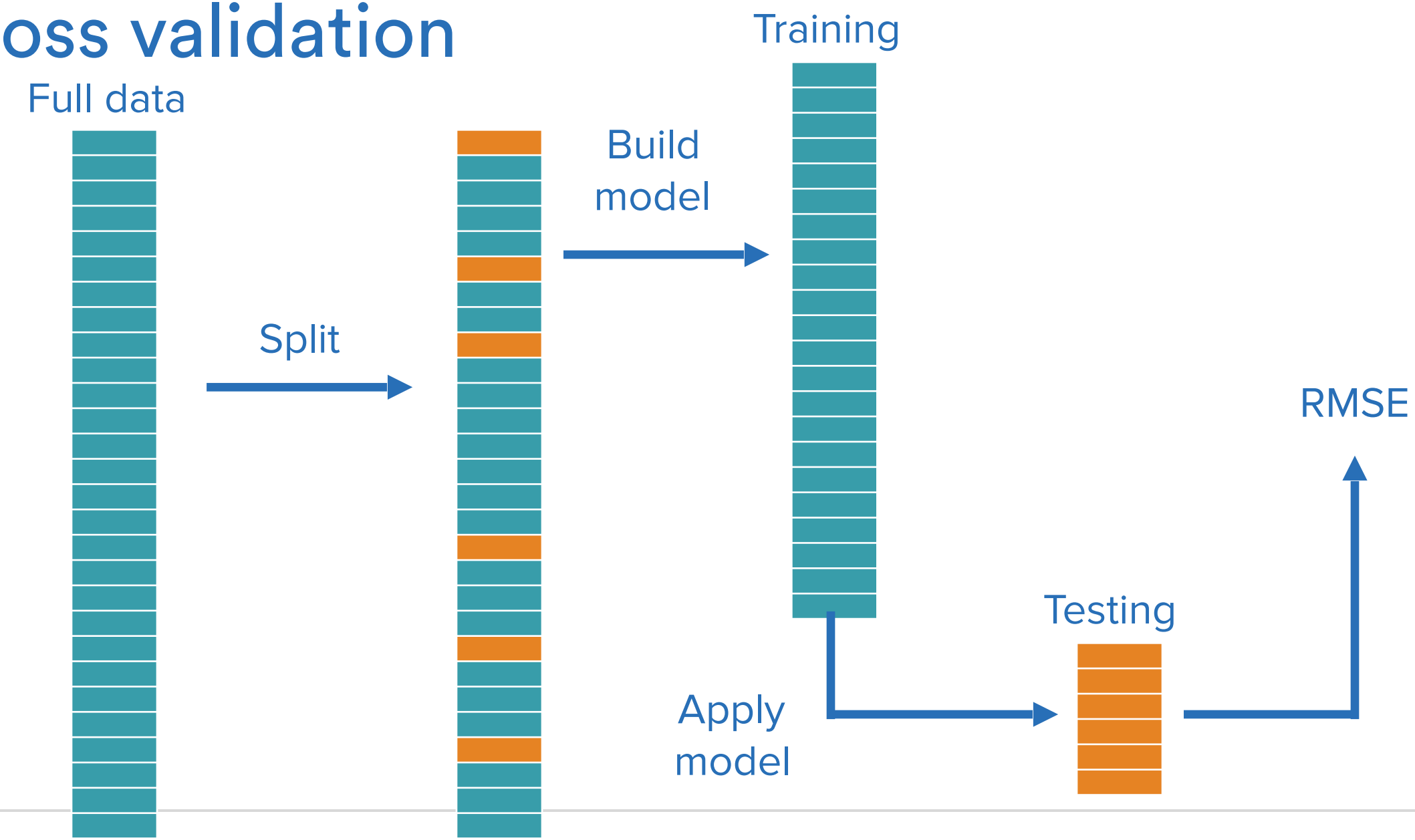
Cross validation

- Randomly split your data into “training” and “testing”
  - “Training” is data you use to build your model
  - “Testing” is data you use to evaluate out-of-sample fit
  - Exact ratio depends on data size, but I like 80 / 20

- Evaluate using root mean squared error :

$$RMSE = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n}}$$

Cross validation



Refinements and variations

- Individual training / testing splits are subject to randomness
- Repeating the process
  - Illustrates variability in prediction accuracy
  - Can indicate whether differences in models are consistent across splits
- I usually repeat the training / testing split

- Folding (5-fold, 10-fold, k-fold, LOOCV) partitions data into equally-sized subsets
  - One fold is used as testing, with remaining folds as training
  - Repeated for each fold as testing
- I don’t do this as often

Cross validation is general

- Can use to compare candidate models that are all “traditional”

- Comes up a lot in “modern” methods
  - Automated variable selection (e.g. lasso)
  - Additive models
  - Regression trees

Prediction as a goal

- In the best case, you have a clear hypothesis you want to test in the context of known confounders
  - I know I already said this, but it's important

- Prediction accuracy matters as well
  - Different goal than statistical significance
  - Models that make poor predictions probably don't adequately describe the data generating mechanism, and that's bad

Tools for CV

- Lots of helpful functions in `modelr`
  - `add_predictions()` and `add_residuals()`
  - `rmse()`
  - `crossv_mc()`

- Since repeating the process can help, list columns and `map` come in handy a lot too :-)