Myths of Data Science

Things You Should and Should Not Believe

Nina Zumel
Win-Vector LLC
http://www.win-vector.com/

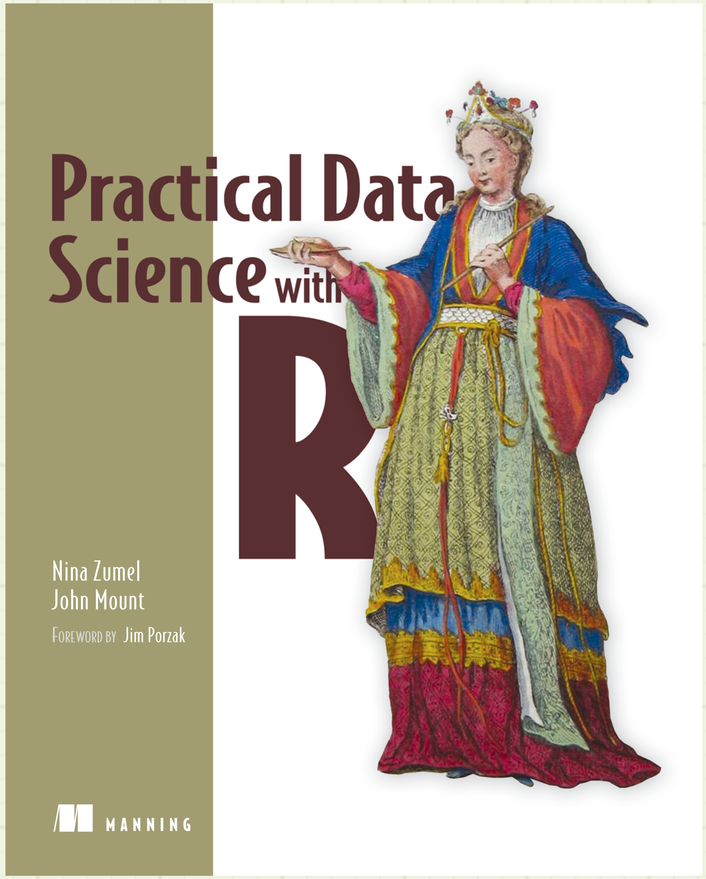
ODSC West 2017

#ODSC

Win-Vector LLC

Who I am

- Nina Zumel
- Principal Consultant at Win-Vector LLC
- One of the authors of *Practical Data Science with R*
- Frequent contributor to the *Win-Vector Blog*



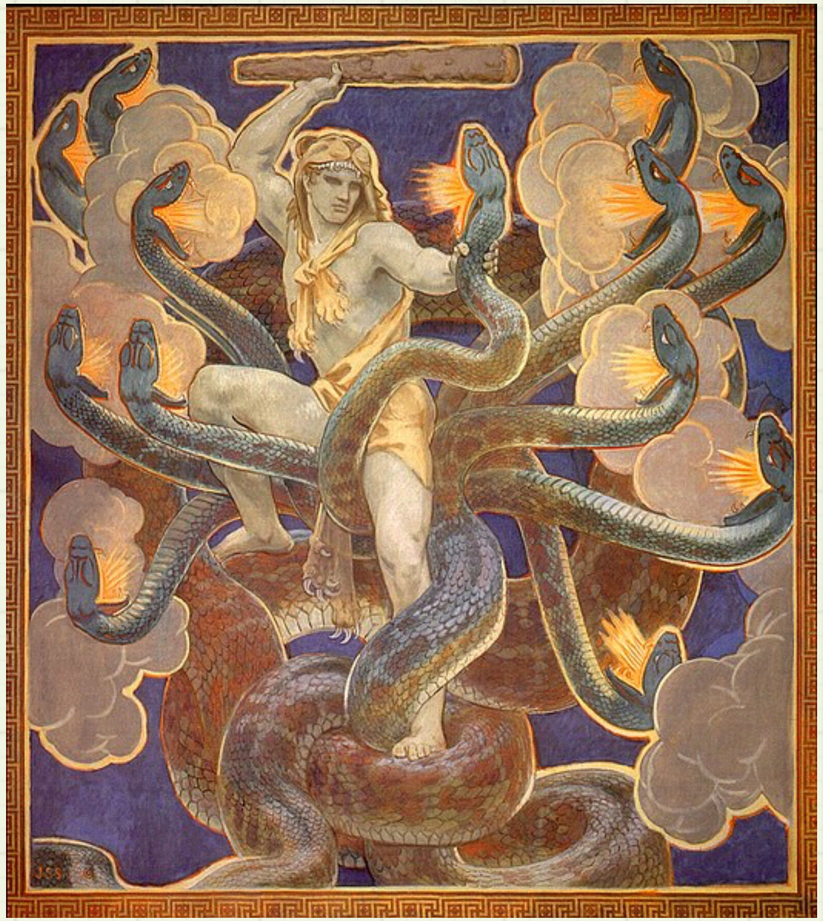
Win-Vector LLC

Goal of Talk

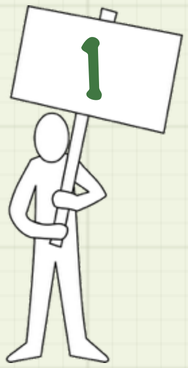
To give an overview of statistical/ ML ~~myths~~ "factoids"

- Things we (unconsciously) believe or that others are apt to tell us
- Often generally true — until they aren't

- Myths about Predictive Modeling
- Myths about Controlled Experiments



Myths about Predictive Modeling



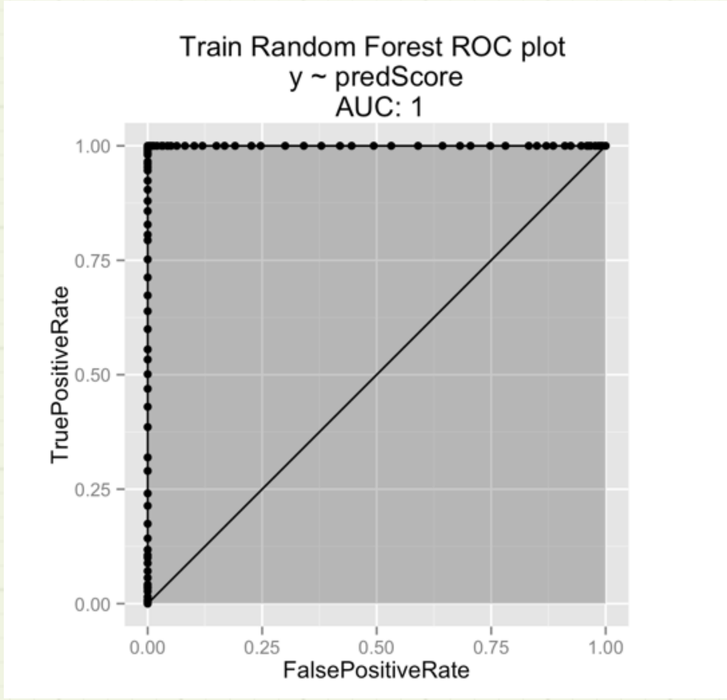
MYTH

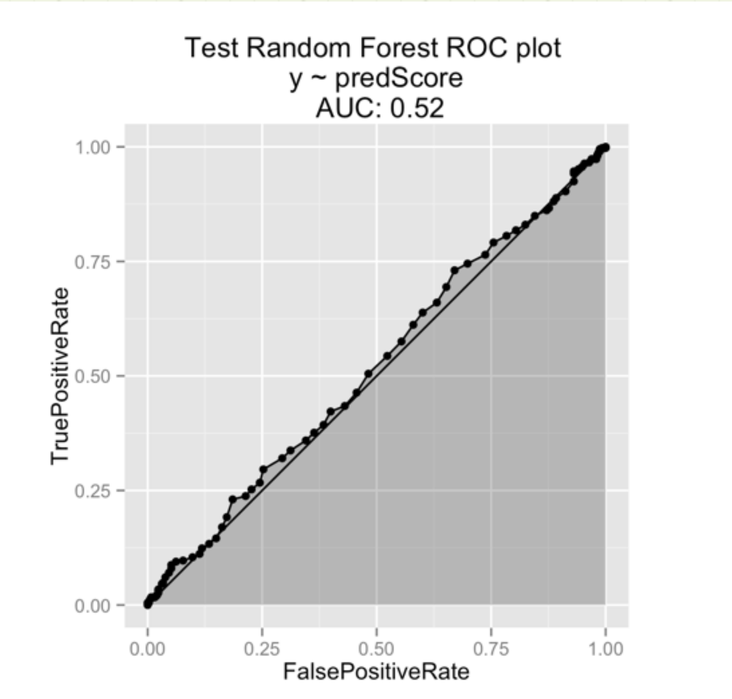
"More variables are better."

"Throw in all the variables you can think of and let the model sort it out"

Counterexample

- 400 variables, 800 rows - no signal
  - Near perfect fit on training data.
  - Model performs like random guessing on new instances.
  - Extreme over fit.





The problem is the wide data, not the learning method

Model	AUC (train)	AUC (test)
Naive Bayes	0.856	0.515
GLM	0.791	0.488
Random Forest	1.000	0.458
SVM	0.976	0.507
Elastic Net	1.000	0.477
Gradient Boosting	0.852	0.511

Fix #1: More data

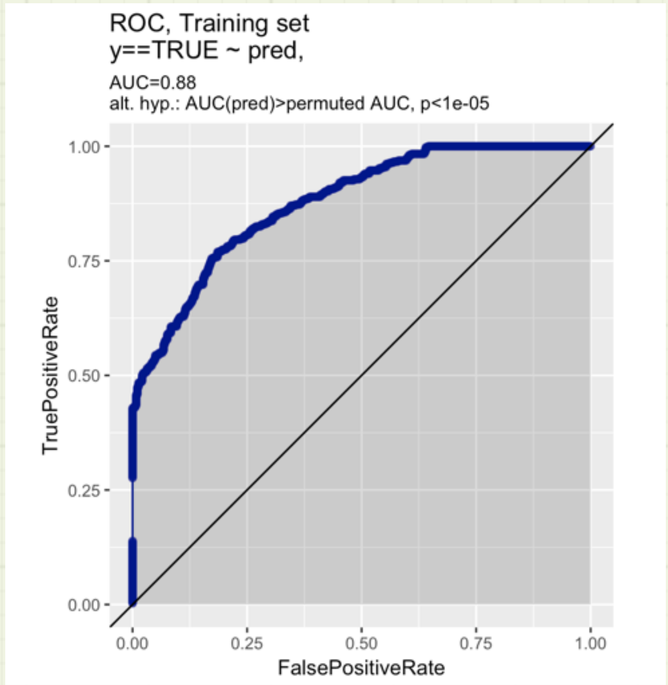
- At least 3 rows/variable
  - More complex models: probably higher ratio
- Categorical variables: Count the levels!!
  - Sneaky source of underconstrained systems

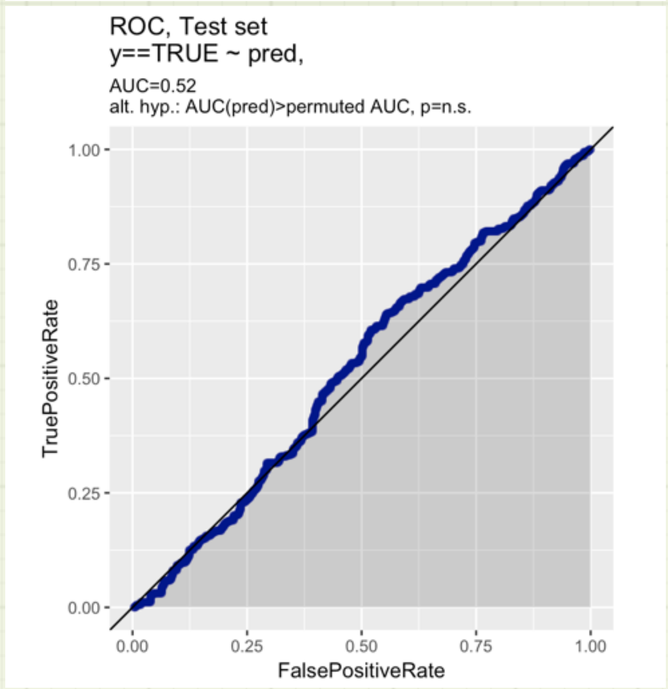
Image: Boris Artzybasheff.

"One" variable overfit

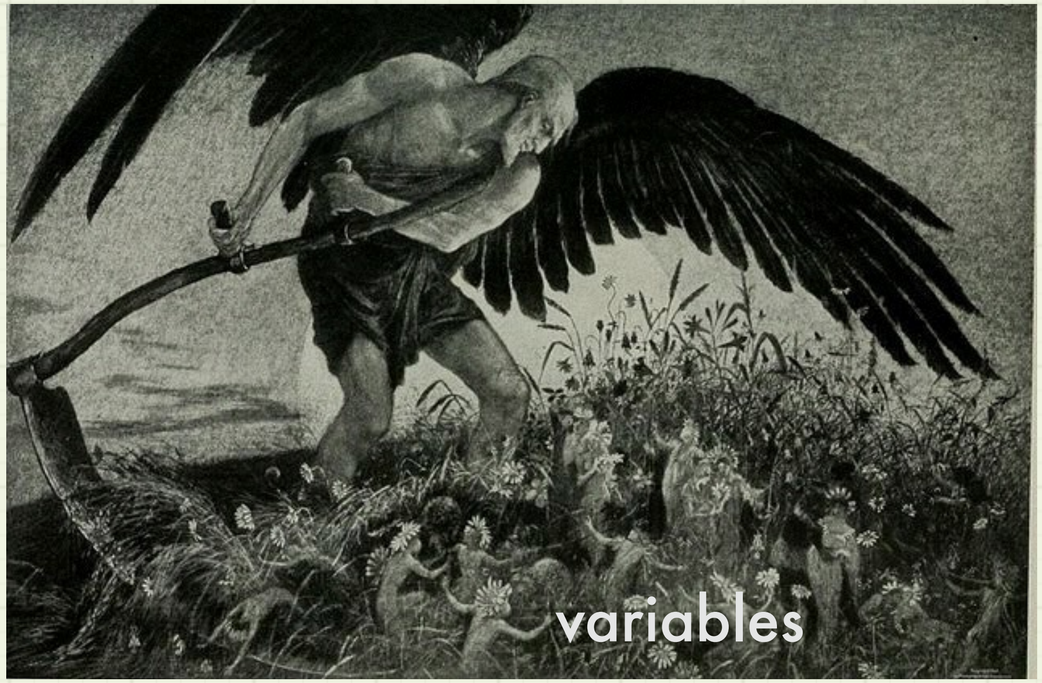
- One categorical variable with 400 levels
- 800 rows - no signal

Dataset	AUC
training	0.878
test	0.52





Fix #2: Variable Pruning



Domain Knowledge

- Identify key variables
- Eliminate irrelevant, redundant ones

Random Forest variable importance

- I like permutation measure

Significance Pruning

- Check significance of one-variable models
  - p = 1/nvar to prune

http://www.win-vector.com/blog/2015/08/how-do-you-know-if-your-data-has-signal/

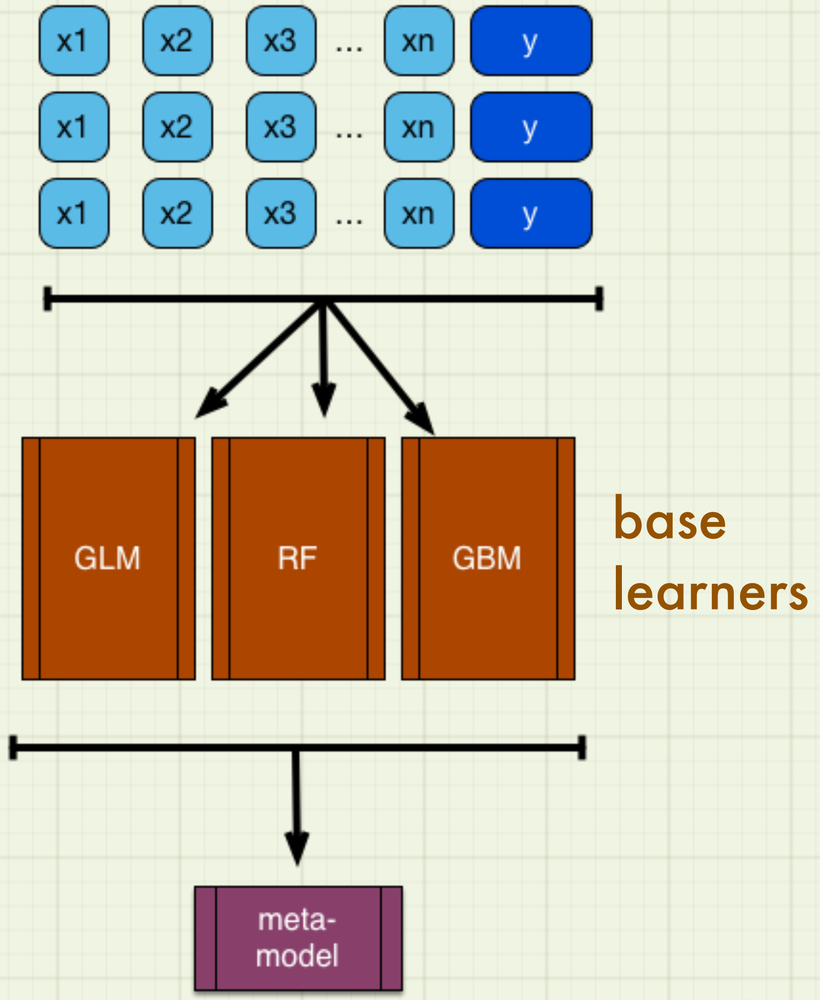
MYTH

"I can combine multiple models to get a better model!"

AKA Stacked Ensembles

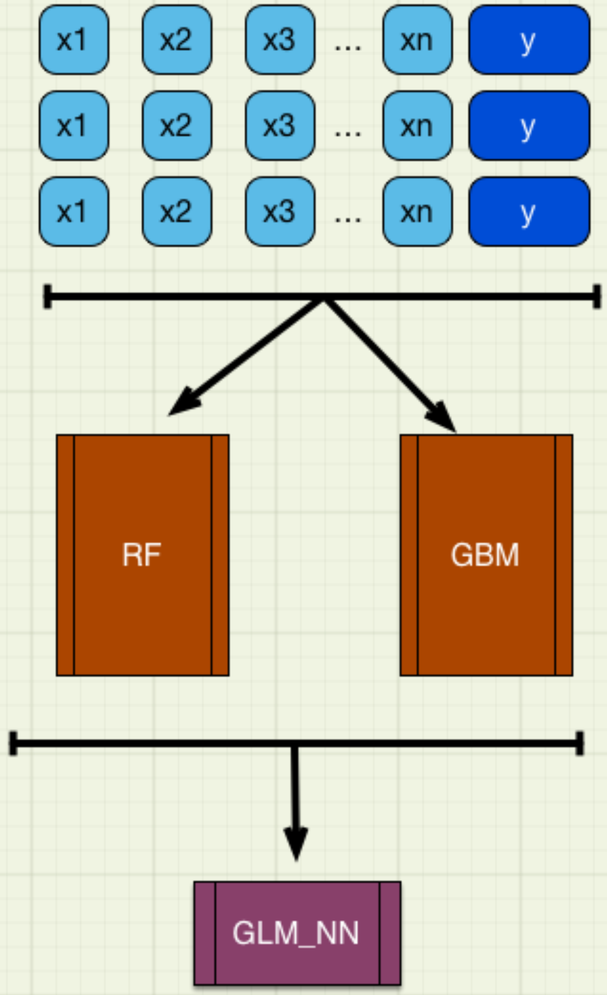
Stacked Models (Naively)

- Use training data to train the base learners
- Combine base learner predictions with a meta-model



Example

- Particle data: distinguish a process that produces Higgs bosons from one that does not.
  - 28 features
  - Training: 10K rows, Test: 5K rows
- Stacked model:
  - Base models: Gradient boosted trees, random forest
  - Metalearner: Logistic regression w/ non-negative coefficients
  - H2O implementations



Results on Holdout Data

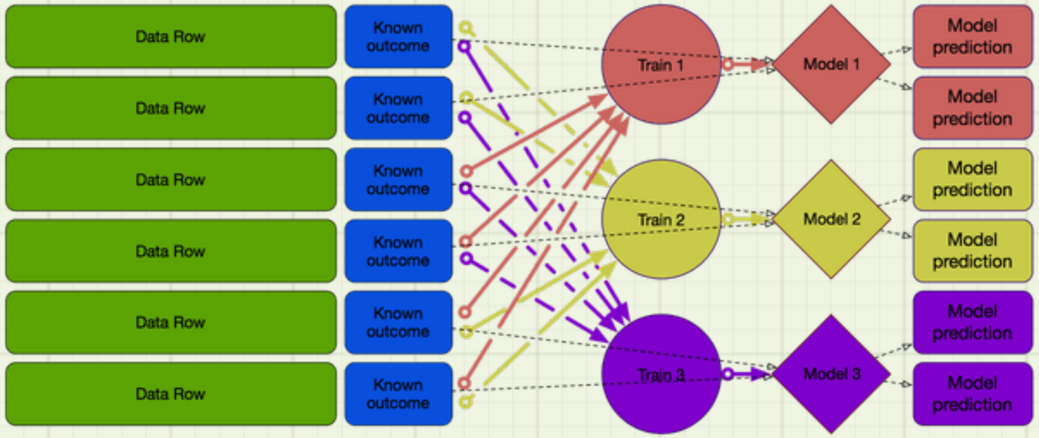
- (smaller is better)
- Ensemble did worse than either base learner!

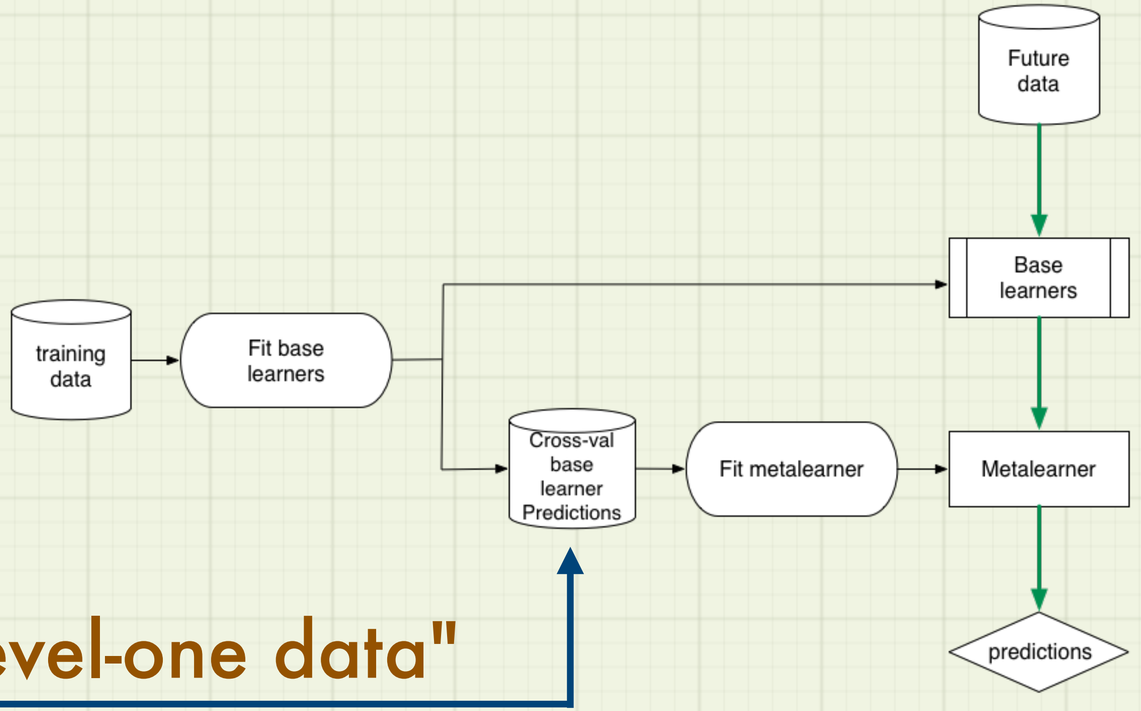
Model	Deviance
GBM	5991.608
RF	5772.693
Naive Ensemble	7762.197

Nested Model Bias

- Base models and metalearner trained on same data
- High complexity base models can memorize training data — look "too good" to metalearner.
- Metalearner learns wrong combination of base learners
    - Holdout data: Overfit base model does not predict as well, so metalearner performs poorly.

Solution: Use Different Datasets (or simulate it)





Wolpert, 1992; Breiman, 1996; van der Laan, 2007

Results on Holdout Data

Model	Deviance
GBM	5991.608
RF	5772.693
Naive Ensemble	7762.197
Super Learner (5-fold cross-val)	5671.564

h2o.stackedEnsemble

Caution: Nested Models are Everywhere!

Any pre-model-fitting task that uses knowledge of outcome is a nested model

- Variable treatment
- Hyperparameter tuning
- Variable selection/stepwise methods
- Y-aware dimension reduction
- Empirical Bayes

MYTH

"Models with more expressive concept spaces are better."

"Sophisticated models are better than simpler ones."

Stacking Continued

- "Linear combination was a good metalearner — Random Forest should be better, right?"
- Wrong

Model	Deviance
GBM	5991.608
RF	5772.693
Naive Ensemble	7762.197
Super Learner (5-fold cross-val)	5671.564
Complex Ensemble	6940.308

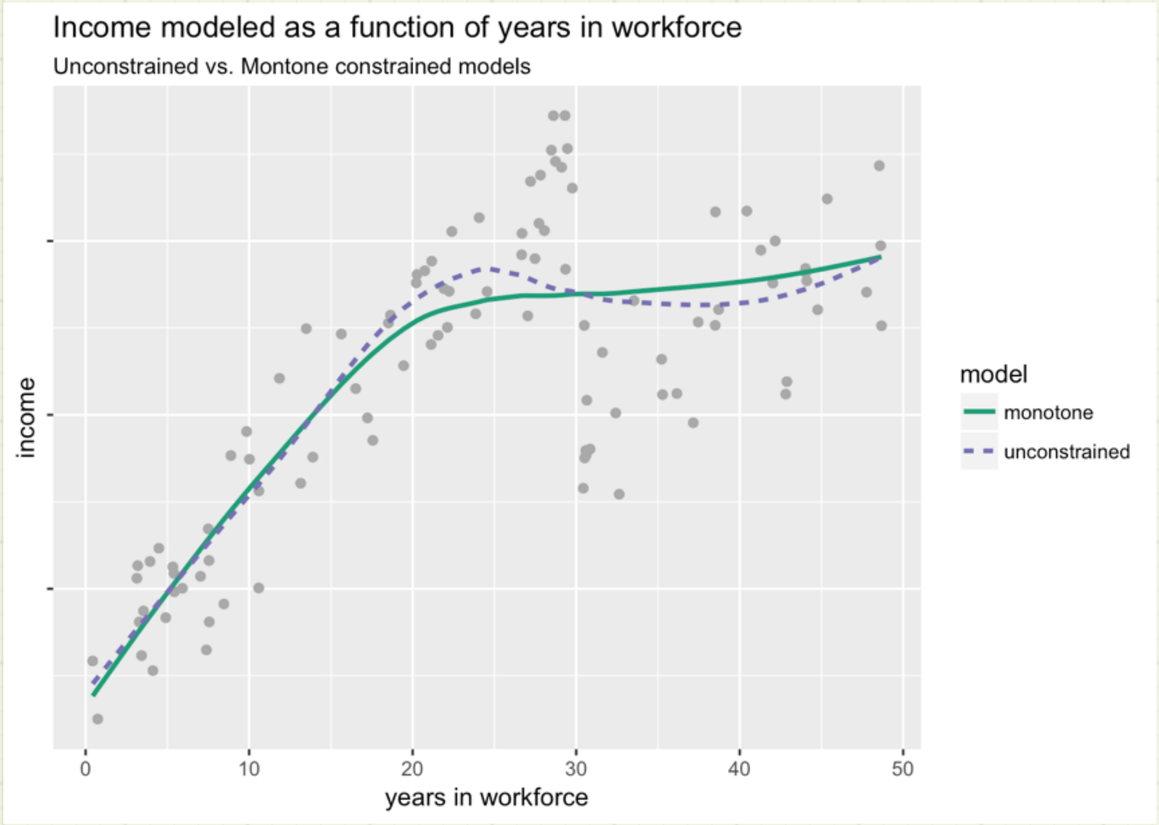
What Happened?

High-complexity models tend to "explain" unexplainable variance

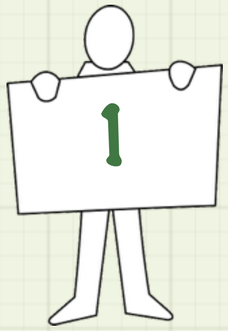
- Need a lot of data to overcome that
- Ideal: Rich enough to learn target concept, but no richer

Strong inductive biases can produce better models in appropriate situations

- Non-negative linear combination
- Monotone (Isotonic) regression
- Multilevel/hierarchical models



Myths about Controlled Experiments



MYTH

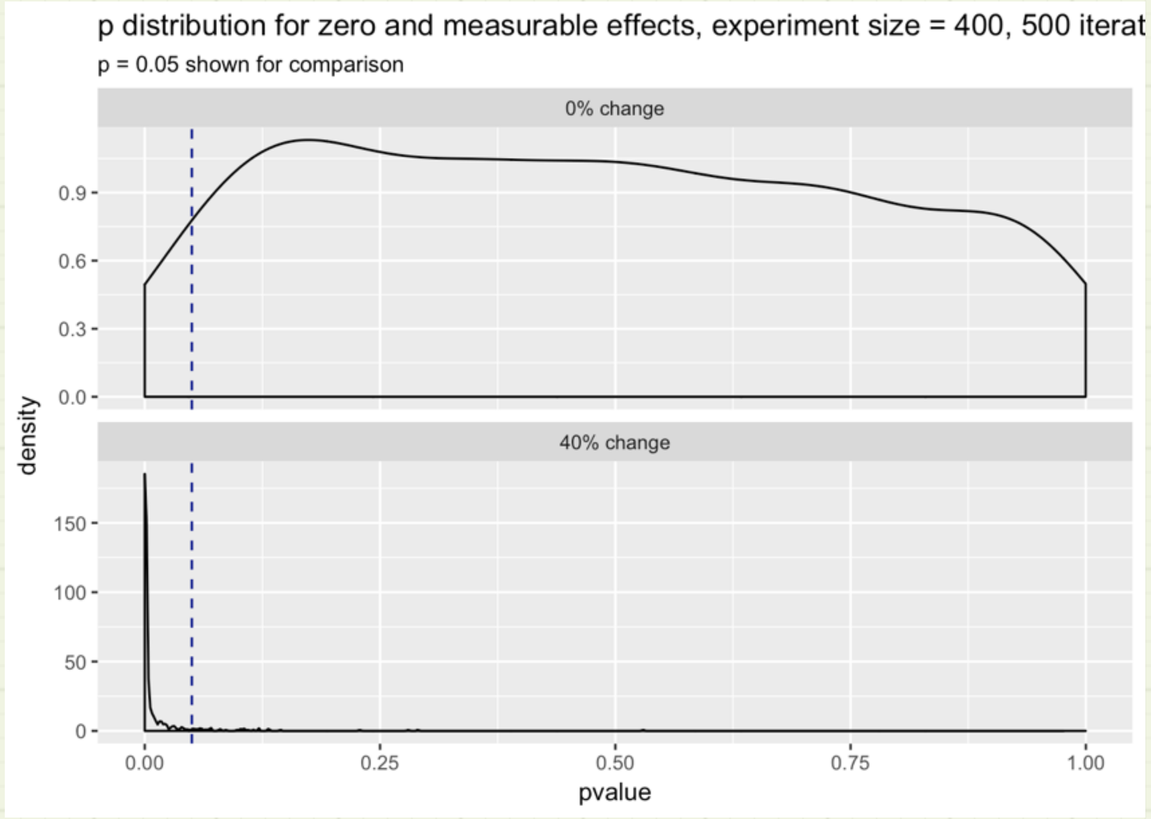
"If there's an effect, *p* is small.
If there's no effect, *p* is large."

Example: Weight Loss Program

- Compare new program to old one
  - Trial with 400 subjects
  - Measure pounds lost by end of trial.
- Two Scenarios:
  - New program is 40% better
  - New program no different from old
  - Simulate trials for each scenario 500 times to get a distribution of results (p-values).

p-value Behavior is Inconsistent

- If there is an effect:
  - $p \to 0$ as $n \uparrow$ (fixed effect size)
  - $p \to 0$ as effect size $\uparrow$ (fixed n)
- But if there is no effect
  - ~~$p \nrightarrow 1$ as $n \uparrow$~~
  - ***p* is uniformly distributed independent of n**
- So if 100 researchers run the same (no-effect) experiment, they don't all get the same result
  - At $p \leq 0.05$, 5 will see a positive result

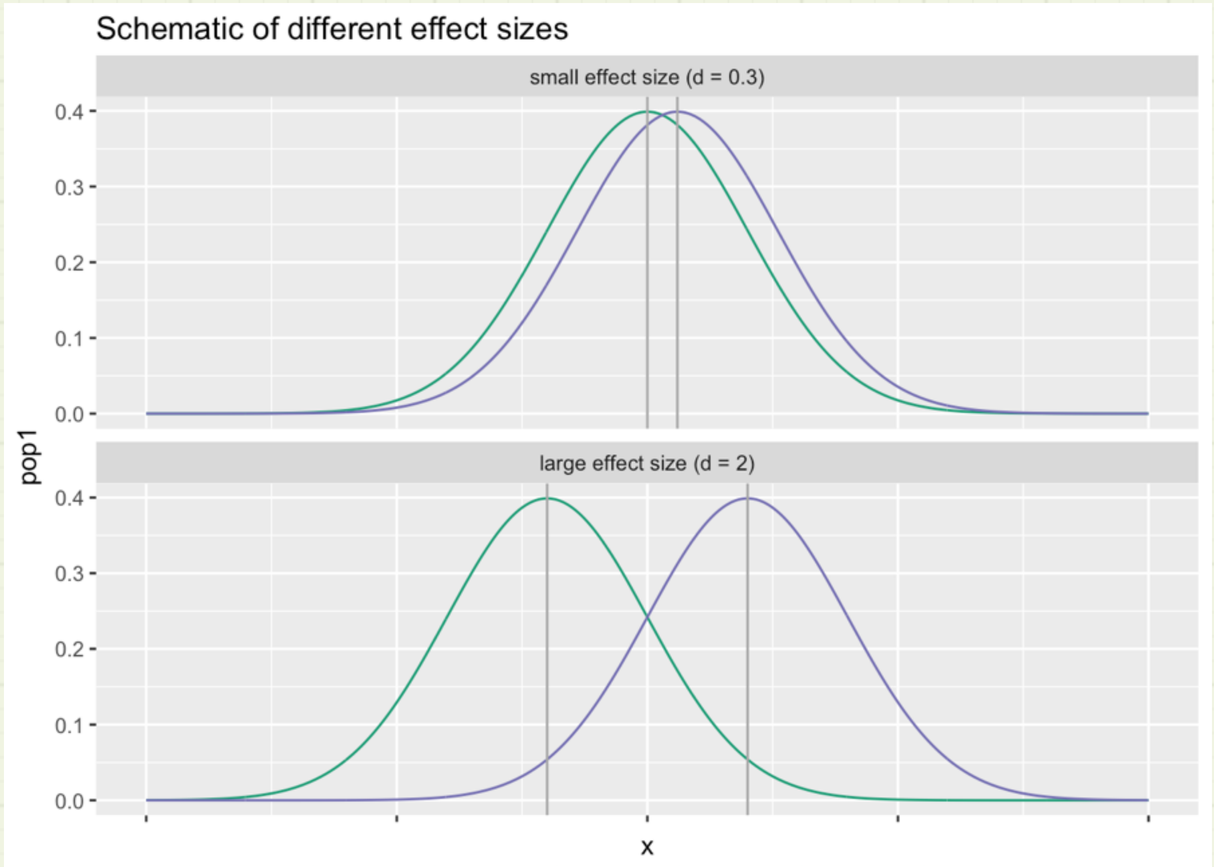


Cohen's *d* is consistent

$$d = (\mu_1 - \mu_2)/\sigma_{pooled}$$

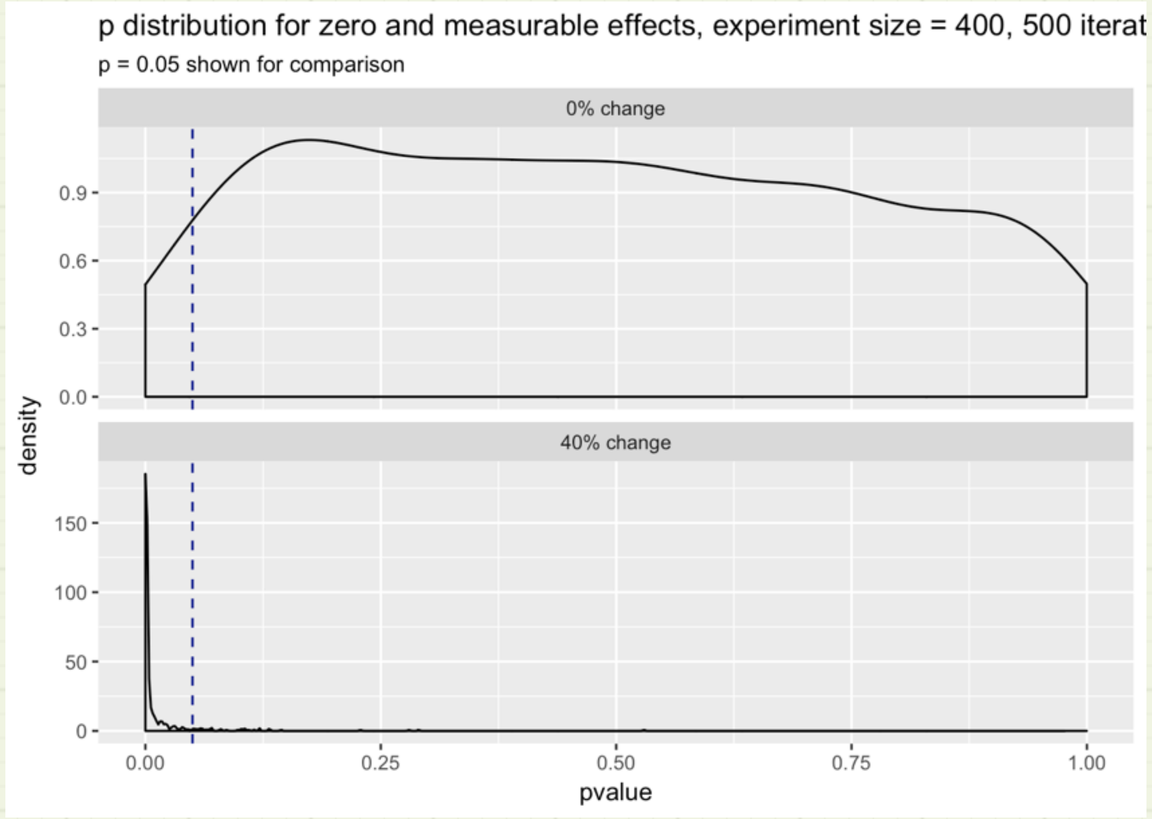
- σ_{pooled} = pooled standard deviation of populations

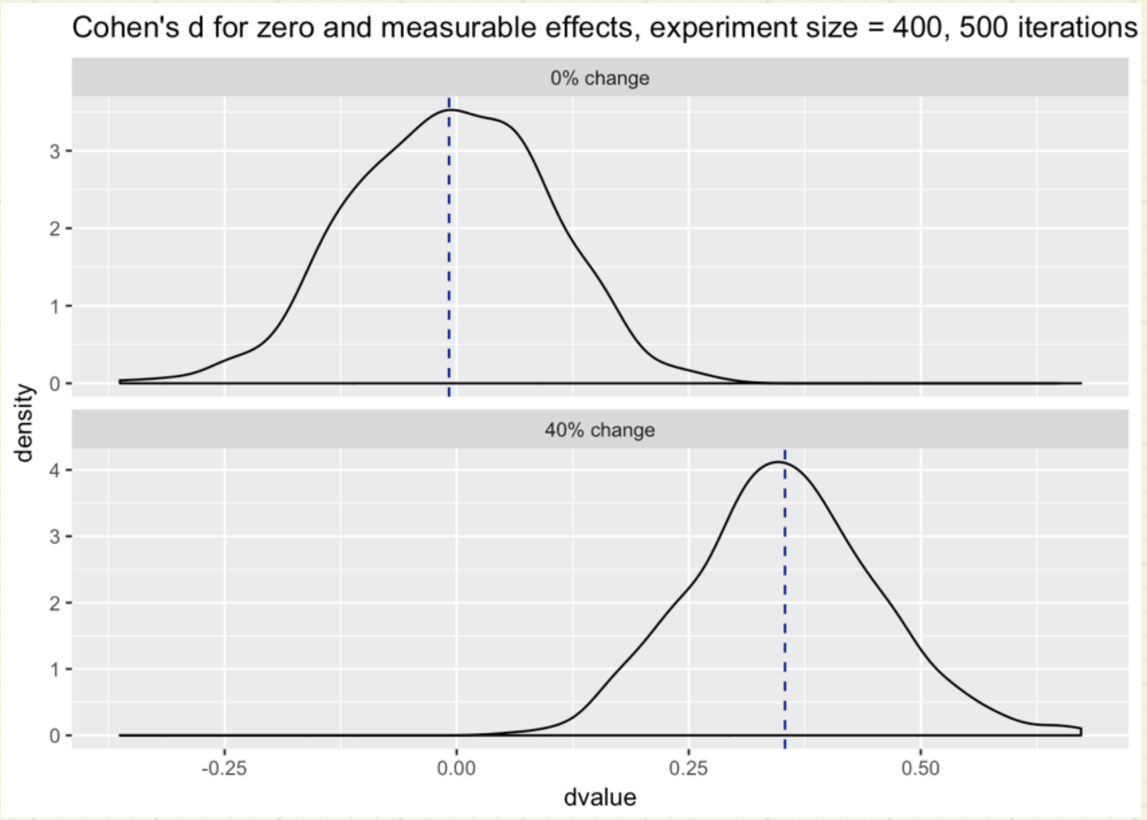
- Traditionally used to plan experiment size
- No effect: $d \rightarrow 0$ as $n \uparrow$
- Effect: $d \rightarrow D$ as $n \uparrow$
- So if 100 researchers run the same (well-designed) study, they should all get similar *d*s
- Comparing rates: Cohen's *h*



Win-Vector LLC

p vs. *d*





Win-Vector LLC

MYTH

"Early Stopping of an A/B test is Free"

Example:
Preference bias by side

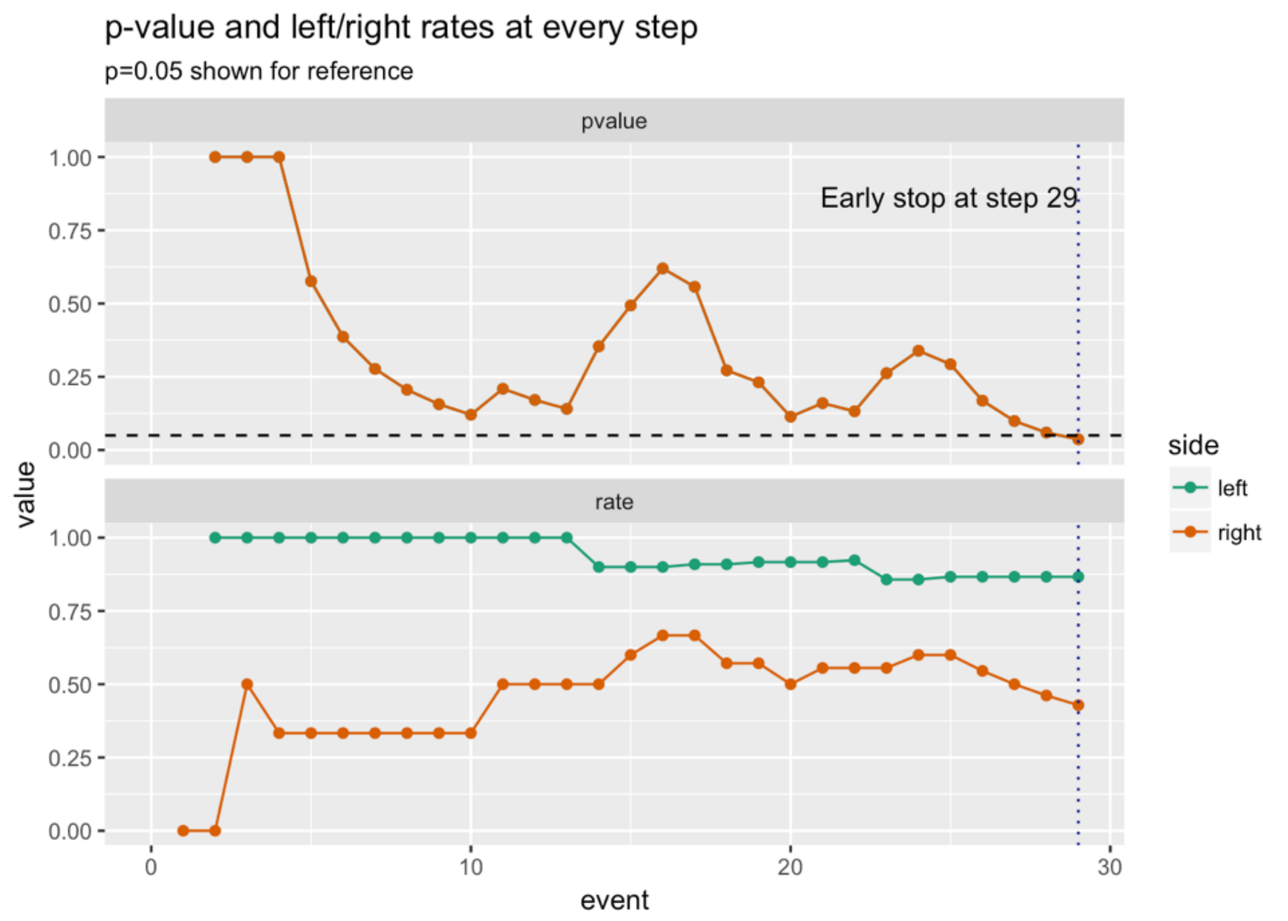
Movie streaming service

- Show two trailers side by side, user picks one to view
- Theory: movies picked more often when trailer is on the right
- Should you pay to place your trailers on the right?

Photo: Miguel Pires da Rosa, Wikipedia

A/B test with early stopping

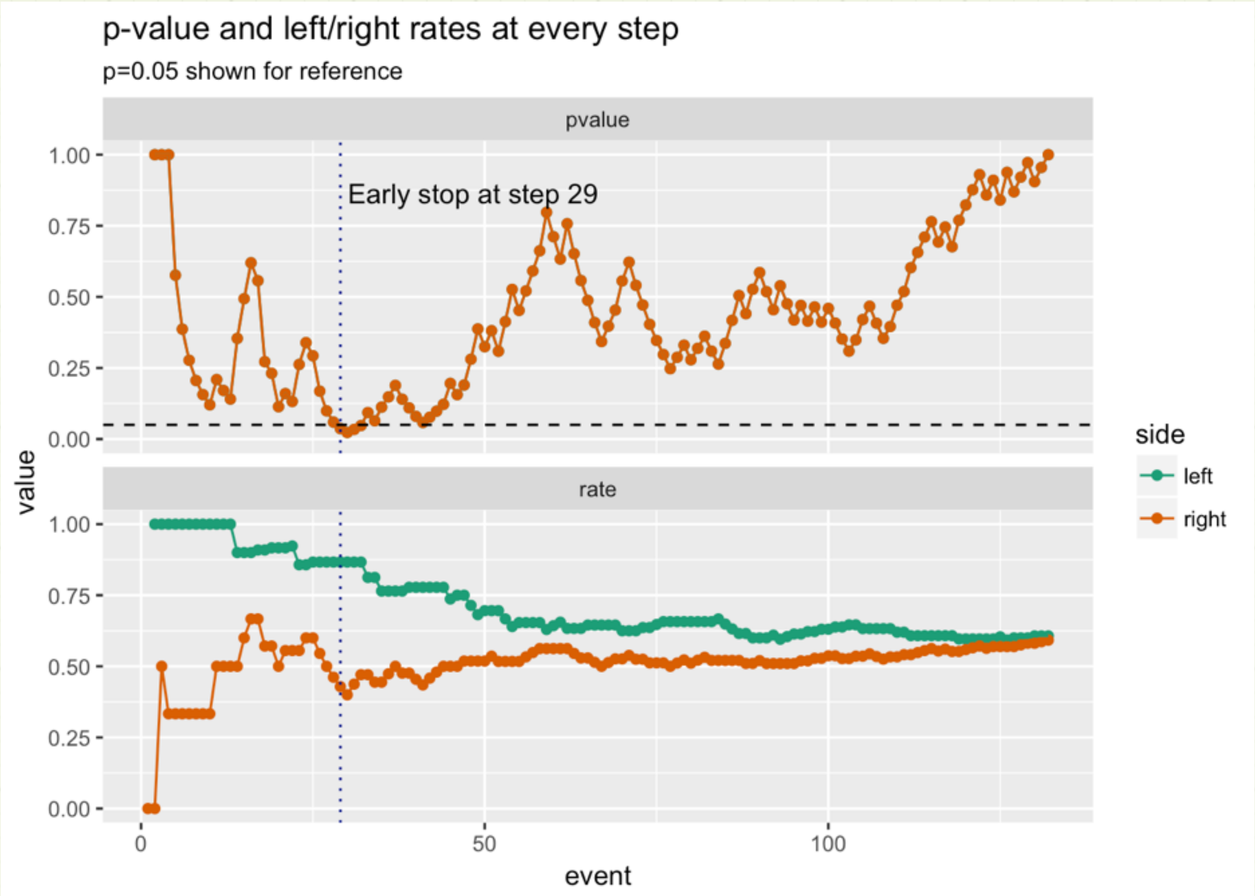
"Every time our movie comes up, calculate left side and right side conversion. Stop when the difference looks significant, and take the winner."



Sometimes the "wrong" side gets lucky (for a while)

Position	Selection rate
left	0.5669
right	0.5835

- In actuality: 1.7% difference
  - after 58,000 views
- ~30,000 views to get p=0.05 significance



Myth

"So, don't ever early stop."

Sequential Analysis

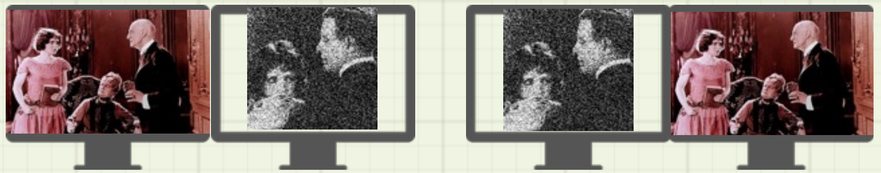
- Wald (1947)
- Stopping rule fixed, experiment size variable
  - "Run until you see a difference of 30 conversions"
- *Sometimes* stops sooner than a non-sequential test
- Comparing two binomial processes:
  - Process efficiencies $k = p/(1-p)$, p = conversion rate

Comparing Two Binomial Processes

Champion/Challenger Model

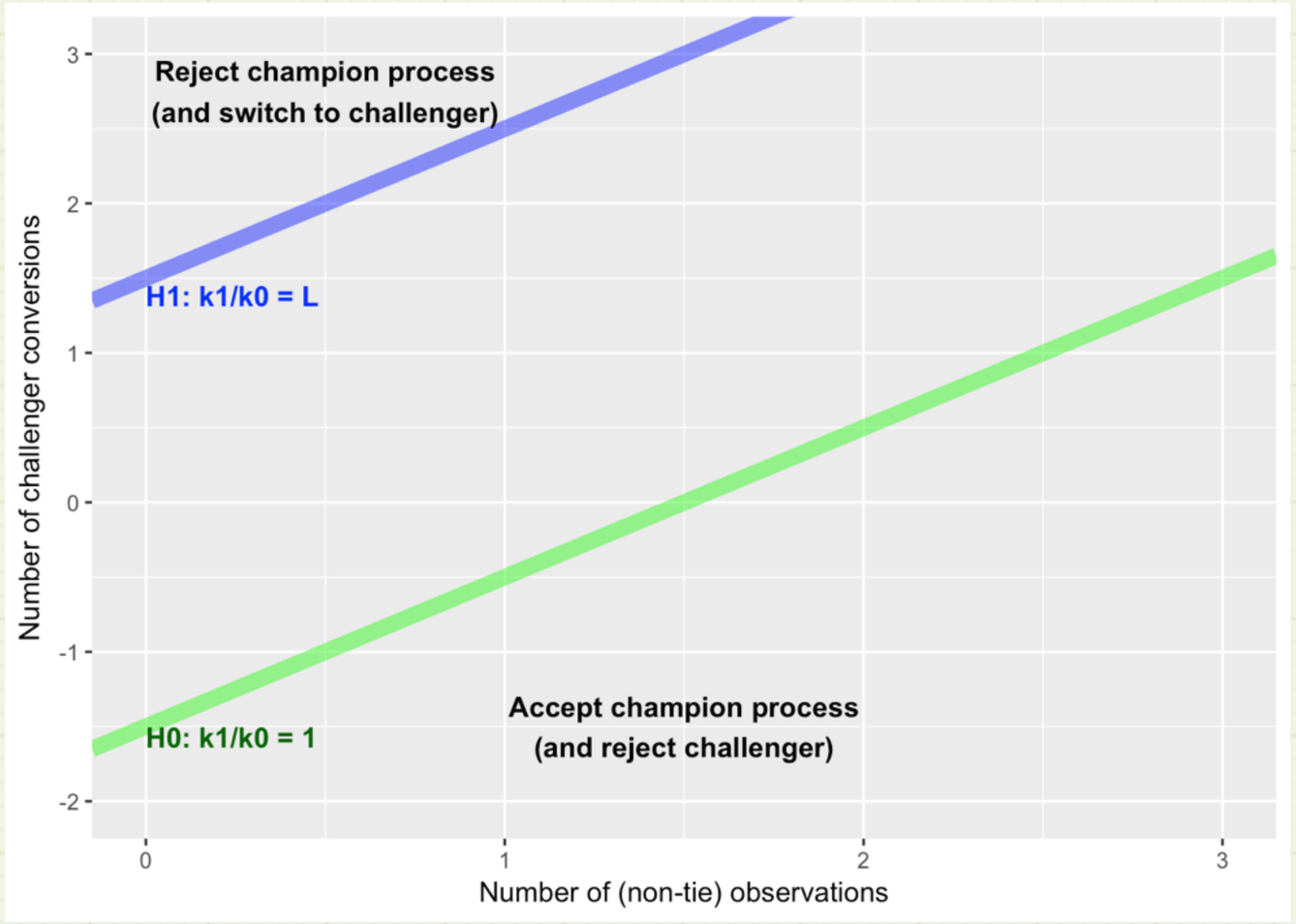
- Champion : Existing Process (P_0)
- Challenger: Process we might switch to (P_1)
- Paired comparisons
- Test process efficiency ratio: k_1/k_0
  - Below ratio H_0: keep P_0
    - $H_0 = 1$ a reasonable choice
    - "Keep P_0 if it looks better."
  - Above ratio H_1: switch to P_1
  - Between H_0 and H_1: keep testing

Process 0 (Champion)	Process 1 (Challenger)
TRUE	FALSE
TRUE	TRUE
FALSE	TRUE
FALSE	FALSE
FALSE	TRUE
TRUE	FALSE



The Idea

Process 0 (Champion)	Process 1 (Challenger)
TRUE	FALSE
TRUE	TRUE
FALSE	TRUE
FALSE	FALSE
FALSE	TRUE
TRUE	FALSE

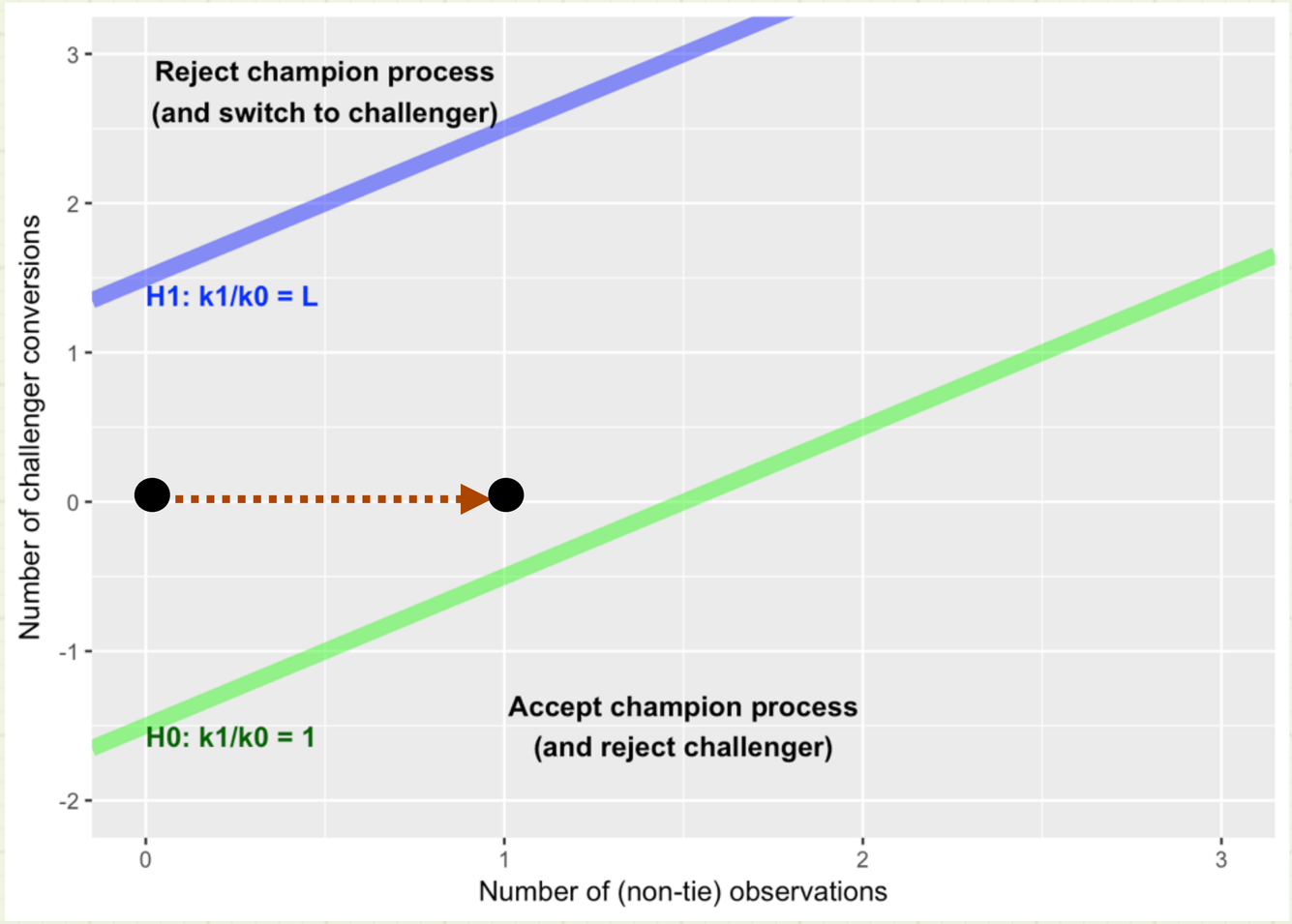


The Idea

Process 0 (Champion)	Process 1 (Challenger)
TRUE	FALSE
TRUE	TRUE
FALSE	TRUE
FALSE	FALSE
FALSE	TRUE
TRUE	FALSE

Start at (0,0)

P_0 converts: move one to the right

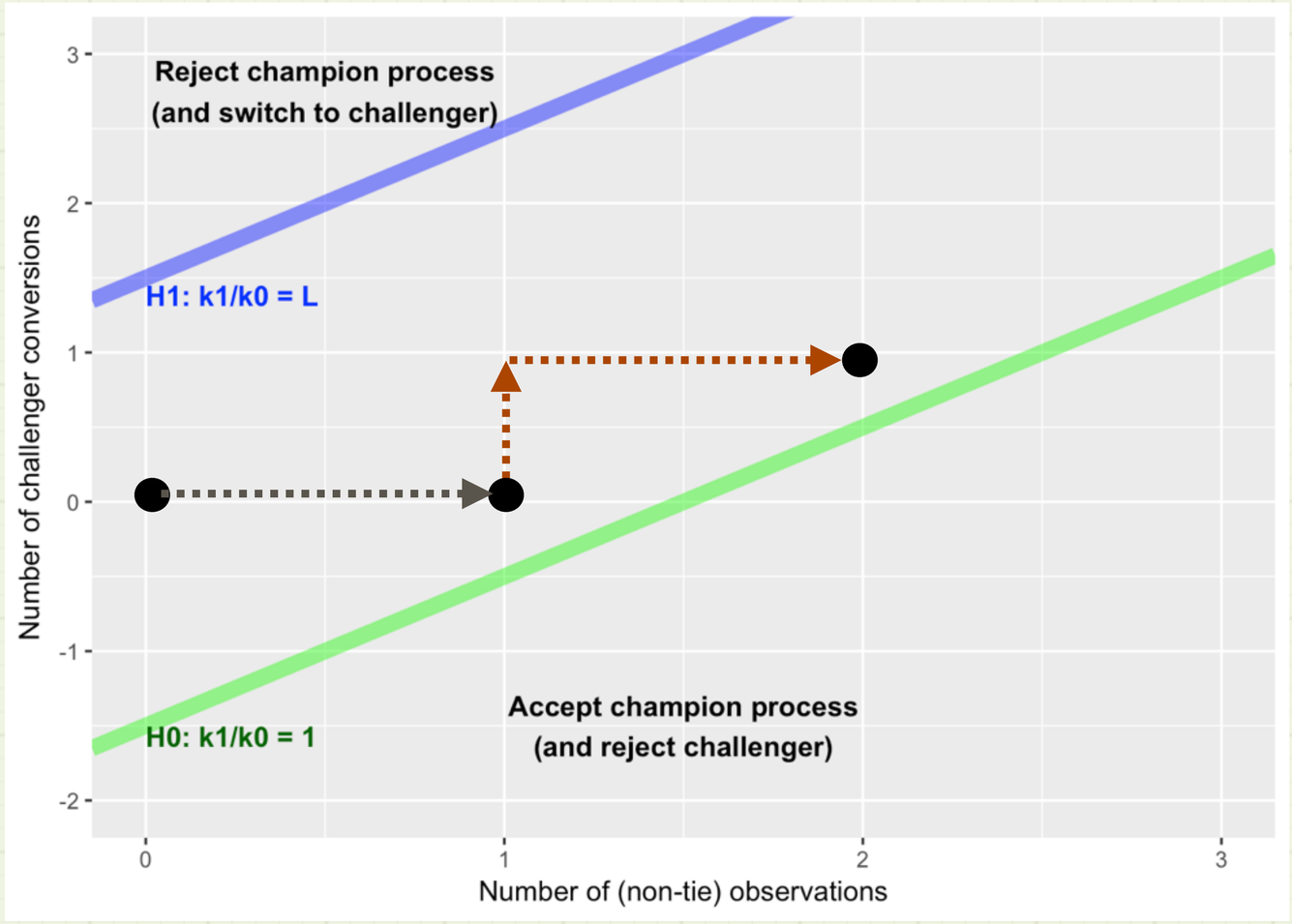


The Idea

Process 0 (Champion)	Process 1 (Challenger)
TRUE	FALSE
~~TRUE~~	~~TRUE~~
FALSE	TRUE
FALSE	FALSE
FALSE	TRUE
TRUE	FALSE

(Skip tie)

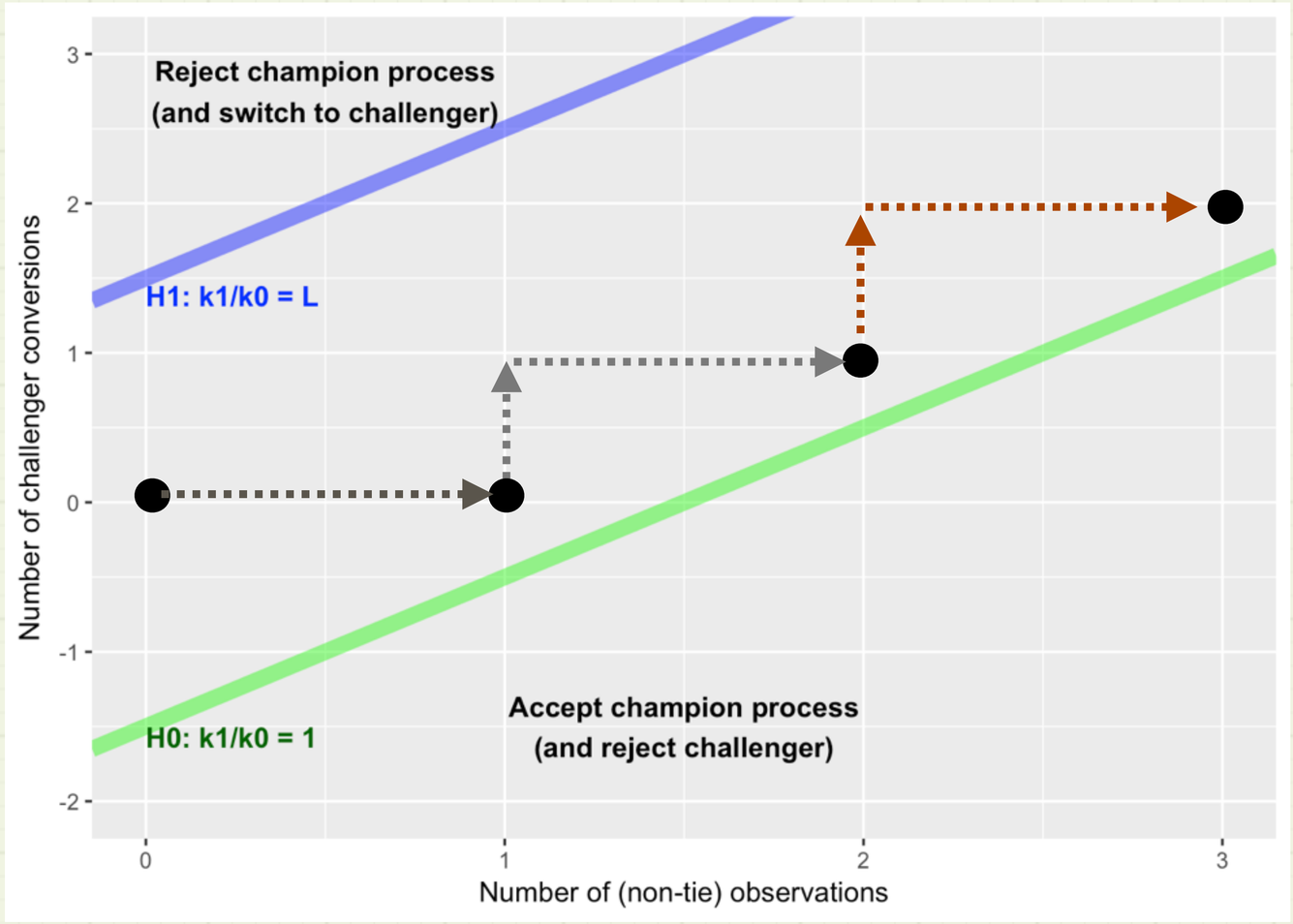
P_1 converts: move one up and one to the right



The Idea

Process 0 (Champion)	Process 1 (Challenger)
TRUE	FALSE
~~TRUE~~	~~TRUE~~
FALSE	TRUE
~~FALSE~~	~~FALSE~~
FALSE	TRUE
TRUE	FALSE

Keep going until exit the channel.
Exit on bottom: keep P_0
Exit on top: switch to P_1

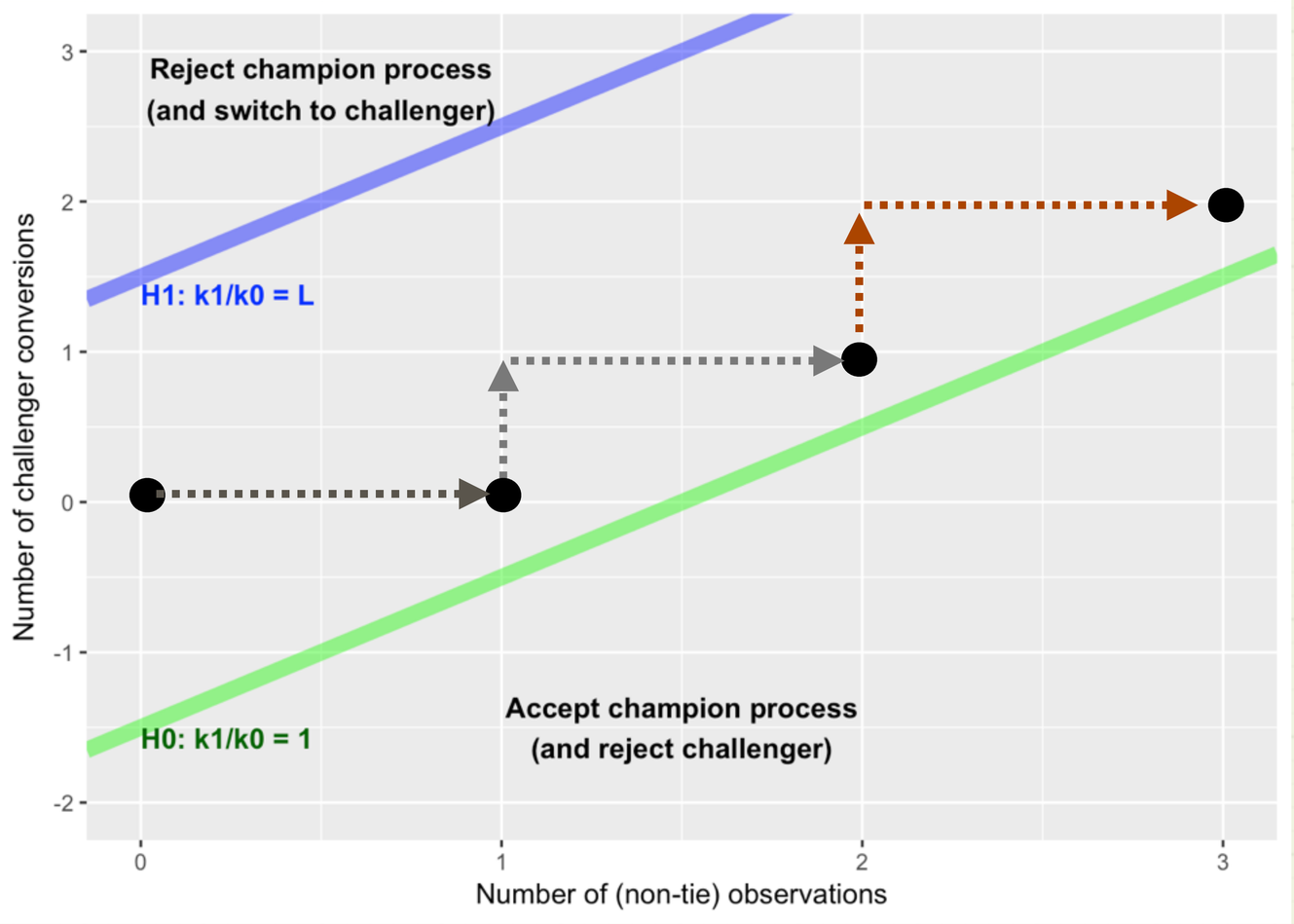


The Idea

If you exit on the bottom...

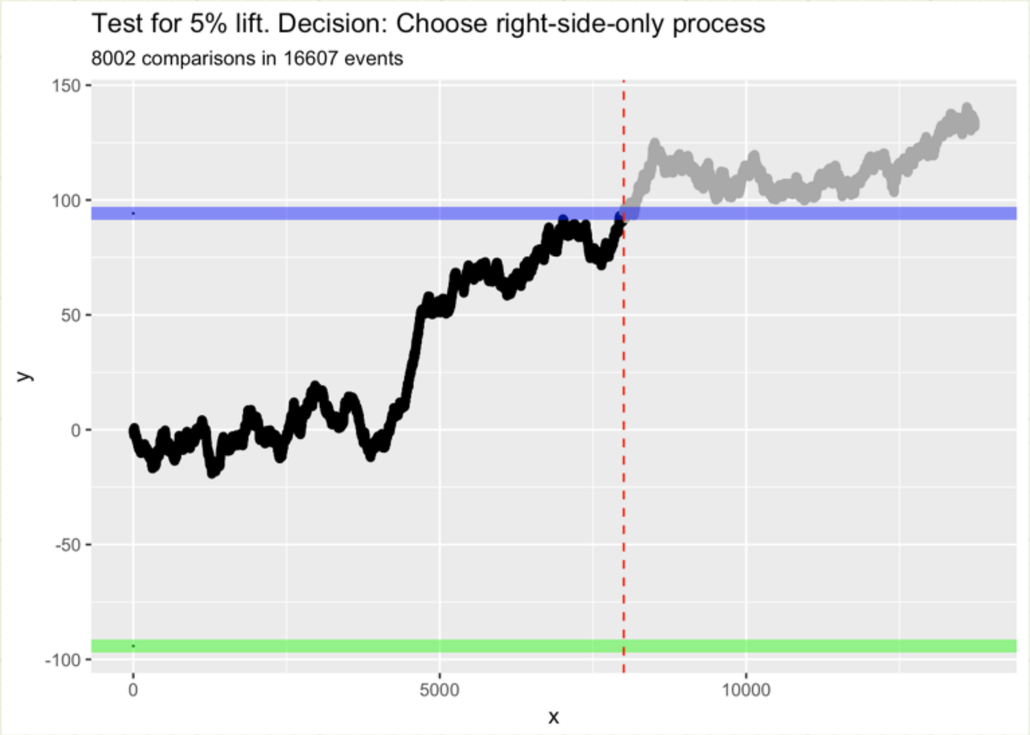
- DOESN'T mean $P_0 > P_1$
- It means P_1 isn't good enough to switch to.

Also, if k_1/k_0 is different than what you expect or what you require, sometimes the test can terminate earlier than a non-sequential test.



Movie Data: Two Scenarios

Scenario 1: Test for 5% lift

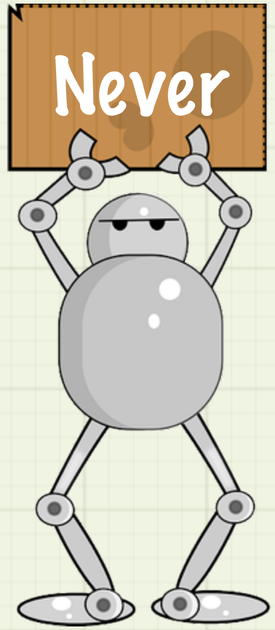


Scenario 2: Test for 20% lift

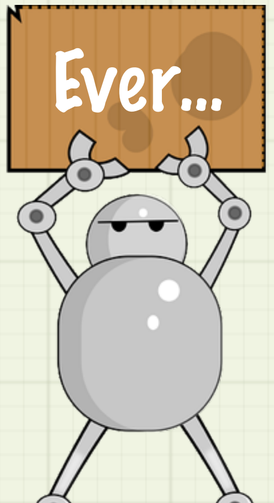
(In reality: total data set has a 7% lift)

Win-Vector LLC

MYTH



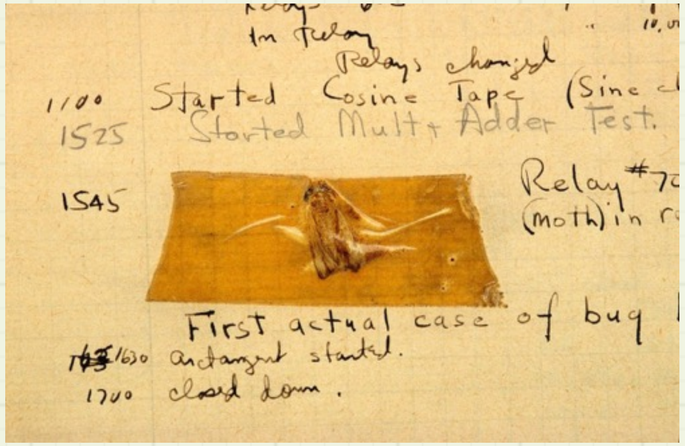
"That will never happen."



Statistics Can't Save You From Bugs and Bad Procedures

Examples

- Potential nested model bias from incorrect y-aware data pre-processing
- Bad decisions from too many A/B tests
  - *p* is the rate of false positives when there's no effect
- Holdout set leakage from too much repeated evaluation
  - "Wacky Boosting" by Moritz Hardt (http://blog.mrtz.org/2015/03/09/competition.html)
  - A sneaky form of overfit



Grace Hopper's bug (1947)

Conclusion

- Our most important data science tools are our theories and methods.
- New algos and technologies are sexy, but the basic laws of statistics still apply.
- Step back, examine your unexamined assumptions

Thank You

Slides:
https://github.com/WinVector/ODSCWest2017/

References

Stacked Learning

- Breiman, Leo. "Stacked Regressions," *Machine Learning, 24* p. 49-64 (1996). http://statistics.berkeley.edu/sites/default/files/tech-reports/367.pdf

- van der Laan, Alan, et.al. *Super Learner*. http://biostats.bepress.com/ucbbiostat/paper222/

Nested Model Bias

- Mount, John (2016). "On Nested Models," *Win-Vector Blog*. http://www.win-vector.com/blog/2016/04/on-nested-models/

Significance and *p*-values

- Mount, John (2017). "Remember, *p*-values Are Not Effect Sizes," *Win-Vector Blog*.
  http://www.win-vector.com/blog/2017/09/remember-p-values-are-not-effect-sizes/

Significance Pruning

- Zumel, Nina (2015). "How Do You Know if Your Data Has Signal?", *Win-Vector Blog*.
  http://www.win-vector.com/blog/2015/08/how-do-you-know-if-your-data-has-signal/
- Significance pruning implementation in `vtreat` R package
  https://winvector.github.io/vtreat/articles/vtreatSignificance.html

Cohen's *d*

- Cohen, Jacob (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.
- Cohen, J (1992). "A power primer". *Psychological Bulletin*. 112 (1): 155–159

Sequential Analysis

- Wald, Abraham (2004). *Sequential Analysis*. Dover.
- https://en.wikipedia.org/wiki/Sequential_probability_ratio_test
- Spiegelhalter, et.al. (2003). "Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery". *International Journal for Quality in Health Care,* Vol. 15, No. 1.
    - Example of single process case
- Sequential analysis packages in `R`:
    - `Sequential`: https://CRAN.R-project.org/package=Sequential
    - `SPRT`: https://CRAN.R-project.org/package=SPRT

Model Evaluation and Testing

- Mount, John (2015) "A deeper theory of testing," *Win-Vector Blog*.

  http://www.win-vector.com/blog/2015/09/a-deeper-theory-of-testing/

- Zumel, Nina (2015) "Random Test/Train Split is Not Always Enough," *Win-Vector Blog*.

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