Machine Learning

How to get the value and avoid the pitfalls

Analytics Frontiers Conference

Bill Kahn Ph.D.
bill.kahn@bankofamerica.com
March 21, 2018

Bank of America

Overview

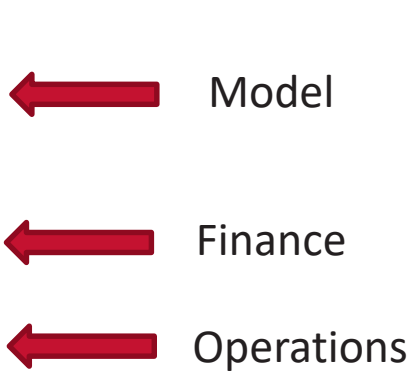
1. What modeling is all about
2. What ML does well
3. Where ML breaks down

What modeling is all about

Running a business, or any endeavor, is about making decisions

Deming 1938:
The object of collecting data is to provide a basis for action.

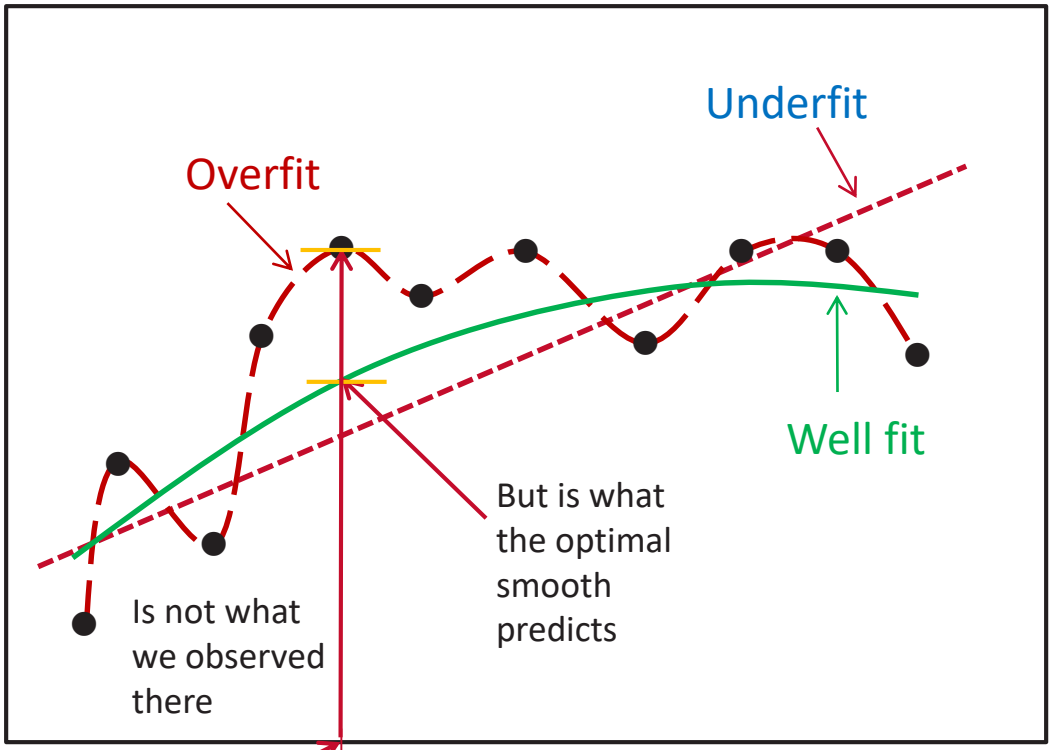
If we do “A” then we’ll see X
If we do “B” then we’ll see Y
X is better than Y
Thus we will do “A”



W. Edwards Deming 1900-1993

The key idea in all Machine Learning—regularization though cross-validation

Modeling is balance



The best prediction for this value of X

Machine Learning does two things

1. Defines a sequence of functional complexity
   - Sequence of basis functions
     Boosting, Bagging, polynomials, and many others
   - Fits the sequence of increasingly squiggly functions
2. On a random hold out (cross validations, bootstrap,…)
   - Evaluates each in the sequence
   - Keeps the best

This simple process enables solving some hard problems:

- Can look through many variables
- Can handle aspects of missing-induced bias
- Enables a good balance between over- and under-fitting
- Predicts within sample well

What ML does not do

1: What does the world need?

Type III errors: Solving the wrong problem (1947 F. David)

Time is precious. Resources finite.

Let's work on what matters.

This is a big part of science—problem selection.

Who will do what differently?
Who cares? Who will celebrate with you?

2: Your data is biased

- **You must understand how the data you have is not representative of the data you need**
- If you have modeled the default probability for all booked loans, how does this relate to the default probabilities for the loans you have historically declined?
- If you have modeled the one-year profitability for people you have marketed, how does this relate to the profitability for those you have not marketed?
- How far back in time is still helpful to guide you forward?

Nothing in ML solves this key question of data representativeness

3: Dependent Variable

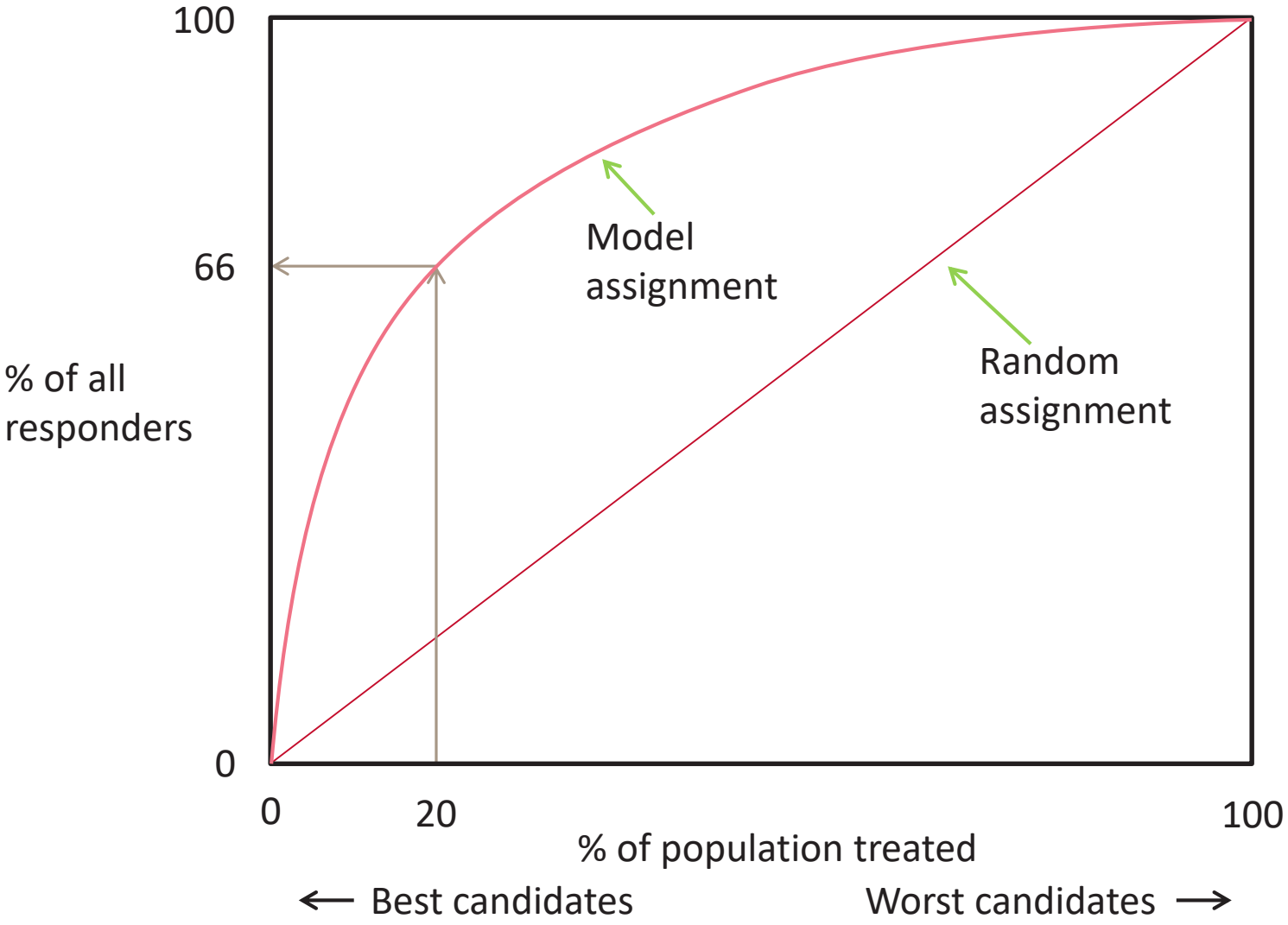
What will you predict?

Even a great model (say a Somers' D of 80%) may have no economic value—it could turn out it is not economic to mail anyone. Or it is economic to mail everyone.

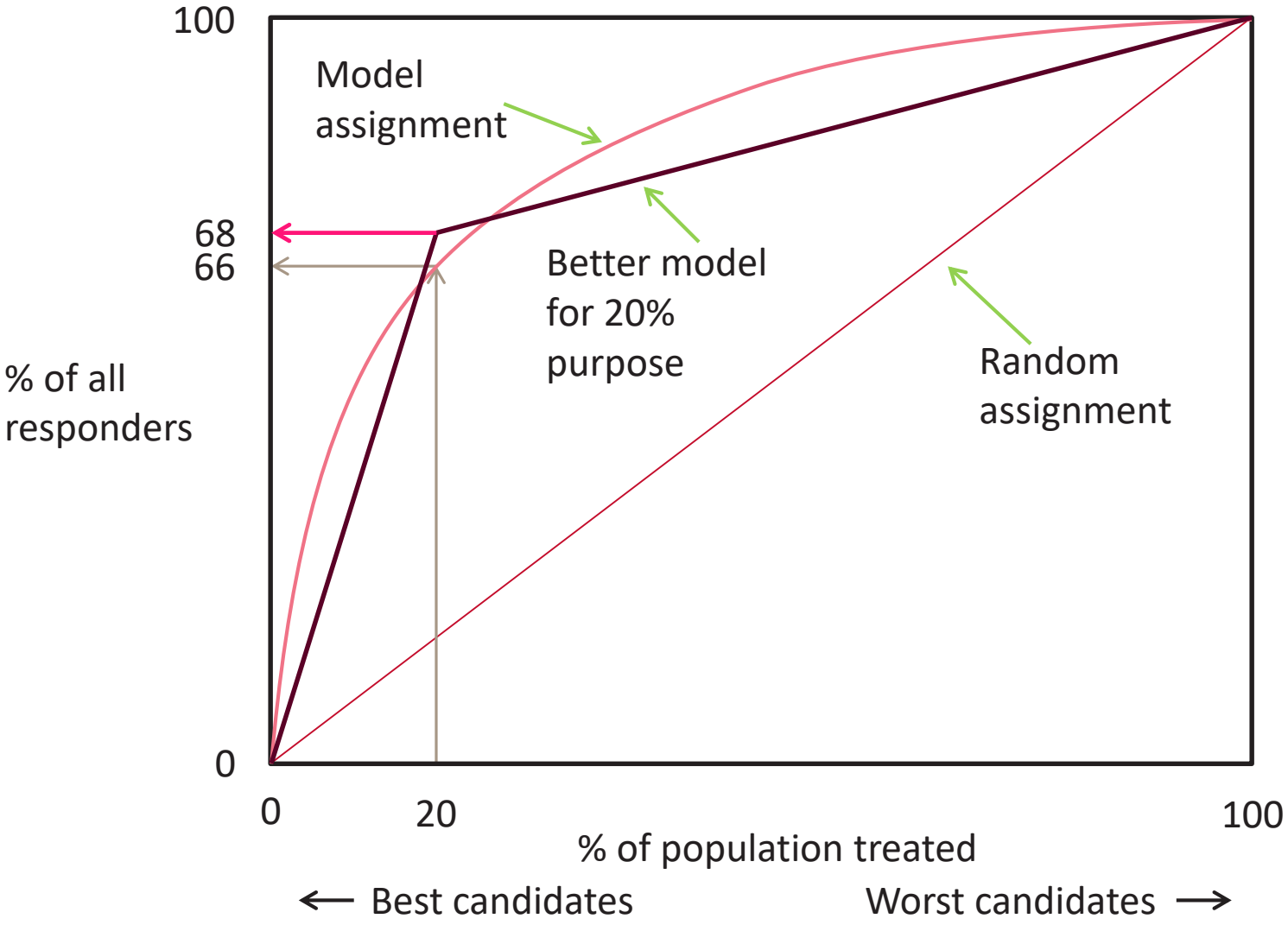
In marketing we tend to acquire customers who are easier to acquire, and individuals who are easier to acquire also are easier to lose, and so do not have average CLTV.

Thus acquisition rate can be a weak proxy for value

The classic lift curve

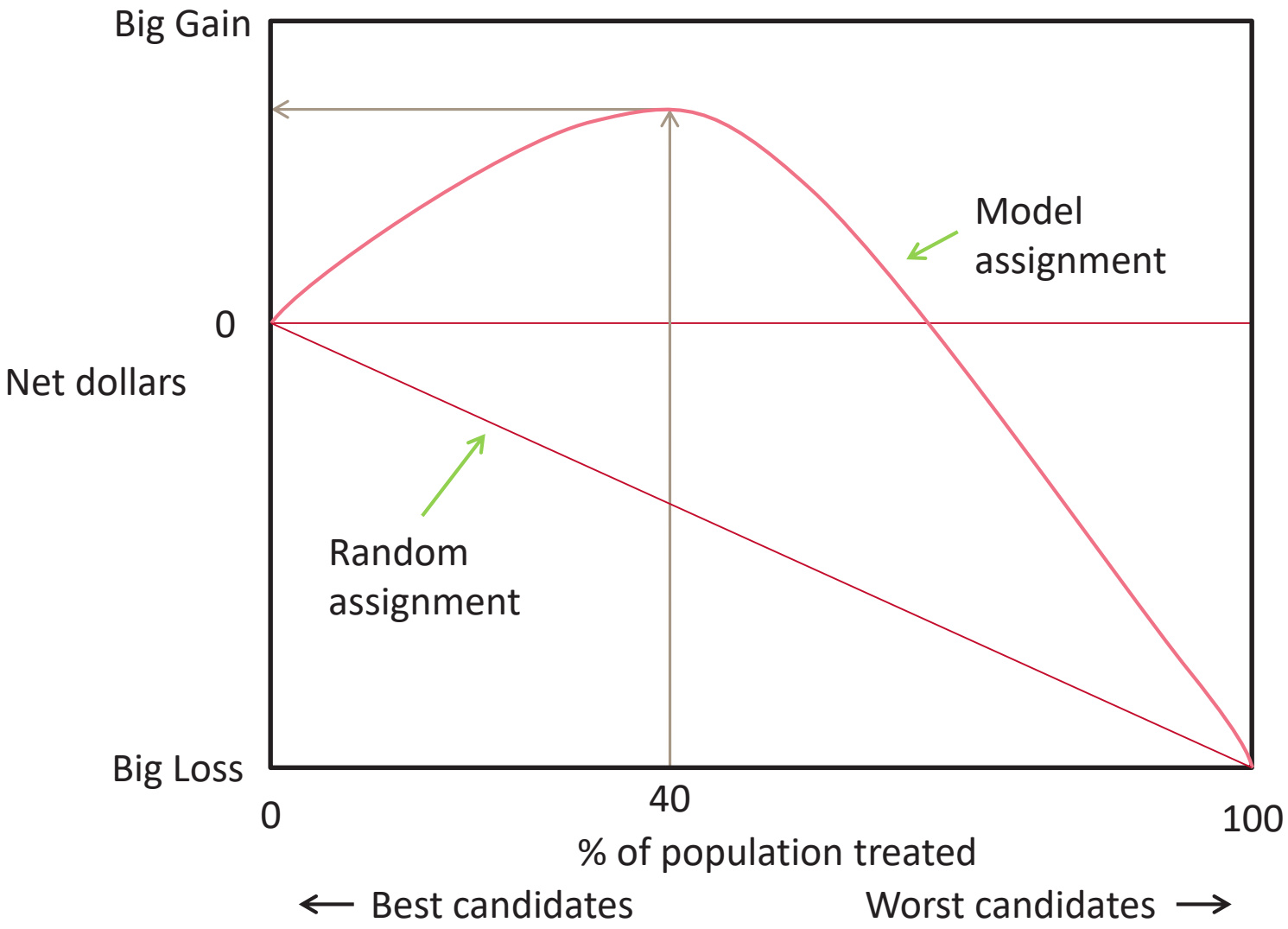


The classic lift curve



The economic lift curve

Assign say $1 cost to everyone mailed
and say $100 CLTV to every responder



The dependent variable matters

- Denomination is important
- Models are designed to improve society
- If we have enough budget to market 20% of the file, then the correct metric is not Somers' D but rather lift at 20%
- When denomination is hard to do rigorously, do approximately
- For example, if what you need is an estimate the 99th percentile only sophisticated use of ML will work

4: Most ML algorithms by default select a wrong loss function

- The typical default is to minimize the classification error rate
- So, if we have a 1% delinquency event rate many algorithms by default will say to approve everyone.
- This decision is right 99% of the time
- And is dumb
- Users of ML must actively control the Type 1 and Type II misclassification costs
- Most tools have some way to do this, but the terminology varies widely
- Failure to actively control the loss function can result in (decision theoretic) awful models

5: What does zero mean?

- Say you have 1,000 business loans in each of five regions
- And last year, in Region A, you had zero defaults
- You put your data into your ML and it predicts **zero** in Region A
- Any other value cross validates poorly
- But, will you assign zero risk to Region A?

- NO!

- **ML needs to be augmented by probability models**
  - How do you shrink your estimates towards something you know to be right?
  - You certainly know important things not available to the ML, so straight trust of ML does not make sense.

5: (continued) Where is the variability?

- All (current) ML assumes that samples are adequately independent
- That is what the cross-validation sampling is doing—assuming it has a million samples from the population you want to make inference to
- Consider you have data on
  - Three years
  - 1 million people per year
  - 1400 variables on each person from a credit bureau
  - The national unemployment rate for each year
- ML thinks it has 3 million df on unemployment rate
- But it only has ***two***
- **ML gives a dramatically wrong result**
- **Understanding the probability sampling process (geographic, temporal, organizational and much more) is essential to getting reliable ML results**
- This is an essential next-gen research project—how to embed investigator knowledge of the sampling process (as in longitudinal models, split-plots, hierarchical) into cross-validation to enable better ML

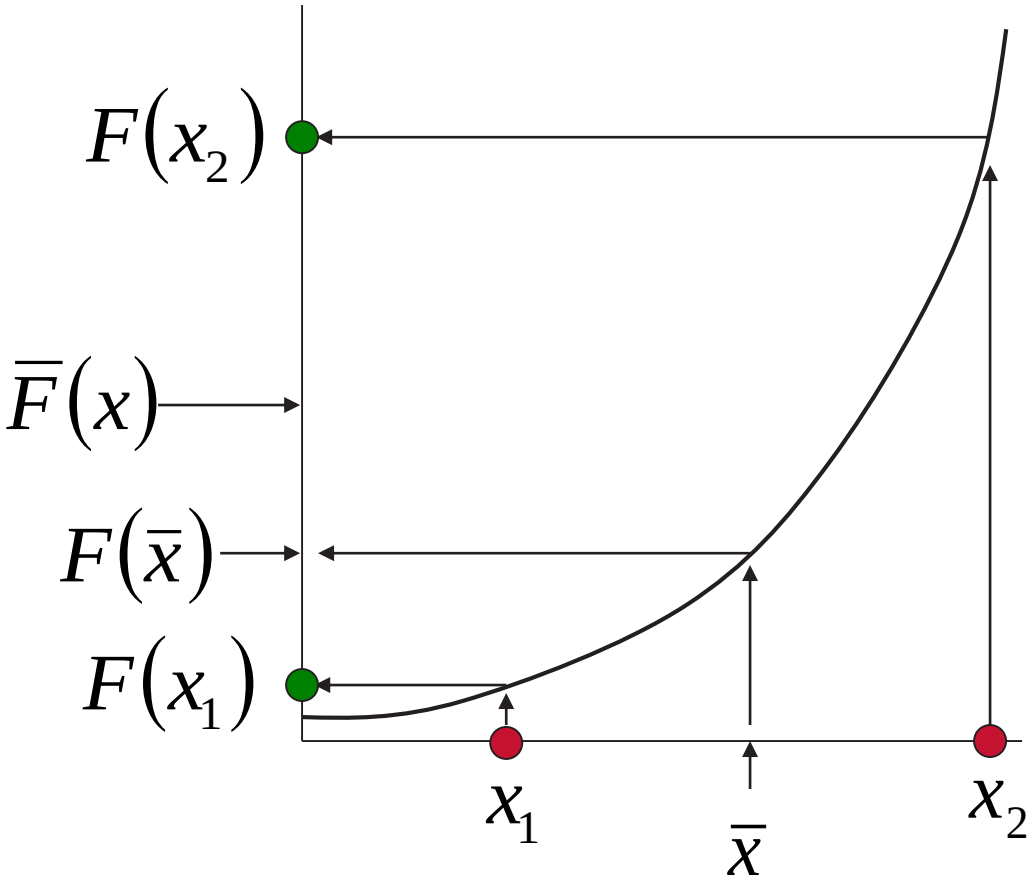
6: Sequencing models is nearly always wrong

Taking the output of one model and feeding that in as the input to another model often produces wrong results.

- The output of the first is typically an expected value and fails to capture the variability
- The second model is often nonlinear (perhaps it is a multiplication).

The combination produces wrong expected values...

Expected values do not propagate through a nonlinear function



Jensen's Inequality:

For F non-linear

$$\overline{F}(x) \neq F(\overline{x})$$

$$E(F(x)) \neq F(E(x))$$

Thus meaning

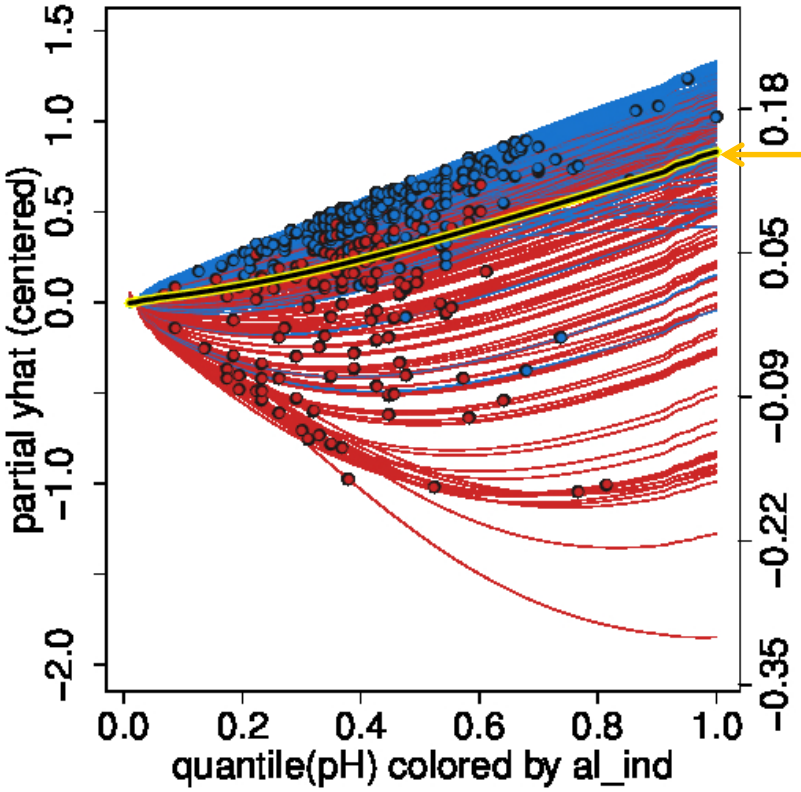
$$E(X * Y) \neq E(X) * E(Y)$$

$$PD * LGD \neq EL$$

For any model, machine learning or not, is the prediction used in a product, ratio, or any other non-linear function?

7: Historical data is not reliably interpretable, ML or not

- Real interpretation, without randomized replicated trials, is way hard
- It is always easy to tell a story about any prediction equation
- **Individual Conditional Expectation plots** (ICE)
  https://arxiv.org/pdf/1309.6392.pdf



The **Partial Dependency Plot** (in yellow) is the average ICE plot

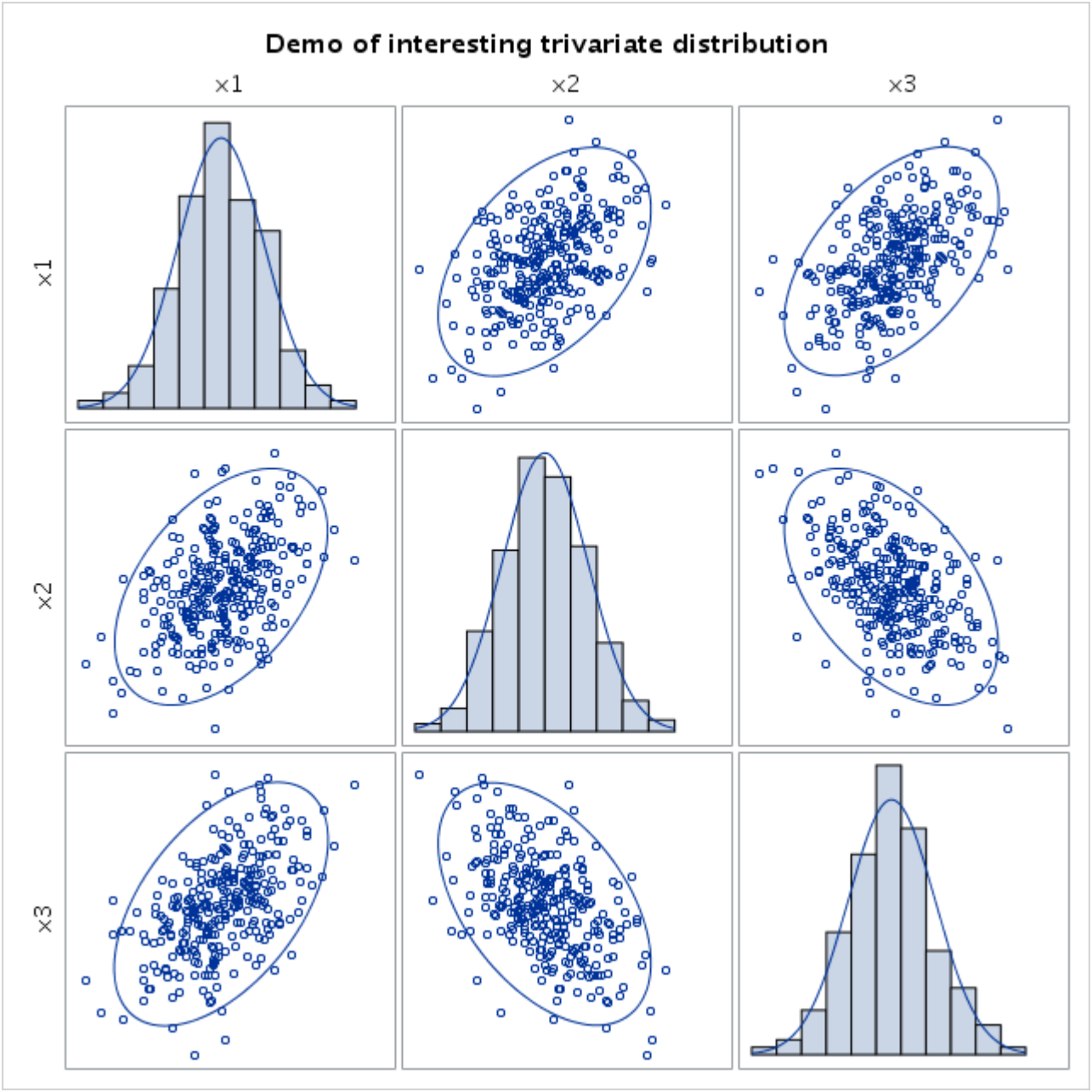
But what does it mean to “interpret,” not just “describe?”

- Interpretation, separate from description, is a **causal attribution**
- In historical data your covariates are confounded in complex ways a human cannot intuit
- For example corr(X,Y)=.5, corr(Y,Z)=.5, and corr(X,Z)=**-.5**
- In ML, try this:
  - Fit your model
  - Simplify it down to the 10 most important drivers
  - Now remove those 10 variables and refit and re-simplify
  - Now see how well the two models correlate—you will find (very often) very well.
  - “The multiplicity of good models”
  - Interpreting one gives a different story than interpreting the other

Correlation is not transitive

Can you puzzle out what these points look like in three-space?

Human intuition is surprisingly weak in certain domains



How get causality, and thus reliable prediction? Experiment.

- Getting to a scientifically supportable **causal** explanation from one piece of historical data is not possible. Causality require a complex scientific sequence.
- As the purpose of modeling is to drive optimal decisions, the best path forward is to go ahead and make alternative decisions.
- Do mostly sensible actions—but now and then try something not predicted to be perfect.
- Thompson Sampling
- This will generate the randomized data that will enable causal interpretation

We must predict conditional on what we do

$$\hat{Y} = f(C, T, C * T)$$

Where C is the covariates that we can observe but not change.

T the treatments we control.

C*T their interaction.

Meaning what works well changes depending on the covariates.

8: Simple empirical exploration is not enough

- Simply exploring multiple ML algorithms is not near enough
  - Basically what makes the algorithms different is the basis vectors explored
  - As nearly all the ML alternatives basically work, you will only find nuanced or specialized advantages of one over the other (compute time, development time…)
- Exploring the hyper-parameter space is important, and not enough
  - Every ML algorithm has dozens of hyper parameters
  - These are stray meta-coefficients that are selected based on how well they happen to work on your data
  - Lousy hyper-parameters can make an otherwise good model lousy
  - Note—default values are often lousy
  - Once in the space of sensible, further refining is typical not a big deal
- **The deep problems remain regardless of the searching**

(When you do these meta searches, do keep a full third-level of hold out. Once you have what you like, evaluate it on Level 3 to learn how well it is actually likely to work. This is just a classic multiple comparison problem that must get managed)

9: Documentation and Reproducibility

- Science must be reproducible
- You need to be able to come back to your raw data and redo all steps through to your final conclusions
- And others have to be able to do so as well
- Much of current ML is a free-for-all of constantly changing open-source software stacks of highly variable quality and Johnny-come-lately commercial vendors
- Like in all high-tech engineering the insides matter and confirming any particular piece of software is adequately reliable is itself a hard project
- And, like all science, and all data analysis, reliably documenting what you did, why, and what you found out, is hard and important. Given the immaturity of the ML world, it is particularly hard.

Summary

- Naïve use of Machine Learning produces a mess
  - Wrong question solved
  - Wrong data used
  - Wrong dependent variable predicted
  - Wrong loss function used
  - Wrong use of the underlying sampling process
  - Wrong sequential chaining of models
  - Wrong interpretation of the results
  - Wrong hyper-parameters
  - Wrong documentation

- Machine Learning helps on a part of the data analysis challenge
  - Enables looking through many variables
  - Enables making good within sample predictions