

# MSiA

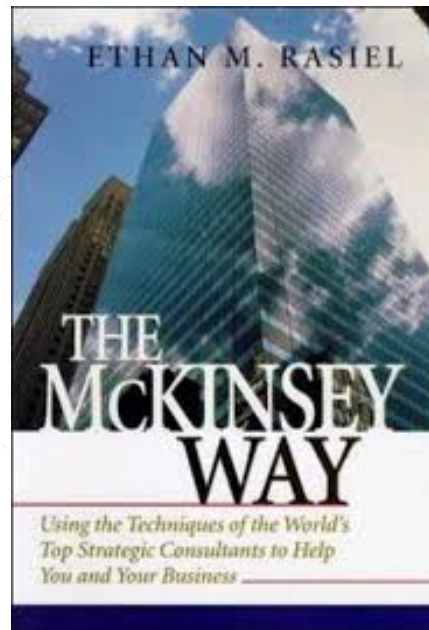
## CLASS 2: MODELS FOR DECISION-MAKING

Joel Shapiro  
Clinical Associate Professor, MEDS

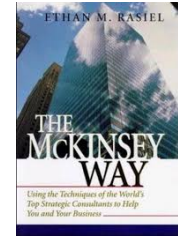
Northwestern | Kellogg

# A FRAMEWORK FOR RIGOROUS DECISION-MAKING IS NOT NEW

**Q: What did YOU  
find helpful (or not)  
about this reading?**



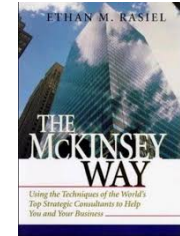
# WHAT I FIND USEFUL...



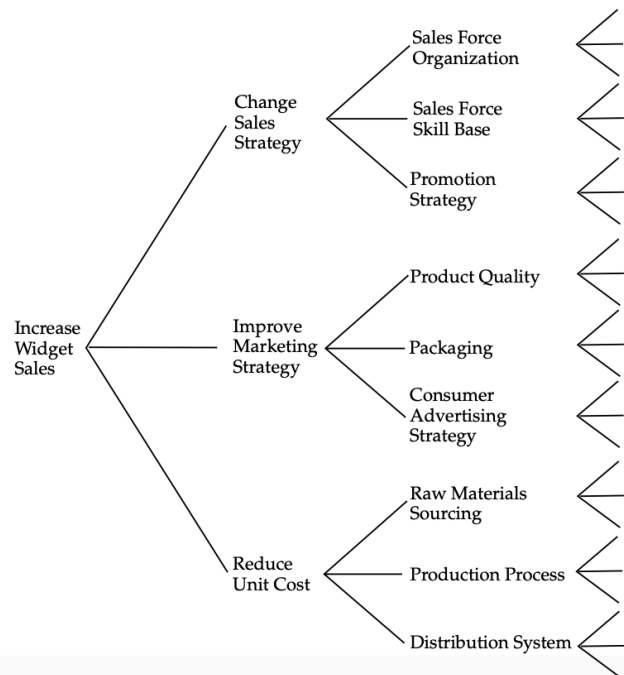
## **“Facts are Friendly” (p4)**

- Analytics reveals stories that are not easily seen otherwise
- Your goal is to seek an "objective truth" through analysis
- But “objectivity” can be elusive!
- What if your initiative boosts NPS but decreases retention?
- Regression analysis can give different results than Random Forest modeling. Same data, both good tools, different “facts.”

# WHAT I FIND USEFUL...



ISSUE TREE FOR ACME WIDGETS



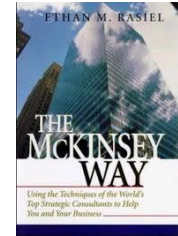
## “Generating hypotheses” (p8-12)

Note the issue tree on p 12

This is an EXCELLENT way to begin an analytics project. Make sure that you know how the results of your analysis will feed into a business decision.

Clearly map out how the analysis will lead to: 1) a decision and 2) business value.

# WHAT I FIND USEFUL...



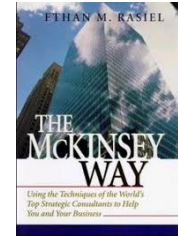
## **“The problem is not always the problem” (p16)**

Sometimes, analysis doesn't answer the stated problem / question / opportunity, but reveals new problems / questions / opportunities.

E.g., you are asked to analyze why your social media messaging isn't as effective as it could be. Your analysis shows that the messaging is effective, but that the process for uploading social media content is slow and unreliable.

It's OK to shift gears! Just be clear about what you're solving at any given time!

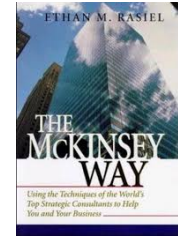
# WHAT I FIND USEFUL...



Make sure your solution fits your client (p22)



# WHAT I FIND USEFUL...

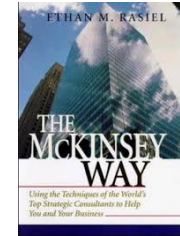


## **Make sure your solution fits your client (p22)**

Play lots of “what-ifs?”

- What if the analysis shows that we should fire / hire people?
- What if the analysis shows that we need to add a new product line?
- What if the analysis shows that promotions don't work?
- What if the analysis shows that our social media spend is ineffective?
- What if the analysis shows that our stores need greater oversight?

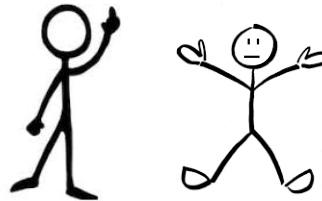
# WHAT I FIND USEFUL...



## Don't boil the ocean (p32) / Hit singles (p39)

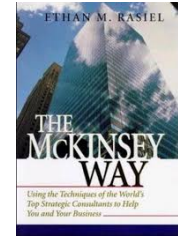
For an analytics project to succeed, it needs specificity and boundaries. Otherwise, you end up analyzing everything at once, becoming overwhelmed, and having no real actionable takeaways for business improvement.

"I need your first project to help us increase revenue, reduce cost, and boost employee satisfaction. Oh, and factor in what happens if we move our HQ across the country. And add a new product line."





# WHAT I DON'T PARTICULARLY LIKE...



## The 80/20 rule (p30)

“One of the great truths of management.”

80% of sales come from 20% of sales people

80% of orders come from 20% of customers

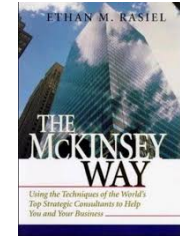
80% of productivity comes from 20% of employees

I see this as a faux-rigorous way of saying “not everyone is exactly equal.”

These “truths” take hold because they are convenient and easy to remember.

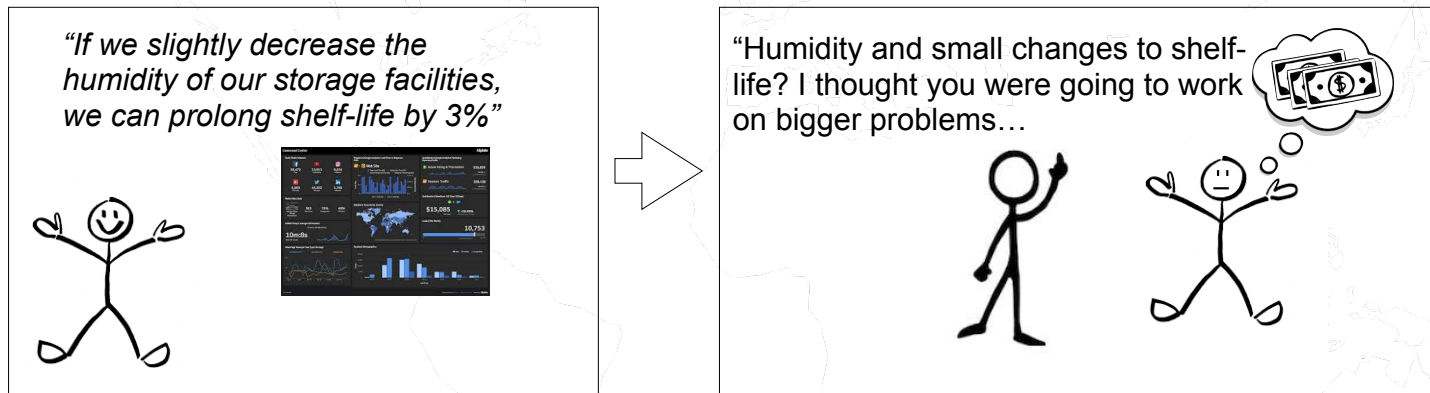
But don't take them as given!!! You have the ability to test them and improve upon them.

# WHAT I DON'T PARTICULARLY LIKE...

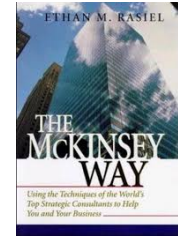


## Focus on the most important drivers (p33)

Sometimes, your modeling will reveal trends / relationships that might seem ancillary, but can yield big results.



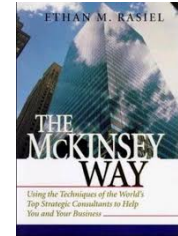
# COMING UP LATER...



## **The elevator test (p34)**

This is entirely about the importance of being a great analytics communicator, which we'll get into later in the class.

# WHY DID WE READ THIS?



*Many analytics tools are new, many concepts are not.*

There are exceptions - being able to craft accurate predictions enables us to do some things differently.

Plus, automation offers lots of cool opportunities to implement analytics at scale.

Your success as a data scientist depends on your ability to do good work, but also to recognize that many analytics concepts are not new to the business world.

The marketing team at your company has a big database of prospective customers, and can predict the likelihood that each customer will convert.

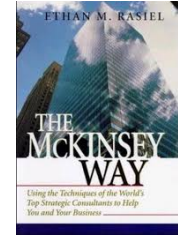
Prospect A is 1% likely to convert

Prospect B is 50% likely to convert

Prospect C is 99% likely to convert

*If you have the resources to send only ONE prospect to the sales team, which one should it be?*

# WHY DID WE READ THIS?



*Many analytics tools are new, many concepts are not.*

There are exceptions - being able to craft accurate predictions enables us to do some things differently.

Plus, automation offers lots of cool opportunities to implement analytics at scale.

Your success as a data scientist depends on your ability to do good work, but also to recognize that many analytics concepts are not new to the business world.

**E.g., a business leader wanting to deploy resources in a way that maximizes marginal benefit is not new. But using prediction with experimentation as a method may be novel.**

The marketing team at your company has a big database of prospective customers, and can predict the likelihood that each customer will convert.

Prospect A is 1% likely to convert

Prospect B is 50% likely to convert

Prospect C is 99% likely to convert

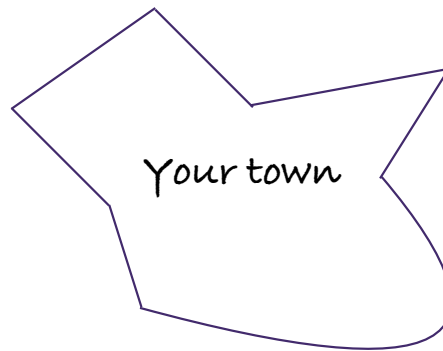
*If you have the resources to send only ONE prospect to the sales team, which one should it be?*

## YOUR TURN... TO HELP DETERMINE WHERE TO PUT EMS SERVICES

You are a council member for a small town in the early 1900s (pre-computers)

Your town has long been using the emergency management services (fire, law enforcement, etc.) of a neighboring town. You feel like you are finally at the point where you can and should build your own.

Q: Where to put it?



# ADVISING ON THESE DECISIONS REQUIRES **PRECISION**

What's important?

What are the goals?

What are the options?

How do you balance the pros and cons of each option?

As data professionals, we build **MODELS** to generate answers.

# MODELS ENABLE DECISION-MAKING

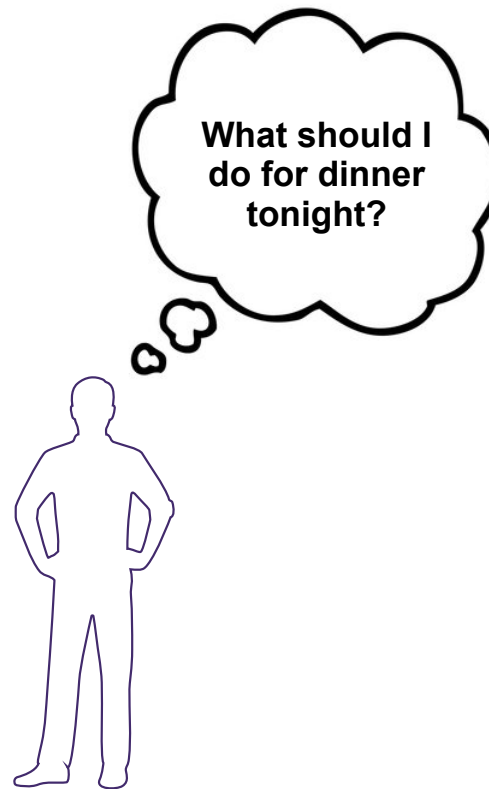
1. Identify universe of options
2. Generate an evaluation mechanism
3. Prioritize among those evaluations

$$\begin{array}{ll} \text{maximize} & \\ \text{minimize} & \end{array} f(x,y,z\dots) \quad \text{s.t. constraints}$$

**You might know these as  
“constrained optimization” problems.**



I WANT TO HELP YOU FIGURE OUT DINNER TONIGHT!



# CONSTRAINED OPTIMIZATION IS A GOOD WAY TO THINK ABOUT WHY YOU BUILD MODELS AND HOW THEY WILL BE USED

**Where should I go for  
dinner tonight?**

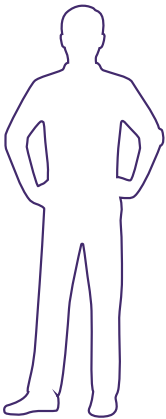
What are my options?

How much value does each have? (taste, fullness, health)

What's my objective?

What's the most I'll spend, farthest I'll drive, etc.?

$$\begin{array}{ll} \text{maximize} & f(x,y,z,\dots) \\ \text{minimize} & \end{array} \quad \text{s.t. constraints}$$



# CONSTRAINED OPTIMIZATION IS A GOOD WAY TO THINK ABOUT WHY YOU BUILD MODELS AND HOW THEY WILL BE USED

**Where should I go for dinner tonight?**

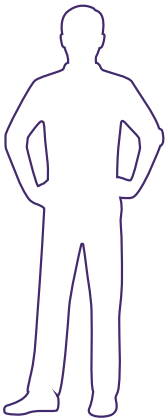
What are my options?

How much value does each have? (taste, fullness, health)

What's my objective?

What's the most I'll spend, farthest I'll drive, etc.?

$$\begin{array}{ll} \text{maximize} & f(x,y,z,\dots) \\ \text{minimize} & \end{array} \quad \text{s.t. constraints}$$



**Things you care about can be expressed as objectives OR constraints.**

What's the difference here?

Maximize fullness, s.t. <\$15

Minimize cost, s.t. must feel full

# CONSTRAINED OPTIMIZATION FORCES PRECISION

I like thinking of problems as constrained optimization problems bc they force decision-makers to articulate **what they care about** and the **relative value of each thing they care about**.

$$\begin{array}{ll} \text{maximize} & f(x,y,z,\dots) \\ \text{minimize} & \end{array} \quad \text{s.t. constraints}$$

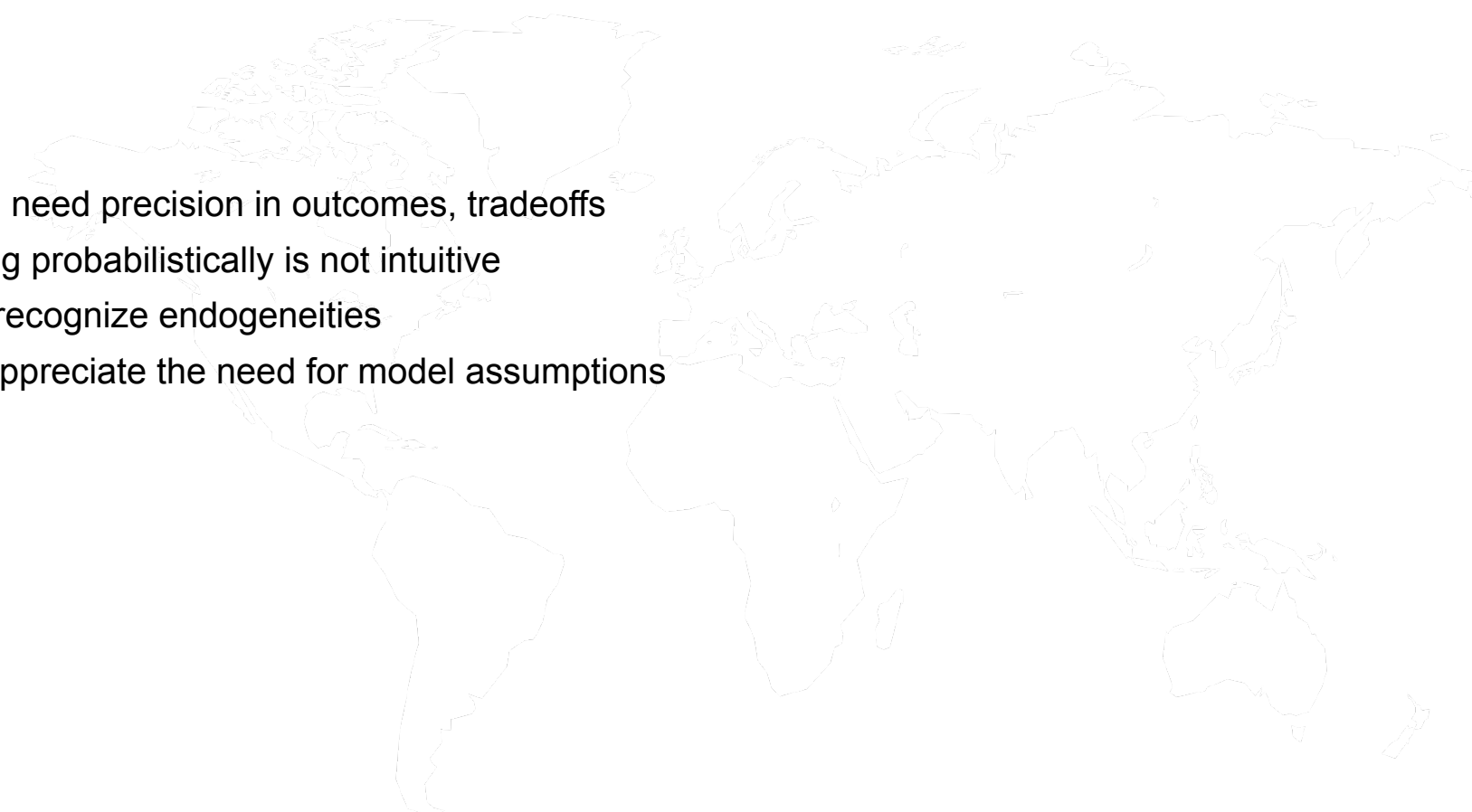
“Advise me on where to go for dinner.”

“Help me figure out which customers I should target.”

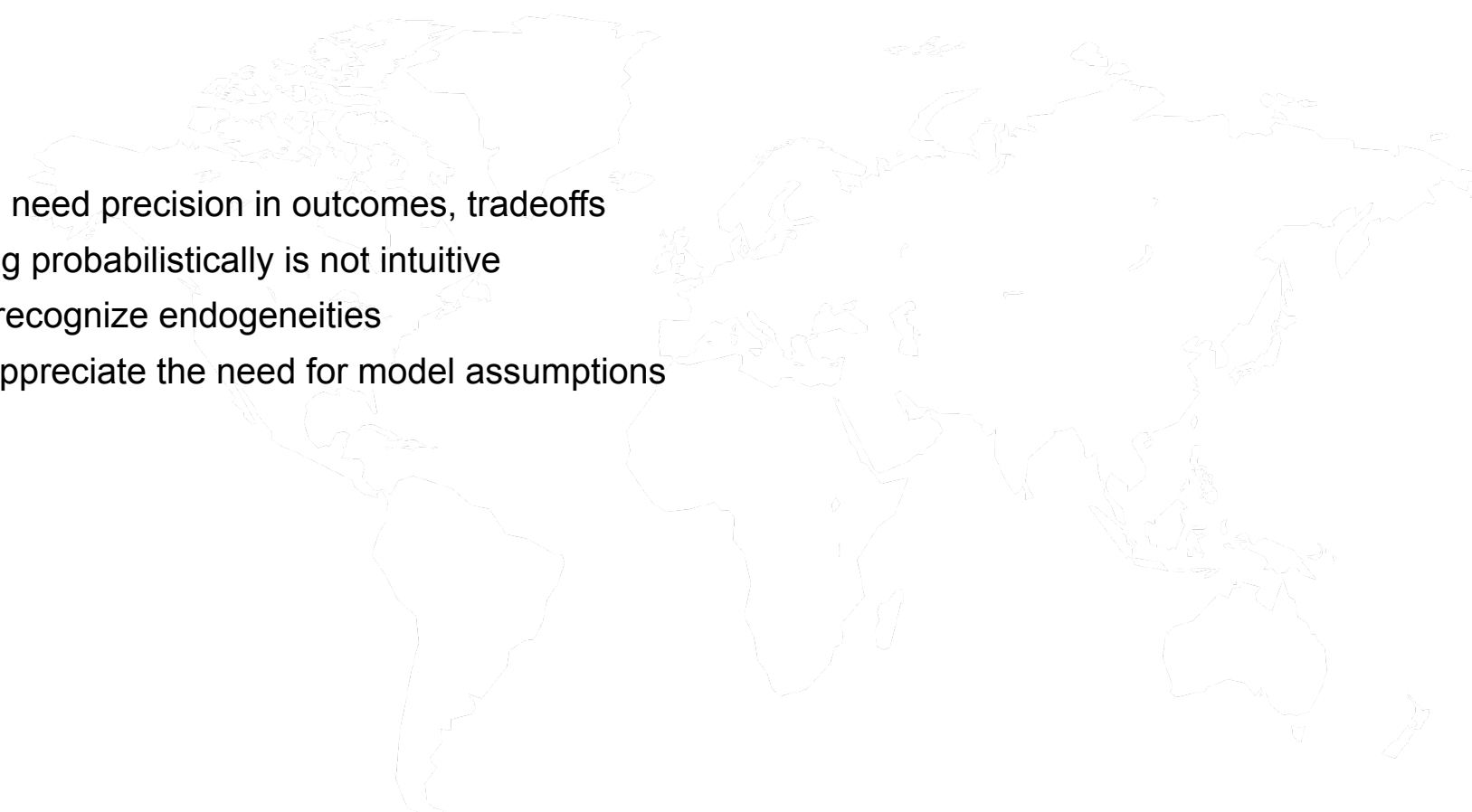
“We have one COVID vaccine left.  
Should we give it to an elderly person with many health problems?  
Or a healthy 25 year old?”

And **precision** means I can provide **better solutions**.

# SO WHY DO BIZ LEADERS STRUGGLE WITH MODELS?

- 
- ✓ 1. Models need precision in outcomes, tradeoffs
  - 2. Thinking probabilistically is not intuitive
  - 3. Fail to recognize endogeneities
  - 4. Don't appreciate the need for model assumptions

# SO WHY DO BIZ LEADERS STRUGGLE WITH MODELS?

- 
- 1. Models need precision in outcomes, tradeoffs
  - ✓ 2. Thinking probabilistically is not intuitive
  - 3. Fail to recognize endogeneities
  - 4. Don't appreciate the need for model assumptions

Was this model wrong?

Three models

☐ Polls-plus forecast  
What polls, the economy and historical data tell us about Nov. 8

☒ Polls-only forecast  
What polls alone tell us about Nov. 8

☐ Now-cast  
Who would win the election if it were held today

National overview

Updates  
National polls

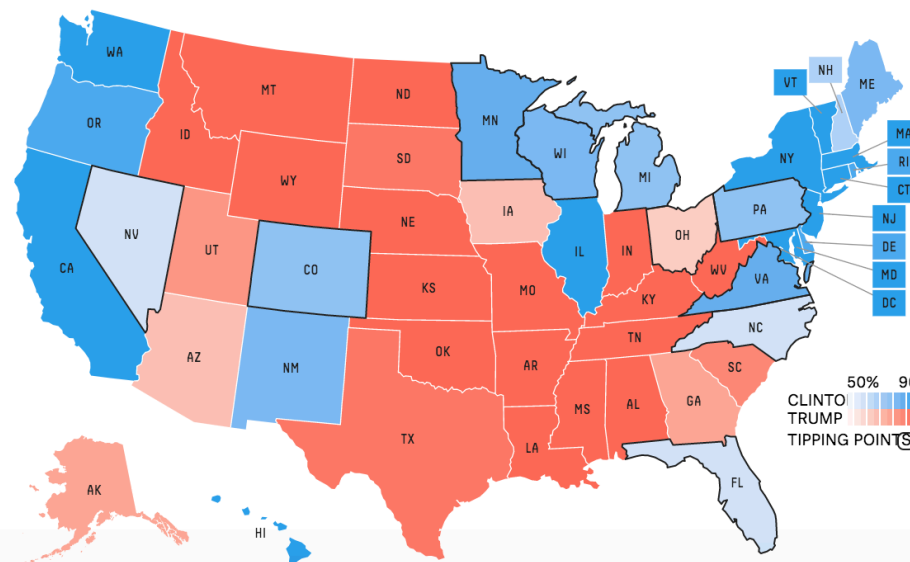
States to watch

Arizona  
Colorado  
Florida  
Georgia  
Iowa  
Maine  
Michigan  
Minnesota  
Nevada  
New Hampshire  
New Mexico  
North Carolina  
Ohio  
Pennsylvania  
Utah

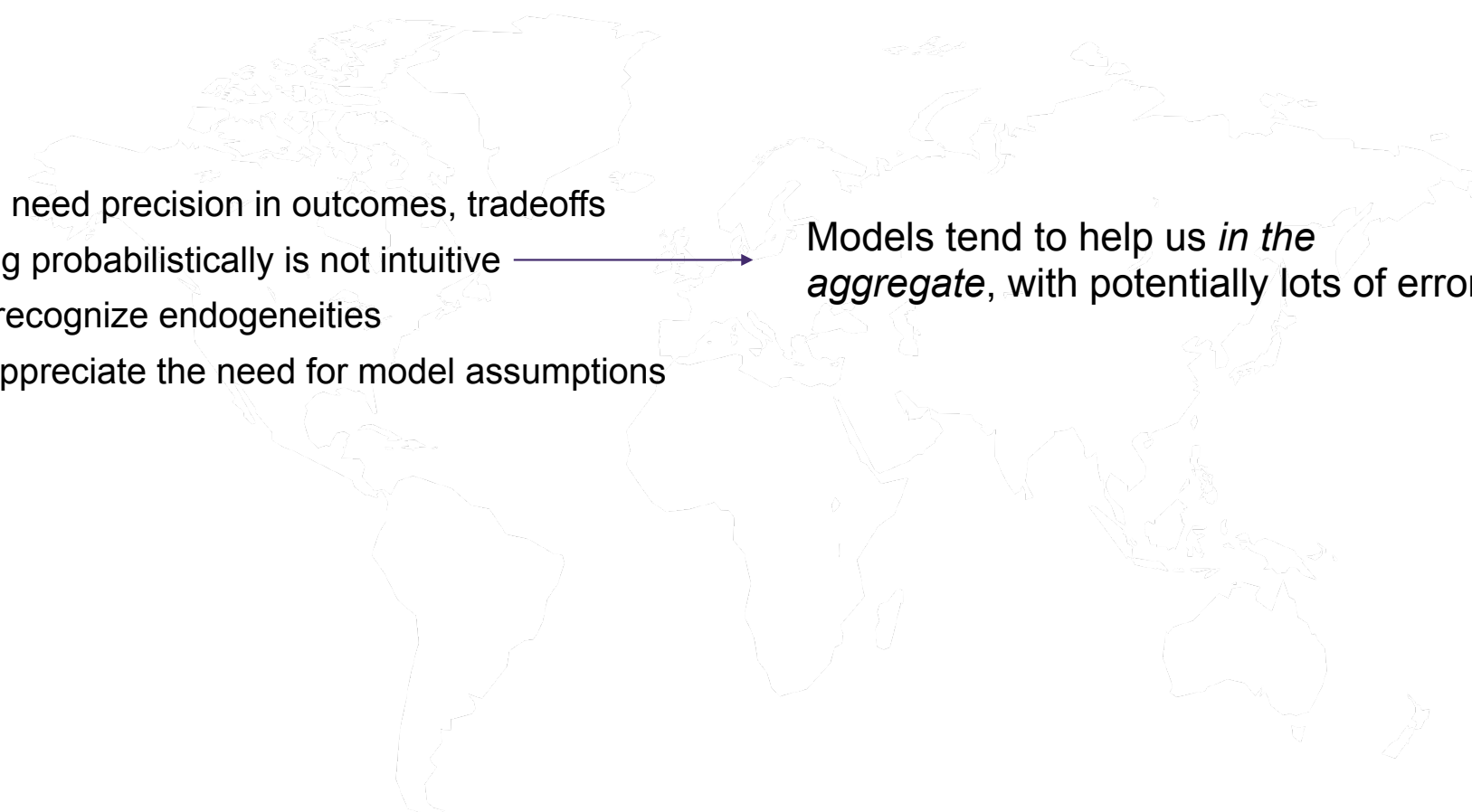
## Who will win the presidency?



### Chance of winning



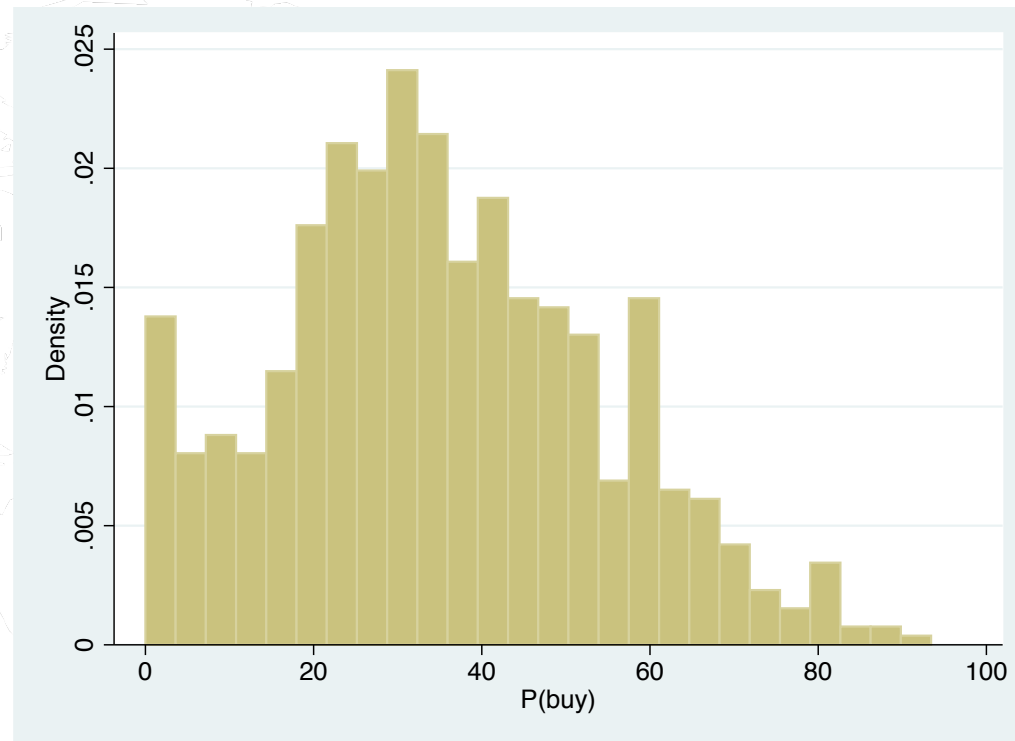
# WHAT'S THE POINT HERE?

- 
1. Models need precision in outcomes, tradeoffs
  2. Thinking probabilistically is not intuitive
  3. Fail to recognize endogeneities
  4. Don't appreciate the need for model assumptions

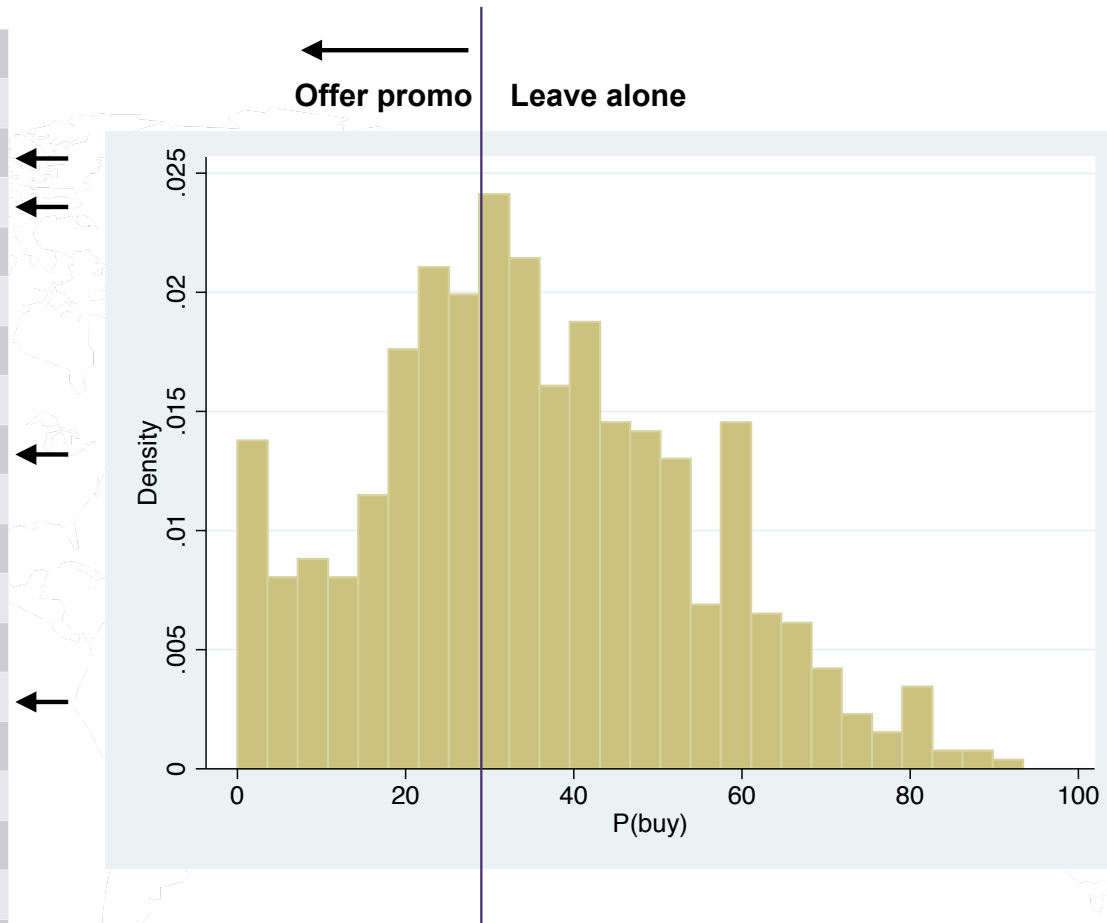
Models tend to help us *in the aggregate*, with potentially lots of error



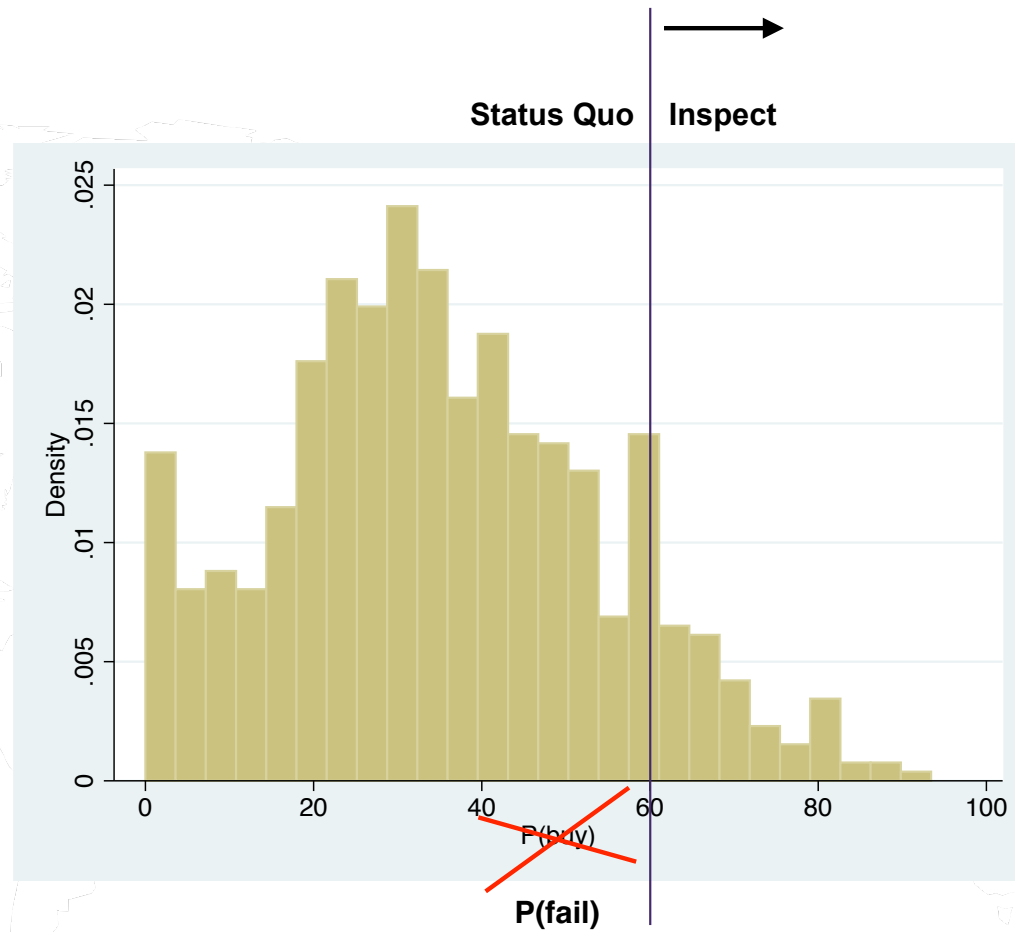
CustomerID	P(buy)
2805341	0.678
6291929	0.297
4560378	0.015
1379279	0.499
2422622	0.673
3571303	0.336
7500176	0.515
7437015	0.041
1245047	0.884
2283893	0.886
1901072	0.686
7263443	0.452
6244271	0.247
9966029	0.382
5391105	0.562
8419109	0.955
2544554	0.382
4706292	0.899



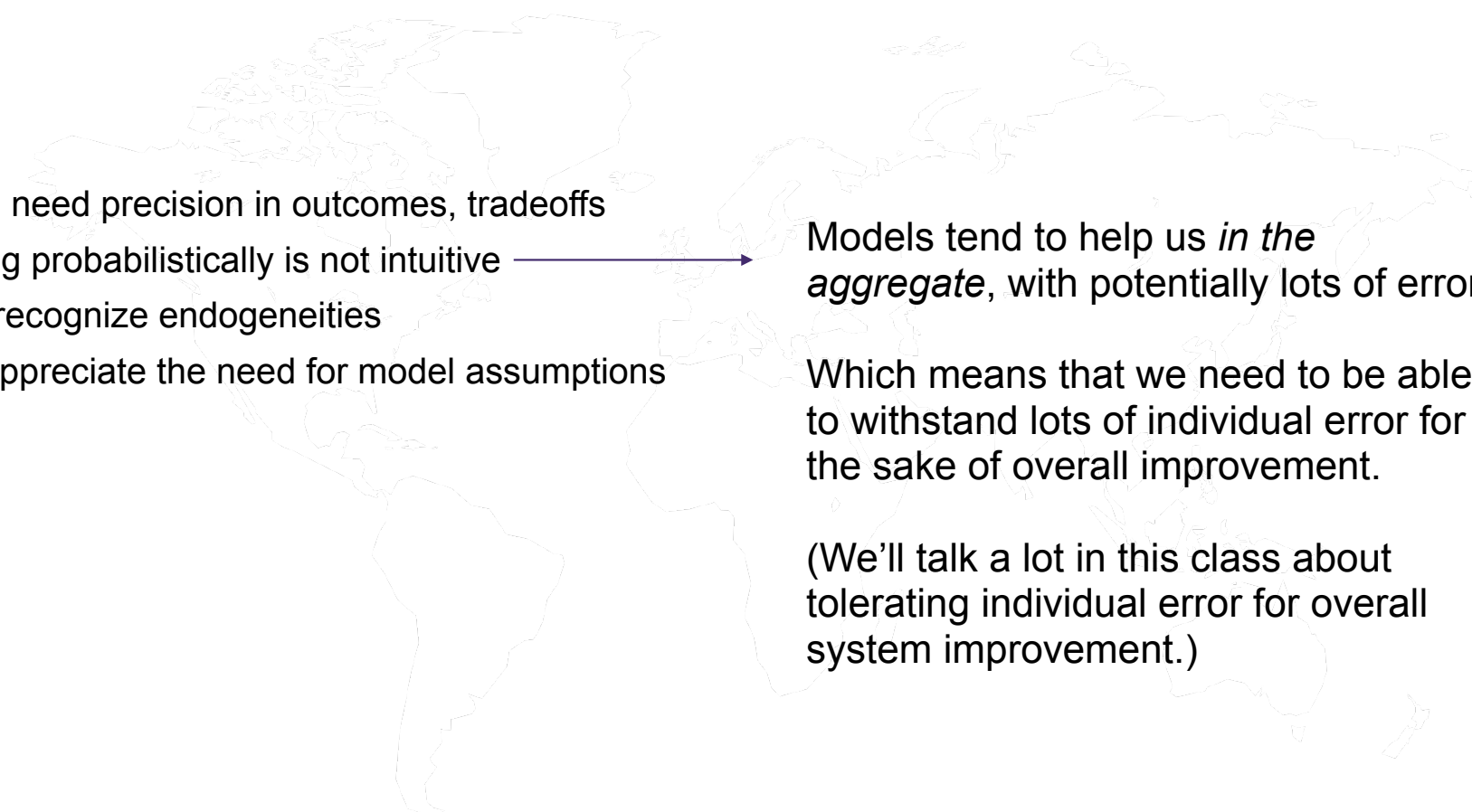
CustomerID	P(buy)
2805341	0.678
6291929	0.297
4560378	0.015
1379279	0.499
2422622	0.673
3571303	0.336
7500176	0.515
7437015	0.041
1245047	0.884
2283893	0.886
1901072	0.686
7263443	0.452
6244271	0.247
9966029	0.382
5391105	0.562
8419109	0.955
2544554	0.382
4706292	0.899



<del>MachineID</del>	<del>P(fail)</del>
CustomerID	P(buy)
2805341	0.678
6291929	0.297
4560378	0.015
1379279	0.499
2422622	0.673
3571303	0.336
7500176	0.515
7437015	0.041
1245047	0.884
2283893	0.886
1901072	0.686
7263443	0.452
6244271	0.247
9966029	0.382
5391105	0.562
8419109	0.955
2544554	0.382
4706292	0.899



# WHAT'S THE POINT HERE?

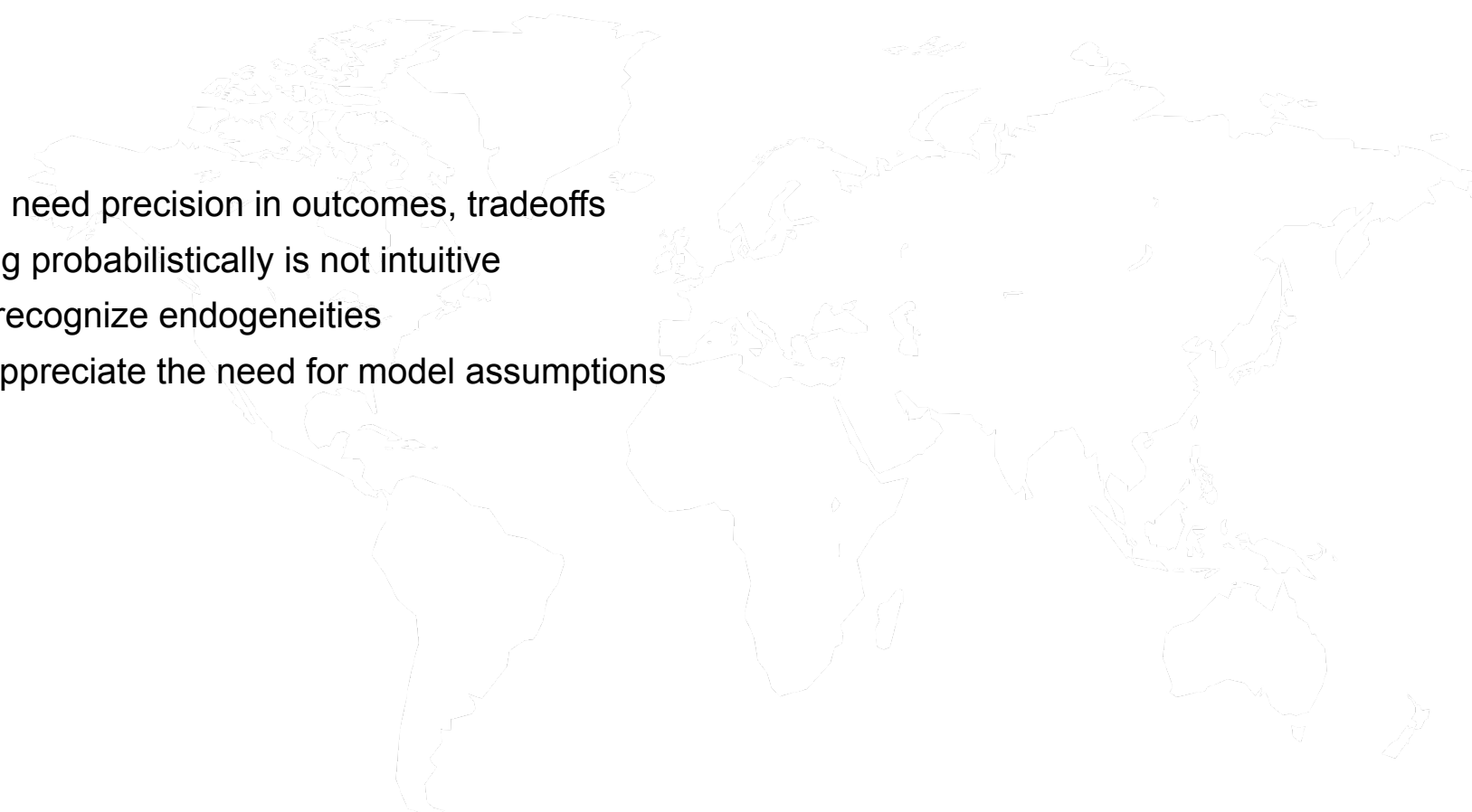
- 
1. Models need precision in outcomes, tradeoffs
  2. Thinking probabilistically is not intuitive
  3. Fail to recognize endogeneities
  4. Don't appreciate the need for model assumptions

Models tend to help us *in the aggregate*, with potentially lots of error

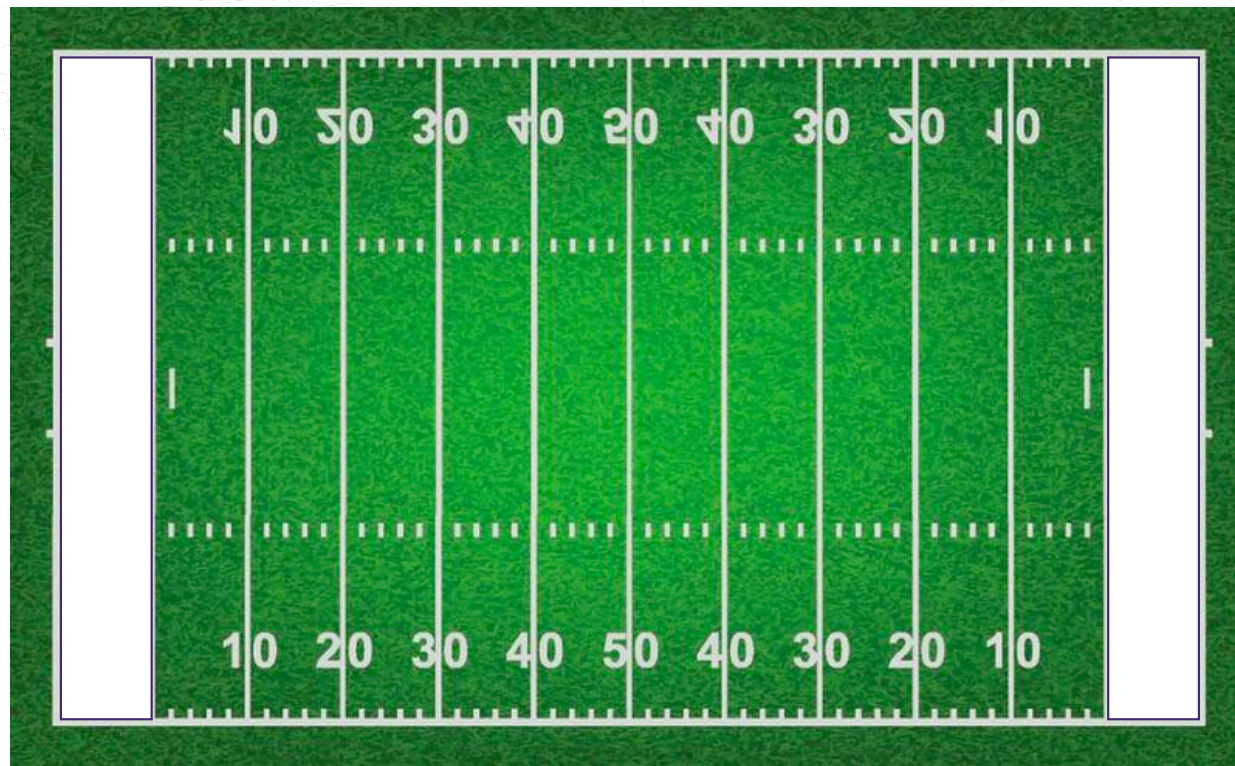
Which means that we need to be able to withstand lots of individual error for the sake of overall improvement.

(We'll talk a lot in this class about tolerating individual error for overall system improvement.)

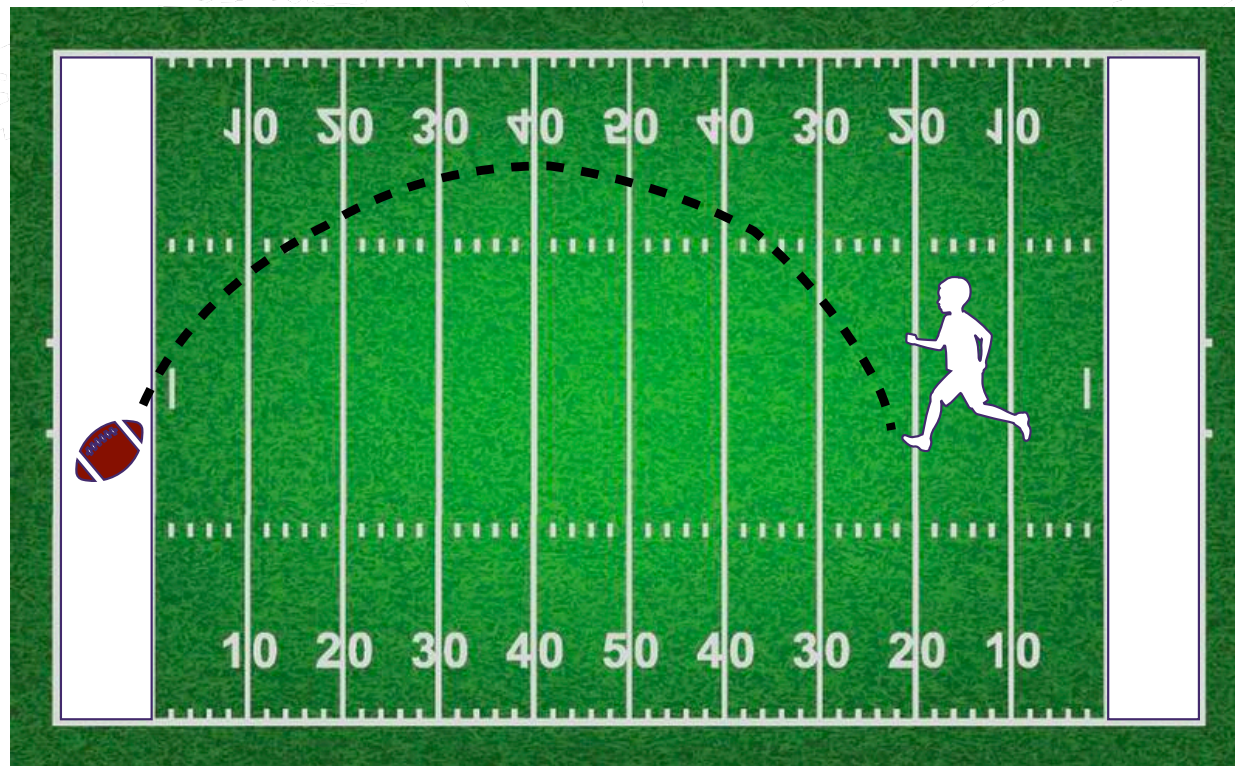
# SO WHY DO BIZ LEADERS STRUGGLE WITH MODELS?

- 
1. Models need precision in outcomes, tradeoffs
  2. Thinking probabilistically is not intuitive
  - ✓ 3. Fail to recognize endogeneities
  4. Don't appreciate the need for model assumptions

# CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL

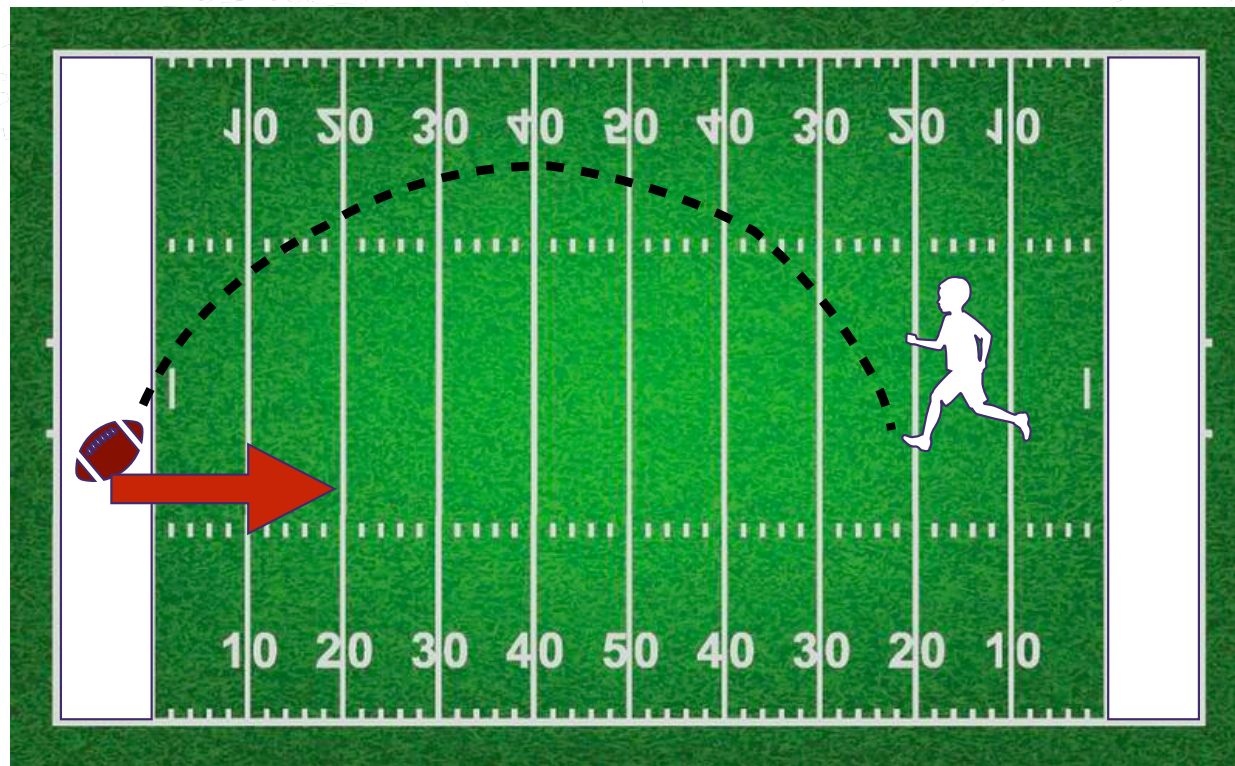


# CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL



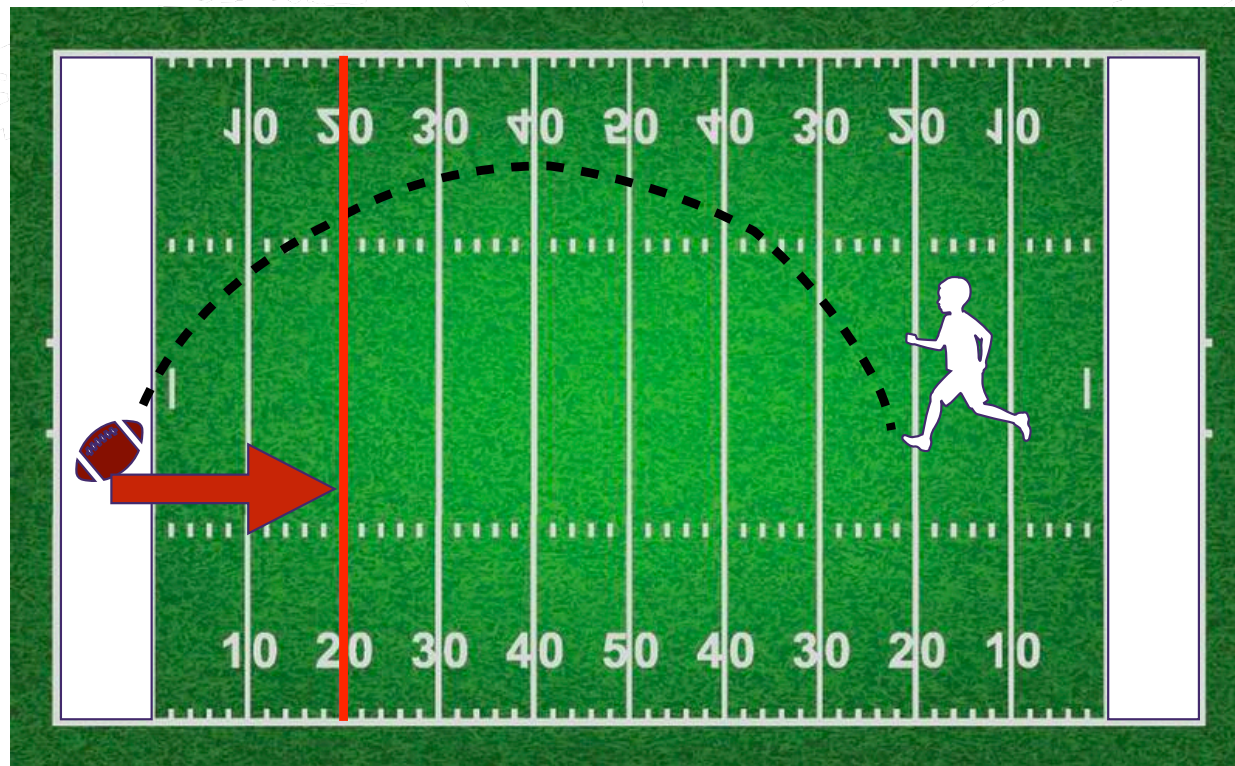


# CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL

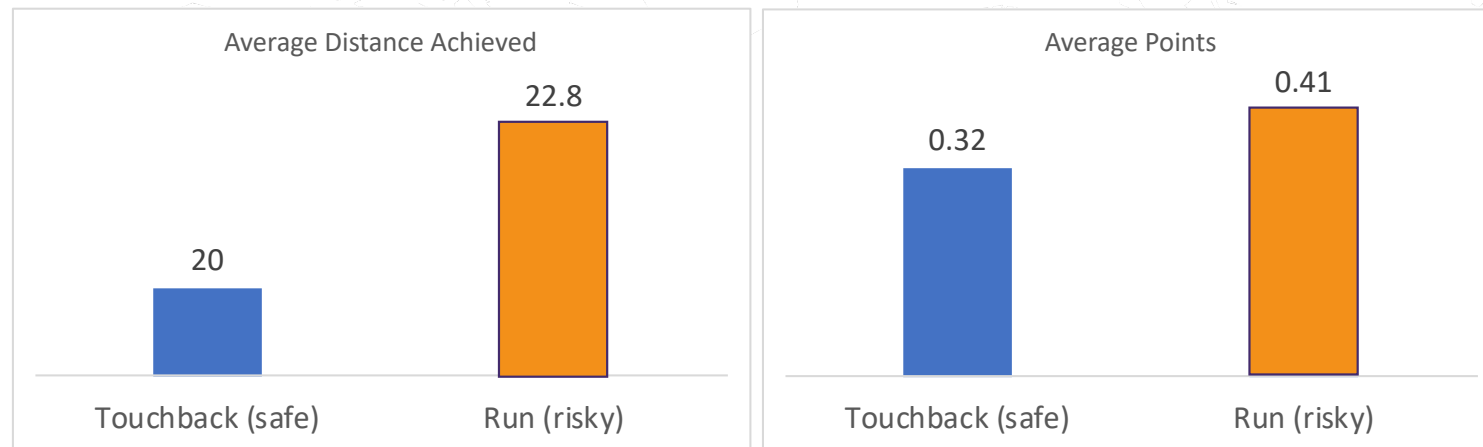




# CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL



# CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL



**The “high risk” decision is consistently better than the safe play.**

# CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL

## We DO know:

Conditions = great!  
Decision = run back (risky)



Conditions = not great!  
Decision = touchback (safe)



## We DON'T know:

Conditions = not great!  
Decision = run back (risky)



**Thus, it isn't reasonable to tell more players to make risky decision!**

# UNRECOGNIZED ENDOGENEITIES / OVB = BIGGEST RISK TO GOOD CRITICAL THINKING WITH DATA & MODELS

**You've encountered endogeneities with statistics:**

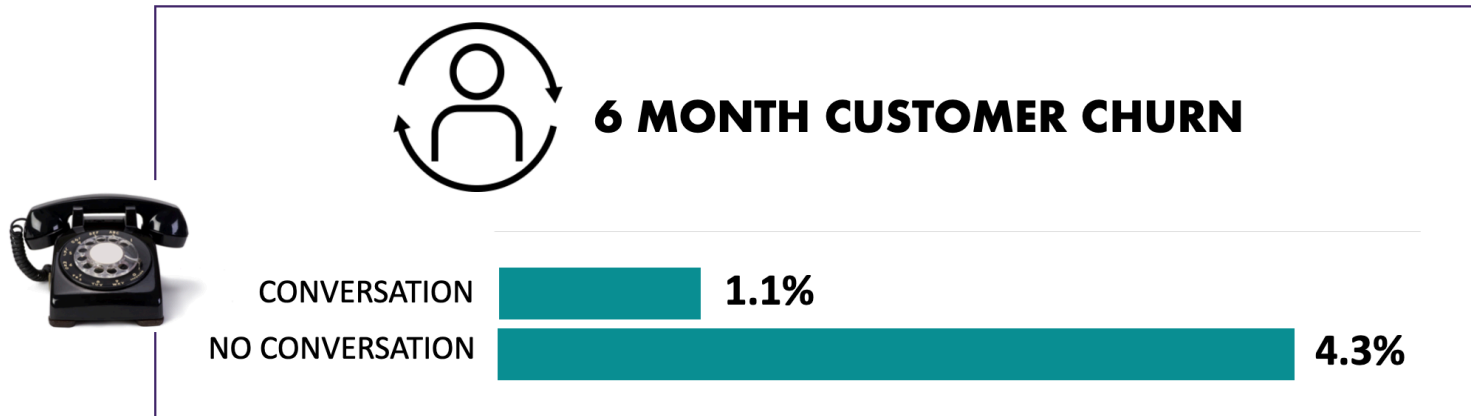
$$y = b_0 + b_1 \cdot x + e$$

Endogeneity if these  
aren't independent

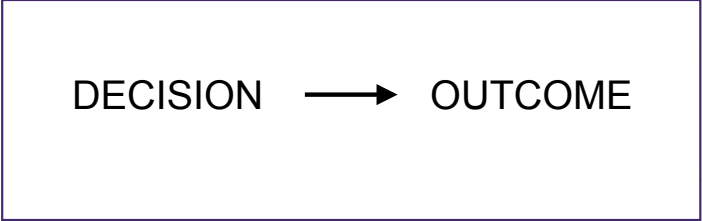
$$\text{Points} = b_0 + b_1 \cdot \text{Touchback?} + e$$

The conditions in which you make  
a decision influence your decision  
AND the outcome, independent of  
your decision

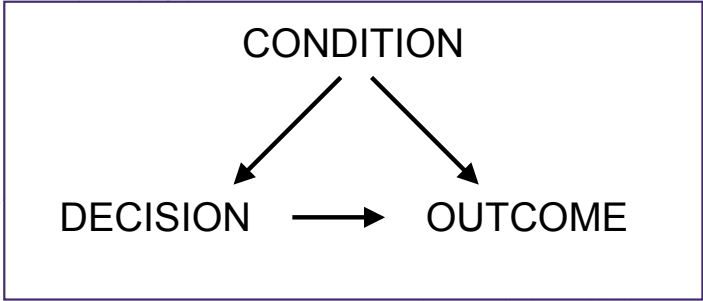
# UNRECOGNIZED ENDOGENEITIES / OVB = BIGGEST RISK TO GOOD CRITICAL THINKING WITH DATA & MODELS



# UNRECOGNIZED ENDOGENEITIES / OVB = BIGGEST RISK TO GOOD CRITICAL THINKING WITH DATA & MODELS



DECISION → OUTCOME

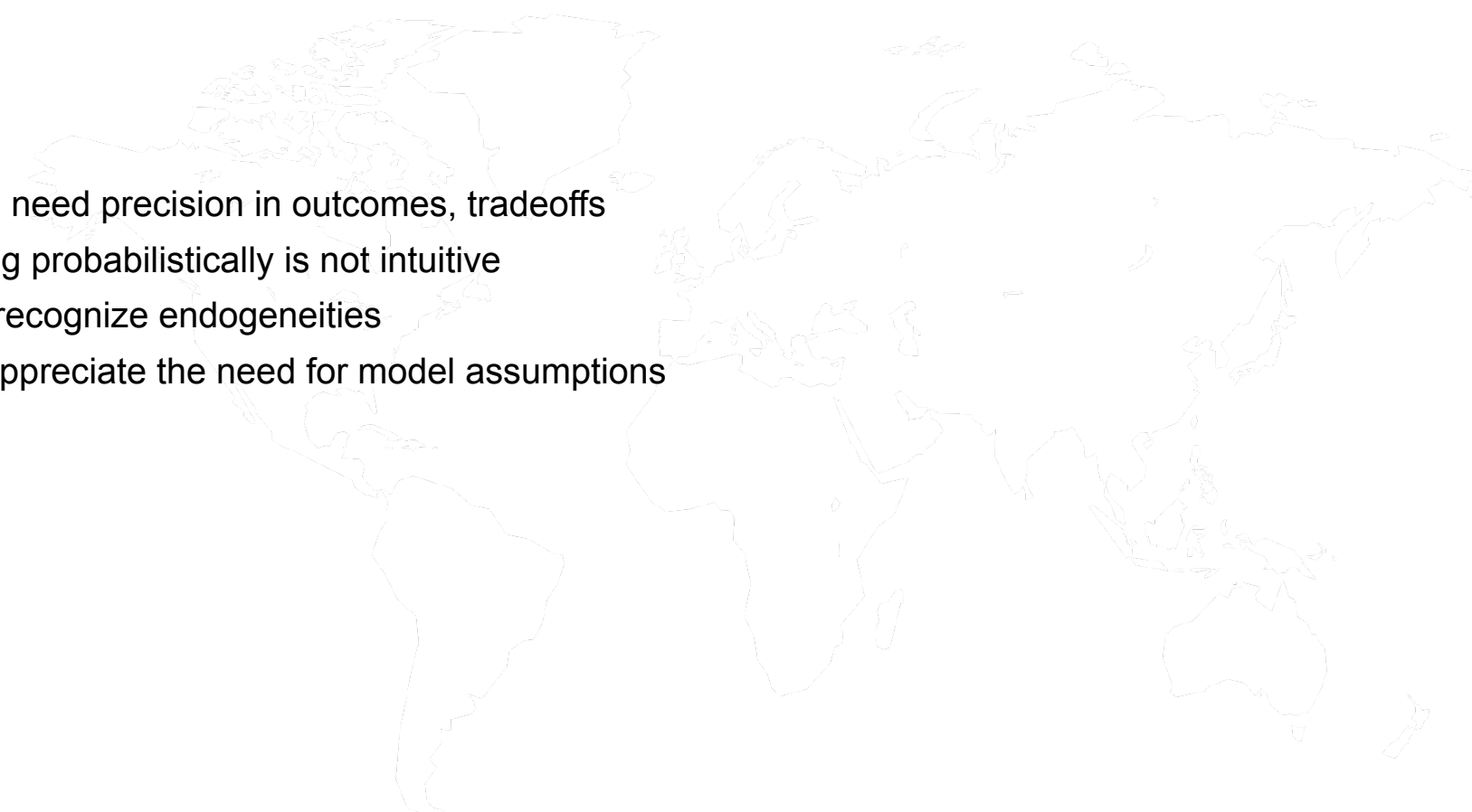


CONDITION  
↓ ↓  
DECISION → OUTCOME

Did decision cause the outcome?  
Or was outcome a byproduct of the condition,  
which caused the outcome?

But this happens ALL THE TIME. Business decision-makers try to make decisions in the conditions where they are most likely to be successful!

# SO WHY DO BIZ LEADERS STRUGGLE WITH MODELS?

- 
1. Models need precision in outcomes, tradeoffs
  2. Thinking probabilistically is not intuitive
  3. Fail to recognize endogeneities
  - ✓ 4. Don't appreciate the need for model assumptions

# YOUR HOMEWORK

## Northwestern | Kellogg School of Management

JOEL SHAPIRO

### **Fly or Drive?**

Your job in Chicago has required that that you go to Milwaukee, WI for a meeting at the Milwaukee airport. Is it safer for you to fly or drive?

Assume the following:

- You live very near Chicago's O'Hare airport
- You always wear a seatbelt and are sober while driving
- The distance to Milwaukee is 71 miles
- Last year, there were 40,100 auto traffic deaths in the U.S.
- 90% of plane accidents occur during the takeoff or landing phase
- There are approximately 0.2 deaths per 10 billion miles flown
- The risk of being in a fatal car crash while sober and wearing a seatbelt is only 20% of the risk for the general population

Provide a complete answer to "is it safer to drive or fly?" by doing the following: 1) clearly answer the question and support your answer with the appropriate calculations 2)



## SAMPLE ANSWER

### RISK OF DEATH, DRIVING

#### General Risk

(12,000 mi) x (300M people)  
=  $3.6 \times 10^{12}$  total miles driven

(40,100 deaths/yr) / ( $3.6 \times 10^{12}$  mile/yr)  
=  $1.114 \times 10^{-8}$  deaths/mi

#### Chi:Mil

( $1.114 \times 10^{-8}$  deaths/mi) x 71 mi x 0.2  
=  **$1.581 \times 10^{-7}$  deaths / trip**

For this 1 trip, I expect 0.0000001581 deaths

### RISK OF DEATH, FLYING

#### General Risk

0.2 deaths /  $1 \times 10^{10}$  mi  
=  $2 \times 10^{-11}$  deaths / mi

#### Chi:Mil

( $2 \times 10^{-11}$  deaths / mi) x 71 mi =  
=  **$1.4 \times 10^{-9}$  deaths / trip**

For this 1 trip, I expect 0.0000000014 deaths

$$p(\text{driving death}) = 113 * p(\text{flying death})$$

# ASSUMPTIONS THAT CONCERN ME

## RISK OF DEATH, DRIVING

### General Risk

(12,000 mi) x (300M people)

=  $3.6 \times 10^{12}$  total miles driven

(40,100 deaths/yr) / ( $3.6 \times 10^{12}$  mile/yr)

=  $1.114 \times 10^{-8}$  deaths/mi

### Chi:Mil

( $1.114 \times 10^{-8}$  deaths/mi) x 71 mi x 0.2

=  $1.581 \times 10^{-7}$  deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

## RISK OF DEATH, FLYING

### General Risk

0.2 deaths /  $1 \times 10^{10}$  mi

=  $2 \times 10^{-11}$  deaths / mi

### Chi:Mil

( $2 \times 10^{-11}$  deaths / mi) x 71 mi =

=  $1.4 \times 10^{-9}$  deaths / trip

For this 1 trip, I expect 0.0000000014 deaths

**Overestimate by factor of 2? 4?**

$$p(\text{driving death}) = 113 * p(\text{flying death})$$

# ASSUMPTIONS THAT CONCERN ME

## RISK OF DEATH, DRIVING

### General Risk

(12,000 mi) x (300M people)

=  $3.6 \times 10^{12}$  total miles driven

(40,100 deaths/yr) / ( $3.6 \times 10^{12}$  mile/yr)

=  $1.114 \times 10^{-8}$  deaths/mi

### Chi:Mil

( $1.114 \times 10^{-8}$  deaths/mi) x **Is this route average riskiness?** x 71 mi =

=  $1.581 \times 10^{-7}$  deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

## RISK OF DEATH, FLYING

### General Risk

0.2 deaths /  $1 \times 10^{10}$  mi

=  $2 \times 10^{-11}$  deaths / mi

### Chi:Mil

=  $1.4 \times 10^{-9}$  deaths / trip

For this 1 trip, I expect 0.0000000014 deaths

$$p(\text{driving death}) = 113 * p(\text{flying death})$$

# ASSUMPTIONS THAT CONCERN ME

## RISK OF DEATH, DRIVING

### General Risk

**(12,000 mi) x (300M people)**

=  $3.6 \times 10^{12}$  total miles driven

(40,100 deaths/yr) / ( $3.6 \times 10^{12}$  mile/yr)

=  $1.114 \times 10^{-8}$  deaths/mi

### Chi:Mil

**( $1.114 \times 10^{-8}$  deaths/mi) x 71 mi x 0.2**

=  $1.581 \times 10^{-7}$  deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

## RISK OF DEATH, FLYING

### General Risk

0.2 deaths /  $1 \times 10^{10}$  mi

=  $2 \times 10^{-11}$  deaths / mi

### Chi:Mil

**( $2 \times 10^{-11}$  deaths / mi) x 71 mi =**

**Is flight risk linearly related to distance?**

For this 1 trip, I expect 0.0000000014 deaths

$$p(\text{driving death}) = 113 * p(\text{flying death})$$

# ASSUMPTIONS THAT CONCERN ME

÷ 4

## RISK OF DEATH, DRIVING

### General Risk

(12,000 mi) x (300M people)  
=  $3.6 \times 10^{12}$  total miles driven

(40,100 deaths/yr) / ( $3.6 \times 10^{12}$  mile/yr)  
=  $1.114 \times 10^{-8}$  deaths/mi

### Chi:Mil

( $1.114 \times 10^{-8}$  deaths/mi) x 71 mi x 0.2  
=  $1.581 \times 10^{-7}$  deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

## RISK OF DEATH, FLYING

### General Risk

0.2 deaths /  $1 \times 10^{10}$  mi  
=  $2 \times 10^{-11}$  deaths / mi

### Chi:Mil

( $2 \times 10^{-11}$  deaths / mi) x 71 mi =  
=  $1.4 \times 10^{-9}$  deaths / trip

For this 1 trip, I expect 0.0000000014 deaths

x 7

~~$p(\text{driving death}) = 113 * p(\text{flying death})$~~

$p(\text{driving death}) = p(\text{flying death})$

# ASSUMPTIONS THAT CONCERN ME

## RISK OF DEATH, DRIVING

### General Risk

(12,000 mi) x (300M people)

=  $3.6 \times 10^{12}$  total miles driven

(40,100 deaths/yr) / ( $3.6 \times 10^{12}$  mile/yr)

=  $1.114 \times 10^{-8}$  deaths/mi

### Chi:Mil

**x 10** ( $1.114 \times 10^{-8}$  deaths/mi) x 71 mi x 0.2

=  $1.581 \times 10^{-7}$  deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

## RISK OF DEATH, FLYING

### General Risk

0.2 deaths /  $1 \times 10^{10}$  mi

=  $2 \times 10^{-11}$  deaths / mi

### Chi:Mil

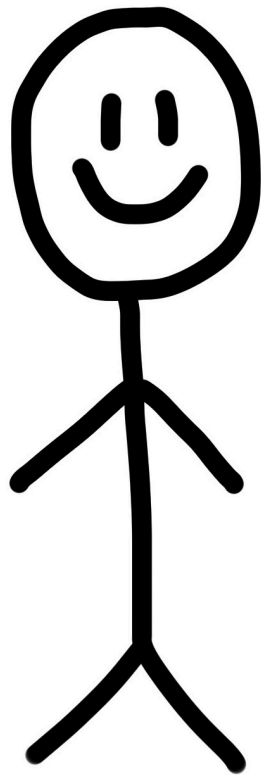
( $2 \times 10^{-11}$  deaths / mi) x 71 mi =

=  $1.4 \times 10^{-9}$  deaths / trip

For this 1 trip, I expect 0.0000000014 deaths

~~$p(\text{driving death}) = 113 * p(\text{flying death})$~~

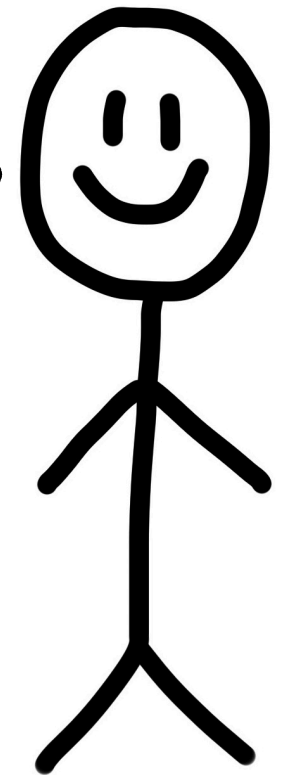
$p(\text{driving death}) = 1000 * p(\text{flying death})$



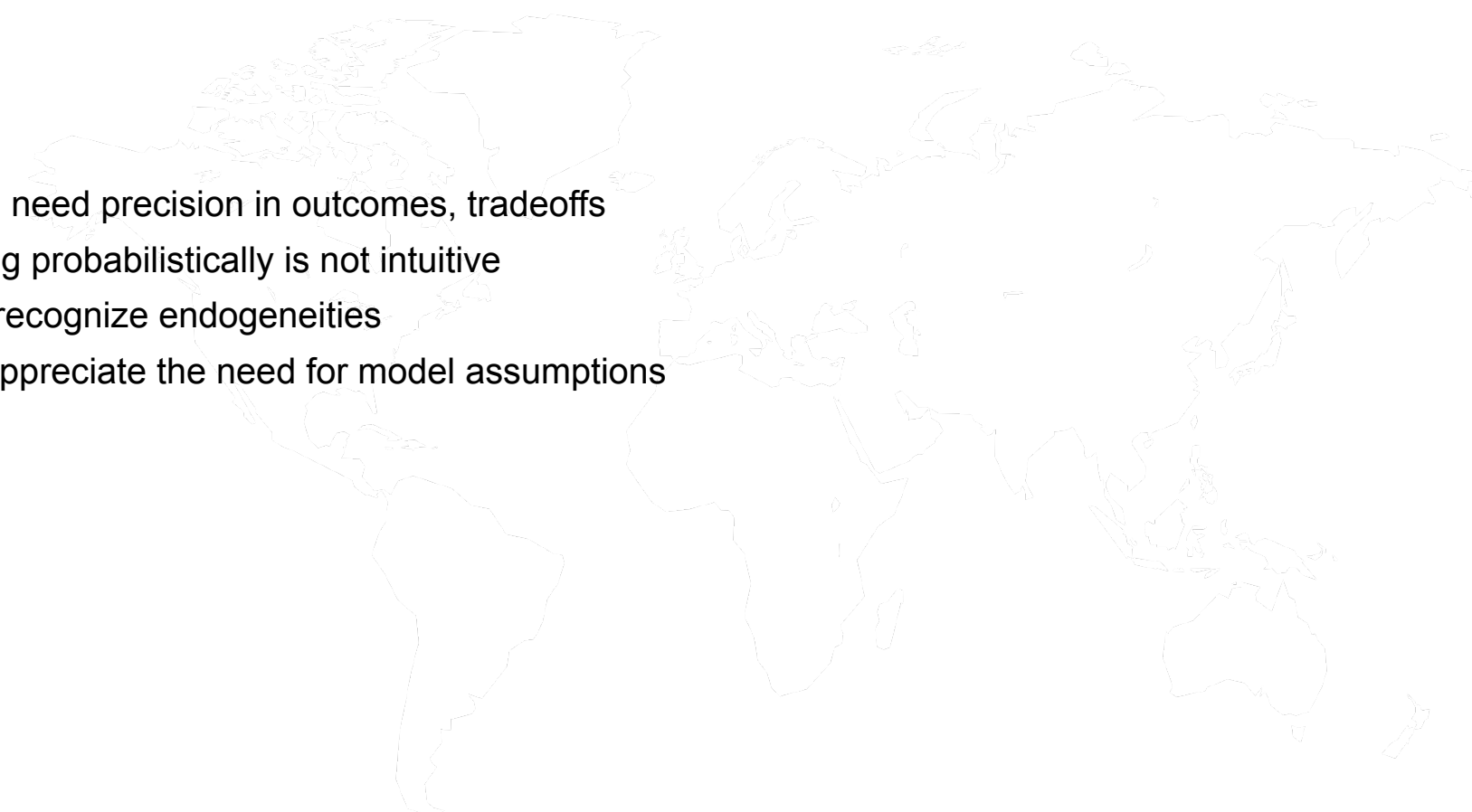
Our new strategy will get us better outcomes for a fraction of the risk!

Really?

Or worse outcomes and more risk. Hard to tell.



# YOUR ROLE AS A GREAT DATA SCIENTIST

- 
1. Models need precision in outcomes, tradeoffs
  2. Thinking probabilistically is not intuitive
  3. Fail to recognize endogeneities
  4. Don't appreciate the need for model assumptions