MSiA

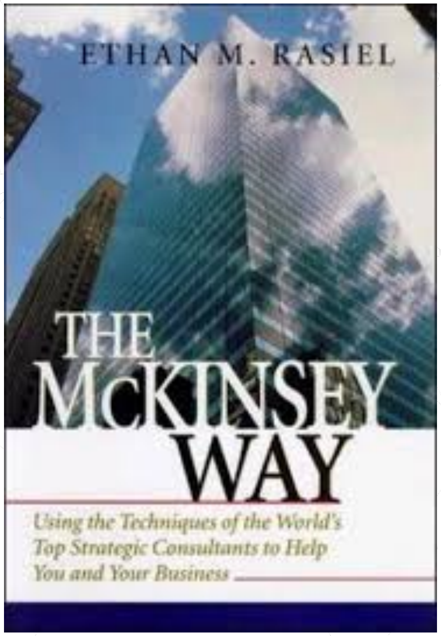
CLASS 2: MODELS FOR DECISION-MAKING

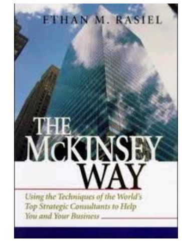
Joel Shapiro
Clinical Associate Professor, MEDS

Northwestern | Kellogg

A FRAMEWORK FOR RIGOROUS DECISION-MAKING IS NOT NEW

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





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 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.

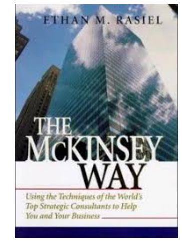
Prospect A is 1% likely to convert

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 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

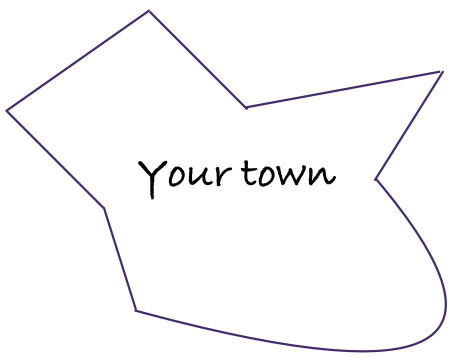
If you have the resources to send only ONE prospect to the sales team, which one should it be?

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.

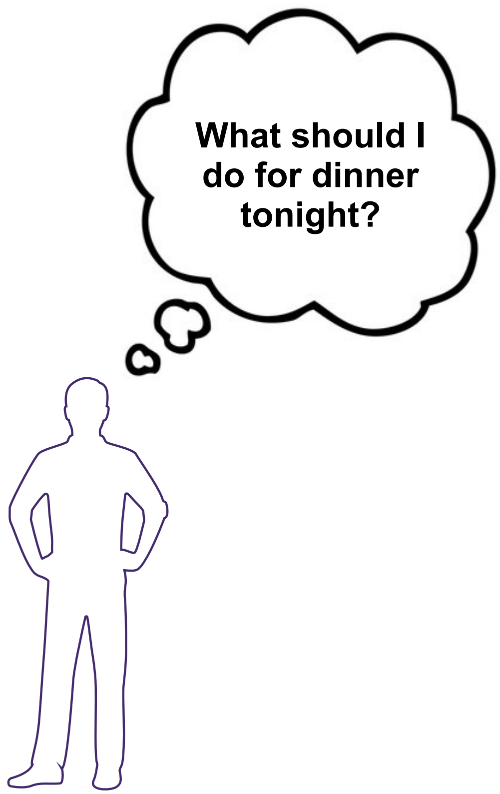
MODELS ENABLE DECISION-MAKING

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

maximize
minimize f (x,y,z…) s.t. constraints

You might know these as "constrained optimization" problems.

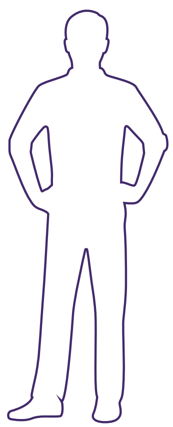
I WANT TO HELP YOU FIGURE OUT DINNER TONIGHT!



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.?



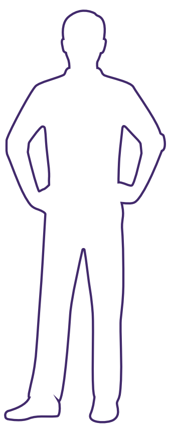
maximize
minimize f (x,y,z…) s.t. constraints

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.?

$$\begin{matrix}\textit{maximize} \\ \textit{minimize}\end{matrix} \quad f\,(x,y,z\ldots) \quad \text{s.t. constraints}$$



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**.

maximize
minimize f (x,y,z…) 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**.

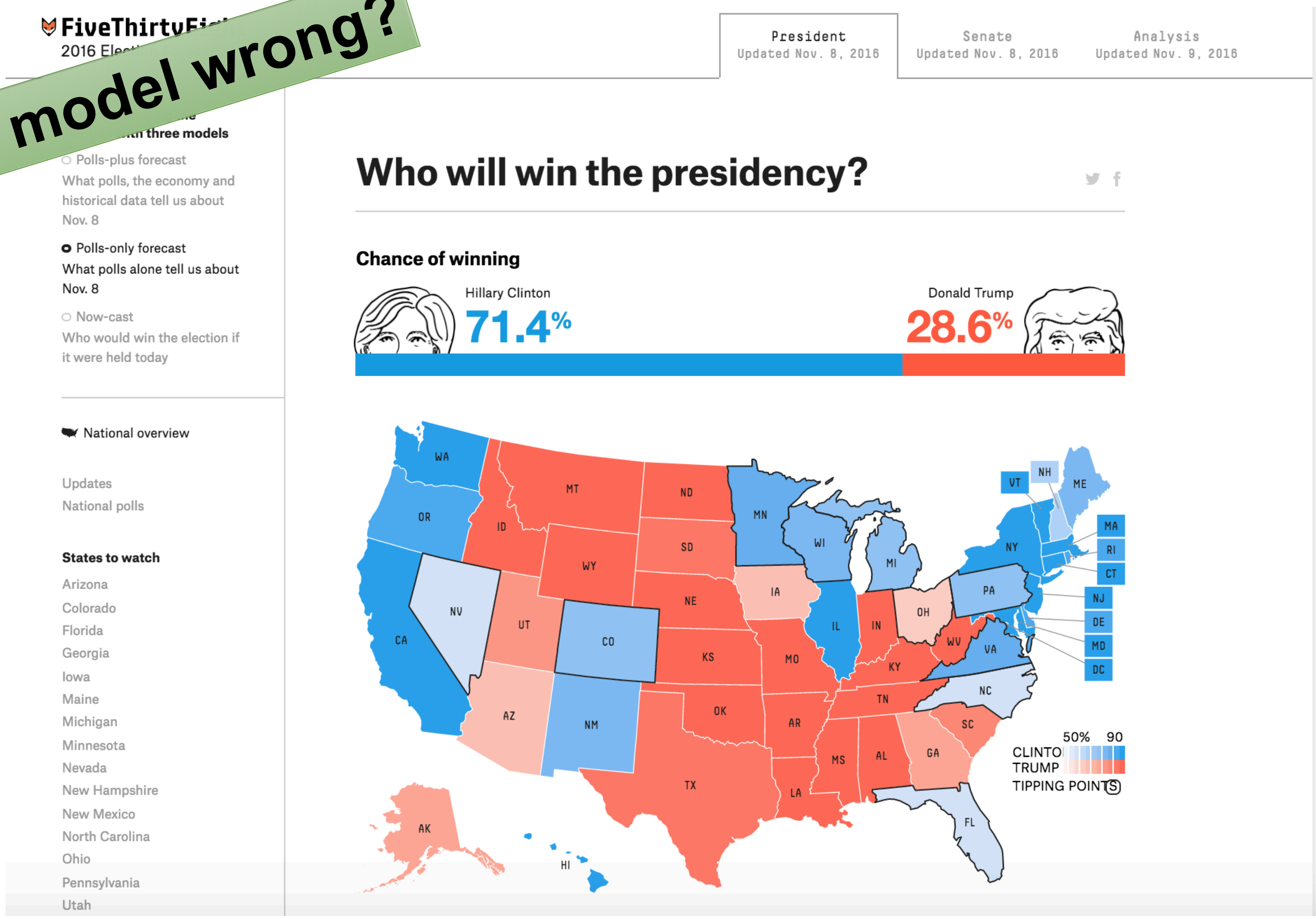
SO WHY DO BIZ LEADERS STRUGGLE WITH MODELS?

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?



CONSIDER THIS EXAMPLE

Cell Block A houses 25 violent criminals, 24 of whom, acting together in the prison yard, kill one of the guards. The 25th prisoner does not participate but cannot be identified by video cameras. An investigation fails to yield additional information because none of the prisoners will talk.

Nevertheless, the District Attorney selects one prisoner ("P") at random and charges him with murder. At trial, the DA presents uncontroverted evidence that:

- P was in the prison yard at that time and
- All but one prisoner participated in the killing.

Would you convict P?

Thanks to Jonathan Koehler, NU Law evidence professor for this example

AND THIS ONE…

Imagine that I've built a new diagnostic predictive tool to help the MSiA admissions team identify which prospective students have lied on their application. The test has a 97% true positive rate and a 97% true negative rate.

In the next round of admissions, the tool identifies someone as a liar.

Should MSiA reject that applicant? What are the chances that person lied?

AND THIS ONE…

Imagine that I've built a new diagnostic predictive tool to help the MSiA admissions team identify which prospective students have lied on their application. The test has a 97% true positive rate and a 97% true negative rate.

In the next round of admissions, the tool identifies someone as a liar.

Should MSiA reject that applicant? What are the chances that person lied?

BAYES THEOREM

$$P(\text{Liar} \mid +) = \frac{P(+ \mid \text{Liar})*P(\text{Liar})}{P(+)} = \frac{P(+ \mid \text{Liar})*P(\text{Liar})}{P(+ \mid \text{Liar})*P(\text{Liar}) + P(+ \mid \text{NotLiar})*P(\text{NotLiar})}$$

AND THIS ONE…

Imagine that I've built a new diagnostic predictive tool to help the MSiA admissions team identify which prospective students have lied on their application. The test has a 97% true positive rate and a 97% true negative rate.

In the next round of admissions, the tool identifies someone as a liar.

Should MSiA reject that applicant? What are the chances that person lied?

BAYES THEOREM

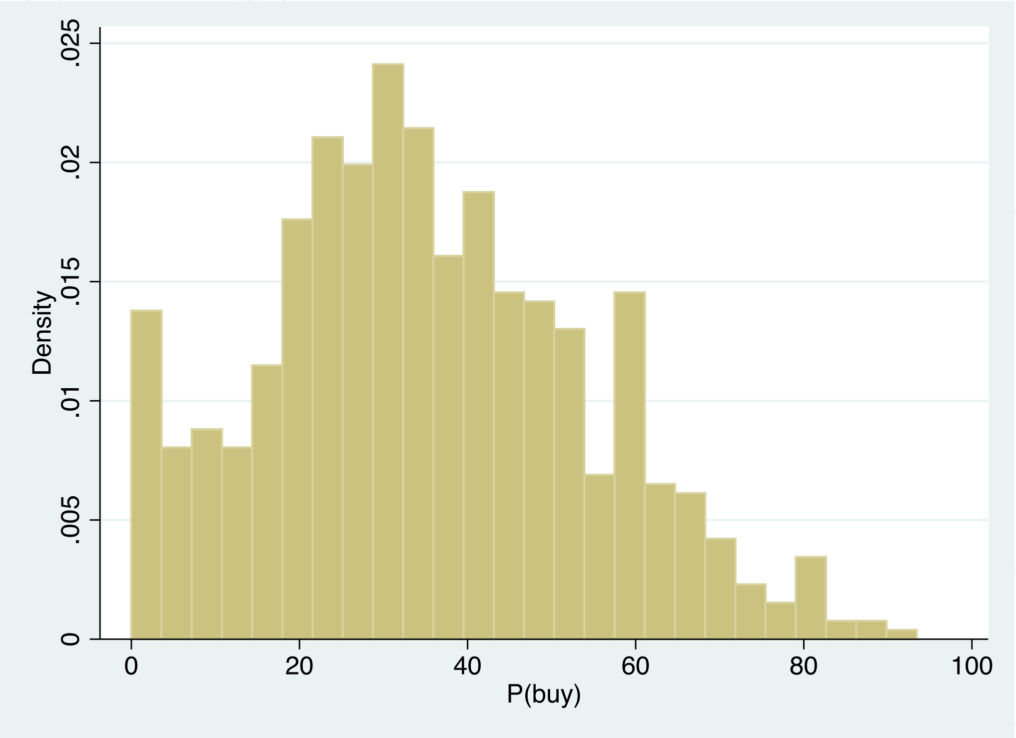
$$P(\text{Liar} \mid +) = \frac{P(+ \mid \text{Liar})*P(\text{Liar})}{P(+)} = \frac{P(+ \mid \text{Liar})*P(\text{Liar})}{P(+ \mid \text{Liar})*P(\text{Liar}) + P(+ \mid \text{NotLiar})*P(\text{NotLiar})}$$

If p(Liar) = 1%, then p(Liar | +) = 25%
If p(Liar) = 10%, then p(Liar | +) = 78%
If p(Liar) = 30%, then p(Liar | +) = 93%

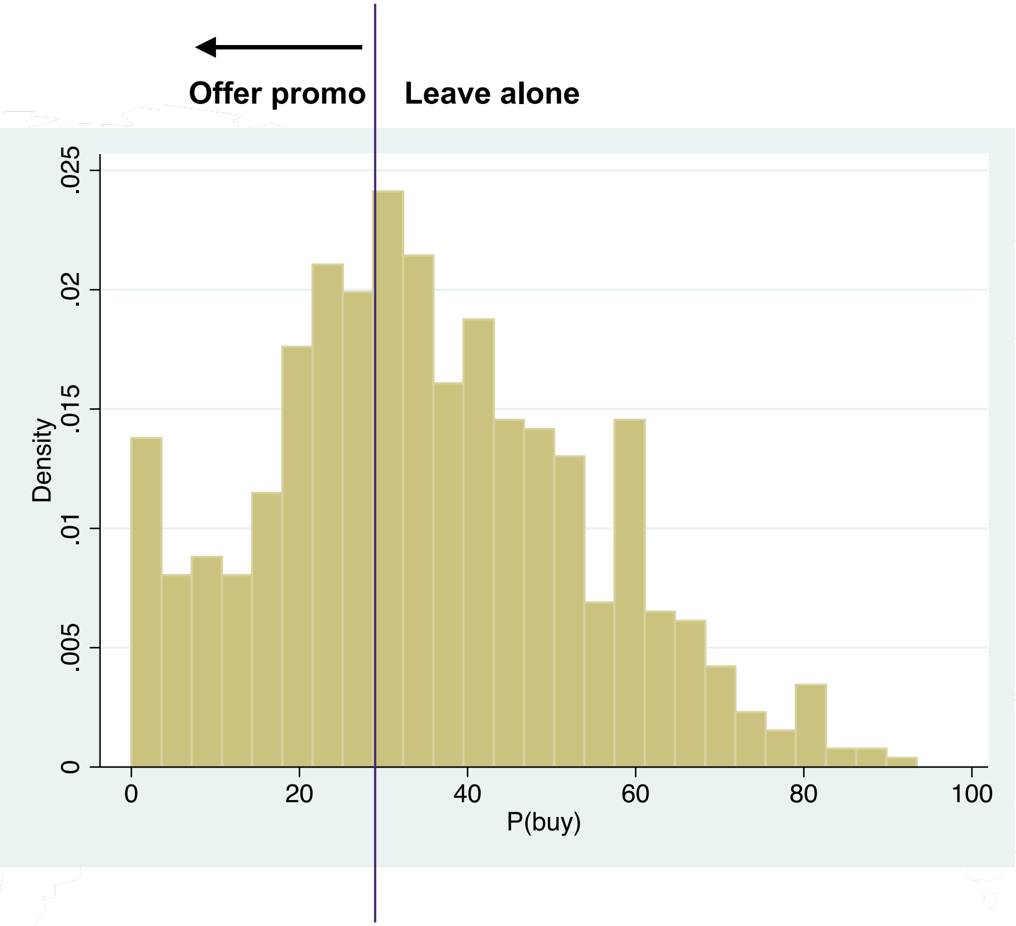
WHAT'S THE POINT HERE?

1. Models need precision in outcomes, tradeoffs
2. Thinking probabilistically is not intuitive → Models tend to help us *in the aggregate*, with potentially lots of error
3. Fail to recognize endogeneities
4. Don't appreciate the need for model assumptions

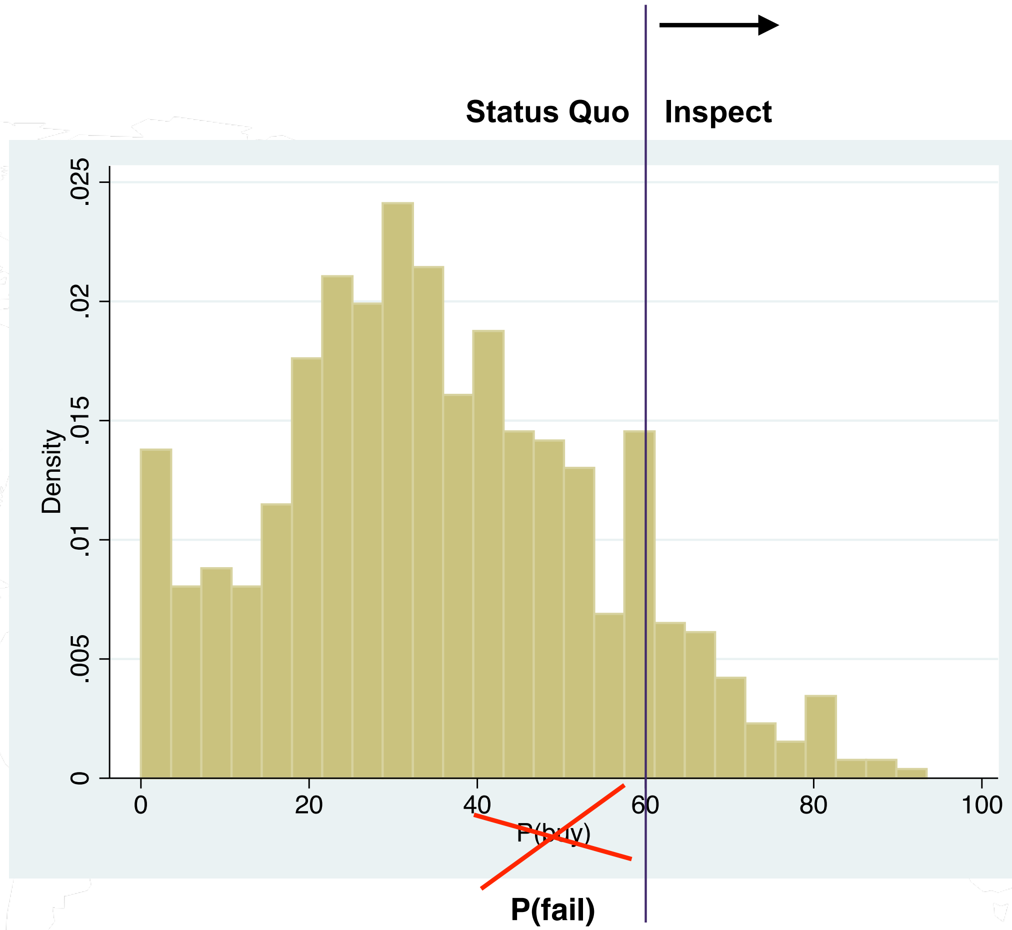
CustomerID	P(buy)
2805341	0.678
6291929	0.297
4560378	0.015
1379279	0.499
2422622	0.673
3571303	0.336
7500176	0.515
7437015	0.041
1245047	0.884
2283893	0.886
1901072	0.686
7263443	0.452
6244271	0.247
9966029	0.382
5391105	0.562
8419109	0.955
2544554	0.382
4706292	0.899



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



MachineID ~~CustomerID~~	P(fail) ~~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

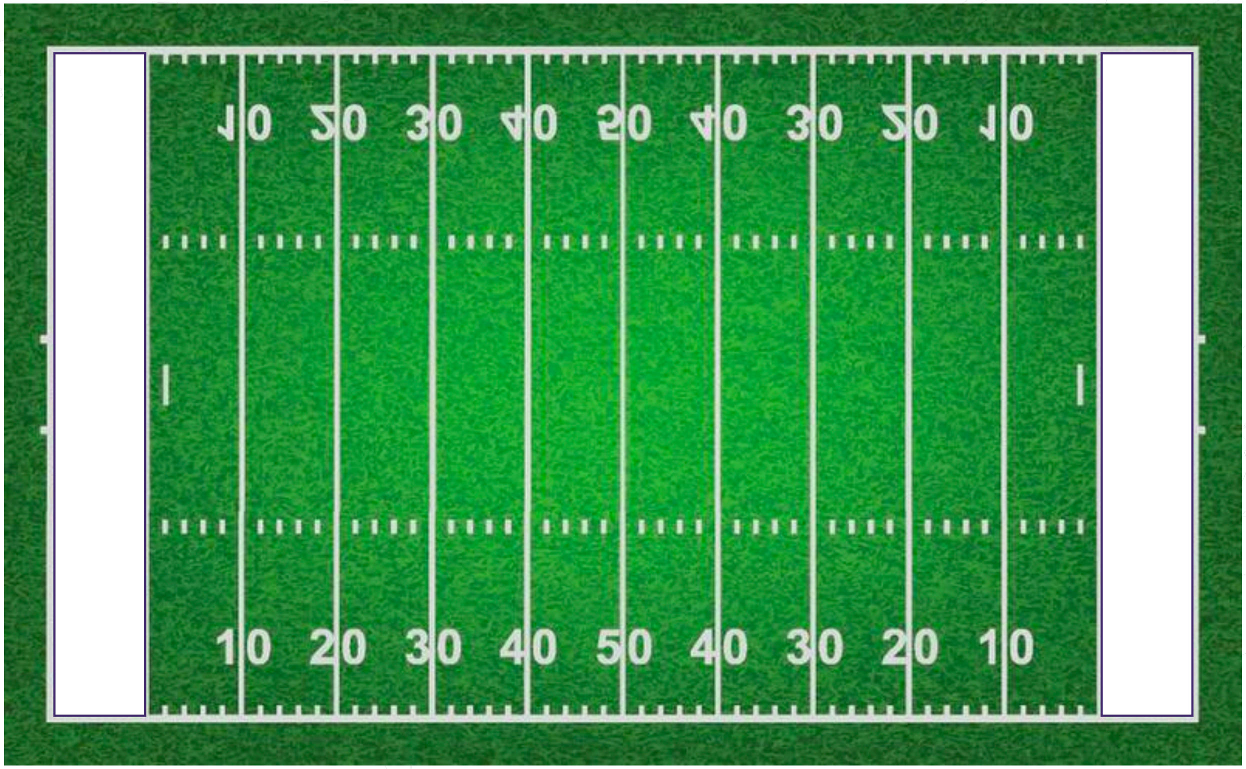
Which means that we need to be able to withstand lots of individual error for the sake of overall improvement.

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

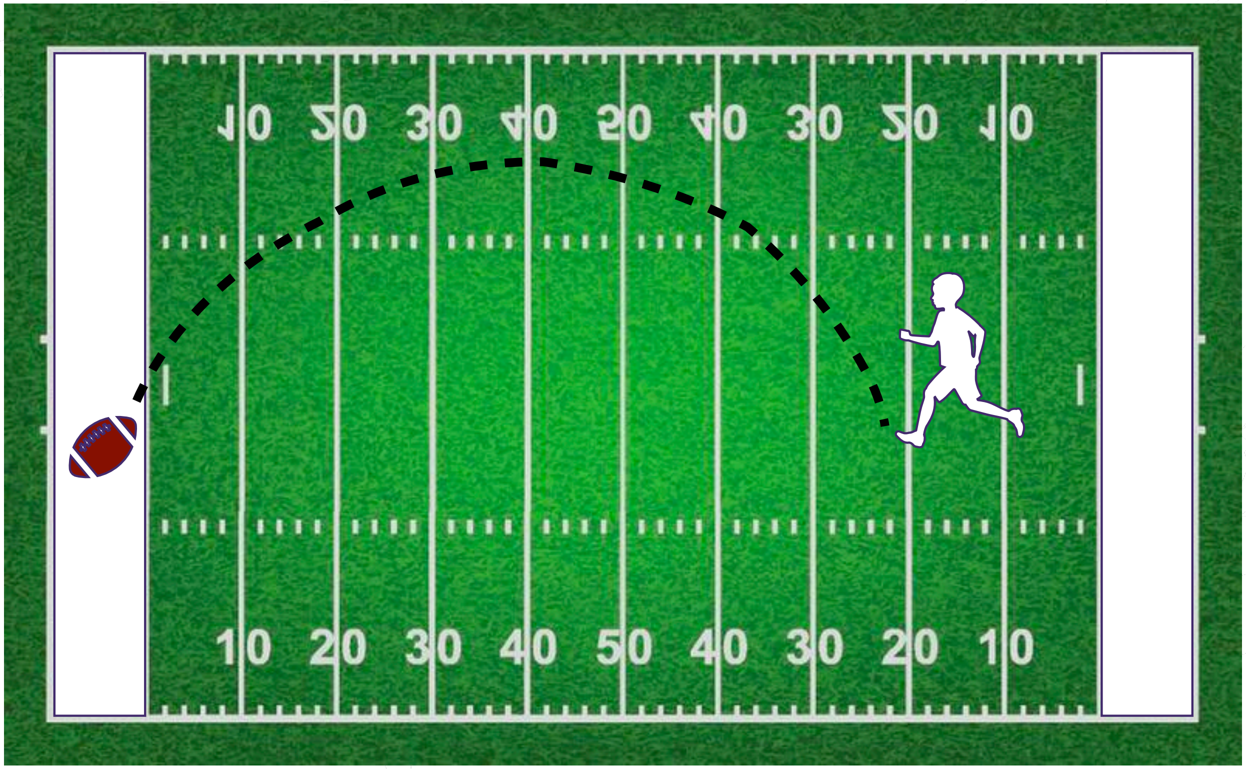
SO WHY DO BIZ LEADERS STRUGGLE WITH MODELS?

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

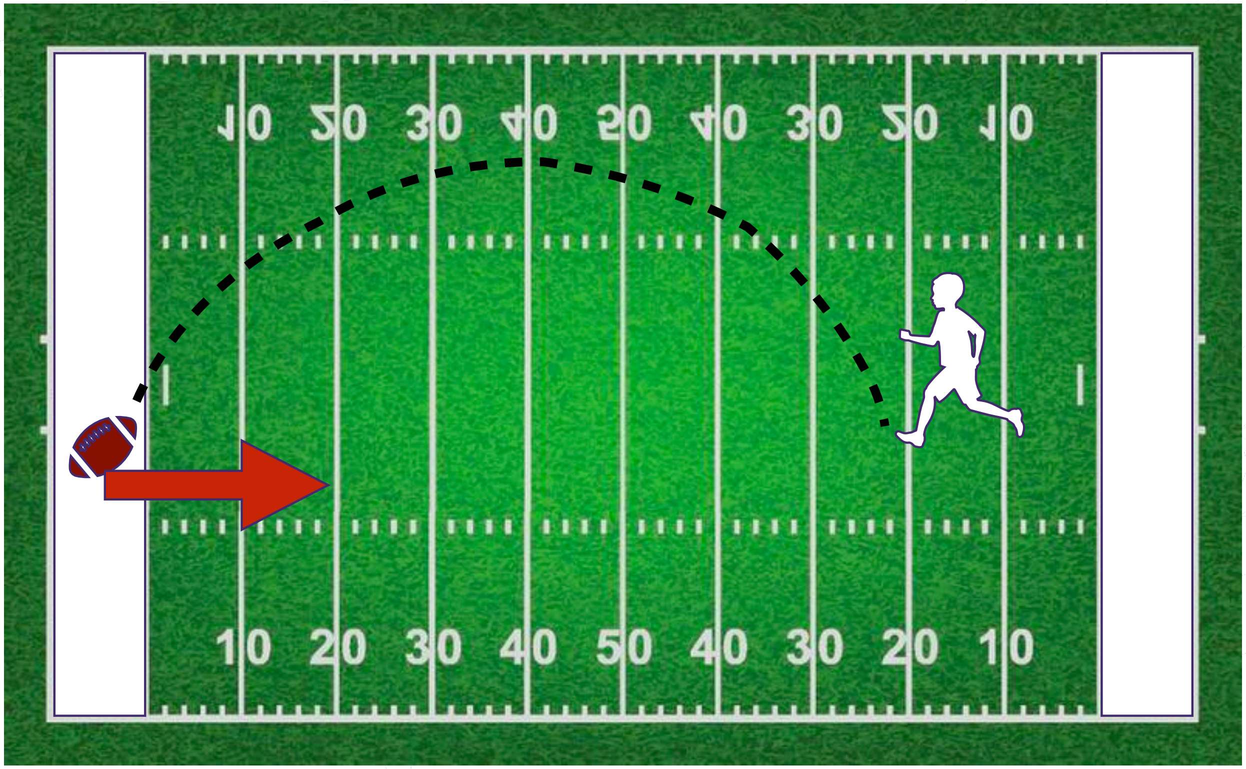
CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL



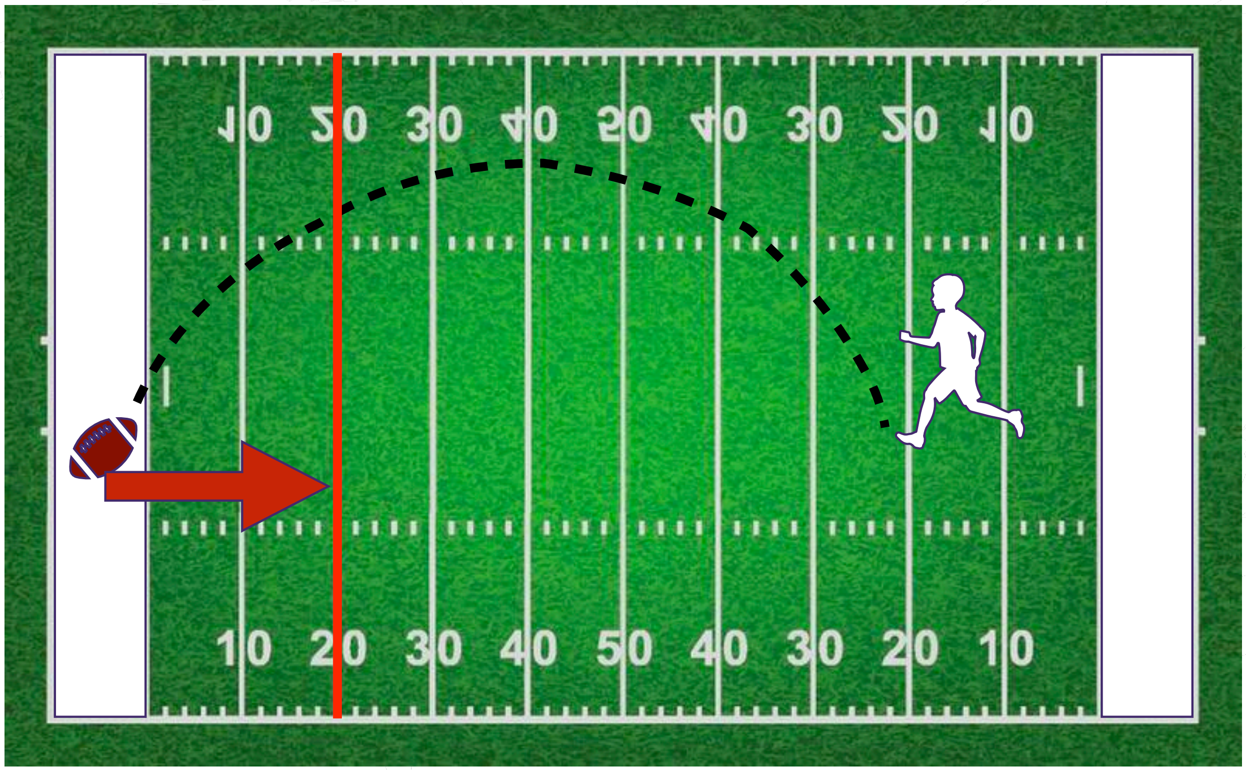
CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL



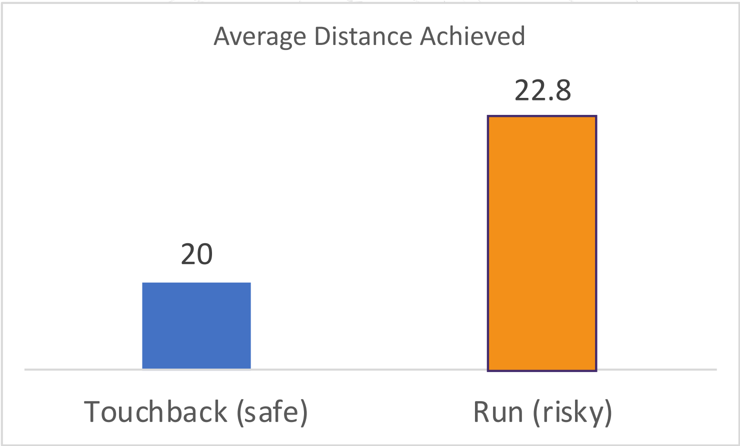
CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL

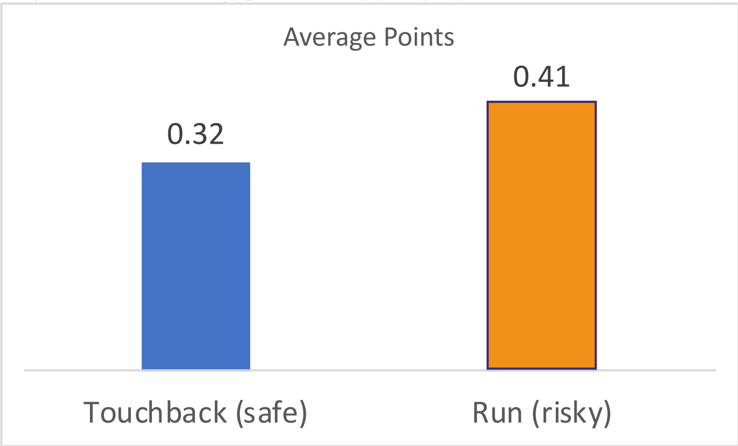


CONSIDER THIS ENDOGENEITY IN AMERICAN FOOTBALL



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:

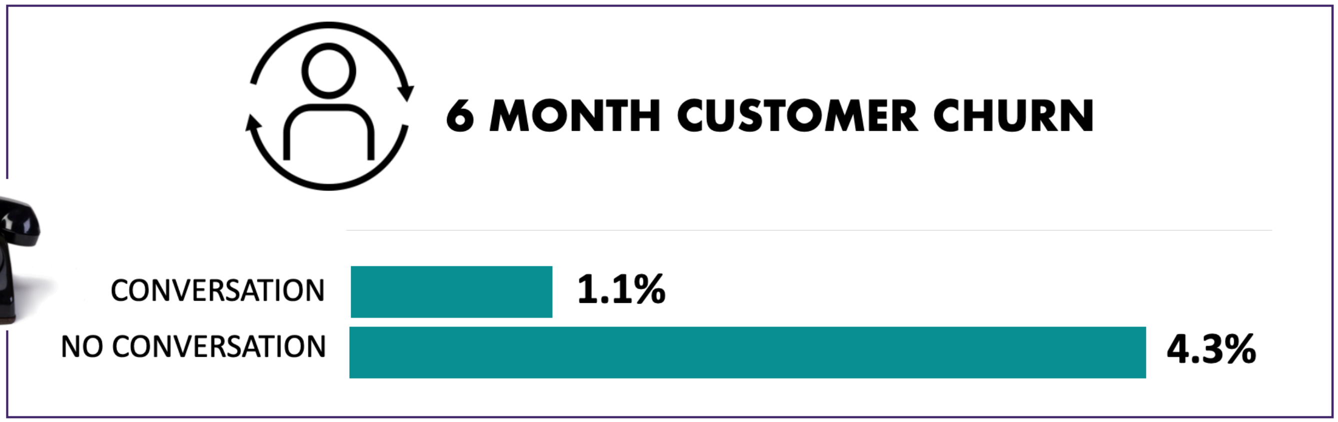
y = b0 + b1*x + e

Endogeneity if these aren’t independent

Points = b0 + b1*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 $\longrightarrow$ OUTCOME

$$\begin{tikzcd}[column sep=small] & \text{CONDITION} \arrow[dl] \arrow[dr] & \\ \text{DECISION} \arrow[rr] & & \text{OUTCOME} \end{tikzcd}$$

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)
= 3.6x10^12 total miles driven

(40,100 deaths/yr) / (3.6x10^12 mile/yr)
= 1.114x10^-8 deaths/mi

Chi:Mil

(1.114x10^-8 deaths/mi) x 71 mi x 0.2
= 1.581 x 10^-7 deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

RISK OF DEATH, FLYING

General Risk

0.2 deaths / 1x10^10 mi
= 2x10^-11 deaths / mi

Chi:Mil

(2x10^-11 deaths / mi) x 71 mi =
= 1.4x10^-9 deaths / trip

For this 1 trip, I expect 0.0000000014 deaths

p(driving death) = 113 * p(flying death)

ASSUMPTIONS THAT CONCERN ME

RISK OF DEATH, DRIVING	RISK OF DEATH, FLYING
General Risk	General Risk
(12,000 mi) x (300M people) = 3.6x10^12 total miles driven	
(40,100 deaths/yr) / (3.6x10^12 mile/yr) = 1.114x10^-8 deaths/mi	0.2 deaths / 1x10^10 mi = 2x10^-11 deaths / mi
Chi:Mil	Chi:Mil
(1.114x10^-8 deaths/mi) x 71 mi x 0.2 **= 1.581 x 10^-7 deaths / trip**	(2x10^-11 deaths / mi) x 71 mi = **= 1.4x10^-9 deaths / trip**
For this 1 trip, I expect 0.0000001581 deaths	For this 1 trip, I expect 0.0000000014 deaths

Overestimate by factor of 2? 4?

p(driving death) = 113 * p(flying death)

ASSUMPTIONS THAT CONCERN ME

RISK OF DEATH, DRIVING

General Risk

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

= 3.6x10^12 total miles driven

(40,100 deaths/yr) / (3.6x10^12 mile/yr)

= 1.114x10^-8 deaths/mi

Chi:Mil

(1.114x10^-8 deaths/mi) x

= 1.581 x 10^-7 deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

RISK OF DEATH, FLYING

General Risk

0.2 deaths / 1x10^10 mi

= 2x10^-11 deaths / mi

Chi:Mil

x 71 mi =

= 1.4x10^-9 deaths / trip

For this 1 trip, I expect 0.0000000014 deaths

Is this route average riskiness?

p(driving death) = 113 * p(flying death)

ASSUMPTIONS THAT CONCERN ME

RISK OF DEATH, DRIVING

General Risk

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

= 3.6x10^12 total miles driven

(40,100 deaths/yr) / (3.6x10^12 mile/yr)

= 1.114x10^-8 deaths/mi

Chi:Mil

(1.114x10^-8 deaths/mi) x 71 mi x 0.2

= 1.581 x 10^-7 deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

RISK OF DEATH, FLYING

General Risk

0.2 deaths / 1x10^10 mi

= 2x10^-11 deaths / mi

Chi:Mil

(2x10^-11 deaths / mi) x 71 mi =

Is flight risk linearly related to distance?

For this 1 trip, I expect 0.0000000014 deaths

p(driving death) = 113 * p(flying death)

ASSUMPTIONS THAT CONCERN ME

RISK OF DEATH, DRIVING

General Risk

÷ 4

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

= 3.6x10^12 total miles driven

(40,100 deaths/yr) / (3.6x10^12 mile/yr)

= 1.114x10^-8 deaths/mi

Chi:Mil

÷ 4

(1.114x10^-8 deaths/mi) x 71 mi x 0.2

= 1.581 x 10^-7 deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

RISK OF DEATH, FLYING

General Risk

0.2 deaths / 1x10^10 mi

= 2x10^-11 deaths / mi

Chi:Mil

(2x10^-11 deaths / mi) x 71 mi =

= 1.4x10^-9 deaths / trip

x 7

For this 1 trip, I expect 0.0000000014 deaths

~~**p(driving death) = 113 * p(flying death)**~~

p(driving death) =p(flying death)

ASSUMPTIONS THAT CONCERN ME

RISK OF DEATH, DRIVING

General Risk

(12,000 mi) x (300M people)
= 3.6x10^12 total miles driven

(40,100 deaths/yr) / (3.6x10^12 mile/yr)
= 1.114x10^-8 deaths/mi

Chi:Mil

x 10

(1.114x10^-8 deaths/mi) x 71 mi x 0.2
= 1.581 x 10^-7 deaths / trip

For this 1 trip, I expect 0.0000001581 deaths

RISK OF DEATH, FLYING

General Risk

0.2 deaths / 1x10^10 mi
= 2x10^-11 deaths / mi

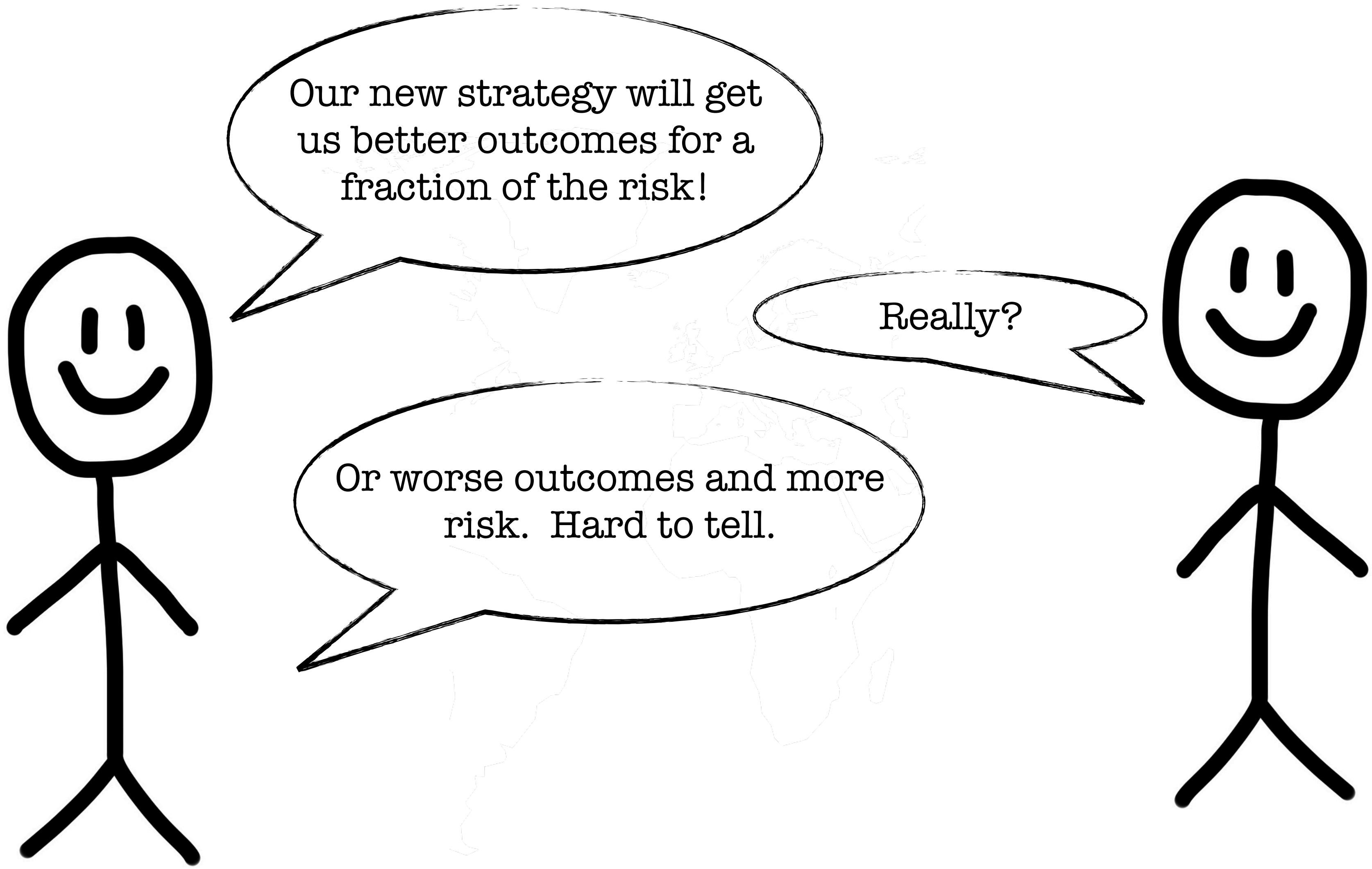
Chi:Mil

(2x10^-11 deaths / mi) x 71 mi =
= 1.4x10^-9 deaths / trip

For this 1 trip, I expect 0.0000000014 deaths

~~p(driving death) = 113 * p(flying death)~~

p(driving death) =1000 * p(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.