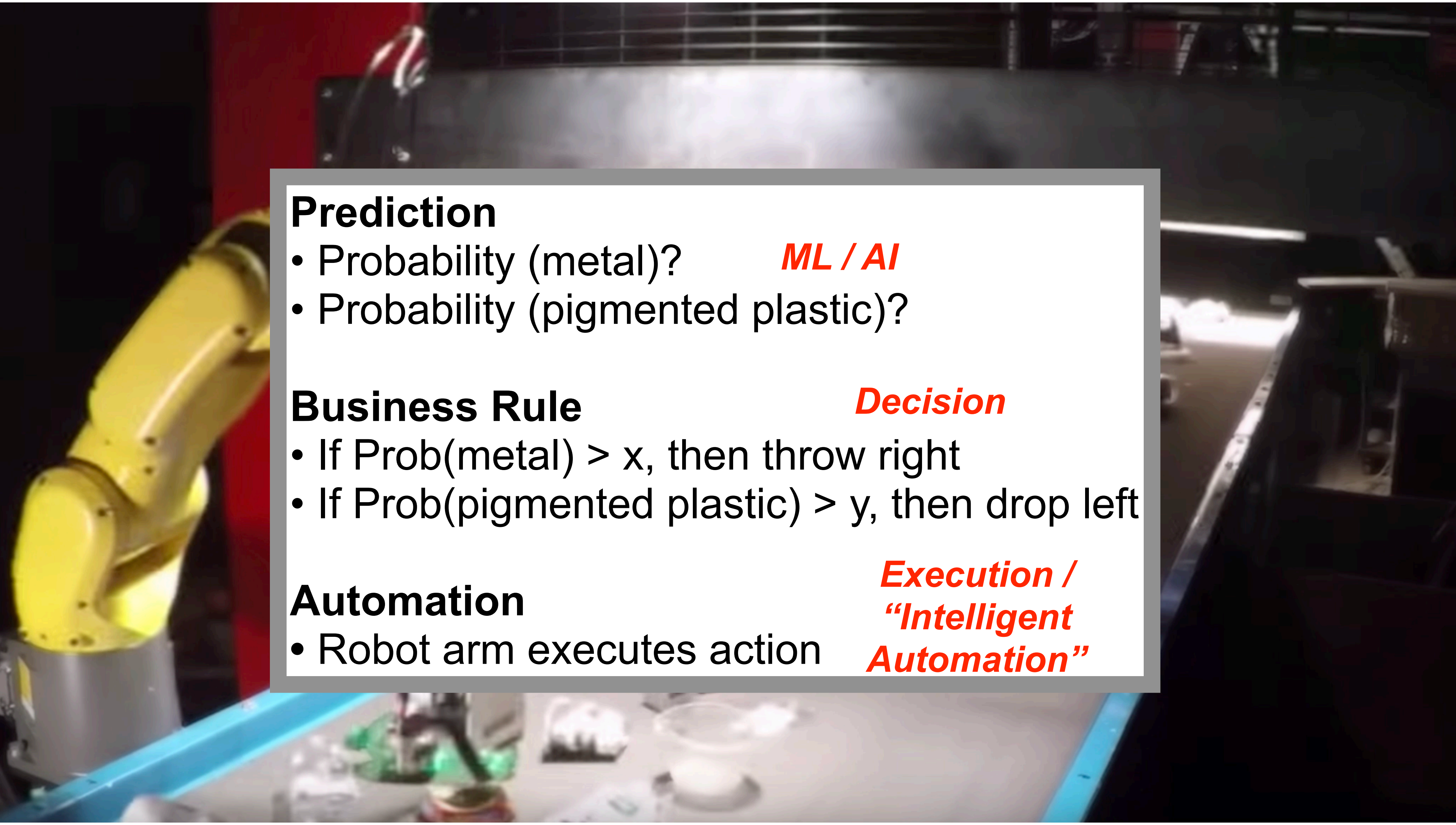
The Importance of Test and Learn

Professor Joel Shapiro
Managerial Economics and Decision Sciences
Lead Faculty, Kellogg Analytical Consulting Lab

Northwestern | Kellogg



Prediction

- Probability (metal)?
- Probability (pigmented plastic)?

ML / AI

Business Rule

- If Prob(metal) > x, then throw right
- If Prob(pigmented plastic) > y, then drop left

Decision

Automation

- Robot arm executes action

Execution / "Intelligent Automation"

PRESCRIPTIVE ANALYTICS: CASE STUDY

Most business decisions are harder than “throw right” and “drop left”

Kellogg School of Management
Northwestern University

Pentathlon:
Promotional E-mail Frequency

Anna Quintero walked out of the monthly Product Department Director meeting feeling pummeled. In her new role as the director of digital marketing for Pentathlon, a leading European sporting goods retailer, Quintero had argued that e-mail access to customers had to be carefully restricted to make sure that customers wanted to continue receiving promotional e-mails from the company. But the department directors just weren’t convinced.

Each department director at Pentathlon, oversaw one of the seven major product categories sold at Pentathlon: Endurance (e.g., running, cycling), Strength and Fitness (e.g., gymnastics, yoga), Water Sports (e.g., sailing, kayaking), Team Sports (e.g., soccer, basketball, rugby), Backcountry (e.g., hiking, climbing), Winter Sports (e.g., skiing, snowboarding), and Racquet (e.g., tennis, squash). Before Quintero joined Pentathlon, digital marketing had been under the

ANNA'S EVIDENCE FOR EMAIL LIMITS

All consumers	Result
"I receive too many promotional e-mails from Pentathlon"	72%
"I receive just the right amount of promotional e-mails from Pentathlon"	21%
"I receive too few promotional e-mails from Pentathlon"	7%

Consumers in 30-44 age group (30.33% of total consumers)	
"I receive too many promotional e-mails from Pentathlon"	74%
"I receive just the right amount of promotional e-mails from Pentathlon"	23%
"I receive too few promotional e-mails from Pentathlon"	3%

Consumers who are among top 25% total euro sales (last 12 months)	
"I receive too many promotional e-mails from Pentathlon"	87%
"I receive just the right amount of promotional e-mails from Pentathlon"	10%
"I receive too few promotional e-mails from Pentathlon"	3%

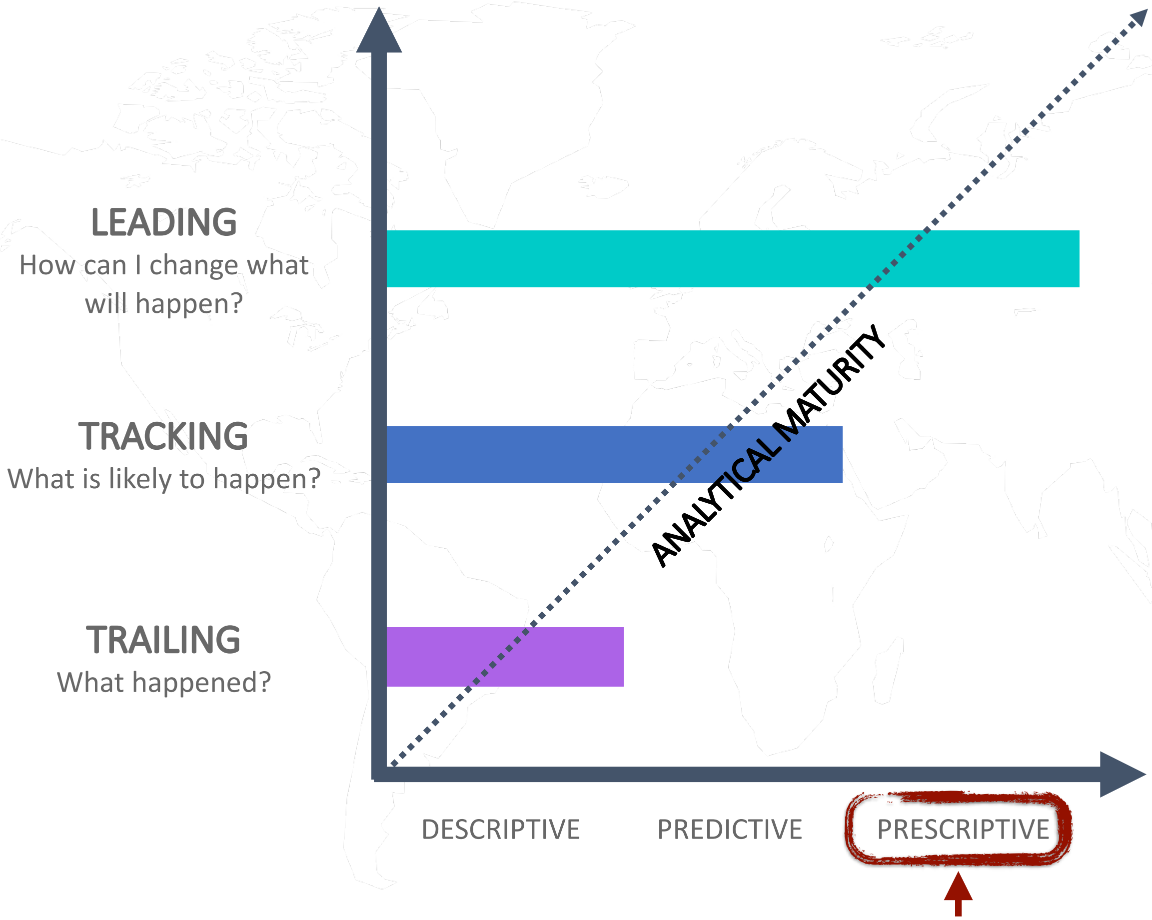
Consumers who are among bottom 25% total euro sales (last 12 months)	
"I receive too many promotional e-mails from Pentathlon"	52%
"I receive just the right amount of promotional e-mails from Pentathlon"	31%
"I receive too few promotional e-mails from Pentathlon"	17%

FRANCOIS' EVIDENCE AGAINST EMAIL LIMITS

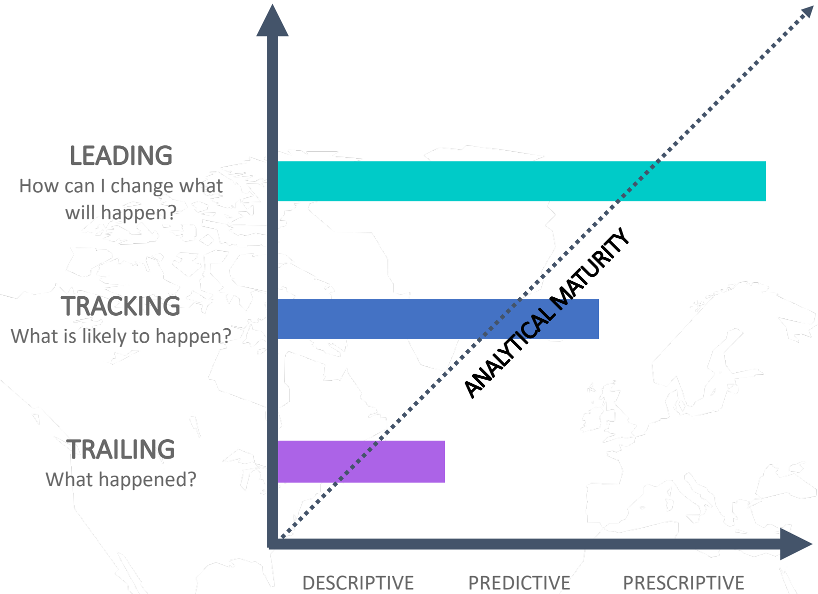
	Number of orders during last 12 months
Average weekly e-mail frequency: 1 - 2.9	3.5
Average weekly e-mail frequency: 3 - 4.9	9.9
Average weekly e-mail frequency: 5 or more	17.3

	Total euro sales during last 12 months
Average weekly e-mail frequency: 1 - 2.9	€ 41.6
Average weekly e-mail frequency: 3 - 4.9	€ 210.4
Average weekly e-mail frequency: 5 or more	€ 498.1

	Average order size during last 12 months
Average weekly e-mail frequency: 1 - 2.9	€ 10.92
Average weekly e-mail frequency: 3 - 4.9	€ 19.97
Average weekly e-mail frequency: 5 or more	€ 27.50



LEADING
How can I change what will happen?
TRACKING
What is likely to happen?
TRAILING
What happened?
ANALYTICAL MATURITY
DESCRIPTIVE
PREDICTIVE
PRESCRIPTIVE

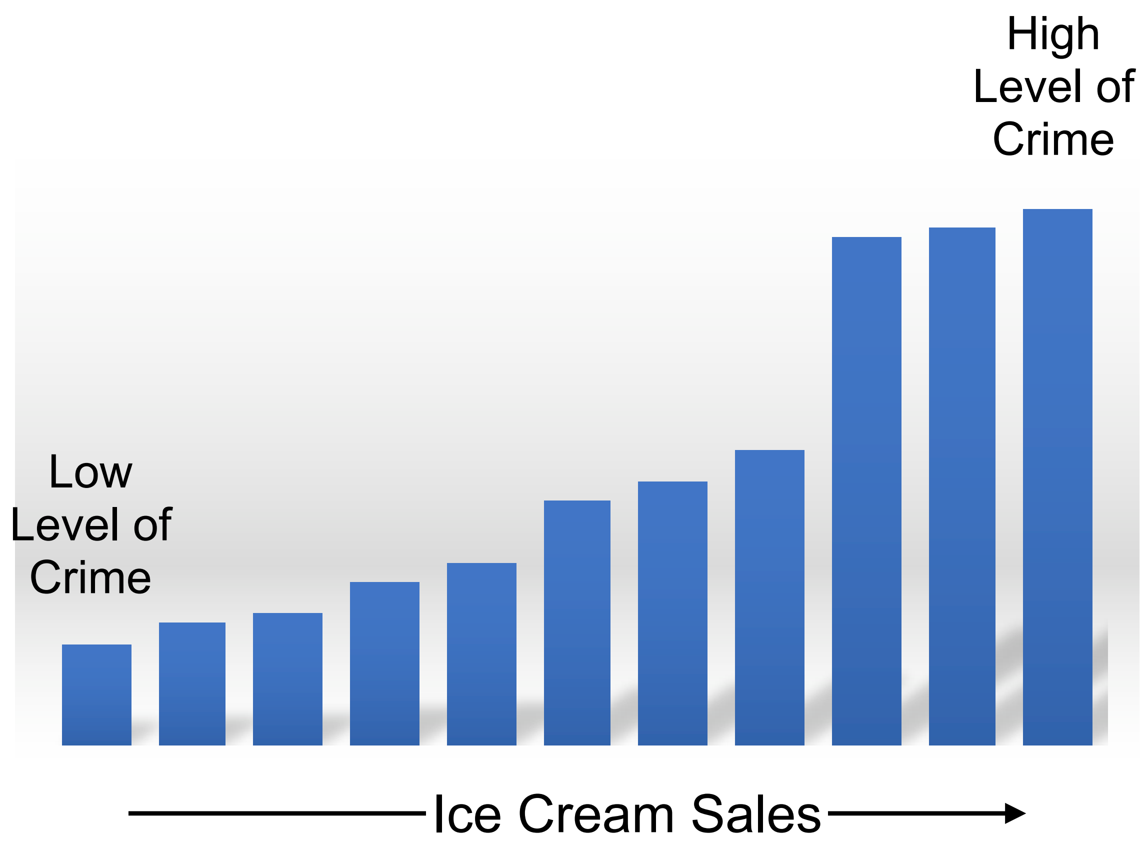


DESCRIPTIVE ANALYTICS
what happened?

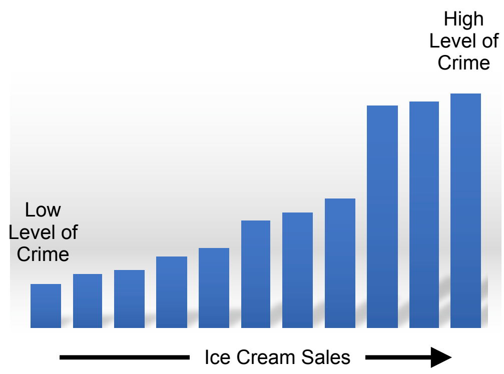
PREDICTIVE ANALYTICS
what will happen in the absence of doing anything differently?

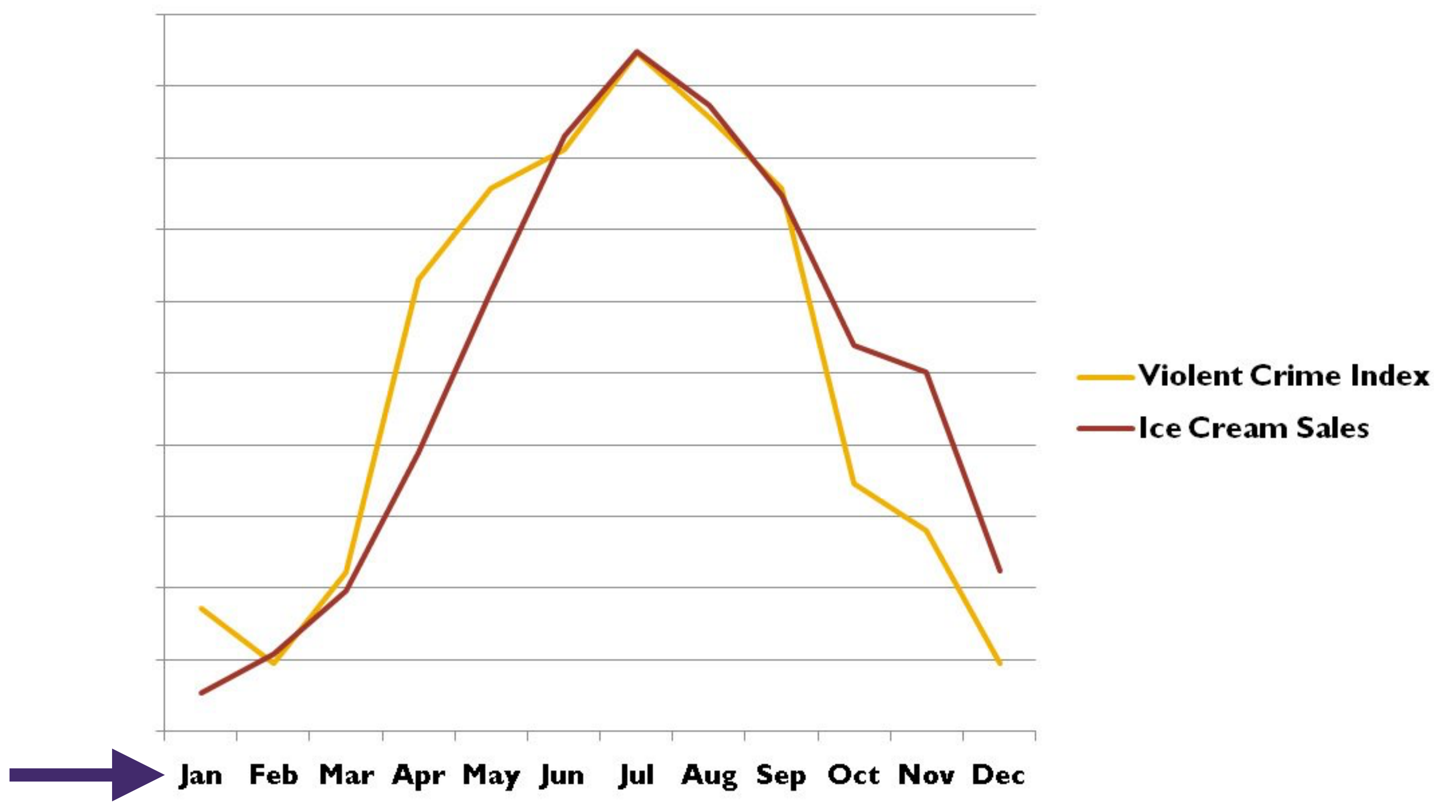
PRESCRIPTIVE ANALYTICS
what will happen if I take some business action?

PRESCRIPTION: FIGHT CRIME WITH ICE CREAM?

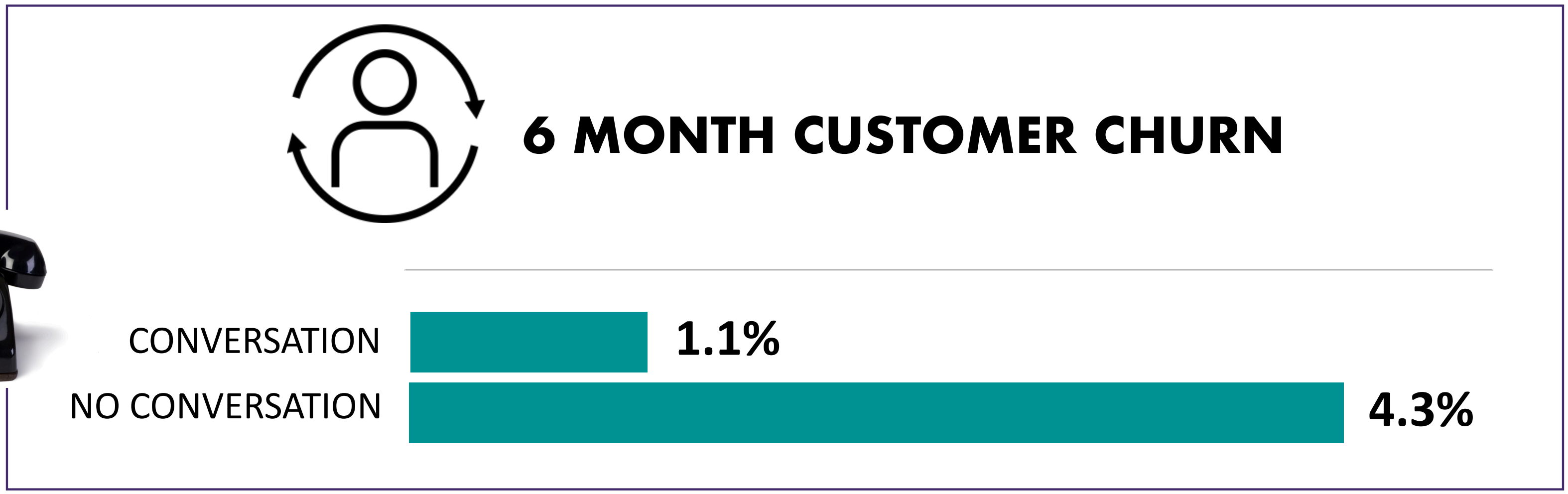


PRESCRIPTION: FIGHT CRIME WITH ICE CREAM?

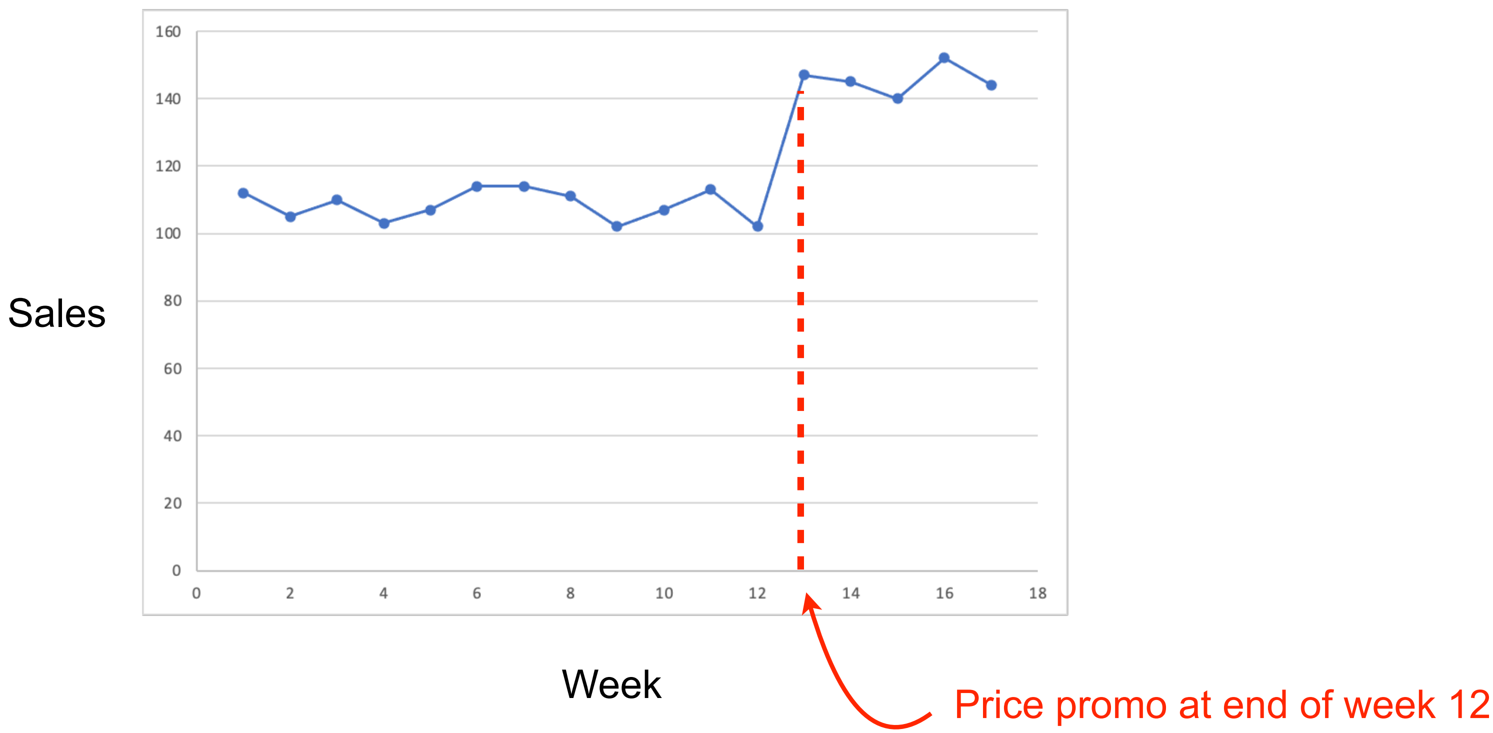




PRESCRIPTION: CALL PEOPLE TO KEEP THEM LOYAL?

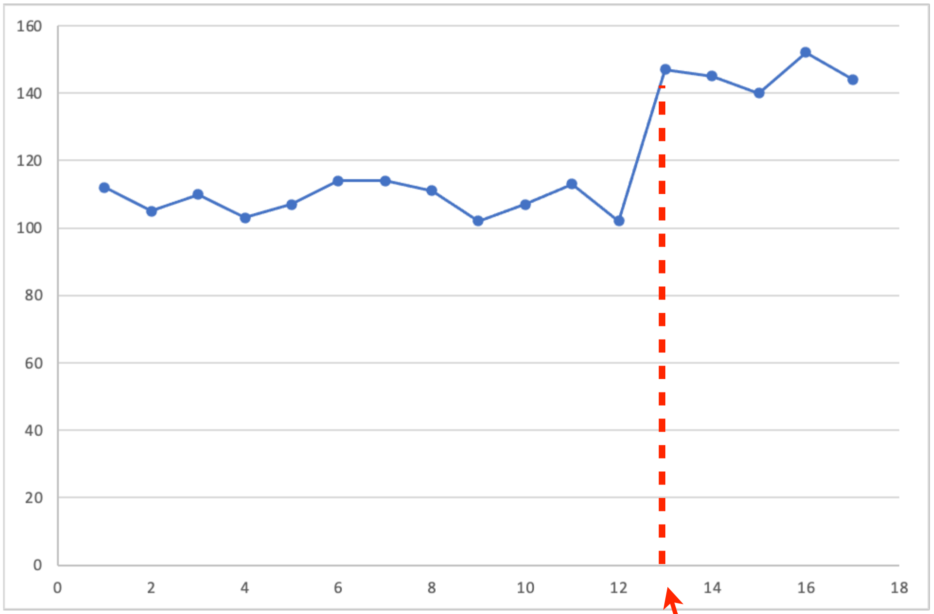


"HEY, OUR PROMO WORKED!"



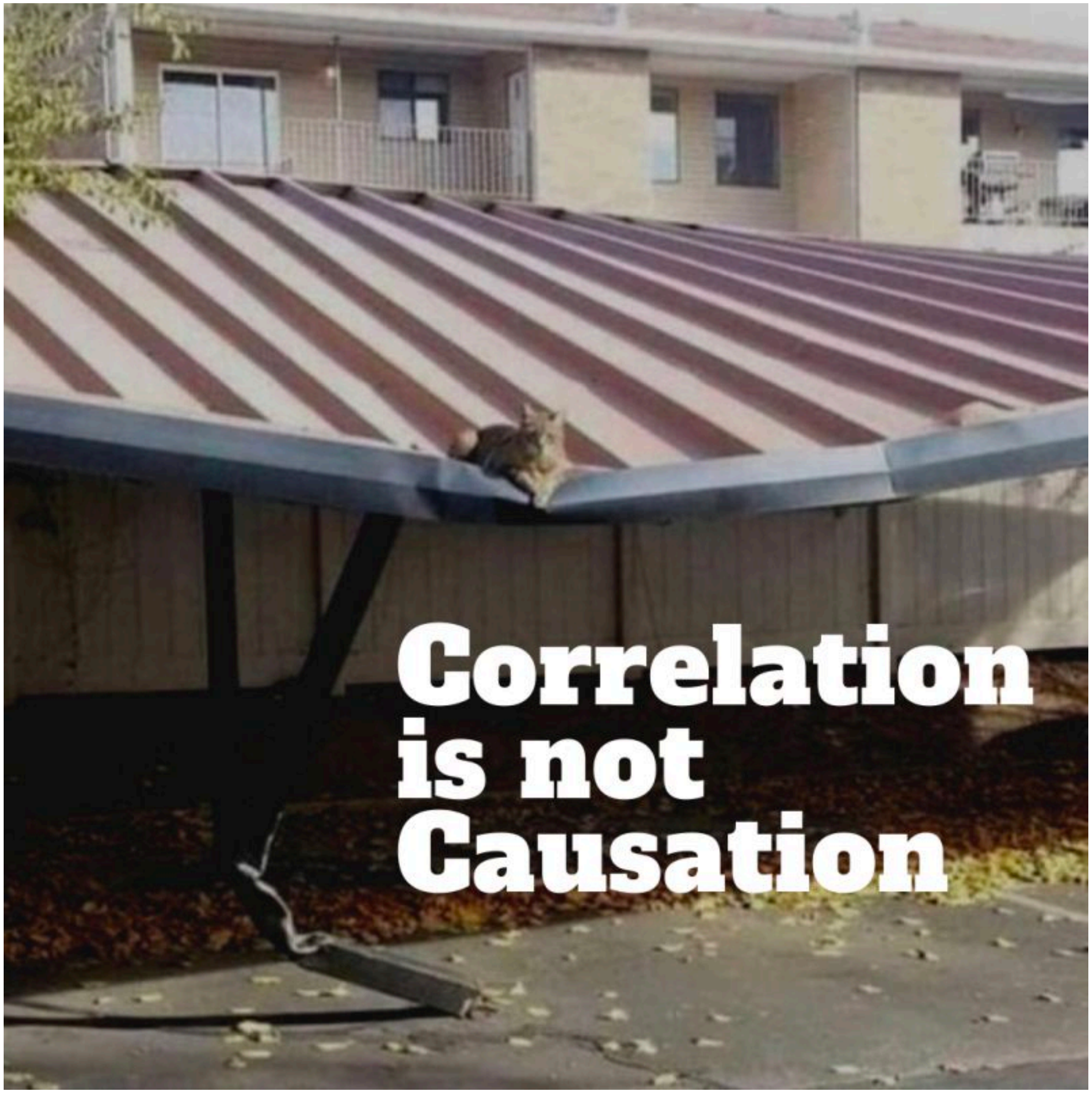
"HEY, OUR WORK FROM HOME POLICY WORKED!"

Productivity



Week

New work from home policy at end of week 12


Correlation
is not
Causation

RECALL THIS EXAMPLE OF A HIGH-END HOTEL'S PREDICTIVE CHURN MODEL

- 32 hotels worldwide
- Average of 90 rooms per hotel
- 36% customer churn, defined as: *no stay within 6 months of last stay*

PREDICTIVE MODEL BEGINS WITH DATA

HotelChurn Joel

Table | Stats

17 / 38 rows < 1 of 3 >

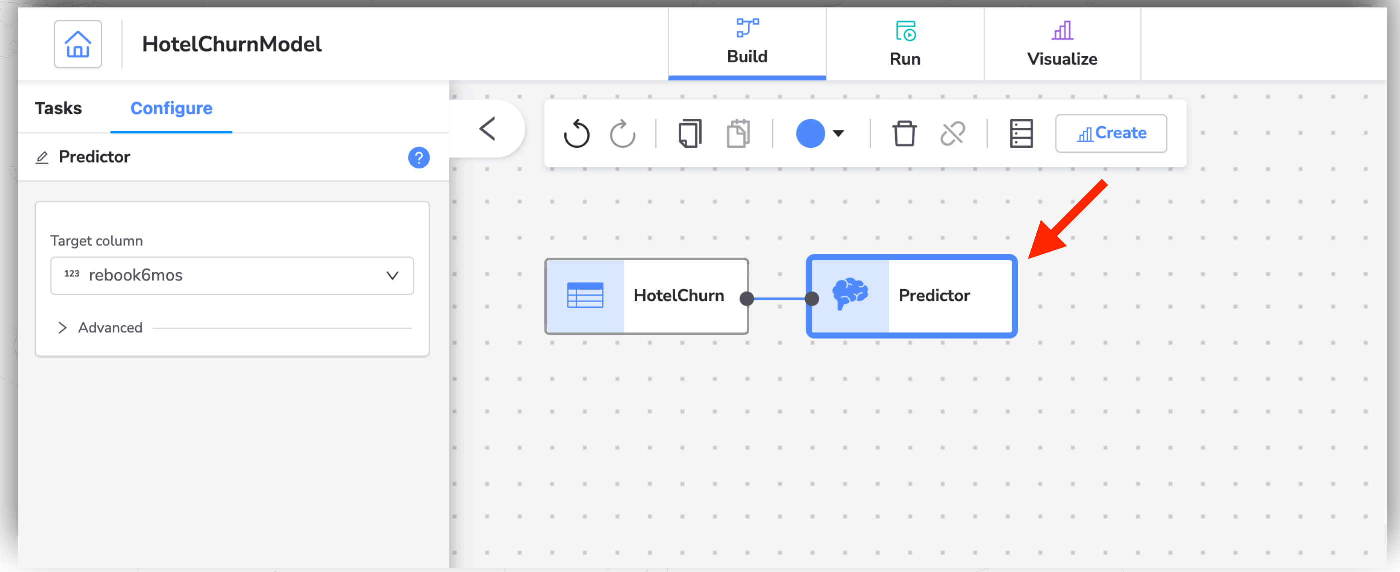
custid	rebook6mos	roomserv	shuttle	fitness	restaurant	check-in	specreq	age	gender	purpose	homecountry
556,287	1	0	1	0	1	in-person	0	31.3	male	leisure	us
222,794	1	0	0	0	4	kiosk	0	50.1	female	leisure	brazil
832,099	0	0	0	0	3	app	0	76.4	male	leisure	spain
616,626	1	1	1	0	7	in-person	0	50.7	male	work	us
219,131	1	0	1	0	4	kiosk	1	39.1	female	work	australia
416,877	0	0	0	0	4	app	1	63.4	male	leisure	china
802,106	1	0	1	0	1	in-person	0	58	female	leisure	us
887,655	0	0	1	0	4	kiosk	0	28.1	male	work	us
529,202	1	0	1	0	0	in-person	0	79.9	male	work	china
537,305	1	1	1	1	3	in-person	0	21	female	work	spain
983,077	0	1	1	1	5	app	0	45.3	female	work	brazil
829,753	1	0	1	1	0	in-person	0	44.7	male	leisure	china
170,204	1	1	1	1	0	in-person	0	34.7	male	leisure	us
561,390	0	0	0	1	6	app	1	71.2	female	work	japan
632,289	1	0	0	1	3	kiosk	1	71.5	male	leisure	brazil
561,612	1	1	0	0	0	in-person	0	43.9	male	leisure	spain
348,779	0	1	0	1	0	in-person	0	70.2	female	work	us

Outcome **Inputs / Predictors**

THERE ARE MANY KINDS OF PREDICTIVE TOOLS

To find the most accurate model, you could just keep trying different algorithms…

But today's consumer-facing tech tools can automatically find the best model for you!



100%

10 / 38 rows < 1 of 4 >

sed on up to 10,000 sampled rows

req	age	gender	purpose	homecountry	Probability	
0	31.3	male	leisure	us	0.927 314 937 ...	Explain
0	50.1	female	leisure	brazil	0.813 864 648 ...	Explain
0	76.4	male	leisure	spain	0.746 664 524 ...	Explain
0	50.7	male	work	us	0.661 574 542 ...	Explain
1	39.1	female	work	australia	0.537 610 173 ...	Explain
1	63.4	male	leisure	china	0.888 415 396 ...	Explain
0	58	female	leisure	us	0.920 725 822 ...	Explain
0	28.1	male	work	us	0.745 765 864 ...	Explain
0	79.9	male	work	china	0.591 027 259 ...	Explain
0	21	female	work	spain	0.978 200 495 ...	Explain

— 100% +

10 / 38 rows < 1 of 4 >

sed on up to 10,000 sampled rows

req	age	gender	purpose	homecountry		Probability	
0	31.3	male	leisure	us	1	0.927 314 937 ...	Explain
0	50.1	female	leisure	brazil	1	0.813 864 648 ...	Explain
0	76.4	male	leisure	spain	0	0.746 664 524 ...	Explain
0	50.7	male	work	us	1	0.661 574 542 ...	Explain
1	39.1	female	work	australia	1	0.537 610 173 ...	Explain
1	63.4	male	leisure	china	0	0.888 415 396 ...	Explain
0	58	female	leisure	us	1	0.920 725 822 ...	Explain
0	28.1	male	work	us	0	0.745 765 864 ...	Explain
0	79.9	male	work	china	1	0.591 027 259 ...	Explain
0	21	female	work	spain	1	0.978 200 495 ...	Explain

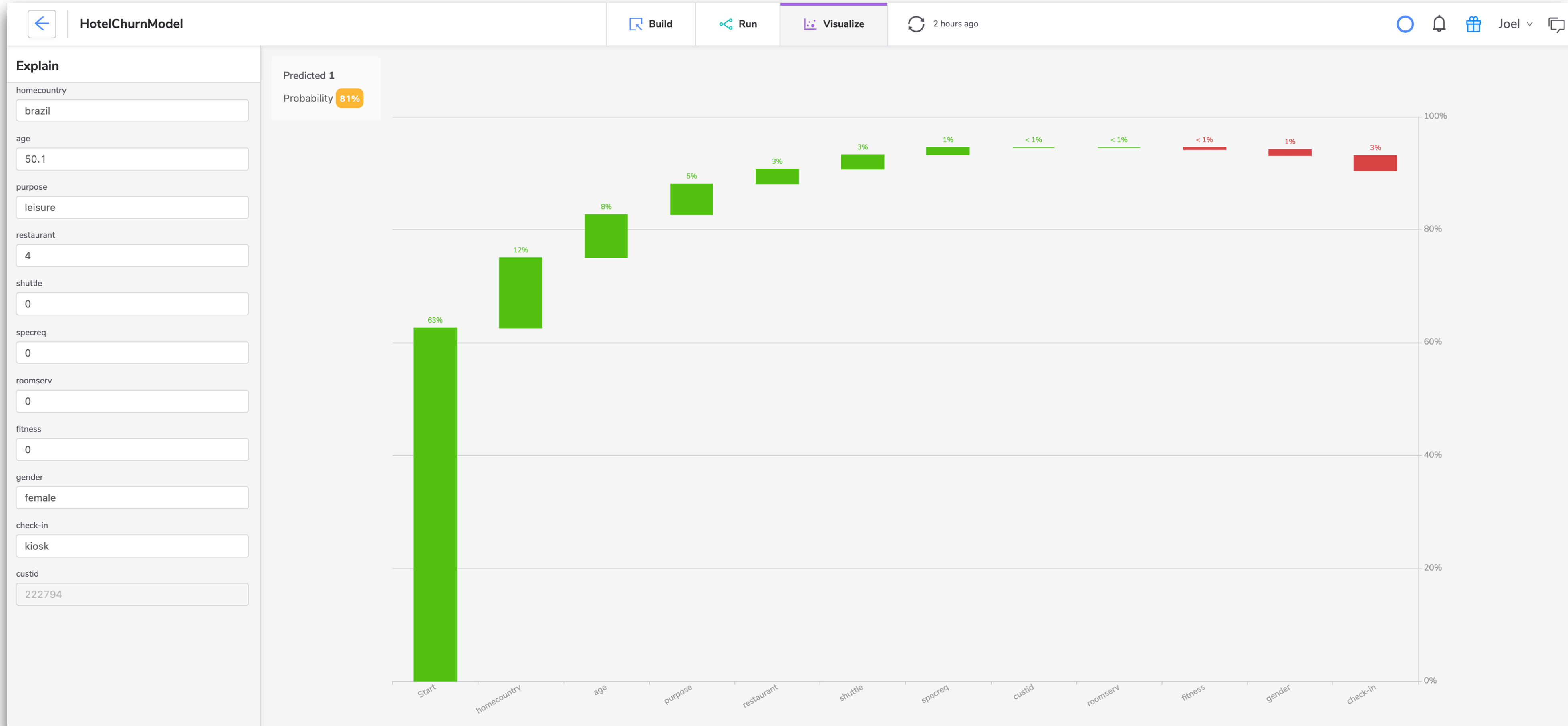
— 100% +

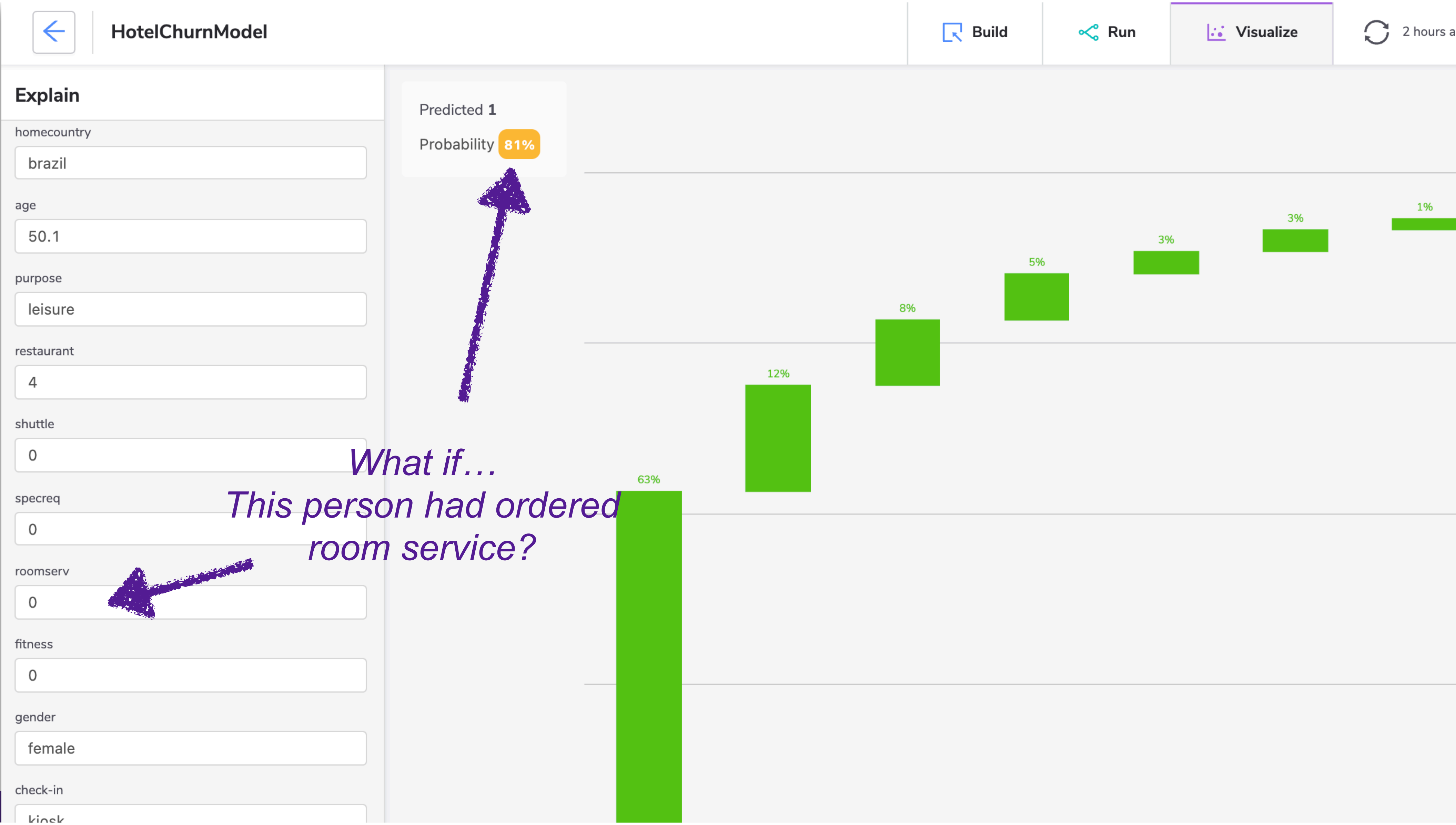
10 / 38 rows < 1 of 4 >

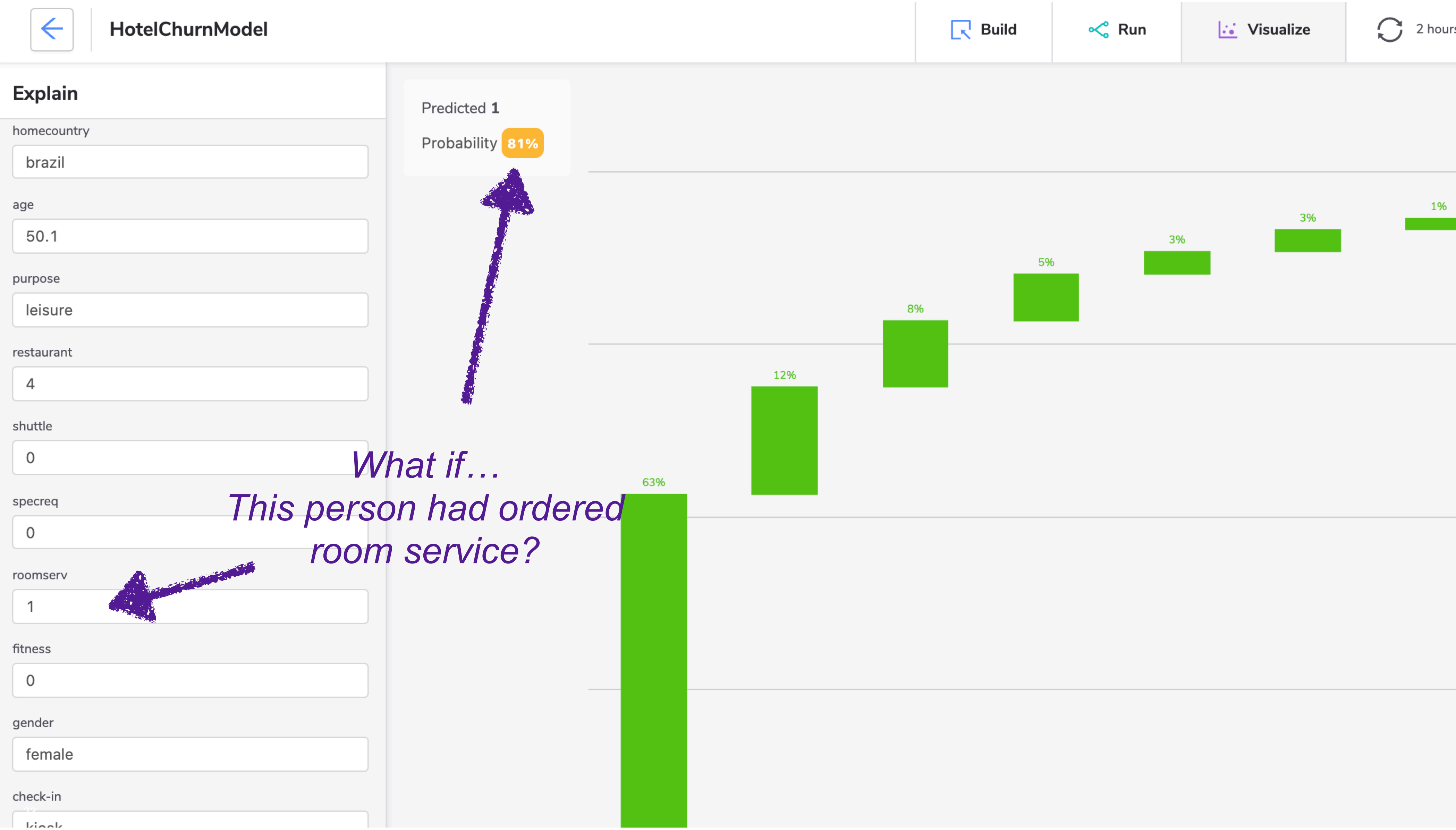
sed on up to 10,000 sampled rows

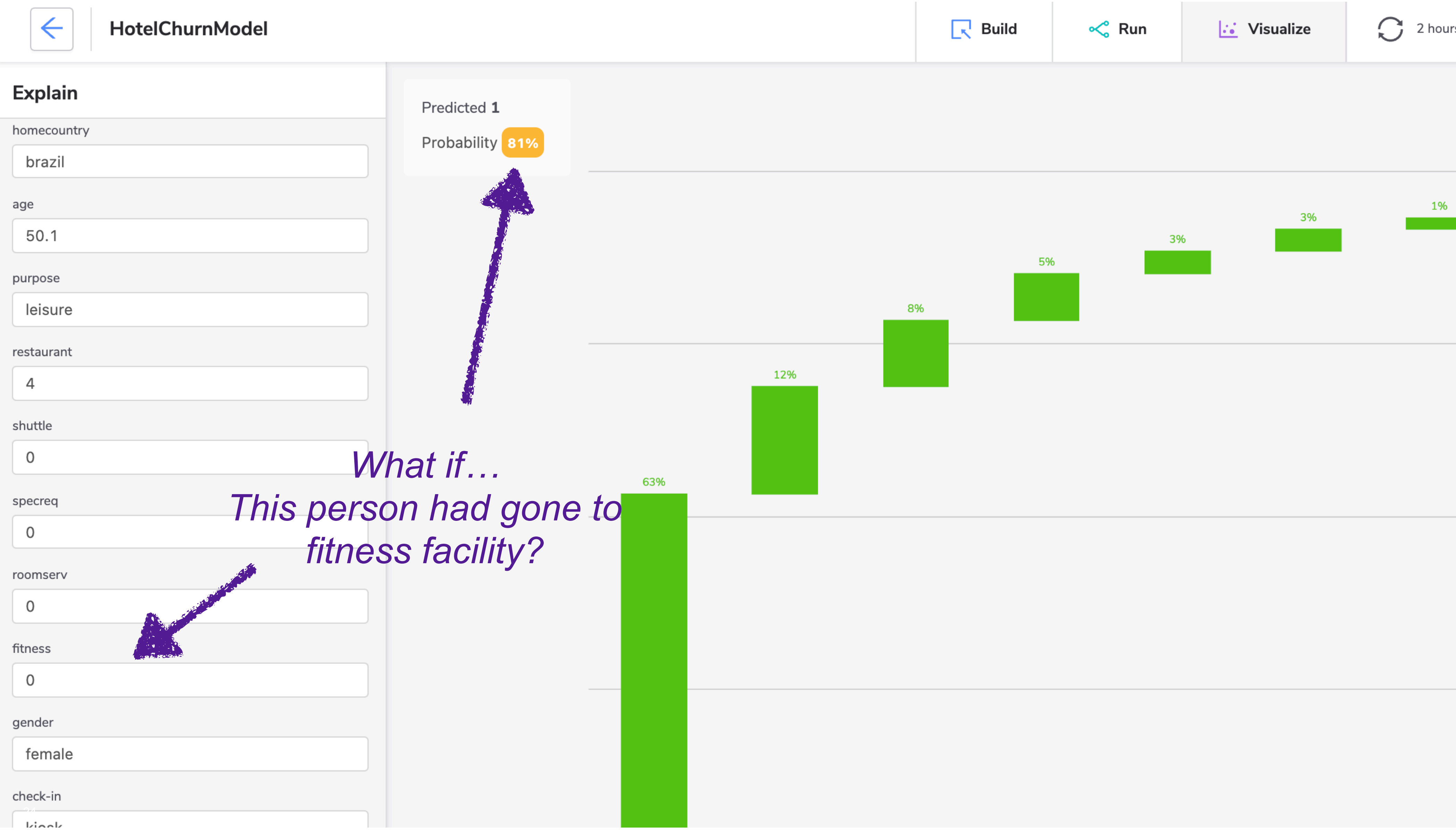
req	age	gender	purpose	homecountry		Probability	
0	31.3	male	leisure	us	1	0.927 314 937 ...	Explain
0	50.1	female	leisure	brazil	1	0.813 864 648 ...	Explain
0	76.4	male	leisure	spain	0	0.746 664 524 ...	Explain
0	50.7	male	work	us	1	0.661 574 542 ...	Explain
1	39.1	female	work	australia	1	0.537 610 173 ...	Explain
1	63.4	male	leisure	china	0	0.888 415 396 ...	Explain
0	58	female	leisure	us	1	0.920 725 822 ...	Explain
0	28.1	male	work	us	0	0.745 765 864 ...	Explain
0	79.9	male	work	china	1	0.591 027 259 ...	Explain
0	21	female	work	spain	1	0.978 200 495 ...	Explain

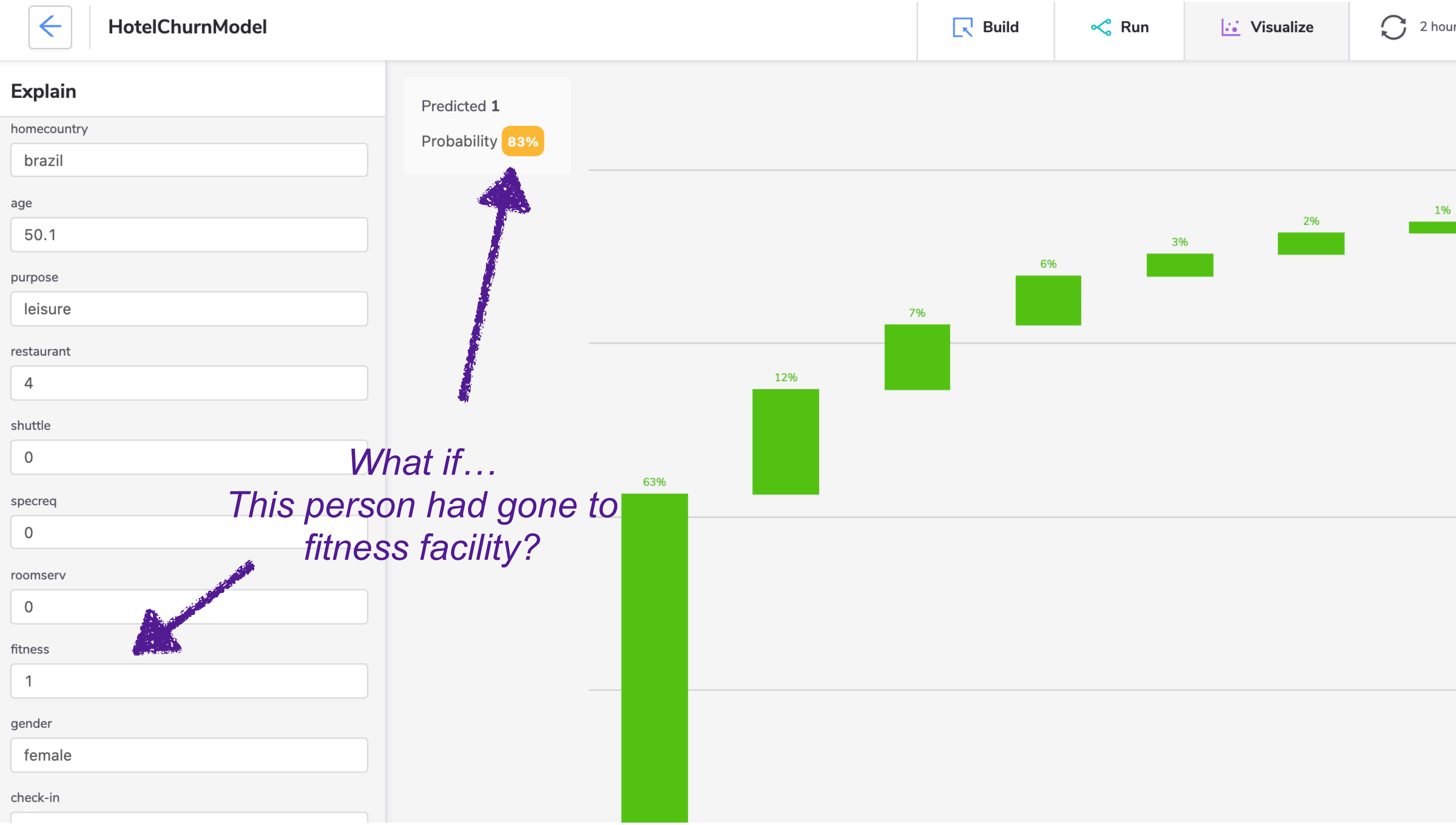
“GLASS BOX” AI!!


HotelChurnModel
Build
Run
Visualize
2 hours ago
Joel
Explain
homecountry
brazil
age
50.1
purpose
leisure
restaurant
4
shuttle
0
specreq
0
roomserv
0
fitness
0
gender
female
check-in
kiosk
custid
222794
Predicted 1
Probability 81%
63%
12%
8%
5%
3%
3%
1%
< 1%
< 1%
< 1%
1%
3%
100%
80%
60%
40%
20%
0%
Start
homecountry
age
purpose
restaurant
shuttle
specreq
custid
roomserv
fitness
gender
check-in


HotelChurnModel
Build
Run
Visualize
2 hours a
Explain
homecountry
brazil
age
50.1
purpose
leisure
restaurant
4
shuttle
0
specreq
0
roomserv
0
fitness
0
gender
female
check-in
kiosk
Predicted 1
Probability 81%
63%
12%
8%
5%
3%
3%
1%
What if…
This person had ordered room service?


HotelChurnModel
Build
Run
Visualize
2 hours a
Explain
homecountry
brazil
age
50.1
purpose
leisure
restaurant
4
shuttle
0
specreq
0
roomserv
1
fitness
0
gender
female
check-in
kiosk
Predicted 1
Probability 81%
63%
12%
8%
5%
3%
3%
1%
What if…
This person had ordered
room service?


HotelChurnModel
Build
Run
Visualize
2 hours
Explain
homecountry
brazil
age
50.1
purpose
leisure
restaurant
4
shuttle
0
specreq
0
roomserv
0
fitness
0
gender
female
check-in
Predicted 1
Probability 81%
63%
12%
8%
5%
3%
3%
1%
What if…
This person had gone to fitness facility?



What if…
This person had gone to fitness facility?

Trying to assess “what works” by looking **backwards** is one of the most difficult tasks you can ask of ANY data professional!

It is much easier and effective to create **forward-looking** measurement strategies.

YOUR DATA WORK WILL, IN ALL LIKELIHOOD, HELP LEADERS KNOW **WHERE TO INVEST**

- Most decisions are harder than “throw right” and "drop left.”
- Being a great leader necessarily means knowing where to invest, and you want to be able to help leaders do more of what works, and less of what doesn’t.
- Knowing “what works” is a foundational management and leadership skill
- Looking backwards at data to know “what works” is one of the hardest tasks for any data professional.

YOUR DATA WORK WILL, IN ALL LIKELIHOOD, HELP LEADERS KNOW **WHERE TO INVEST**

- Most decisions are harder than “throw right” and "drop left.”
- Being a great leader necessarily means knowing where to invest, and you want to be able to help leaders do more of what works, and less of what doesn’t.
- Knowing “what works” is a foundational management and leadership skill
- Looking backwards at data to know “what works” is one of the hardest tasks for any data professional.

TEST AND LEARN

TEST AND LEARN

TEST AND LEARN

The process of trying an initiative for the purpose of rigorous measurement. Informs whether initiative was successful, whether it was worthwhile, how to improve it, how to move forward.

TEST AND LEARN

TEST AND LEARN

The process of trying an initiative for the purpose of rigorous measurement. Informs whether initiative was successful, whether it was worthwhile, how to improve it, how to move forward.

Evaluating "what works" = causal inference

Testing to build predictive models (e.g., for targeting / IPFA)

TEST AND LEARN

TEST AND LEARN

The process of trying an initiative for the purpose of rigorous measurement. Informs whether initiative was successful, whether it was worthwhile, how to improve it, how to move forward.

Evaluating "what works"
= causal inference

Testing to build predictive models
(e.g., for targeting / IPFA)

EXPERIMENTS

QUASI-
EXPERIMENTS

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
3. How will you establish **internal validity (causation)?**
4. To whom / what will your results **generalize**?
5. What will be your **criterion** for decision-making?

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**

Purposefully create groups that **look similar** so that the initiative is the only difference

If we can randomize, great = **experiment**

If we can NOT randomize, might need “second best” = **quasi-experiment**

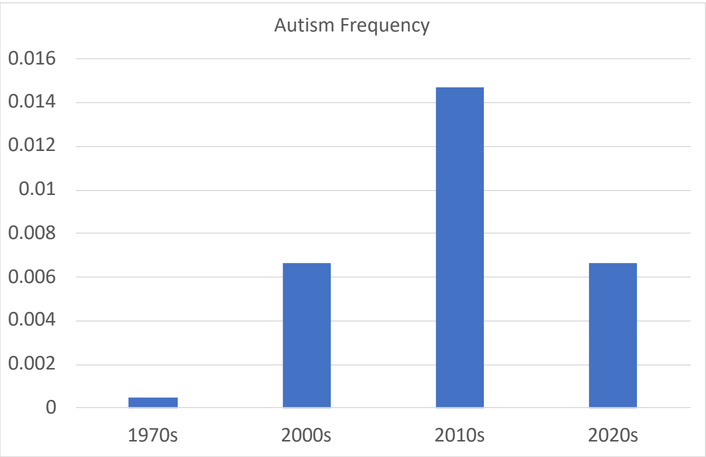
Search for groups / circumstances in the past that looked similar prior to being exposed to the initiative, then infer the effect of the initiative.

= quasi-experiment

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
    a. Precisely defined

INCIDENCE OF AUTISM IN RECENT YEARS



1 in 2,000 *1 in 150* *1 in 68* *1 in 150*

What caused the increase?

CDC has posited some risk factors, including:

- Genetics
- Age of parents
- Ingestion of chemicals, *in utero*

KEYS TO SUCCESSFUL EVALUATION

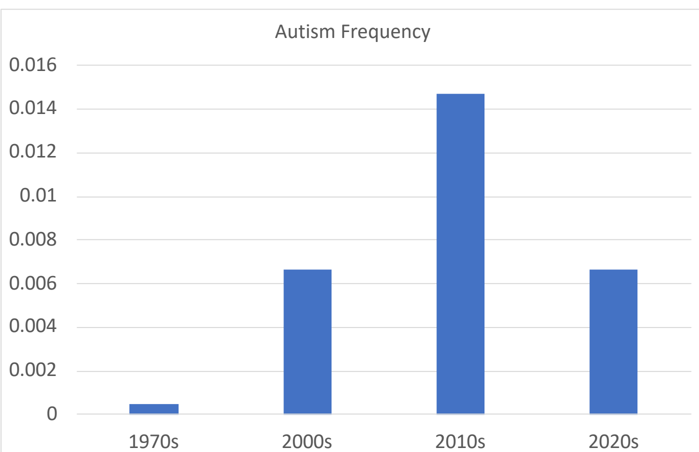
1. Wi

2. Wh

But a big part of this "increase" is that the definition of autism has changed, not the human condition.

1940s: Originated as a version of schizophrenia
1980s: definition refined, symptoms added, but limited to kids under 3
2010s: broadened to include kids of any age
2020s: all kids, but symptoms refined

INCIDENCE OF AUTISM IN RECENT YEARS



1 in 2,000 *1 in 150* *1 in 68* *1 in 150*

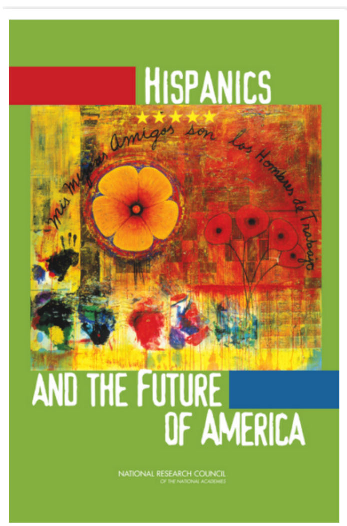
What caused the increase?

CDC has posited some risk factors, including:

- Genetics
- Age of parents
- Ingestion of chemicals, *in utero*

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
   a. Precisely defined
   b. Actually represent the thing we care about (construct validity)



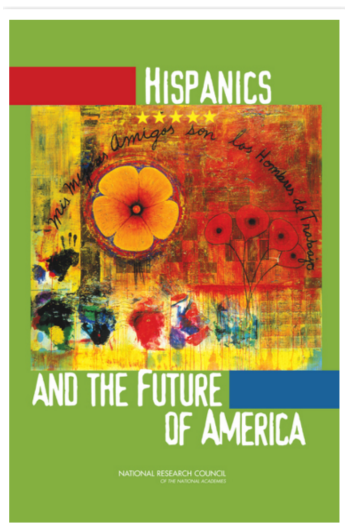
National Research Council, 2006

Identified key challenges, including Hispanics' "lack of access to health care."

1. How would you measure "access to health care?"

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
    a. Precisely defined
    b. Actually represent the thing we care about (construct validity)



National Research Council, 2006

Identified key challenges, including Hispanics' "lack of access to health care."

1. How would you measure "access to health care?"
2. Does an increase in your measure mean that things have improved?

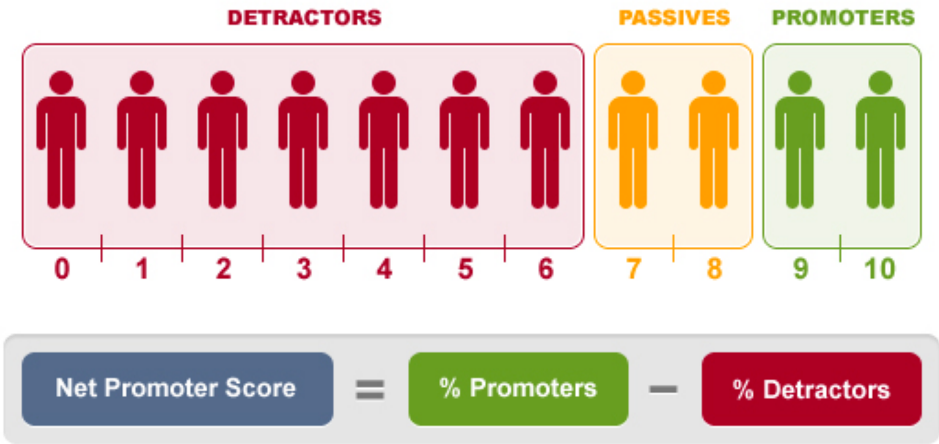
KEYS TO SUCCESSFUL EVALUATION

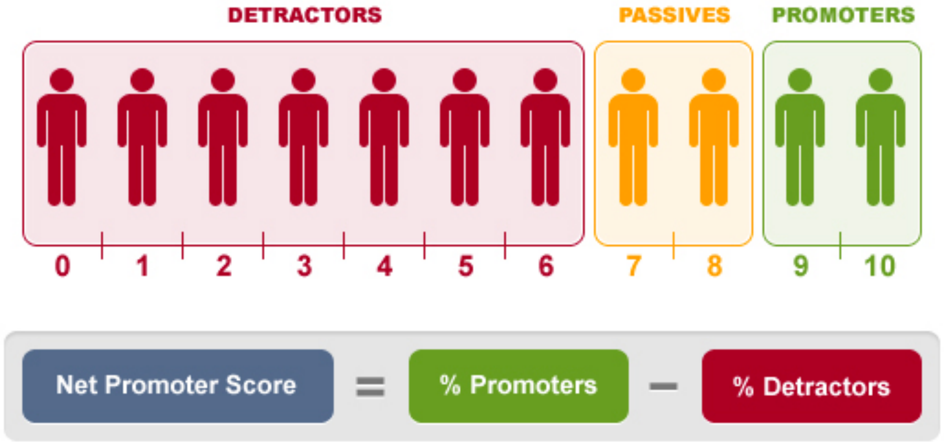
1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
   a. Precisely defined
   b. Actually represent the thing we care about (construct validity)

This is one of the least appreciated barriers to good analytics:

Are we measuring the right things?

“On a scale of 0 to 10, how likely are you to recommend us to a friend or colleague?”





0	1	2	3	4	5	6	7	8	9	10	TOTAL
100	100	100	100	100	100	100	100	100	100	100	**-45**
1100	0	0	0	0	0	0	0	0	0	0	**-100**
0	0	0	0	0	0	1100	0	0	0	0	**-100**
0	0	0	0	0	0	0	1100	0	0	0	**0**
0	0	0	0	0	0	0	0	0	1100	0	**100**

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
    a. Precisely defined
    b. Actually represent the thing we care about (construct validity)
    c. Be aware of unintended consequences

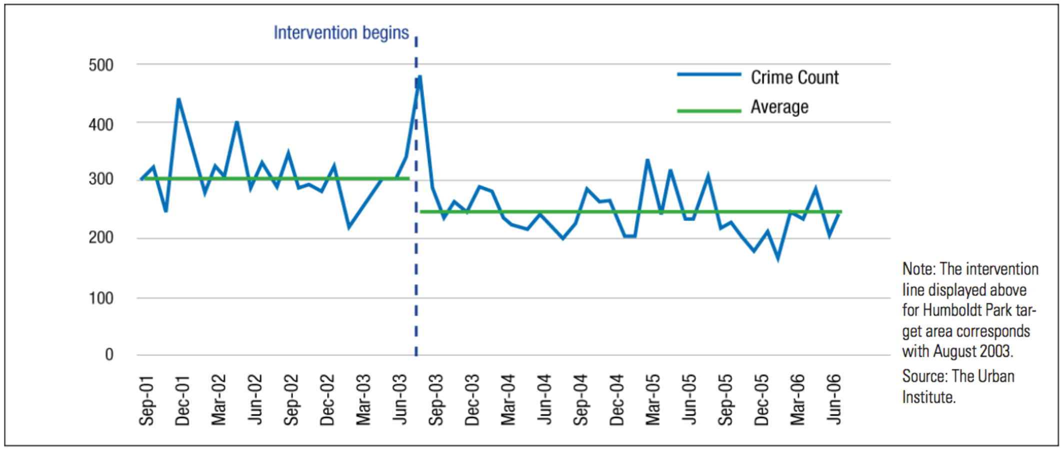


Figure 5.2: Crime Trend in Humboldt Park Area, Chicago, 2001–2006

Do surveillance cameras reduce crime?

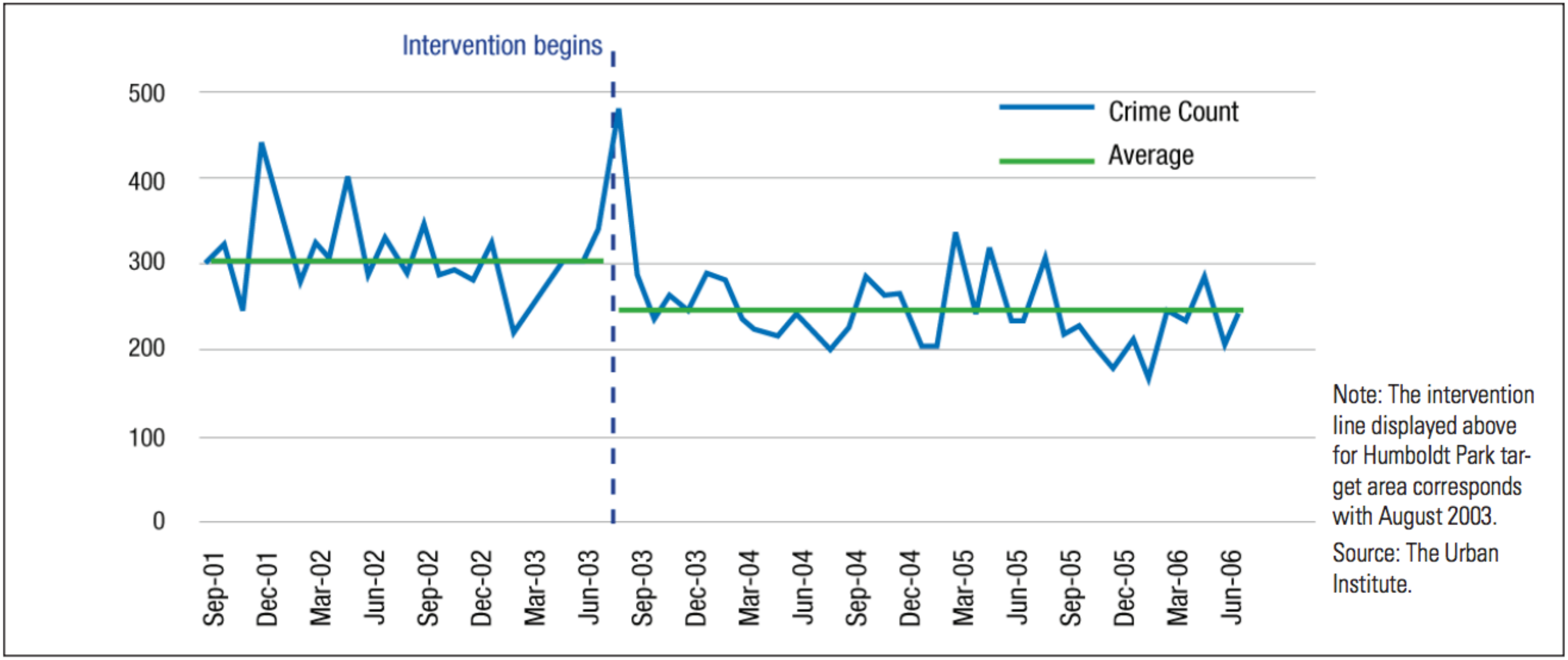


Figure 5.2: Crime Trend in Humboldt Park Area, Chicago, 2001–2006

A Nasty, Nafta-Related Surprise: Mexico's Soaring Obesity

Few predicted when Mexico joined the free-trade deal that it would transform the country in a way that would saddle millions with diet-related illnesses.

The Nasty Surprise

Obesity Rate

1980	2016
7%	20.3%

Diabetes now Mexico's "top killer"

The Nasty Surprise

Obesity Rate

1980	2016
7%	20.3%

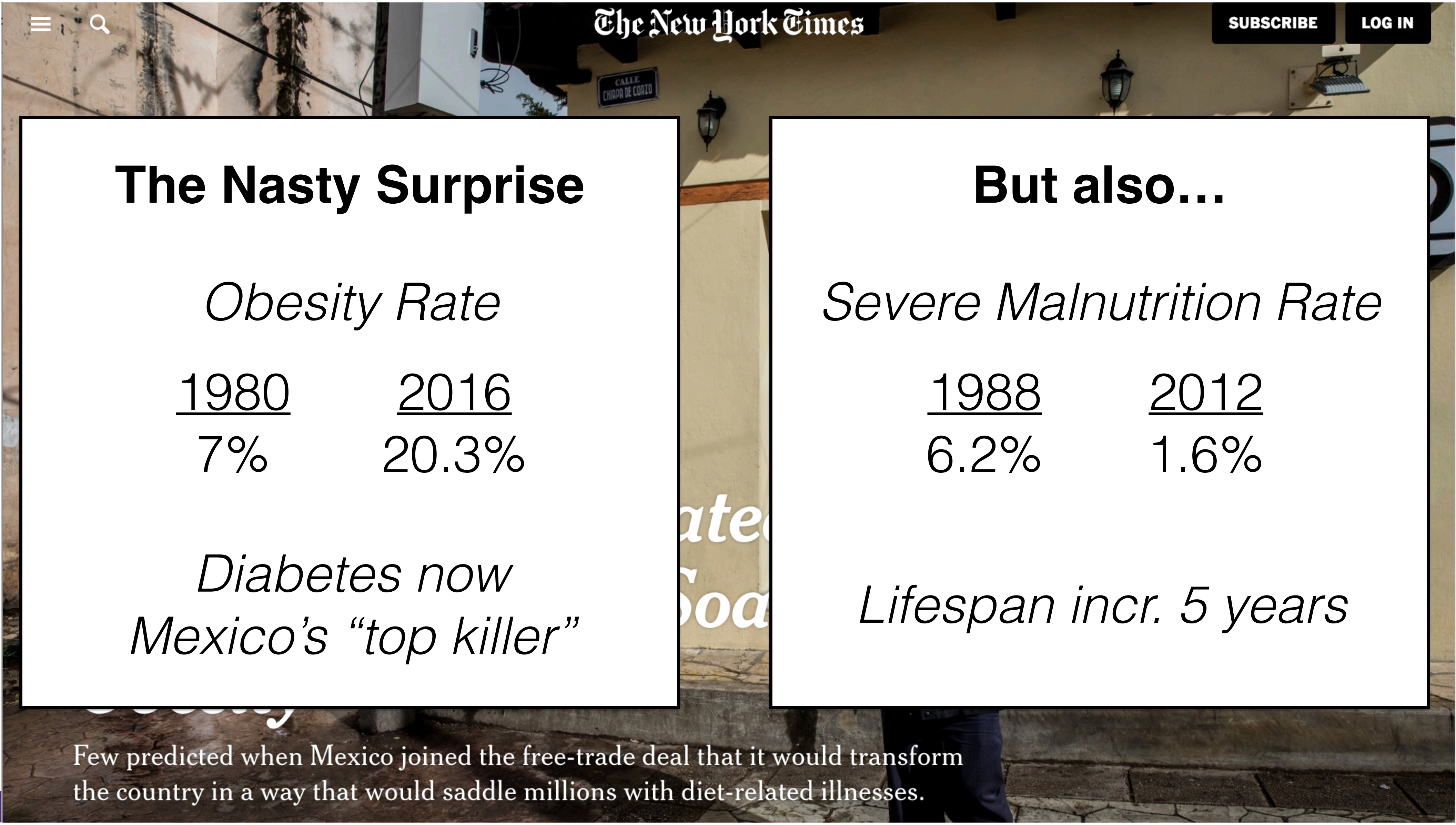
Diabetes now Mexico's "top killer"

But also…

Severe Malnutrition Rate

1988	2012
6.2%	1.6%

Lifespan incr. 5 years



Innovation

The Discipline of Business Experimentation

Increase your chances of success with innovation test-drives. by Stefan Thomke and Jim Manzi

From the Magazine (December 2014)

**A new flatbread at Wawa did really well in various tests, but the chain decided not to move forward with it.
WHY???**

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
3. How will you establish **internal validity (causation)? (MORE SHORTLY)**

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
3. How will you establish **internal validity (causation)? (MORE SHORTLY)**
4. To whom / what will your results **generalize**?

The directionality of generalizability is key

Backwards:

Group was subject to the treatment → Application of results constrained by characteristics of the group

E.g., we ran a promotion for beer at the Chicago Bears' game…
Results only generalize to Bears' fans?

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
3. How will you establish **internal validity (causation)? (MORE SHORTLY)**
4. To whom / what will your results **generalize**?

The directionality of generalizability is key

Backwards:

You know the population to which you want the results to generalize → Strategically choose who get your initiative / are you in your study

KEYS TO SUCCESSFUL EVALUATION

1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
3. How will you establish **internal validity (causation)? (MORE SHORTLY)**
4. To whom / what will your results **generalize**?
5. What will be your **criterion** for decision-making?

   Is your initiative **effective**?

   And is it **effective enough** to be worth the cost?

KEYS TO SUCCESSFUL EVALUATION

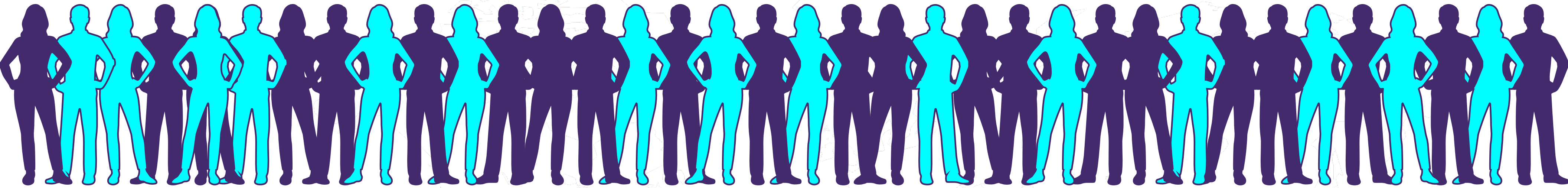
1. Will your evaluation be **forward-** or **backward-looking?**
2. What **outcome** will you use, and is it a good proxy for what you care about?
3. How will you establish **internal validity (causation)? (MORE SHORTLY)**
4. To whom / what will your results **generalize**?
5. What will be your **criterion** for decision-making?

CAUSAL INFERENCE

For evaluation, the gold standard = **business experimentation**

CAUSAL INFERENCE

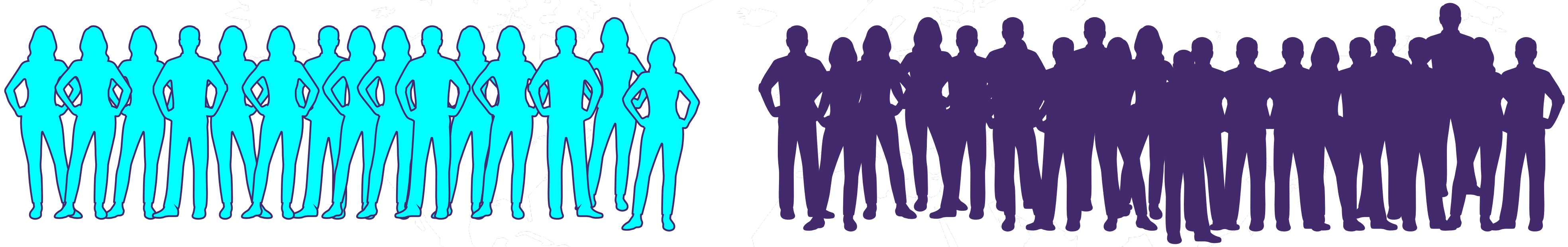
CAUSAL INFERENCE



Initiative

No Initiative

CAUSAL INFERENCE



Initiative

No Initiative

Outcome 1 ← ? → Outcome 2

If Outcome 1 > Outcome 2 by 10%,
= Initiative is responsible for a **10% lift in outcome!**

CAUSAL INFERENCE

An "initiative" can be anything you want to assess:

- **Price promo**
- **Sales training program**
- **Vaccine**
- **Harrah's free steak dinner**

No Initiative

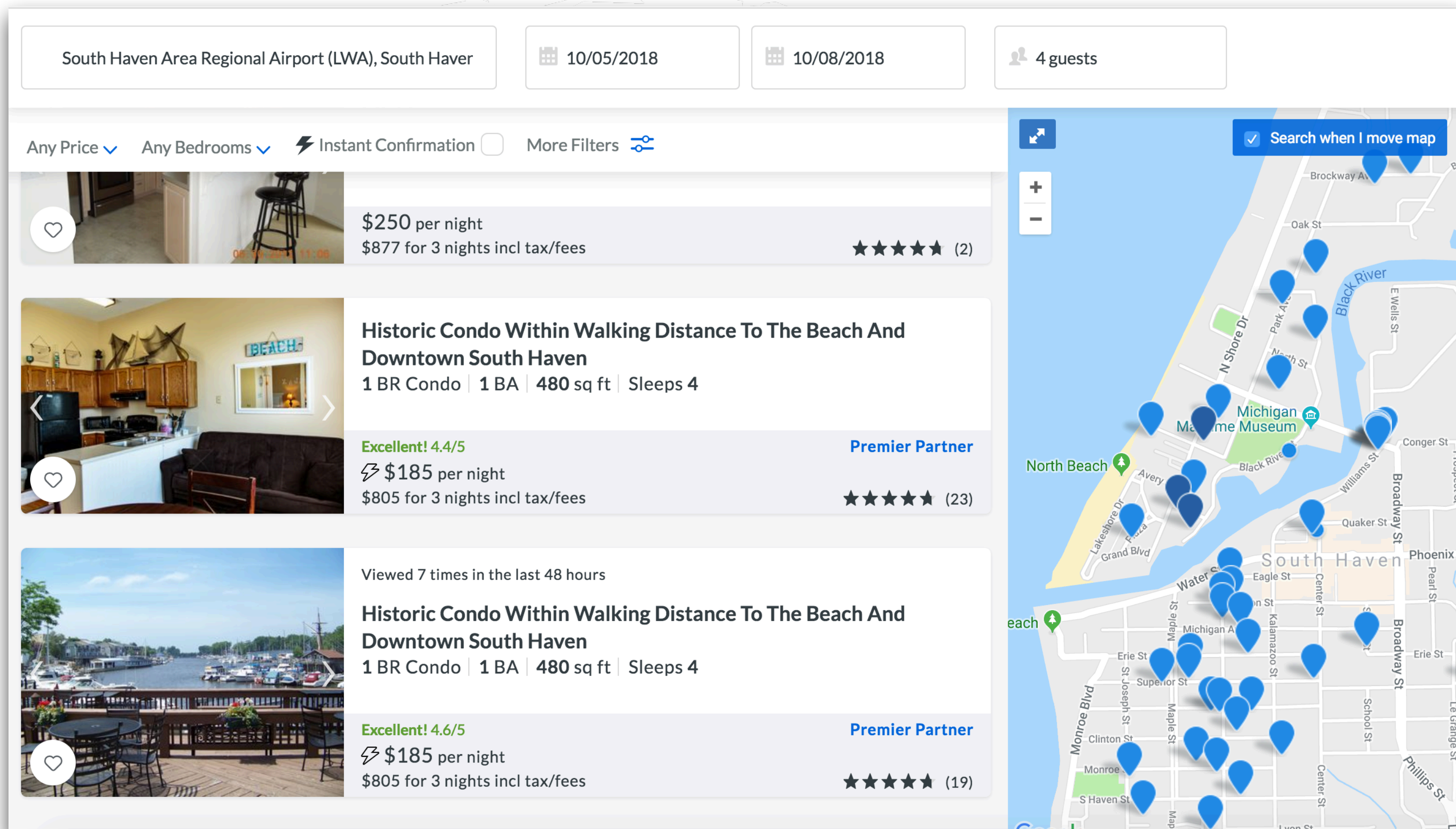
Outcome 1 ← ? → Outcome 2

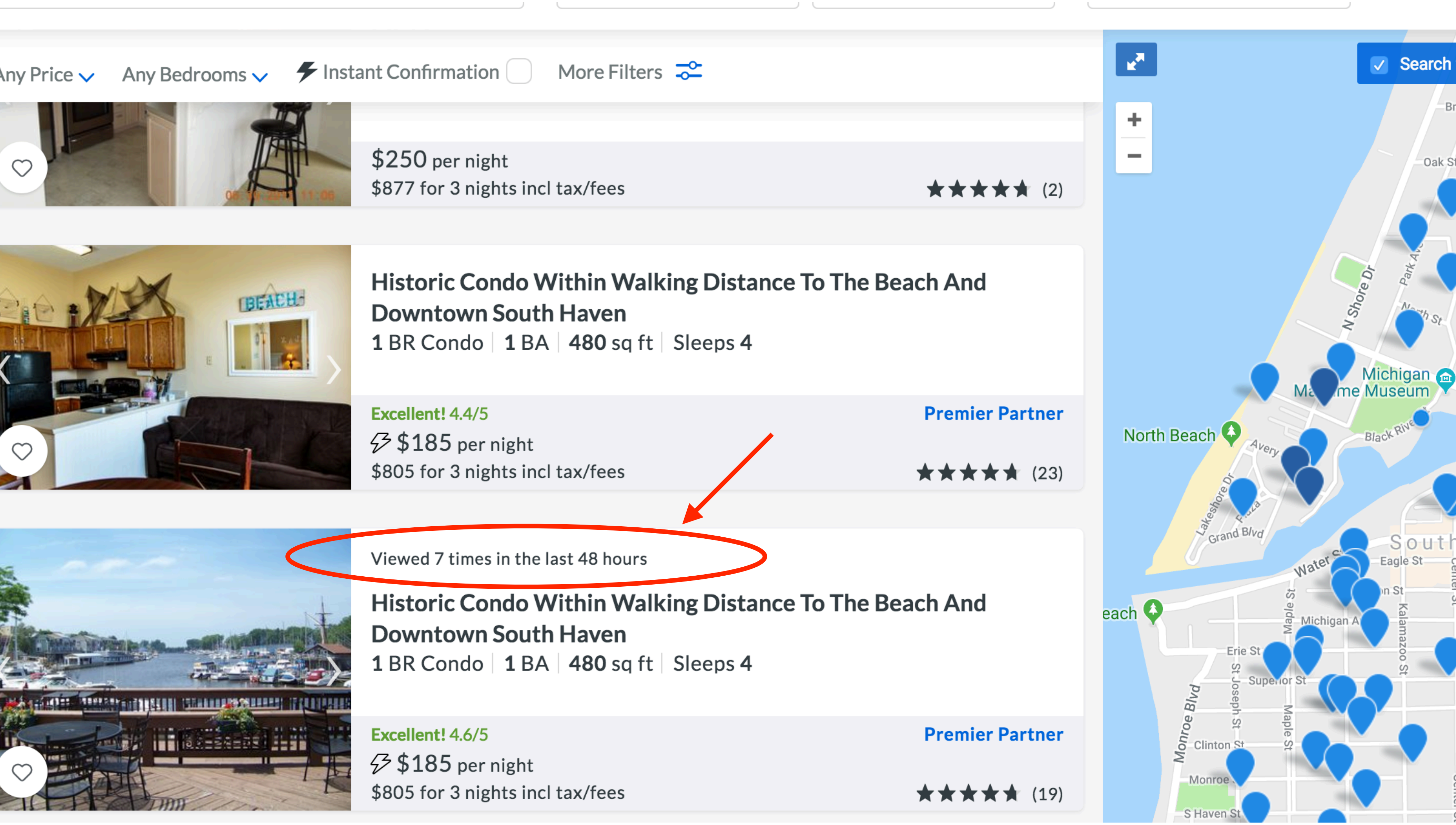
If Outcome 1 > Outcome 2 by 10%,
= Initiative is responsible for a **10% lift in outcome!**

HOW TO RUN EXPERIMENTS

1. Define your business **initiative**
2. Define the **outcome** that will be measured
3. Choose your **group** to be analyzed, split randomly into a **treatment** and **control** sub-group
4. Implement initiative with treatment group
5. Compare outcomes

A CLEAN AND SIMPLE EXPERIMENT




Any Price
Any Bedrooms
Instant Confirmation
More Filters
$250 per night
$877 for 3 nights incl tax/fees
(2)
Historic Condo Within Walking Distance To The Beach And Downtown South Haven
1 BR Condo | 1 BA | 480 sq ft | Sleeps 4
Excellent! 4.4/5
Premier Partner
$185 per night
$805 for 3 nights incl tax/fees
(23)
Viewed 7 times in the last 48 hours
Historic Condo Within Walking Distance To The Beach And Downtown South Haven
1 BR Condo | 1 BA | 480 sq ft | Sleeps 4
Excellent! 4.6/5
Premier Partner
$185 per night
$805 for 3 nights incl tax/fees
(19)
Search
North Beach
N Shore Dr
Michigan Maritime Museum
Black River
Grand Blvd
Eagle St
Michigan A
Erie St
Superior St
Clinton St
Monroe
S Haven St
Monroe Blvd
Maple St
Kalamazoo St
BEACH

A CLEAN AND SIMPLE EXPERIMENT

1,000 users

randomize

500
No message

500
“Viewed 7 times in the last 48 hours"

22% more clicks
17% more reservations

EXPERIMENTS ARE THE FOUNDATION OF **KNOWING WHAT WORKS**

They are how analytically mature teams do **everyday business**

- Which version of website causes most engagement / purchases?
- Which customer experience causes highest retention / NPS?
- Which product features cause most customer engagement?
- What retail store layout causes the most purchases / shortest check-out time?
- What is the right cadence for customer outreach?

EXPERIMENTS ARE THE FOUNDATION OF **KNOWING WHAT WORKS**

They are how analytically mature teams do **everyday business**

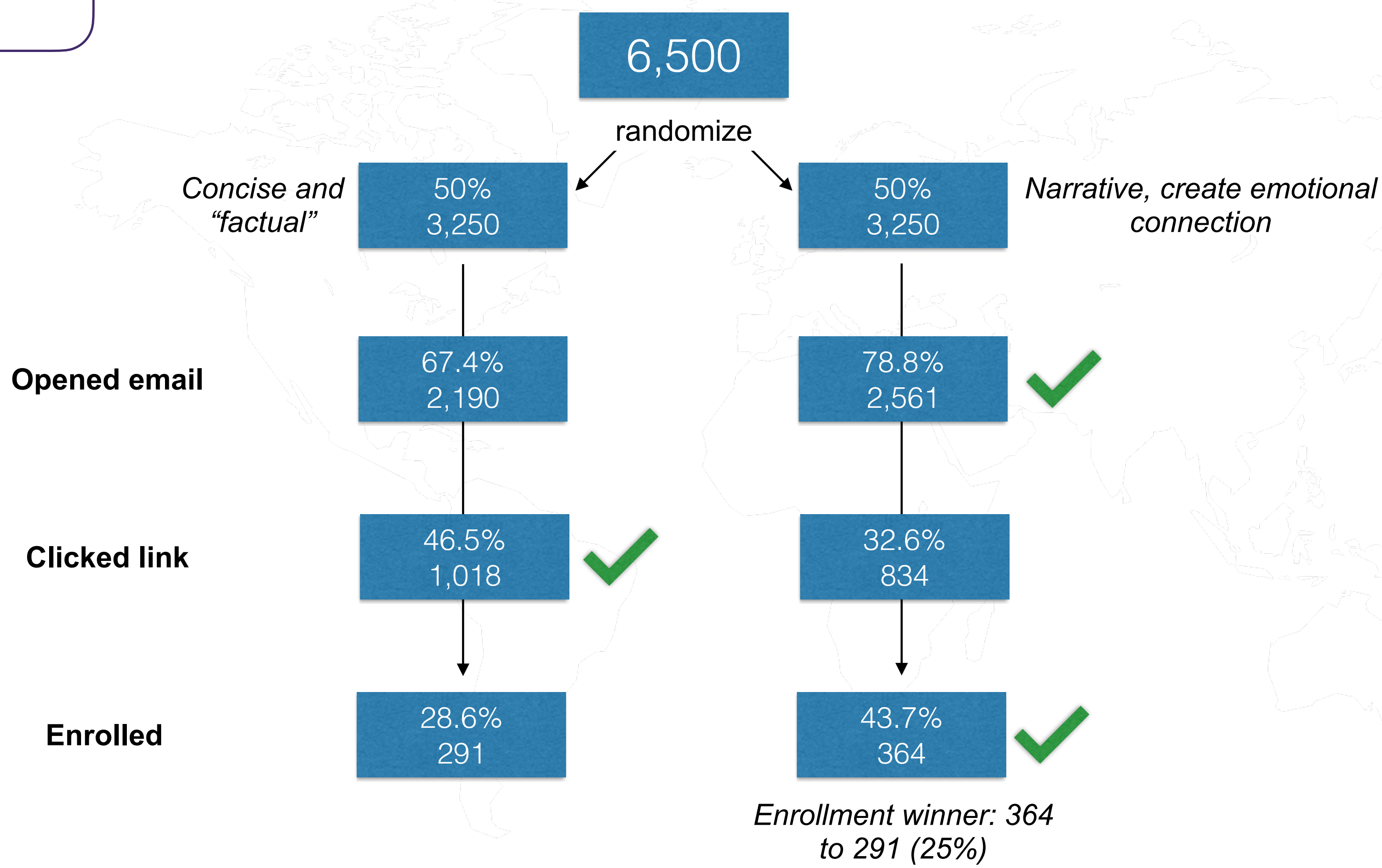
They can be used to measure **more than one outcome**

- A company wanted to encourage employees to enroll in a charitable "give-back" program. There was disagreement about how to frame the email solicitation.

Concise and "factual" or *Narrative, create emotional connection*

- 65,000 employees - how would you execute this experiment?

Pilot first, then choose winner for roll-out



EXPERIMENT CHECKLIST

Innovation

The Discipline of Business Experimentation

Increase your chances of success with innovation test-drives. by Stefan Thomke and Jim Manzi

From the Magazine (December 2014)

Purpose

- Does the experiment focus on a specific action?
- What will people learn?

Buy-In

- What changes will be made?
- How will the org ensure the results aren't ignored?

Feasibility

- What is the required sample size?
- Can the experiment be run at the test locations for the required duration?

Reliability

- Does the randomization work?
- Would others conducting same test get similar results?

Value

- Has the org considered a targeted rollout (customers, markets, segments) to concentrate investments for max payback?
- Does the org know what variables are causing what effects?

EXPERIMENTS ARE **NOT** FOOL-PROOF

SCIENCEINSIDER | FUNDING

NIH pulls the plug on controversial alcohol trial

Director Francis Collins acting on report, ends $100 million study of moderate drinking

15 JUN 2018 • BY MEREDITH WADMAN

Study found that moderate drinking caused:

- Statistically significant benefits, such as prevention on cardiovascular disease and diabetes
- No statistically significant increases in cancer risks

Allegations = that *n* had been deliberately chosen to be just barely large enough to detect benefits, but not large enough to detect cancer risks.

More at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7077768/

AN EXPERIMENT AT THAI TEA


Northwestern | Kellogg

ThaiTea

Thai Tea:
Experimenting with a Mobile App

Cam Nguyen felt overwhelmed. As the newly appointed director of customer engagement for Thai Tea, one of eastern Asia's fastest-growing tea and coffee chains, she saw tremendous opportunity to grow the business even further. But how to prioritize the many possibilities before her was a daunting task.