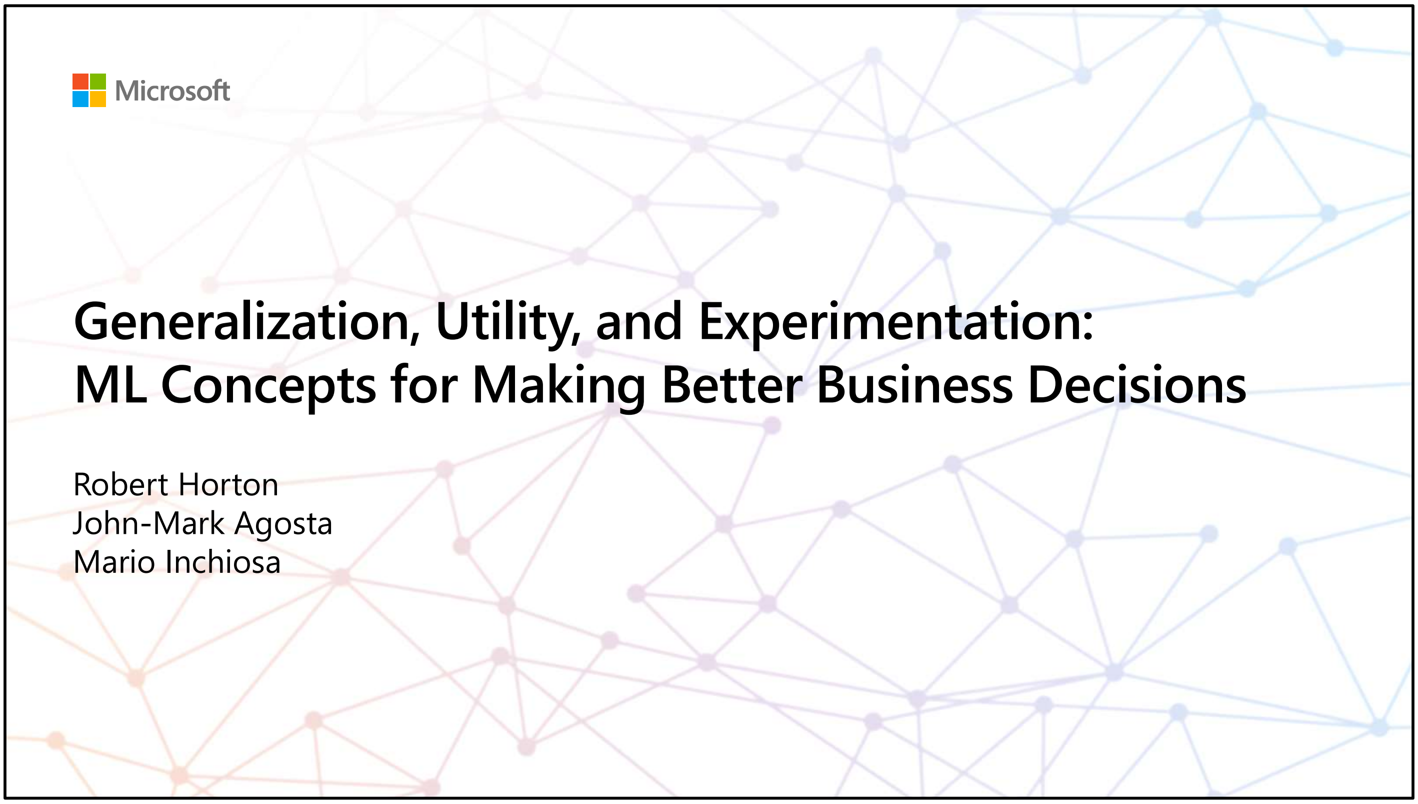
Microsoft

MLADS

MACHINE LEARNING, AI,
AND DATA SCIENCE CONFERENCE

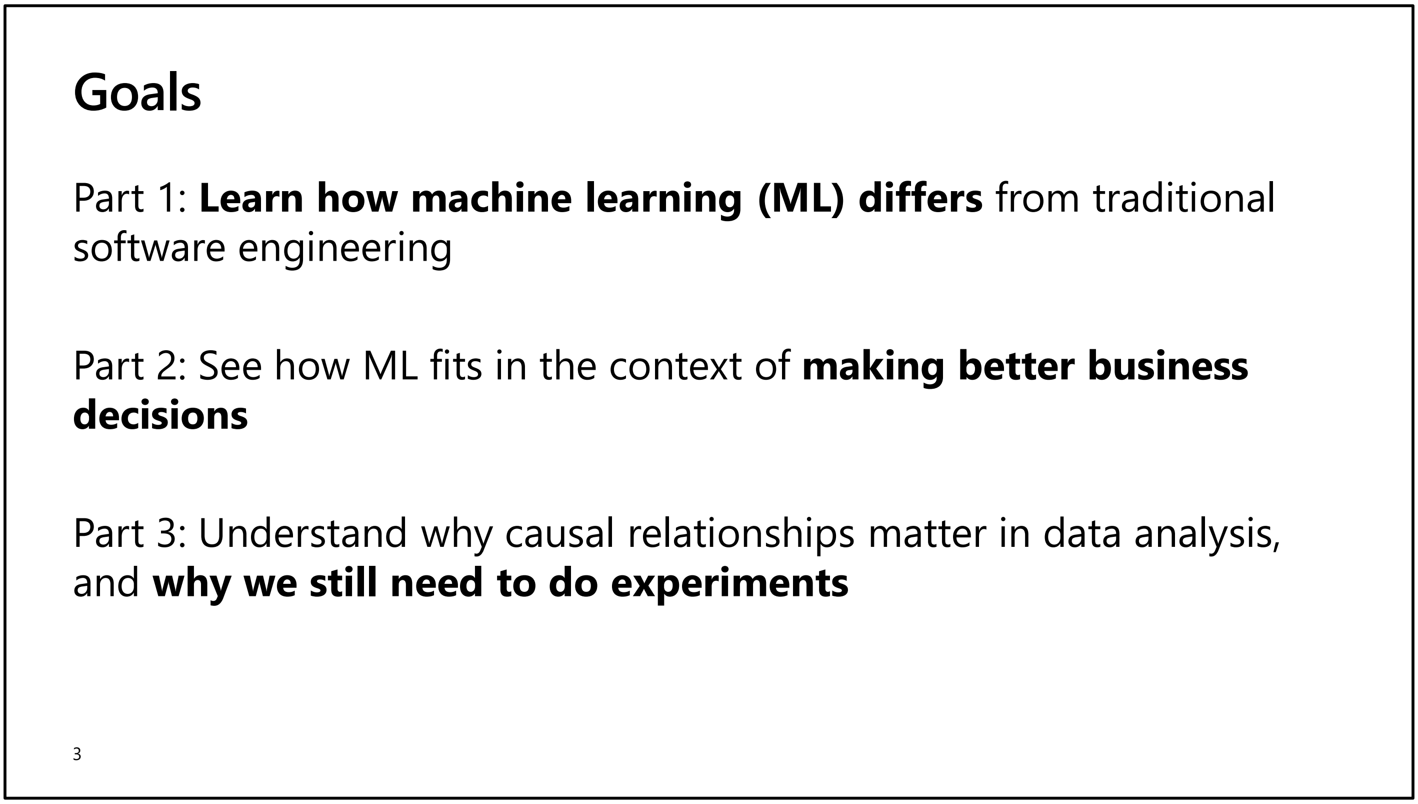
JUNE, 2020



Microsoft

Generalization, Utility, and Experimentation: ML Concepts for Making Better Business Decisions

Robert Horton
John-Mark Agosta
Mario Inchiosa



Goals

Part 1: **Learn how machine learning (ML) differs** from traditional software engineering

Part 2: See how ML fits in the context of **making better business decisions**

Part 3: Understand why causal relationships matter in data analysis, and **why we still need to do experiments**

3

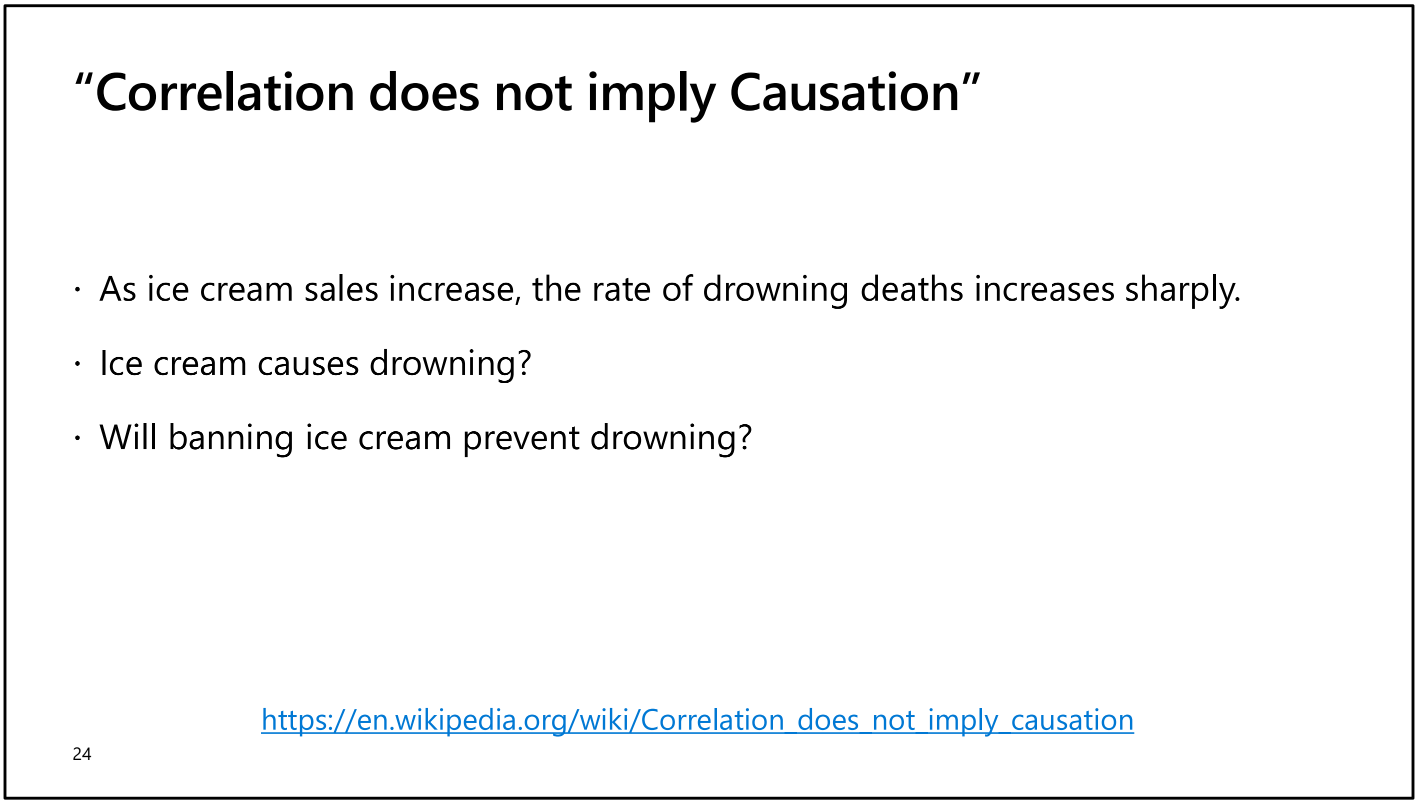
Part 3: Causality and Other Cautionary Tales

Mario Inchiosa
AI Customer Engineering (ACE) Team
Azure AI Platform

First, we will discuss how drawing conclusions about cause and effect from correlations between variables in observational data is risky.

Next, we will consider how to address this by using interventional techniques such as A/B Testing and Reinforcement Learning.

In particular, we will learn about the Azure Personalizer Service, which relies on a form of Reinforcement Learning.



“Correlation does not imply Causation”

- As ice cream sales increase, the rate of drowning deaths increases sharply.
- Ice cream causes drowning?
- Will banning ice cream prevent drowning?

https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

24

You may have heard the saying “Correlation does not imply Causation”.

In fact, there is an entire Wikipedia entry devoted to this phrase:
https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

To paraphrase Wikipedia, “Correlation does not imply Causation” refers to the fact that one cannot legitimately deduce a cause-and-effect relationship between two variables solely on the basis of an observed association or correlation between them.

For example, suppose we observe that ice cream sales and the rate of drowning deaths are correlated – when ice cream sales are low, drowning deaths are low, and when ice cream sales are high, drowning deaths are high.

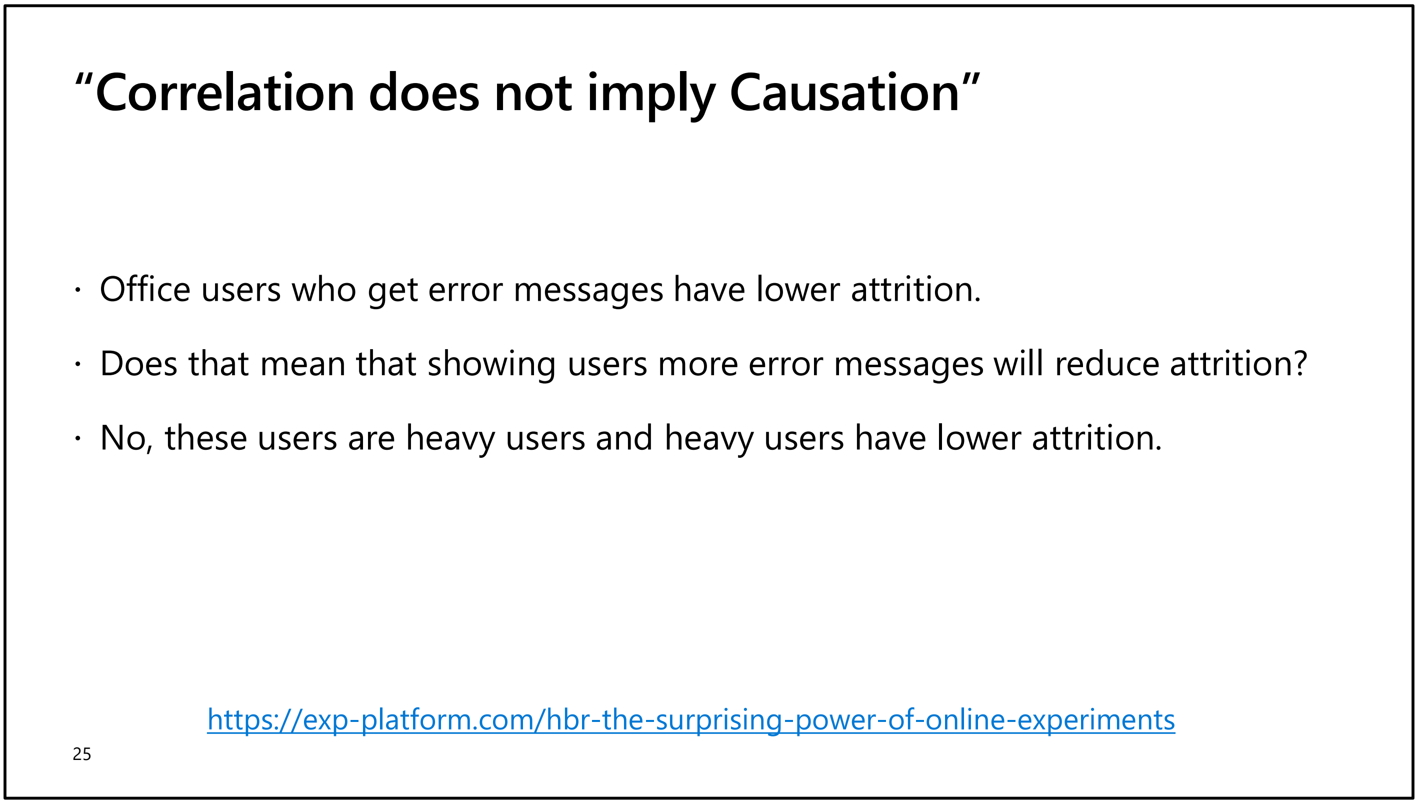
Can we conclude that there is a cause-and-effect relationship, and that high ice cream sales are causing a high rate of drowning deaths, or vice versa?

If there were a cause-and-effect relationship, with ice cream sales causing drowning deaths, then banning ice cream would indeed help prevent drowning.

Of course, there is no such cause and effect.

The true explanation is that there is a “lurking third variable”, namely the outdoor air temperature. When it is hot out, people buy more ice cream and people swim more, hence drown more.

To tease out whether banning ice cream would actually prevent drowning, we could reason about the causal relationships between ice cream sales, drowning, and outdoor temperature,
or we could do an experiment such as banning ice cream.



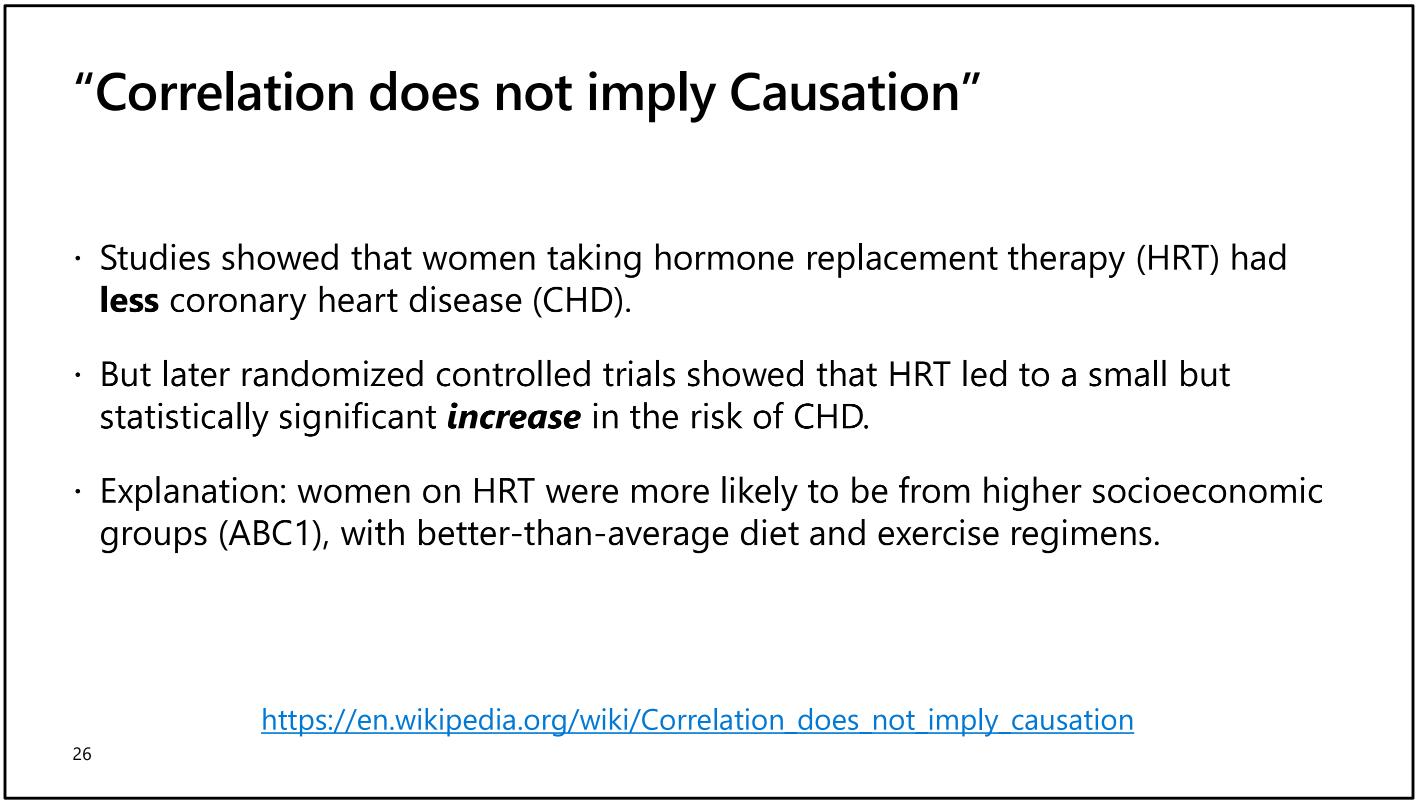
“Correlation does not imply Causation”

- Office users who get error messages have lower attrition.
- Does that mean that showing users more error messages will reduce attrition?
- No, these users are heavy users and heavy users have lower attrition.

https://exp-platform.com/hbr-the-surprising-power-of-online-experiments

Just because more error messages is correlated with lower attrition does not mean that showing all users more error messages will reduce attrition.

In this case, the “lurking third variable” is whether the user is a heavy, power user. Such users are more likely to push software to its limits and therefore encounter error messages, but they are also highly invested in the tool and therefore less likely to abandon it.

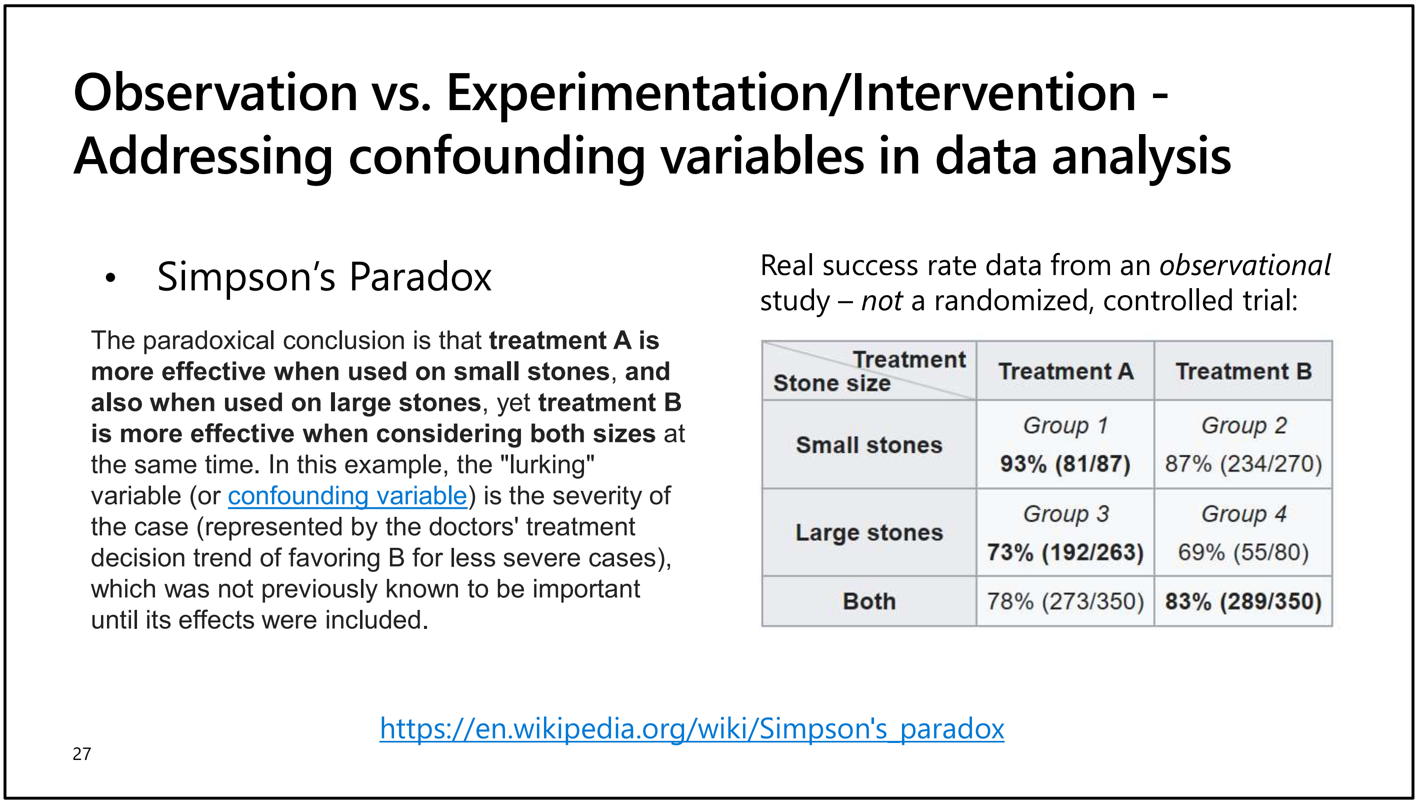


“Correlation does not imply Causation”

- Studies showed that women taking hormone replacement therapy (HRT) had **less** coronary heart disease (CHD).
- But later randomized controlled trials showed that HRT led to a small but statistically significant ***increase*** in the risk of CHD.
- Explanation: women on HRT were more likely to be from higher socioeconomic groups (ABC1), with better-than-average diet and exercise regimens.

https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Here the lurking third variable was socioeconomic status. It took an experiment, specifically a “randomized controlled trial” (RCT), to discover that more HRT would actually increase the risk of CHD.



Observation vs. Experimentation/Intervention - Addressing confounding variables in data analysis

- Simpson's Paradox

The paradoxical conclusion is that **treatment A is more effective when used on small stones**, **and also when used on large stones**, yet **treatment B is more effective when considering both sizes** at the same time. In this example, the "lurking" variable (or confounding variable) is the severity of the case (represented by the doctors' treatment decision trend of favoring B for less severe cases), which was not previously known to be important until its effects were included.

Real success rate data from an *observational* study – *not* a randomized, controlled trial:

Treatment Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

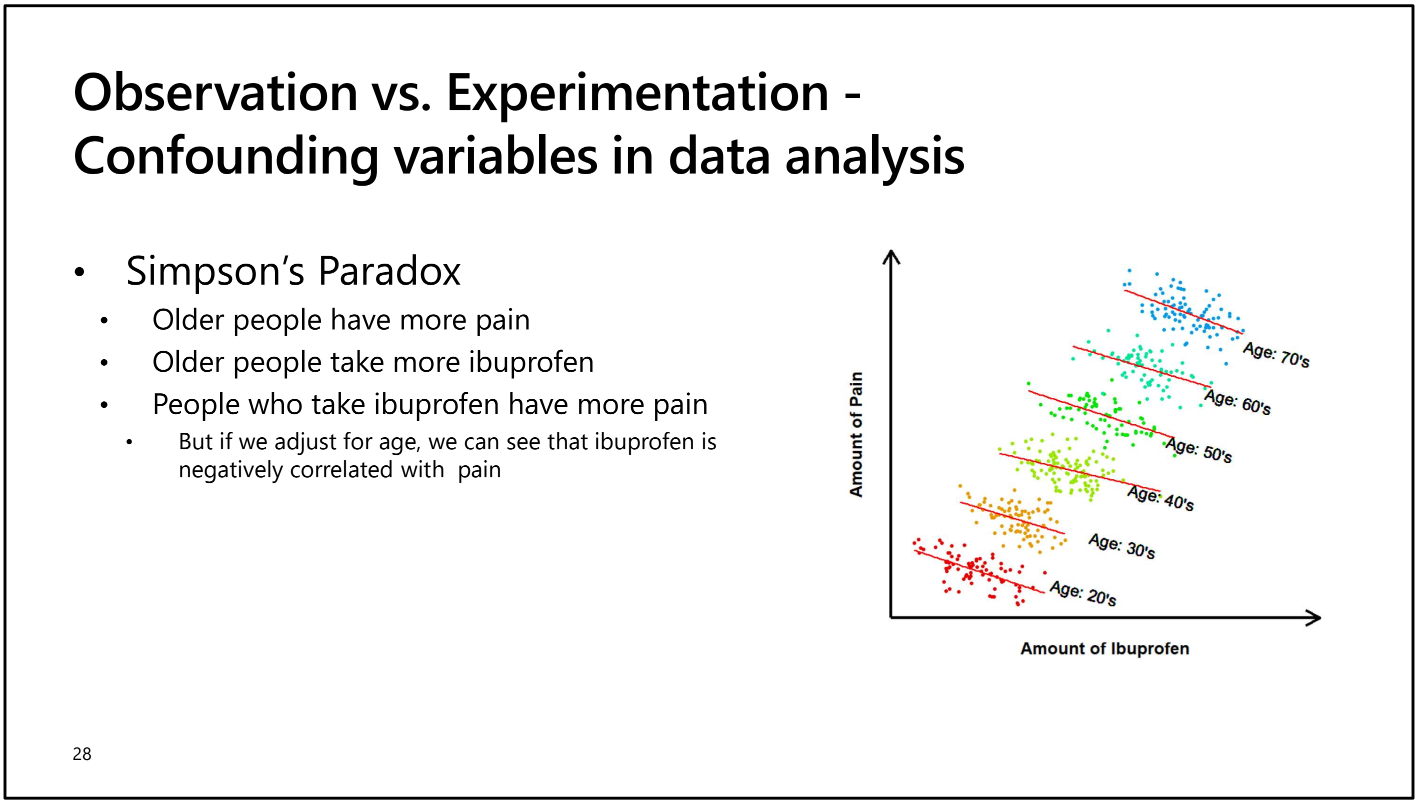
https://en.wikipedia.org/wiki/Simpson's_paradox

Simpson's Paradox is a particularly mind-bending example of how one can be led astray by purely observational data.

One sure remedy is to run an experiment where one intervenes, rather than simply observes. The gold standard is the "randomized, controlled trial."

Treatment A includes all open surgical procedures and Treatment B involves only a small puncture.

"When the less effective treatment (B) is applied more frequently to less severe cases, it can *appear* (incorrectly) to be a more effective treatment."



Observation vs. Experimentation - Confounding variables in data analysis

- Simpson’s Paradox
  - Older people have more pain
  - Older people take more ibuprofen
  - People who take ibuprofen have more pain
    - But if we adjust for age, we can see that ibuprofen is negatively correlated with pain

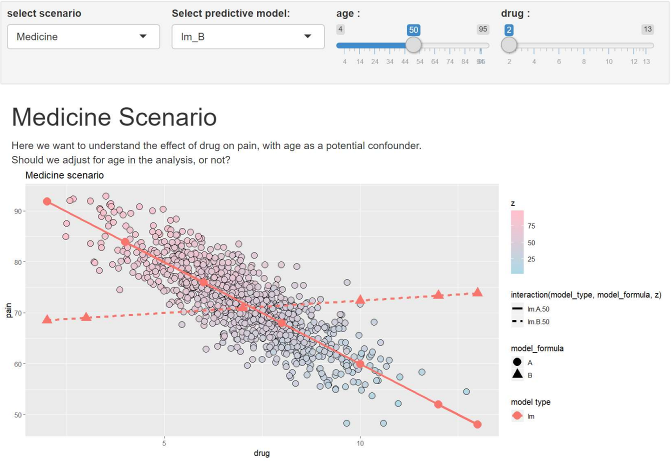
Another example of Simpson’s Paradox (note: the data plotted are notional).

The plot shows that if we consider only amount of ibuprofen and amount of pain, then more ibuprofen is correlated with more pain.

Is this saying that ibuprofen actually *causes* pain? Probably not.

When we group the data by age, we see that more ibuprofen is correlated with *less* pain, within an age group.

Simpson's Paradox – Interactive Explorer



select scenario: Medicine

Select predictive model: lm_B

age : 50

drug : 2

Medicine Scenario

Here we want to understand the effect of drug on pain, with age as a potential confounder.
Should we adjust for age in the analysis, or not?

https://ml4managers.shinyapps.io/effects_of_x_and_z

There is an interactive app for exploring Simpson's Paradox at
https://ml4managers.shinyapps.io/effects_of_x_and_z

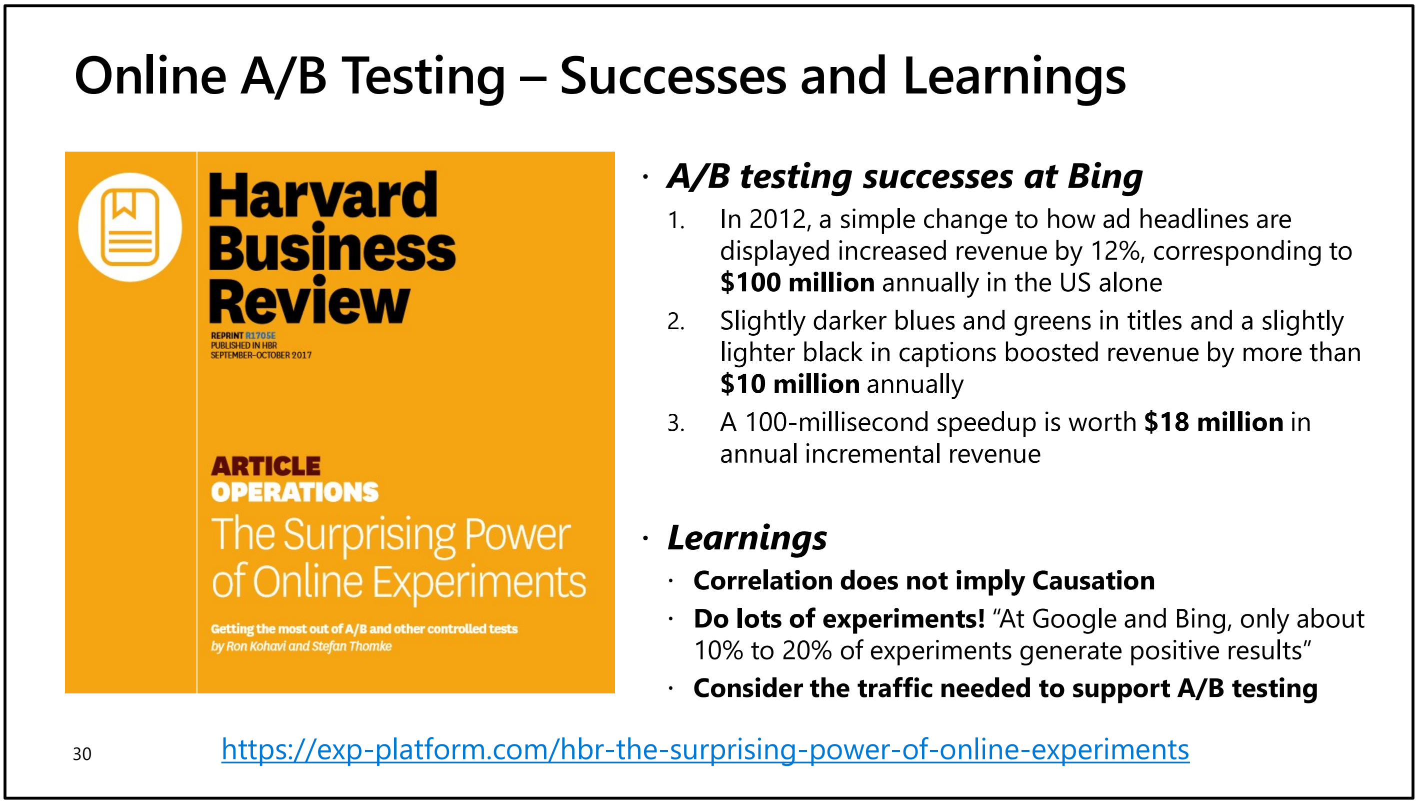
Instructions:

1. Start with the default scenario, Medicine, and model, "lm_A". Slide the "drug" slider between 1 and 13 to see how pain is correlated with drug for the whole population. Note that the more drug, the less pain.
2. Now, change the model to "lm_B". Leave the "age" slider at 52. Move the "drug" slider between 1 and 13. Note that for 52 year olds, the more drug, the more pain. This is the opposite of what we saw for the population as a whole!
3. Next, try changing the "age" slider to 75. Move the "drug" slider between 1 and 13. Note that for 75 year olds, the more drug, the more pain, just like for 52 year olds. In fact, we can see the same behavior for any specific age group.

Conclusion:

- Whether a drug produces a desired effect generally cannot be concluded by looking at data from observational studies alone. Instead, teasing out cause and effect requires domain understanding and/or randomized controlled trials (experiments).

For a deeper dive into Simpson's Paradox, you can read the following blog post:
https://blog.revolutionanalytics.com/2015/11/fun-with-simpsons-paradox-simulating-confounders.html



Online A/B Testing – Successes and Learnings

Harvard Business Review
REPRINT R1705E
PUBLISHED IN HBR
SEPTEMBER–OCTOBER 2017
ARTICLE
OPERATIONS
The Surprising Power of Online Experiments
Getting the most out of A/B and other controlled tests
by Ron Kohavi and Stefan Thomke

- ***A/B testing successes at Bing***
  1. In 2012, a simple change to how ad headlines are displayed increased revenue by 12%, corresponding to **$100 million** annually in the US alone
  2. Slightly darker blues and greens in titles and a slightly lighter black in captions boosted revenue by more than **$10 million** annually
  3. A 100-millisecond speedup is worth **$18 million** in annual incremental revenue

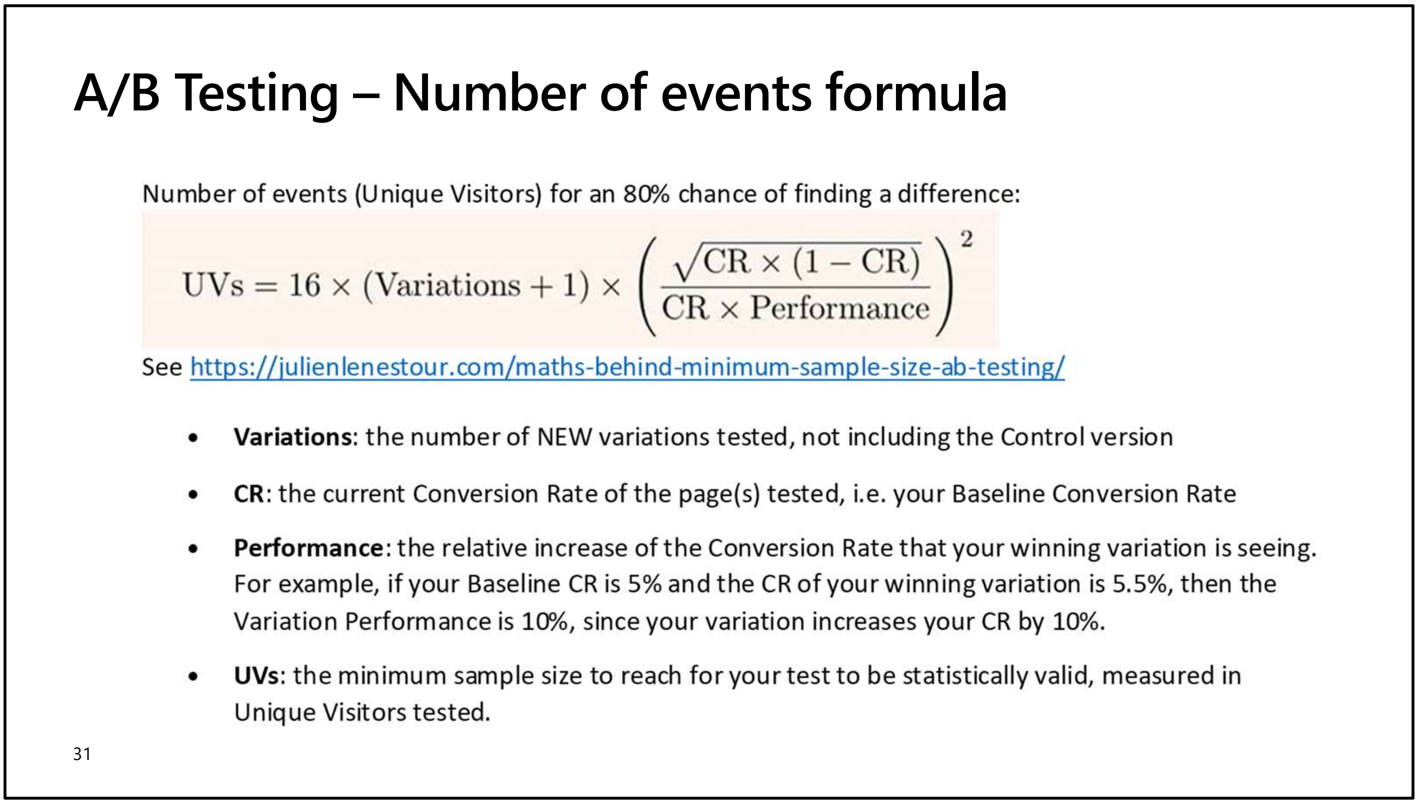
- ***Learnings***
  - **Correlation does not imply Causation**
  - **Do lots of experiments!** "At Google and Bing, only about 10% to 20% of experiments generate positive results"
  - **Consider the traffic needed to support A/B testing**

30

https://exp-platform.com/hbr-the-surprising-power-of-online-experiments

Online A/B Testing is a type of randomized controlled trial or experiment.

Given enough user interactions (e.g. web site visits), online A/B testing provides reliable, statistically significant conclusions about cause and effect.



A/B Testing – Number of events formula

Number of events (Unique Visitors) for an 80% chance of finding a difference:

$$\mathrm{UVs} = 16 \times (\mathrm{Variations} + 1) \times \left(\frac{\sqrt{\mathrm{CR} \times (1 - \mathrm{CR})}}{\mathrm{CR} \times \mathrm{Performance}} \right)^2$$

See https://julienlenestour.com/maths-behind-minimum-sample-size-ab-testing/

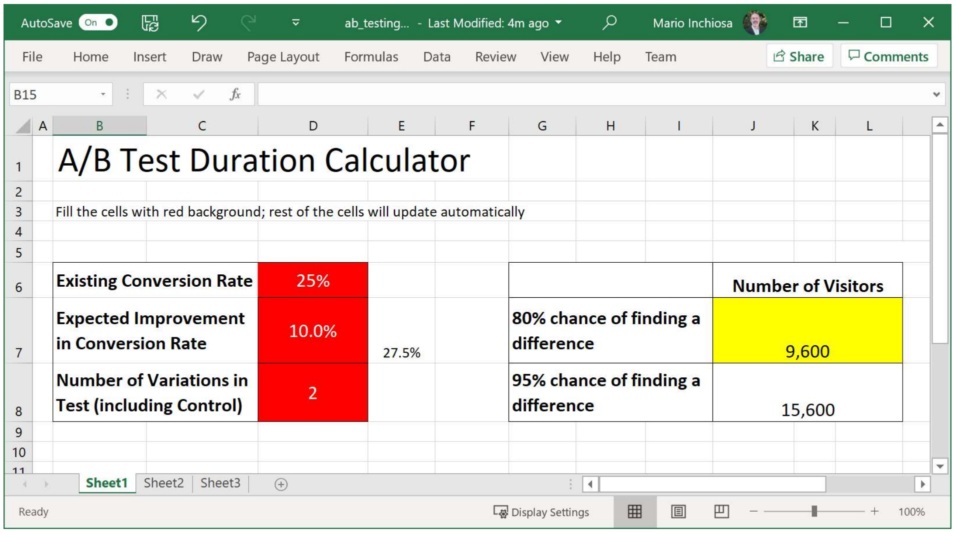
- **Variations**: the number of NEW variations tested, not including the Control version
- **CR**: the current Conversion Rate of the page(s) tested, i.e. your Baseline Conversion Rate
- **Performance**: the relative increase of the Conversion Rate that your winning variation is seeing. For example, if your Baseline CR is 5% and the CR of your winning variation is 5.5%, then the Variation Performance is 10%, since your variation increases your CR by 10%.
- **UVs**: the minimum sample size to reach for your test to be statistically valid, measured in Unique Visitors tested.

The key take-away from this formula is that detecting a subtle change in Conversion Rate requires much more data than detecting a larger change.

For example, reliably detecting a 1% increase in Conversion Rate requires 100 time more events than would be required to detect a 10% increase in Conversion Rate.

Mathematically, the number of unique visitors required depends on 1/(performance^2), where “performance” is the relative increase in Conversion Rate, e.g. performance = 0.1 for a 10% increase and 0.01 for a 1% increase.

A/B Testing – Design of Experiments



A/B Test Duration Calculator

Fill the cells with red background; rest of the cells will update automatically

Existing Conversion Rate	25%			Number of Visitors
Expected Improvement in Conversion Rate	10.0%	27.5%	80% chance of finding a difference	9,600
Number of Variations in Test (including Control)	2		95% chance of finding a difference	15,600

Test Duration Calculator (80%):

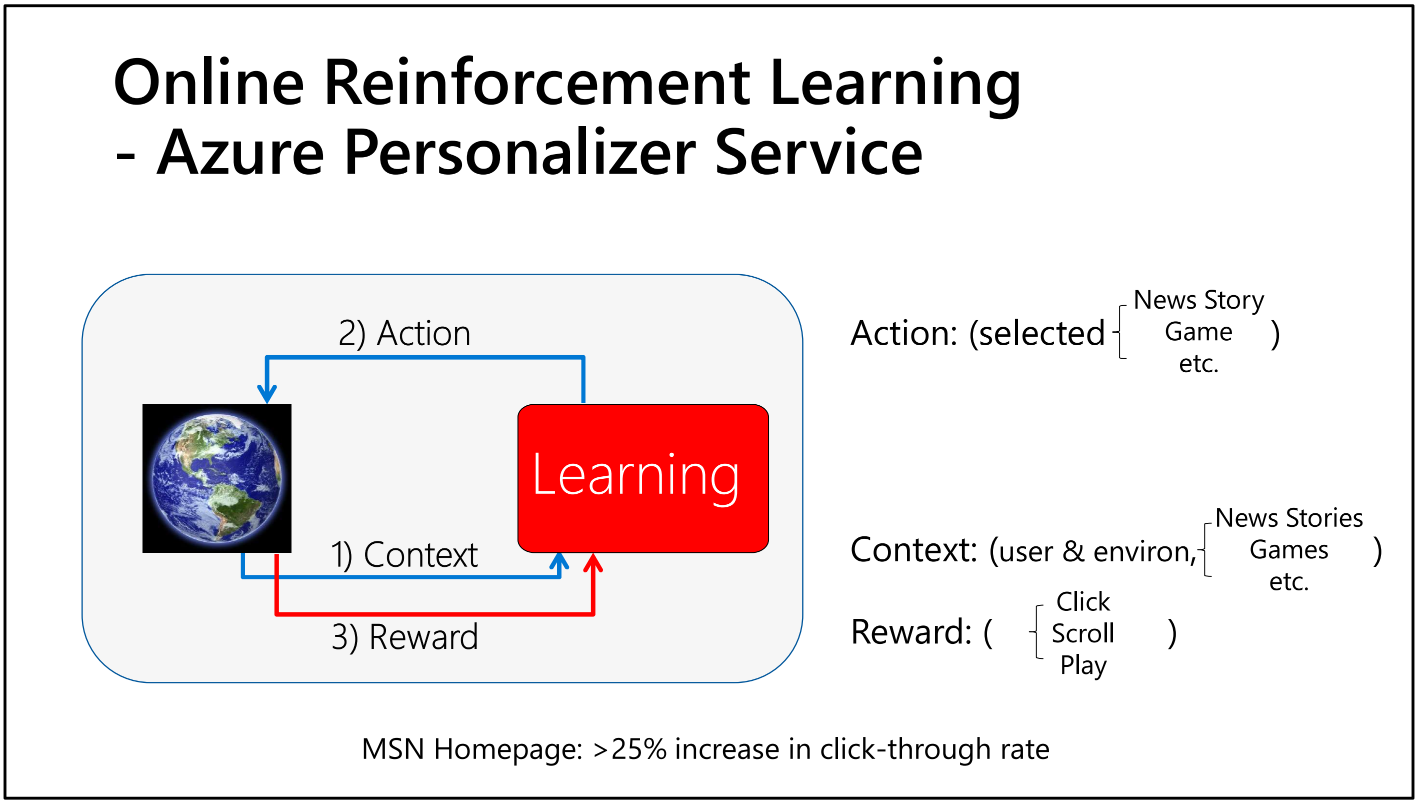
https://aka.ms/ab-duration

Test Significance Calculator:

https://aka.ms/ab-significance

32

Try https://aka.ms/ab-duration – first use the default values, and then try changing "**Minimum improvement in conversion rate you want to detect (%)**" from the default of 20% to 2%. Note how the number of days to run the test increases from 28 to **2,800!**



Online Reinforcement Learning - Azure Personalizer Service

Action: (selected { News Story / Game / etc.)

Context: (user & environ, { News Stories / Games / etc.)

Reward: ({ Click / Scroll / Play)

MSN Homepage: >25% increase in click-through rate

Azure Personalizer Service - https://azure.microsoft.com/en-us/services/cognitive-services/personalizer/

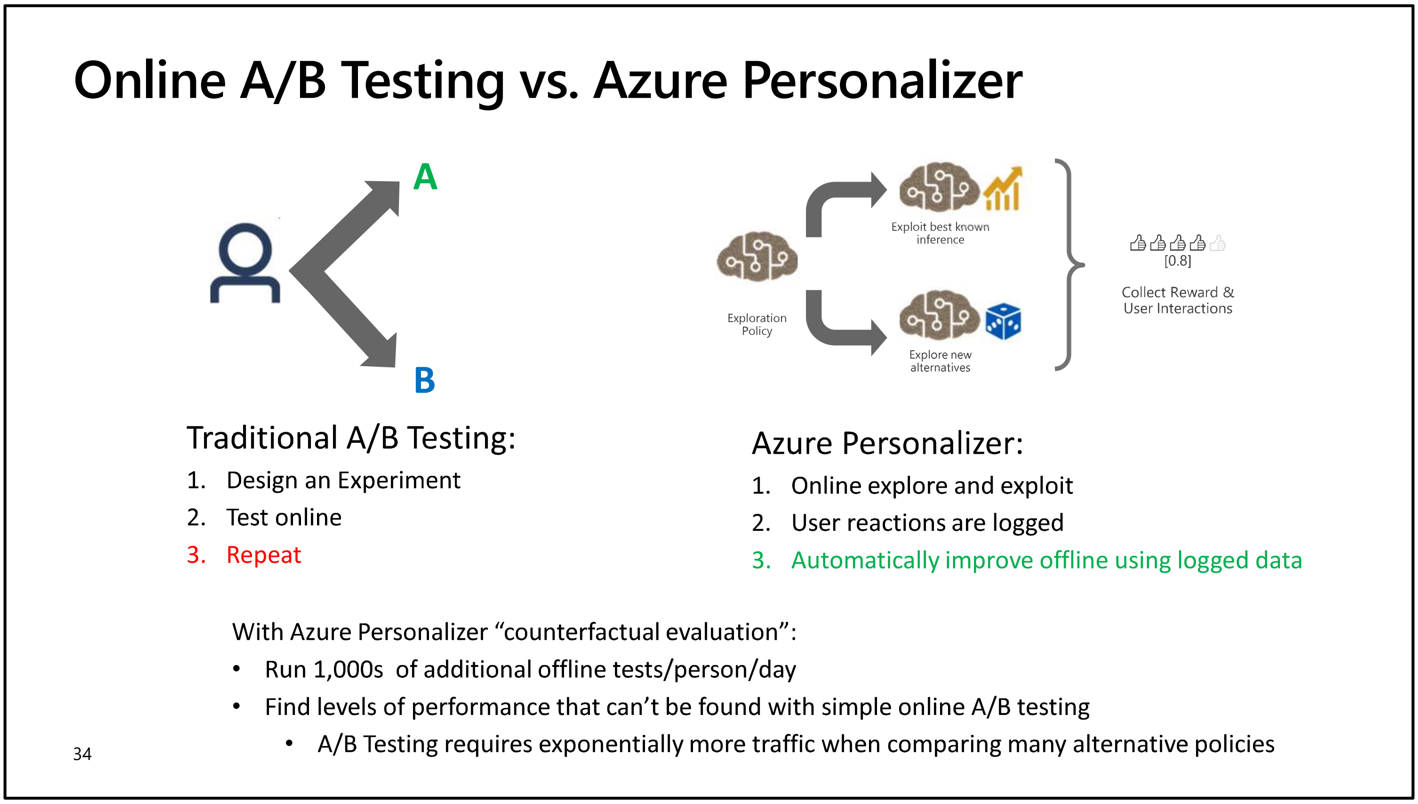
Azure Personalizer Service continuously and automatically seeks to optimize "reward."

For a given **context** (information about, e.g., a web site visitor, their browser, their approximate location, time of day, day of week, etc.),

Azure Personalizer Service will either recommend the best action learned so far for the context or will recommend a randomly chosen action.

One can configure how often Personalizer recommends a random action, e.g. 20% of the time. This is called the exploration percentage.

Exploration allows Personalizer to keep learning and adapting as conditions change.


A
B
Exploit best known inference
Exploration Policy
Explore new alternatives
[0.8]
Collect Reward & User Interactions

Online A/B Testing vs. Azure Personalizer

Traditional A/B Testing:

1. Design an Experiment
2. Test online
3. Repeat

Azure Personalizer:

1. Online explore and exploit
2. User reactions are logged
3. Automatically improve offline using logged data

With Azure Personalizer “counterfactual evaluation”:

- Run 1,000s of additional offline tests/person/day
- Find levels of performance that can’t be found with simple online A/B testing
  - A/B Testing requires exponentially more traffic when comparing many alternative policies

Personalizer can use data logged from online exploration to test additional, alternative policies via “offline evaluations”:

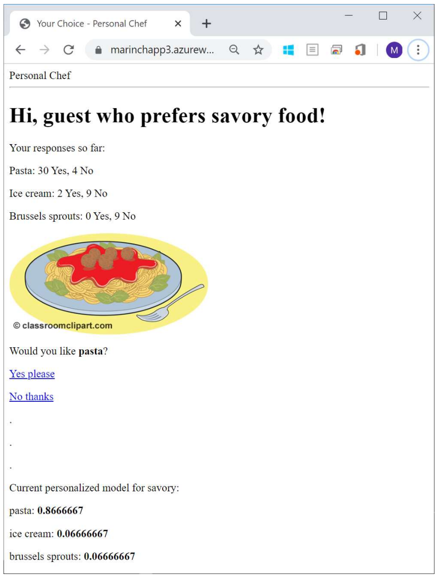
https://docs.microsoft.com/en-us/azure/cognitive-services/personalizer/concepts-offline-evaluation#how-offline-evaluations-are-done

Example A/B experiment:

1. Control group (Group A) sees the same text regardless of geographic location
2. Treatment group (Group B) sees one of 10 different versions of the text based on geographic location

Traditional online A/B testing can only test one policy of assigning versions to locations at a time.
Using offline evaluation, Personalizer can test many alternative policies.

Lab exercise – let’s train a Personal Chef!



Your Choice - Personal Chef

marinchapp3.azurew...

Personal Chef

Hi, guest who prefers savory food!

Your responses so far:

Pasta: 30 Yes, 4 No

Ice cream: 2 Yes, 9 No

Brussels sprouts: 0 Yes, 9 No

Would you like **pasta**?

Yes please

No thanks

.

.

.

Current personalized model for savory:

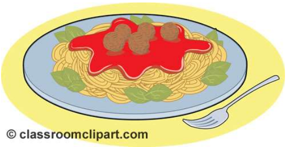
pasta: **0.8666667**

ice cream: **0.06666667**

brussels sprouts: **0.06666667**

https://aka.ms/personal-chef

Savory



Sweet

This is a simple example with one binary context feature (savory or sweet preference) and three possible actions (pasta, ice cream, and Brussels sprouts), but Personalizer can handle multiple context features, actions, and action features.

Let’s train a personal chef powered by Azure Personalizer Service!

Point your browser to https://aka.ms/personal-chef

On the first page, “**Choose Personalizer Instance,**” simply click the “**Submit**” button.

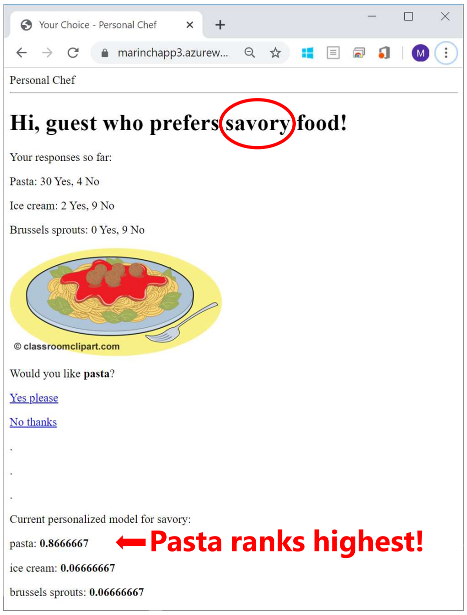
(Note: for instructions on training your *own* Personalizer instance, rather than using the shared instance, skip down two slides.)

On the next page, choose your food preference, SAVORY or SWEET.

You will be offered either pasta, ice cream, or Brussels sprouts.

Sometimes you will be offered a choice that is not the top ranked choice for your food preference. This is an example of Personalizer exploring alternative actions.

Let’s check on https://aka.ms/personal-chef …



Personal Chef

Hi, guest who prefers savory food!

Your responses so far:

Pasta: 30 Yes, 4 No

Ice cream: 2 Yes, 9 No

Brussels sprouts: 0 Yes, 9 No

Would you like **pasta**?

Yes please

No thanks

.

.

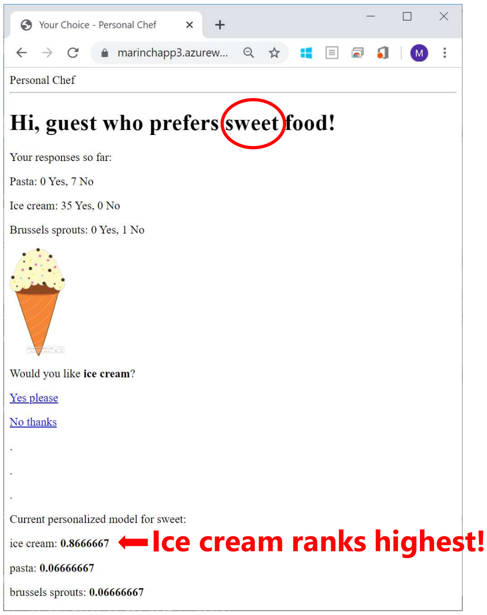
.

Current personalized model for savory:

pasta: **0.8666667** ⬅Pasta ranks highest!

ice cream: **0.06666667**

brussels sprouts: **0.06666667**



Personal Chef

Hi, guest who prefers sweet food!

Your responses so far:

Pasta: 0 Yes, 7 No

Ice cream: 35 Yes, 0 No

Brussels sprouts: 0 Yes, 1 No

Would you like **ice cream**?

Yes please

No thanks

.

.

.

Current personalized model for sweet:

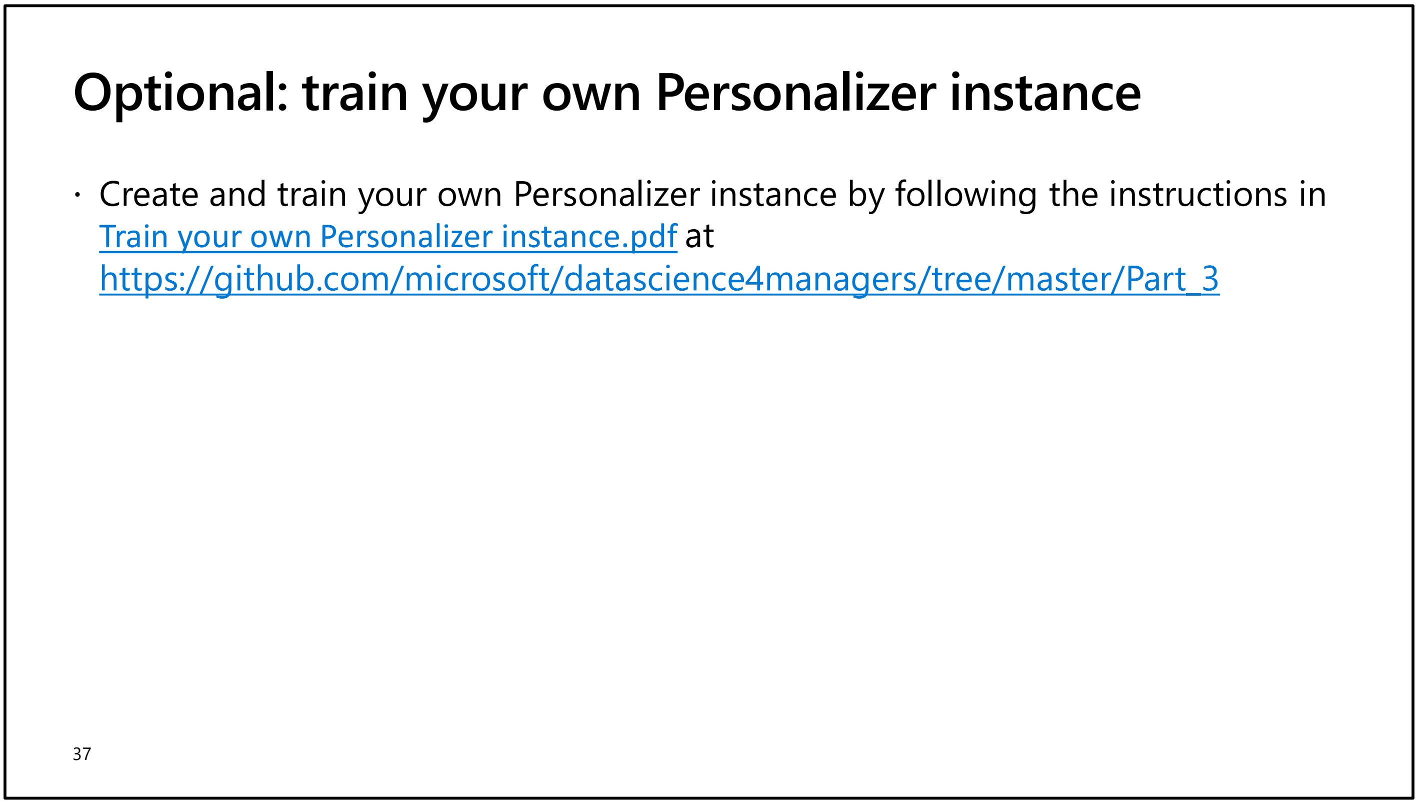
ice cream: **0.8666667** ⬅Ice cream ranks highest!

pasta: **0.06666667**

brussels sprouts: **0.06666667**

36

By scrolling to the bottom of the Personal Chef web page, you can see the current personalized rankings.

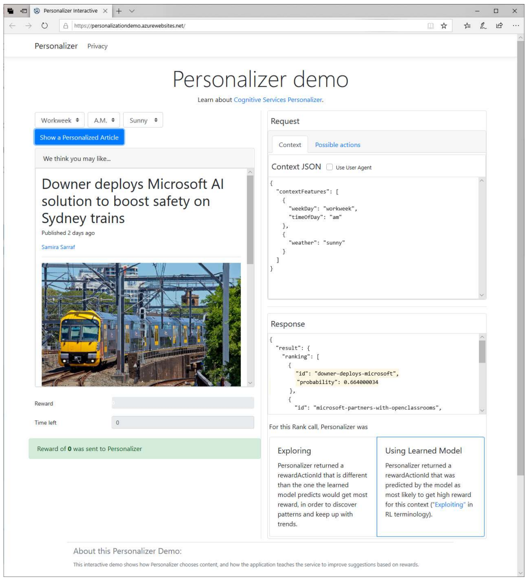


Optional: train your own Personalizer instance

- Create and train your own Personalizer instance by following the instructions in Train your own Personalizer instance.pdf at https://github.com/microsoft/datascience4managers/tree/master/Part_3

Follow the instructions here if you want to try creating and training your own Personalizer instance from scratch!

Azure Personalizer Service – Interactive Demo



Personalizer demo

Learn about Cognitive Services Personalizer.

Workweek | A.M. | Sunny

Show a Personalized Article

We think you may like...

Downer deploys Microsoft AI solution to boost safety on Sydney trains

Published 2 days ago

Samira Sarraf

Reward

Time left 0

Reward of **0** was sent to Personalizer

Request

Context | Possible actions

Context JSON □ Use User Agent

```
{
  "contextFeatures": [
    {
      "weekDay": "workweek",
      "timeOfDay": "am"
    },
    {
      "weather": "sunny"
    }
  ]
}
```

Response

```
{
  "result": {
    "ranking": [
      {
        "id": "downer-deploys-microsoft",
        "probability": 0.664000034
      },
      {
        "id": "microsoft-partners-with-openclassrooms",
```

For this Rank call, Personalizer was

Exploring

Personalizer returned a rewardActionId that is different than the one the learned model predicts would get most reward, in order to discover patterns and keep up with trends.

Using Learned Model

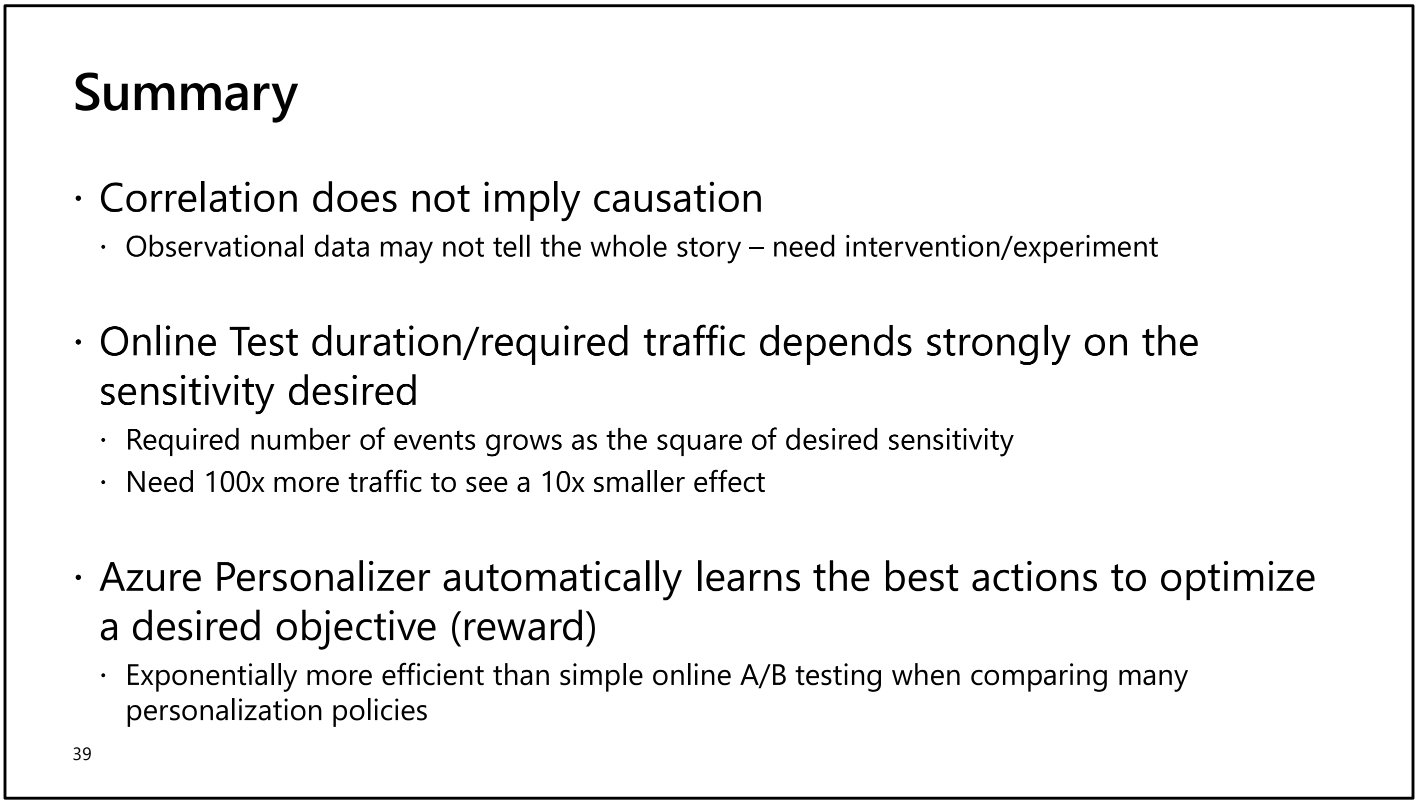
Personalizer returned a rewardActionId that was predicted by the model as most likely to get high reward for this context ("Exploiting" in RL terminology).

About this Personalizer Demo:

This interactive demo shows how Personalizer chooses content, and how the application teaches the service to improve suggestions based on rewards.

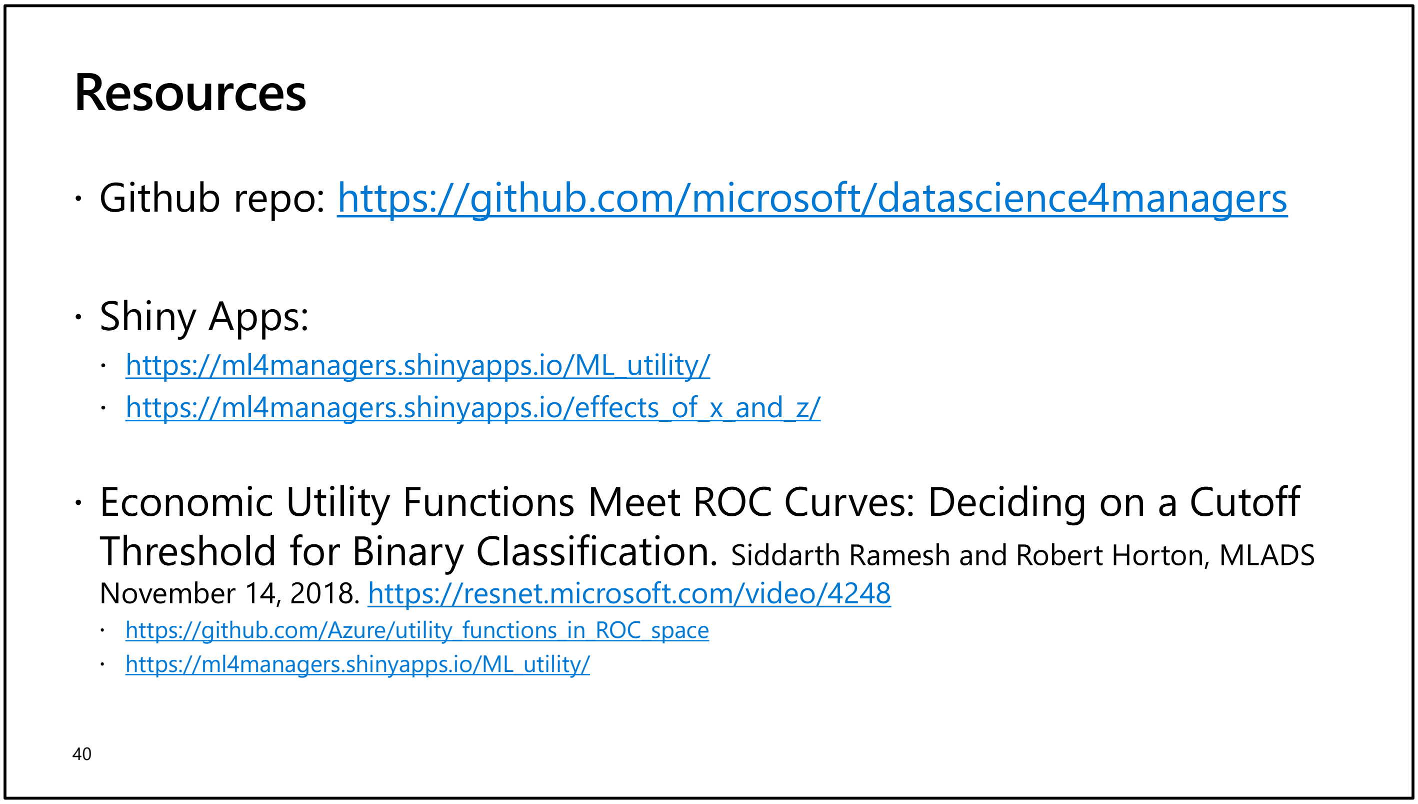
https://aka.ms/personalizer-demo

With this demo, we can peek behind the covers and see the context features and action list sent to Personalizer, along with Personalizer’s response.



Summary

- Correlation does not imply causation
  - Observational data may not tell the whole story – need intervention/experiment
- Online Test duration/required traffic depends strongly on the sensitivity desired
  - Required number of events grows as the square of desired sensitivity
  - Need 100x more traffic to see a 10x smaller effect
- Azure Personalizer automatically learns the best actions to optimize a desired objective (reward)
  - Exponentially more efficient than simple online A/B testing when comparing many personalization policies

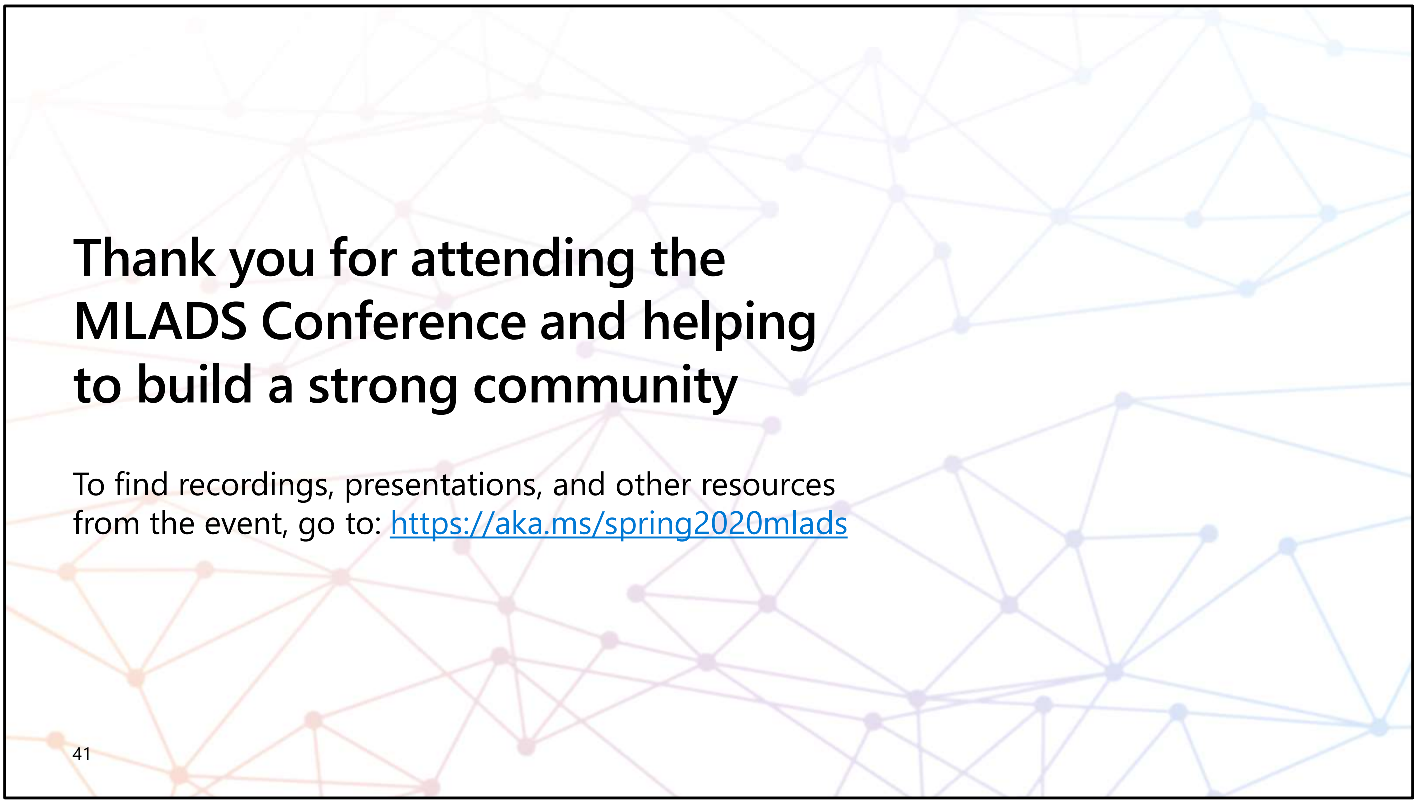


Resources

- Github repo: https://github.com/microsoft/datascience4managers

- Shiny Apps:
  - https://ml4managers.shinyapps.io/ML_utility/
  - https://ml4managers.shinyapps.io/effects_of_x_and_z/

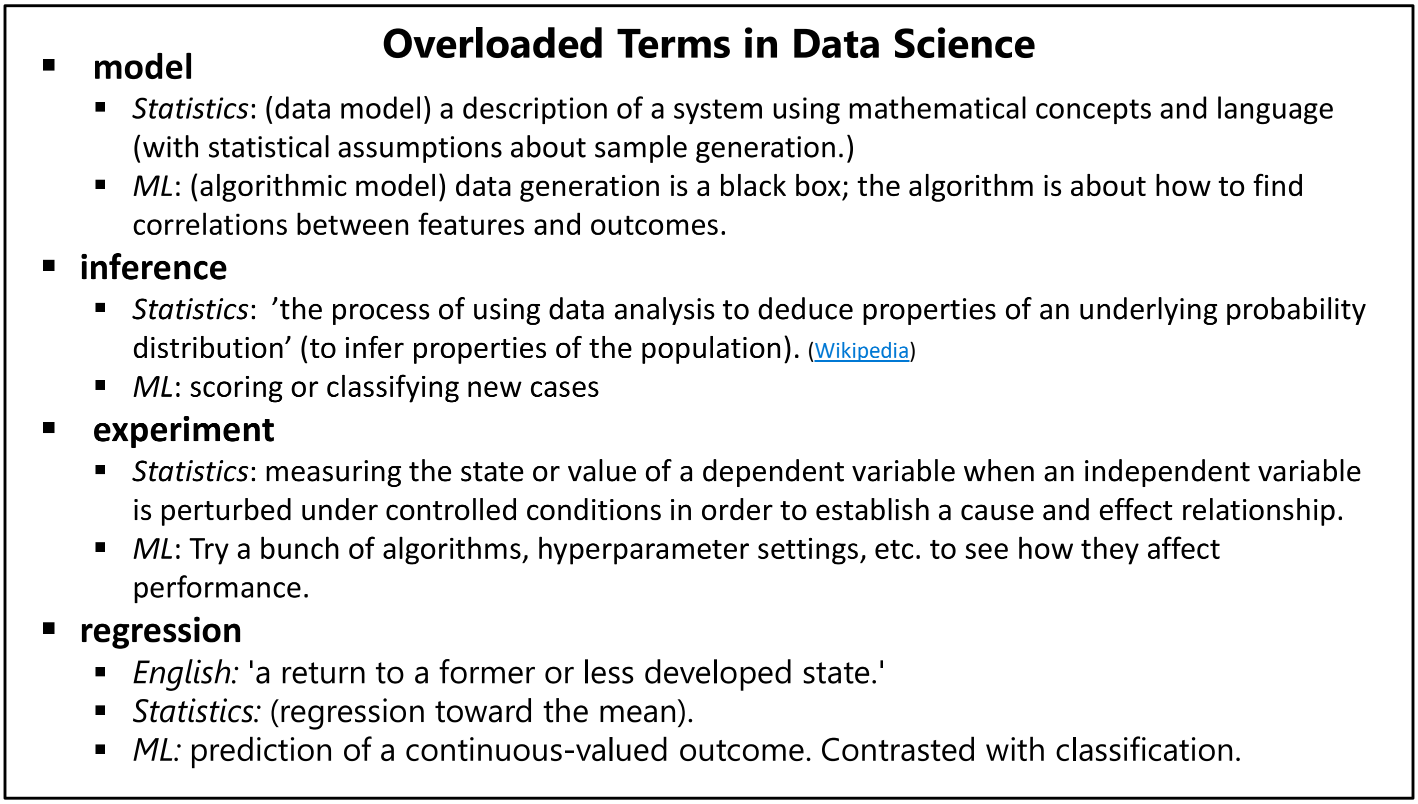
- Economic Utility Functions Meet ROC Curves: Deciding on a Cutoff Threshold for Binary Classification. Siddarth Ramesh and Robert Horton, MLADS November 14, 2018. https://resnet.microsoft.com/video/4248
  - https://github.com/Azure/utility_functions_in_ROC_space
  - https://ml4managers.shinyapps.io/ML_utility/



Thank you for attending the MLADS Conference and helping to build a strong community

To find recordings, presentations, and other resources from the event, go to: https://aka.ms/spring2020mlads

Microsoft



Overloaded Terms in Data Science

- **model**
  - *Statistics*: (data model) a description of a system using mathematical concepts and language (with statistical assumptions about sample generation.)
  - *ML*: (algorithmic model) data generation is a black box; the algorithm is about how to find correlations between features and outcomes.
- **inference**
  - *Statistics*: 'the process of using data analysis to deduce properties of an underlying probability distribution' (to infer properties of the population). (Wikipedia)
  - *ML*: scoring or classifying new cases
- **experiment**
  - *Statistics*: measuring the state or value of a dependent variable when an independent variable is perturbed under controlled conditions in order to establish a cause and effect relationship.
  - *ML*: Try a bunch of algorithms, hyperparameter settings, etc. to see how they affect performance.
- **regression**
  - *English:* 'a return to a former or less developed state.'
  - *Statistics:* (regression toward the mean).
  - *ML:* prediction of a continuous-valued outcome. Contrasted with classification.

Wikipedia: "**regression toward** (or **to**) **the mean** is the phenomenon that arises if a random variable is extreme on its first measurement but closer to the mean or average on its second measurement and if it is extreme on its second measurement but closer to the average on its first."

"Indeed, the statistician David Freedman used to say that if the topic of regression comes up in a criminal or civil trial, the side that must explain regression to the jury will lose the case." **-** from *Thinking, Fast and Slow* by Daniel Kahneman

More terms:

Precision:
Science or engineering: the number of decimal places to which a measurement is made.
Data science: positive predictive value.