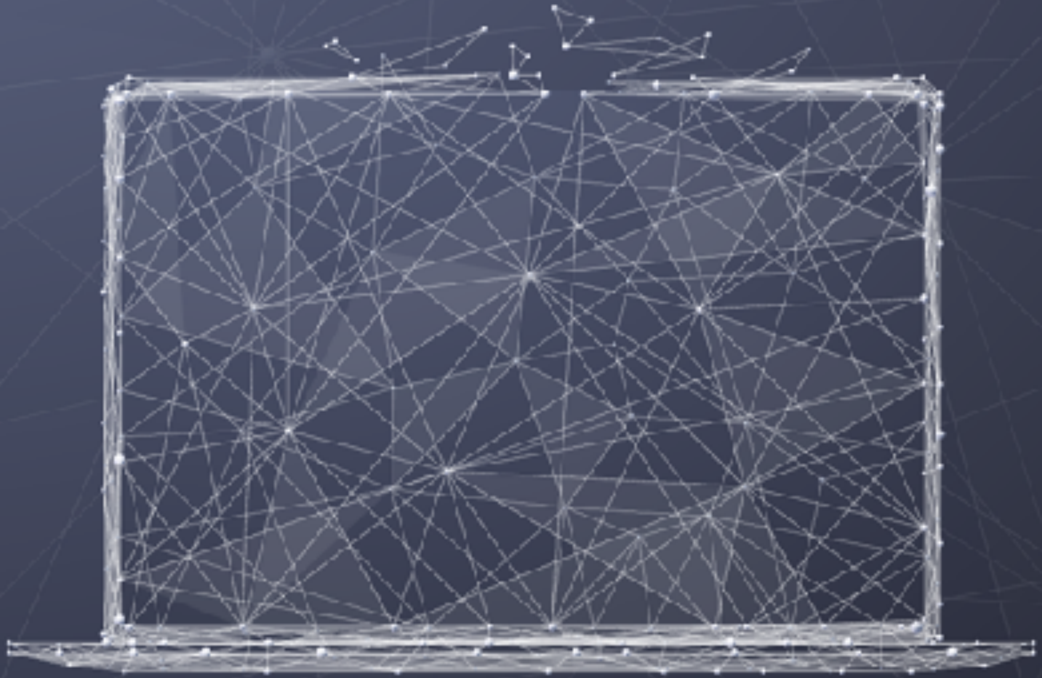
Data Science
Foundations of Decision Making

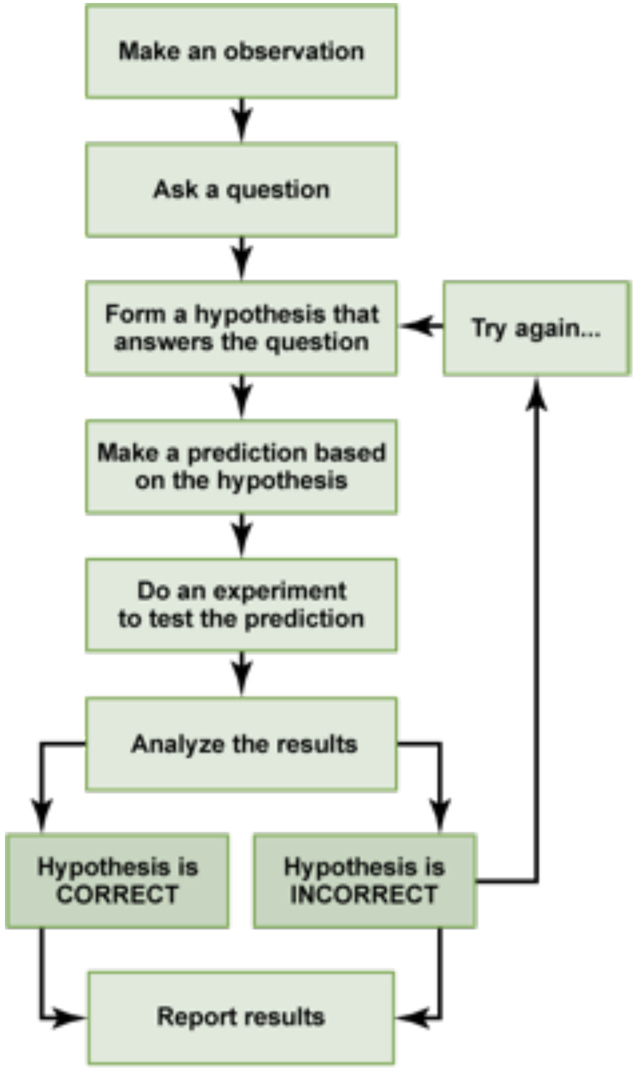
AB testing



PURDUE UNIVERSITY | College of Science

Conjecture and test

- The best data scientists are skeptics
- “If a statistic/figure looks interesting or unusual it is probably wrong.” Twyman’s Law
- If you don’t formulate a conjecture about your data/process/model and then test your idea, you will make mistakes
- Example: Bing experimentation platform for online A/B testing


Make an observation

Ask a question

Form a hypothesis that answers the question

Try again...

Make a prediction based on the hypothesis

Do an experiment to test the prediction

Analyze the results

Hypothesis is CORRECT

Hypothesis is INCORRECT

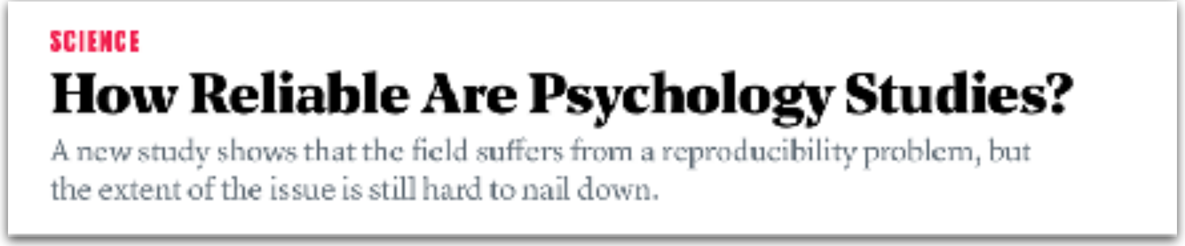
Report results

Testing conjectures in data science

- Making a claim about discovered pattern or estimated model?
  - What population are you generalizing to?
- Making a claim about a model/algorithm within a single domain?
  - Do you want to predict model/algorithm accuracy or choose between methods?
- Making a claim about a model/algorithm across multiple domains?
  - What representative data characteristic is key to method success?

Testing conjectures in data science

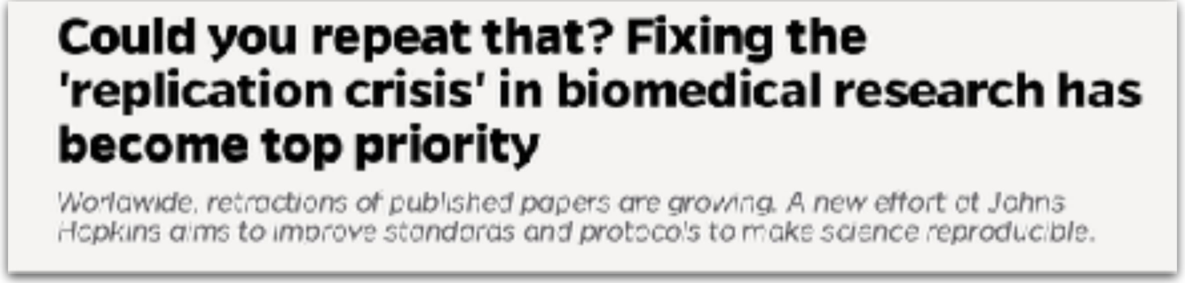
- Making a claim about discovered pattern or estimated model?
  - What population are you generalizing to?
- Making a claim about a model/algorithm within a single domain?
  - Do you want to predict model/algorithm accuracy or choose between methods?
- Making a claim about a model/algorithm across multiple domains?
  - What representative data characteristic is key to method success?



SCIENCE

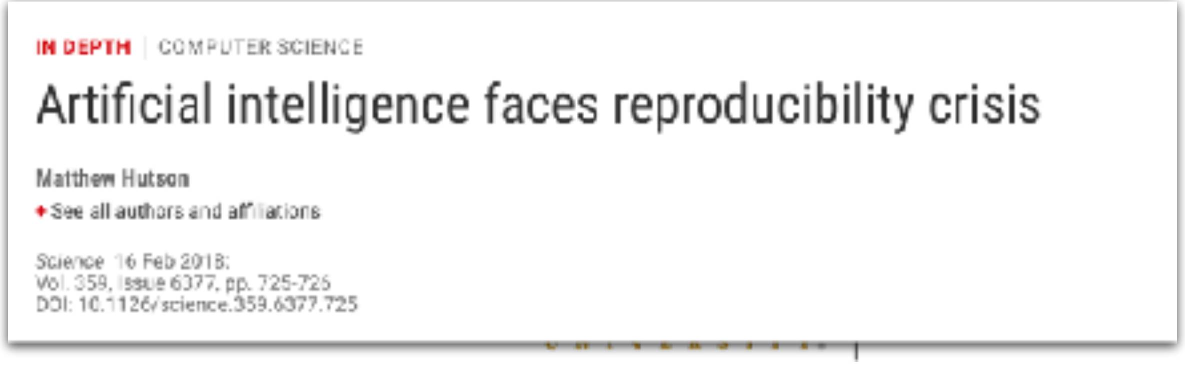
How Reliable Are Psychology Studies?

A new study shows that the field suffers from a reproducibility problem, but the extent of the issue is still hard to nail down.



Could you repeat that? Fixing the 'replication crisis' in biomedical research has become top priority

Worldwide, retractions of published papers are growing. A new effort at Johns Hopkins aims to improve standards and protocols to make science reproducible.



IN DEPTH | COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Matthew Hutson

+ See all authors and affiliations

Science 16 Feb 2018:
Vol. 359, Issue 6377, pp. 725-726
DOI: 10.1126/science.359.6377.725

Testing conjectures in data science

- Making a claim about discovered pattern or estimated model?
  - What population are you generalizing to?
- Making a claim about a model/algorithm within a single domain?
  - Do you want to predict model/algorithm accuracy or choose between methods?
- Making a claim about a model/algorithm across multiple domains?
  - What representative data characteristic is key to method success?

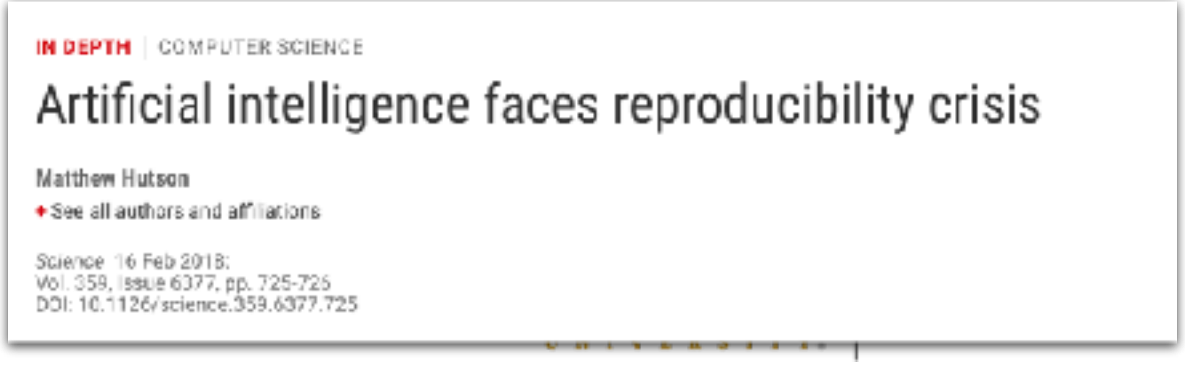
SCIENCE

How Reliable Are Psychology Studies?

A new study shows that the field suffers from a reproducibility problem, but the extent of the issue is still hard to nail down.

Could you repeat that? Fixing the 'replication crisis' in biomedical research has become top priority

Worldwide, retractions of published papers are growing. A new effort at Johns Hopkins aims to improve standards and protocols to make science reproducible.



IN DEPTH | COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Matthew Hutson

+ See all authors and affiliations

Science 16 Feb 2018:
Vol. 359, Issue 6377, pp. 725-726
DOI: 10.1126/science.359.6377.725

The life of an idea (Ronny Kohavi, Microsoft Research)

- Microsoft team proposed to change the way ad titles were displayed on Bing (2012)
  - One of hundreds of ideas proposed, other features were ranked as more valuable
  - Implementation was delayed for >6 months
  - Engineer decided it was trivial to implement in a few days, so started a controlled experiment (A/B test) to evaluate

Source: Ronny Kohavi

The life of an idea (Ronny Kohavi, Microsoft Research)

- Microsoft team proposed to change the way ad titles were displayed on Bing (2012)
  - One of hundreds of ideas proposed, other features were ranked as more valuable
  - Implementation was delayed for >6 months
  - Engineer decided it was trivial to implement in a few days, so started a controlled experiment (A/B test) to evaluate

Source: Ronny Kohavi

The life of an idea (Ronny Kohavi, Microsoft Research)

- Result
  - When running A/B test, a system alert fired that Bing was making too much money… the idea increased Bing’s revenue by 12% without hurting other metrics
  - Hundreds of engineers work on Bing Ads and increase revenue by 1.5% in a good month. Thus, simple change to titles was worth the equivalent of over 100 person years of work
- Takeaway: We are terrible at assessing the value of ideas. The best revenue-generating idea in Bing’s history was badly rated and delayed for months!

Source: Ronny Kohavi

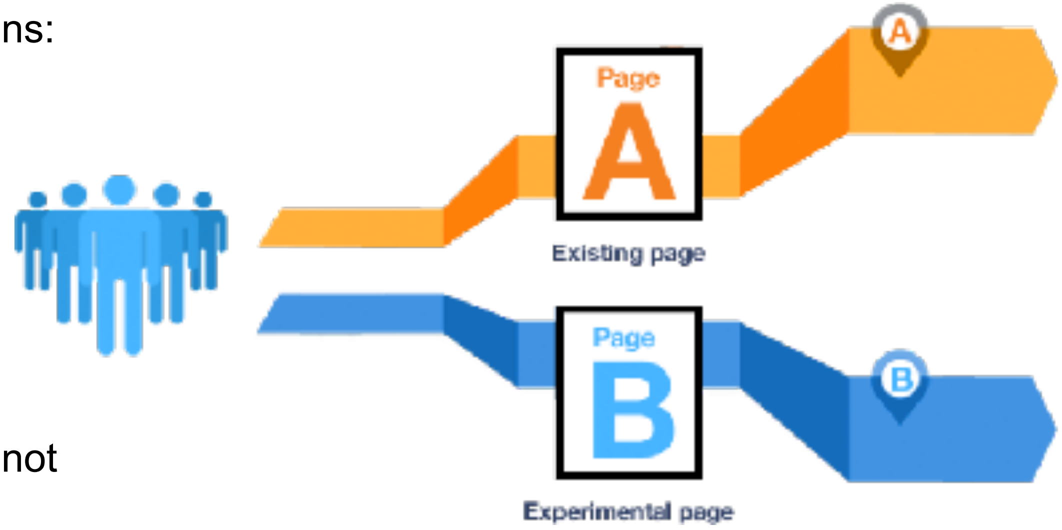
The life of an idea (Ronny Kohavi, Microsoft Research)

- Result
  - When running A/B test, a system alert fired that Bing was making too much money… the idea increased Bing's revenue by 12% without hurting other metrics
  - Hundreds of engineers work on Bing Ads and increase revenue by 1.5% in a good month. Thus, simple change to titles was worth the equivalent of over 100 person years of work
- Takeaway: We are terrible at assessing the value of ideas. The best revenue-generating idea in Bing's history was badly rated and delayed for months!

Source: Ronny Kohavi

A/B Testing

- Randomly split traffic between two (or more) versions:
  - A (Control) vs. B (Treatment)
  - Collect metrics of interest, and analyze
- A/B test is the simplest controlled experiment,
  - A/B/n refers to multiple treatments
- Must run statistical tests to confirm differences are not due to chance
- Best scientific way to prove causality, i.e., the changes in metrics are caused by changes introduced in the treatment(s)



A/B Testing at Bing (Ronny Kohavi, Microsoft Research)

- Features are built because teams believe they are useful. But most experiments show that features fail to improve metrics they were designed for

- Experiments at Microsoft shows that only 1/3 of ideas improve performance and 1/3 actually decrease performance

- Bing success rate is lower. The low success rate has been documented many times across multiple companies

Source: Ronny Kohavi

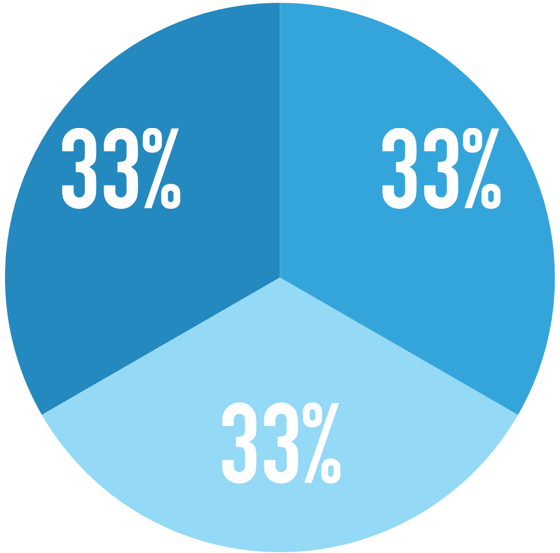
A/B Testing at Bing (Ronny Kohavi, Microsoft Research)

- Features are built because teams believe they are useful. But most experiments show that features fail to improve metrics they were designed for
- Experiments at Microsoft shows that only 1/3 of ideas improve performance and 1/3 actually decrease performance
- Bing success rate is lower. The low success rate has been documented many times across multiple companies

Source: Ronny Kohavi

A/B Testing at Bing (Ronny Kohavi, Microsoft Research)

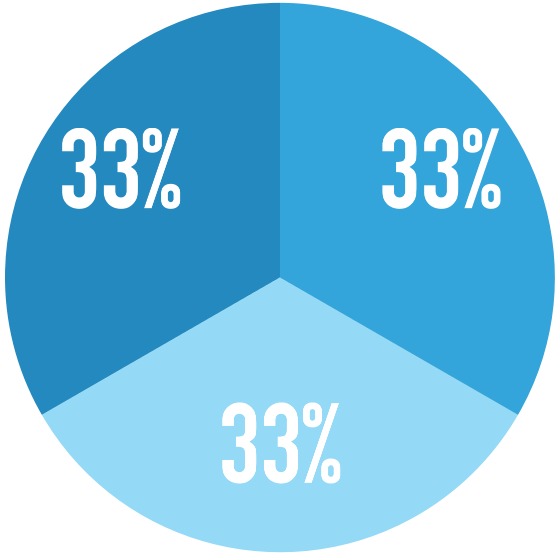
- Features are built because teams believe they are useful. But most experiments show that features fail to improve metrics they were designed for
- Experiments at Microsoft shows that only 1/3 of ideas improve performance and 1/3 actually decrease performance
- Bing success rate is lower. The low success rate has been documented many times across multiple companies



Source: Ronny Kohavi

A/B Testing at Bing (Ronny Kohavi, Microsoft Research)

- Features are built because teams believe they are useful. But most experiments show that features fail to improve metrics they were designed for
- Experiments at Microsoft shows that only 1/3 of ideas improve performance and 1/3 actually decrease performance
- Bing success rate is lower. The low success rate has been documented many times across multiple companies



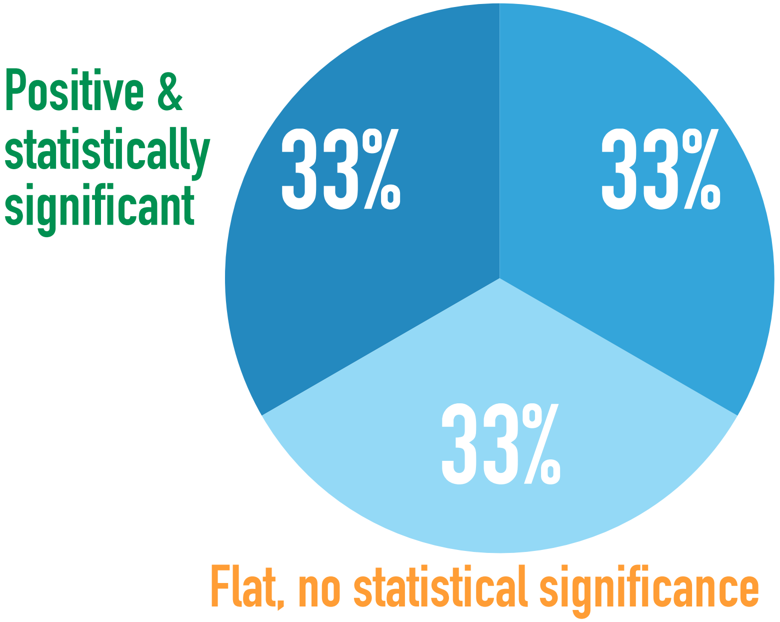
Flat, no statistical significance

PURDUE UNIVERSITY | College of Science

Source: Ronny Kohavi

A/B Testing at Bing (Ronny Kohavi, Microsoft Research)

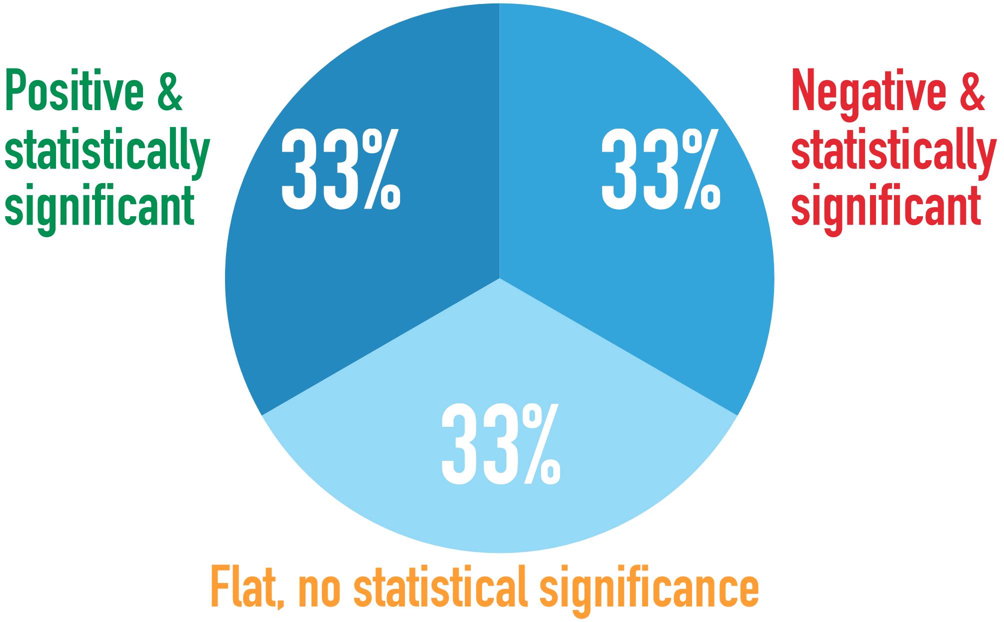
- Features are built because teams believe they are useful. But most experiments show that features fail to improve metrics they were designed for
- Experiments at Microsoft shows that only 1/3 of ideas improve performance and 1/3 actually decrease performance
- Bing success rate is lower. The low success rate has been documented many times across multiple companies



Source: Ronny Kohavi

A/B Testing at Bing (Ronny Kohavi, Microsoft Research)

- Features are built because teams believe they are useful. But most experiments show that features fail to improve metrics they were designed for
- Experiments at Microsoft shows that only 1/3 of ideas improve performance and 1/3 actually decrease performance
- Bing success rate is lower. The low success rate has been documented many times across multiple companies

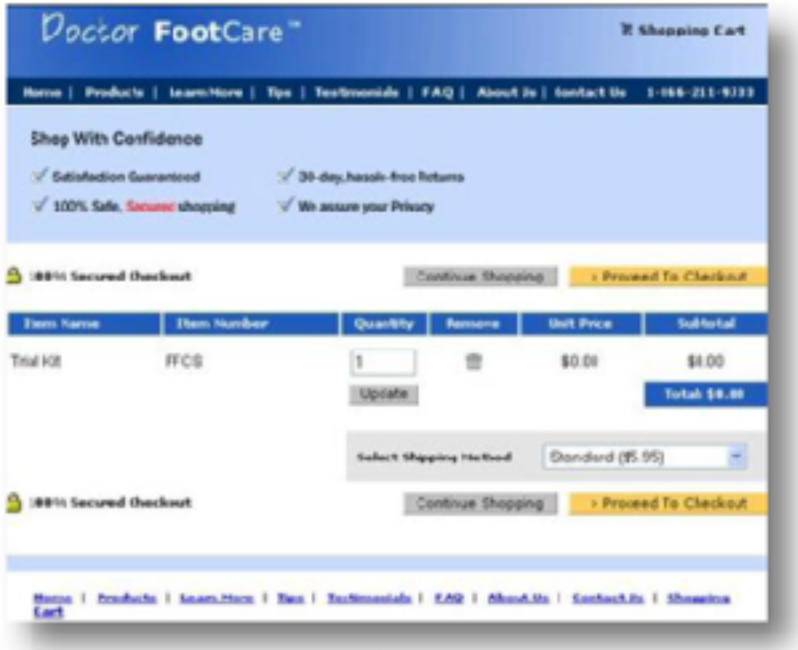


Purdue University | College of Science

Source: Ronny Kohavi

AB testing example

- Can you guess which page has a higher conversion rate and whether the difference is significant?



Doctor FootCare™ | Shopping Cart

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | [illegible]

Shop With Confidence

✓ Satisfaction Guaranteed ✓ 30-day, hassle-free Returns

✓ 100% Safe, Secure shopping ✓ We assure your Privacy

100% Secured Checkout | Continue Shopping | › Proceed To Checkout

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	1	🗑	$0.00	$4.00

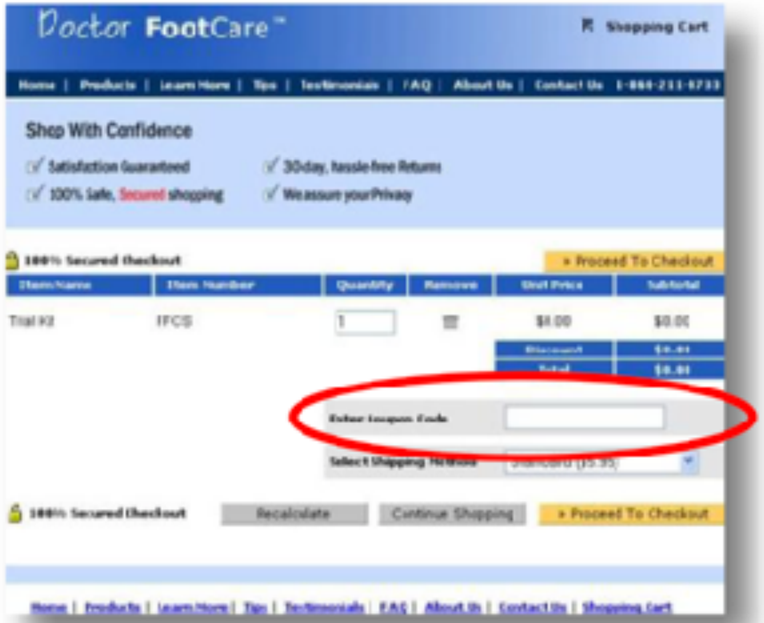
Update | Total: $[illegible]

Select Shipping Method: Standard ($5.95)

100% Secured Checkout | Continue Shopping | › Proceed To Checkout

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | Shopping Cart

A



Doctor FootCare™ | Shopping Cart

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | [illegible]

Shop With Confidence

✓ Satisfaction Guaranteed ✓ 30-day, hassle-free Returns

✓ 100% Safe, Secured shopping ✓ We assure your Privacy

100% Secured Checkout | » Proceed To Checkout

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	1	🗑	$4.00	$0.00
				Discount	$[illegible]
				Total	$[illegible]

Enter Coupon Code:

Select Shipping Method: Standard ($5.95)

100% Secured Checkout | Recalculate | Continue Shopping | » Proceed To Checkout

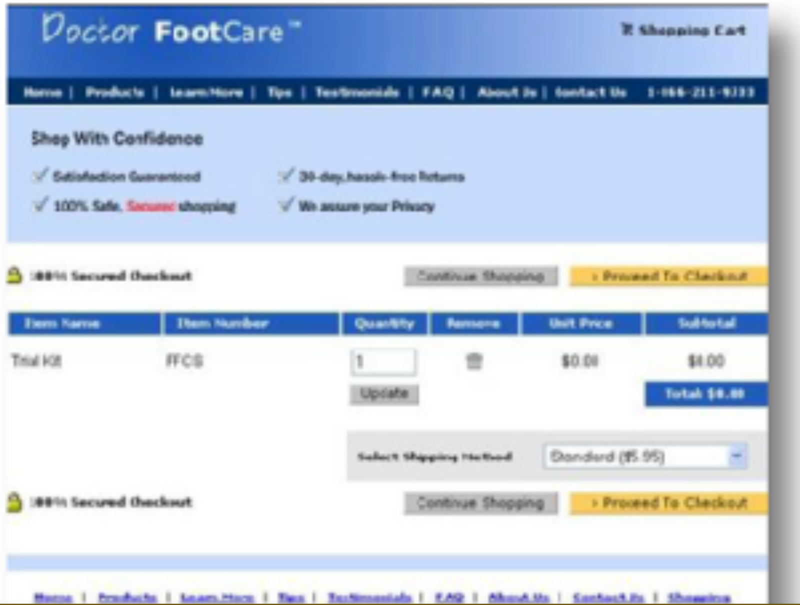
Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | Shopping Cart

B

Kumar et al. 2009

AB testing example

- Can you guess which page has a higher conversion rate and whether the difference is significant?



Doctor FootCare™

Shopping Cart

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us 1-866-211-9333

Shop With Confidence

✓ Satisfaction Guaranteed ✓ 30-day, hassle-free Returns

✓ 100% Safe, Secured shopping ✓ We assure your Privacy

100% Secured Checkout Continue Shopping › Proceed To Checkout

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	1		$0.00	$4.00

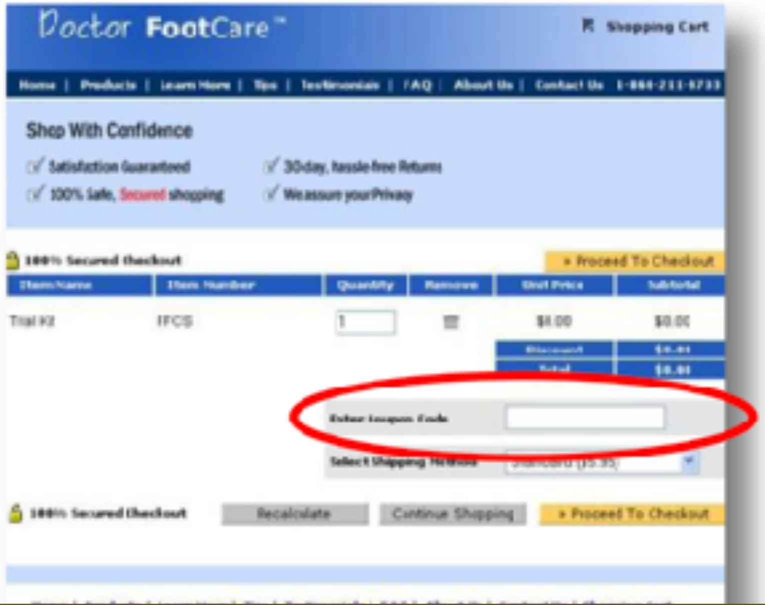
Update

Total: $0.00

Select Shipping Method Standard ($5.95)

100% Secured Checkout Continue Shopping › Proceed To Checkout

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | Shopping



Doctor FootCare™

Shopping Cart

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us 1-888-211-8733

Shop With Confidence

✓ Satisfaction Guaranteed ✓ 30-day, hassle-free Returns

✓ 100% Safe, Secured shopping ✓ We assure your Privacy

100% Secured Checkout » Proceed To Checkout

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	1		$8.00	$0.00
				Discount	$0.00
				Total	$0.00

Enter Coupon Code

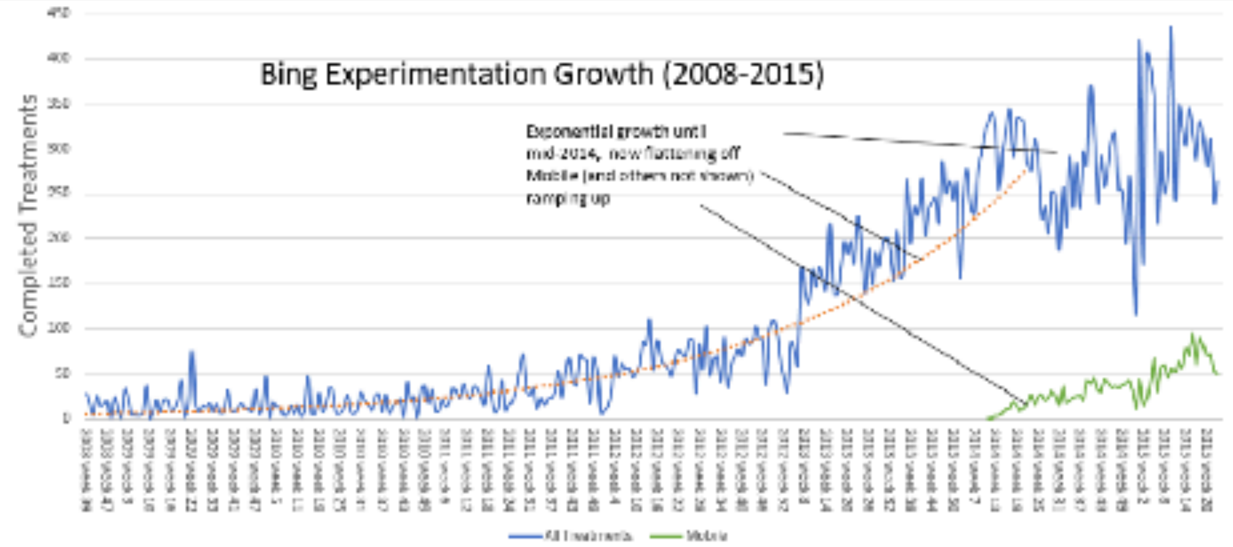
Select Shipping Method Standard ($5.95)

100% Secured Checkout Recalculate Continue Shopping » Proceed To Checkout

Using version B the site lost 90% of their revenue. Why?
“There maybe discount coupons out there that I do not have.The price may be too high…” (Kumar et al. 2009)

Experimentation at scale (Ronny Kohavi, Microsoft Research)

- ~300 experiment treatments are completed at Bing every week
- Each variant is exposed to between 100K and 10M users
- 90% of eligible users are in experiments (10% are a global holdout changed once a year)
- There is no single Bing. Since a user is exposed to 15 concurrent experiments, they get one of 5^{15} = 30 billion variants



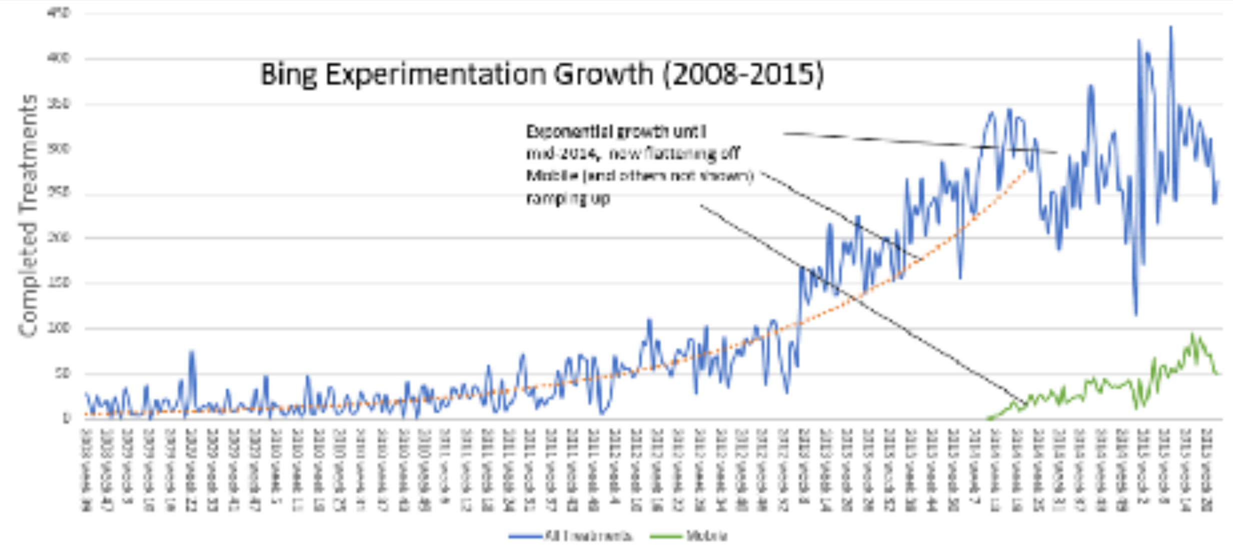
Bing Experimentation Growth (2008-2015)

Exponential growth until mid-2014, now flattening off

Mobile (and others not shown) ramping up

Experimentation at scale (Ronny Kohavi, Microsoft Research)

- ~300 experiment treatments are completed at Bing every week
- Each variant is exposed to between 100K and 10M users
- 90% of eligible users are in experiments (10% are a global holdout changed once a year)
- There is no single Bing. Since a user is exposed to 15 concurrent experiments, they get one of 5^{15} = 30 billion variants



Advantage of controlled experiments

- Controlled experiments test for causal relationships, not simply correlations
  - The gold standard in science
  - The only way to prove efficacy of drugs in FDA drug tests
- When the variants run concurrently, only two things can explain differences:
  - The “feature(s)” (A vs. B)
  - Random chance
- All other effects are the same in both the conditions
- To control for random chance, statistical tests are used to test for significance

First controlled experiment for medical purposes

- Scurvy is a disease that results from vitamin C deficiency
- Killed over 100,000 people in the 16th-18th centuries, mostly sailors
  - E.g., Lord Anson’s circumnavigation voyage from 1740 to 1744 started with 1,800 sailors and only about 200 returned; most died from scurvy
- Dr. James Lind noticed lack of scurvy in Mediterranean ships
  - Gave some sailors limes (treatment), others ate regular diet (control)
  - Experiment was so successful, British sailors are still called limeys

Lind’s experimental details

- Lind’s hypothesis was that scurvy was due to putrefaction of the body which could be helped by acids
    - The experiment was done on 12 sailors split into 6 pairs
    - Each pair got a different treatment: cider, elixir vitriol, vinegar, sea-water, nutmeg+barley water, oranges+lemon
    - The sailors given two oranges and one lemon per day and recovered
- Lind didn’t understand the reason and tried treating Scurvy with concentrated lemon juice called “rob.” But the lemon juice was concentrated by heating it, which destroyed the vitamin C.

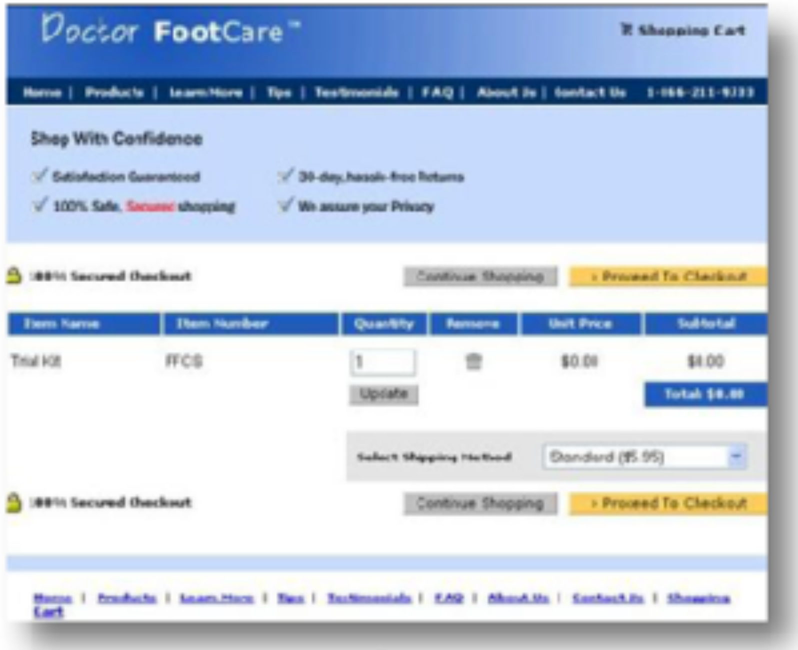
Lind’s experimental details

- Lind’s hypothesis was that scurvy was due to putrefaction of the body which could be helped by acids
  - The experiment was done on 12 sailors split into 6 pairs
  - Each pair got a different treatment: cider, elixir vitriol, vinegar, sea-water, nutmeg+barley water, oranges+lemon
  - The sailors given two oranges and one lemon per day and recovered
- Lind didn’t understand the reason and tried treating Scurvy with concentrated lemon juice called “rob.” But the lemon juice was concentrated by heating it, which destroyed the vitamin C.

Lesson: Even when you find a significant effect, the reasons are often not understood. Controlled experiments tell you which variant won, not why.

AB test example

- Sample 5000 customers for each treatment
- Measure number of shopping carts that are "converted" to purchases in conditions A and B


Doctor FootCare™
Shopping Cart
Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us 1-866-211-9333
Continue Shopping
› Proceed To Checkout
Update
Select Shipping Method
Continue Shopping
› Proceed To Checkout
Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | Shopping Cart

Shop With Confidence

✓ Satisfaction Guaranteed ✓ 30-day, hassle-free Returns
✓ 100% Safe, Secure shopping ✓ We assure your Privacy

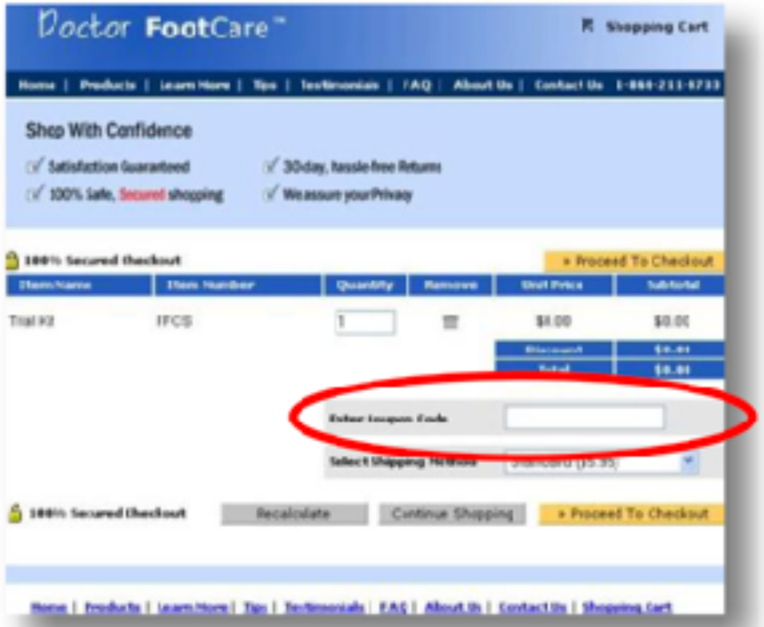
100% Secured Checkout

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	1	🗑	$0.0[illegible]	$4.00
					Total: $[illegible]

Select Shipping Method: Standard ($5.95)

100% Secured Checkout

A


Doctor FootCare™
Shopping Cart
Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us 1-888-211-9733
» Proceed To Checkout
Select Shipping Method
Recalculate
Continue Shopping
» Proceed To Checkout
Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | Shopping Cart

Shop With Confidence

✓ Satisfaction Guaranteed ✓ 30-day, hassle-free Returns
✓ 100% Safe, Secured shopping ✓ We assure your Privacy

100% Secured Checkout

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	1	🗑	$4.00	$0.0[illegible]
				Discount	$[illegible]
				Total	$[illegible]

Enter Coupon Code:

Select Shipping Method: Standard ($5.95)

100% Secured Checkout

B

Kumar et al. 2009

AB test example

	Not converted	Converted
A	4461	539
B	4522	478

```
from scipy.stats import chi2_contingency

obs = np.array([[4461,539],[4522,478]])
chi2_contingency(obs)[:2]
```

```
(3.9405799942664563, 0.047134524006671369)
```

AB test example

	Not converted	Converted
A	4461	539
B	4522	478

```
from scipy.stats import chi2_contingency

obs = np.array([[4461,539],[4522,478]])
chi2_contingency(obs)[:2]
```

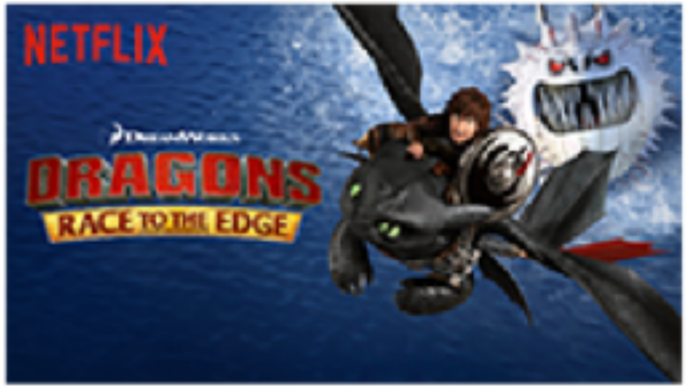
```
(3.9405799942664563, 0.047134524006671369)
```

Pvalue <0.05 so conclude effect is significant, but in this case conversions(B) < conversions(A) so impact is negative

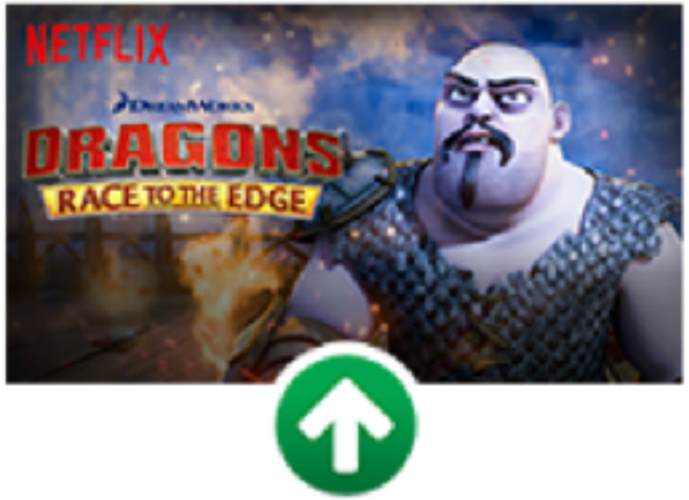
AB test example

- Sample 100,000 users for each treatment
- Measure streaming hours per user in each treatment

A



B



AB test example

	Mean	Std
A	6.730	2.5
B	6.762	2.5

```
from scipy.stats import ttest_ind

# generate pseudo data randomly
dA = np.random.normal(loc=6.730, scale=2.5, size=100000)
dB = np.random.normal(loc=6.762, scale=2.5, size=100000)

ttest_ind(dA, dB)
statistic=-2.826054368705794, pvalue=0.004712993526951O335
```

PURDUE UNIVERSITY. | College of Science

AB test example

	Mean	Std
A	6.730	2.5
B	6.762	2.5

```
from scipy.stats import ttest_ind

# generate pseudo data randomly
dA = np.random.normal(loc=6.730, scale=2.5, size=100000)
dB = np.random.normal(loc=6.762, scale=2.5, size=100000)

ttest_ind(dA, dB)
statistic=-2.826054368705794, pvalue=0.0047129935269510335
```

Pvalue <0.05 so conclude effect is significant