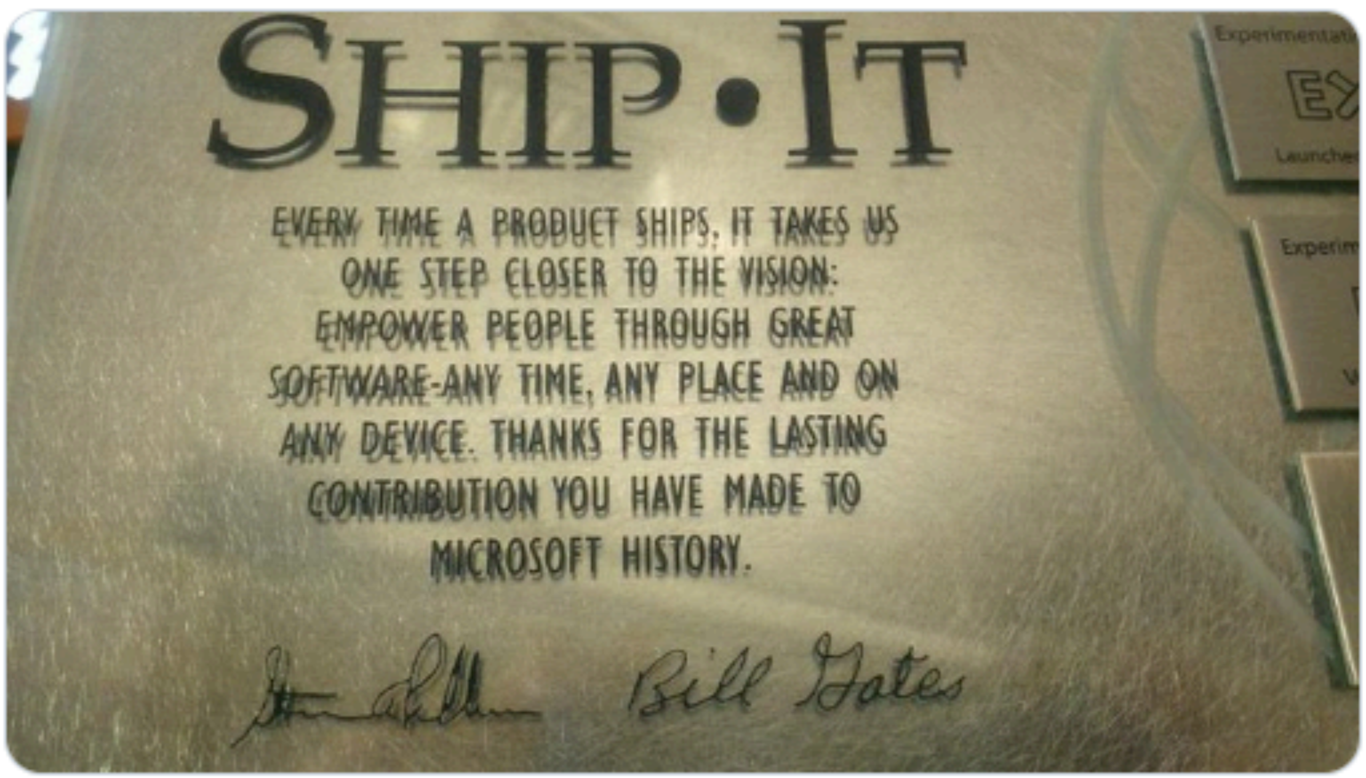
[320] Web 5: A/B Testing

Tyler Caraza-Harter

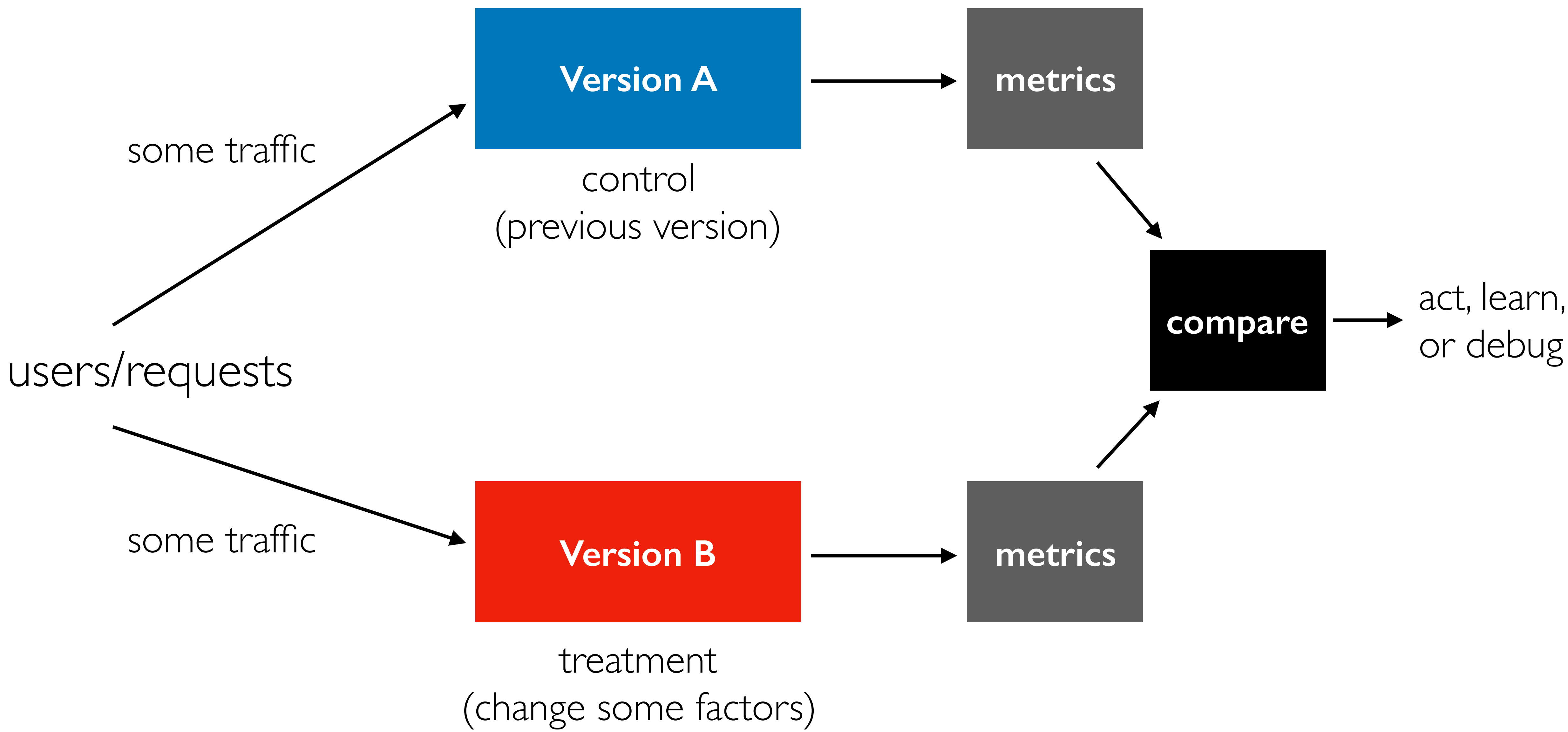
Source for Examples/Lessons

Ronny Kohavi Keynote Talk at KDD conference (Knowledge Discovery and Data Mining)
Title: Online Controlled Experiments: Lessons from Running A/B/n Tests for 12 years
Video: https://exp-platform.com/kdd2015keynotekohavi/

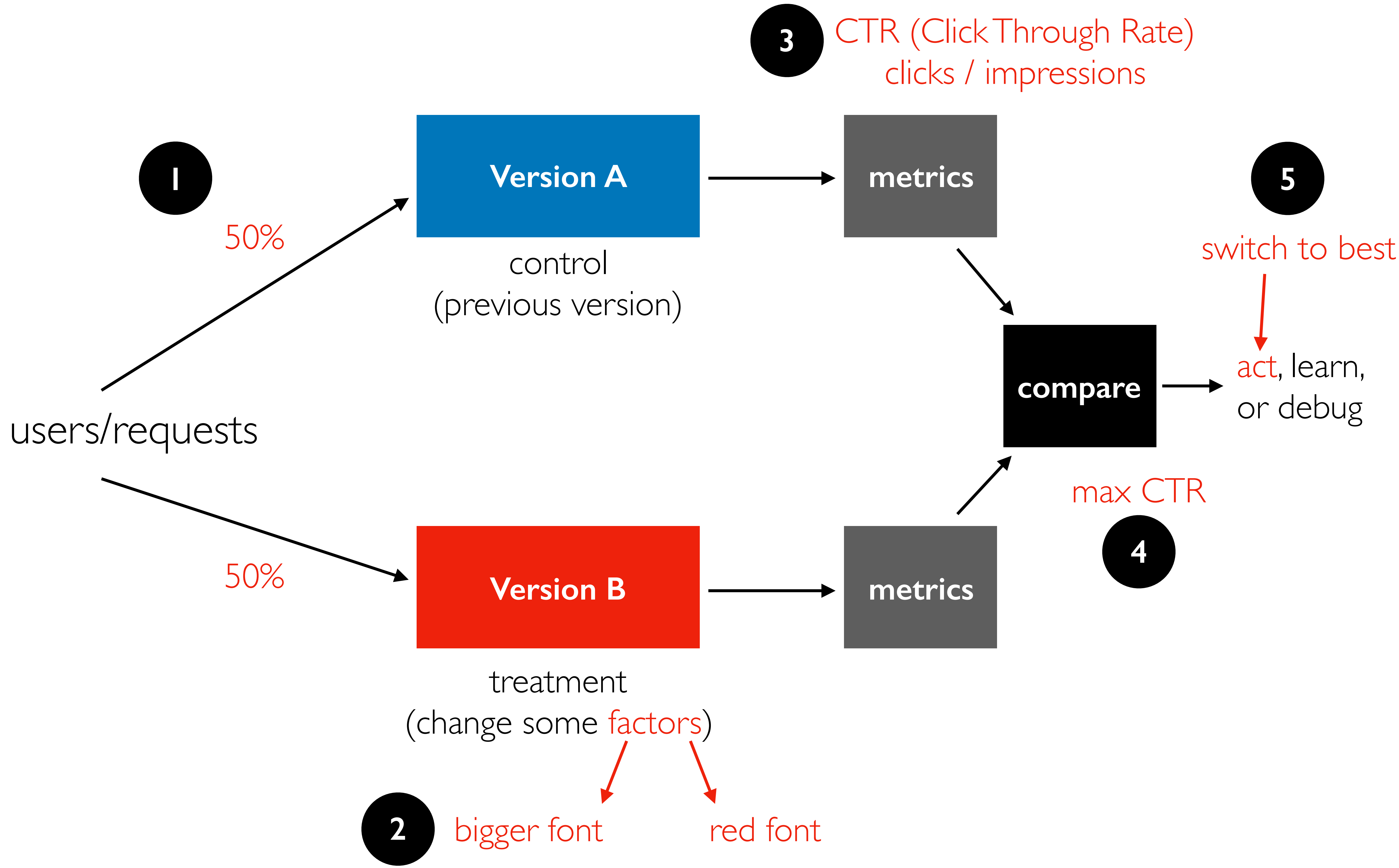
Ronny Kohavi @ronnyk · Nov 7, 2014
Microsoft stopped ship-it-awards today! With #abtesting, it's about user-impact; NOT shipping is often better!



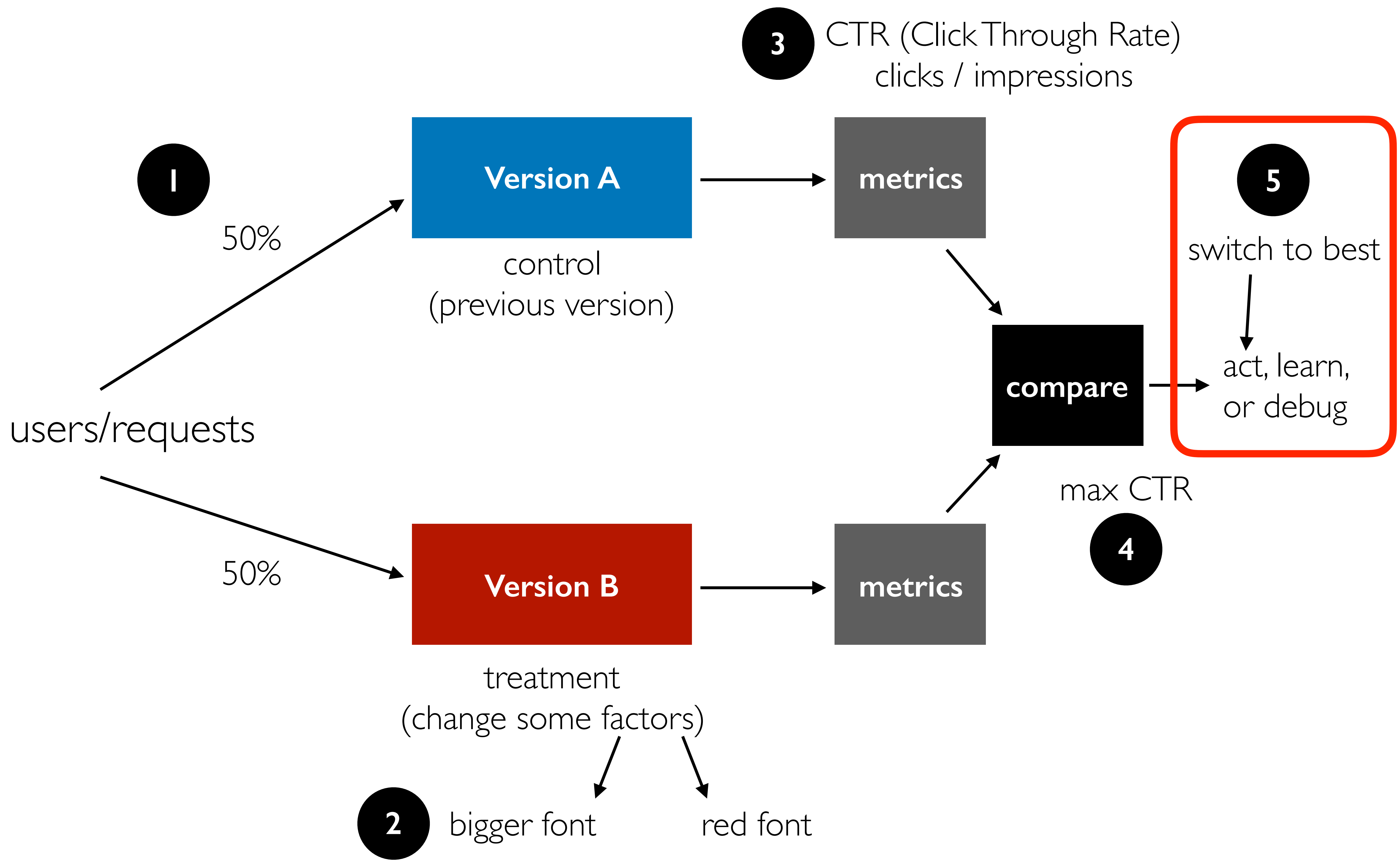
A/B Test Overview



Example 1: Link to Donation Page

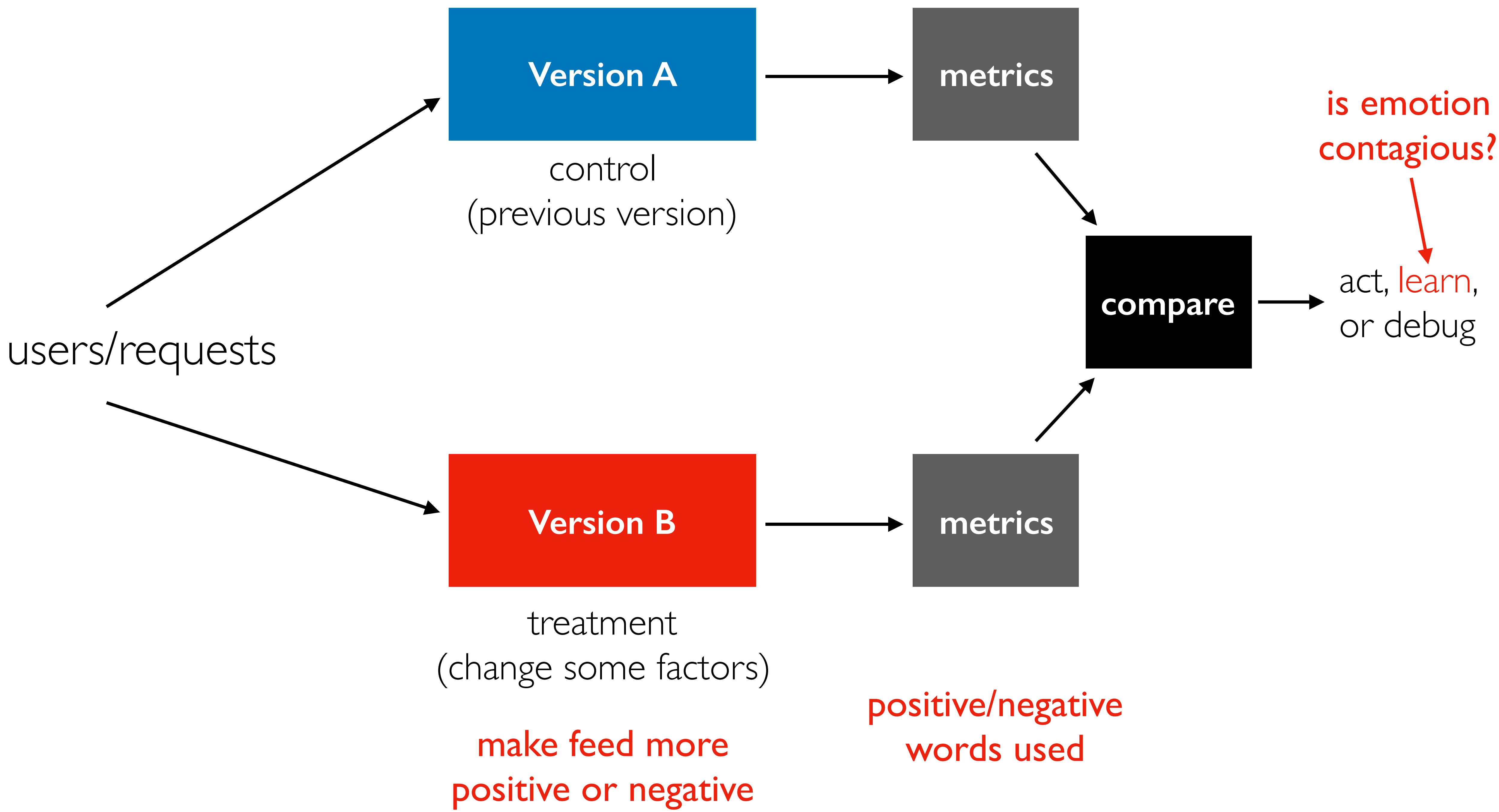


Lecture Outline



Example 2: Facebook Emotional Contagion Study

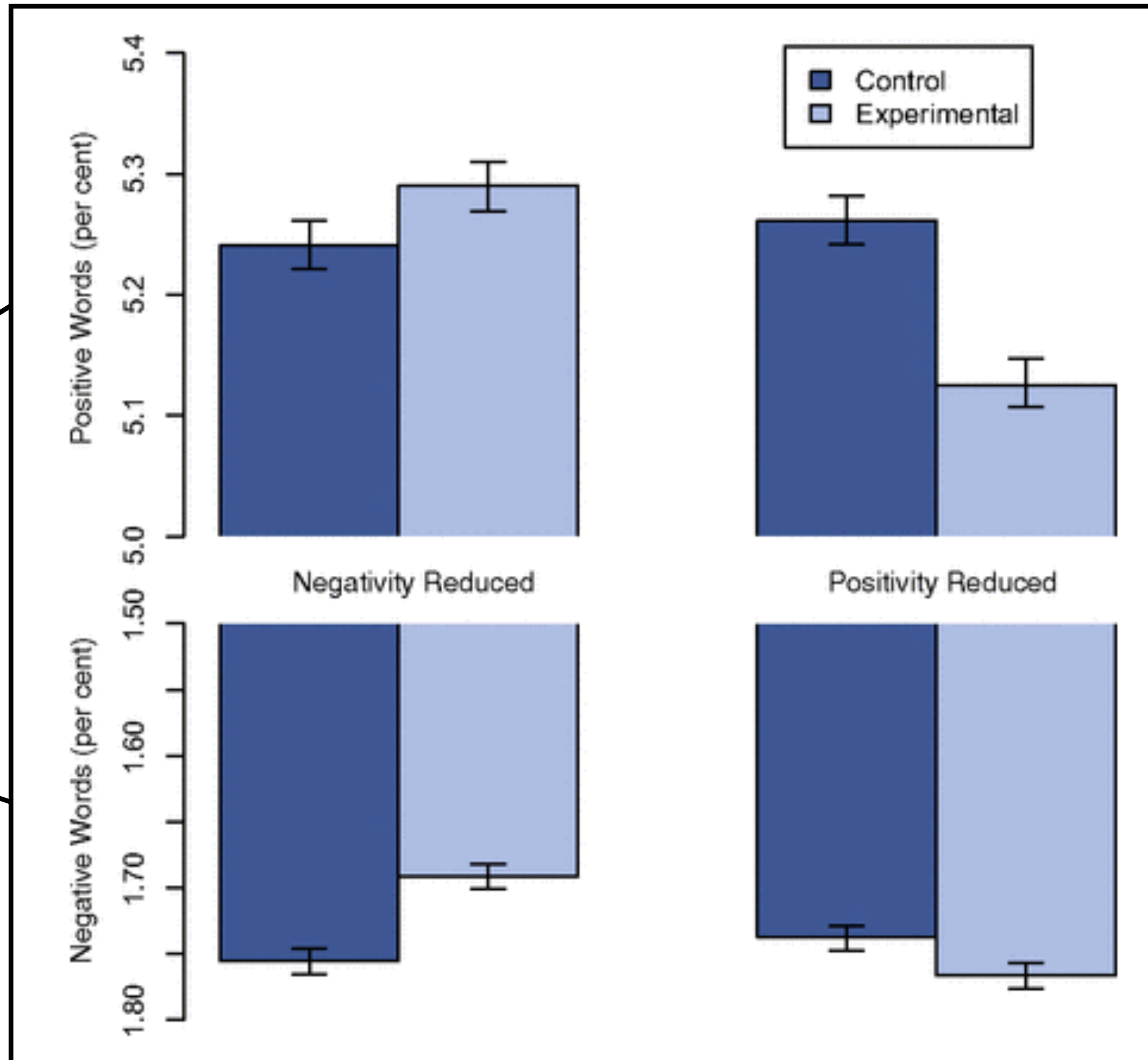
Reading: https://techcrunch.com/2014/06/29/ethics-in-a-data-driven-world/



Example 2: Facebook Emotional Contagion Study

Reading: https://techcrunch.com/2014/06/29/ethics-in-a-data-driven-world/

users/requests



ompare

act, learn, or debug

is emotion contagious?

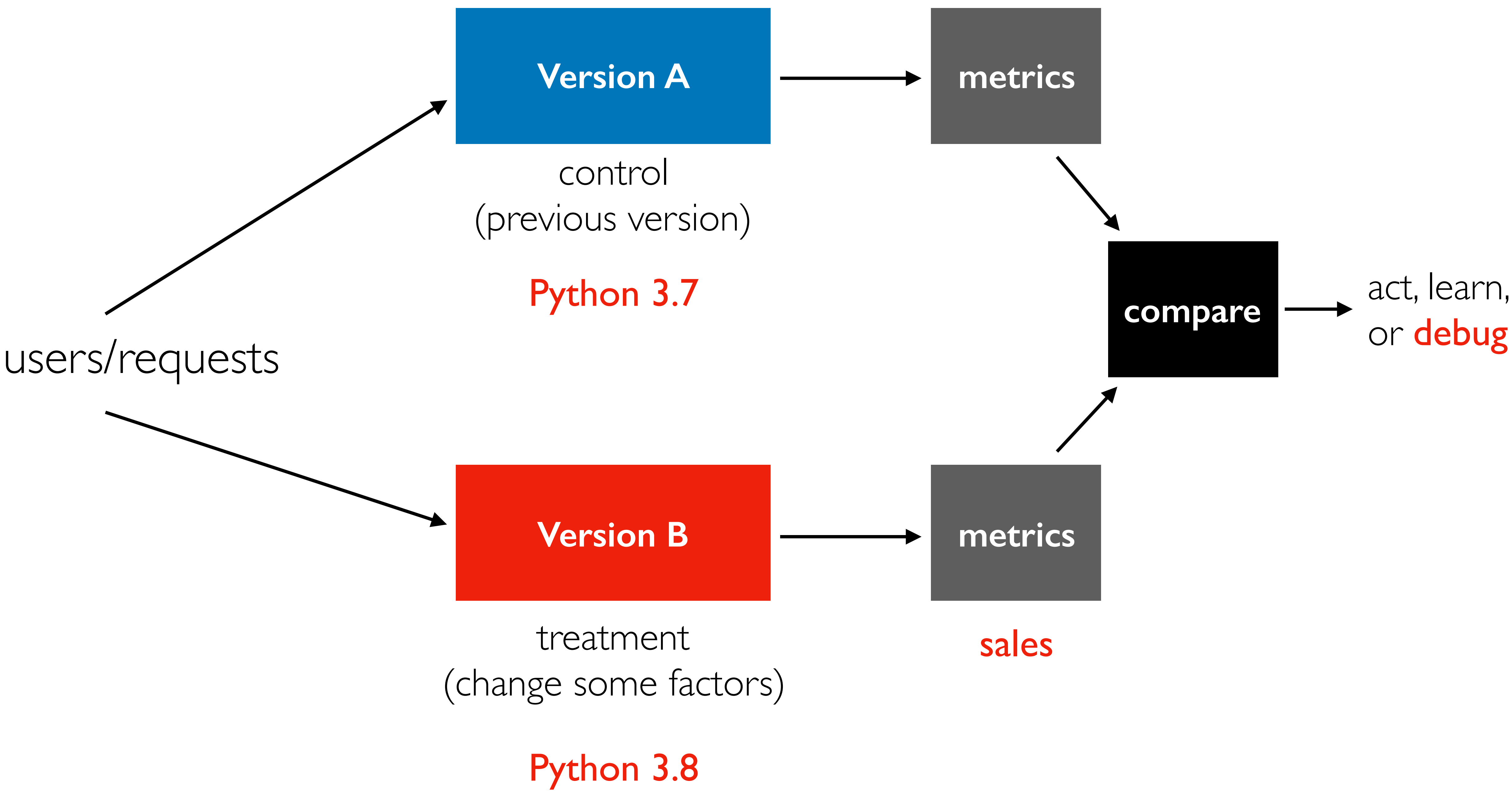
(change some factors)

make feed more positive or negative

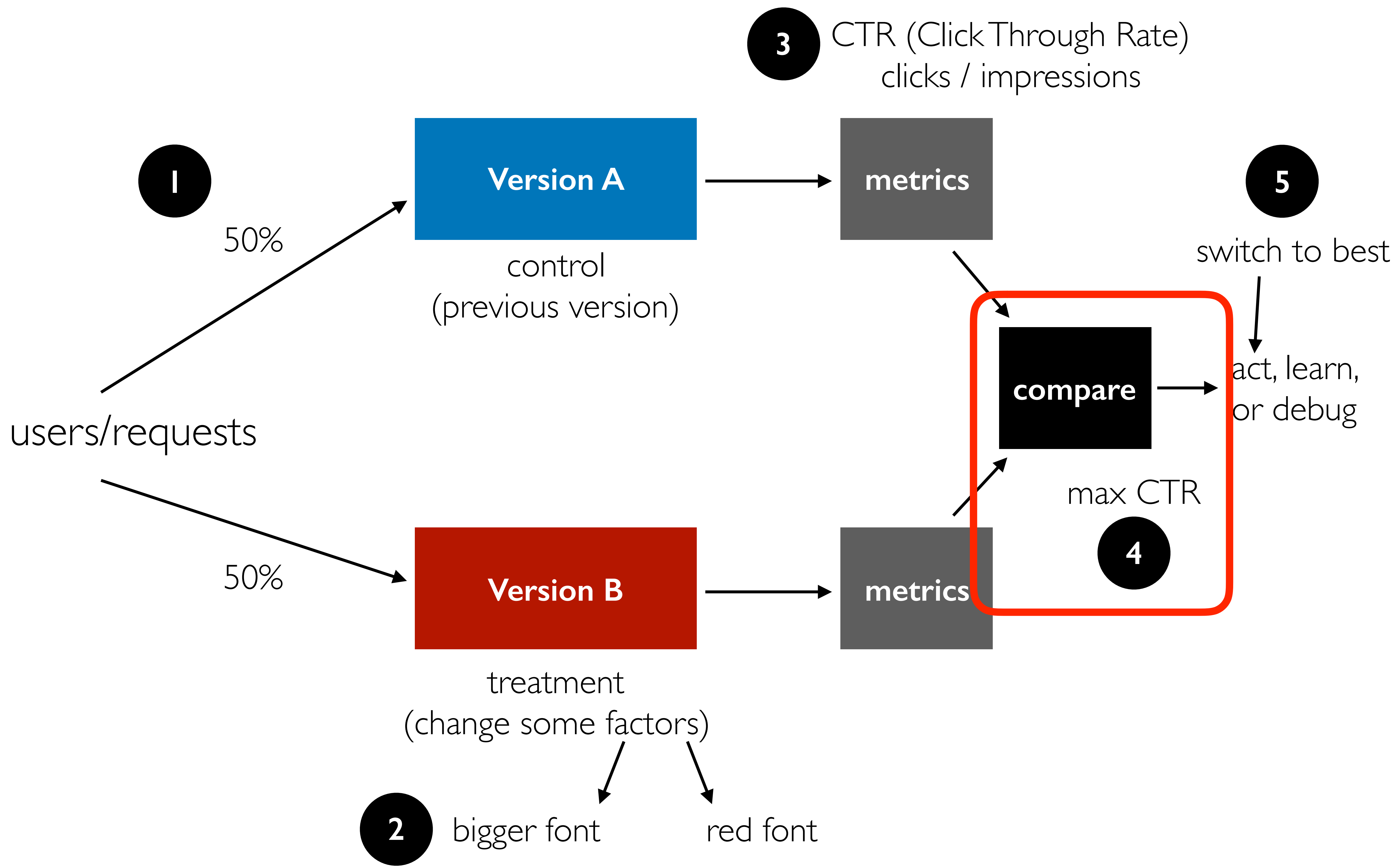
positive/negative words used

didn't need to submit to the IRB (Institutional Review Board) -- *when should it be required?*

Example 3: Update Python Version



Lecture Outline



Comparisons

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

	click	no-click
A	12	68
B	6	14

df: contingency table

how many B **impressions** were there?
what was B's **CTR**?

Comparisons

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

	click	no-click
A	12	68
B	6	14

df: contingency table

how many B **impressions** were there? 20
what was B's **CTR**? 6/20 = 30%

Comparisons

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

	click	no-click
A	12	68
B	6	14

df: contingency table

```
df["click"] / (df["click"] + df["no-click"])
```

```
A    0.15
B    0.30
dtype: float64
```

is the improvement noise?

Comparisons

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

	click	no-click
A	12	68
B	6	14

```
df["click"] / (df["click"] + df["no-click"])
```

```
A     0.15
B     0.30
dtype: float64
```

df: contingency table

pip3 install scipy

```
import scipy.stats as stats
_, pvalue = stats.fisher_exact(df)
pvalue
```

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html

```
0.18864434784714 97
```

Comparisons

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

p-value is probability of seeing a difference this extreme (or more) if both ratios were generated by the same underlying model (the one most likely to generate this)

	click	no-click
A	12	68
B	6	14

df: contingency table

"significant" means p-value is less than some threshold (e.g., 5%)

false positive means it is significant even though underlying model is same

```
import scipy.stats as stats
_, pvalue = stats.fisher_exact(df)
pvalue
```

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html

```
0.18864434784714 97
```

Comparisons

out of 200 neutral changes, how many will falsely show up as significant if we set our p-value threshold to 5%?

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

p-value is probability of seeing a difference this extreme (or more) if both ratios were generated by the same underlying model (the one most likely to generate this)

	click	no-click
A	12	68
B	6	14

df: contingency table

"significant" means p-value is less than some threshold (e.g., 5%)

false positive means it is significant even though underlying model is same

```
import scipy.stats as stats
_, pvalue = stats.fisher_exact(df)
pvalue
```

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html

```
0.18864434784714 97
```

Comparisons

out of 200 neutral changes, how many will falsely show up as significant if we set our p-value threshold to 5%?

10

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

p-value is probability of seeing a difference this extreme (or more) if both ratios were generated by the same underlying model (the one most likely to generate this)

	click	no-click
A	12	68
B	6	14

df: contingency table

"significant" means p-value is less than some threshold (e.g., 5%)

false positive means it is significant even though underlying model is same

```
import scipy.stats as stats
_, pvalue = stats.fisher_exact(df)
pvalue
```

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html

```
0.18864434784714 97
```

Comparisons

out of 200 neutral changes, how many will falsely show up as significant if we set our p-value threshold to 5%?

10

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

occasionally run A/A tests

	click	no-click
A	12	68
B	6	14

df: contingency table

```
import scipy.stats as stats
_, pvalue = stats.fisher_exact(df)
pvalue
```

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html

```
0.18864434784714 97
```

Comparisons

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

	click	no-click
A	12	68
B	6	14

df: contingency table

3 outcomes, based on CTRs and significance

- A is significantly better
- B is significantly better
- *neither wins*

what to do?

- ideas???

Comparisons

Example Metric: **CTR** (Click-Through Rate)

CTR = clicks / impressions

"Impression" means user saw it

	click	no-click
A	12	68
B	6	14

df: contingency table

3 outcomes, based on CTRs and significance

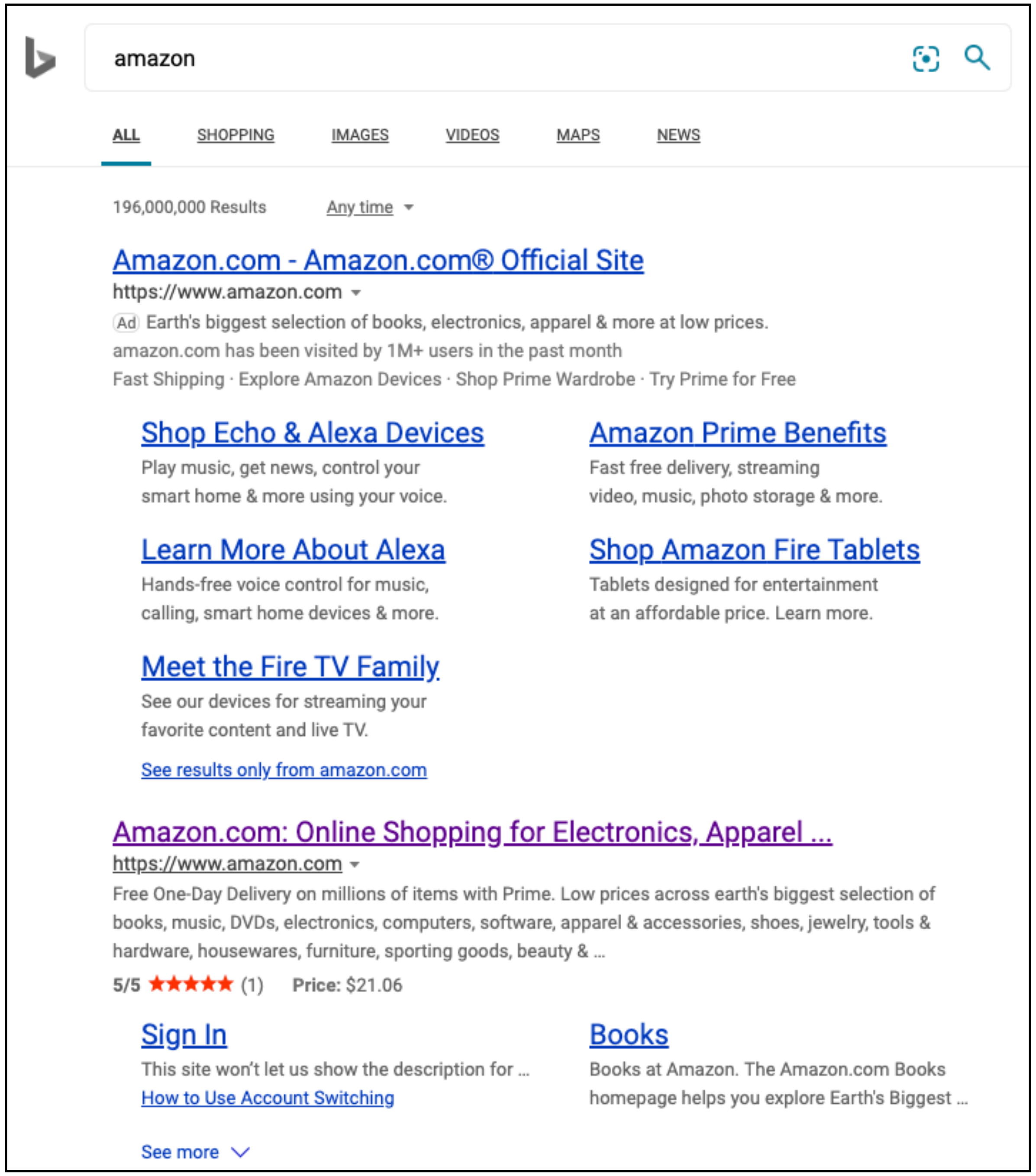
- A is significantly better
- B is significantly better
- *neither wins*

what to do?

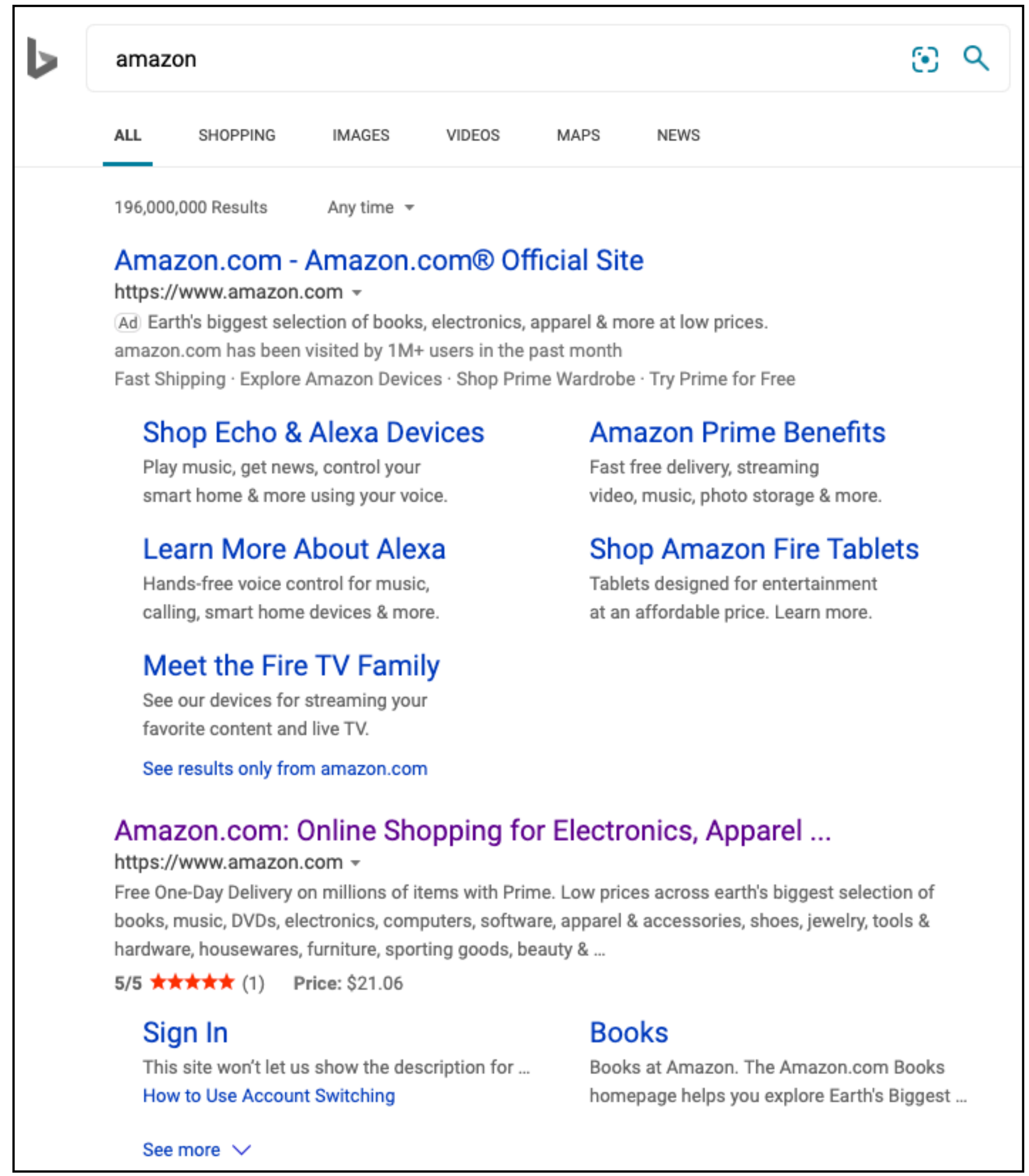
- collect more data
- ignore significance, just look at CTR (indecision may be the worst decision)
- choose previous version A (probably fewer bugs)
- choose new version B (for simplicity or other merits)

Which Version Has Higher Whole-page CTR?

Version A



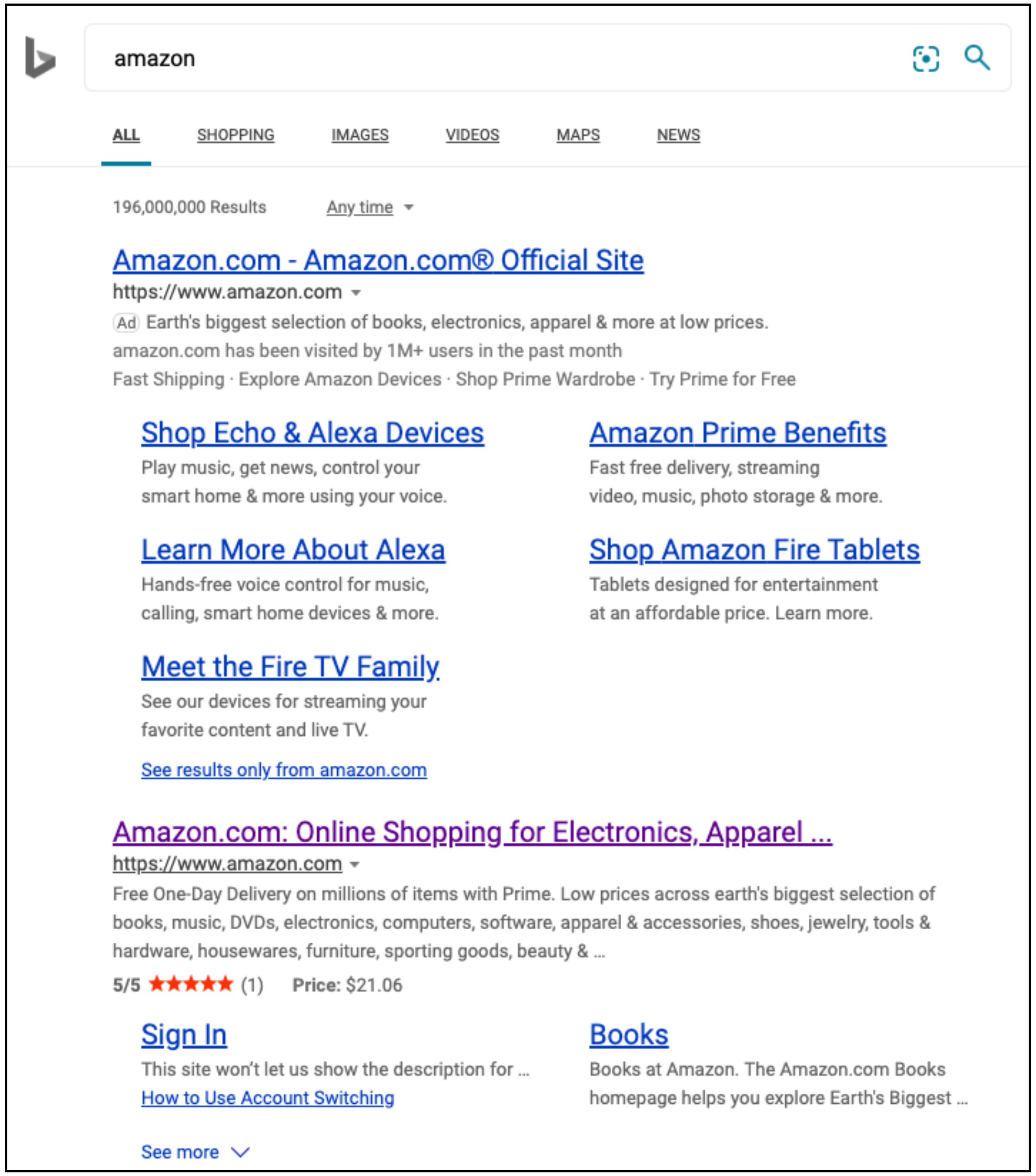
Version B



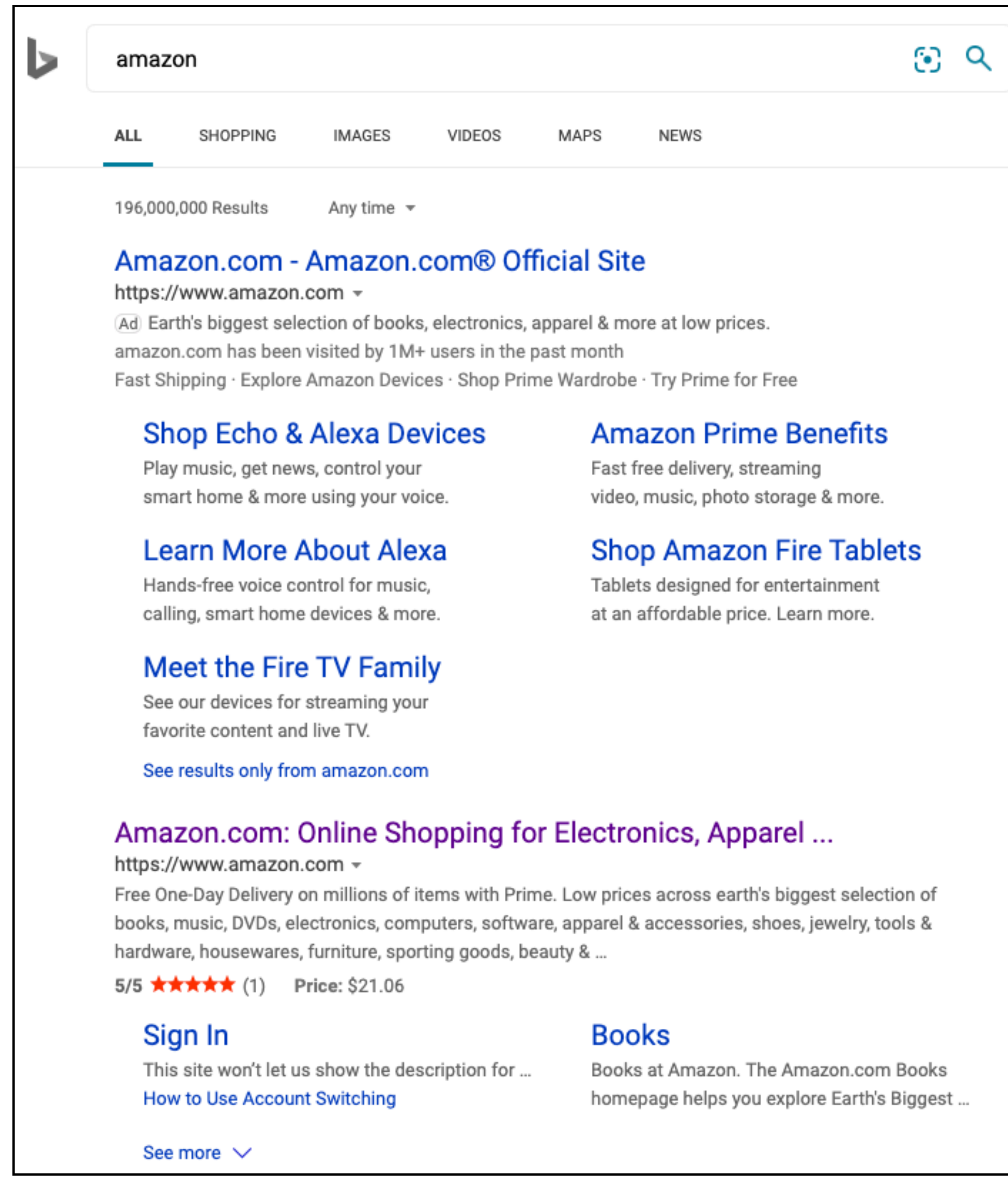
https://youtu.be/qtboCGd_hTA?t=2873

Which Version Has Higher Whole-page CTR?

Version A



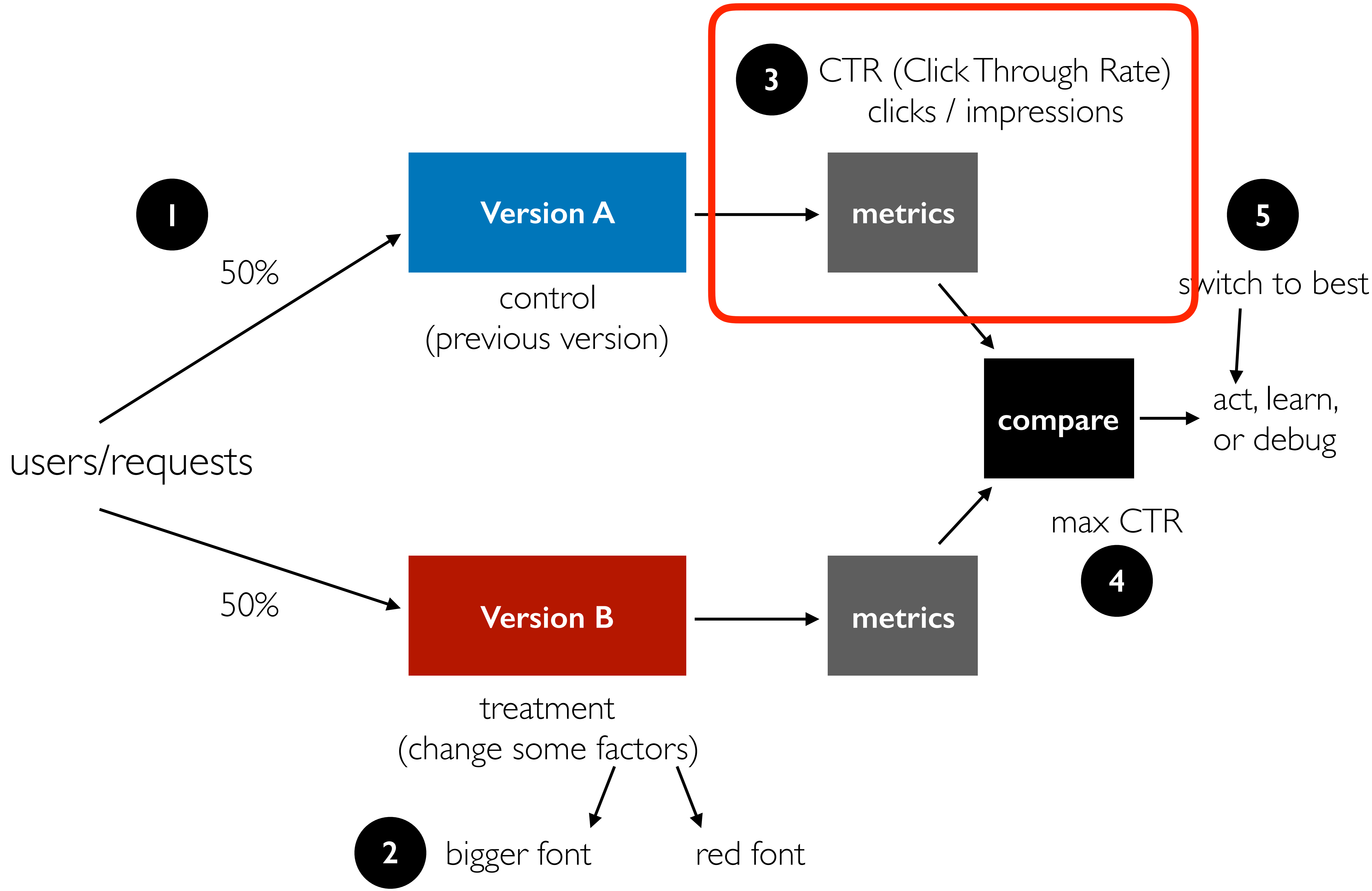
Version B



Lesson: metrics should inform humans, not directly determine decisions

https://youtu.be/qtboCGd_hTA?t=2873

Lecture Outline



Metrics

Things to measure:

- clicks -- when are they bad?

Metrics

Things to measure:

- clicks
- scroll (did they read it?)
- subscribe/unsubscribe
- other ideas? discuss with your neighbour

Metrics

Things to measure:

- clicks
- scroll (did they read it?)
- subscribe/unsubscribe
- purchases/returns
- hover (did they think about it?)
- shares
- likes/upvotes
- comments

Metrics

Things to measure:

- clicks
- scroll (did they read it?)
- subscribe/unsubscribe
- purchases/returns
- hover (did they think about it?)
- shares
- likes/upvotes
- comments

combos: Bing measures how often people click a result link and don't hit back within 30 seconds

Metrics

Things to measure:

- clicks
- scroll (did they read it?)
- subscribe/unsubscribe
- purchases/returns
- hover (did they think about it?)
- shares
- likes/upvotes
- comments

combos: Bing measures how often people click a result link and don't hit back within 30 seconds

what is the effect of B?
B is **send twice as many spammy emails**

Metrics

Things to measure:

- clicks
- scroll (did they read it?)
- subscribe/unsubscribe
- purchases/returns
- hover (did they think about it?)
- shares
- likes/upvotes
- comments

combos: Bing measures how often people click a result link and don't hit back within 30 seconds

what is the effect of B?
B is **send twice as many spammy emails**

Metrics

Things to measure:

- clicks
- scroll (did they read it?)
- subscribe/unsubscribe
- purchases/returns
- hover (did they think about it?)
- shares
- likes/upvotes
- comments

combos: Bing measures how often people click a result link and don't hit back within 30 seconds

what is the effect of B?
B is **send twice as many spammy emails**

what is the effect of B?
B is **remove price from product page link**

Metrics

Things to measure:

- clicks
- scroll (did they read it?)
- subscribe/unsubscribe
- purchases/returns
- hover (did they think about it?)
- shares
- likes/upvotes
- comments

combos: Bing measures how often people click a result link and don't hit back within 30 seconds

what is the effect of B?
B is **remove price from product page link**

Lesson: it's easy to shift clicks

what is the effect of B?
B is **send twice as many spammy emails**

Lesson: it's hard to measure long-term effects (noisy!), so use common sense

Metrics

Things to measure:

- clicks
- scroll (did they read it?)
- subscribe/unsubscribe
- purchases/returns
- hover (did they think about it?)
- shares
- likes/upvotes
- comments

combos: Bing measures how often people click a result link and don't hit back within 30 seconds

what is the effect of B?
B is **send twice as many spammy emails**

what is the effect of B?
B is **remove price from product page link**

Decide beforehand on one OEC metric: Overall Experiment Criterion

- Bing has thousands of debug metrics, but only 4 OECs.

Metrics

Things to measure:

- clicks
- scroll (did they read it?)
- subscribe/unsubscribe
- purchases/returns
- hover (did they think about it?)
- shares
- likes/upvotes
- comments

combos: Bing measures how often people click a result link and don't hit back within 30 seconds

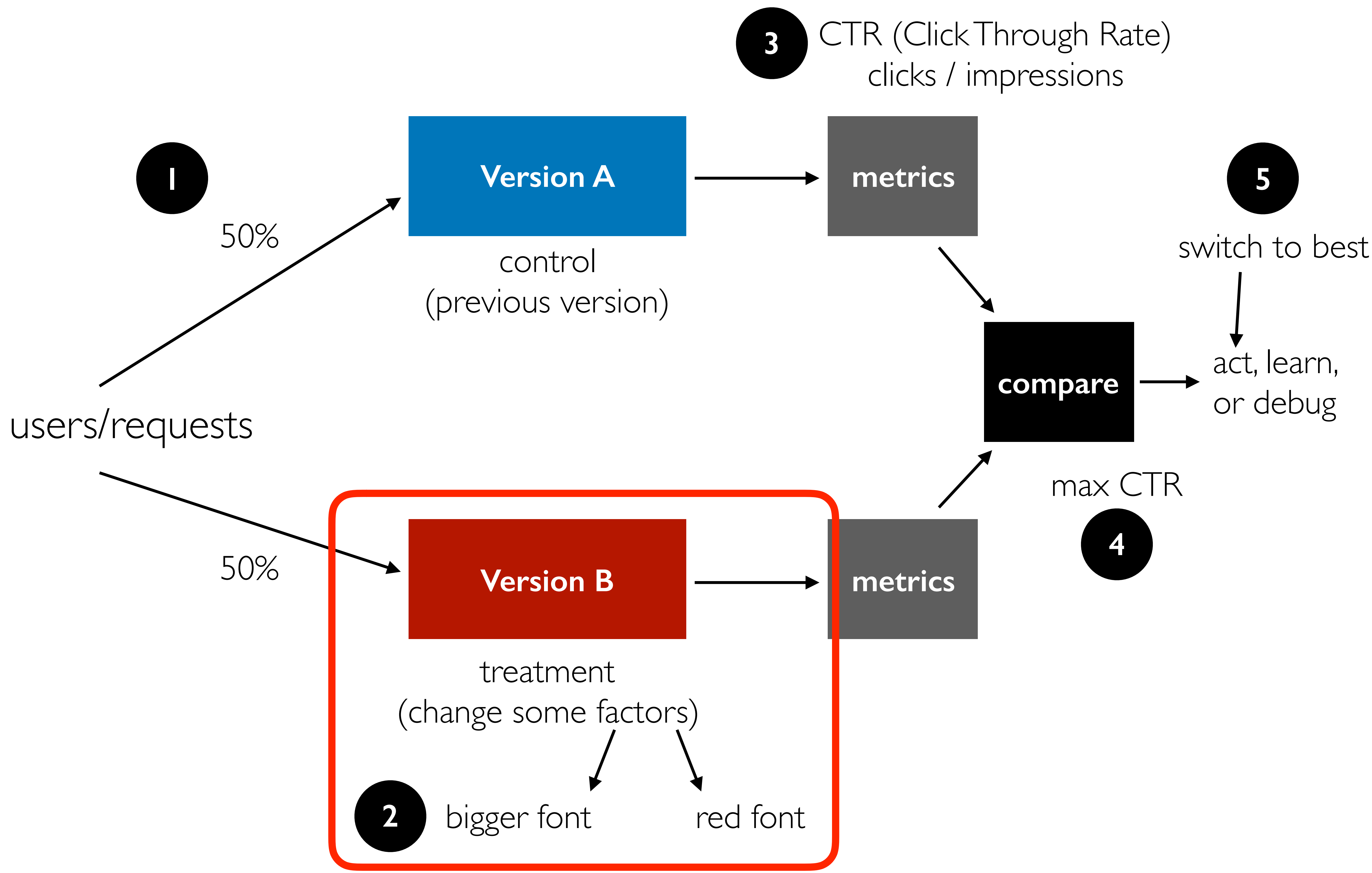
what is the effect of B?
B is **send twice as many spammy emails**

what is the effect of B?
B is **remove price from product page link**

Decide beforehand on one OEC metric: Overall Experiment Criterion

- Bing has thousands of debug metrics, but only 4 OECs. Try to consider cost as well as benefit!
- As a rule of thumb, "*if you make something bigger, more people will click on it*" ~ Ron Kovani
- Making part of the site better could hurt other parts if you have a naive OEC

Lecture Outline



Lecture Outline