Analytics Vidhya | DATAHACK SUMMIT

Hack Session

How to build an in-house platform to conduct thousands of parallel A/B experiments

Speaker

Rama Badrinath

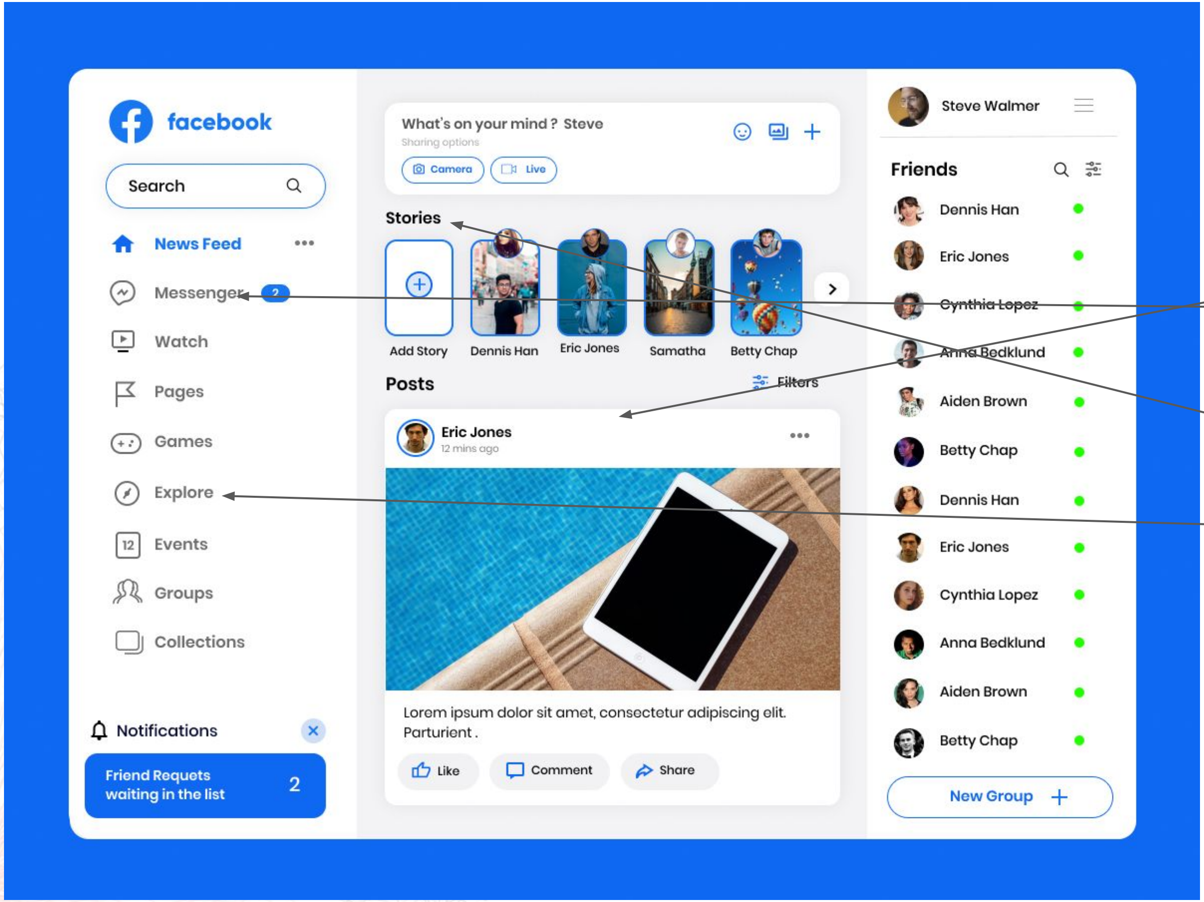
Principal Data Scientist, Meesho

Speaker

Rama Badrinath
Principal Data Scientist @ Meesho

Previously, DS @ Sharechat, Microsoft
Masters in Machine Learning from IISc
[Rama Badrinath](#)

Social media app (Ex: Facebook)



Important Widgets

1. News Feed
2. Messenger
3. Groups
4. Games
5. Stories
6. Pages
7. Explore
8. Events
9. Collections

…

Social media app metrics

Metric	Example value
Number of clicks / User	15
Number of shares / User	7
Number of downloads / User	2
Clicks / Views (CTR)	0.03
Shares / Views	0.02
Downloads / Views	0.01
Time spent in a session / User	10 mins
No. of sessions / User	3
D7 Retention	40%

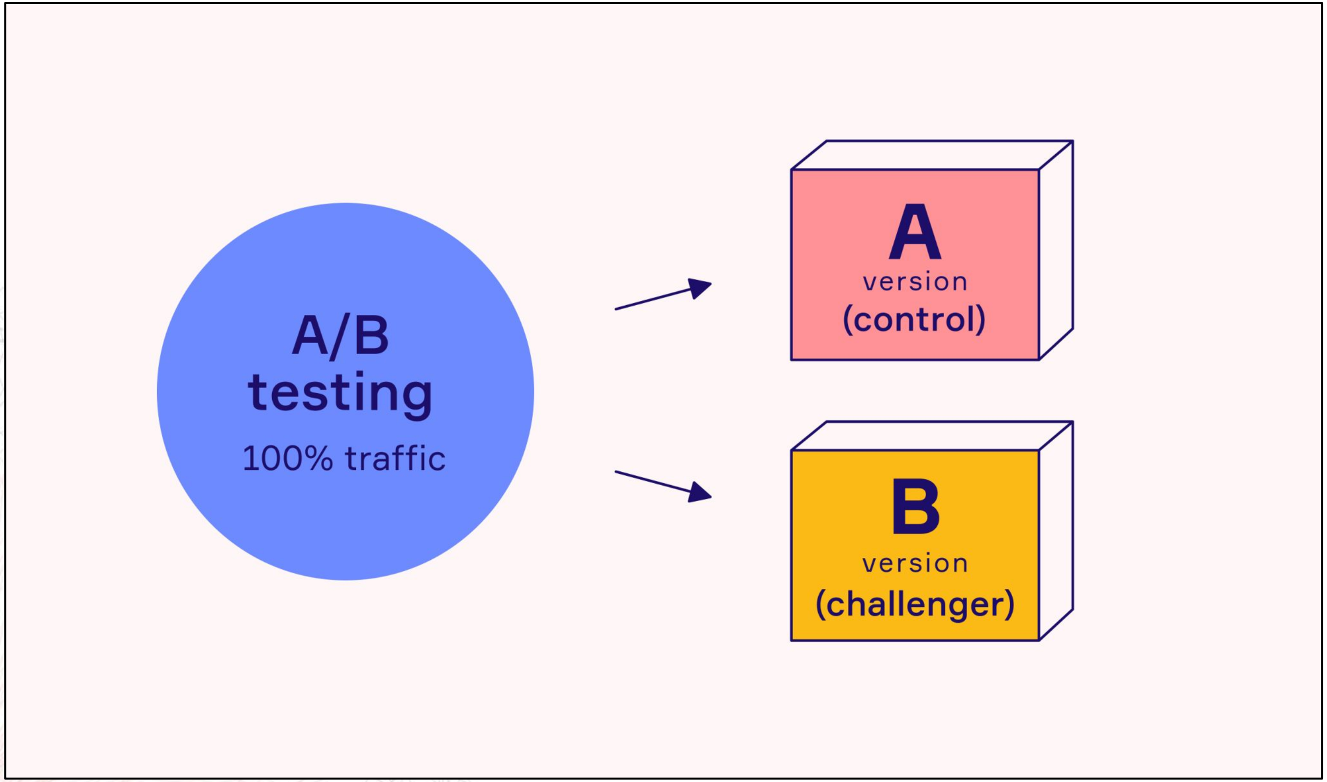
Scaling without testing is dangerous!

Metric	Metrics before scaling up	Metrics after scaling to 100% without testing	% Delta
Number of clicks / User	15	12	-20%
Number of shares / User	7	6	-14.3%
Number of downloads / User	2	1	-50%
Clicks / Views (CTR)	0.03	0.02	-33.33%
Time spent in a session / User	10 mins	9 mins	10%
No. of sessions / User	3	2	-33.33%
D7 Retention	40%	36%	-10%

Marketing team spends crores of $$ to improve DAU by 10% over a period time, but we have brought down D7 retention by 10% in 1 day!!

What is A/B testing?

Launch it for a small audience & look at the metrics!



Agenda

01	What should be the primary metric in an A/B test?	• Examples of bad vs good metrics
02	Introduction to hypothesis testing	• Null hypothesis testing • One-tailed vs Two-tailed tests • Type-1 error vs Type-2 error
03	Sample size calculation	• Sample size calculation • Minimum Detectable Effect
04	Audience creation to conduct 1000s of parallel A/B tests	• Issues with mutually exclusive users • Hashing to achieve random user set • User homogenization
05	Statistical Significance calculation	• Statistical significance calculation • T-statistic to P-value conversion • Scale up decision based on P-value
06	Overall architecture	• Overall architecture • Default A/B configs that are common in industry

1. Primary metric

Analytics Vidhya | DATAHACK SUMMIT

- Agree early on what you are optimizing
- The primary metric should correlate well with the long-term goals
- Generate many other metrics to understand why the primary metric moved / didn't move

Social media (Ex: Facebook)

Bad metric: Number of clicks
Reason: Clickbait images can easily improve clicks.

Good metric: D7 retention (**Proxy:** Time spent on the app)
Reason: Users are returning to the app after 7 days. Shows users are genuinely interested in the app.

E-commerce (Ex: Amazon)

Bad metric: Number of orders
Reason: Easy to move this metric by placing 1000 orders a day & cancelling the next day.

Good metrics: Net orders & CLTV
Reason: Users are not cancelling & doing repeated purchases.

Search Ads (Ex: Google)

Bad metric: Ads revenue alone
Reason: Easy to move this metric by increasing the number of ad slots, but users will eventually churn.

Good metric: Ads revenue
Check metric: Sessions / User
Reason: Users are not churning & still generating revenue.

2. A/B testing & Statistical tests

Is treatment really better than control?

- What's the proof that the increase in time spent is because of the change we introduced?
- It could be because of pure chance also !!

Time spent in Control for different users (in mins)	Time spent in Variant for different users (in mins)
9	21
13	17
10	18
7	25
16	16
6	11
8	10
5	12
Mean = 9.25 mins	**Mean = 16.25 mins**

2.1 Hypothesis testing

- **Null Hypothesis**
  - Time spent in treatment = Time spent in control
  - The two groups are from the same population

$H_0: \mu_1 = \mu_2$

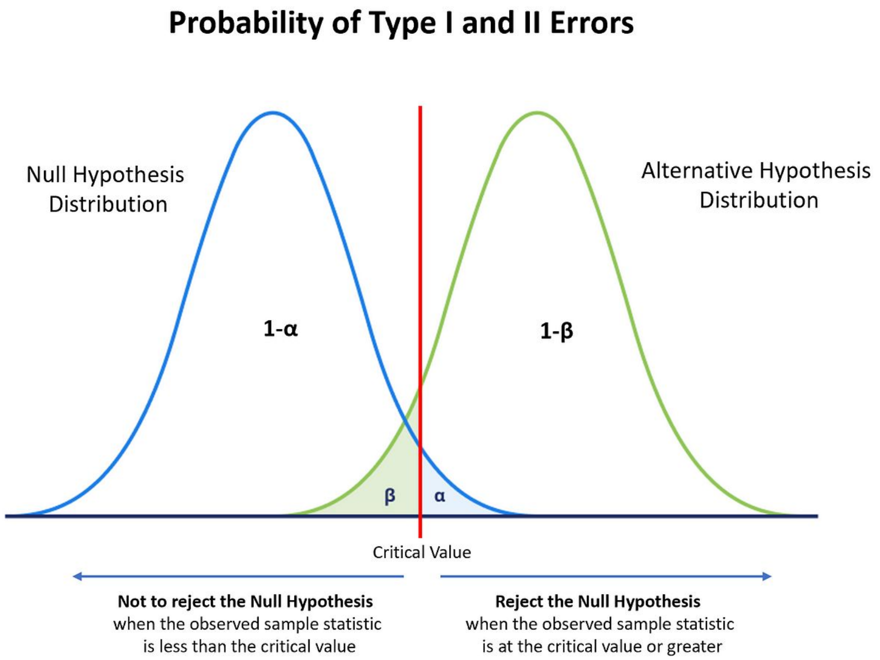
- **Alternate Hypothesis in case of Two-tailed test**
  - Time spent in treatment != Time spent in control
  - The two groups are from different populations

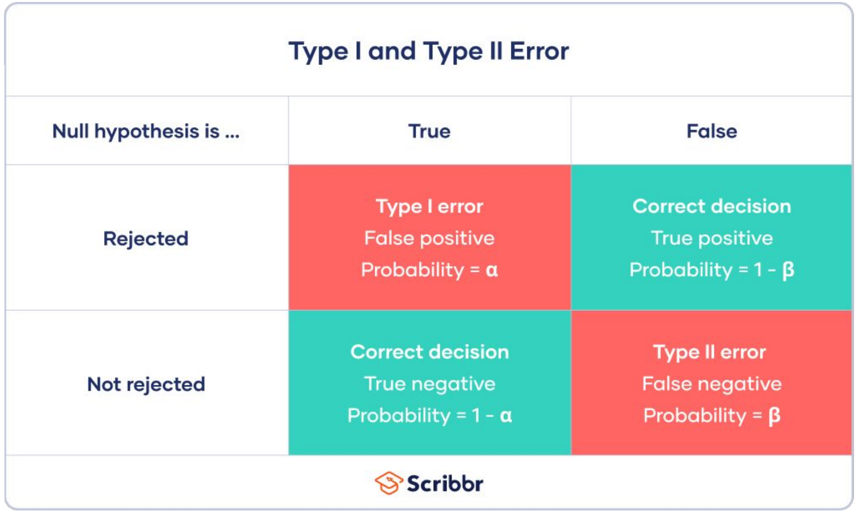
$H_0: \mu_1 \neq \mu_2$

- **Alternate Hypothesis in case of One-tailed test**
  - Time spent in treatment > Time spent in control
  - The two groups are from different populations and mean of treatment > mean of control

$H_0: \mu_1 > \mu_2$

2.2 Type-1 error (α) vs Type-2 error (β)





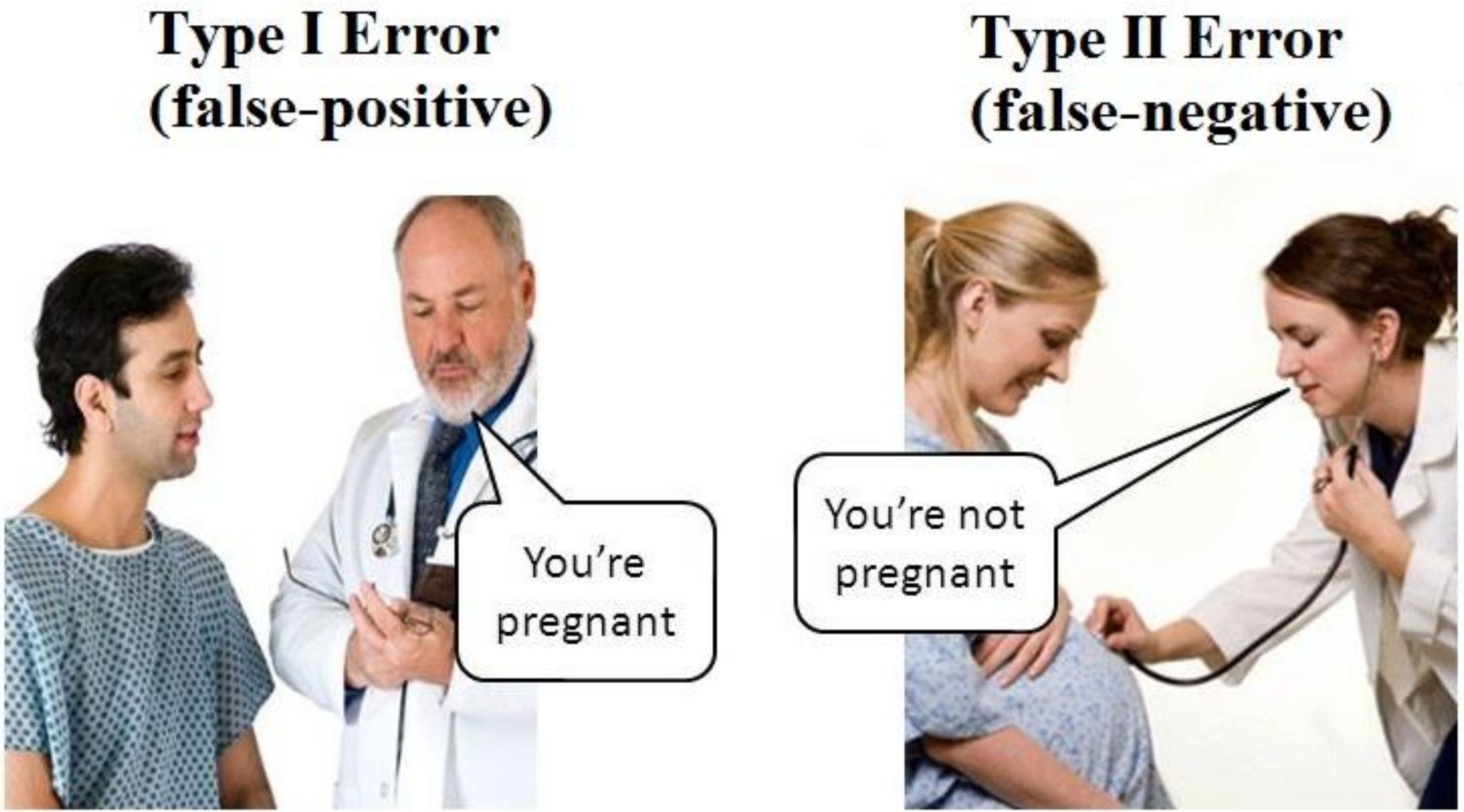
Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

Scribbr

Error	Default
Type-1	5%
Type-2	20%

2.3 Type-1 error ($\underline{\alpha}$) vs Type-2 error (β)



3. Sample size

Two-tailed test

$$N = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2}{\delta^2}$$

$$Z_{1-\alpha/2} = 1.96, Z_{1-\beta} = 0.84$$

$$N = \frac{15.7 * \sigma^2}{\delta^2}$$

One-tailed test

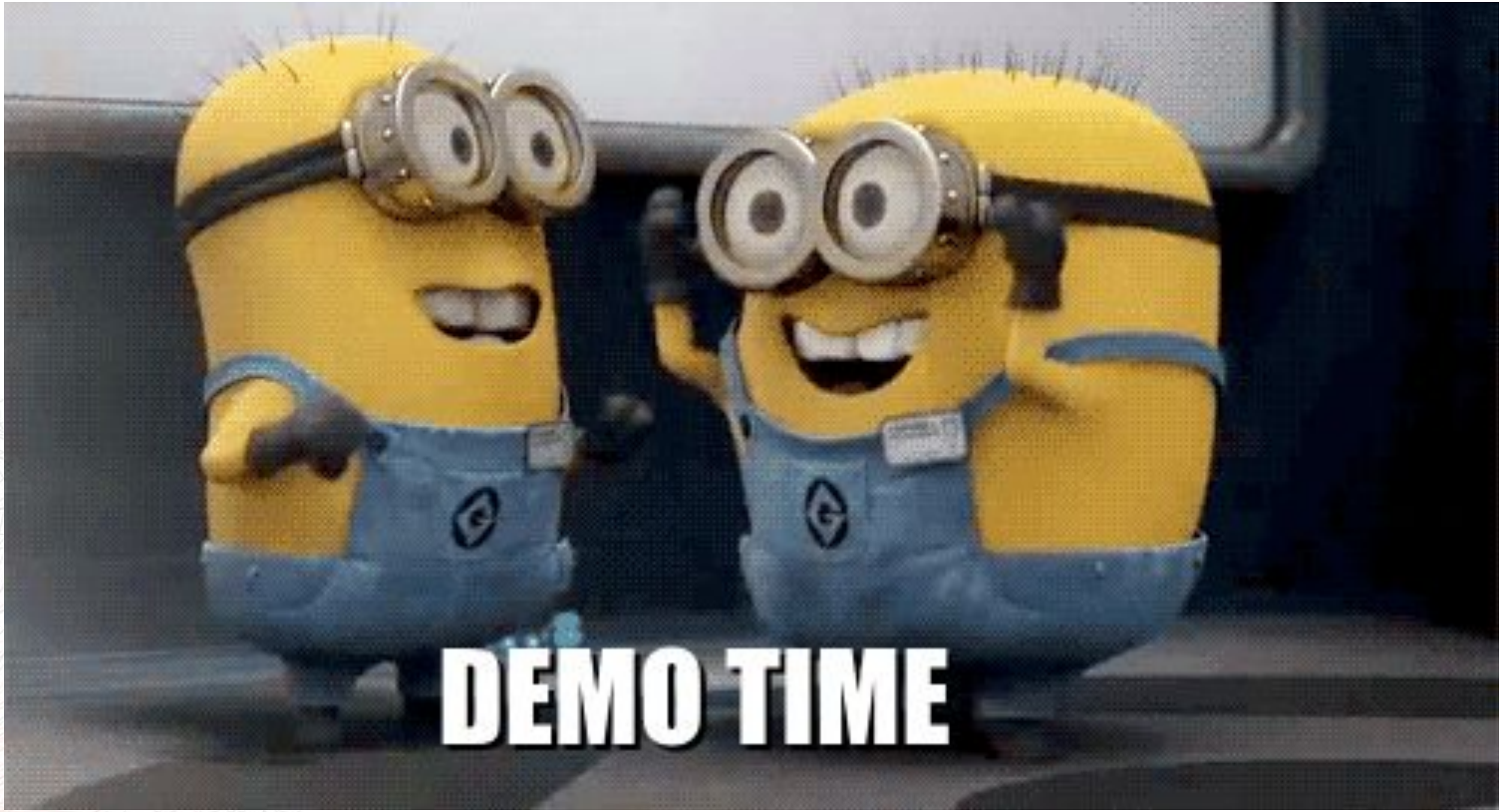
$$N = \frac{2(Z_{1-\alpha} + Z_{1-\beta})^2 \sigma^2}{\delta^2}$$

$$Z_{1-\alpha} = 1.645, Z_{1-\beta} = 0.84$$

$$N = \frac{12.35 * \sigma^2}{\delta^2}$$

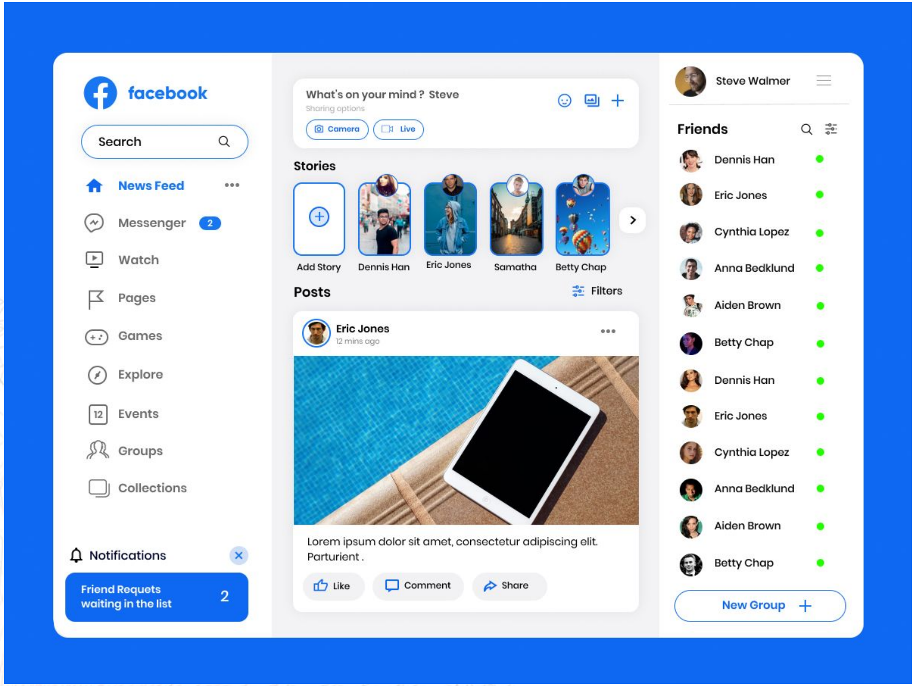
N	Number of users required per variant
σ	Variance of the metric
δ	Mean of the metric * Minimum Detectable Effect (MDE)
MDE	Minimum detectable effect
Assumption	Alpha = 0.05, Beta = 0.2

3.1 Sample size demo



4. How to assign users to experiments?

Daily Active Users (DAU) = 30 million



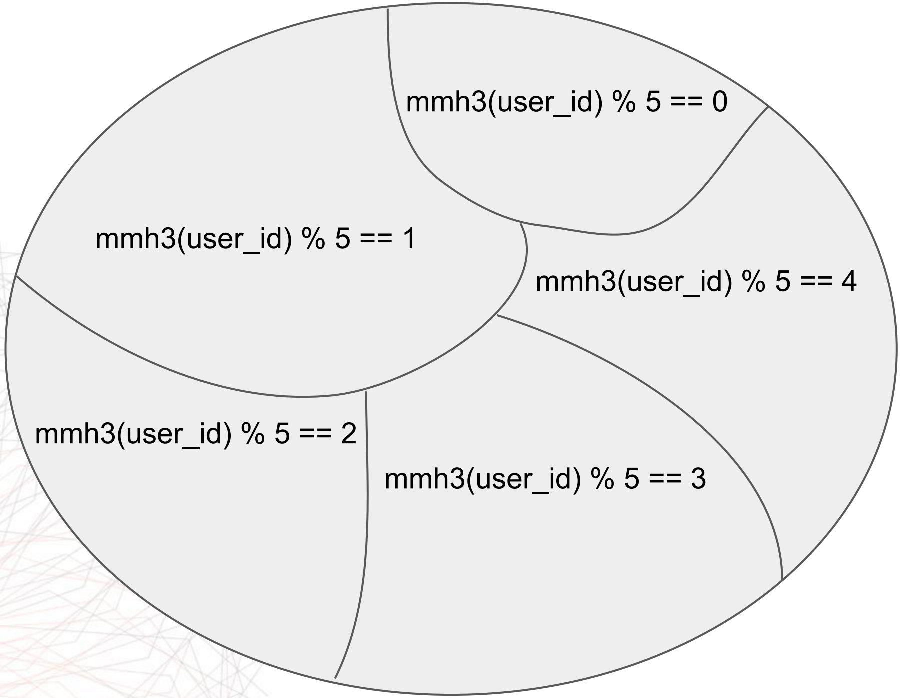
Widget	Number of experiments	Sample size for all experiments assuming mutual exclusion
News feed	5	12 million
Messenger	5	10 million
Groups	5	10 million
Games	5	15 million
Pages	5	13 million
Total	**25**	**60 million**

Total users needed for experimentation = 60 million, but DAU = 30 million only

~~Mutual Exclusion~~

Overlapping audiences based on user ids can screw up results

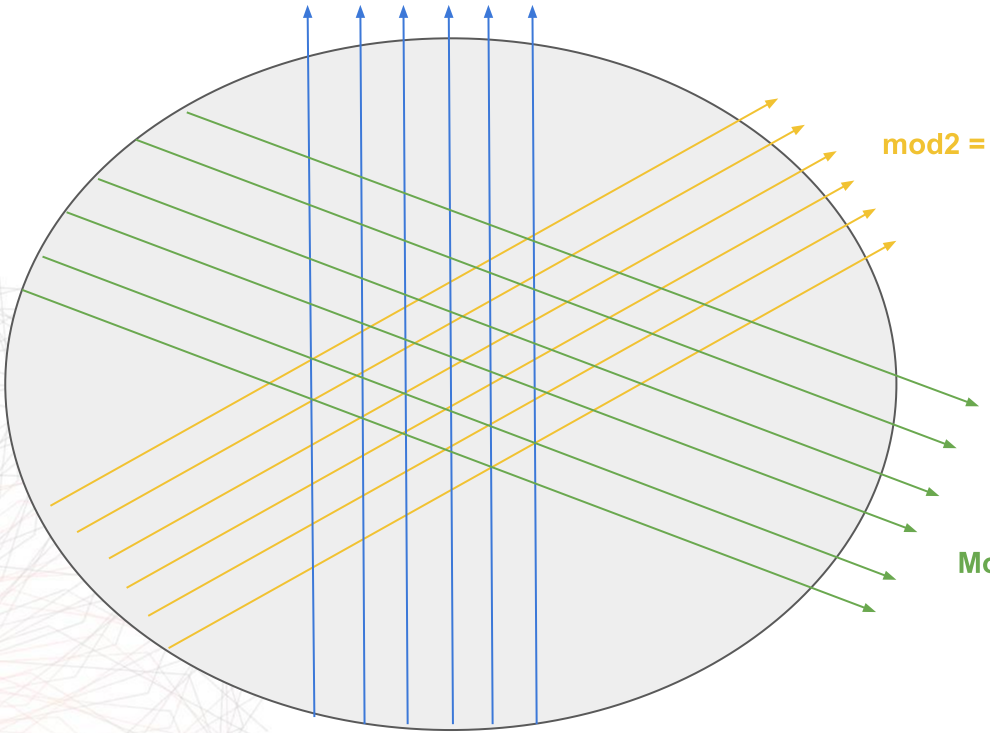
4.1 Mutual exclusion example



4.2 Replicating user traffic

Add salt before computing hash!

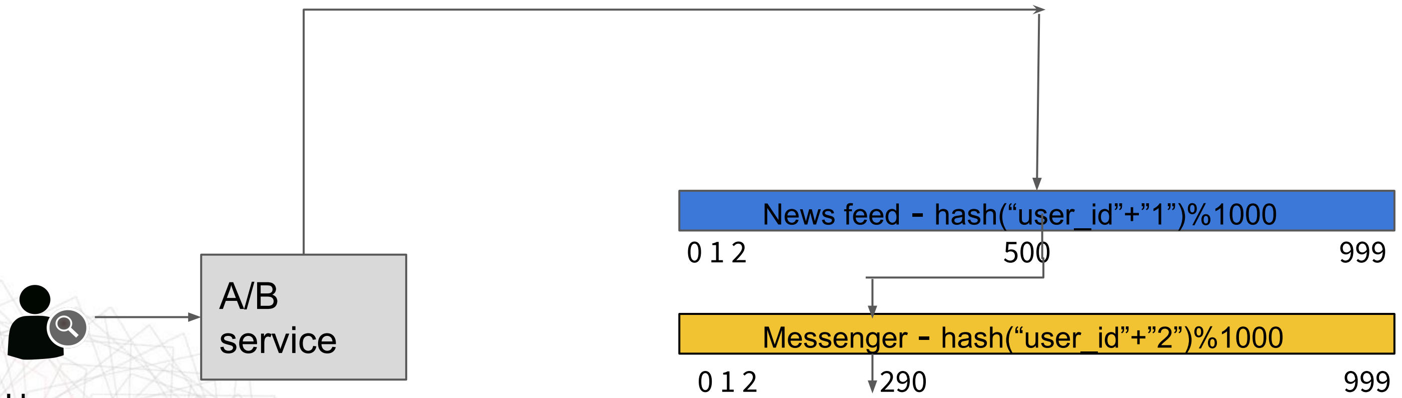
News feed
mod1 = murmur3_hash("user_id" + "1")%1000

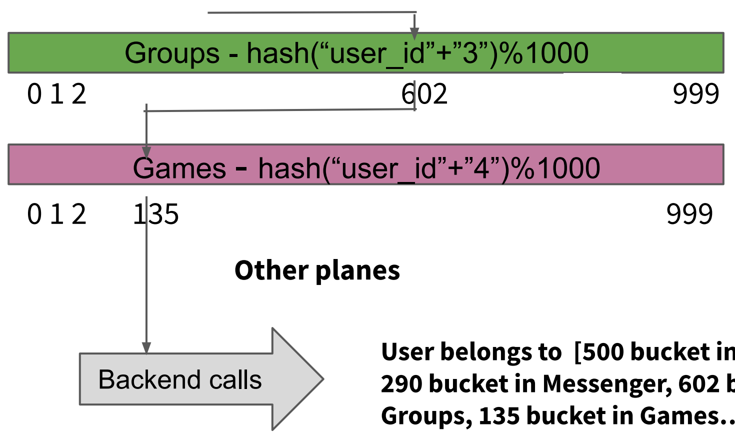


Messenger
mod2 = murmur3_hash("user_id" + "2")%1000

Groups
Mod3 = murmur3_hash("user_id" + "3")%1000

4.3 Overlapping audience creation





100% traffic

100% traffic

100% traffic

100% traffic

1. **A user belongs to a particular bucket in each plane**
2. **Each bucket in a plane will be in either control or some experiment.**

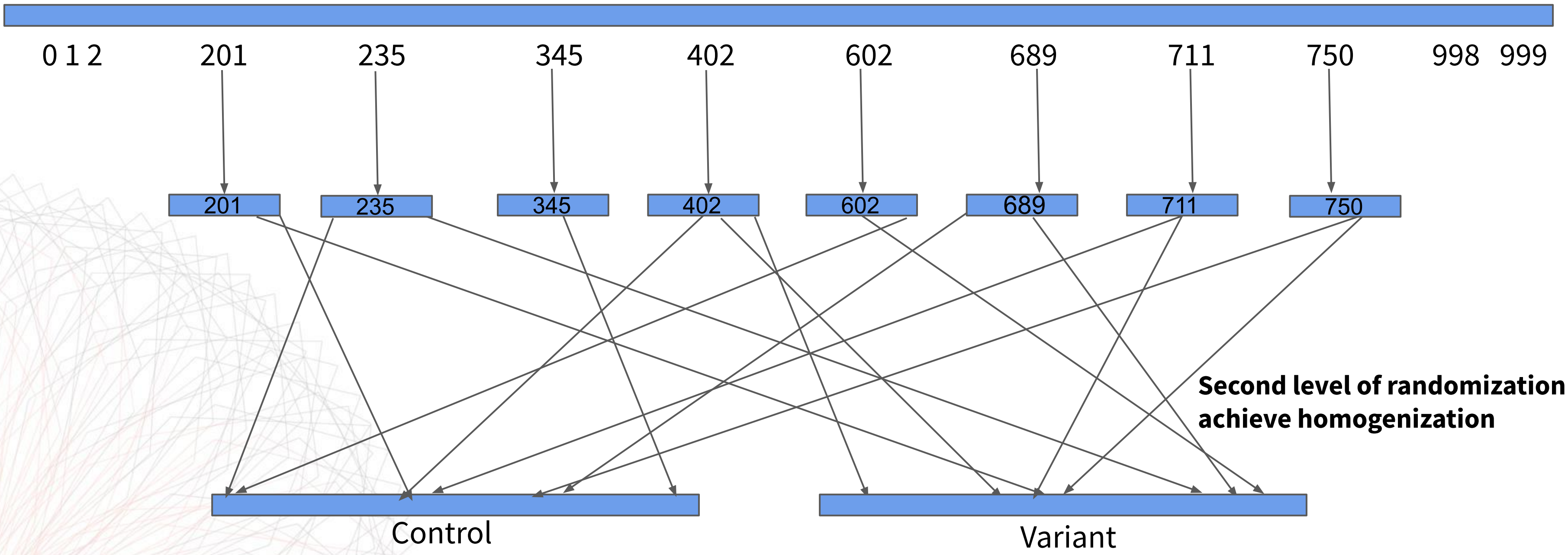
User belongs to [500 bucket in News feed, 290 bucket in Messenger, 602 bucket in Groups, 135 bucket in Games…]

4.4 Avoiding carry-over effect

Analytics Vidhya | DATAHACK SUMMIT

User bucket = mmh3(user id + “plane_salt”)%1000

First level of randomization to find the experiment bucket



Second level of randomization to achieve homogenization

Control = mmh3(user id + “plane_salt” + “exp”)%2 == 0

Variant = mmh3(user id + “plane_salt” + “exp”)%2 == 1

4.5 Audience creation summary

1. How to split the traffic so that users can be assigned to experiments dynamically?

User ID ⟷ Experiment

Solution: Murmur3_hash(user_id)%1000. Every mod value could correspond to an experiment

2. How to conduct parallel A/B experiments across different widgets?

Control	Exp 1	Exp 2	Exp 3

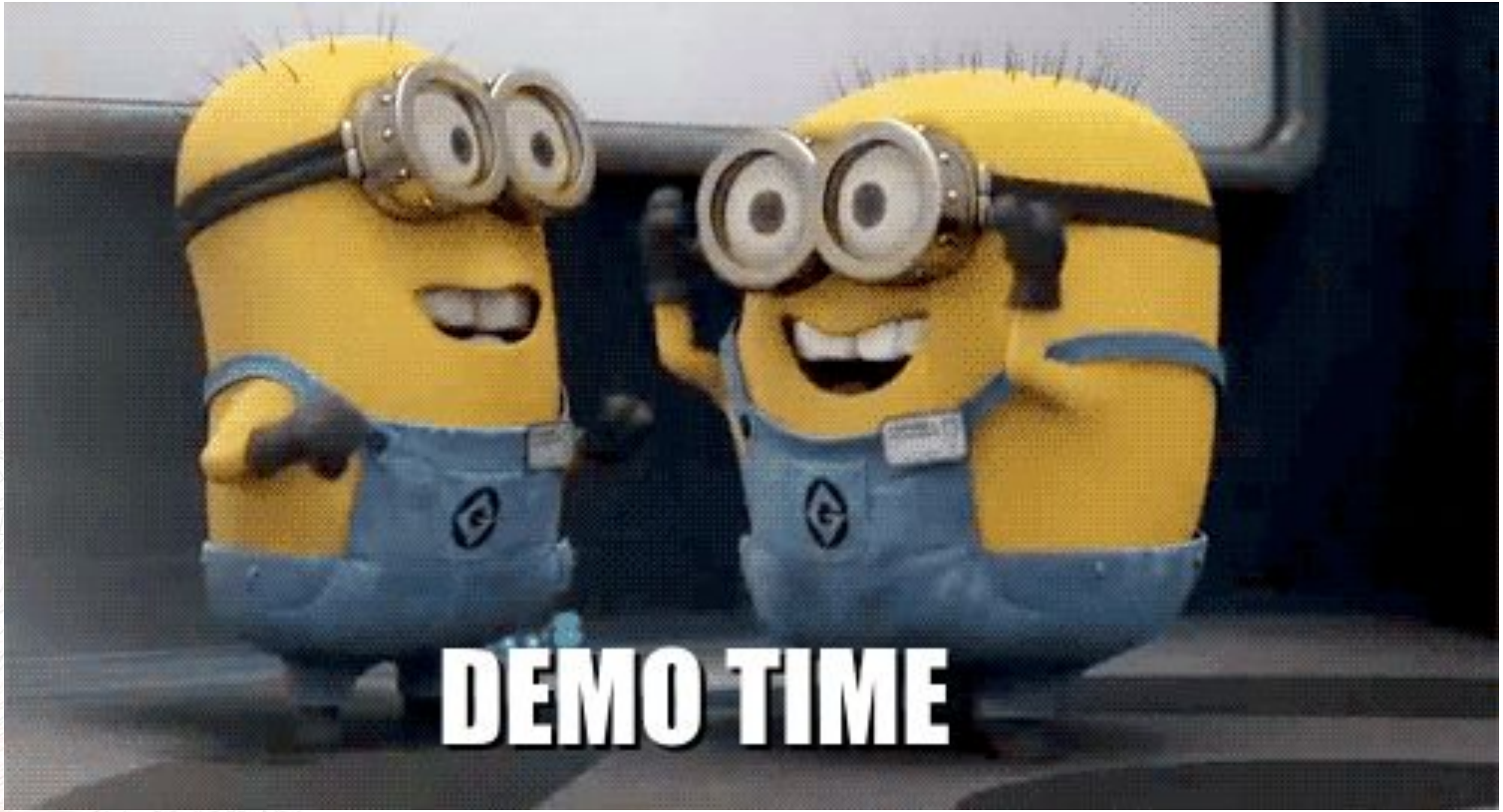
Solution: Each widget forms a plane & users can belong to multiple experiments across planes

3. How to assign users to variants so that they are homogeneous?

Control	Exp1 Variant	Exp1 Control	Exp2	Exp3

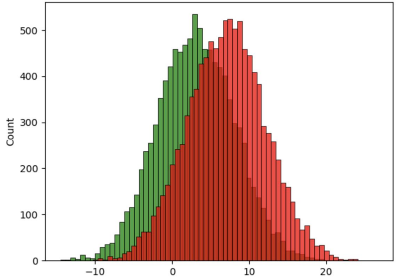
Solution: Re-randomization to avoid carry-over effect & make groups homogeneous

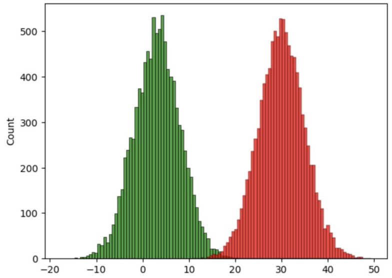
4.5 User assignment demo



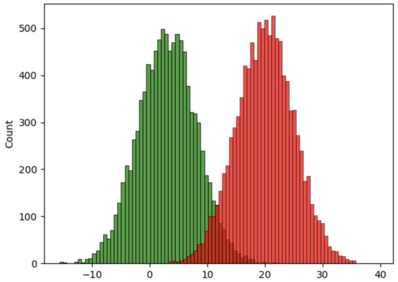
5.1 Statistical significance calculation

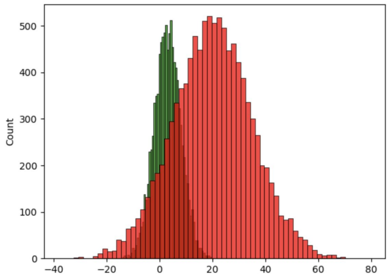
When are two distributions different?



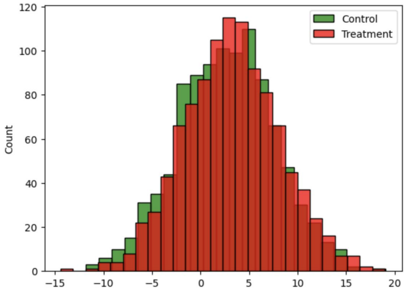


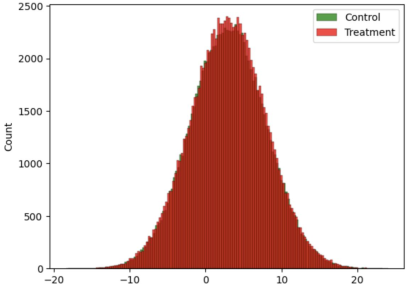
The difference in the mean is high





The standard deviation is low





The number of samples is high

5.2 Why do we talk about t-distribution?

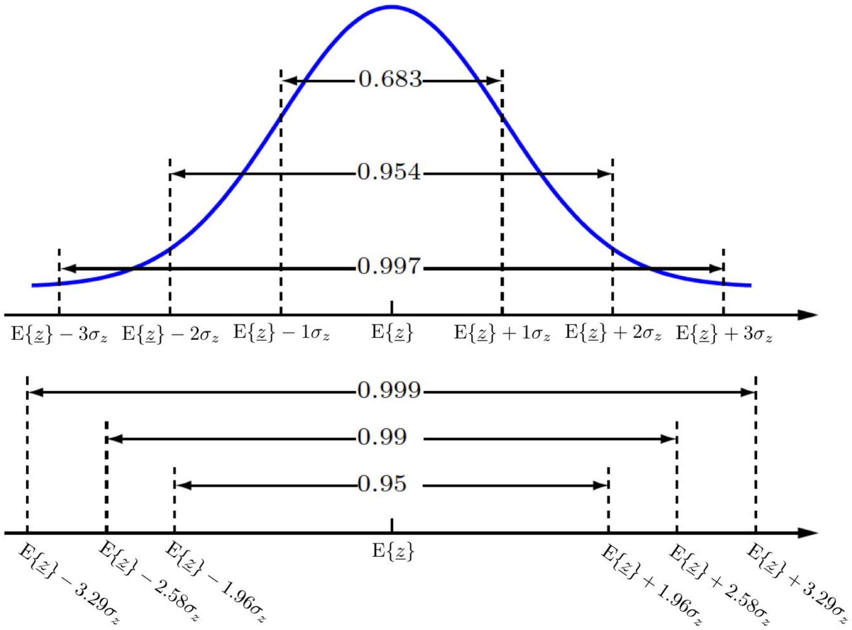
1	The difference in the mean is high	The higher the difference, the higher the confidence in two populations being different
2	The standard deviation of the samples is high	The higher the standard deviation, the lower the confidence in two populations being different
3	The number of samples is high	The higher the number of samples we collect, the higher the confidence in two populations being different

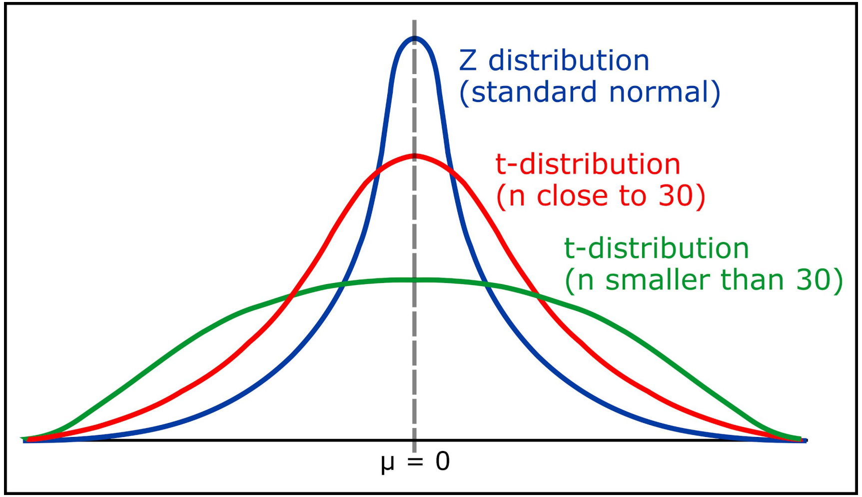
$$\frac{Signal}{Noise} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} = t - statistic$$

Welch's t-test = Unpaired two-sample t-test with unequal variances

1. Numerator = Sampling distribution of sample mean -> Normal distribution -> Difference of two normal distribution is again a normal distribution
2. Denominator = Sampling distribution of sample variance -> Chi-squared distribution -> Sum of two Chi-squared distribution is again a Chi-squared distribution
3. Ratio = Normal distribution divided by Chi-square distribution -> t-distribution

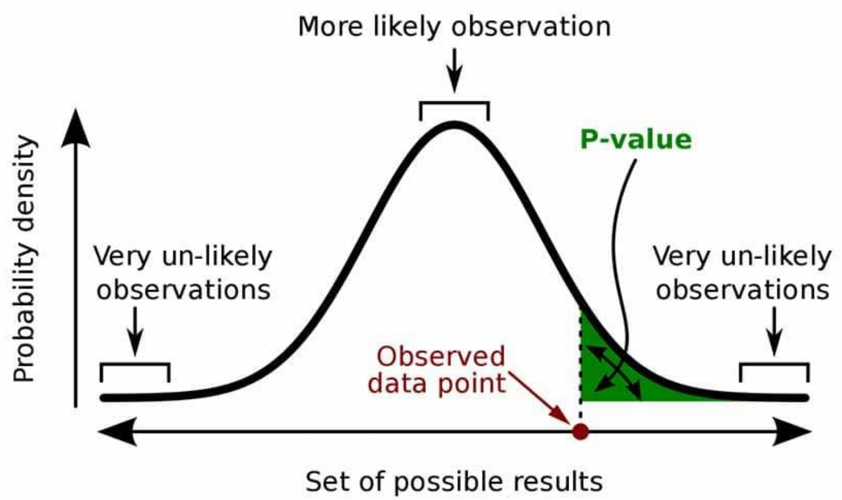
5.3 Normal distribution vs T-distribution



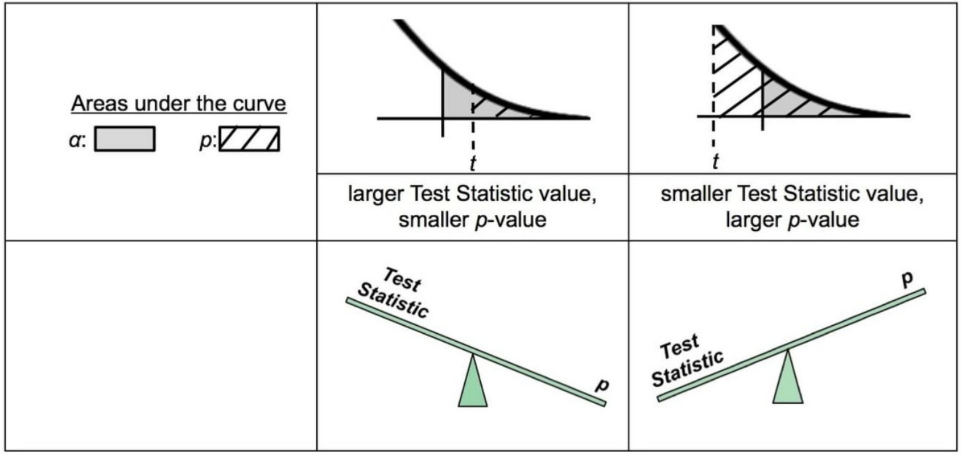


5.4 From t-statistic to p-value

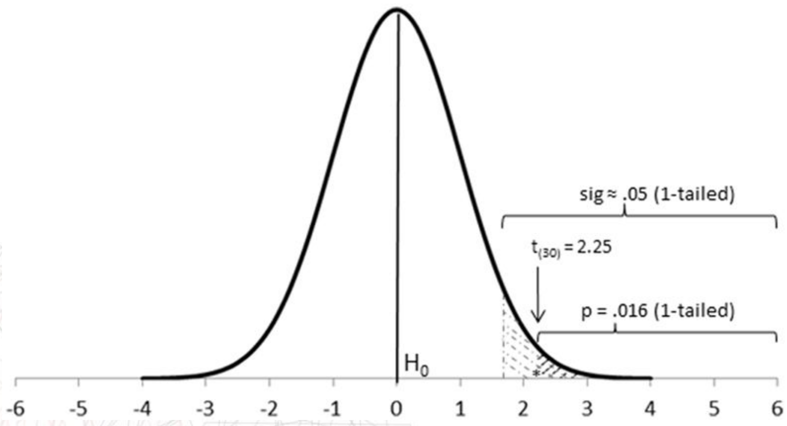
P-value = P(Obtaining a result >= than observed value | Null hypothesis is true)



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.



5.5 Taking decision based on p-value



P-value < 0.05

Decision: Reject the null hypothesis (H0)

Conclusion: Significant.

Scale up the treatment & make it as new control

P-value >= 0.05

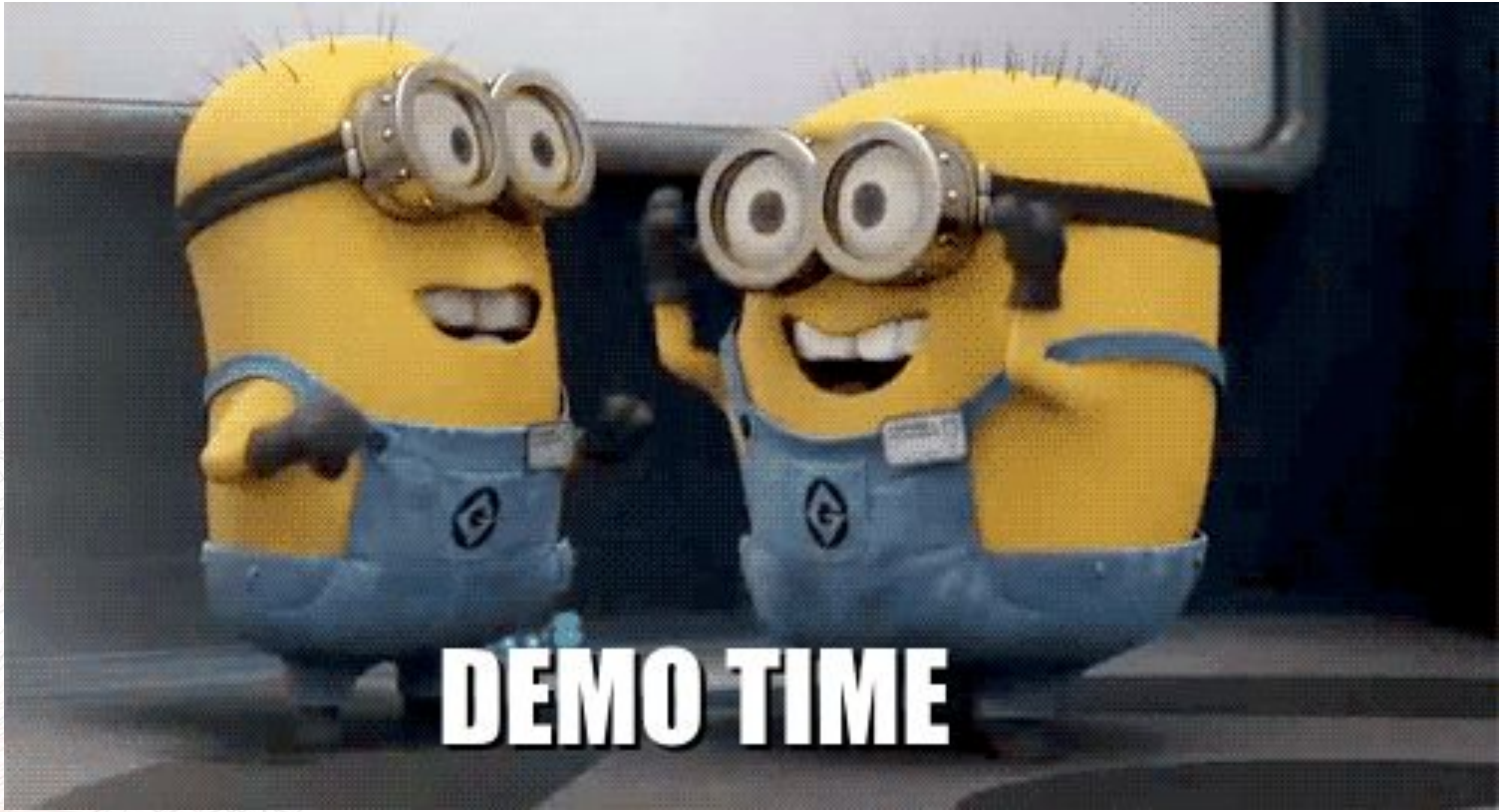
Decision: Do not reject the null hypothesis (H0)

Conclusion: Not significant.

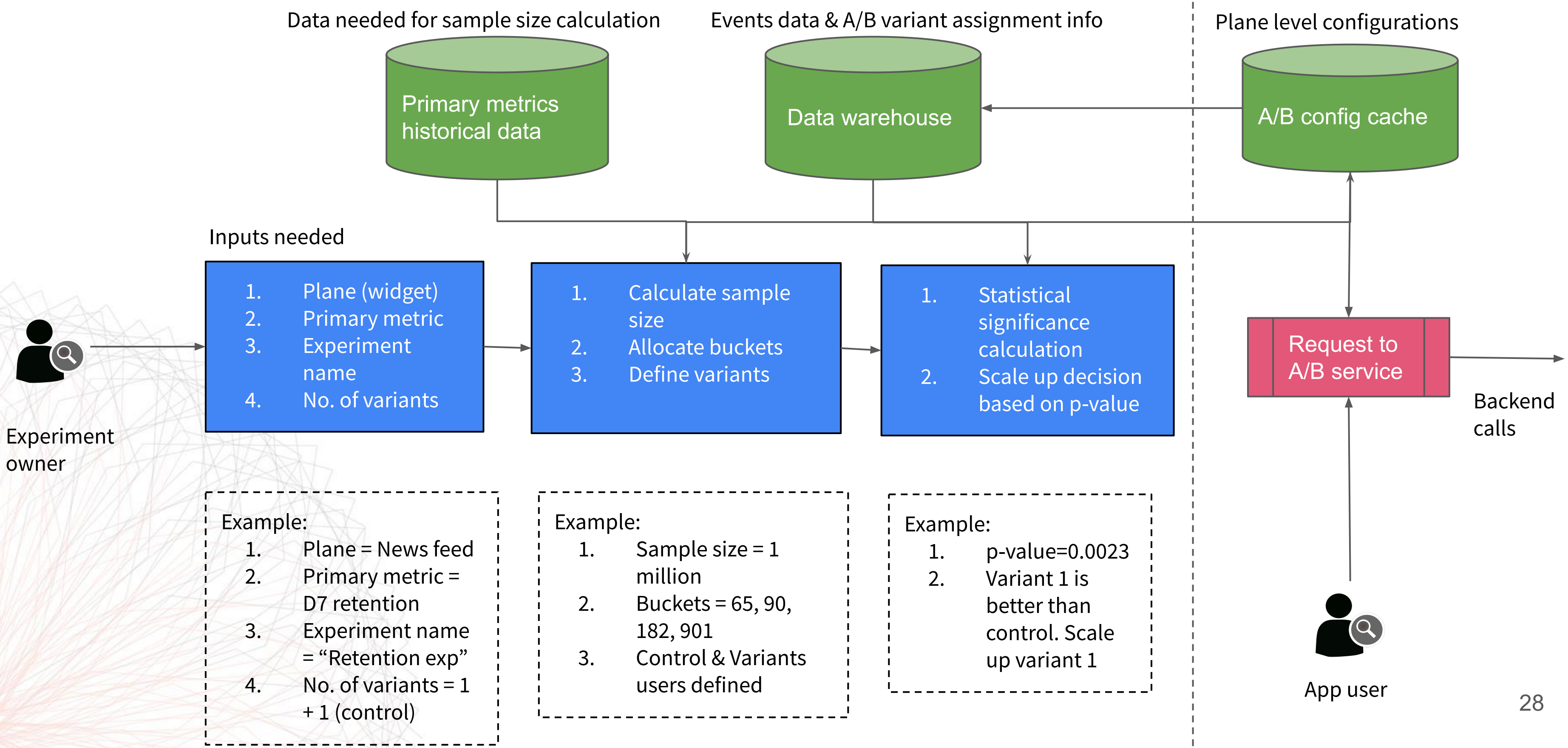
Do not scale up treatment

We compare against 0.05 because our Alpha (Type-1 error) is 0.05

5.6 P-value calculation demo



6.1 Overall architecture



6.2 A/B testing defaults

1	Experiment duration	1-2 weeks (to prevent day of the week effect)
2	Alpha (Type-1 error), Beta (Type-2 error)	Alpha = 0.05, Beta = 0.2, Confidence level = 95%, Power = 80%
3	Sample size	Two-tailed test: 15.7 * $\boldsymbol{\sigma}^2$ / δ^2 One-tailed test: 12.35 * $\boldsymbol{\sigma}^2$ / δ^2
4	Static audience or Dynamic audience	Dynamic audience
5	Assigning users to experiments in widgets	**Hashing with salt:** mmh3(user_id + “plane_salt”)%1000
6	Avoid carry-over effect	**Second level of randomization via Hashing:** mmh3(user id + “plane_salt” + “exp”)%2 == 0

Deck & Code

How to build an in-house platform to conduct thousands of parallel A/B experiments
https://github.com/ramab1988/dhs2023

Pyabtest package
https://github.com/ramab1988/pyabtest
https://pypi.org/project/pyabtest/

Thank You!

References:

- Overlapping Experiment Infrastructure: More, Better, Faster Experimentation
- Controlled experiments on the web: survey and practical guide