KPIs and Experimentation

Day 1 Lecture

Transcript:

0:00
thinking like a product manager is critical for being a good data engineer and what does that look like it looks
0:05
like you build pipelines that actually impact the business that actually change business decision and that can be in a
0:11
couple different ways one is you build good metrics that change business decision- making so learning how to
0:17
develop good kpis and good metrics is a very powerful way to do that another
0:23
side of it is you build metrics that are impacted by experiments so if you can learn how to do both of these things
0:29
build good kpis and understand how experiments work and how to plug your metrics into experiments you will become
0:35
a much more impactful data engineer and in this 2hour course we are going to be covering all of that so I'm excited for
0:42
you to check this out and make sure to like follow And [Music]
0:48
[Applause] subscribe so I think this is an important question why do metrics matter
0:54
at all it really depends on the company and all sorts of different things on like how much metric matter and how much
1:01
they actually impact things I know for example like when I like when I worked at Airbnb for example uh metrics
1:10
mattered less there uh than at Facebook for example like Facebook was a very
1:16
very very data driven company they were all about that data and I'm not saying Airbnb wasn't I'm just saying that like
1:22
a lot of times at Airbnb they'd be like well the data is pointing this way but
1:28

we have a design kind of situation where we think this is a better design anyways and so one of the
1:35
things that can happen there is it really does depend on who the founder of the company is of of like how much
1:42
metrics actually matter because the founder is going to set the culture of the company and if you look at the
1:47
difference between uh Mark Zuckerberg and Brian chesky then it becomes very
1:52
clear why there's a difference in the importance of metrics where you know Zuckerberg Computer Science Guy very
2:00
like nerdy you know he's like I am a robot right Zuckerberg gets that Vibe right it's like if he didn't care about
2:06
like like metrics that would be like very out of character for Zuckerberg right but then Brian chesk is more of
2:12
like a design guy he's more of like an artsy fartsy guy like I feel like if chesky didn't found Airbnb he totally
2:19
would have been like a starving artist type and so like very very different vibes from those guys and then like so
2:25
how they lead can make a big difference and so keeping in keeping that in mind when you're like working with metrics
2:31
and working with things like depending on the company you might need to have another layer there where at for example
2:40
at Airbnb one of the big things was it was it wasn't just about like this metric went up this metric went down it
2:47
was more like you needed to come up with a story behind it like oh like we made these changes and this happened and then
2:53
maybe it's like a a six-month Arc or a one-year Arc and you talk about like all the buildup to these changes and these
3:00
differences as opposed to like at Facebook you could just be like this number went down and that's bad or this
3:06

number went up and that's good right and that's where you can have depending on the company like how much they put in
3:13
stock in metrics versus storytelling can can tell you a lot about the company but
3:18
also metrics are important metrics provide visibility metrics ex they they explain the world especially as you get
3:24
more and more and more of them the the visibility that you get into your data
3:29
is just better so uh and just into your business because if you have different
3:36
types of metrics that are measuring different types of things you can make more clear decisions because you have
3:42
more visibility into the business environment and the more clarity and visibility you have the fewer icebergs
3:49
you're going to crash into because you can see a little bit further into the future and we we we're going to talk a
3:57
little bit more about that as we kind of go into more metrics but keeping in mind metrics are very important so what are we talking about
4:03
today well we have about five things that we're going to talk about uh so have metrics play a huge part in data
4:10
modeling um they in a lot of ways they kind of inform data modeling I'm sure if
4:16
y'all uh finished the homework uh from week three on the spec one of the things
4:23
you'll notice is that a lot of the metrics that you defined in the spec and the aggregations that you were looking
4:30
for really did inform what the data table should look like and that's a big
4:37
deal because sometimes you can have really funky metrics that people really want and they're just very hard to get
4:45
or they're very weird to like create a data table for especially like if you
4:50
are getting if you're getting Beyond like simple aggregations and counts and stuff like that it it can get very
4:57

interesting to get the right data model for that kind of stuff especially like if you're doing like ratios and denominators where you have a over b or
5:05
A over B plus C or like any of any other like anything that's more than just like a simple count or aggregation metrics
5:12
can get kind of funky so also metrics uh make sure uh metrics can't be gamed uh
5:18
we'll talk a little bit more in details there uh I I I ran into a couple different cases where I've seen metrics
5:23
be gamed at Big in big tech companies and how you can avoid that uh there's
5:29
definitely a lot of different examples there because metric metric go up does not mean good like it it's more complicated
5:36
than that and that's why you need to have uh some balancing that needs to be put in place we're also going to be
5:42
talking about the the inner workings of experimentation uh so how does an
5:47
experiment work how do you you know specify test and control and we're also
5:53
going to talk about the difference between feature Gates and experimentation those are slight they
5:58
are very related but also different so we'll we'll talk a bit more about those
6:04
and then to wrap it up we're going to talk about how to plug metrics into in experimentation Frameworks so uh for
6:13
example um we're going to be using stat Sig for the lab today and in stat Sig
6:19
you can put web loggers in your API and stuff like that you can definitely do
6:25
that with stat Sig um but the problem is is like you might need other metrics
6:30
that are not easily logged like uh they might be more from like a database or
6:36
more from like state or something like that that need to be more like eted and we're going to talk more about how to
6:42
get those metrics into your experimentation Frameworks as well so yeah let's uh let's let's get into
6:49
this all right so key thing here

6:55
uh when you're doing that spec the spec stuff that we were talking about in week
7:00
three and you're having back and forths with your stakeholders on what are the business metrics that you should be
7:06
looking at this is something that you should be considering at all times is if
7:13
they are asking for something that's more than like a count or a sum or something that's you know really just
7:20
like a group by like simple thing like that's going to be mostly something that
7:27
you should push back on at least in my experience like whenever the data scientist is like hey we need this funky
7:32
rolling average sum uh percentile thing uh generally speaking as a data engineer
7:39
you should push back on that that's like not your job that should be their job and like you should just give them the raw Aggregates and let them let the
7:46
analytics people give uh do whatever sort of analytics and numbers that they want on top of it so key thing here
7:54
don't let the data scientists dictate your data models
7:59
with extremely complex metrics to make you think like you have to add all these other crazy metrics into your data model
8:06
because that's not the case that's not the case and nor should it be that's just going to like it's going to make your life miserable so let's yeah let's
8:14
go to the next slide okay so what are the types of metrics so you know in the last slide I
8:20
was talking about there's Aggregates and counts those metrics are great uh there's also ratios you can think of
8:26
like percent of users in different countries talked about like compositions before in in a previous lecture uh you
8:33
can also have ratio which is like percent still active if you remember from last week we talked a lot about the
8:41

uh uh retention metrics right or it's like percent of users still active like the survivorship bias and those numbers

8:48

those ones are uh interesting and like because that's going to be like you have A over B in those cases and so those can

8:55

be other metrics generally speaking as a data engineer you should only be supplying Aggregates and counts

9:03

and what grain you supply those at uh is usually at like the entity grain

9:11

you know in like uh week one and two we talked about daily metrics where you have like user ID and then you have

9:17

metric name and then you have the value of the metric those are going to be generally speaking the metrics that you're going

9:23

to supply to data science and if they're asking for ratios or they're asking for percentiles or asking for anything like

9:30

that like they're just trying to give your job to you give their job to you

9:36

and that's not cool I mean I used to like as a data engineer that was a big mistake that I made was I felt like a

9:43

very confident data engineer where I was like I can do data science too so I'll just do it because they asked me to and I'm a good I can do data science and

9:50

data engineering I'm a freaking Wizard and uh for the most part like I learned that one you don't get recognition for

9:56

that in your performance reviews for the most part because your performance reviews are based on how well you do

10:02

data engineering not based on how well you do data science so that is a big minus uh and also you're going to be

10:09

asked a lot of weird stat statistical questions around the metrics that you supply that you probably don't want to

10:14

answer because you aren't a data scientist and that's more their wheelhouse and they and I don't know

10:20

like I'm fine with answering those questions because I have the statistical foundations to like do data science but
10:27
like I don't enjoy those those questions I don't it's not like I'm like oh my God like I just I'm so excited for you to
10:33
ask about freaking whether or not this is like a normally distributed freaking metric or like all the different things
10:40
that can happen when they're talking about metrics and definitions and stuff like that like so the key thing to remember is that if you define a metric
10:47
you should be able to ask answer any question about that metric and so that's why you want to keep them simple
10:53
because that's not like data Engineers aren't meant to be statistical Wizards
10:59
it's not our job like like I look at statistics as something that a data engineer needs to be like two out of 10
11:07
or maybe three out of 10 at to be successful on the job so you don't have to be a crazy wizard at it but this is
11:14
the main idea you have these three types of metrics and we're going to talk more about like why we have these types of
11:20
metrics because they're all pretty useful um so Aggregates and counts uh
11:26
they're the Swiss army knife uh you should be using these a lot as a data engineer uh you can think of like the
11:31
number it's like counting events like you know fact data is a very common one
11:37
like the number of times someone's logged in the number of friends someone has added the number of notifications
11:44
that someone has responded to like a lot of those are really powerful metrics but
11:50
also other ones are like more like they're more they're like in the middle
11:56
between like a metric and a dimension like are the is the user active and like
12:02
did did they send any um messages today did they like any posts today so it's
12:08

more like a daily active thing like a one or a zero kind of thing and those can be very useful especially when uh
12:16
you have uh a skewed demographic like for example uh with notifications if you
12:22
make an extremely viral post the number of notifications that you'll receive is a lot higher like an order of magn itude
12:29
or two orders of magnitude higher than the average user but if you have just
12:34
the notification sent or notification received yes or no it doesn't matter if you went super viral or not because
12:41
whether you got one or a million it doesn't matter and we're counting people not actions and that's a big thing to
12:48
remember here is that there's the two different types right you can have uh metrics that are based on the number of
12:53
people and then you have also ones that are based on the number of actions and those two different types of Aggregates
12:59
can be very very important and we're going to go more in depth into those different types of Aggregates uh later
13:06
later in the lab and in this presentation so keeping in mind uh you're going to have a lot of Aggregates
13:12
here um another thing to remember is uh you're going to want to bring in uh Dimensions here as well there will
13:18
probably be some specific cuts that people care about I know uh at Facebook
13:24
they really cared about the like the Young Generation trying to get them off of Snapchat and Tik Tok and getting them
13:30
on to uh Facebook obviously given The Branding that it doesn't seem like they were very successful but that was
13:36
something that they definitely tried to do and then you have uh other other ones like there's a big strong push to grow
13:43
in places like India and uh Brazil and other countries so there's probably a
13:49

geographic Dimension that you want to add and sometimes uh when you run experiments they can be statistically
13:55
significant overall but not statistically significant in some countries so we're get we'll we'll go a
14:02
little bit more into detail there uh as we kind of go through this presentation but the key thing to remember about this
14:08
slide is when you're working with metrics uh as a data engineer you generally want to give your your
14:15
analytics Partners simple metrics okay then you have ratios ratios
14:21
are important uh and they are uh something that mostly
14:28
data scientists should work with so that because like you don't want to have to be a data engineer and have to think
14:33
about like the numerators and denominators and the inclusions and exclusions of different users in the
14:38
numerators and the denominators of these uh functions that's stuff that should be done by the data scientist because
14:45
they're going to be better at that like statistical knowledge than uh you are as a data engineer um unless you're a super
14:52
awesome data science data engineer person then I don't know then maybe do both but even then I I'm more of the
14:59
opinion of having that be owned by the analytics people just so that you aren't burdened I mean in my case when I've
15:05
owned metrics like this over my career the problem that happens is that like I end up doing both roles and I'm like
15:12
answering business questions about these metrics and I'm like why am I doing this
15:17
I want to just write schola and do ETL and stuff like that and then it's like then I just do both and get kind of
15:23
burnt out obviously doing both is a good way to get impact but uh we can talk about that more later but anyways these
15:30

ratio metrics you can think of like clickthrough rate the number of people who saw your web page versus the number of people who sign up you also have a
15:36
purchase rate the number of people who signed up to the number of people who gave you money and then you also have
15:42
things like cost to acquire a customer like you might spend you know $1,000 on Google ads and then that gives you 10
15:49
users and then it's $100 a user to get uh to acquire a user so you can think of
15:55
like there's a bunch of different ratio metrics that you can have that uh like
16:00
are very common and ratio metrics generally speaking they measure not the
16:07
like amount of something but they usually measure the quality of something like the conversion rate is the quality
16:14
of your web page you know if you only convert 0.00001% of your users that visit your
16:19
web page it's probably terrible right but if you convert 100% like which doesn't happen but if you converted 100%
16:26
of the users that probably means that you had one traffic and it was you and you signed up and that was it right
16:34
because 100% also is like you should be skeptical uh but generally speaking a good conversion rate is like like more
16:40
than three or 4% if you can get that many people to sign up that's great so that's kind of the idea behind ratio
16:46
metrics and these metrics are going to be also they can also be cut by
16:52
different dimensions and this is where it's important to be careful because
16:58
what dimension you cut by makes a difference in how you aggregate uh
17:03
because if you have for example uh say you wanted to look at the number of you
17:09
the number of unique users or like like so you're looking at like unique website user hits div or
17:17
like it's a conversion right but like the bottom the denominator is the number of unique users who visited and the

17:23
numerator is the number of unique users who was there were there right and then if you look at it from if you break it
17:29
out by uh say like device operating system then the numbers aren't going to
17:35
necessarily add up uh to the right things because those like you can have someone who's on Android and iPhone at
17:42
the same time and so that's where you get this nonadditive sort of thing where it's like if you look at the conversion
17:48
rate of all the operating systems and like try to average it out to the conversion rate like and you do like a
17:54
weighted conversion rate of all the operating systems it might not equal the overall because of this additive versus
18:01
non-additive uh property of ratio metrics and like of certain Dimensions like operating system like interface
18:08
there's there's a couple of them that are like that we talked about that earlier in uh previous lecture but so
18:14
Rao metrics are all about you know indicating quality um okay so percentile
18:20
metrics percentile metrics are very useful so when I worked at Netflix uh one of the things that I worked on was
18:27
reliability of like uptime one of the things we we prided ourselves in at Netflix was we wanted to have
18:34
99.99% uptime so there was only like it was like 25 minutes a year that you get
18:40
where Netflix can be down and we were trying our best to reach that and we did a pretty good job so let's talk about
18:47
the the differences here of like what okay so if you had a metric that's called P99 latency what that means is
18:56
okay when our website loads for the slowest
19:01
1% how fast is the website so it's like so it's like essentially it's it's asking like how fast is our worst or
19:10

yeah how fast is our worst experience and so it's really important uh if you're trying to optimize the tales of
19:17
your distribution and capture more audience and capture more share I know that this uh these types of metrics were
19:23
also very very important at Facebook so one of the things when I was working at Facebook was there was was a big Push To
19:29
Go Global a big push to get people across the entire world and one of the
19:35
things that we realized was across the entire world like people not everyone had iPhones not everyone had or Samsung
19:43
Galaxy s37 or whatever like they all had like most people in the world like their
19:49
phones are kind of lowbudget and one of the things that we realized looking at
19:54
some of these latency metrics and some of these uh engagement metrics for people in these developing countries was
20:01
that they their phones couldn't handle Facebook because Facebook was like
20:06
either too data intensive too Ram intensive or like there was just so many different like things like the the
20:12
hardware of their phones just couldn't handle it so what these metrics helped
20:18
leadership understand was that they needed to make another version of Facebook and that's what they did they
20:23
made another version of Facebook called Facebook light that was uh it it took it
20:29
used I think like 20 or 20% the same amount of data as regular Facebook and
20:35
the install used like it was a 90% smaller install and it also used a lot
20:40
less Ram it was also stripped down obviously it's funny because like I remember at least like Facebook light
20:46
when I was like using Facebook light for a while I actually I shifted to using Facebook light for a while because it
20:52

was like a worse experience and I'm like I you know I I I'm already chronically on social media and I want to like not
20:59
be on social media so much so I like to downgrade my experience to make it a little bit less addictive but uh anyways
21:06
it's it it's funny because like Facebook light back then really reminds me of like threads now I'm like are they just
21:13
making the same app again like like what is going on with Facebook but anyways back to percentile metrics like because
21:19
Facebook light was a massive success like it got like hundreds of millions of users and it was able to it was able to
21:25
get that tail end of users who were not on good enough Hardware to enjoy regular Facebook app so you can have you can
21:34
have a you can look at the tail on that side but you can also look at the tail on the other side of like okay P10
21:39
engagement of active users so in this case like what are like what are we doing for people who are like lurking on
21:46
Facebook the people who are like they show up they scroll the feed but they never send any messages they never like
21:52
any posts they just like scroll Facebook and one of the things that uh there's been studies that show this that like uh
22:00
how you use social media makes a big difference on how it impacts your mental health and for example if you're like a
22:07
a lurker and all you do is passively consume then it impacts your mental health a lot more than if you're a
22:14
Creator and engaging and being and being more social like if you if you're more social on social media it's not as
22:21
negative to your mental health so that's another great example of like okay what about uh if we're we're looking at the
22:29
engagement metrics for the those bottom like lurker people like what's going on with them and like how like how can we

22:36
make maybe make those metrics better what what experiments could we run to try to engage the least engaged people
22:43
so these metrics matter a lot because they they are looking at the Tails this
22:49
is where you're going to have a lot of your incremental impact because of the fact that your average user like like
22:57
for example Facebook average user like is not going to leave they're just going to stay because of the fact that they
23:03
are nope Lulu come here and uh they're just going to stay and because like
23:09
they're in the mey part of the bail curve and they're like their needs are already being satisfied and they already are having a pretty good experience and
23:17
so that's where you really do want to look at these kind of tail metrics because that's where you're going to
23:22
find people who you can actually make their experience better and actually uh help them come back and engage more so
23:31
that's how percentile metrics work as a data engineer like you should probably not be passing these metrics to uh your
23:37
data scientist you should be passing them the the the daily Aggregates and then the data scientist people can do
23:44
their percentile magic and stuff like that on the data on top of that uh they can plug it into their experimentation
23:51
Frameworks All That Jazz okay shifting gears here a little
23:56
bit so metrics can be gamed especially with experimentation
24:04
so experiments can move metrics up shortterm but down longterm I remember
24:09
one of the things that happened when I first worked at Facebook there was an experiment that ran that was like what happens if we 10x the notification
24:18
volume in a small subset of the Facebook users and then they noticed like oh wow
24:23

if we 10x the volume of notifications we get more growth and they called it like blast like notification blasts and then
24:31
they were like wow we should probably do this more often and I'm I'm sure a lot of y'all on this call are like no please
24:37
don't like please I'll I will turn you off I will silence you don't do it and so like and that could be a big thing
24:44
that can happen where you can do these short-term like gimmicks and games and all sorts of things that can uh still
24:53
like make your numbers and make your metrics go up because of like some sort of like shortterm for some short-term
25:00
reason but then it's not like helping you look at the long term so the big one for me right was in notifications you
25:07
send more you get more use you get more users in the short term in the long term you lose that handle you lose that lever
25:15
because what happens is more people turn off settings and then you can't access those users anymore so what you need to
25:22
do is you need to create other metrics that can help you figure this stuff out
25:30
like for example um like in notifications I created the metric called reachability which was the
25:37
percent of users that we can send a notification to so then when they run experiments where they blast all these
25:43
users they can see that it's actually like detrimental so that made it so that
25:48
like and then we had a rule in notifications after that metric be came online where the only experiments that
25:56
got shipped were experim ments that increased growth and either were neutral
26:02
or increased reachability they they had to have both of them going up otherwise
26:07
like it was going to be uh an experiment that was uh considered not successful so that's a great point when
26:15

I was talking about earlier like why do metrics matter in those cases like it makes it so that you're not flying in
26:21
the dark where you're like wow uh like you see now we can see that we're we're
26:26
making notifications better increas growth and they're more relevant because fewer people are turning off their
26:32
settings so then you can actually get like a a a win-win you can set up your experiments to be win-win not just like
26:38
gimmicks and games and so other things you can do right with metrics right is you can you can fiddle with stuff you
26:45
can fiddle with the numerator or the the denominator where it's like oh what metrics are you looking at um for
26:52
example like you could say like I know a lot of them was like uh in engagement
26:59
per uh daily active user or engagement per monthly active user and you can look
27:05
at these different kind of uh crosssections of different users and I I
27:11
always thought it was interesting because like that was one of the one of our goals uh when I was working there
27:16
was I worked a little bit more on like email and SMS notifications which are used more to bring people back on like a
27:22
monthly basis than on a daily basis for Facebook and so it's like okay but if we
27:29
look at the engagement per daily active user then the metric the the experiment might look like a success but that
27:35
wasn't our goal to begin with because we're looking we're trying to boost the engagement of monthly active users where
27:41
the effect might be uh more diminished because of which people are in which
27:47
populations and so you can you can fiddle with the numerator and denominator of your metrics and like you
27:53
can essentially like get experiments to tell you whatever you want so uh that's where you want to be very

28:00
careful with doing stuff like that and not going too deep into like fiddling with stuff and like really having clear
28:07
hypotheses when you are on the out like when you're starting your experiment and that you can test those specifically and
28:14
look at those metrics specifically so that you aren't like trying just trying
28:19
to I don't know if y'all have ever heard of packing packing is a a great term that I recommend y'all look up and it's
28:26
it's essentially a way to make experiments kind of to show uh significance in whichever way you want
28:31
them to show significance or not and so then other things um you have a novelty
28:38
effects so for example if you introduce a new feature or you introduce new
28:43
things then users will uh generally speaking they're going to
28:50
uh be excited by this difference if you add a new tab you freaking uh change
28:55
your notifications I know this was an interesting one that uh that I I saw was like if you
29:02
added an an emoji to Notifications uh there was like this ma
29:07
like at the beginning of the experiment there was this massive lift it was like a 15 or 20% lift in a notification
29:13
conversion rate if you added an emoji and people were like in the notification team like that first week we thought we
29:20
had like we were all going to get promoted and like oh my God like we just we just struck gold we just had to put a
29:26
freaking smiley face in the not ification and who would have thought right oh my God right but then like as
29:32
the experiment kept running uh we noticed that that uh the the lift that
29:37
we saw at the beginning ended up starting to die down as the novelty of the new feature wore off so and emojis
29:46

were not the slam dunk to notification conversion that we thought they were and so that can happen though and that's why
29:53
like you need to be aware when you're running experiments for how long you're experiment should run um another uh
30:01
another example I had was when I was working at Netflix uh they were doing these feed refreshes where they were
30:08
trying to see like if we refresh the uh the movie feed uh uh in the background
30:15
while uh someone's not using the app and we do a background refresh does it increase retention because then the
30:21
movie feed is as fresh as possible and uh they tested it out like at different intervals like do we refresh every 8
30:28
hours 12 hours 24 hours right different like intervals and they noticed like the more often they refreshed uh the more
30:35
retention they got which was great but like what about the downside of that
30:41
like if you're only caring about retention then like obviously you would pick the the the most frequently
30:48
refreshing uh experience but then it's like if that adds millions and millions of dollars to your AWS bill because
30:55
you're just you know causing all all all sorts of additional requests then that
31:00
might not be the right play so another thing to think about uh that's why you
31:05
need counter metrics and that's what I did at at Netflix was I added another counter metric which was the AWS cost of
31:13
uh of an AB test so then they could look and they could see like then they could balance it out they they can then do an
31:19
ROI analysis of like we get this much increased retention at this cost because
31:24
a lot of times like you get this thing called diminishing returns where like like you do get a little bit more value
31:30
as like you keep uh increasing that refresh but it costs substantially more

31:36
for every uh interval at some point you get this like nice little Parabola effect and um and so oh no I mean a
31:44
logistic effect where you hit that kind of like carrying capacity where you have to increase the cost a lot more to get a
31:51
smaller and smaller benefit so metrics can be gamed we're going to talk a lot more about this on Wednesday as well
31:59
okay so now we're going to kind of go back to the fundamentals here of like how does an experiment work so there's
32:06
really four pieces to uh an experiment this is very similar to the scientific
32:11
method if y'all remember that from school uh your first step is make a hypothesis then you want to do group
32:18
assignment you can think of that as like test control you might have multiple test groups there could be two or three
32:24
or four test groups as well and then you want to collect data and then you want to look at the differences between the groups so that's
32:32
going to be the main things and we're going to go over each one of these in detail so that y'all can be ready when
32:38
you are making your own experiments okay so hypothesis testing
32:43
uh the the null hypo so hypo hypothesis testing is where you have What's called
32:50
the null hypothesis and then you have the alternative hypothesis and the null hypothesis is always the same which
32:56
essentially says there is no difference between test and control there's no
33:02
statistical difference like making this change will have no statistical difference and the alter the alternative
33:09
hypothesis is there is a significant difference from the changes that you introduce like the test and control
33:14
cells so those are the two um types
33:20

of um ways that or two hypothesis and these are like you can see how these are
33:25
like only one can be accept accepted or and and
33:31
so let's go into a little bit more details around this because this is an important kind of nuance
33:38
so you never prove the alternative hypothesis right so in this case you can
33:44
reject you reject or fail to reject the null so in the case when there is like
33:50
there isn't a statistical difference you fail to reject the null hypothesis which
33:56
says that there is none but if there is a statistically significant difference you reject the null hypothesis because
34:03
your alternative hypothesis might not be right it because it might have more details in like what you think is like
34:10
the reasoning behind it and it's it's similar in in science how like you know you have like hypotheses and then you
34:16
have like theorems and you only really get to that proof level when you have like a massive body of work right and so
34:24
you don't really prove hypothesis you just support them and that's going to be a key thing to remember here because of
34:30
the fact that you can have all sorts of other things that could be going on so you want to always be that kind of like
34:36
hesitant kind of handwavy data scientist in that case Okay so group testing uh this is an
34:46
important Point uh this is essentially how we assign test and control or uh to
34:53
our group members or and in this case our group members are our users and then and we got to ask a bunch of questions
35:01
so one of the big questions to ask is are these users in a long-term hold out
35:06
like for example at Facebook when I worked in notifications they had this thing called long-term holdouts which is
35:13
they wanted to measure the effectiveness of notifications as a product and the
35:18

way they did that was they took a small percentage of Facebook users who we just never sent them notifications ever they
35:25
got no notifications it's like a small percent of users get nothing and the reason why we do that is then we can
35:31
compare the growth of that group to the grow the growth of everybody else and
35:36
then we can see the impact that notification has on a percentage basis and that like incremental growth
35:42
comparing those two groups was a very critical me metric for Facebook but with that being said if the user's in a
35:49
long-term holdout and I'm doing a notification experiment that user is not eligible for that experiment because
35:55
they can't get notifications so you can have these long-term holdouts that are very important because they're
36:01
like long running experiments and you want long-term holdouts because they
36:06
measure the effectiveness of various parts of your application and and the
36:13
health of those parts of the application because you'll have users who don't have that feature and then you can see like oh are these users like like how many
36:20
more users are we getting back because of this feature and that could be huge as well and so like make sure that like
36:28
that's the first thing you want to test is like are they in a long-term holdout another thing to test is like you might
36:33
be doing an experiment that is geographic so maybe I'm doing like a
36:39
sale in the US and then it's like okay are the are are my users in India uh
36:45
eligible no because it's it's a sale in the US so like uh a lot of times your experiment groups are not 100% of your
36:53
users it can be some fraction of your users based on different criteria that
36:59
you have for them another one might be like oh I want to do an experiment on Chrome to see uh if these users have

37:08
this bookmark installed or to have this thing installed on or like uh an
37:13
extension installed or different things like that like and then then then your population is just people who use Chrome
37:19
not everybody so you can have experiments that have all sorts of different like criteria to be involved
37:26
in and we're going to go over how to set some of that stuff up in stat Sig uh today in the lab and then another one is
37:33
like what percentage like of users you want to experiment on because another thing is is like what if your experiment
37:39
sucks what if your experiment has a lot of downside to it you don't know and that's where it's like if you're a big
37:45
company like Facebook and Google like you can actually really limit the the downsized impact where it's like you
37:50
might start with like 1% of users or half a percent of users and then if the experiment shows promise then you open
37:58
it up to a wider uh number of users to see like how the power and like the statistical trends that they'll probably
38:05
stay the same because you have a a strong enough statistical sample but they they could shift as well so that's
38:12
where like what percent of users do you want to experiment on uh for for my uh experiments that I'm running on my
38:18
website I just do 100% because I don't have 10 trillion users like Facebook and
38:24
Google so I for me to get enough statistical power out of my tests and out of my experiments I need pretty much
38:31
every user I can get so that's where it's important to look at what the differences are there and like how big
38:39
Tech versus a smaller company is going to do experiments and a big one here is going to be the percent allocation
38:45

because big Tech gets that luxury where they can experiment on a smaller percentage of users and see what happens
38:51
before they open it up to everybody else whereas at a smaller company if you did that you're going to get like two two
38:57
users and it's like okay um you don't have enough statistical power to prove
39:03
that there was an impact or not so you want to make sure that your tests are powered we're going to talk more about
39:09
that in another slide here okay so let's talk about group
39:15
assignment so there's uh two ways that you can be
39:20
essentially uh assigned to a group in uh in in an
39:26
experiment the first question is are you logged in and if you're logged in you can have a
39:33
logged in experience and uh and your your user ID for your app will be the
39:41
identifier for that user and logged in exper experiments are more powerful than
39:46
logged out ones in some regards because of the fact that you have a lot more
39:51
information on your logged in users right like for example pretty much everyone in this call filled that intake
39:58
form so I know all like the time zone that y'all are all in I know all sorts
40:03
of different things about y'all right and so if I run experiments on exactly.com that are logged in exper
40:09
experiments I can use that and assign groups and bring in more data and richer
40:16
data about y'all because I have so many more like dimensions and different things on like how y'all interact with
40:22
uh my website but on the flip side like uh you also probably want to do logged
40:28
out experiments to see how people convert so that you can get customers and usually speaking if you're going to
40:34
be doing a logged out experiment you're going to want to use one of two identifiers like you can either use the

40:40
uh an IP identifier but you want to Hash it if you hash your IP identifier then
40:46
that will uh be a way to get someone to kind of have a stable identification the
40:52
problem there though is like if someone's on like they're visiting your website and they're on mobile and then
40:58
they go from Mobile to Wi-Fi back to mobile then that's going to be uh treated as two different users even
41:05
though it was one user that's where stat Sig is pretty cool they have this other thing called a stable device ID which is
41:12
going to be a more consistent way to hold on to users that you can also use
41:19
they uh they have their different trade-offs and benefits right and the stable device ID can also be pretty powerful as well so uh and also know
41:26
that like the if you use IP another thing that can happen is if you have two
41:31
people who are two people on the same Wi-Fi that are visiting uh a certain
41:37
website they're going to have the same they're going to be the same user so you also get the you G have one user become
41:43
two but you can also have two users become one and so you can actually have both experiences happen like when if you
41:49
use just IP and so that's where you can that being said IP is also very simple
41:54
to use and actually in the lab that's what we're going to be using so what happens is you pass your
41:59
identifiers to stat Sig stat Sig will then say okay uh server use the blue
42:05
send the blue environment or the red environment so you can think of like the different colors as different
42:11
experiences that users can uh get and Stat Sig will essentially tell the server that and then the server will
42:18
send to the client the the the the changed view like whether it's a blue button or a red button or you you you
42:25

say like data engineering a Academy versus data engineering course or different things like that and you
42:31
change the words up so then you can measure the impact of your experim experiments over the long
42:37
term and another big important thing that you have to do is you need to track
42:43
your events so if I send a client with a blue button I need to track all the
42:50
events like so that I know that this like they got the blue button and then they signed up and then they bought from
42:56
me right have the whole funnel of events that they did uh and so those events can
43:02
happen either on the client or on the server depending on where you want to do your logging and Stat Sig uh offers apis
43:08
for both right you can have a client you going do client Side Event tracking or server side um for the lab today we're
43:15
going to do all server side because it's simpler um so this is kind of the idea
43:20
behind how to run experiments there is one uh one piece in here that is missing
43:26
and that is what about other metrics that you that
43:31
aren't logged on stat Sig but you want to add like for example you might have a user pipeline that you want to add a
43:38
bunch of other extra Dimensions to so you can have like you can imagine like a separate ETL process that dumps into SAT
43:45
stat Sig as well that gives you like user level dimensions and metrics as well that you can also add so uh and
43:53
we'll talk a little bit more about that at the very end of this presentation okay so that's the idea behind group
43:59
assignment that's how they get into test to control it's based on those identifiers then you collect data you
44:06
see okay you collect data for a while until you get a a statistically
44:11

significant result generally speaking in big Tech that meant you waited at least a month so you held you ran your
44:18
experiment for a month to see what was going to happen so and what I like what I like I was saying in the previous
44:24
slide you want to use stable identifiers uh whether that be your IP address IP
44:30
address or a stable ID uh in stat Sig and then there's also um you can also
44:36
use uh your user ID like your logged in user ID if you're doing a logged in
44:42
experiment we're going to be doing a logged out experiment in the lab today because setting up logged in would be
44:47
just a lot of work so collect a lot of data until you get a until you get a
44:52
statistically significant result but keeping in mind you might not especially if if uh you're looking for either a
44:59
very high level of significance or the change is very nuanced and they're
45:05
actually isn't a statistically significant result generally speaking the longer you collect data the more
45:11
likely you're going to get a statistically significant result okay so how it works right is the
45:18
smaller the effect the longer you're going to have to wait so if there's a big change and like one one like uh one
45:24
version of your test is just atically better than other versions then great
45:31
awesome and then uh you're you'll get your statistically significant result very quickly and that can be a very it
45:38
can be a good feeling too because then you're like oh awesome like my hypothesis was correct and I got my data
45:43
really quickly and that can often times be a sign that you have like a slam dunk experiment and uh on the flip side like
45:51
if it's been a month and you still don't have a statistically significant result it might not ever be right or might it
45:57

might the effect might be so small that like it doesn't matter so keep it keep that in mind when you're like collecting
46:03
data uh another big thing do not underpower your experiments So today
46:10
we're going to be talking about how to have multiple test cells so you can have uh test cells one two and three and then
46:16
your control cell you can have different test cells and uh the problem is like if
46:22
you have too many test cells then you're going to run into problems because you're going to have
46:29
underpowered because imagine if you had 10 test cells and that means that like each cell is only going to get like 10%
46:35
of the data and if each cell gets 10% of the data then in order to get enough
46:41
data to to measure if there's a statistically significant result is very
46:47
high and you don't want to do that because you if if you're working in a
46:53
startup you definitely don't want to do that because you don't have enough data Google is a little bit different right they they had they ran an experiment
46:59
where they tested 41 different shades of blue and they were a able to send like 2 and a half% of the traffic to each uh
47:07
shade of blue and then they were able to get enough data back on that and they didn't have to wait a million years and
47:12
the reason why they didn't have to wait a million years is because they just have an insane amount of data at their disposal and they can do a lot more
47:18
experiments that's one of the things that's beautiful about having a lot of data is that you can run experiments a
47:24
lot quicker because you have like just more incoming uh input from people and
47:31
remember this this is critical is that like when before you start your experiment you want to make sure that
47:37
you have logging set up for all of your events and dimensions you want to measure so that you can check and like
47:44

you don't have to be like oh halfway through the experiment be like oh man we forgot to log this other U metric and
47:50
now we and then cuz then like you're going to miss out on a lot and then it's going to mess up the the differences cuz
47:56
like you're going to have like a skewed experiment where half the people were in the experiment and you didn't log the metric and then the other half you
48:02
logged it halfway through and it's a problem so that's a big important thing to remember before launching experiments
48:08
is make sure you have all your logging in place okay now we're getting to the cool
48:13
parts of this so you want to look at the data and here's an example of an experiment that
48:19
I'm running on my website on exactly.com and you're going to when you're running an experiment you're going to look and
48:26
you're going to get charts like this that are different metrics and you can see like the statistical significance uh
48:33
like kind of day over day or like the different this is not the this is just like comparing test and control so in
48:39
this experiment uh I'm doing Red versus Blue buttons and you'll see if if
48:45
there's if there's a bar that overlaps zero so if you have negative and positive options then it's not a
48:52
statistically significant result so the only like the only metric that actually makes a difference here is whether or
48:59
not someone visits the signup page and you'll see that I mean there's not a lot
49:04
of data here yet and that's why these uh the air bars are so freaking huge so
49:09
you'll see in this case this uh the impact here is 21% plus or minus 20% so
49:16
that means that like the Red versus Blue Button could be it could impact signups between one you get impact sence between
49:22
1% and 41% that's like the the range of like
49:28

what could possibly be here keeping in mind that I've only been running this experiment for two days and the the
49:34
confidence interval right you see the confidence interval here is only 80% which is a pretty weak confidence
49:39
interval that's not the uh industry standard in data science I would like so
49:44
technically like if you use industry standard this experiment shows that like whether I use a red button or a blue
49:50
button there's no statistically uh significant difference so uh and let me
49:55
show you that like as talk more about P values so in this case the P value would be 0.2 which is a very high P value um
50:04
that's what we want to look at so P values so generally speaking P values
50:09
are uh the the the standard one is uh 0.05 so what that says is there's a 90
50:17
like if the P value is less than 0.05 that means that there's a 95% chance
50:23
that the effect that we're looking at is not due to chance and is due to some
50:29
other Factor right and it's not due to Randomness and uh and so in the last
50:34
slide you see when it's like this 80% confidence interval that means that for for this metric this visited sign up the
50:40
one that's statistically significant down here that means that there's an 80% chance that this is not random but a 20%
50:48
chance that it is random so keeping that in mind when you're like looking at confidence intervals and the the lower
50:54
the P value the higher certainty you have that the randomness is uh it's not
51:00
due to Randomness but due to some sort of underlying change so um P values are very important
51:08
if you haven't looked at them I highly recommend looking into them they're a statistical tool that are used to pick
51:15
out change and they're one of the most important things when running an experiment to determine whether or not

51:20
we can say that there was an actual change or not okay uh but just because something
51:27
is statistically significant doesn't mean that it's a it's valuable uh and one of the reasons that is the case is
51:34
the statistically significant change might be tiny like for example in that last slide you saw how like I was like
51:41
oh there might be a one% lift if I use a red button instead of a blue button and
51:46
it's like if it's like a 1% lift is that even like is that I mean it's valuable but like what if it's like a 0.001% lift
51:53
then it's like okay that's essentially the same even though there is a a statistically significant difference
51:58
it's the same and that can happen especially for experiments that are running for a long time because as you
52:04
collect more and more data more and more uh metrics the differences in them will
52:09
become statistically significant but the the Deltas might be tiny and if they're tiny like they're probably worthless so
52:17
that's a a key thing to remember about statistical significance okay so I want to talk
52:24
about some other important gotas with statistical significance um if you have
52:30
tested control uh here's an example for you like so say you're measuring notifications received and like that's
52:36
your uh that's the metric you're looking at in your experiment and uh in the test
52:42
group you have a bunch of average people and then in the control group you have like Beyonce and you have uh Justin
52:49
Bieber and uh and then a bunch of also a bunch of you know average people if you
52:55
look at notifications received uh Beyonce and Bieber are probably going to have a few more notifications than
53:01
everybody else and that's going to really skew the average up a lot where it's kind of like how you know in a room

53:08
of me uh and uh Warren Buffett the average between us is a billionaire even
53:14
though I'm pulling him down a lot he just has so much money that it doesn't matter and so it's the same idea here
53:20
where if you have extreme outliers you want to be careful and uh one of the things you can do is this it's a
53:25
technique called winsorization so if the outliers are so extreme what you can do is you can clip
53:32
them to a less extreme value and generally speaking that's like the
53:37
99.9th percentile is what you want that's what the like the standard kind of winsorization is so that like if you
53:43
have like just very very extreme outliers that will make it so that you don't have as much skew but another
53:49
thing you can do is instead of looking at um event counts look at user counts
53:54
because that can uh but it's depends on like what you're trying to measure but like a lot of times user counts is going
54:00
to give you a better view because in that case it's like okay uh Bieber and
54:05
uh Beyonce both received a notification so they count as one and I received a
54:10
notification so I'm at the same level as Bieber and Beyonce right so um anyways
54:16
that's kind of the idea when you're looking at statistical significance remember that skew can cause problems uh
54:21
winsorization is often built into different experimentation platforms to help uh avoid this problem but it is a
54:28
technique that you should be aware of okay we are getting real close so so
54:34
statsd can create metrics uh we're going to go over how they're logged and all sorts of things um in uh in the lab
54:40
today and uh how about adding your own so this is a very common problem and
54:46

this is a big problem that I did at Facebook was like uh as a data engineer you often are going to be adding new
54:52
metrics into experimentation platforms so that data scientists can look at how
54:57
they change via experimentation and so that's a big thing to remember as like uh you're
55:03
building stuff out is that like a lot of these daily metrics and daily Dimensions or slowly changing dimensions are going
55:10
to be added into your experimentation platform so that data scientists can go nuts to see like if there's like uh any
55:19
dimensional impact on an experiment or like if this metric is important or what counter metrics you need so very common
55:27
very common thing I've done this at literally every company I've worked for in Big Tech congrats on getting to the end of the day one lecture if you're
55:34
taking this class for credit make sure to switch to the other tab so that you get credit