This is an automated transcript without any guarantee of accuracy

good morning everyone
or good afternoon or good evening as the case might be
depending on where you're located around the world
my name is Martin Tingley
I work on the centralized experimentation platform
and Netflix and today
I'll be talking broadly on the themes of demo
democratizing
decision making and experimentation at Netflix
and we'll hit on this theme of democratization
three times throughout this talk
and will summarize them at the end first
how we make decisions second
what we try in our product experience then third
how we scale this process throughout the company
to kick us off
let's go back in time about a decade to 2010
this is what the Netflix television application
look like at that
at that time pretty primitive user experience
from today's standards pretty static
some limited interactivity just generally looks
looks pretty dated at this point
but it served as well about a decade ago
so let's fast forward to today or almost today
this is basically the current state of our
our TV application
to much more video forward experience
a much richer
more immersive experience lot more interactivity
more efficient use of space
and overall would like to believe
a better experience for our members
so the question is
how do we make all of the decisions required
to take us from that static
primitive experience from a decade ago
this video forward uh
more immersive
and interactive experience that you'll find on the TV
the Netflix TV application today
so how do we make the right decisions to involve
evolve the product and this extends beyond the UI
this extends to
to everything we think about when it comes to the
the Netflix product experience
well

there's a lot of options on how we might make decisions
as a product organization
in Netflix we could ask a hippo
a hippo here stands for highest paid person's opinion
right just say hey
you know whoever's making the most money in the room
will let them decide they must know best
we can hire a bunch of experts
and just do what they tell us
some some great graphical designers
some great UX designers
you know great great product managers all of that
we can step a giant debate within the company
about what we should do and you know
maybe whoever is persuasive
most charismatic during that debate
maybe their opinion will will carry the day
and more and more these days
being a streaming video on demand service
we could simply copy the competition
we all know
there's enough competition out there that we can copy
at this point
so of course it's a trick question
we don't do any of these as
use any of these as the basis for decision making
instead
we AB test every idea before exposing it to all of our
our paying members so what is an AB test
very simply we take our Netflix members
who we always hope are happy and dancing
and then we take two versions of our product experience
version a or the control experience version B
the test experience the treatment experience
what you're called as well
and here the difference is that middle row of images
version B introduces this mobile preview experience
which you can now find on our mobile app
we take a subset of our members
divide that subset into two
one subset one random sample gets version a
one sample gets version B
and we expose members to these
two different experiences
we compare some business outcomes
is the basic framework of a test
later on in this talk
will go into a bit more of the technical details
okay so why does Netflix believe in experimentation
as a way to make decisions

well simply put
it's our belief that experimentation
enables us to make better decisions
about how to evolve our service
so ultimately we can deliver more joy to our members
is really
is about maximizing the value of the experience
we give to our members
and we believe experimentation is the way
we can do this well
and this
this brings us to this first democratization theme
AB experimentation is scales
and by that
I mean many members are voting on proposed experiences
if we think about these these different paradigms
we could use to make decisions
starting with the hippo over there
and which is one person
and ending on the far right with millions of voices
getting to contribute to that decision
we start mapping some of these decision paradigms onto
that you know
if Netflix were to rely on the hippo internal experts
group debates
maybe somewhere between one and a few thousand voices
Max would participate in that decision making process
we were to do qualitative customer research
somewhere between hundreds
maybe tens of thousands
with AB experimentation
we can we can reach a lot more people
we can run very large experiments
at the scale of Netflix
you can really democratize the input
to that decision making process
we can give each each of the members in that test
to vote on how we should evolve our product experience
and they vote with their behavior
if one of these product experiences leads to more
engagement will conclude hey
this is this is a great thing for our members
it's allowing them to
to derive more value from our service
so that's the marketization theme one
how we decide we use AB tests
and really that means
we allow our members to vote with their actions
on how to evolve our product to deliver more joy
okay

so whenever I talk about AB experimentation
groups like this one of the questions that comes up is
why not just roll out a new feature to everyone
and just measure what happens right
take the data before we roll out that feature
take the data after that feature
and see if there's a difference
in and use that to conclude that this
this new product experience is good or bad
so let's talk about what could go wrong if we do that
and
and we'll use kind of a simple toy example to
to illustrate so on the left we have
you know
our current product experience will call it product a
which is a standard boxard
it's getting those bottom two rows
and product B uses upside down box art
someone has this idea hey
if we
we use upside down box art will increase engagement
that's the driving hypothesis behind this test
and as an outcome metric
will look at something like streaming hours
just general engagement with the service
okay so say say we
we we just roll out that new upside down box
our product experience product B
and say we roll it out on December 2021
2018 very precise date
and this is the data we collect blue that product a
the regular box art red that product B
the new upside down box art
and wow look at that engagement
engagement metric spike
when we roll out that new experience
this this looks great
okay
so the the question for for the
The room is based on these results
when you roll out this product experience
with upside down box art
on the one hand you know
when we think about intuition
we think about design
this is a pretty terrible experience
but on the other hand
we have this data that suggests hey
you know engagement spikes when we roll this out
what what should we do

well of course
it's a trick question
and the issue here is that
correlation tempts us to infer causality
when perhaps no causality exists
so it turns out that Netflix launched the movie
Bird Box
on the same date of this hypothesized experiment
about launching product B
Bird Box is a pretty big hit for us
drove a lot of viewing so the challenge now is
we don't know why streaming hours spiked
starting on December 21st
was it this new product experience or was it this big
big film that launched and you might say well
you know just don't launch your new product experience
on the day a new
a big movie comes out and the challenge there is
we don't really know what movies are gonna be big
there's stuff we can't control
there's always what we call confounders
kicking around in the background
so instead if we had run a test
and had these two product experiences
the regular box art in the upside down box art running
running in parallel for different members
across the launch period of Bird Box
we might see data that looks like this where yeah
basically every every day that product a
the regular box art
result in more engagement than product B
and then when Bird Box comes out
both of these experiences
uh see
see increased viewing that we
we think is probably caused by Bird Box
not the product
uh so the point here is because
because we've now randomized the assignment
because we have these two product experiences
running a parallel
we can really conclude that product a
is what is causing the higher screaming hours
it's not Bird Box
it's that different product experience
and we sort of controlled for the presence of Bird Box
by comparing engagement
under these two product experiences
okay
so that's why we we run experiments

rather than just roll things out
and do what you might call a pre post analysis
so what's interesting about experimentation
certainly at Netflix
I believe this holds more generally
is despite having expert colleagues
and all kinds of discipline to generate
you know explore
test experiment
with tons of creative ideas to improve our service
most experiments do not win
we have a lot of ideas and most are not successful
it's relatively small fraction of experiment
where you see that light bulb really going off
and our members through your action saying hey
you know putting up my hand here
this is a great new experience
roll this out to everyone
most things are either you know
flat or just a negative experience or
and by running experience experiments
we actually let our members tell us
which of these potential products
experiences are good
and this brings us to our second democratization theme
which is the democratization of ideation
ah because we are in this
environment
where we will test any proposed product change
because most proposed product changes are
are actually not winners
there's a real hunger and thirst for new ideas
how can we make this product better
what are we missing what do our members want etcetera
as a result great product ideas can come from anyone
yes
many come from our product managers from our designers
but we also see user facing product innovation
ideas coming from engineers
from scientists from executives uh
and and
because so many of our ideas don't pan out in tests uh
you know there's
there's a real opportunity to
to sort of put your hand up and say hey
I've got an idea and and you know
all you gotta do is convince someone to test it and
and we'll see
so experimentation
really permits for the democratization of

of ideas about what to try
because we test and because most ideas are not winners
better innovation ideas really can come from anywhere
I I've seen many MMO at Netflix where the
the product manager and partner
charge of that part of the experience
no take the original idea here came from
from an engineer or
or or from from some other type of function uh
so the marketplace for ideas is really super open
once you move to experimentation
um it's kind of like an ego free situation
when it's done well
because it's less about who has what title
and more who has an idea that we think is gonna work
we hope it's gonna work
and we test it out on our members
and let them help us decide
okay so
what we're going to do now
is take a little bit more time to go through
how to experiment works
some of the technical nitty gritty
and then I'll talk a bit more about what my team does
in Netflix
and that will bring out
this third theme around scaling experimentation
so experiments all start with an idea
we have on the left our current product
and on the right we have a new
a new idea for the product here again
it's this mobile previous experience
you can sort of click into this
it will expand to a vertical trailer
and you can swipe through the trailers
and sort of a richly immersive video based experience
uh so we have this idea that this
this mobile preview product feature is good
for our members we need to do
is convert that into a testable hypothesis
and identify metrics that will help us determine if
if it is in fact an improvement for our members
so the the way we think about this is
if we make change X at that mobile previous experience
it will affect member behavior
in a way that makes metric y improve
or we can measure metric y
so in this particular experiment
the hypothesis was something like
presenting a row of short previews will increase

awareness of these titles
and make it easier for members
to find something to watch
increasing our core engagement metrics
so there's this nice causal relationship um
members will engage with this row
that will help them find stuff to watch
and as a result
we'll see more we'll see more engagement
we generally spend time thinking about
if experiments are even worth running
so some questions you might ask is
you know would you release or not release this feature
regardless of baby test results
or certainly anything to do with account security
we will roll out rather than test because you know
we just want to keep those accounts secure
same for things potentially like parental controls
you know just doing that is just the right thing to do
second we think about
if the potential
results will actually be meaningful to our business
and we will really deliver more value to our members
and and if not
might say hey
why are we spending our time exploring these
these sort of product innovation ideas
why don't we do something more impactful
so in the case of that mobile preview experience
it is a new feature it's differentiate to us a bit
and the underlying hypothesis is one that
that is directly relevant to our business will
will increase engagement
then the third part is
is it even possible to validate the hypothesis
through an experiment
is there like a super well defined causal relationship
so in that mobile previous experiment
you know really was like
we'll add this feature folks will engage with it
you'll help them find something to watch
that will improve these core engagement metrics
or very well defined causal
instead of causal relations there
uh if all those questions turn to the right way
we'll run the experiment again
we take our Netflix members
you'll hear this referred to as the target population
I will take a random sample from that
target population divided into two buckets

each of those buckets will get a different experience
version a control version B test
then we'll compare business outcomes
so that's the statistical analysis of of metrics
and critically we hold everything else constant
across these two experiences
any big content release will impact both
control and test the same way
so when we compare control and test weeks
we still have sort of a valid causal causal read
okay so we run our test
we gather a whole bunch of data
then we have to analyze the results
and in this case
when we compare the behavior of members
who saw the new previous experience
with those who saw the original experience
we found based on our our metric hierarchy
or our internal metric hierarchy that
you know primary metric didn't really move
our first secondary metric didn't really move
but our second secondary metric
saw a statistically significant improvement
and and this is the evidence
we used to conclude that this was a
good experience for our members
let's take a little bit of time to
to talk about and build intuition about
the statistical terminology and concepts
we use to help manage and understand uncertainty
with AV tests to help us do this
we're going to come back to that age old question
on the internet
how do you identify photos that have cats in them
okay so
here we have a photo of a cat and a photo of a not cat
and there's two ways to correctly identify
if a photo has a cat in it
that the photo with the cat can say I am a cat
and the photo that is clearly not a cat can say
I'm not a cat
so the language you'll hear is
is this is a true positive
a sort of a positive
identification of something that's actually a cat
and a true negative
the correct identification of something that's not
a cat
and likewise there's two ways to make an error
the cat can can be mislabeled as not a cat

and the not a cat can be mislabeled as a cat
so we call these a false negative
we failed to find that cat in a false positive
we claim we find a cat when we didn't
so there's four possible outcomes
two ways to make the right decision
and two ways to make the wrong decision
and the same core ideas show up
when we think about the results of
AB tests
so to illustrate supposed
we saw a 1% increase in our primary metric
in our AB test
this result is still uncertain
it could be a false positive
you'll also hear this called a type 1 error
so what does that mean
means in the context of that experiment
we claim there's an effect from the experiment
from this new product feature
but really there isn't so it's sort of like that
not cat being identified as a cat
uh and likewise
supposed
we saw no change to our primary metric in our AB test
this result is also uncertain
it could be a false negative
it could be in a language for example
it could be that it actually is a cat
and we've just failed to identify it as a cat
so here we don't think there's an effect
due to our product intervention
but in reality there is and we just haven't
haven't identified the existence of that effect
so this second type of uncertainty
this
this type 2 error or or that rate of false negatives
there's sort of three ways
three levers we have to help quantify and manage
and mitigate that uncertainty
the first is effect size
and by effect size I mean
the difference in the metric value between the control
experience and the test experience
and the core intuition here is
the larger
the difference in the metrics between these two
experience experiences
the easier it is for us to
correctly identify that there is a difference

so it kind of reduces that false negative rate
so in the in the context of product innovation uh
one of the things my my team will talk about is
if you're going to test an idea
push the limits a little bit test big bold ideas
if you have a hypothesis
and this is not a great hypothesis
that increasing the size of the Netflix ribbon
logo will increase engagement
don't just increase it by a little bit
like really max out the size of that logo
in hopes of getting a bigger difference
between your control and treatment experiences
so
we can have a better chance of correctly identifying
if there is a difference
so larger differences lead to easier
more reliable detection
second is sample size basically
the number of members we put into each test cell
or test experience
and here larger samples lead to easier
more reliable detection
it's just more people we put into tests
the smaller the effects that we can correctly identify
so we talk a lot internally
you know how many people
how many members should we use in this
this experiment
given the type of effect sizes that we expect
and then the third
which is something my team does a lot of work on
is to think about variability
how disparate versus consistent are the metrics
among the population participating in the experiment
it's when the contact of Netflix
we know some of our members are super heavy streamers
super high users of our platform
some are far lighter users of our service
and then there's a bunch in the middle
so we can shift our thinking from hey
you know let's just
let's just look at all of our members
in the true experience
all of our members in the treatment experiences
and compare some average to like a modeling approach
where we say okay
not all of our members are the same
going into this experience
some are heavy streamers some are light

some are medium
let's look at changes in each of those buckets
so those sort of statistical modeling approaches
I can help us reduce variants and
and help us with reliable detection
and we use these 3 knobs
to reduce the occurrence of these false negatives
where again
your your little mental image of the false negative
is the cat saying it's not a cat
uh
and in the the the stats parlance
this is called statistical power
statistical power
is the probability that we correctly identify
true effects we aim to be able to
to correctly identify those true effects
most of the time
and then second we
we need to choose a tolerance level
for these false positives
so we just talk a little bit about false negatives
and power now
we have to talk about false positives and significance
so we measure statistical significance
and we accept that general about 5% of the time
statistically significant results are just noise
so this is that five magic arbitrary 5% number
if something has less than a 5% chance of
of occurring through chance alone
then then we'll sort of say hey
we're willing to say that
that statistically significant
there's a trade off here
between controlling false positives
and false negatives like I said
we're not really eliminating uncertainty
we're just managing and understanding the uncertainty
from our tests
so how do we interpret statistical significance
well here's here's some some some made up numbers uh
that it sort of shows you the results
you might get from a test
we have our version a our version B
we have some metric we care about uh
in version a it's 87.2 and then version B we see a 0
0.7 increase in the value of that metric
then we have something called the p value
which here is 0.026 the p value means that the um
the probability of seeing a result

at least as extreme as this point
seven difference
that we observe in this test is only 0.026
and because we rely on that 5% value
for statistical significance we would conclude
that there is a statistically significant difference
between these experiences
and that the reason this product
version B is doing better
I'm sorry we conclude that the
the test treatment product
version B is the reason why this metric has increased
okay
so as we wrap up this
this more detailed overview of testing
I wanna end by saying that statistics help
helps us reduce and understand error rates and
and generally make good decisions
in the face of uncertainty
we don't eliminate uncertainty
and we don't eliminate the problem
the possibility that we might make a mistake
can feel a little bit uncomfortable
um and in particular
there's no way to know
whether the results of a specific experiment
is a false positive
if it comes out as a positive result
or if you were a false negative
if it comes out as a negative result
so what we try to do is mix the statistical results
with a fair amount of judgment
as as we think about making decisions
so do the results and here are some questions that
that help bring judgment to the interpretation of test
results so do the results align with the hypothesis
let's go back to that mobile preview example
if we found some big increase in engagement
when we we expose members to that feature
we actually find that
none of our members are interacting with that feature
that might give us pause
so that sort of getting it does
does the metric story hang together
can we
sort of measure the steps through that causal chain
from the product intervention to these
these core metrics that we care about
we also look for supporting or refuting evidence
across cells

we might test several variants of that mobile
previous experience they all consistently give
give us
show a similar response in the metrics will have more
gives us more more reason to believe that we've
we've truly improved or changed member behavior
and finally
do do results repeat if we rerun the experiment
do we get a similar result
so Netflix we run a lot of experiments um
in fact
test results are expected for for most decisions
as a result we've invested a great deal of resource
into an internal platform
to support our experimentation program call it XP
blazes are fronted UI some of the materials
I've talked about
today are from our intro to experimentation
internal class really
this is about uh democratizing access
and contributions to experimentation
so
I just wanna spend a few minutes here before I wrap up
talking about our experimentation platform
in the role that my team plays
so our experimentation platform
I think of as a truly interdisciplinary collaboration
the collaboration between
bunch of different types of engineer uh
between data scientists and statisticians
there's a major numerical computing piece
that allows us to do statistics on the
the sizes of tests that we run
and the number of tests that we run
and finally
there's a product design and product management piece
as we think about our platform as an internal tool
and internal product within the company
and there's three pillars that our platform relies on
first is trustworthyness
if any data shows up on our platform
we need to ensure that it's reliable and trustworthy
so our decision makers can confidently use that data
to inform product decisions
second we talk a lot on my team about inclusivity
we're serving a lot of different audiences
from executives product managers
data scientists engineers who run tests
data scientists
who use languages that are different from engineers

we need to be inclusive of all of these disciplines
for this to work the third is scalability again
we went a lot of experiments
many of them very large and today for a few minutes
I just wanna talk about scalability
and how it interacts with this
theme of democratization
and here specifically how
how we think as a platform team about scaling
the scientists that we support
the scientists who actually run experiments
interpret and interpret results
then how we scale decision makers
be they product managers be they
engineers
and sometimes some case the scientist themselves
so in scaling the scientist
we've made some very deliberate bets
as a platform team and in particular
we have thought about how to build a modular platform
that democratizes contributions
and access to to
to different elements
specifically we've disentangled this idea
these disentangled
the platform into these three distinct modules
the first is where you would go to define a new metric
that's important for your experiment
the second is where you would go to define
the statistical analysis
is required to analyze that metric and experiment
and then the third is is where you
would go to
define visualizations that are necessary to surface
the results of that statistical test
and then what's what's great
and what really links this to the theme of
democratization is that
not only can folks contribute to each of these modules
their contributions are then available to everyone
and as a result
many many possible analysis workflows are supported
you can kind of just wire these up pay for this test
here are the metrics I need
they're already defined
I want to wire them up to to this type of cause a model
and then wire the results up to these visualizations
I want that to flow through to both the internal UI
as well as a notebook environment
that gives me more flexibility

and finally in this is more on inclusivity
all of this can be done in our Python
to the language of data science
rather than the language of engineering
even though this is a pretty involved
engineering system in the background
so scientists can come
they can they can access the results of their tests
both in notebook environments like you're seeing here
very quickly and just a few lines of code
reproduce the same type of visualizations
that you actually see on our internal UI
but with more flexibility
okay and then briefly
how do we think about scaling decision makers
well the the first is kind of counterintuitive
we we scale decision makers by scaling scientists
so by by really emphasizing
an efficient and sciencecentric platform
one of the things we're trying to do
is build automation
deeper into the workflows of test analysts
so the metal image I have here is that
if our analysts and Netflix are doing like
run on the Hampshire wheel type of work
you know I've got to write some query
I've got to grab this data
I've got to use some package
and Python to fit some model
just like super super wrote stuff
we should be automating that
we should be taking that burden on from the platform
and really freeing up the analyst
to focus more on creative
problem solving
exploratory work to generate hypotheses
research projects like
the super
high value work that these data scientists can do
so I really think a big goal of our platform is
is workflow improvement and workflow automations
so anything that the data scientists
you know need to do more than once we take on
and second is through UI access
so on the right I showed that image before
this is the interactive notebook environment
that's more targeted at data scientists
then on the left is our blaze our internal UI
the goal here is to develop accessible and intuitive
you eyes that allow that that present our results

and permit for confident decision making
for a wide variety of different types of individual
different types of stakeholder
I think if I'm being honest here
we're a long way from being finished here
there's a lot of room for improvement
and how we're presenting these results
so when I think about democratizing
and this is our third theme
democratizing access
and contributions to an experimentation platform
to really help a scale
this is kind of the mental image I have
our platform directly support scientists
who then contribute back to the platform
in this uh
this positive feedback loop
that's a
directly support decision makers through our UIs
and then in success
because of these workflow improvements
that the platform delivers
we actually allow the scientists to scale
their support of the decision makers
often product managers or engineers
so in summary
not quite in summary
so this is our third theme about democratization
how we scale so
our platform level investments
allow scientists to tribute directly in power
a variety of decision makers
so in success no one is ever blocked by the platform
they are enabled by the platform to serve themselves
and to contribute back
okay so what we talk about today is decision making
an experimentation at Netflix um
democratization three ways
so the first is how we decide by using AB tests
we allow our members to vote with their actions
on how to evolve the product experience
second is what we tried
because we test and because most ideas are not winners
the really is this open marketplace for product
innovation ideas
and we see great ideas coming from
from a whole bunch of different types of people with
within the business
and the third how we scale
this is the result of our platform level investments

which allows scientists to contribute directly
really empower that large variety of decision makers
and that is all for me
thank you for your time and attention