PRODUCTION AS AN EXPERIMENT LAB

FROM TEST IN PRODUCTION TO EXPERIMENT IN PRODUCTION

WHO?

- Software Engineer
- Working on Data Science teams as the fool
- Exposed to “proper science”
- CTO of a startup that accidentally got big
- Shifted mentality from “i can fix it” to having to understand what engineers were doing

@rmn

fuzzbox

 hyperr

 hieroglyph

 sonar

 supercollider

 errorsmith

 echoplex

@rmn

WHAT ARE WE TALKING ABOUT

@rmn

TEST IN PROD

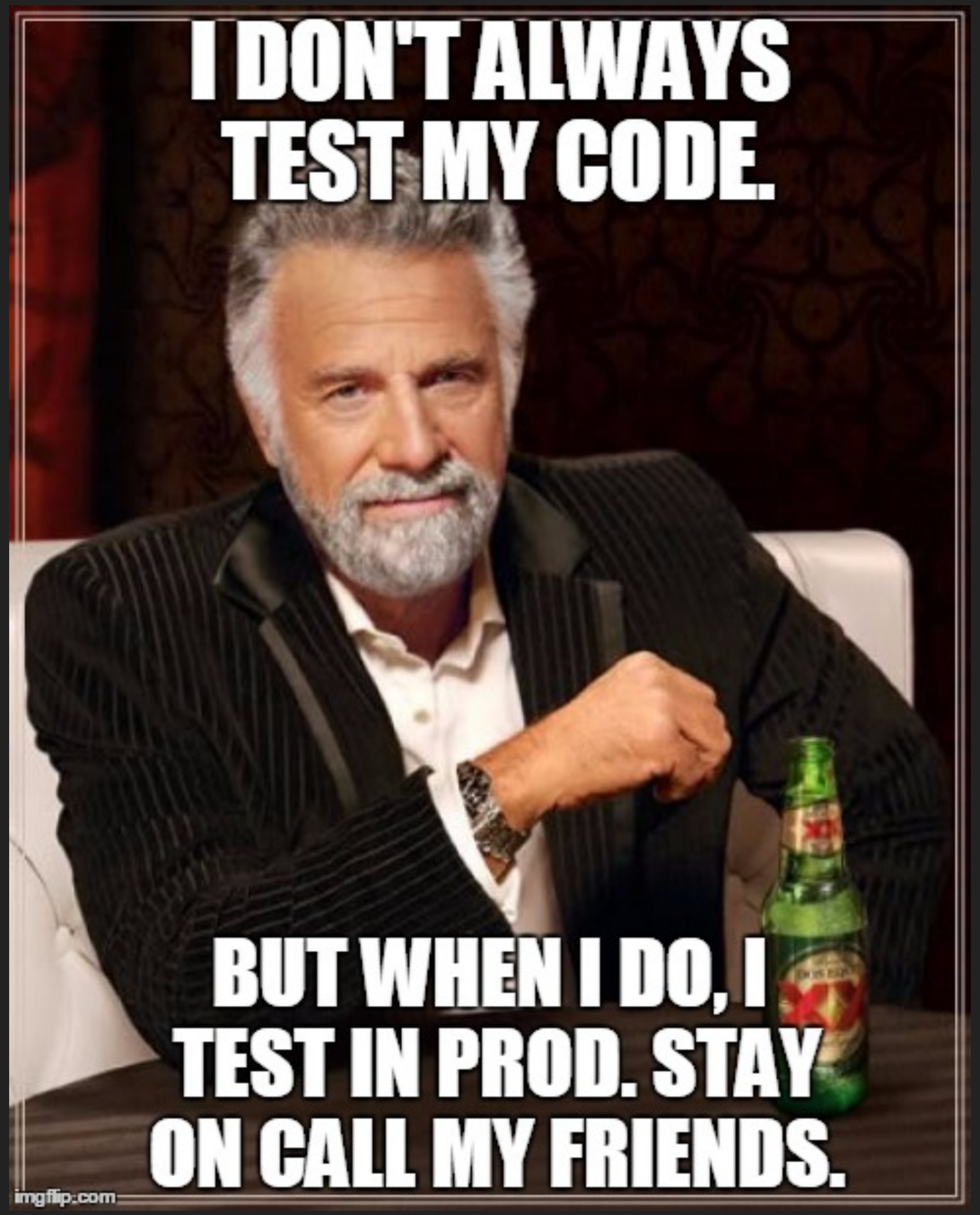
PROGRESSIVE DELIVERY

ERROR BUDGETS

@rmn

TEST IN PROD

@rmn


I DON'T ALWAYS
TEST MY CODE.
BUT WHEN I DO, I
TEST IN PROD. STAY
ON CALL MY FRIENDS.
imgflip.com

TEST IN PROD

- Stop: Go read/watch anything by Charity Majors (@mipsytipsy) and be enlightened
- Single handedly advanced this concept beyond a developer joke
- Attempting to clone production is foolish
- If you are small enough to clone, stay simple, if you are a big enough, attempting to clone production is foolish and waste of cycles
- “Real users, real traffic, real scale, real unpredictabilities”

@rmn

TEST IN PROD DOESN'T MEAN RELEASE WITHOUT TESTING

@rmn

STAGING IS JUST

“IT WORKS ON OUR MACHINE”

@rmn

TESTING IN PROD MEANS
EXTENDING THE SOFTWARE
DEVELOPMENT LIFECYCLE
BEYOND RELEASE

@rmn

“REAL USERS, REAL TRAFFIC, REAL SCALE, REAL UNPREDICTABILITIES”

@rmn

PROGRESSIVE DELIVERY

@rmn

"PROGRESSIVE DELIVERY IS CONTINUOUS DELIVERY WITH FINE-GRAINED CONTROL OVER THE BLAST RADIUS."

James Governor, RedMonk (@monkchips)

SEPARATE DEPLOY FROM RELEASE

@rmn

HANG ON

THIS LOOKS LIKE A FIELD EXPERIMENT

@rmn

ERROR BUDGETS

@rmn

2 THREATS TO AVAILABILITY

THE SOFTWARE CHANGES
THE ENVIRONMENT CHANGES

@rmn

MAYBE THE NATURAL
DISTRIBUTION OF FAILURE
HAS SPARED YOU

@rmn

YOU MIGHT HAVE SOME 9S TO PLAY WITH

@rmn

ULTIMATELY OUR JOB ISN'T (EXCLUSIVELY) SHIPPING CORRECT/ WORKING SOFTWARE

IT'S MAKING MONEY

@rmn

AVAILABILITY EXPERIMENTS

PERFORMANCE
COST
CAPACITY
FAULT INJECTION
NEW ARCHITECTURES

@rmn

NOW WE HAVE A VOCABULARY FOR PRODUCTION EXPERIMENTATION

@rmn

QUICK CONFESSION

I LIED

@rmn

FIELD EXPERIMENTS

SHIFT FROM CONTROLLED ENVIRONMENT TO EXCLUDABILITY AND NON INTERFERENCE

@rmn

STAGING

LAB

TEST

PRODUCTION FIELD EXPERIMENTS

@rmn

PRODUCTION FIELD EXPERIMENTS

@rmn

FROM EXPERIMENTS THE LAB TO FIELD EXPERIMENTS

- Starting to look like our reasons for testing in prod
- Staging to prod is the same as lab to field
- We are in the business of increasing value
- This means safely trying things, inside your error budget
- Software just happens to be our medium currently
- Why wouldn’t you experiment in production?

@rmn

WHAT IS THE DIFFERENCE BETWEEN A TEST AND AN EXPERIMENT?

@rmn

TEST

VERIFY

CONFIRM

INTEGRATION

EXPERIMENT

DISCOVER

LEARN

ONGOING CHANGE

@rmn

EXPERIMENTS LEAD TO
NEW KNOWLEDGE
TESTS DON'T

@rmn

EXPERIMENTS AREN'T JUST CHANGING STUFF

ORCHESTRATING VALID EXPERIMENTS IS HARD

@rmn

CHAOS EXPERIMENTS

HOW I MIGHT EXPERIMENT WITH A DIFFERENT TYPE OF HAT

@rmn

EXPERIMENTS AREN'T JUST CHANGING STUFF

ORCHESTRATING VALID EXPERIMENTS IS HARD

REALLY HARD

@rmn

TO CALL IN THE STATISTICIAN AFTER THE EXPERIMENT IS DONE MAY BE NO MORE THAN ASKING HIM TO PERFORM A POST-MORTEM EXAMINATION: HE MAY BE ABLE TO SAY WHAT THE EXPERIMENT DIED OF.

– Ronald Fisher

CHALLENGES OF PERFORMING GOOD EXPERIMENTS

- Validity (does this test what I think it tests)
- Bias (this one weird trick)
- Hawthorne effect (modifying behavior because I am in test)
- Self fulfilling prophecy (IQ - Rosenthal and Jacobson)
- Contamination (i’m gonna keep looking)
- Complicated (we’re running many overlapping experiments)
- Follow on experiments

@rmn

HOW CAN WE DO THIS WITH SYSTEMS

THINK ABOUT A REQUEST AND HOW IT MOVES THROUGH YOUR SYSTEM AS A FOUNDATION TO ORCHESTRATE EXPERIMENTS AND MEASURE THEM

@rmn

HOW NOT TO RUN AN EXPERIMENT

- A/B testing of search scores results
- Switching in code based on state assigned to user in db
- Bias in choosing a population
- Indexes double the data
- Tested a performance change as well as ranking algorithm change
- INVALID experiment

@rmn

SCENARIO A: AFFILIATE E COMMERCE

- System generating millions of dollars. No one from original team around. Why was it so provisioned.
- Deterministic assignment to segment via randomization of parameters/units
- Ability to rebuild segment deterministically and with diff population
- Small population
- Route to unique variant
- Observe behavior for weeks or months and be able to reconstruct and effectively replay based on historical samples of requests
- Observe traces via proxy from different infra population to see changes in shape of successful request and remove dependencies

@rmn

SUCCESS: SMALLER, ITERATIVE EXPERIMENTS

- #1 Smaller cluster
- #2 observe audience and proxy to gather trace data
- #3 expose new audiences to new code

- For each, resample old population and compare distribution of behavior

@rmn

SCENARIO B: ARCHITECTURAL CHANGE, LONG RUNNING

- Situation: Pricing data very dynamic, complexity cost penalty for high consistency
- We want to expose some population of our users to minimally stale data
- Classify a population based into groups based on actions from trace, then parameterizing based on distribution of trace data based actions
- Challenge: value of a customer understood over months
- Is it up and fault tolerant? Might be yes, but are we losing money very slowly though?
- Understand this in terms of the metrics of our 2 sided marketplace, which means the unit economics of our business (new signups, acquired customers, those at risk of churn)
- Why? Can we reduce cost and not reduce revenue

@rmn

WAIT: MULTI ARM BANDITS

- Usually have to let an experiment run to collect enough exposures of a segment to a variant to be significant
- Reinforcement Learning: Bandit algorithm can self optimize the variants, to find a result faster. Resulting in shorter experiments
- Lets think about Chaos Engineering again…

SCENARIO C: MULTI ARM BANDIT SICK CANARY

- automated chaos, slow, gradual fault injection
- A/B testing in reverse, tests optimize for worst case scenario
- Introduce variants with increasing latency, measure against steady state metrics, if you survive, bootstrap a new variant
- Find the breaking point much quicker than a single experiment
- Much quicker time to value
- Experiments need to run in less time to collect significant data

SUMMARY

- Production is just begging for field experiments
- We can apply this whole methodology to pushing the SDL past deployment and really take progressive delivery to a new place of progressive experimentation
- Almost completely unexplored new land of tooling and practices
- Thank you to the pioneers
- Thank you to you!

@rmn

IF IT WAS COMPLETELY SAFE TO EXPERIMENT IN PRODUCTION, WHY WOULDN'T YOU?

@rmn

FURTHER READING

- https://corecursive.com/019-test-in-production-with-charity-majors/
- https://opensource.com/article/17/8/testing-production
- https://www.infoq.com/presentations/testing-production-2018
- https://redmonk.com/jgovernor/2018/08/06/towards-progressive-delivery/#comment-2241326
- https://medium.com/@njones_18523/chaos-engineering-traps-e3486c526059
- https://en.wikipedia.org/wiki/Field_experiment
- https://www.youtube.com/watch?v=NU-fTr-udZg

@rmn