

# QUALITY ASSESSMENT IN PRODUCTION

Christian Kaestner

Required Reading:

- Hulten, Geoff. "[Building Intelligent Systems: A Guide to Machine Learning Engineering.](#)" Apress, 2018, Chapters 14 and 15 (Intelligence Management and Intelligent Telemetry).

Suggested Readings:

- Alec Warner and Štěpán Davidovič. "[Canary Releases.](#)" in [The Site Reliability Workbook](#), O'Reilly 2018
- Kohavi, Ron, Diane Tang, and Ya Xu. "[Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.](#)" Cambridge University Press, 2020.



Changelog  
@changelog



“Don’t worry, our users will notify us if there’s a problem”



2:03 PM · Jun 8, 2019



[Read the full conversation on Twitter](#)



2.2K



Reply



Copy link

[Read 12 replies](#)

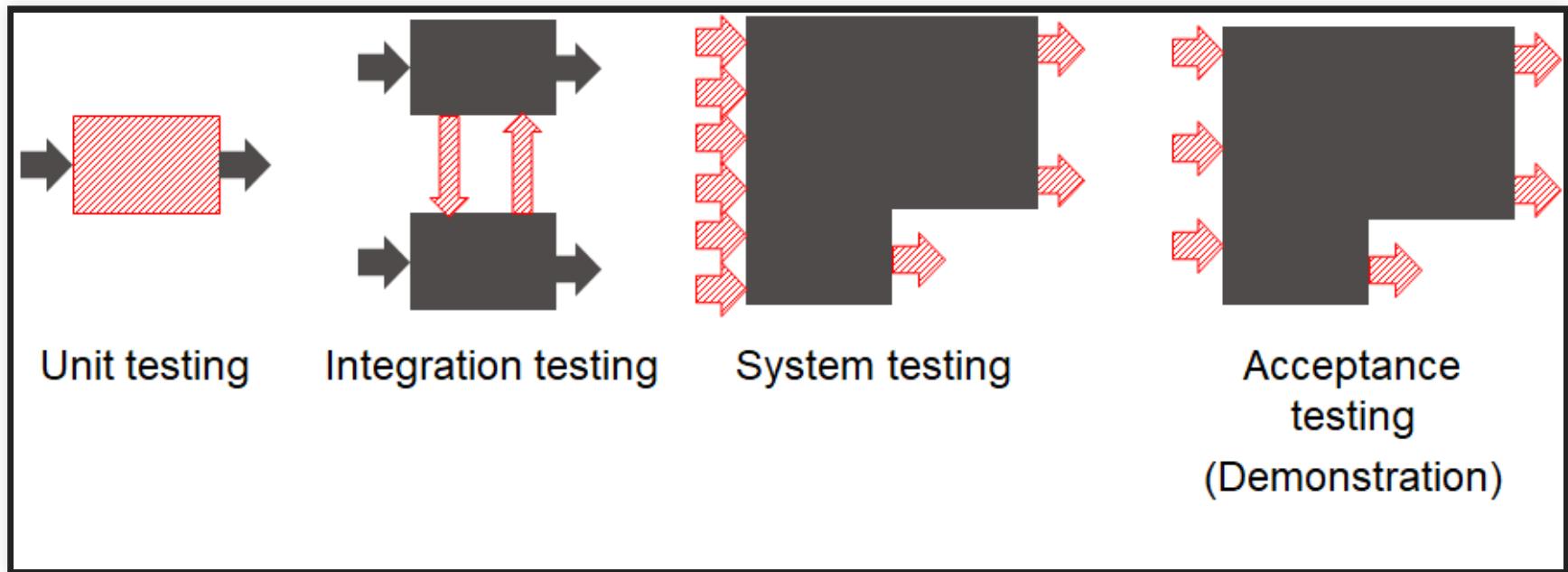
# LEARNING GOALS

- Design telemetry for evaluation in practice
- Understand the rationale for beta tests and chaos experiments
- Plan and execute experiments (chaos, A/B, shadow releases, ...) in production
- Conduct and evaluate multiple concurrent A/B tests in a system
- Perform canary releases
- Examine experimental results with statistical rigor
- Support data scientists with monitoring platforms providing insights from production data

# FROM UNIT TESTS TO TESTING IN PRODUCTION

*(in traditional software systems)*

# UNIT TEST, INTEGRATION TESTS, SYSTEM TESTS



## Speaker notes

Testing before release. Manual or automated.

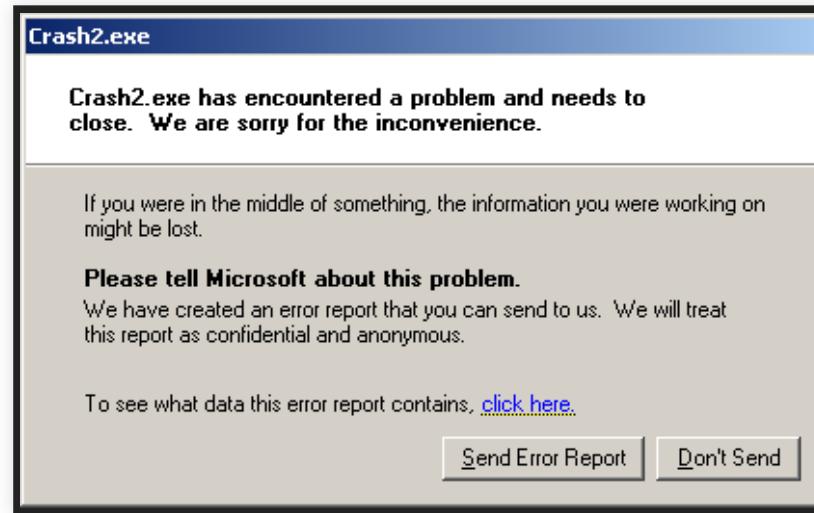
# BETA TESTING



## Speaker notes

Early release to select users, asking them to send feedback or report issues. No telemetry in early days.

# CRASH TELEMETRY



## Speaker notes

With internet availability, send crash reports home to identify problems "in production". Most ML-based systems are online in some form and allow telemetry.

# A/B TESTING

Original: 2.3%

The Groove website landing page features a large image of a smiling man in a plaid shirt. Text on the left reads: "SaaS & eCommerce Customer Support. Managing customer support requests in Groove is so easy. Way better than trying to use Gmail or a more complicated help desk." A quote from "Griffin, Customer Champion at Allocacoo" is included. Below the image are four navigation links: "How it works", "What you get", "What it costs", and "How we're different". A green button says "Learn More >". At the bottom, a call-to-action states: "You'll be up and running in less than a minute."

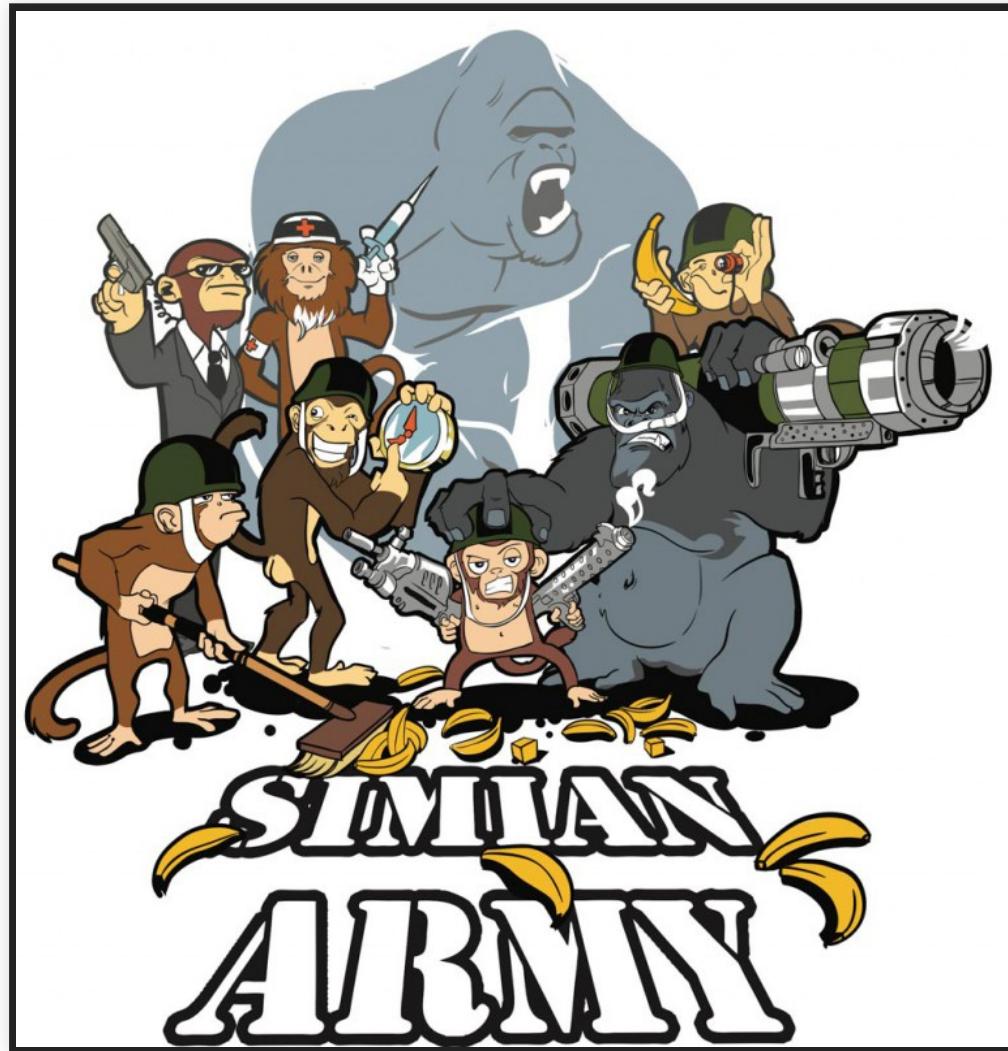
Long Form: 4.3%

The Groove website landing page features a large video player showing a man named Allan speaking. To the right, text reads: "Everything you need to deliver awesome, personal support to every customer." Below the video, a testimonial says: "Assign support emails to the right people, feel confident that customers are being followed up with and always know what's going on." A sidebar lists "WHAT YOU'LL DISCOVER ON THIS PAGE" with several bullet points. At the bottom, logos for "BuySellAds", "HUBSPOT", "METALAB", and "StatusPage.io" are displayed.

## Speaker notes

Usage observable online, telemetry allows testing in production. Picture source: <https://www.designforfounders.com/ab-testing-examples/>

# CHAOS EXPERIMENTS



## Speaker notes

Deliberate introduction of faults in production to test robustness.

# MODEL ASSESSMENT IN PRODUCTION

Ultimate held-out evaluation data: Unseen real user data

# LIMITATIONS OF OFFLINE MODEL EVALUATION

- Training and test data drawn from the same population
  - **i.i.d.: independent and identically distributed**
  - leakage and overfitting problems quite common
- Is the population representative of production data?
- If not or only partially or not anymore: Does the model generalize beyond training data?

# IDENTIFY FEEDBACK MECHANISM IN PRODUCTION

- Live observation in the running system
- Potentially on subpopulation (A/B testing)
- Need telemetry to evaluate quality -- challenges:
  - Gather feedback without being intrusive (i.e., labeling outcomes), without harming user experience
  - Manage amount of data
  - Isolating feedback for specific AI component + version

# DISCUSS HOW TO COLLECT FEEDBACK

- Was the house price predicted correctly?
- Did the profanity filter remove the right blog comments?
- Was there cancer in the image?
- Was a Spotify playlist good?
- Was the ranking of search results good?
- Was the weather prediction good?
- Was the translation correct?
- Did the self-driving car break at the right moment? Did it detect the pedestrians?



## Speaker notes

### More:

- SmartHome: Does it automatically turn off the lights/lock the doors/close the window at the right time?
- Profanity filter: Does it block the right blog comments?
- News website: Does it pick the headline alternative that attracts a user's attention most?
- Autonomous vehicles: Does it detect pedestrians in the street?

Skype for Business

## How was the call quality?

Good

**Audio Issues**

- Distorted speech
- Electronic feedback
- Background noise
- Muffled speech
- Echo

**Video Issues**

- Frozen video
- Pixelated video
- Blurry image
- Poor color
- Dark video

blog post demo

Privacy Statement

Submit Close

Matt Millman  
Because I'm happy 😊

Settings

Help and feedback

Report a problem

RECENT CHATS

Besties 10/10/2018

EN Elena Nilsson, Anna Davie... 7/27/2018  
It was great talking to all of ...

Anna Davies 6/26/2018  
coffee awaits!

Maarten Smenk 5/25/2018  
Missed call

MS Maarten Smenk, Anna Davie... 5/21/2018  
Hi, happy Monday!

## Speaker notes

Expect only sparse feedback and expect negative feedback over-proportionally

A screenshot of a flight search interface. At the top, there's a green line graph icon followed by the text "DFW ↔ SFO" and "Nov 16". Below this, it says "1659 of 1687 flights" and "Wednesday". A red oval highlights a yellow callout box containing the following text:

**Prices may fall within 7 days – Watch**

Our model strongly indicates that fares will fall during the next 7 days. This forecast is based on analysis of historical price changes and is not a guarantee of future results.

The interface includes a "Create a price alert" button, a "Stops" section with checkboxes for "nonstop", "1 stop", and "2+ stops" (all checked), and a "Times" section with a dropdown menu showing "Take-off Dallas" and "Arrival San Francisco".

## Speaker notes

Can just wait 7 days to see actual outcome for all predictions

A screenshot of a transcription software interface. At the top, there's a header with the project name 'the-changelog-318', a link to 'Dashboard', and a 'Quality' setting at 'High'. To the right are buttons for 'Last saved a few seconds ago', three dots for more options, and a yellow 'Share' button. Below the header is a timeline bar with markers at 00:00, Offset, 00:00, and 01:31:27. Underneath the timeline are four buttons: 'Play', 'Back 5s', '1x Speed', and 'Volume'. The main area contains transcribed text.

## NOTES

Write your notes here

## Speaker 5 ► 07:44

Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to measure a, the amount of time I did data entry from my father's business and I was on windows at the time and there wasn't a function called time dot [inaudible] time, uh, which I needed to parse dates to get back to time, top of representation, uh, I figured out a way to do it and I gave it to what's called the python cookbook because it just seemed like something other people could use. So it was just trying to be helpful. Uh, subsequently I had to figure out how to make it work because I didn't really have to. Basically, it bothered me that you had to input all the locale information and I figured out how to do it over the subsequent months. And actually as a graduation gift from my Undergrad, the week following, I solved it and wrote it all out.

## Speaker 5 ► 08:38

And I asked, uh, Alex Martelli, the editor of the Python Cookbook, which had published my original recipe, a, how do I get this into python? I think it might help

How did we do on your transcript?

## Speaker notes

Clever UI design allows users to edit transcripts. UI already highlights low-confidence words, can

# MANUALLY LABEL PRODUCTION SAMPLES

Similar to labeling learning and testing data, have human annotators



# SUMMARY: TELEMETRY STRATEGIES

- Wait and see
- Ask users
- Manual/crowd-source labeling, shadow execution
- Allow users to complain
- Observe user reaction

# BREAKOUT: DESIGN TELEMETRY IN PRODUCTION

Discuss how to collect telemetry (Wait and see, ask users, manual/crowd-source labeling, shadow execution, allow users to complain, observe user reaction)

Scenarios:

- Front-left: Amazon: Shopping app feature that detects the shoe brand from photos
- Front-right: Google: Tagging uploaded photos with friends' names
- Back-left: Spotify: Recommended personalized playlists
- Back-right: Wordpress: Profanity filter to moderate blog posts

(no need to post in slack yet)

# MEASURING MODEL QUALITY WITH TELEMETRY

- Three steps:
  - Metric: Identify quality of concern
  - Telemetry: Describe data collection procedure
  - Operationalization: Measure quality metric in terms of data
- Telemetry can provide insights for correctness
  - sometimes very accurate labels for real unseen data
  - sometimes only mistakes
  - sometimes delayed
  - often just samples
  - often just weak proxies for correctness
- Often sufficient to *approximate* precision/recall or other model-quality measures
- Mismatch to (static) evaluation set may indicate stale or unrepresentative data
- Trend analysis can provide insights even for inaccurate proxy measures

# BREAKOUT: DESIGN TELEMETRY IN PRODUCTION

Discuss how to collect telemetry, the metric to monitor, and how to operationalize

Scenarios:

- Front-left: Amazon: Shopping app feature that detects the shoe brand from photos
- Front-right: Google: Tagging uploaded photos with friends' names
- Back-left: Spotify: Recommended personalized playlists
- Back-right: Wordpress: Profanity filter to moderate blog posts

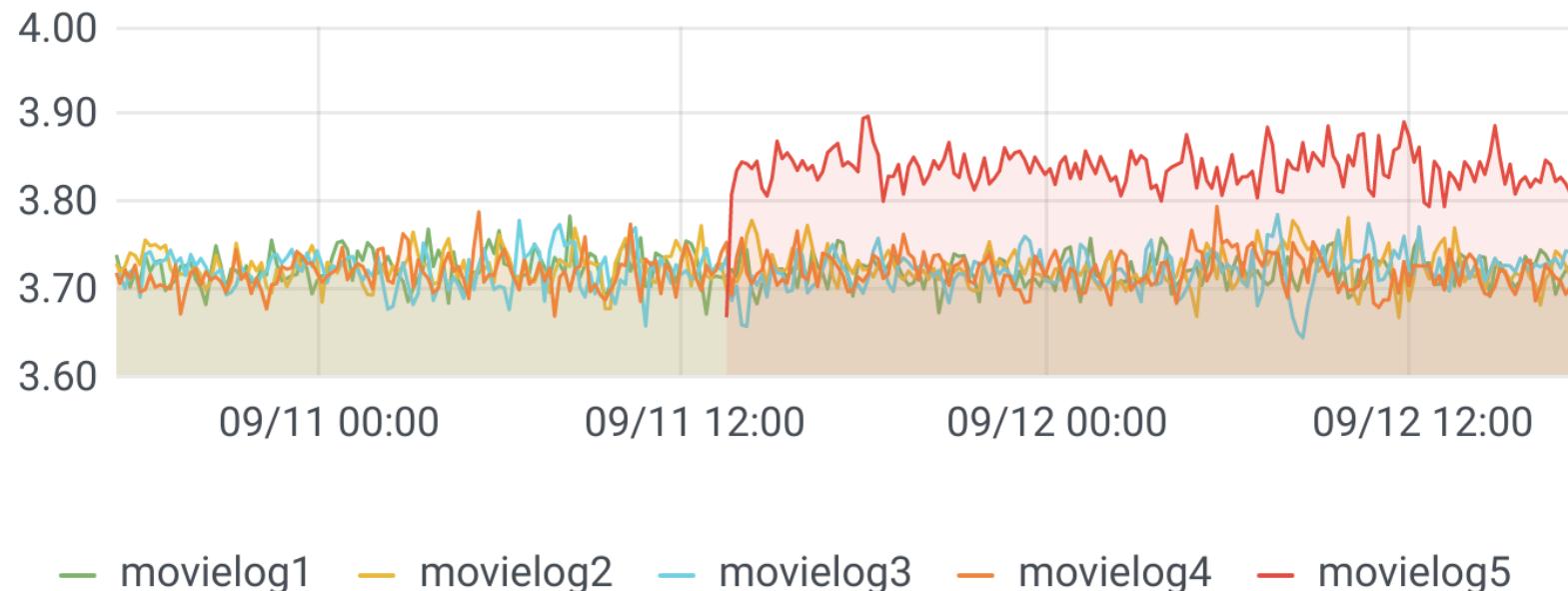
Post in slack in #lecture:

- Data to collect:
- Quality metric:
- Operationalization:
- AndrewId:

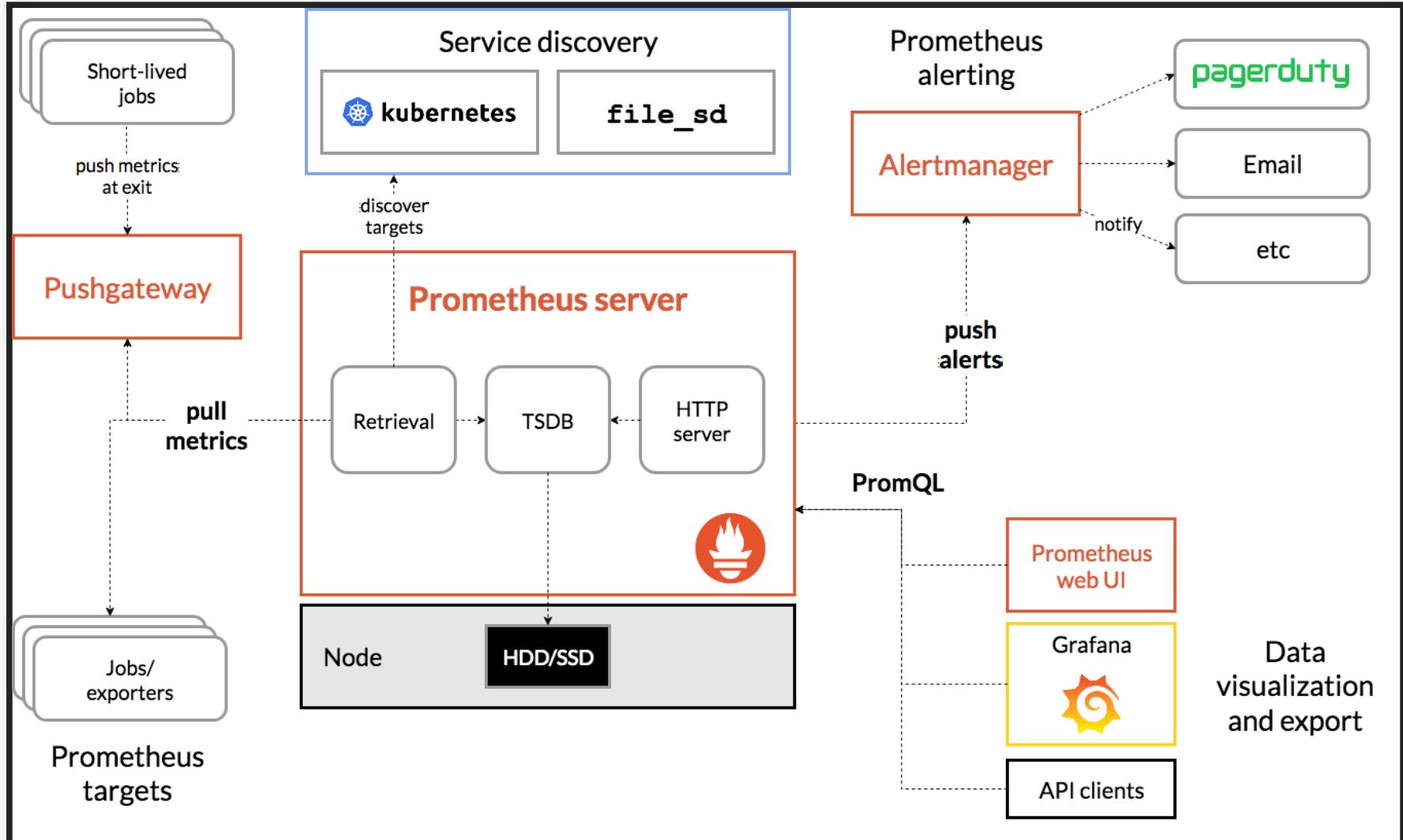
# MONITORING MODEL QUALITY IN PRODUCTION

- Monitor model quality together with other quality attributes (e.g., uptime, response time, load)
- Set up automatic alerts when model quality drops
- Watch for jumps after releases
  - roll back after negative jump
- Watch for slow degradation
  - Stale models, data drift, feedback loops, adversaries
- Debug common or important problems
  - Monitor characteristics of requests
  - Mistakes uniform across populations?
  - Challenging problems -> refine training, add regression tests

## Average rating last 15min



# PROMETHEUS AND GRAFANA







Website Overview

Zoom Out

Last 3 hours



Logins

**190**

Sign ups

**269**

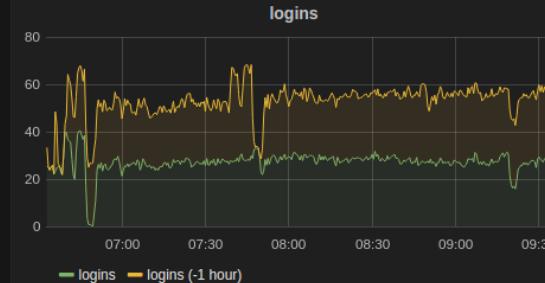
Sign outs

**273**

Memory / CPU



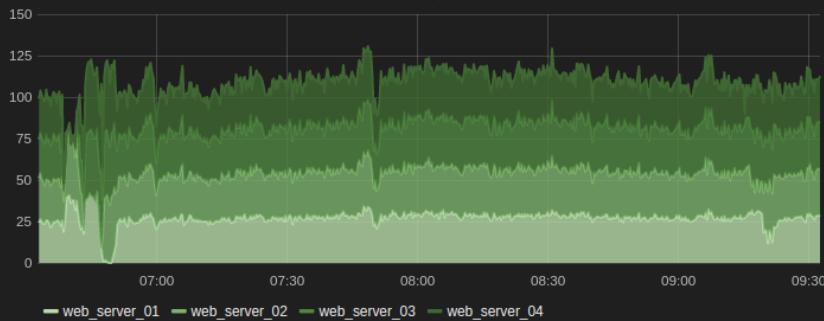
logins



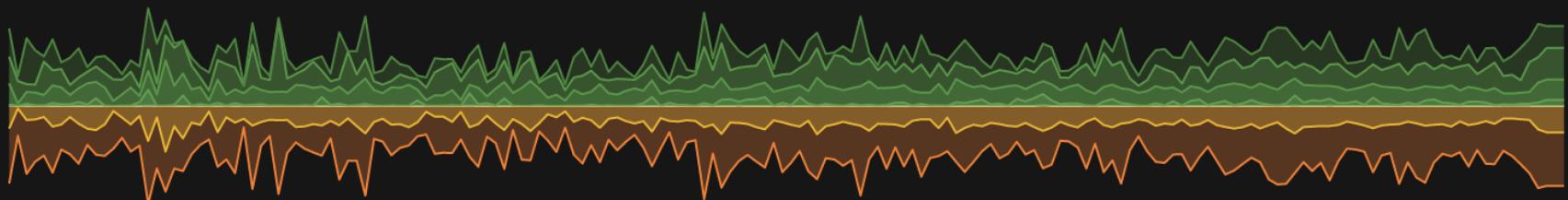
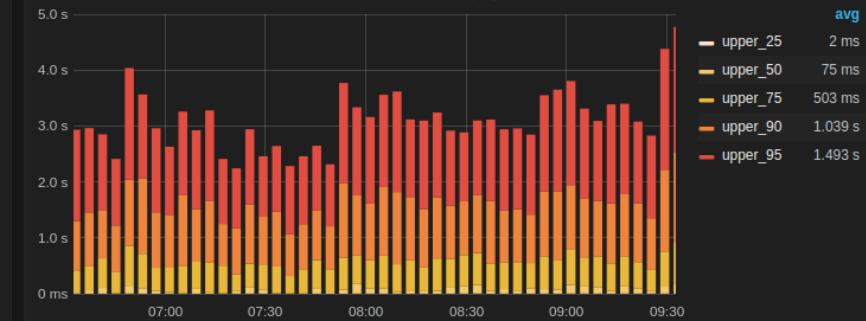
Memory / CPU



server requests

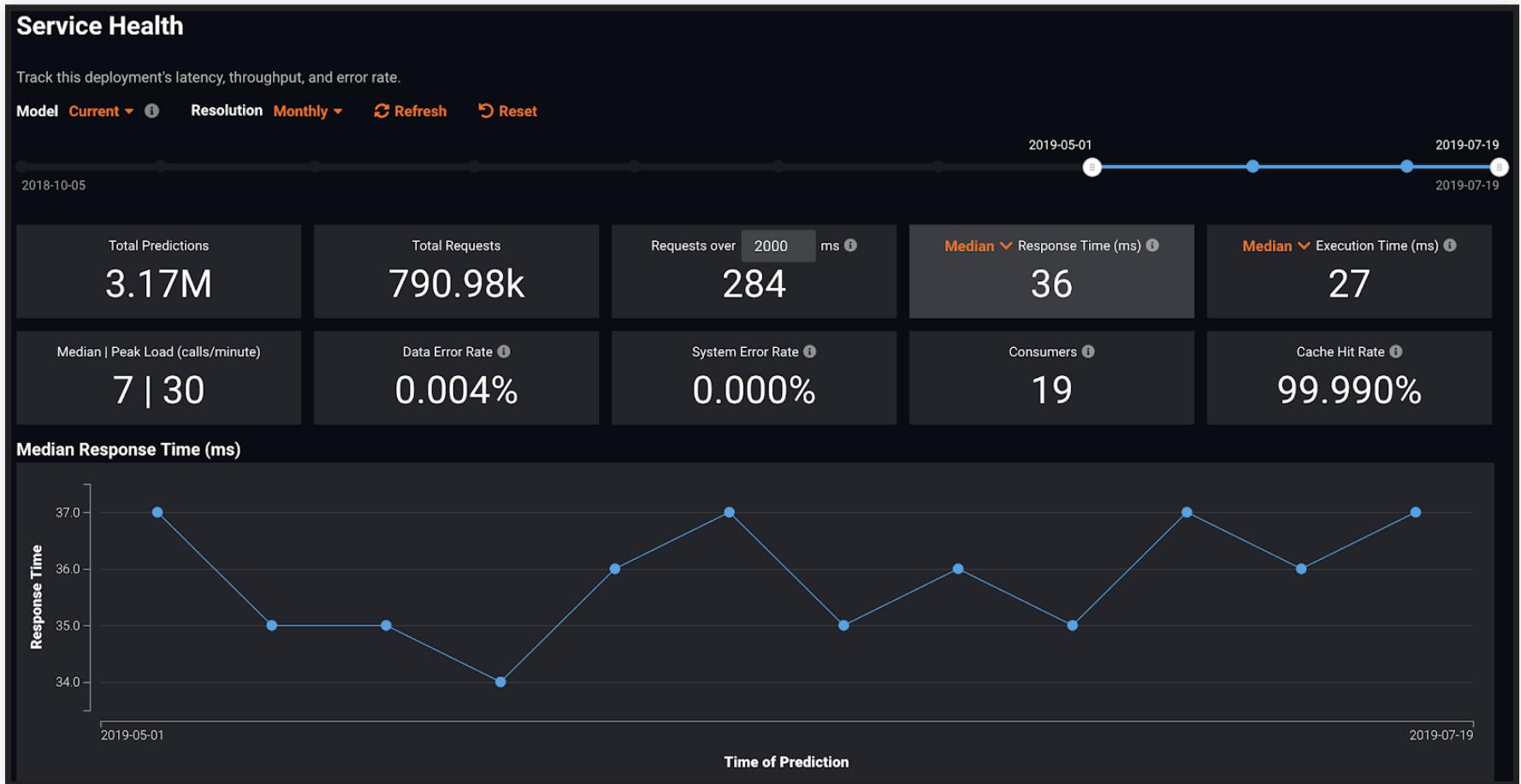


client side full page load





# MANY COMMERCIAL SOLUTIONS



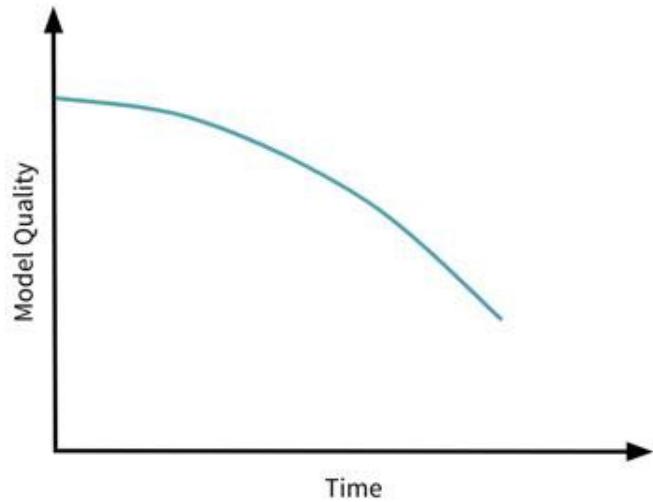
e.g. <https://www.datarobot.com/platform/mlops/>

Many pointers: Ori Cohen "Monitor! Stop Being A Blind Data-Scientist." Blog 2019



# DETECTING DRIFT

Static models



Refreshed models

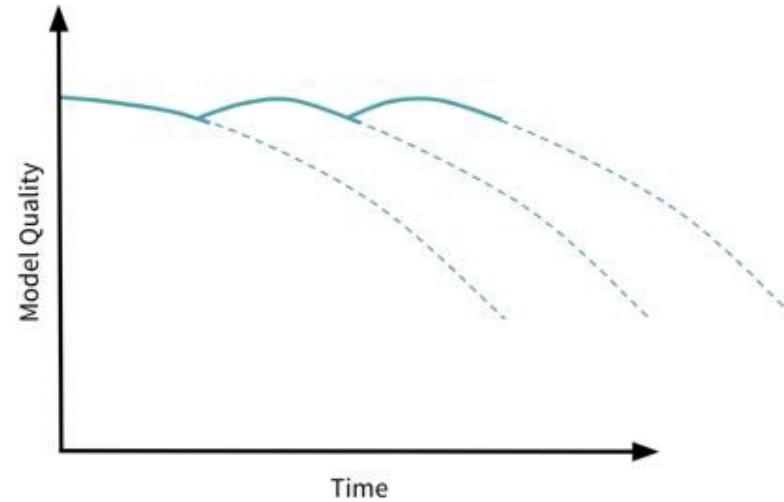


Image source: Joel Thomas and Clemens Mewald. [Productionizing Machine Learning: From Deployment to Drift Detection](#). Databricks Blog, 2019

# ENGINEERING CHALLENGES FOR TELEMETRY

TRENDING

Buying Guides

Note 10

Best Laptops

iOS 13

Best Phones

## Amazon Alexa stores voice recordings for as long as it likes (and shares them too)

By Olivia Tambini 21 days ago Digital Home

A letter from Amazon reveals all



# ENGINEERING CHALLENGES FOR TELEMETRY

- Data volume and operating cost
  - e.g., record "all AR live translations"?
  - reduce data through sampling
  - reduce data through summarization (e.g., extracted features rather than raw data; extraction client vs server side)
- Adaptive targeting
- Biased sampling
- Rare events
- Privacy
- Offline deployments?

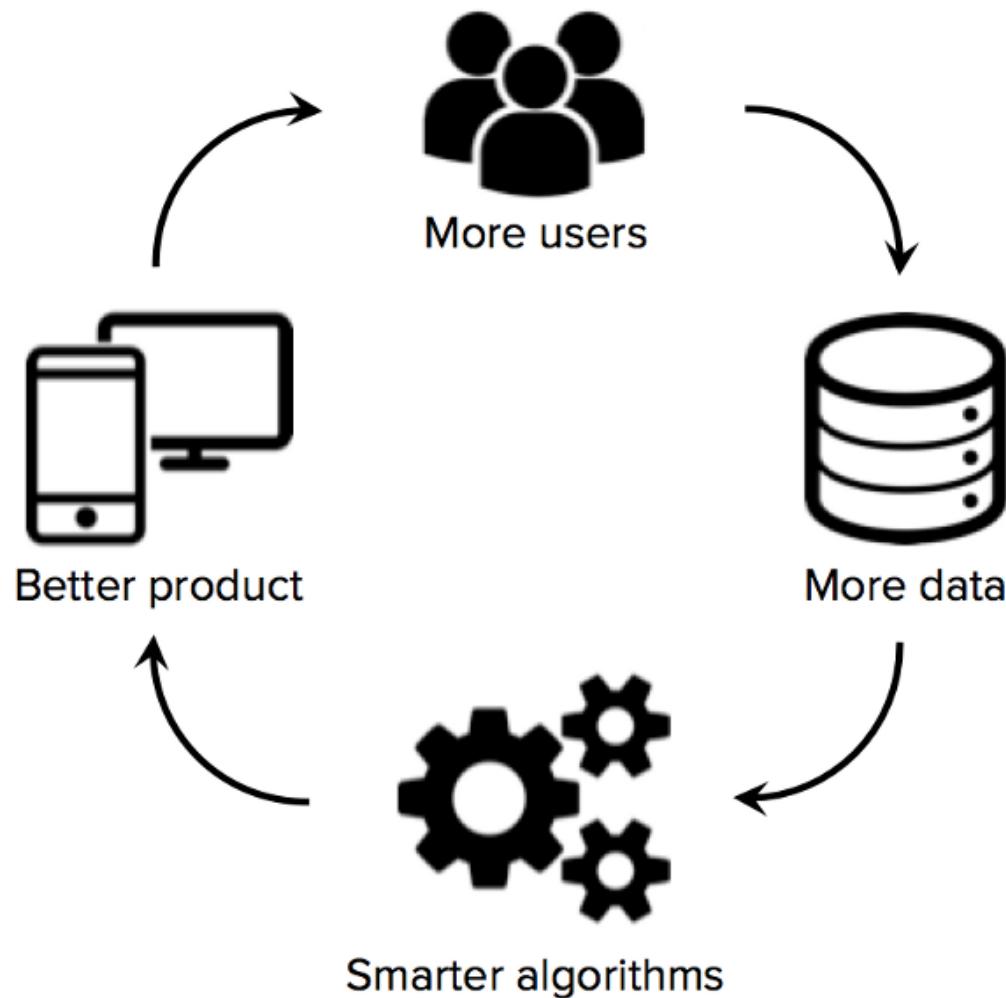
# BREAKOUT: ENGINEERING CHALLENGES IN TELEMETRY

Discuss: Cost, privacy, rare events, bias

Scenarios:

- Front-left: Amazon: Shopping app feature that detects the shoe brand from photos
- Front-right: Google: Tagging uploaded photos with friends' names
- Back-left: Spotify: Recommended personalized playlists
- Back-right: Wordpress: Profanity filter to moderate blog posts  
  
(can update slack, but not needed)

# TELEMETRY FOR TRAINING: THE ML FLYWHEEL



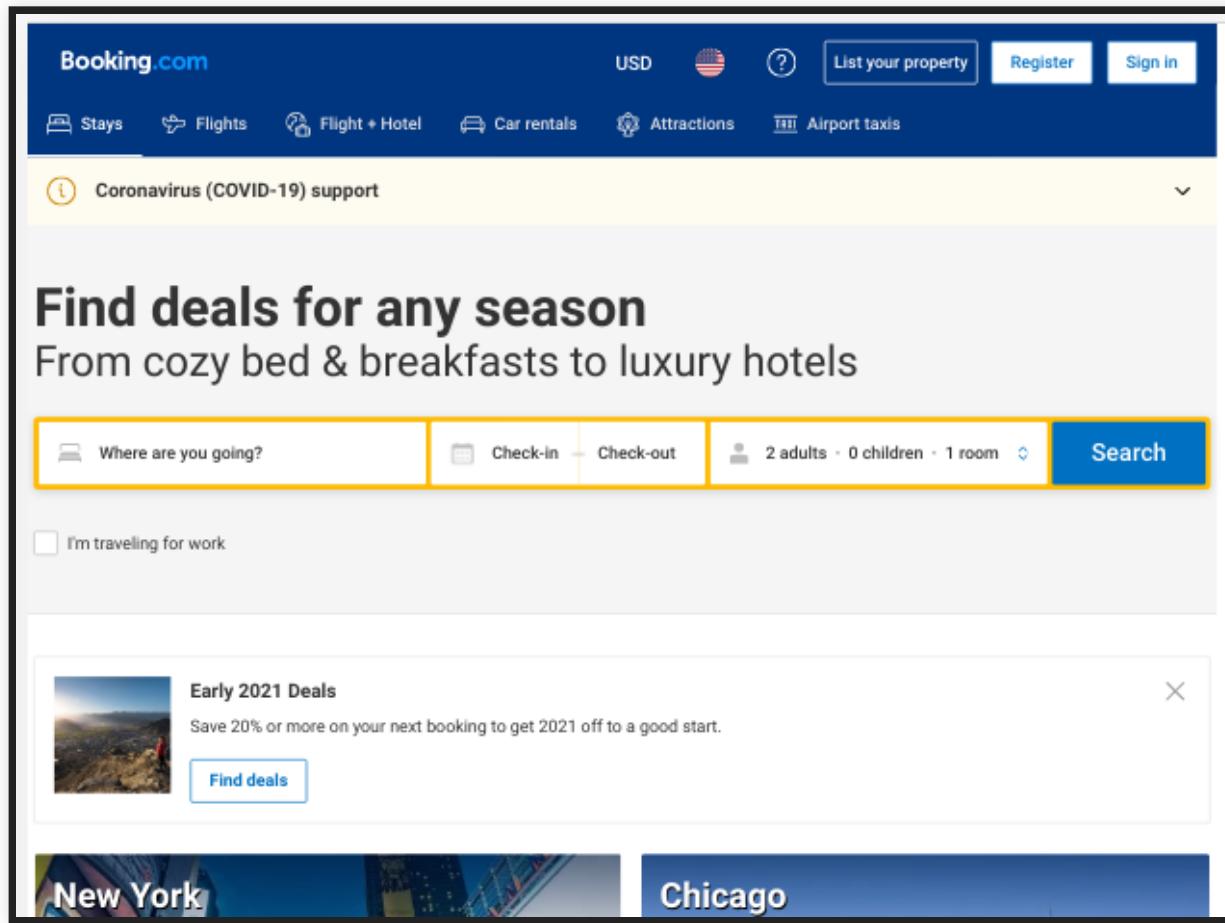
graphic by [CBInsights](#)

# MODEL QUALITY VS SYSTEM GOALS

# MODEL QUALITY VS SYSTEM GOALS

- Telemetry can approximate model accuracy
- Telemetry can directly measure system qualities, leading indicators, user outcomes
  - define measures for "key performance indicators"
  - clicks, buys, signups, engagement time, ratings
  - operationalize with telemetry

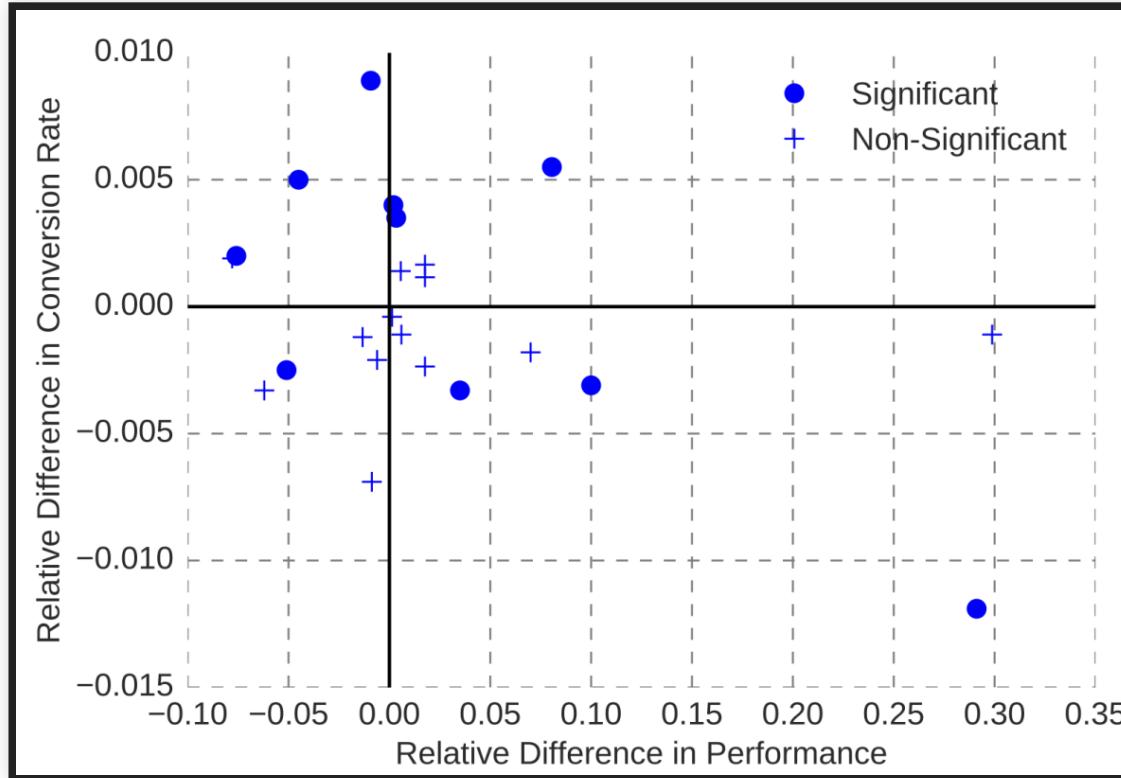
# MODEL QUALITY VS SYSTEM QUALITY



Bernardi, Lucas, Themistoklis Mavridis, and Pablo Estevez. "150 successful machine learning models: 6 lessons learned at Booking.com." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1743-1751. 2019.



# MODEL QUALITY VS SYSTEM QUALITY



Possible causes?

Bernardi et al. "150 successful machine learning models: 6 lessons learned at Booking.com." In Proc KDD, 2019.

## Speaker notes

hypothesized

- model value saturated, little more value to be expected
- segment saturation: only very few users benefit from further improvement
- overoptimization on proxy metrics not real target metrics
- uncanny valley effect from "creepy AIs"

# BREAKOUT: DESIGN TELEMETRY IN PRODUCTION

Discuss: What key performance indicator of the *system* to collect?

Scenarios:

- Front-left: Amazon: Shopping app feature that detects the shoe brand from photos
- Front-right: Google: Tagging uploaded photos with friends' names
- Back-left: Spotify: Recommended personalized playlists
- Back-right: Wordpress: Profanity filter to moderate blog posts

(can update slack, but not needed)

# EXPERIMENTING IN PRODUCTION

- A/B experiments
- Shadow releases / traffic teeing
- Blue/green deployment
- Canary releases
- Chaos experiments



Changelog  
@changelog



“Don’t worry, our users will notify us if there’s a problem”



2:03 PM · Jun 8, 2019



[Read the full conversation on Twitter](#)



2.2K



Reply



Copy link to Tweet

[Read 12 replies](#)

# A/B EXPERIMENTS

# WHAT IF...?

- ... we had plenty of subjects for experiments
- ... we could randomly assign subjects to treatment and control group without them knowing
- ... we could analyze small individual changes and keep everything else constant
  - ▶ Ideal conditions for controlled experiments



SEE SOMETHING NEW, EVERY DAY.

TAKE A LOOK



amazon  
Try Prime

All ▾



Deliver to  
Pittsburgh 15213

Your Amazon.com

Today's Deals

Gift Cards

Whole Foods

Registry

Sell

EN



Hello, Sign in  
Account & Lists ▾

Orders

Try Prime ▾

Cart 0

## Valentine's Day deals

Amazon Devices



echo



fire tv stick 4K  
2-pack



echospot



fire 7  
3-pack

\$99<sup>99</sup> \$69<sup>99</sup>

\$99<sup>99</sup> \$84<sup>98</sup>

\$129<sup>99</sup> \$99<sup>99</sup>

\$149<sup>92</sup> \$109<sup>97</sup>

### Bargain finds



### Explore live plants



### New year, new records



Sign in for the best  
experience

Sign in securely

[https://www.amazon.com/Low-Price-With-Free-Shipping/bb?category=/electronics&ref=bb\\_bb\\_a77114\\_in\\_db\\_w\\_ur\\_en\\_US&linenumberAsins=B078H77S6K&of\\_rd\\_p=39b36ea0-aa36-484b-adef-e16d0468b38f&of\\_rd\\_r=6R617PT5...](https://www.amazon.com/Low-Price-With-Free-Shipping/bb?category=/electronics&ref=bb_bb_a77114_in_db_w_ur_en_US&linenumberAsins=B078H77S6K&of_rd_p=39b36ea0-aa36-484b-adef-e16d0468b38f&of_rd_r=6R617PT5...)

# A/B TESTING FOR USABILITY

- In running system, random sample of X users are shown modified version
- Outcomes (e.g., sales, time on site) compared among groups

Original: 2.3%



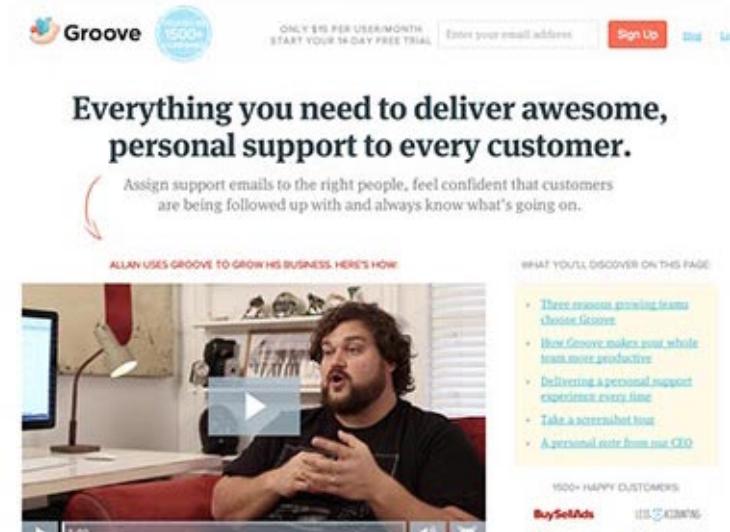
SaaS & eCommerce Customer Support.  
"Managing customer support requests in Groove is so easy. Way better than trying to use Gmail or a more complicated help desk."  
- Griffin, Customer Champion at Allocacoo  
97% of customers recommend Groove.

Learn More >

How it works    What you get    What it costs    How we're different

You'll be up and running in less than a minute.

Long Form: 4.3%



ONLY \$15 PER USER/MONTH  
START YOUR 14-DAY FREE TRIAL

Enter your email address  Sign Up

Everything you need to deliver awesome, personal support to every customer.

Assign support emails to the right people, feel confident that customers are being followed up with and always know what's going on.

ALLAN USES GROOVE TO GROW HIS BUSINESS. HERE'S HOW.

WHAT YOU'LL DISCOVER ON THIS PAGE:

- Three reasons growing teams choose Groove
- How Groove makes your whole team more productive
- Delivering a personal support experience every time
- Take a screenshot tour
- A personal note from our CEO

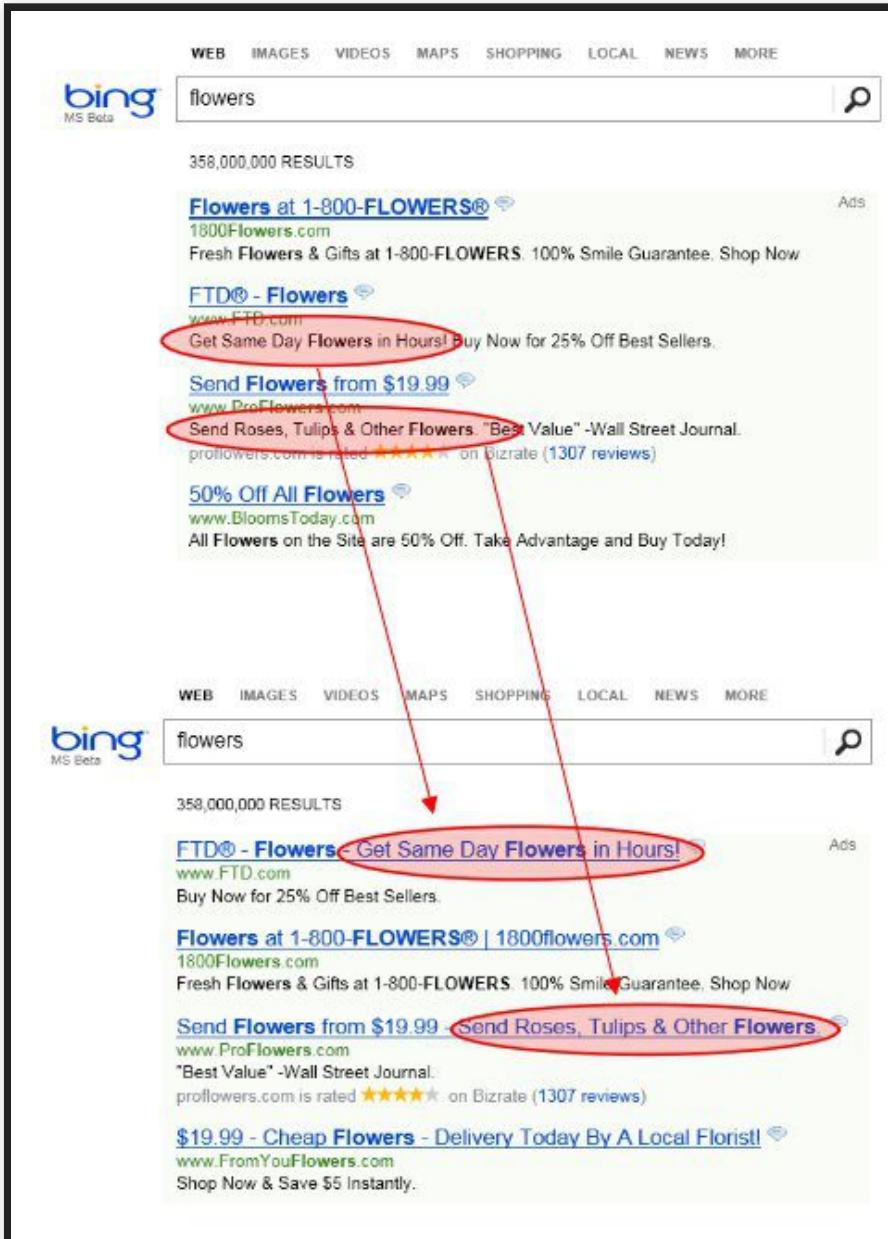
150+ HAPPY CUSTOMERS

BuySellAds    Hootsuite    METACRAFTER    StatusPage.io

Speaker notes

Picture source: <https://www.designforfounders.com/ab-testing-examples/>

# BING EXPERIMENT



- Experiment with Ad Display at Bing
- Suggestion prioritized low
- Not implemented for 6 month
- Ran A/B test in production
- Within 2h *revenue-too-high* alarm triggered suggesting serious bug (e.g., double billing)
- Revenue increase by 12% - \$100M annually in US
- Did not hurt user-experience metrics

From: Kohavi, Ron, Diane Tang, and Ya Xu. "["Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing."](#) Cambridge University Press, 2020.



# A/B EXPERIMENT FOR AI COMPONENTS?

- New product recommendation algorithm for web store?
- New language model in audio transcription service?
- New (offline) model to detect falls on smart watch



# EXPERIMENT SIZE

- With enough subjects (users), we can run many many experiments
- Even very small experiments become feasible
- Toward causal inference



# IMPLEMENTING A/B TESTING

- Implement alternative versions of the system
  - using feature flags (decisions in implementation)
  - separate deployments (decision in router/load balancer)
- Map users to treatment group
  - Randomly from distribution
  - Static user - group mapping
  - Online service (e.g., [launchdarkly](#), [split](#))
- Monitor outcomes *per group*
  - Telemetry, sales, time on site, server load, crash rate

# FEATURE FLAGS

```
if (features.enabled(userId, "one_click_checkout")) {  
    // new one click checkout function  
} else {  
    // old checkout functionality  
}
```

- Boolean options
- Good practices: tracked explicitly, documented, keep them localized and independent
- External mapping of flags to customers
  - who should see what configuration
  - e.g., 1% of users sees `one_click_checkout`, but always the same users; or 50% of beta-users and 90% of developers and 0.1% of all users

```
def isEnabled(user): Boolean = (hash(user.id) % 100) < 10
```

**Treatments** ⓘ | 2 treatments, if Split is killed serve the default treatment of "off"

Treatment	Default	Description
on		The new version of registration process is enabled.
off		The old version of registration process is enabled.

[+ Add treatment](#) | [Learn more about multivariate treatments](#).

**Whitelist** ⓘ | 0 user(s) or segments individually targeted.

[+ Add whitelist](#)

**Traffic Allocation** ⓘ | 100% of user included in Split rules evaluation below.

Total Traffic Allocation: 100 % total User in Split

**Targeting Rules** ⓘ | 2 rules created for targeting.

```

graph TD
    If1((if)) --> Cond1["user is in segment qa"]
    Cond1 --> Then1["Then serve on"]
    ElseIf2((else if)) --> Cond2["user is in segment beta_testers"]
    Cond2 --> Then2["Then serve percentage"]
    Then2 --> On50["on 50"]
    Then2 --> Off50["off 50"]
  
```

[+ Add rule](#)

**Default Rule** ⓘ | Serve treatment of "off".

serve off

# **CONFIDENCE IN A/B EXPERIMENTS**

(statistical tests)

# COMPARING AVERAGES

## Group A

*classic personalized content  
recommendation model*

2158 Users

average 3:13 min time on site

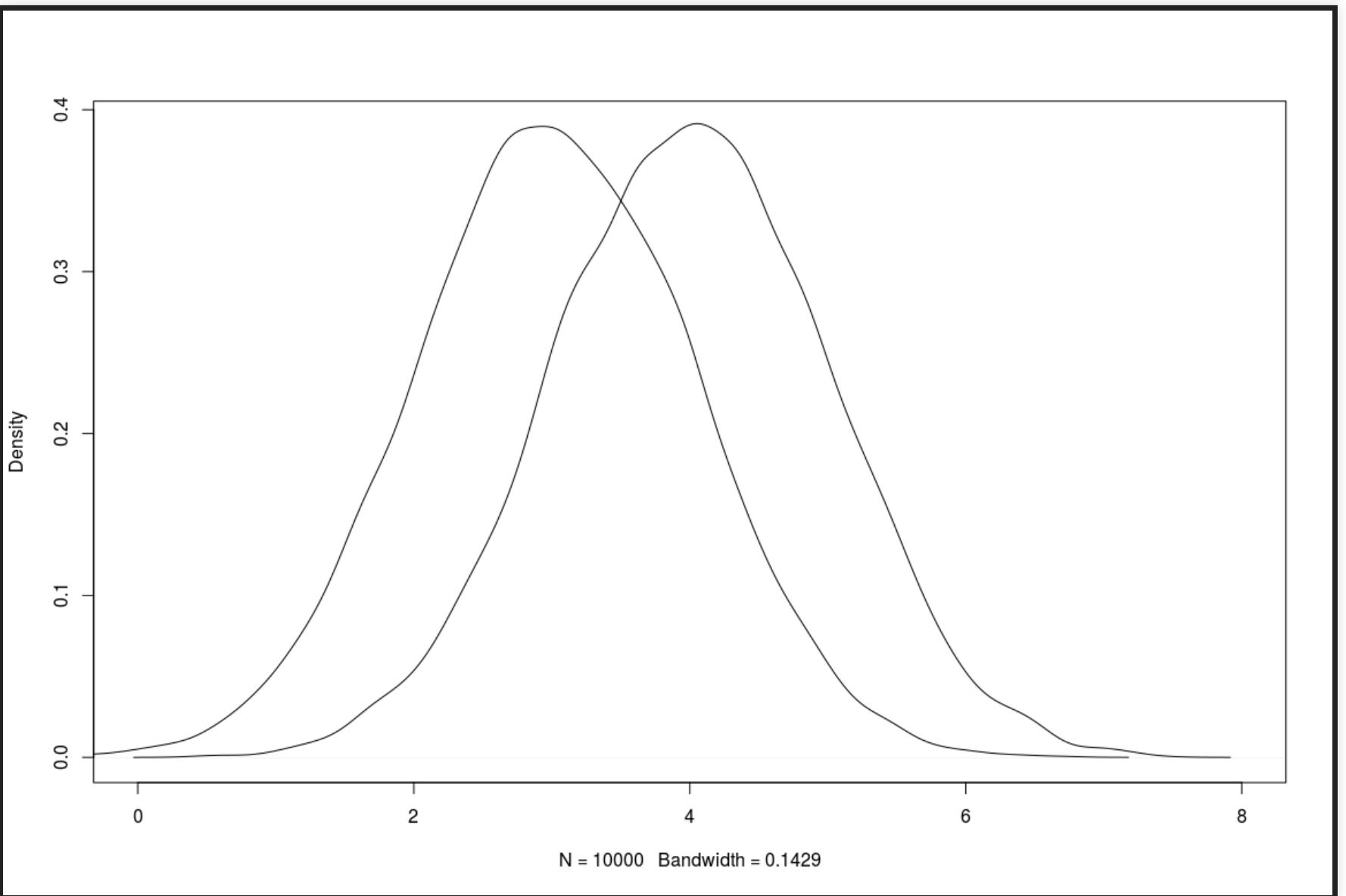
## Group B

*updated personalized content  
recommendation model*

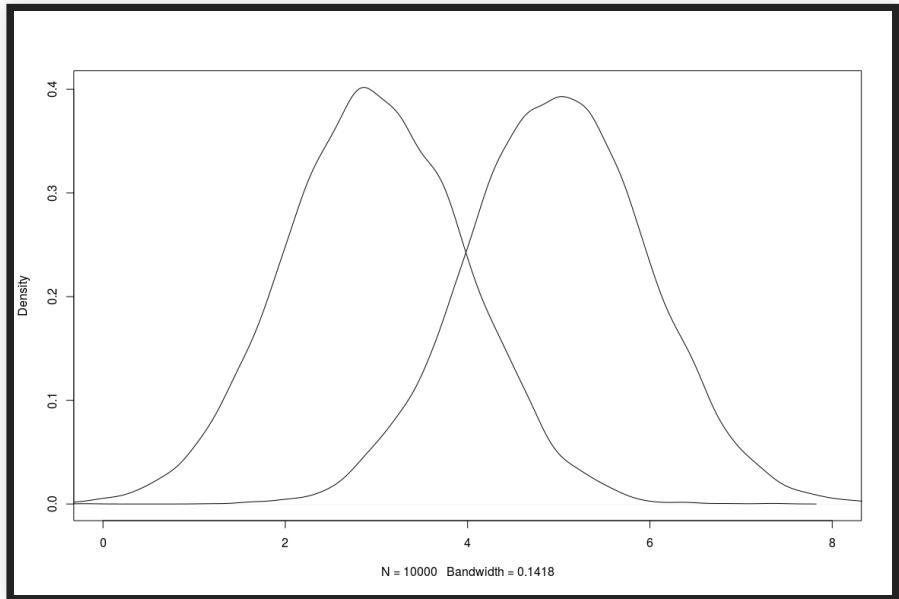
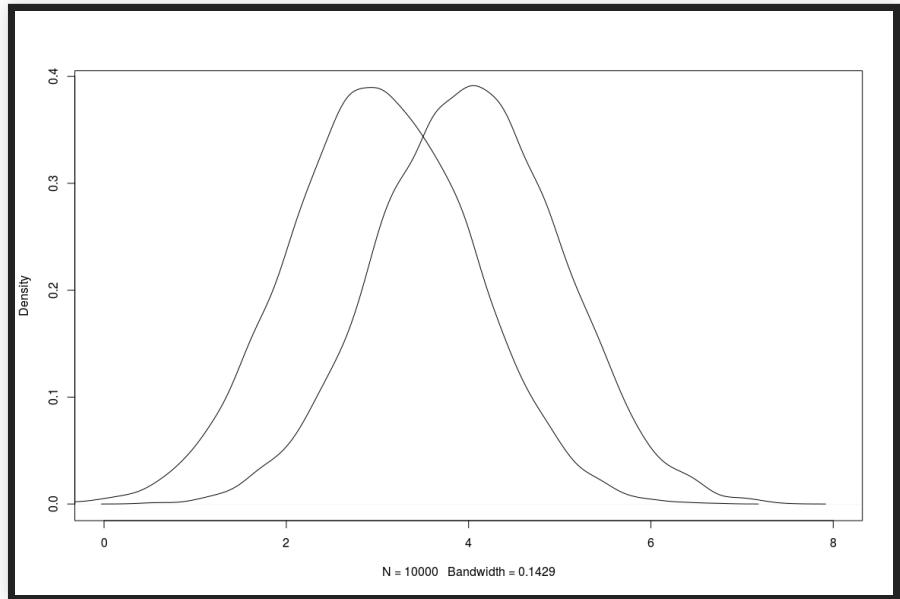
10 Users

average 3:24 min time on site

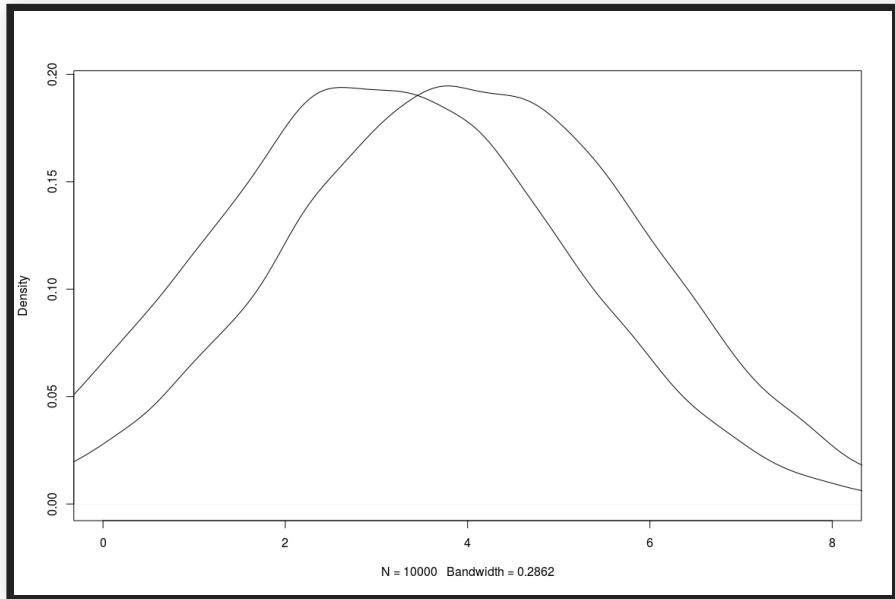
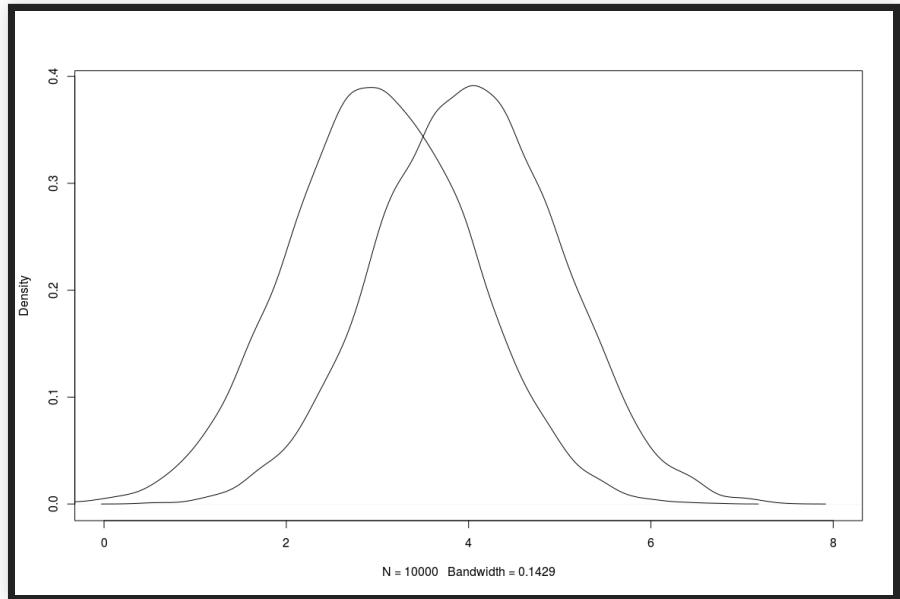
# COMPARING DISTRIBUTIONS



# DIFFERENT EFFECT SIZE, SAME DEVIATIONS



# SAME EFFECT SIZE, DIFFERENT DEVIATIONS



Less noise --> Easier to recognize

# DEPENDENT VS. INDEPENDENT MEASUREMENTS

- Pairwise (dependent) measurements
  - Before/after comparison
  - With same benchmark + environment
  - e.g., new operating system/disc drive faster
- Independent measurements
  - Repeated measurements
  - Input data regenerated for each measurement

# SIGNIFICANCE LEVEL

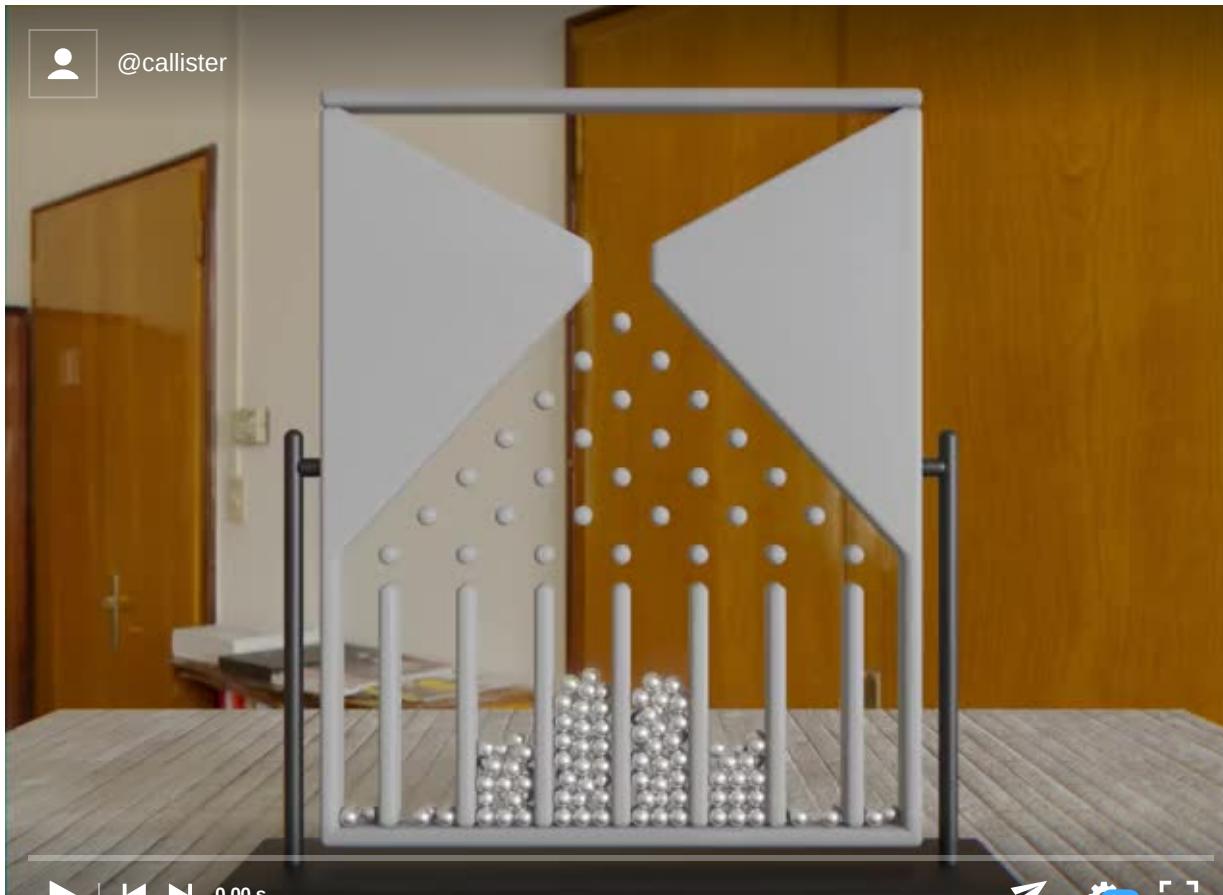
- Statistical change of an error
- Define before executing the experiment
  - use commonly accepted values
  - based on cost of a wrong decision
- Common:
  - 0.05 significant
  - 0.01 very significant
- Statistically significant result  $\Rightarrow$  proof
- Statistically significant result  $\Rightarrow$  important result
- Covers only alpha error (more later)

# INTUITION: ERROR MODEL

- 1 random error, influence +/- 1
- Real mean: 10
- Measurements: 9 (50%) und 11 (50%)
  
- 2 random errors, each +/- 1
- Measurements: 8 (25%), 10 (50%) und 12 (25%)
  
- 3 random errors, each +/- 1
- Measurements : 7 (12.5%), 9 (37.5), 11 (37.5), 12 (12.5)



@callister



▶ | ← ▶ 0.00 s



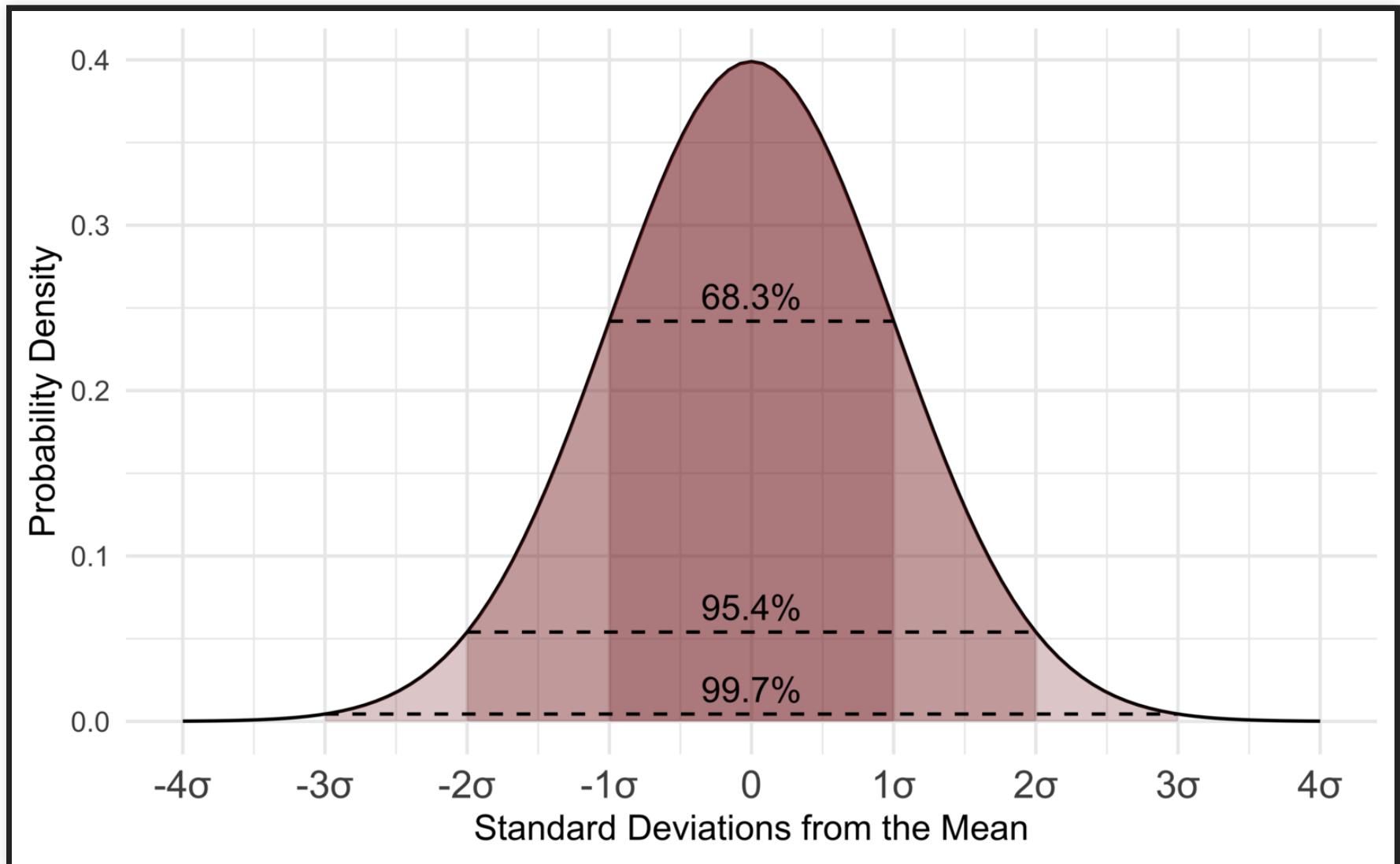
SD



16K views

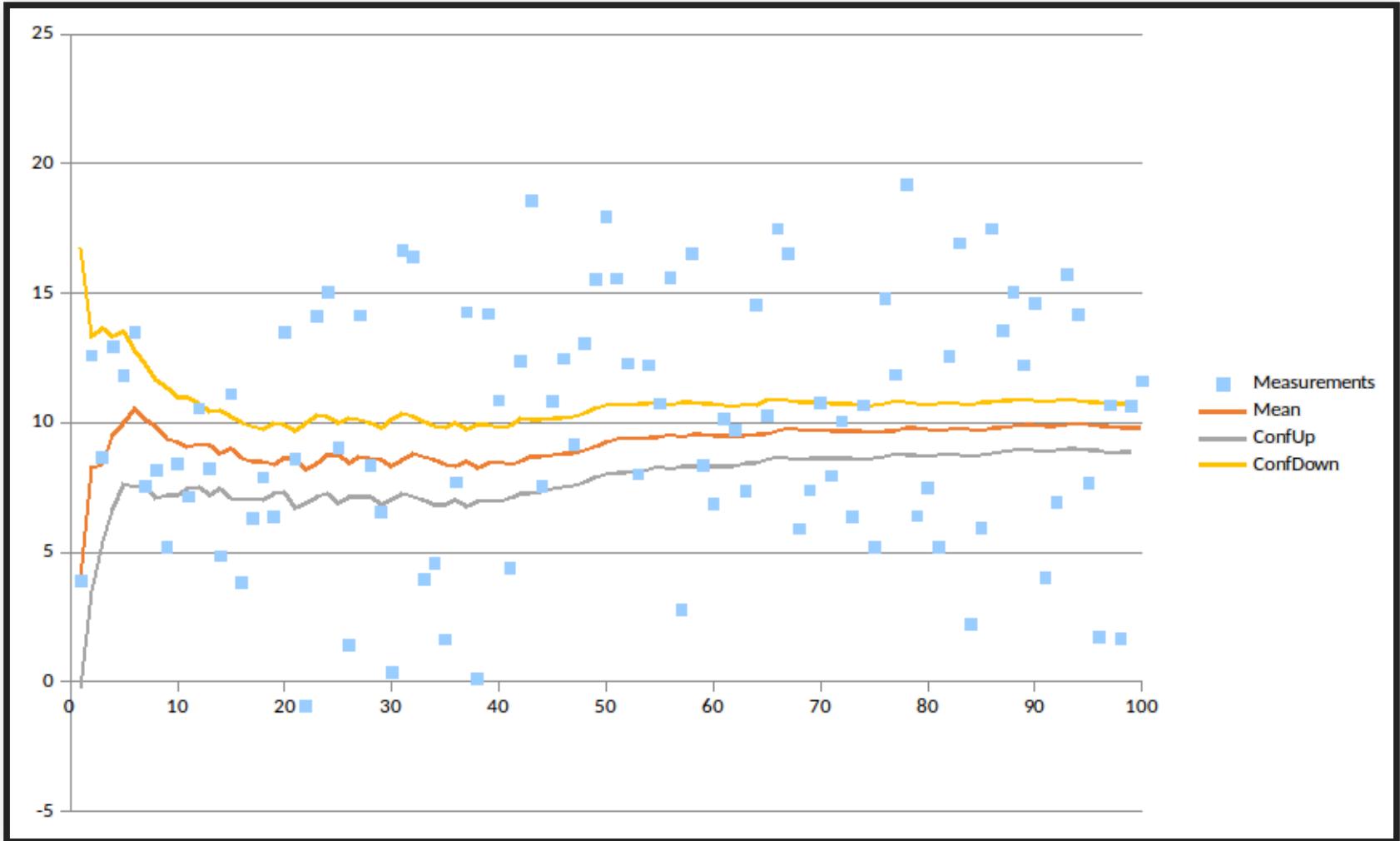
gfycat

# NORMAL DISTRIBUTION



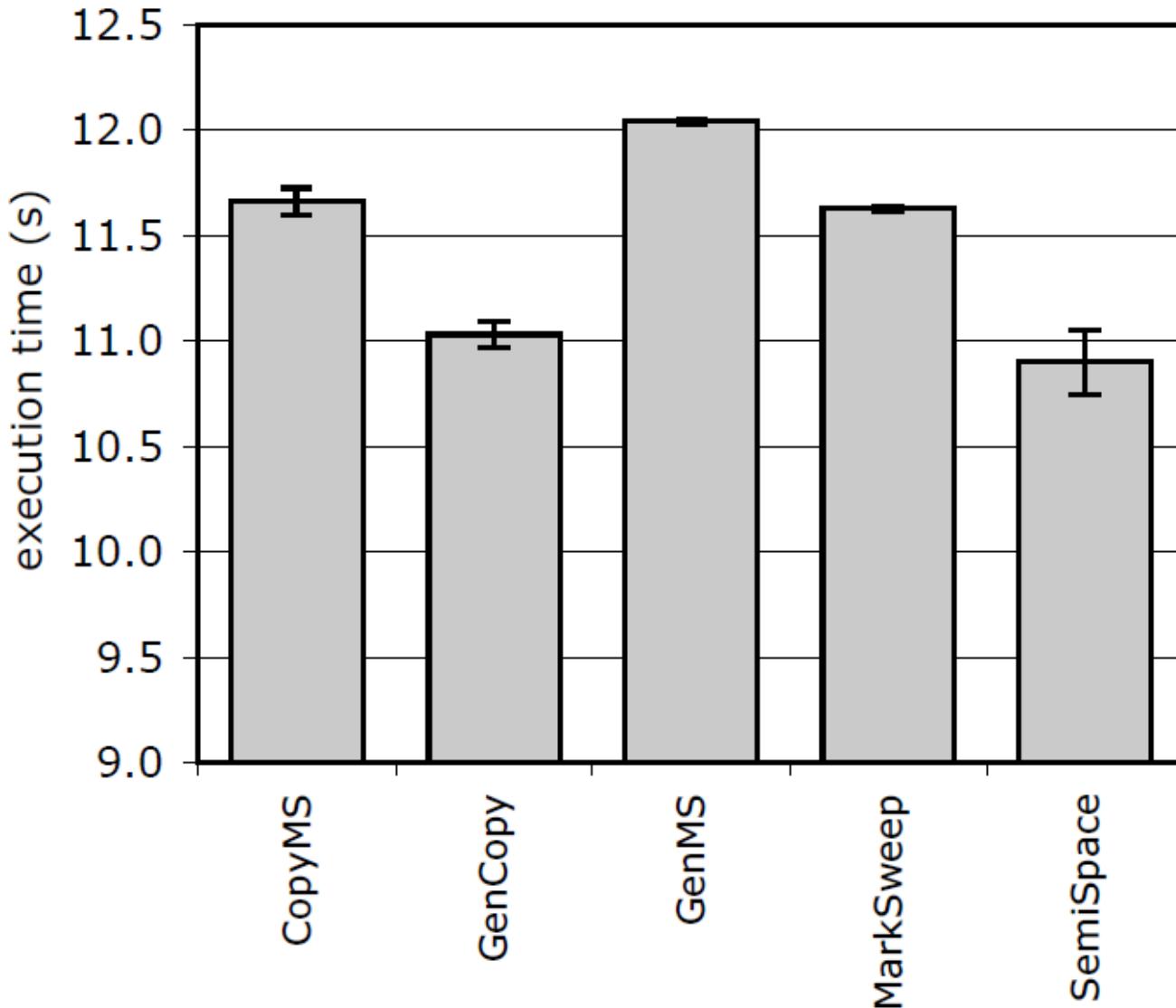


# CONFIDENCE INTERVALS



# COMPARISON WITH CONFIDENCE INTERVALS

*mean w/ 95% confidence interval*



# T-TEST

```
> t.test(x, y, conf.level=0.9)

Welch Two Sample t-test

t = 1.9988, df = 95.801, p-value = 0.04846
alternative hypothesis: true difference in means is
not equal to 0
90 percent confidence interval:
 0.3464147 3.7520619
sample estimates:
mean of x mean of y
 51.42307 49.37383

> # paired t-test:
> t.test(x-y, conf.level=0.9)
```

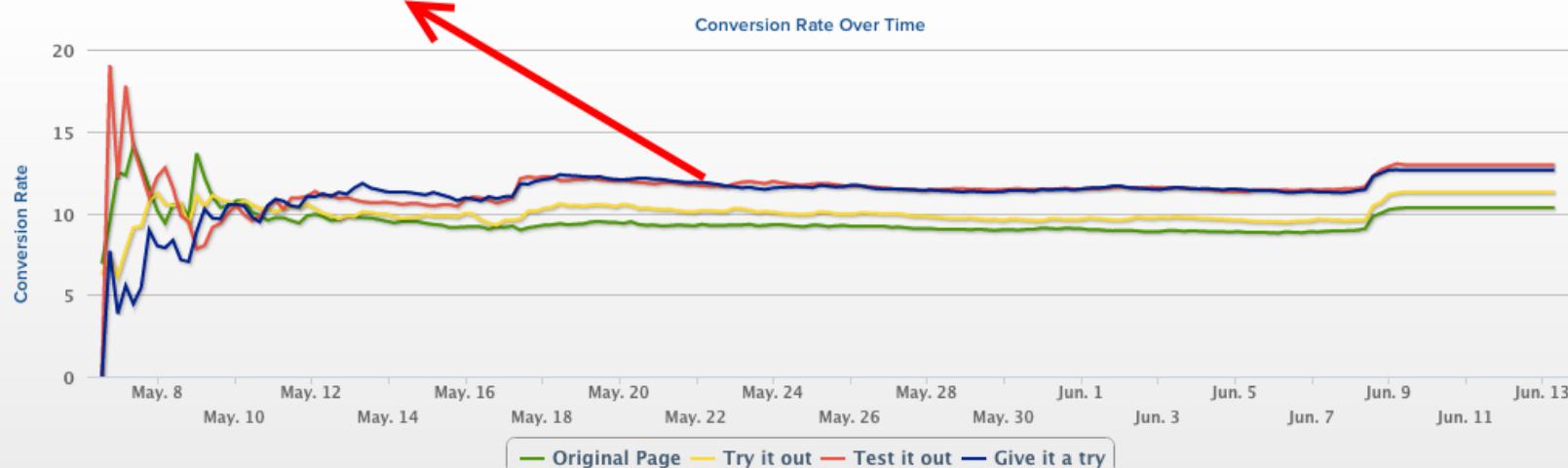
## Experiment Created

[Edit](#) [Remove](#) [Delete](#)

✓ Test it out is beating Original Page by +25.4%.

The percentage of visitors who clicked on a tracked element.

Variations	Conversions / Visitors	Conversion Rate	Baseline	Chance to beat Baseline	Improvement
Experiment	Conversions / Visitors	Conversion Rate			
Test it out	462 / 3,568	12.9% ( $\pm 1.1\%$ )		 100.0%	+25.4%
Give it a try	440 / 3,479	12.6% ( $\pm 1.1\%$ )		 99.9%	+22.5%
Try it out	395 / 3,504	11.3% ( $\pm 1.0\%$ )		90.2%	+9.2%
Original Page	378 / 3,662	10.3% ( $\pm 1.0\%$ )		 ***	---

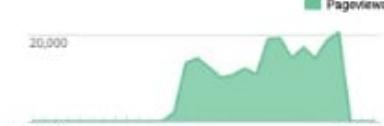
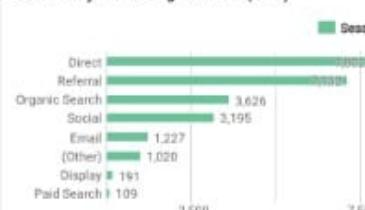
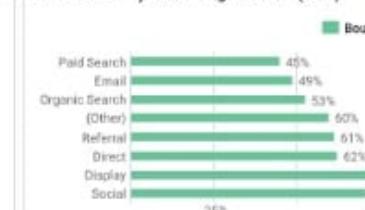
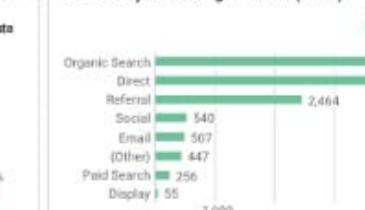
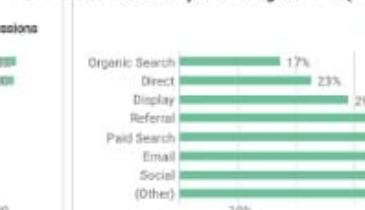


Source: <https://conversionsciences.com/ab-testing-statistics/>





01/23/2017 - 01/31/2017

V1 (Old Design)		V2 (New Design)	
Sessions (OLD) <b>24,301</b>	Pageviews (OLD) <b>59,987</b>	Sessions (NEW) <b>13,091</b>	Pageviews (NEW) <b>100,623</b>
New Visitor vs Returning Visitor (OLD)	Pageviews Trend (OLD)	New Visitor vs Returning Visitor (NEW)	Pageviews Trend (NEW)
 New Visitor: 5,856 (24%) Returning Visitor: 18,445 (76%)	 Pageviews 5,000 01/11/2017 02/09/2017	 New Visitor: 1,009 (8%) Returning Visitor: 12,082 (92%)	 Pageviews 20,000 01/11/2017 02/09/2017
Bounce Rate (OLD) <b>62.19%</b>	Time on Site in seconds (OLD) <b>187</b>	Bounce Rate (NEW) <b>25.03%</b>	Time on Site in seconds (NEW) <b>443</b>
Pages / Session (OLD) <b>2.47</b>	Sessions w. Search (OLD) <b>5.61%</b>	Pages / Session (NEW) <b>7.69</b>	Sessions w. Search (NEW) <b>42.23%</b>
List Through Rate (OLD) <b>2.10%</b>	Contact Through Rate (OLD) <b>4.12%</b>	List Through Rate (NEW) <b>2.09%</b>	Contact Through Rate (NEW) <b>5.43%</b>
Sessions by Marketing Channel (OLD)	Bounce Rate by Marketing Channel (OLD)	Sessions by Marketing Channel (NEW)	Bounce Rate by Marketing Channel (NEW)
 Direct: 18,445 Referral: 3,195 Organic Search: 3,626 Social: 3,195 Email: 1,227 (Other): 1,020 Display: 191 Paid Search: 109	 Paid Search: 45% Email: 49% Organic Search: 53% (Other): 60% Referral: 61% Direct: 62% Display: 74% Social: 75%	 Organic Search: 13,091 Direct: 12,082 Referral: 2,464 Social: 540 Email: 507 (Other): 447 Paid Search: 256 Display: 55	 Organic Search: 17% Direct: 23% Display: 29% Referral: 33% Paid Search: 36% Email: 36% Social: 40% (Other): 42%
Sessions by Landing Page (OLD)	Bounce Rate by Landing Page (OLD)	Sessions by Landing Page (NEW)	Bounce Rate by Landing Page (NEW)
LANDING PAGE GROUP 1 VIP: 14,792 eVIP: 3,206 Homepage: 2,464 ActivateAdSuccess: 1,080	SESSIONS Login: 21.05% ResultsBrowse: 23.77% Homepage: 30.32% ResultsSearch: 35.31% (NULL): 49.32%	SESSIONS VIP: 4,329 Homepage: 2,582 ResultsSearch: 2,666 ResultsBrowse: 1,598	SESSIONS UserRegistrationForm: 0.00% EditAdForm: 4.39% Homepage: 6.94% Login: 14.47% MyAds: 15.15%
Search Pages / Session (OLD)	Search Pageviews Trend (OLD)	Search Pages / Session (NEW)	Search Pageviews Trend (NEW)

# HOW MANY SAMPLES NEEDED?

Too few?

Too many?



# A/B TESTING AUTOMATION

- Experiment configuration through DSLs/scripts
- Queue experiments
- Stop experiments when confident in results
- Stop experiments resulting in bad outcomes (crashes, very low sales)
- Automated reporting, dashboards

Further readings:

- Tang, Diane, et al. [Overlapping experiment infrastructure: More, better, faster experimentation.](#)  
Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.  
ACM, 2010. (Google)
- Bakshy, Eytan, Dean Eckles, and Michael S. Bernstein. [Designing and deploying online field experiments.](#)  
Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014. (Facebook)

# DSL FOR SCRIPTING A/B TESTS AT FACEBOOK

```
button_color = uniformChoice(  
    choices=['#3c539a', '#5f9647', '#b33316'],  
    unit=cookieid);  
  
button_text = weightedChoice(  
    choices=['Sign up', 'Join now'],  
    weights=[0.8, 0.2],  
    unit=cookieid);  
  
if (country == 'US') {  
    has_translate = bernoulliTrial(p=0.2, unit=userid);  
} else {  
    has_translate = bernoulliTrial(p=0.05, unit=userid);  
}
```

Further readings:

- Bakshy, Eytan, Dean Eckles, and Michael S. Bernstein. [Designing and deploying online field experiments](#). Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014. (Facebook)

# CONCURRENT A/B TESTING

- Multiple experiments at the same time
  - Independent experiments on different populations -- interactions not explored
  - Multi-factorial designs, well understood but typically too complex, e.g., not all combinations valid or interesting
  - Grouping in sets of experiments (layers)

Further readings:

- Tang, Diane, et al. [Overlapping experiment infrastructure: More, better, faster experimentation.](#)  
Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.  
ACM, 2010. (Google)
- Bakshy, Eytan, Dean Eckles, and Michael S. Bernstein. [Designing and deploying online field experiments.](#)  
Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014. (Facebook)

# OTHER EXPERIMENTS IN PRODUCTION

- Shadow releases / traffic teeing
- Blue/green deployment
- Canary releases
- Chaos experiments

# SHADOW RELEASES / TRAFFIC TEEING

- Run both models in parallel
- Report outcome of old model
- Compare differences between model predictions
- If possible, compare against ground truth labels/telemetry

Examples?

# BLUE/GREEN DEPLOYMENT

- Provision service both with old and new model (e.g., services)
- Support immediate switch with load-balancer
- Allows to undo release rapidly

Advantages/disadvantages?

# CANARY RELEASES

- Release new version to small percentage of population (like A/B testing)
- Automatically roll back if quality measures degrade
- Automatically and incrementally increase deployment to 100% otherwise



# CHAOS EXPERIMENTS



# CHAOS EXPERIMENTS FOR AI COMPONENTS?



## Speaker notes

Artifically reduce model quality, add delays, insert bias, etc to test monitoring and alerting infrastructure

# ADVICE FOR EXPERIMENTING IN PRODUCTION

- Minimize *blast radius* (canary, A/B, chaos expr)
- Automate experiments and deployments
- Allow for quick rollback of poor models (continuous delivery, containers, loadbalancers, versioning)
- Make decisions with confidence, compare distributions
- Monitor, monitor, monitor

# BONUS: MONITORING WITHOUT GROUND TRUTH

# INVARIANTS/ASSERTIONS TO ASSURE WITH TELEMETRY

- Consistency between multiple sources
  - e.g., multiple models agree, multiple sensors agree
  - e.g., text and image agree
- Physical domain knowledge
  - e.g., cars in video shall not flicker,
  - e.g., earthquakes should appear in sensors grouped by geography
- Domain knowledge about unlikely events
  - e.g., unlikely to have 3 cars in same location
- Stability
  - e.g., object detection should not change with video noise
- Input conforms to schema (e.g. boolean features)
- And all invariants from model quality lecture, including capabilities

Kang, Daniel, Deepti Raghavan, Peter Bailis, and Matei Zaharia. "Model Assertions for Monitoring and Improving ML Model." In Proceedings of MLSys 2020.

# SUMMARY

- Production data is ultimate unseen validation data
- Telemetry is key and challenging (design problem and opportunity)
- Monitoring and dashboards
- Many forms of experimentation and release (A/B testing, shadow releases, canary releases, chaos experiments, ...) to minimize "blast radius"
- Gain confidence in results with statistical tests