

# QUALITY ASSESSMENT IN PRODUCTION

Christian Kaestner

Required Reading: Alec Warner and Štěpán Davidovič. "[Canary Releases.](#)" in [The Site Reliability Workbook](#), O'Reilly 2018



Suggested Reading: Georgi Georgiev. "[Statistical Significance in A/B Testing – a Complete Guide.](#)" Blog 2018



Changelog

@changelog



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



♡ 2,336 2:03 PM - Jun 8, 2019



💬 706 people are talking about this



# LEARNING GOALS

- Design telemetry for evaluation in practice
- 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

# RECALL: MODEL QUALITY

# CONFUSION/ERROR MATRIX

	Actually A	Actually B	Actually C
AI predicts A	10	6	2
AI predicts B	3	24	10
AI predicts C	5	22	82

Accuracy = correct predictions (diagonal) out of all predictions

$$\text{Example's accuracy} = \frac{10+24+82}{10+6+2+3+24+10+5+22+82} = .707$$

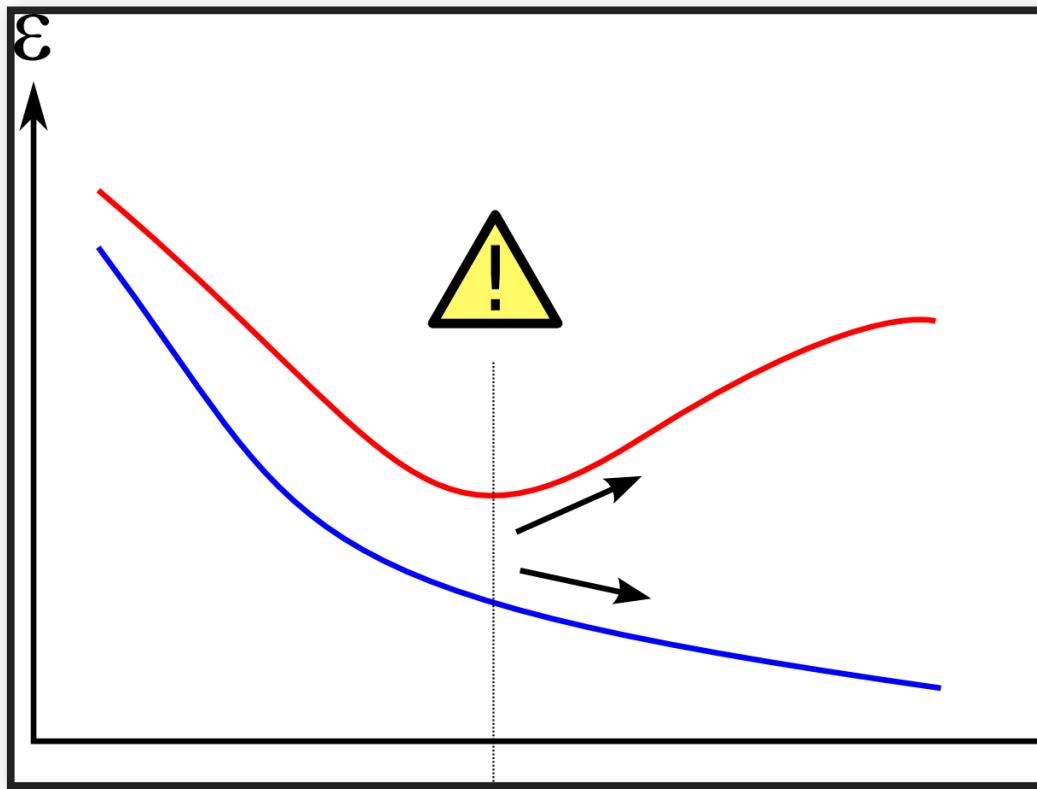
# AREA UNDER THE CURVE

Turning numeric prediction into classification with threshold ("operating point")



# DETECTING OVERFITTING

Change hyperparameter to detect training accuracy (blue)/validation accuracy (red) at different degrees of freedom



(CC SA 3.0 by [Dake](#))

# SEPARATE TRAINING, VALIDATION AND TEST DATA

Often a model is "tuned" manually or automatically on a validation set  
(hyperparameter optimization)

In this case, we can overfit on the validation set, separate test set is needed for  
final evaluation

```
train_xs, train_ys, valid_xs, valid_ys, test_xs, test_ys =  
    split(all_xs, all_ys)  
  
best_model = null  
best_model_accuracy = 0  
for hyperparameters in candidate_hyperparameters:  
    candidate_model = learn(train_xs, train_ys, hyperparameter)  
    model_accuracy = accuracy(model, valid_xs, valid_ys)  
    if (model_accuracy > best_model_accuracy):  
        best_model = candidate_model  
        best_model_accuracy = model_accuracy  
  
accuracy_test = accuracy(model, test_xs, test_ys)
```

# VIOLATING INDEPENDENCE OF TEST DATA





## Speaker notes

Many examples:

- Stock prediction (trained on future data)
- Detecting distracted drivers (multiple pictures per driver)
- Detecting horse breeds (copyright marks)
- Detecting severity of cancer (different scanners)
- Detecting tanks in photographs (sunny vs cloudy days)
- Left steering on rainy days of self-driving cars (cloudy skies)



# VALIDATION DATA REPRESENTATIVE?

- Validation data should reflect usage data
- Be aware of data drift (face recognition during pandemic, new patterns in credit card fraud detection)
- "*Out of distribution*" predictions often low quality (it may even be worth to detect out of distribution data in production, more later)

# ACADEMIC ESCALATION: OVERFITTING ON BENCHMARKS



(Figure by Andrea Passerini)

# MODEL ASSESSMENT IN PRODUCTION

Ultimate held-out evaluation data: Unseen real user data

# IDENTIFY FEEDBACK MECHANISM IN PRODUCTION

- Live observation in the running system
- Potentially on subpopulation (AB testing)
- Need telemetry to evaluate quality -- challenges:
  - Gather feedback without being intrusive (i.e., labeling outcomes), 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 "Create a price alert" button. At the bottom, there are date and time dropdown menus for "Take-off Dallas" and "Arrival San Francisco".

## Speaker notes

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



A transcription interface with a timeline at the top showing 00:00, Offset, 00:00, and 01:31:27. Below the timeline are four buttons: Play, Back 5s, 1x Speed, and Volume.

## 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



# MEASURING MODEL QUALITY WITH TELEMETRY

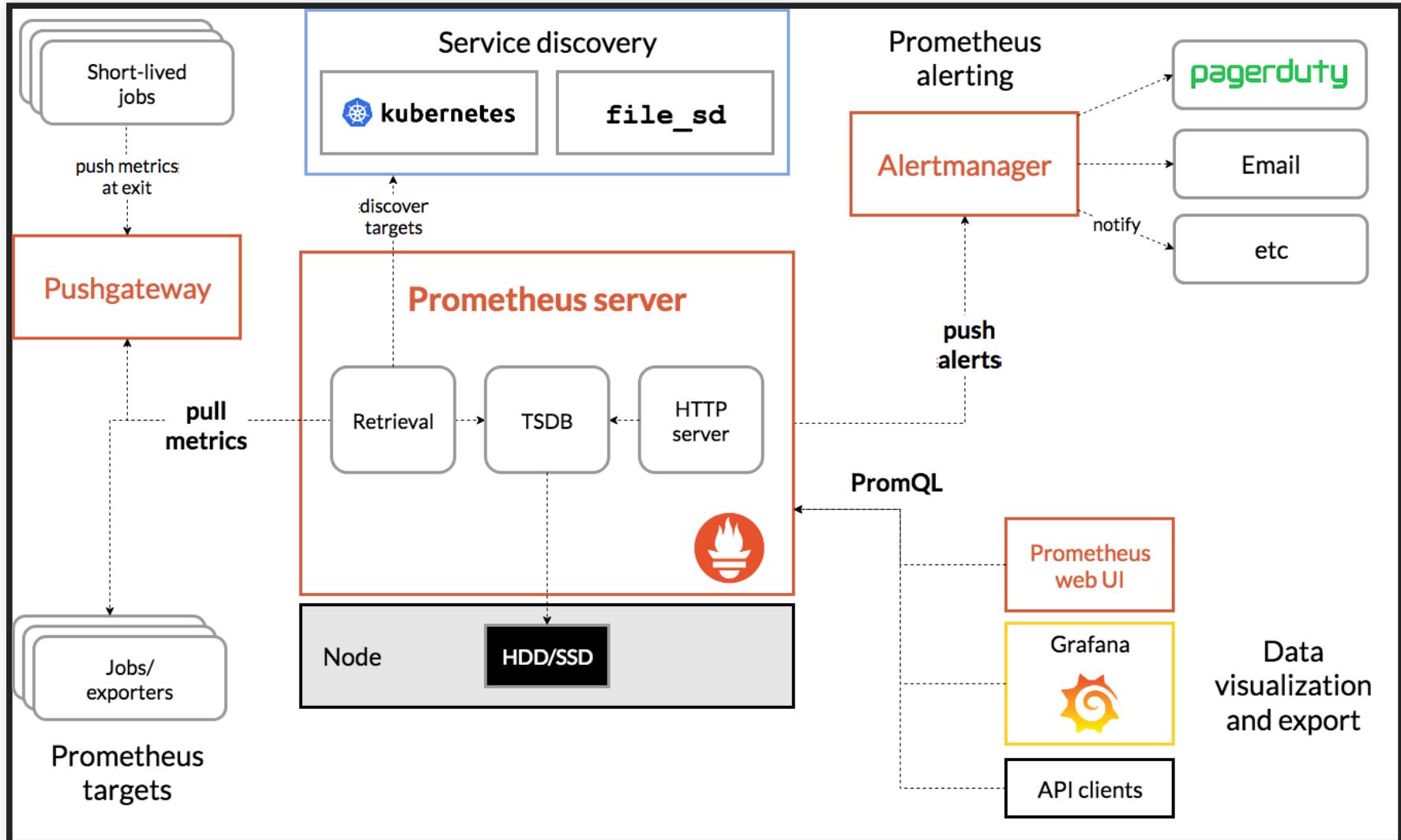
- Three steps:
  - Metric: Identify quality of concern
  - Telemetry: Describe data collection procedure
  - Operationalization: Measure quality metric in terms of data
- Telemetry can provide insights for correctness
  - sometimes very accurate labels for real unseen data
  - sometimes only mistakes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



# 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

# PROMETHEUS AND GRAFANA







Website Overview

Zoom Out

Last 3 hours



Logins

**190**

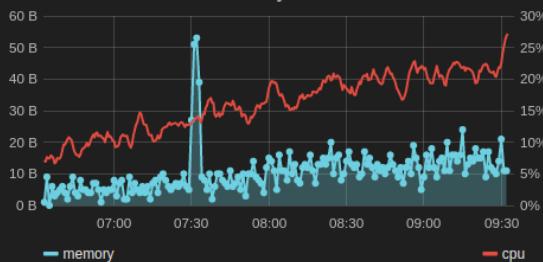
Sign ups

**269**

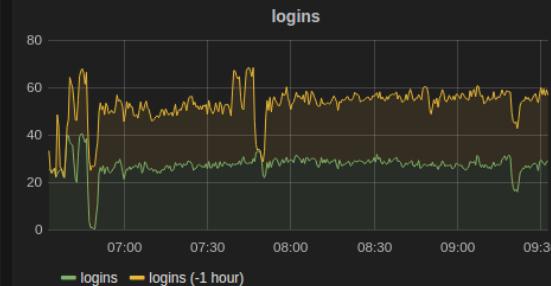
Sign outs

**273**

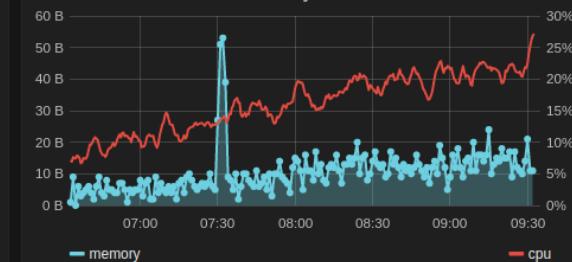
Memory / CPU



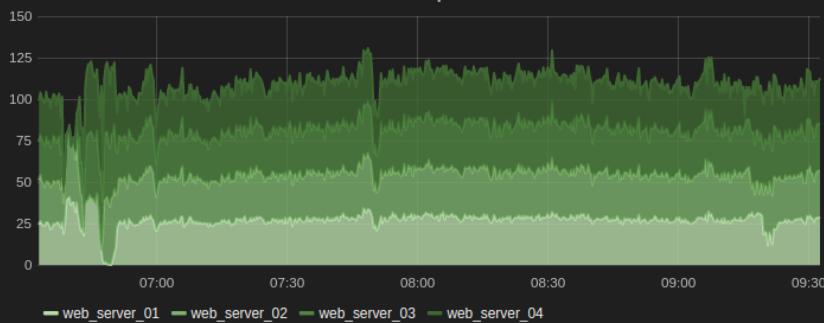
logins



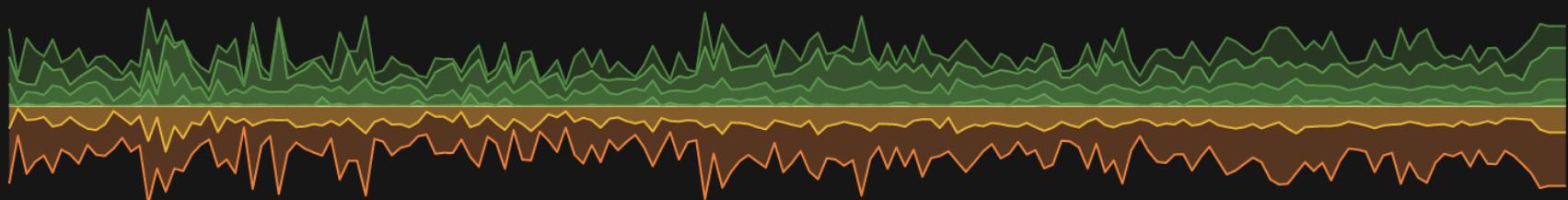
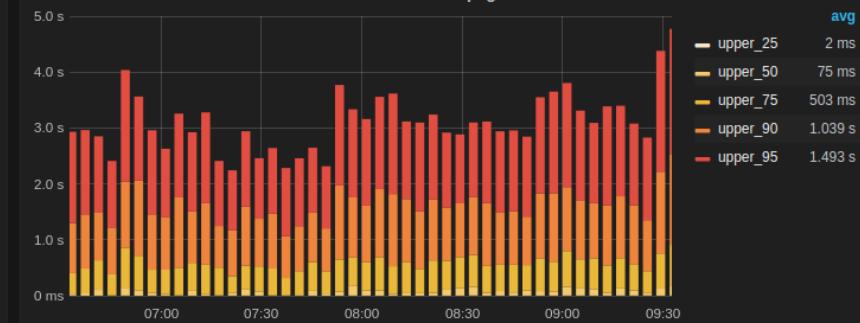
Memory / CPU



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



# 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?



# EXERCISE: DESIGN TELEMETRY IN PRODUCTION

*Scenario: Injury detection in smart home workout (laptop camera)*

Discuss: Quality measure, telemetry, operationalization, false positives/negatives, cost, privacy, rare events





# 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”



♡ 2,336 2:03 PM - Jun 8, 2019



💬 706 people are talking about this



# 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

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



fire tv stick 4K  
2-pack

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



echospot

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



fire 7  
3-pack

\$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&linenumber=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&linenumber=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

Blog    Learn

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

1500+ HAPPY CUSTOMERS

BuySellAds    Hootsuite    StatusPage.io

METACRAFTER

## Speaker notes

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



## Save on prescription drugs - over \$3,637\* a year!

Last year, Humana's Medicare Advantage plan members saved, on average, \$3,637\* on prescription drugs!

Choose your Humana Medicare Advantage plan and you could enjoy savings on prescription drugs, plus:

- Hospital, doctor AND drug coverage combined into one easy-to-use plan
- Extra benefits not offered by Original Medicare
- Affordable or no monthly plan premiums

[Shop 2014 Medicare Plans](#)

Control



## Explore Humana's Medicare plans

Let us help you determine the Humana plan that's best for your needs.

[Get started now](#)



Treatment



## Speaker notes

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



# 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

**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)) --> Then1[Then serve on]
    ElseIf2((else if)) --> Then2[Then serve percentage]
    style If1 fill:#337ab7,color:#fff
    style ElseIf2 fill:#337ab7,color:#fff
    style Then1 fill:#337ab7,color:#fff
    style Then2 fill:#337ab7,color:#fff
  
```

[+ 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

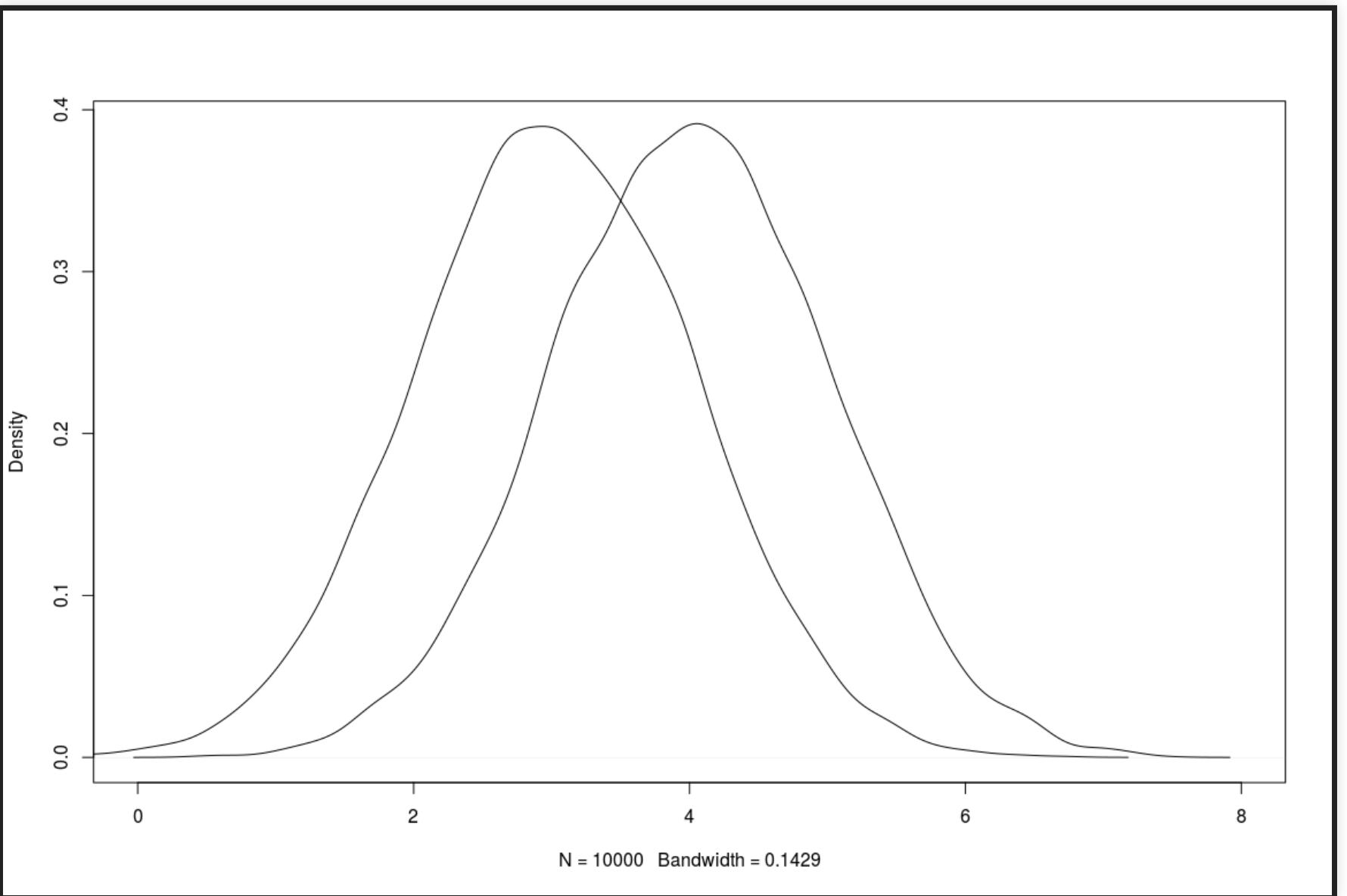
*updated personalized content  
recommendation model*

10 Users

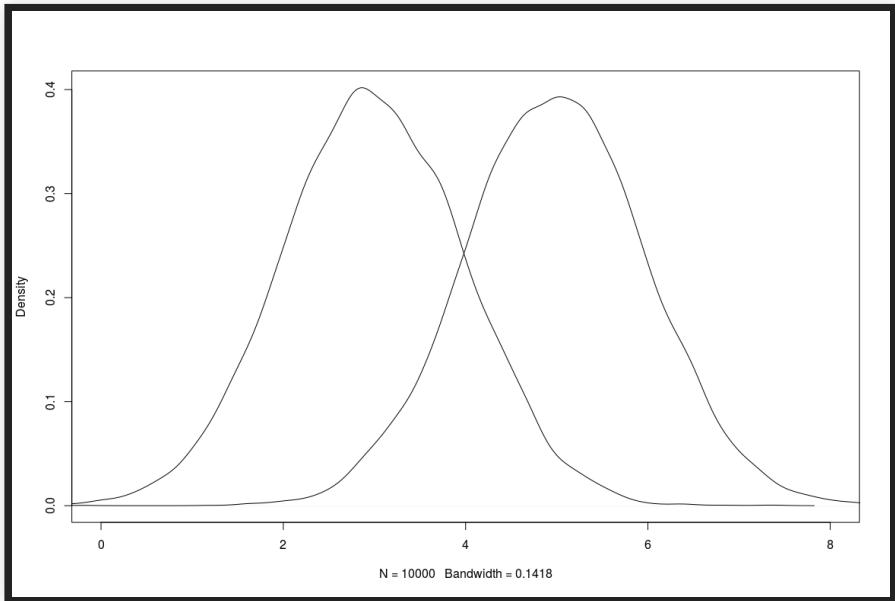
average 3:24 min time on site

# COMPARING DISTRIBUTIONS

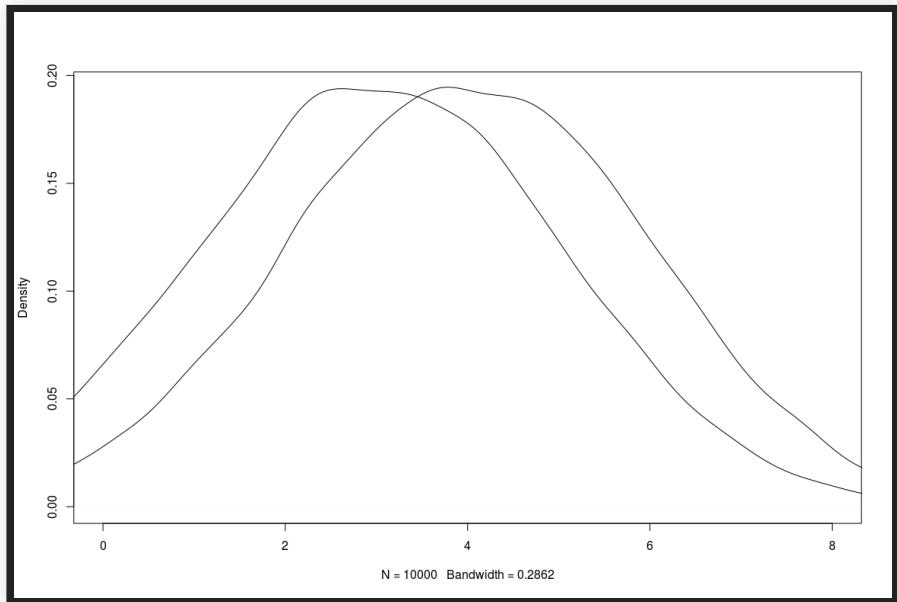
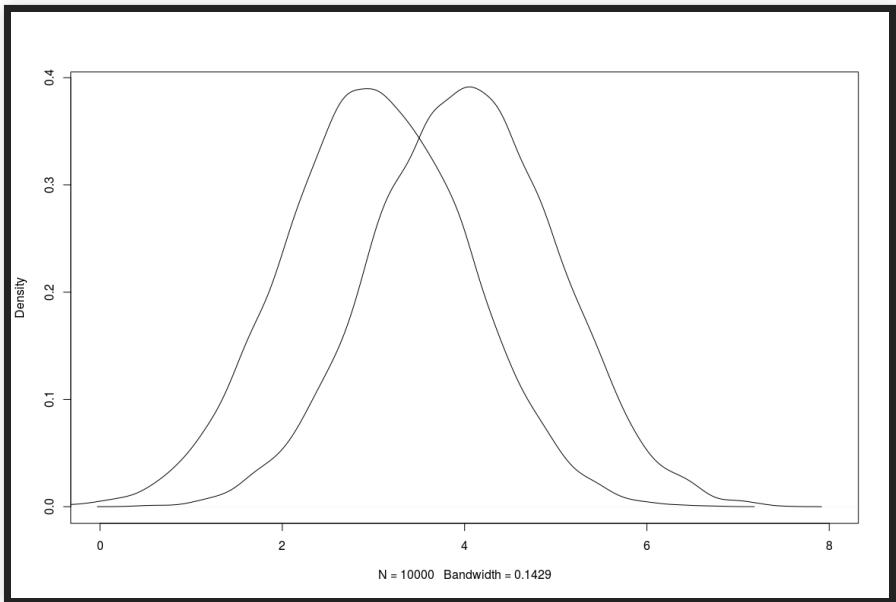




# DIFFERENT EFFECT SIZE, SAME DEVIATIONS



# 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

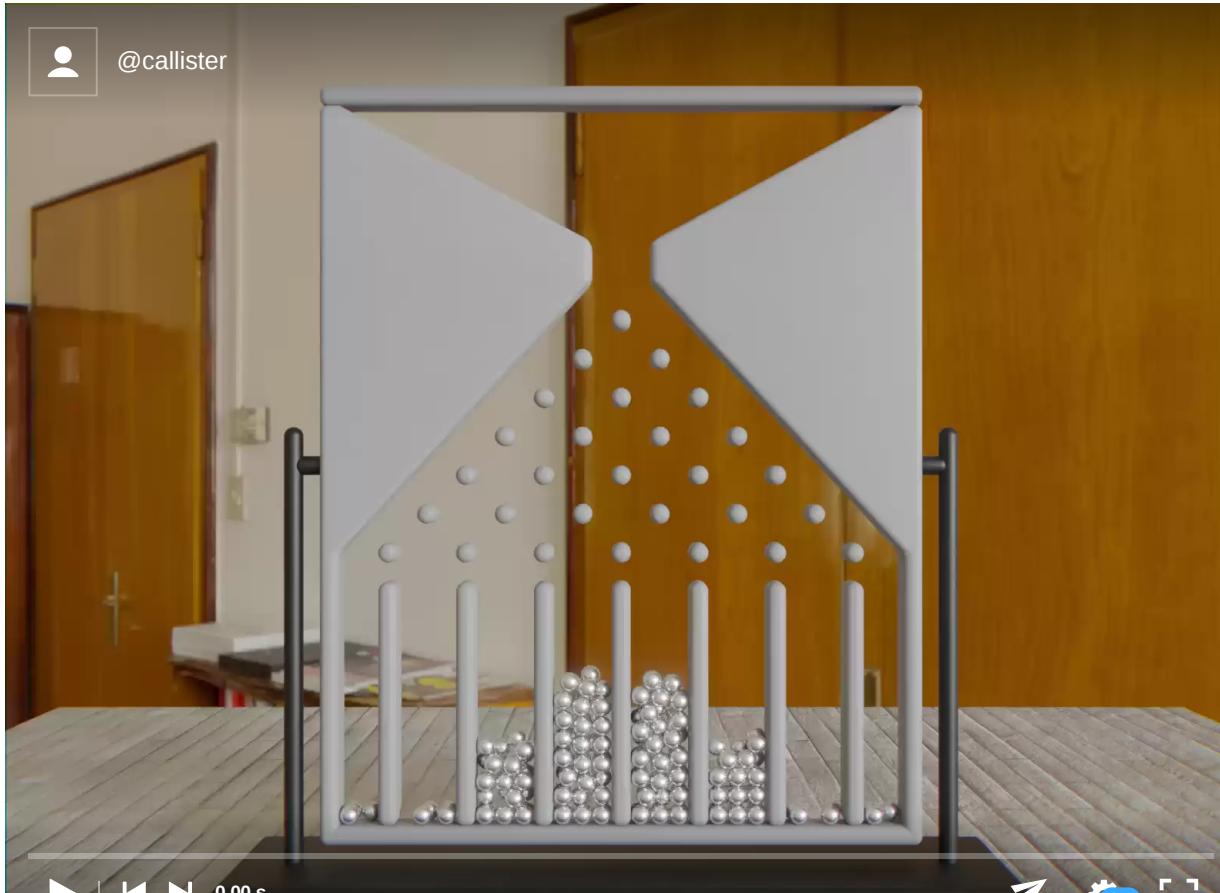
- Statistical change of an error
- Define before executing the experiment
  - use commonly accepted values
  - based on cost of a wrong decision
- Common:
  - 0.05 significant
  - 0.01 very significant
- Statistically significant result =!> proof
- Statistically significant result =!> important result
- 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 [ ]

10.1K views

gfycat

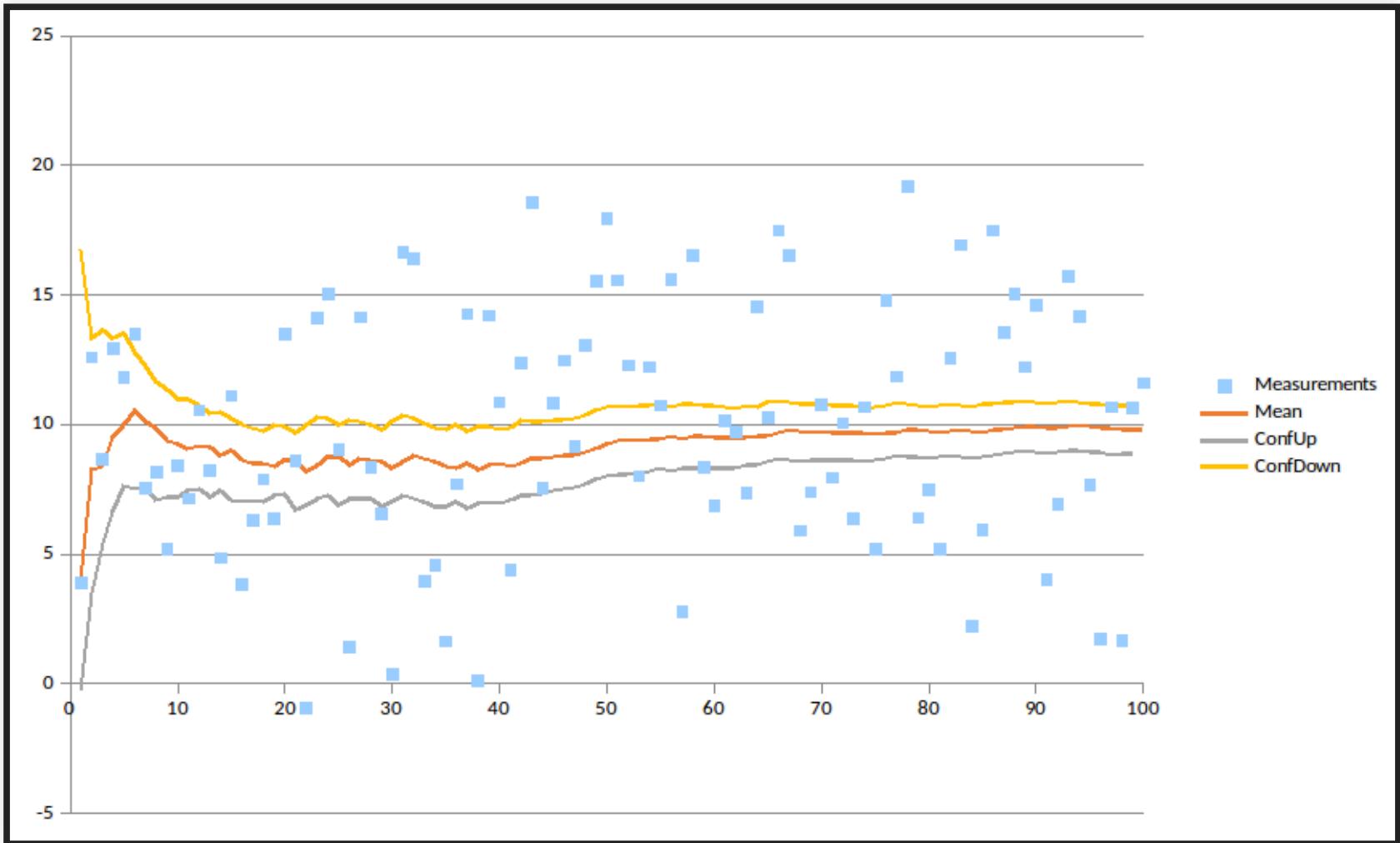


# NORMAL DISTRIBUTION





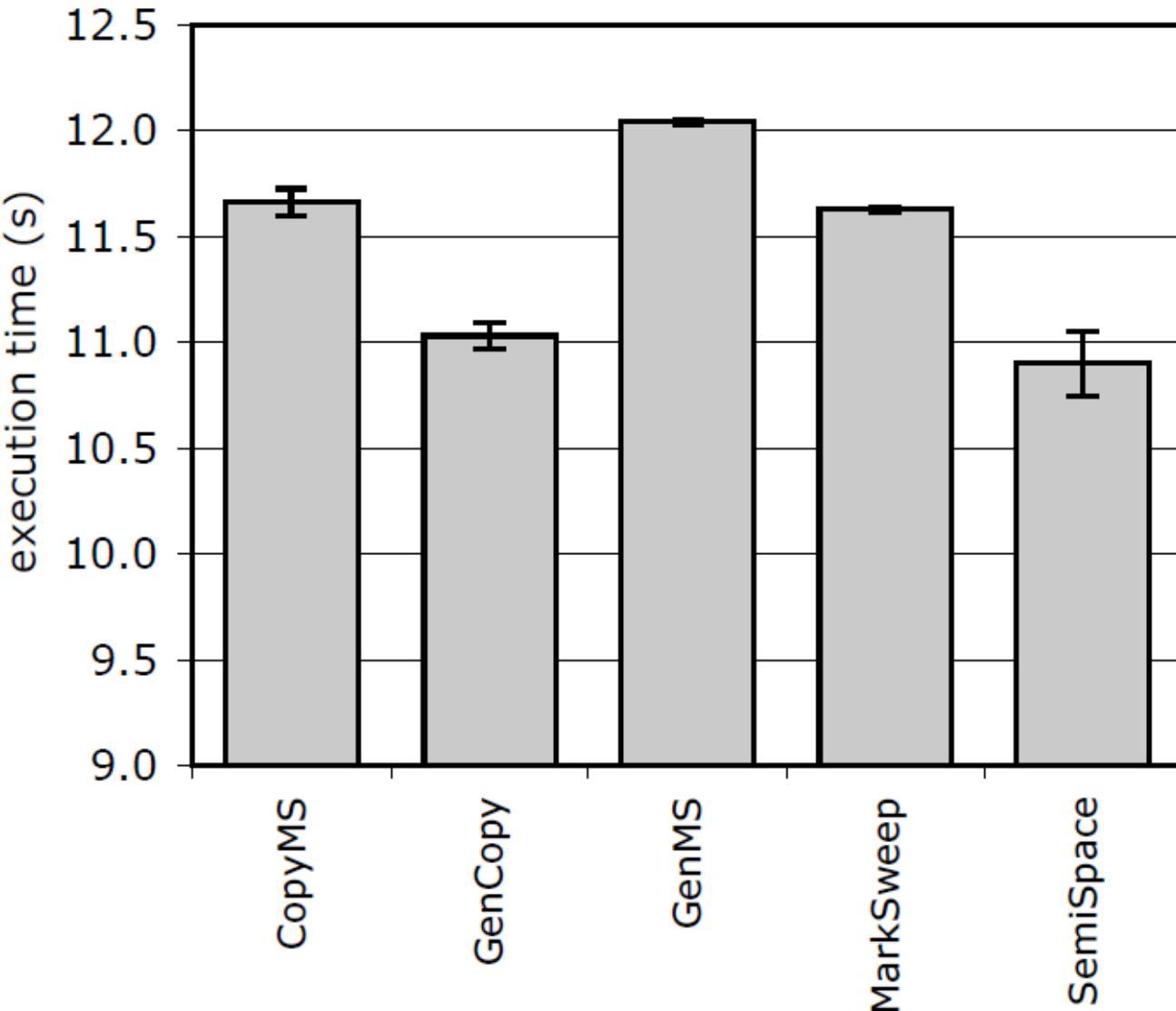
# CONFIDENCE INTERVALS



# 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)
```



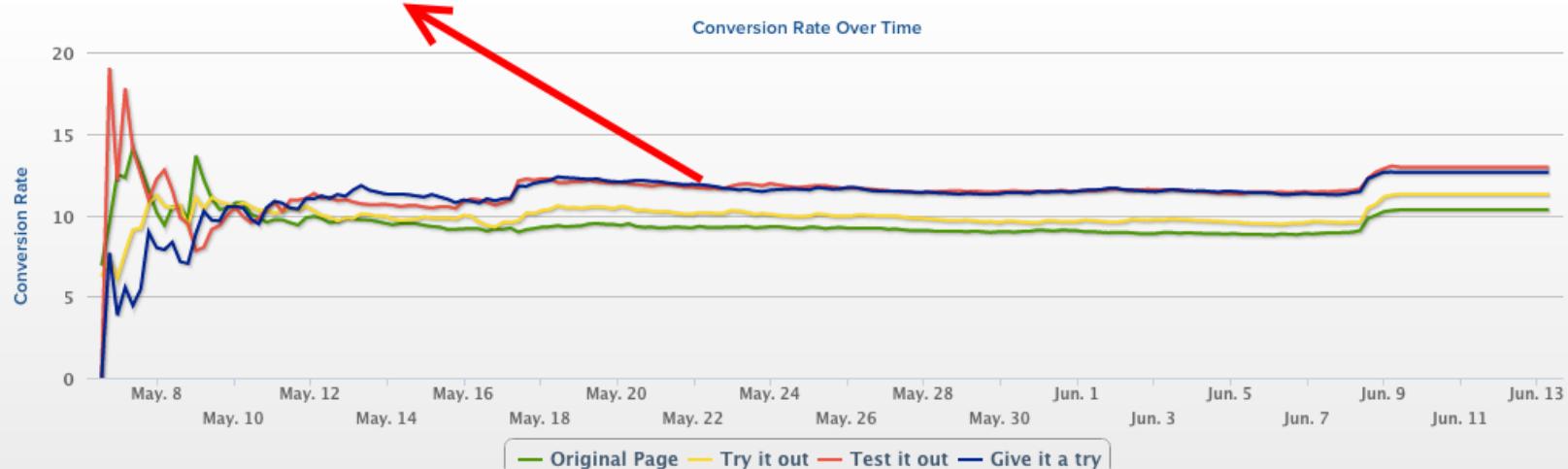
## Experiment Created

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



# 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

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

# INTERACTING WITH AND SUPPORTING DATA SCIENTISTS





A Venn diagram consisting of two overlapping circles. The left circle is light green and contains the text "Data Scientists". The right circle is light orange and contains the text "Software Engineers". The two circles overlap in the center.

**Data  
Scientists**

**Software  
Engineers**

# LET'S LEARN FROM DEVOPS



Distinct roles and expertise, but joint responsibilities, joint tooling

# SUPPORTING DATA SCIENTISTS

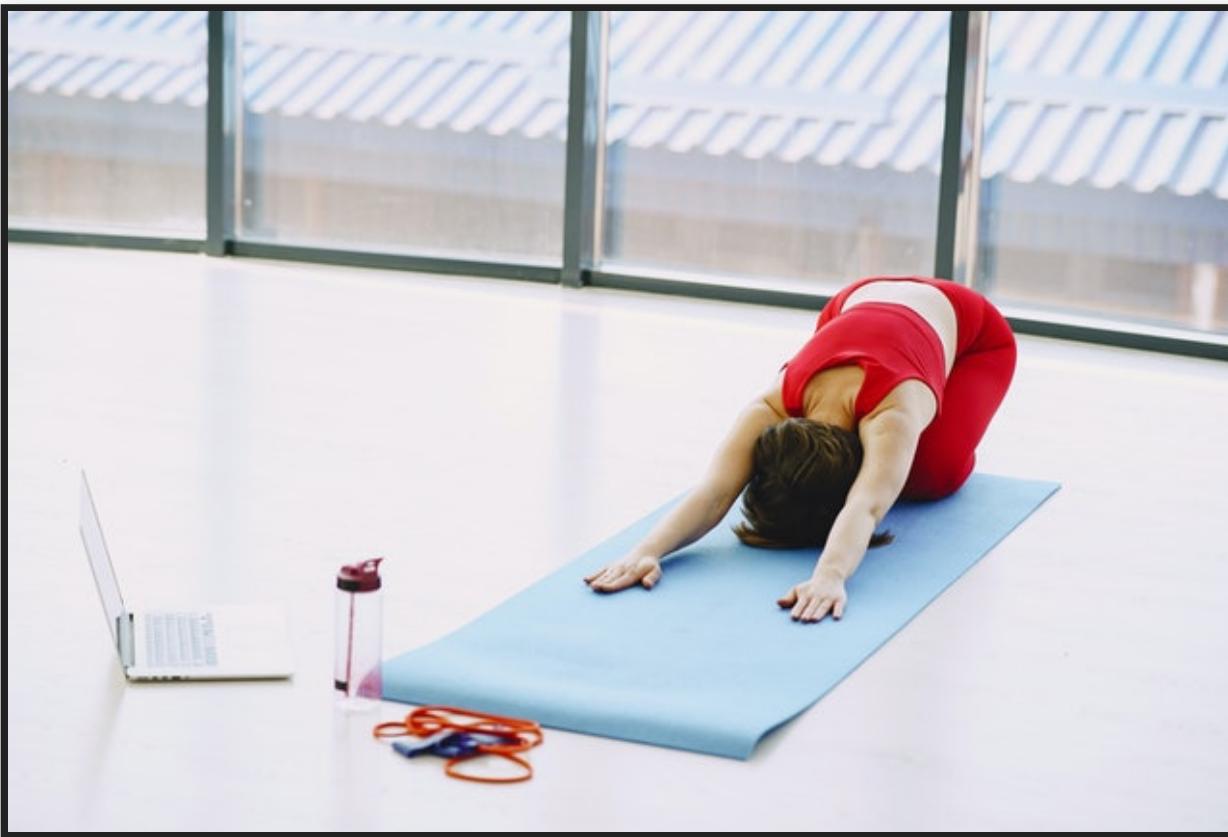
- From evaluating with static datasets to testing in production
- Provide CI/CD infrastructure for testing in production
  - make it easy to deploy and test models
- Provide access to telemetry data and dashboards
- Encourage modeling infrastructure and versioning beyond notebooks



# EXERCISE: INFRASTRUCTURE DESIGN

*Scenario: Injury detection in smart home workout (laptop camera)*

Discuss: Deployment and infrastructure decisions for A/B experiments -- how to divide users, how to implement A/B testing, what access to give to data scientists?





# 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
- DevOps-like infrastructure to support data scientists