



Decision model, meet the real world

Testing optimization models for use in production environments

September 20, 2023

Hello and welcome



Ryan O'Neil
CTO at Nextmv



Nicole Misek
Engineering VP at Nextmv

Stories to get started

We'll set the stage with stories from the field

Why testing is hard

Reasons to test and the challenges that arise

Testing and CI/CD for decision models

What an opinionated testing experience looks like

Q&A time

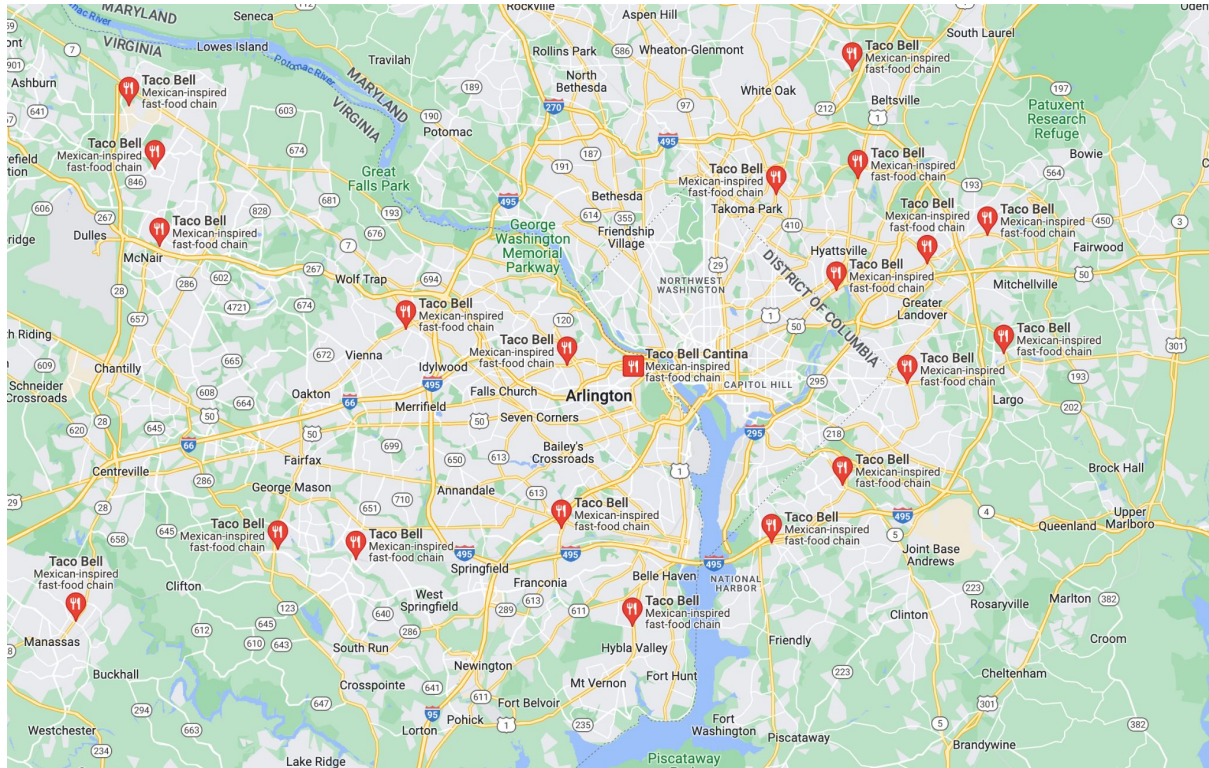
Your time to shine: ask questions, give feedback!



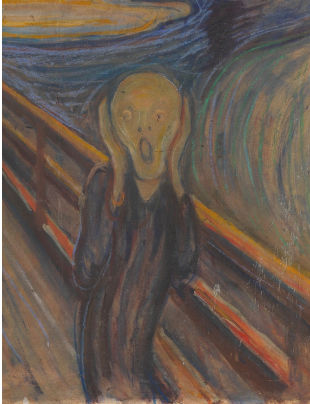
First, story time



A tale of tacos and fried chicken



One Friday night in Metropolis...



Testing is hard



What is decision model testing?

A set of techniques for **reducing risk and building confidence** to achieve stakeholder buy-in with decision models and drive improving solutions.



Why does decision model testing matter?

- Teams need a clear, repeatable path to production
- Stakeholders require confidence in outcomes
- Mistakes are expensive



Why is decision model testing hard?

- Technical setup and maintenance (SRE stakeholder)
- Complexity of analysis (OR stakeholder)
- Buy-in and alignment (Product stakeholder)
- Problem drift (Operational stakeholder)





Systems behave differently depending on environment and traffic patterns. Since the behavior of utilization can change at any time, sampling real traffic is the only way to reliably capture the request path. To guarantee both authenticity of the way in which the system is exercised and relevance to the current deployed system, **Chaos strongly prefers to experiment directly on production traffic.**

Source: [Principles of Chaos Engineering](#)



How we think about testing



Common types of decision model testing

Shadow testing

Run candidate model alongside production model

Switchback testing

Switch between treatment and control models over time

Batch experiments

Run exploratory experiments on one or more models

Scenario testing

Identify outcomes for different inputs, models, or decisions

Acceptance testing

Determine if business KPIs are met by a new model

Benchmarking

Compare scale, speed, performance of models and solvers



Decision model testing framework

“I want to compare models in using production data”



Production testing

Go/no-go for production rollout
Ex: Shadow, switchback

“I want to compare models using historical data”



Historical testing

Validate model performance on sample inputs
Ex: Batch, acceptance, scenario, benchmarking

“I want to do systematic, repeatable testing”



Input sets

Model management, sharing

Run history

Infrastructure

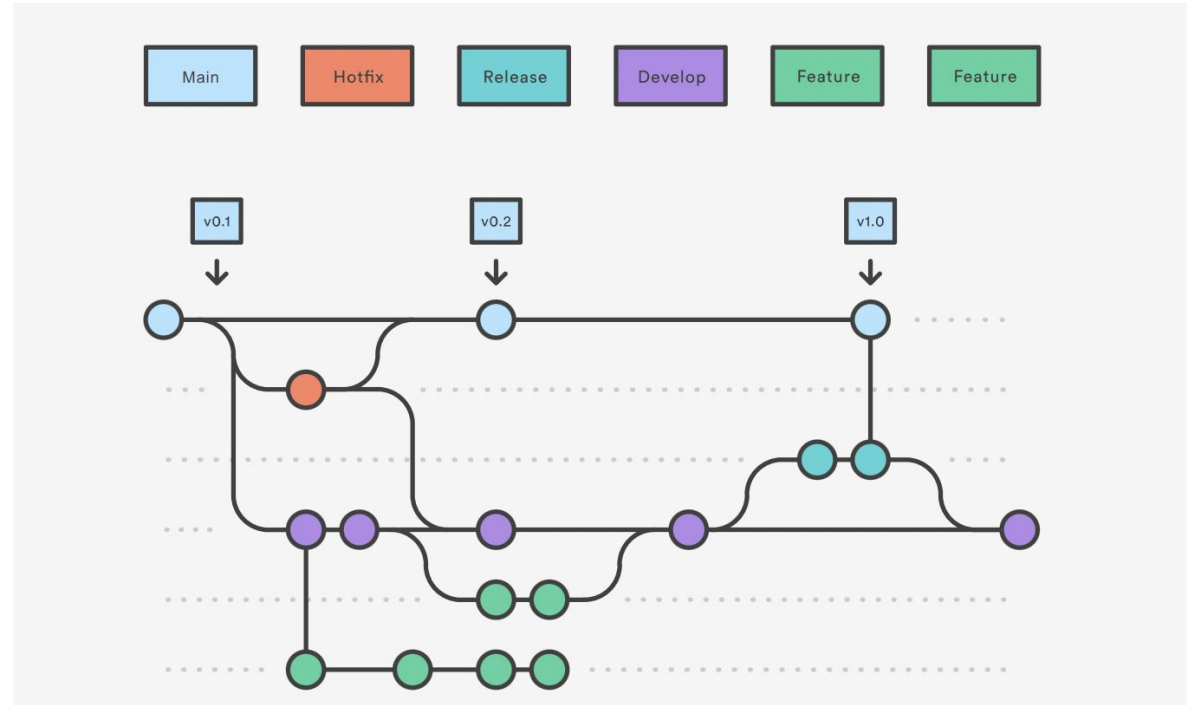
Observability

Stats Analysis

Randomizer



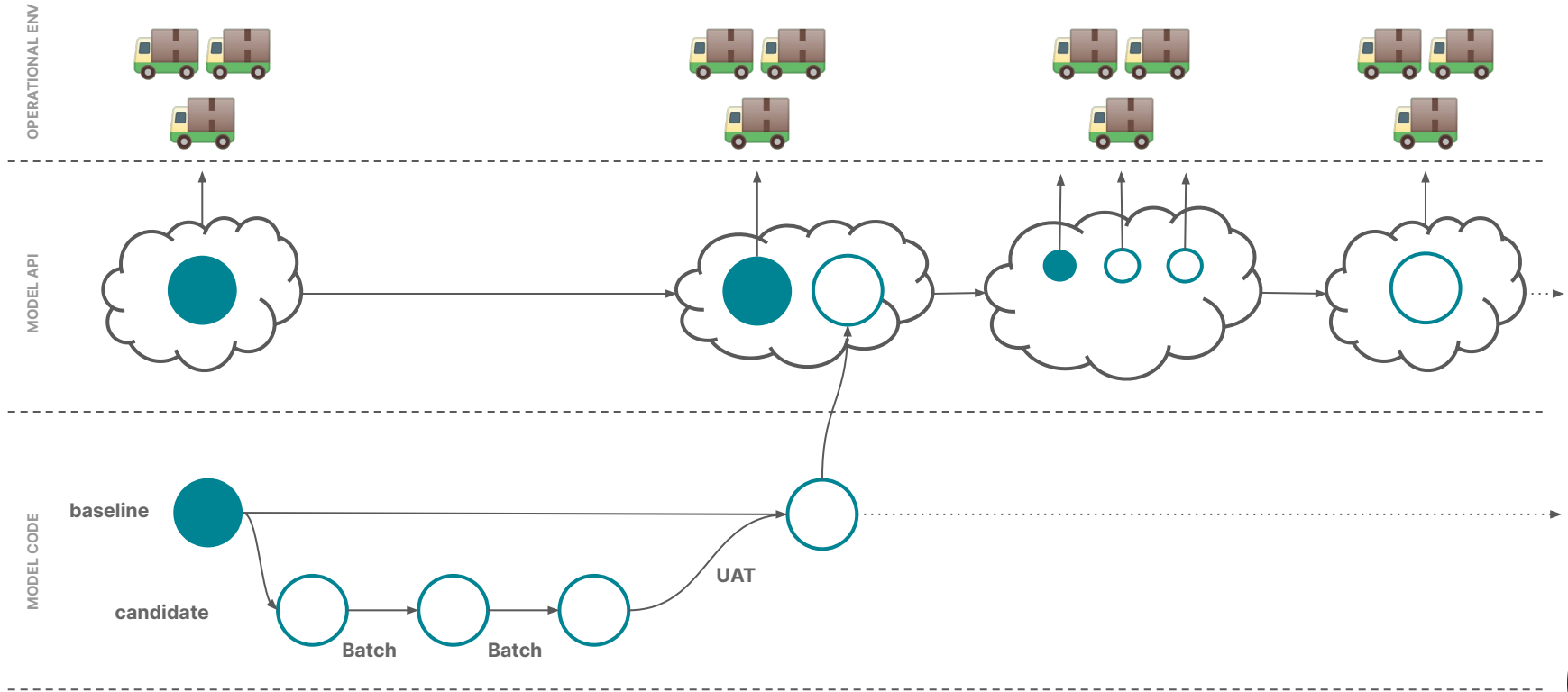
Inspiration: Git Flow



Source: Atlassian













DecisionOps model development workflow




DecisionOps test workflow in practice

benchmark



succeeded last week in 4m 30 s


- >  Set up job 2s
- >  Clean up 0s
- >  Clone 3s
- >  Clone console repository 6s
- >  Setup SSH Keys and known_hosts 0s
- >  Set go version to be latest supported 0s
- >  Set up go 0s
- >  Install CLI 6s
- >  Configure CLI 29s
- >  Run benchmark 3 m 37s

Yesterday ▾

 **Benchmark Degradation** APP 11:49 PM

Performance warning: new geometric mean is worse than 2% of the old geometric mean! (environment: dev, old version/value: 901534b / 9047118.95, new version/value: a665d59 / 9269395.81) [See details >>](#)

 1 

 [11 replies](#) Last reply today at 2:27 AM



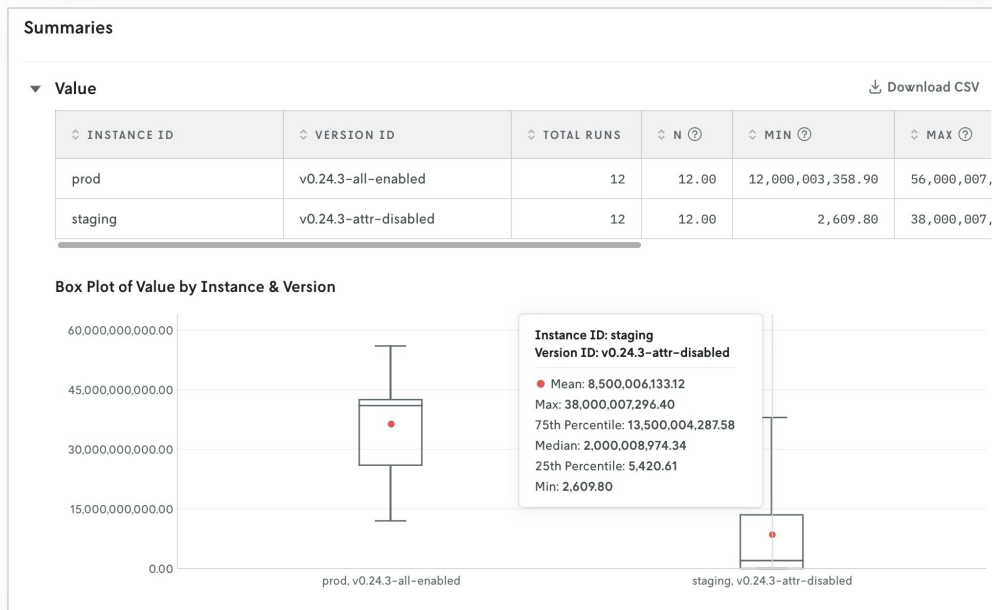
Batch experiments

Early, ad hoc exploration of model change impact across output metrics

CHARACTERISTICS

- ✗ Operational decisions
- ✗ Production conditions
- ✗ Online data inputs
- ✗ Acceptance criteria
- ✓ Historical data inputs

CONFIDENCE



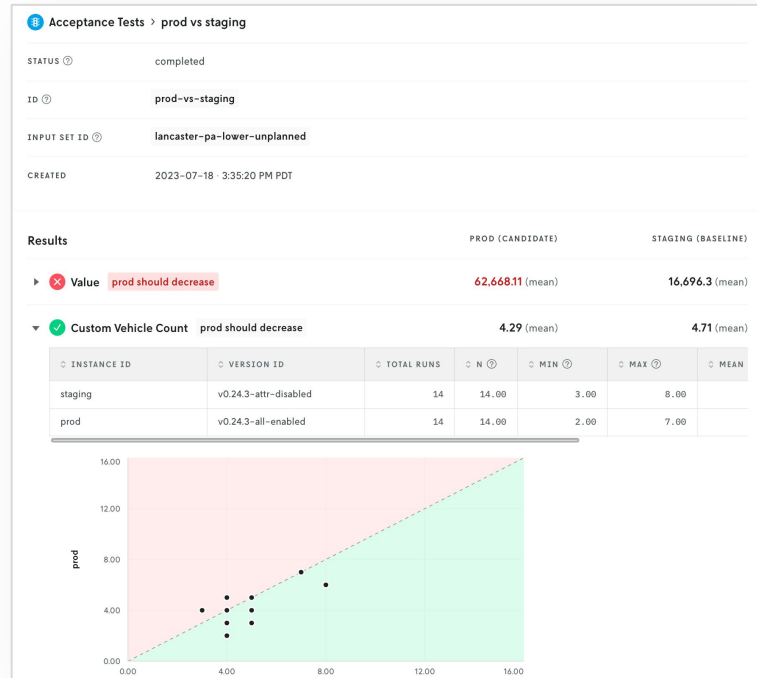
Acceptance testing

Determines if business requirements/KPIs are met by a new model

CHARACTERISTICS

- ✗ Operational decisions
- ✗ Production conditions
- ✗ Online data inputs
- ✓ Acceptance criteria
- ✓ Historical data inputs

CONFIDENCE



Scenario testing

Identify outcomes for a range of model inputs or configuration

CHARACTERISTICS

- ✗ Operational decisions
- ✗ Production conditions
- ✗ Online data inputs
- ✓ Acceptance criteria
- ✓ Historical data inputs

CONFIDENCE



```
...  
"vehicles":  
  {  
    "capacity": "20"  
  }  
...
```

```
...  
"vehicles":  
  {  
    "capacity": "300"  
  }  
...
```



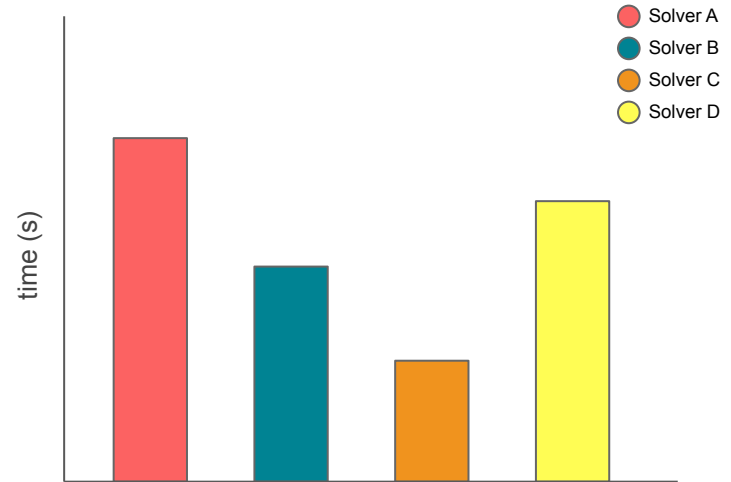
Benchmarking

Compare scale, speed, performance of models and solvers

CHARACTERISTICS

- ✗ Operational decisions
- ✗ Production conditions
- ✗ Online data inputs
- ✓ Acceptance criteria
- ✓ Historical data inputs

CONFIDENCE



Shadow testing


Run candidate model alongside production model, without production impact

CHARACTERISTICS

- ✗ Operational decisions
- ✓ Production conditions
- ✓ Online data inputs
- ✓ Acceptance criteria
- ✓ Historical data inputs

CONFIDENCE



 Shadow Test > Create new test

NAME
For reference only

ID ⓘ

DESCRIPTION
(optional)

BASELINE INSTANCE ⓘ

CANDIDATE INSTANCE ⓘ

END CRITERIA ⓘ
Maximum runs

START CRITERIA ⓘ
(optional)
Start date: Aug 30, 2023 @ 3:46:07 PM PDT
Specify in RFC3339 format, e.g. 2023-08-30T15:46:07-07:00



Switchback testing

Test treatment and control models over time/regions

CHARACTERISTICS


- ✓ Operational decisions
- ✓ Production conditions
- ✓ Online data inputs
- ✓ Acceptance criteria
- ✓ Historical data inputs

CONFIDENCE



| | Denver | New York |
|---------|--------|----------|
| 2:00 PM | App A | App B |
| 3:00 PM | App A | App A |
| 4:00 PM | App B | App A |

 App A - Staging

 App B - Production



Summary

- We shared some stories from the field
- We talked about why testing decision models is hard
- We presented a framework for how testing can fit into your decision model development workflow



QUESTIONS?



The logo for nextmv features a stylized white icon of two leaves or petals on the left, followed by the text "nextmv" in a bold, lowercase, sans-serif font. The logo is centered horizontally across the image.

nextmv

Components to test on

APPS



~ REPOS

VERSIONS



~ RELEASE SNAPSHOTS

INSTANCES



~ BRANCHES

