

DATA-DRIVEN SIMULATION IN E-COMMERCE FULFILMENT

From Simulation Twinning to Factory Control

A nine-month journey from model build to production evidence

AnyLogic Conference 2026 | Simulation & Digital Twins | 15 September 2026

20-minute presentation | Questions at the end

gousto



Two organisations – one joint team

From operational challenge to model engineering and production evidence

gousto

Tech-led meal-kit operator

Live operations • data • decision services • deployment

ONE POD

DECISION LAB



Applied decision intelligence

Simulation & Digital Twin Engineering • AnyLogic Experts



Dr Fabrice Durier

Senior Principal Data Scientist
Gousto

Business context, OR framing & future value



Peter Riley

Head of Applied Solutions
Decision Lab

Platform choice, delivery & architecture



Will Seymour

Senior Simulation Scientist
Gousto

Validation, experimentation & production impact



Jacob Whyte

Simulation Engineer
Decision Lab

Inventory storage & routing logic



Andrew Nestor

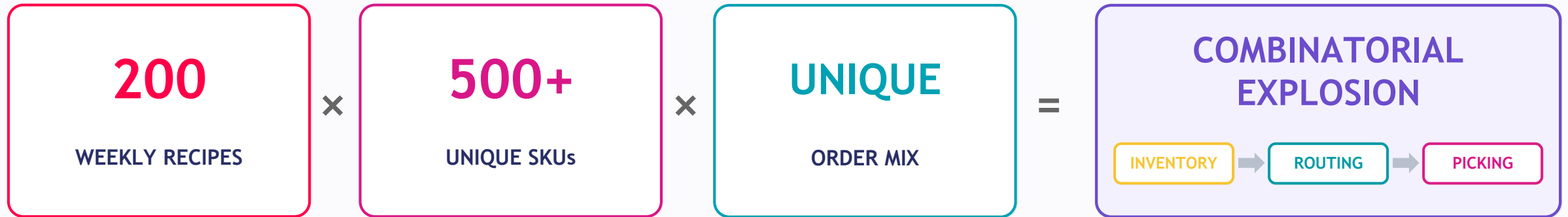
Simulation Engineer
Decision Lab

Mechanical reliability

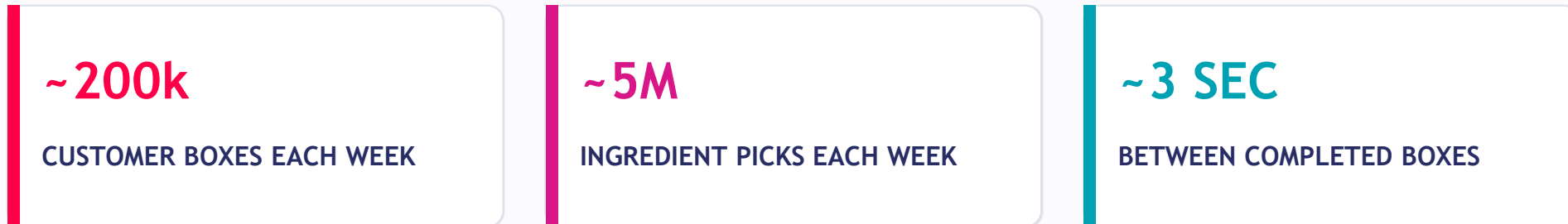
Operational knowledge + specialist simulation engineering + a direct route to implementation

Customer choice creates exponential operational complexity

Every customer box has its own recipe mix — yet Operations must deliver millions of picks at pace.



While the Operations must deliver



Every Decision, whether Strategic or Tactical can have a large impact on our Operational performances

The Latest Chapter in Four Years of Simulation Development

A mature red-box model had already proved the value of simulation; *ToteSim* closed the inventory-supply gap.



100s-1,000s
simulated per hour

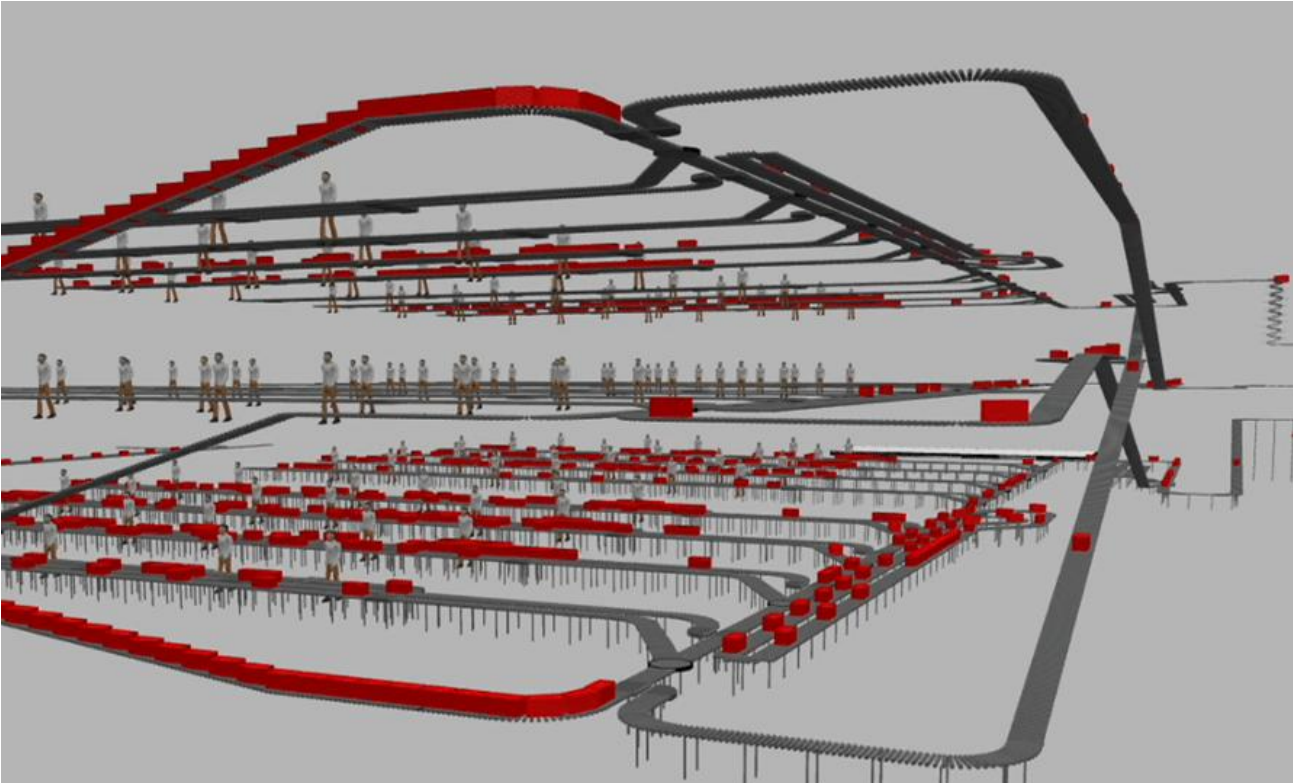
Several thousand
runs in a typical month

Hundreds of millions
customer boxes simulated

Tens of billions
ingredient picks simulated

The customer red box was only half the system

Our mature picking model still relied on simplified assumptions about the inventory flow that fed it.



LIVE OPERATIONS + HISTORICAL DATA + CONTROL LOGIC

18,000

reserve locations

9,800

active pick slots

20

vertical lifts

270

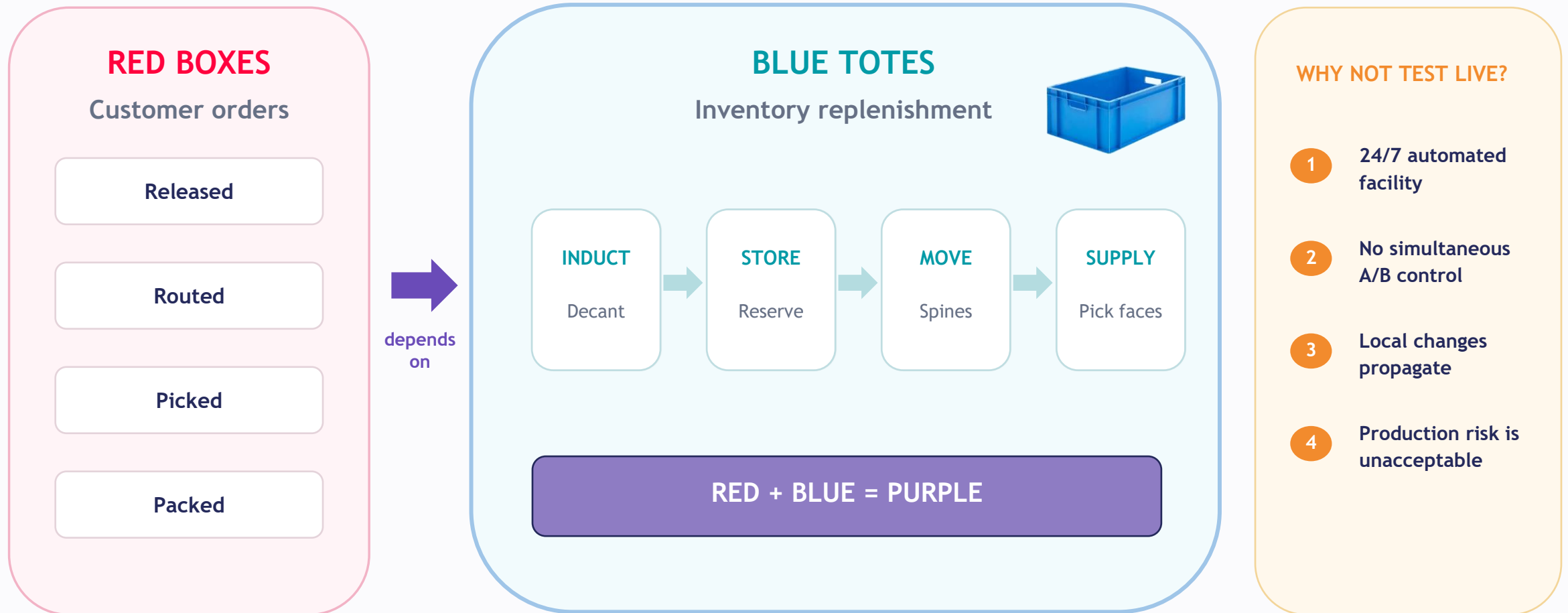
automated shuttles

THE REAL DIFFICULTY

Reproducing how equipment, inventory and stock decisions interact over time.

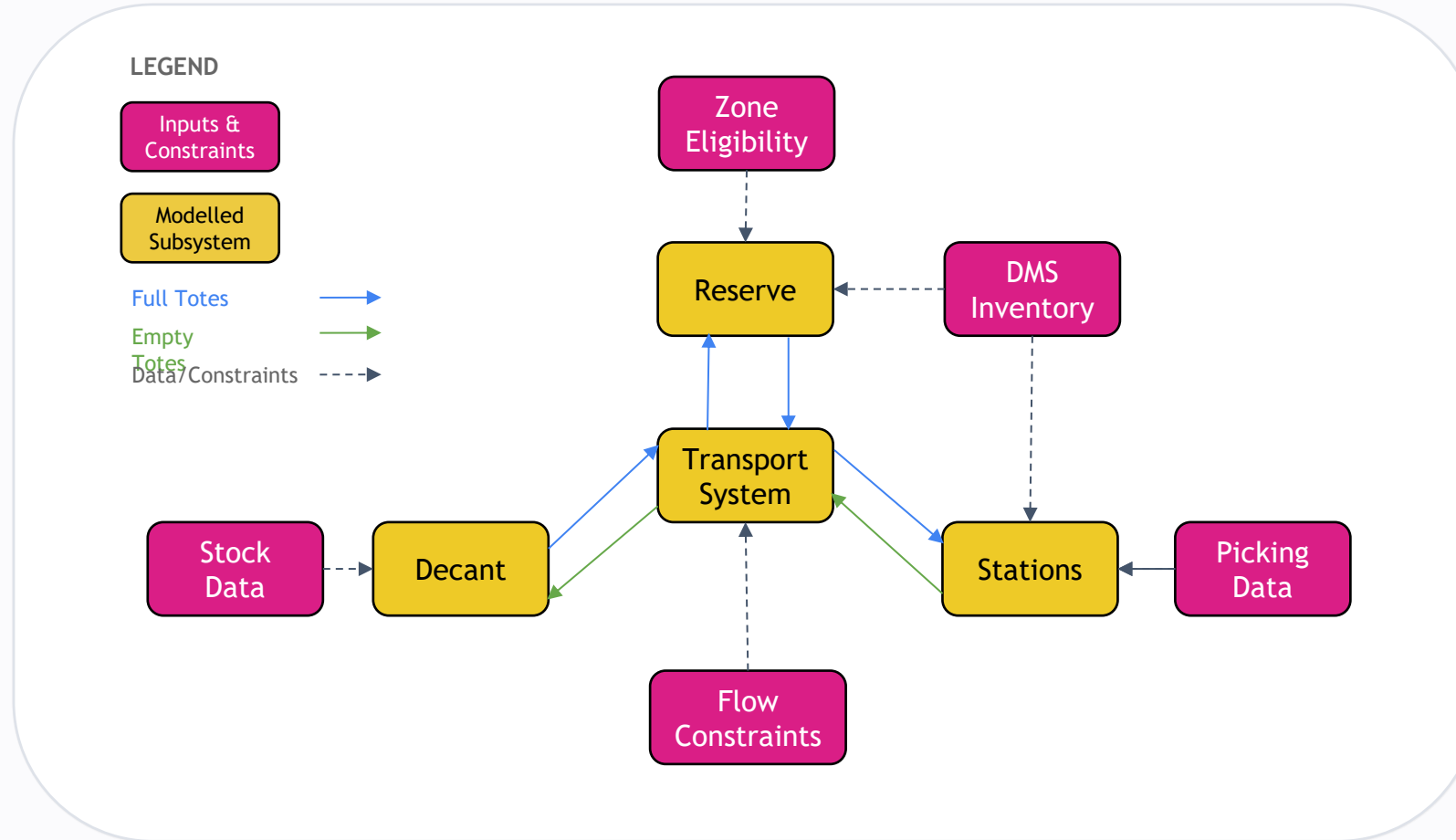
The missing problem coupled live operations, data and control logic

A full transport system—not a single queue or machine—had to be represented as one decision environment.



Conceptualising the system as a 'Golden Triangle'

Decant, reserve and stations formed the model boundary; the shared transport system connected them.



BUILD

Three bounded operational domains

EXPERIMENT

Isolate or recombine subsystems

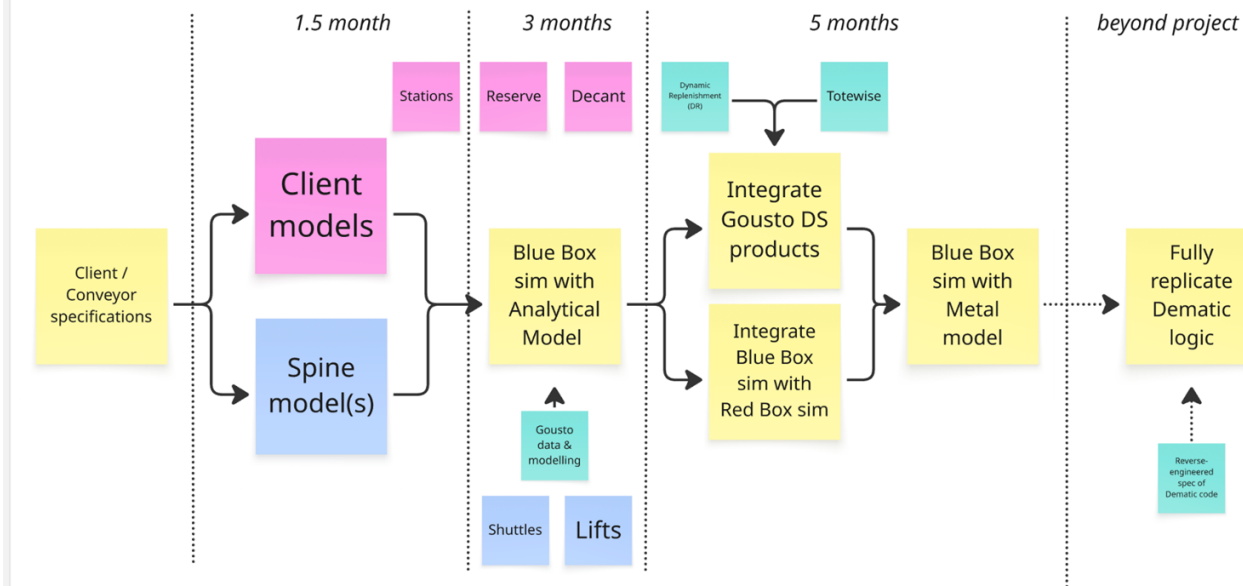
VALIDATE

Replay each edge, then the whole system

Collaborative delivery made a five-month build possible

The speed came from an integrated pod, deliberate decomposition, and modular workstreams.

Milestone Deliverables



Strong initial vision + iterated delivery plan + specialist simulation skillset

5 months

rapid build

24 epics

delivery decomposition

300+

technical tasks

21

user stories

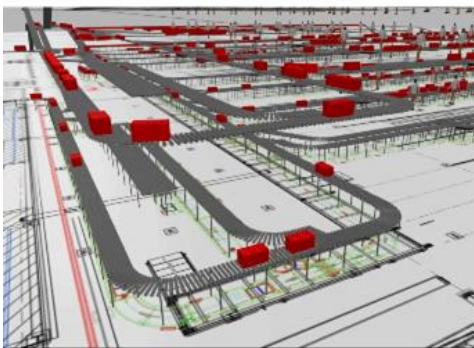
ONE INTEGRATED POD

Shared Slack, daily working rhythm, fortnightly retrospectives, common code, and robust challenge.

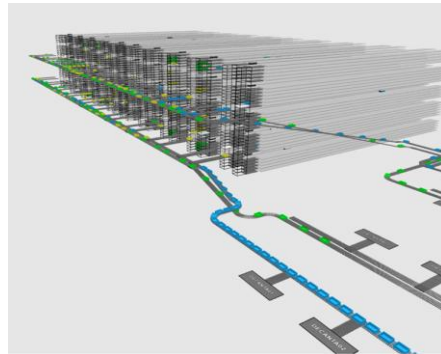
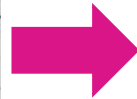
AnyLogic was the natural extension of a proven platform

Continuity with the mature red-box model came first; product capabilities and Decision Lab expertise accelerated the build.

PLATFORM CONTINUITY



Mature red-box model



Extended with blue-tote model



Same platform, with shared
knowledge and model interoperability

1

Material-handling operations

Built-in conveyor, routing, capacity, and merge behaviour reduced low-value custom engineering.

2

Agent extensibility

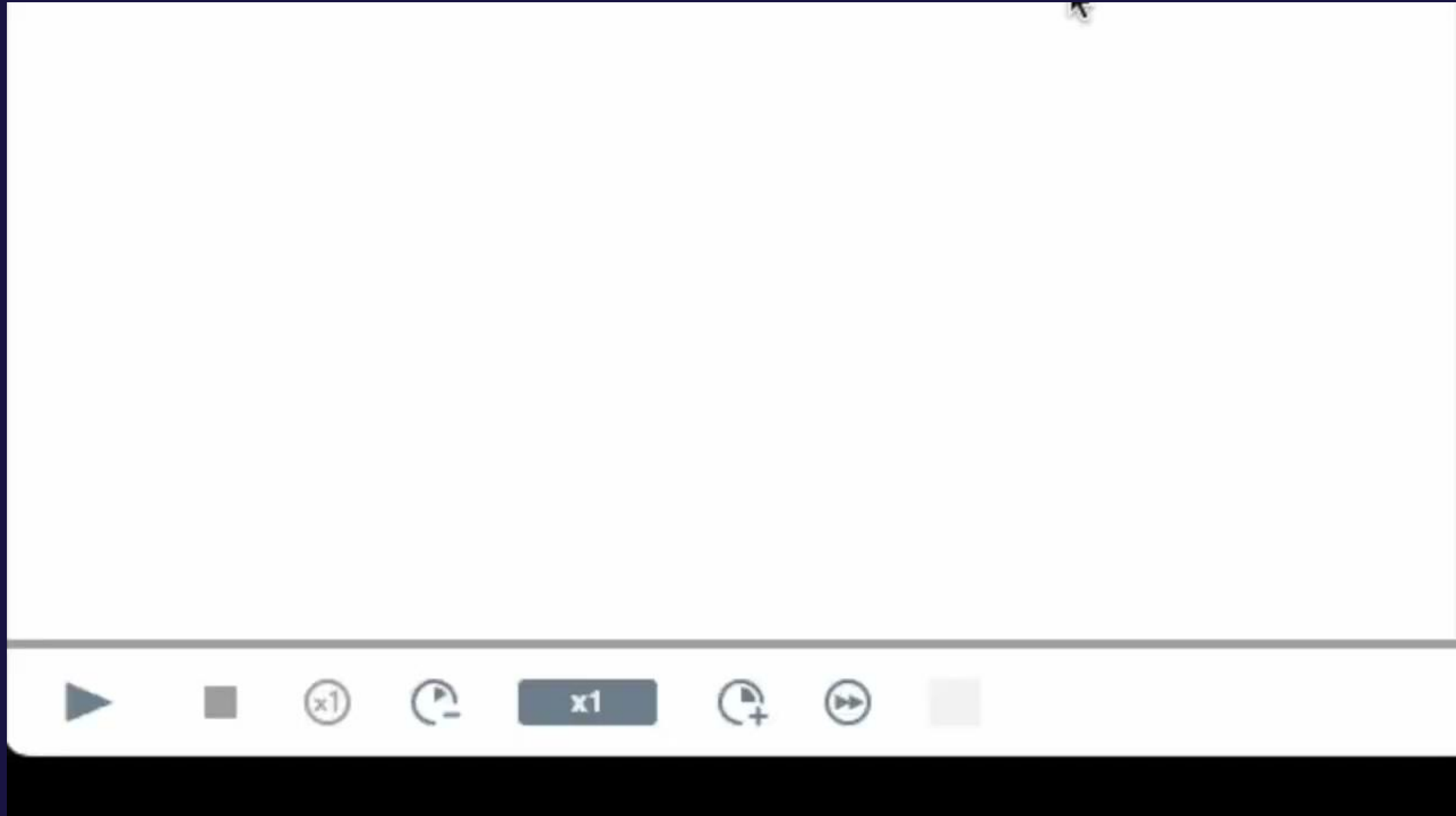
Lifts, shuttles, and failure behaviour could be modelled with bespoke Gousto logic.

3

Enterprise integration

Java execution connected the model with Databricks & Python services, and enabled scalable experiments.

A full *ToteSim* flyover



Full totes • empty totes • reserve aisles • lifts • shuttles • shared spines • pick-station supply

Inventory storage is a decision system, not just geometry

ToteSim represents 27,800 reserve and pick locations – and the logic that decides where inventory goes.

18,000 reserve locations

9,800 active pick slots

Aisle

Level

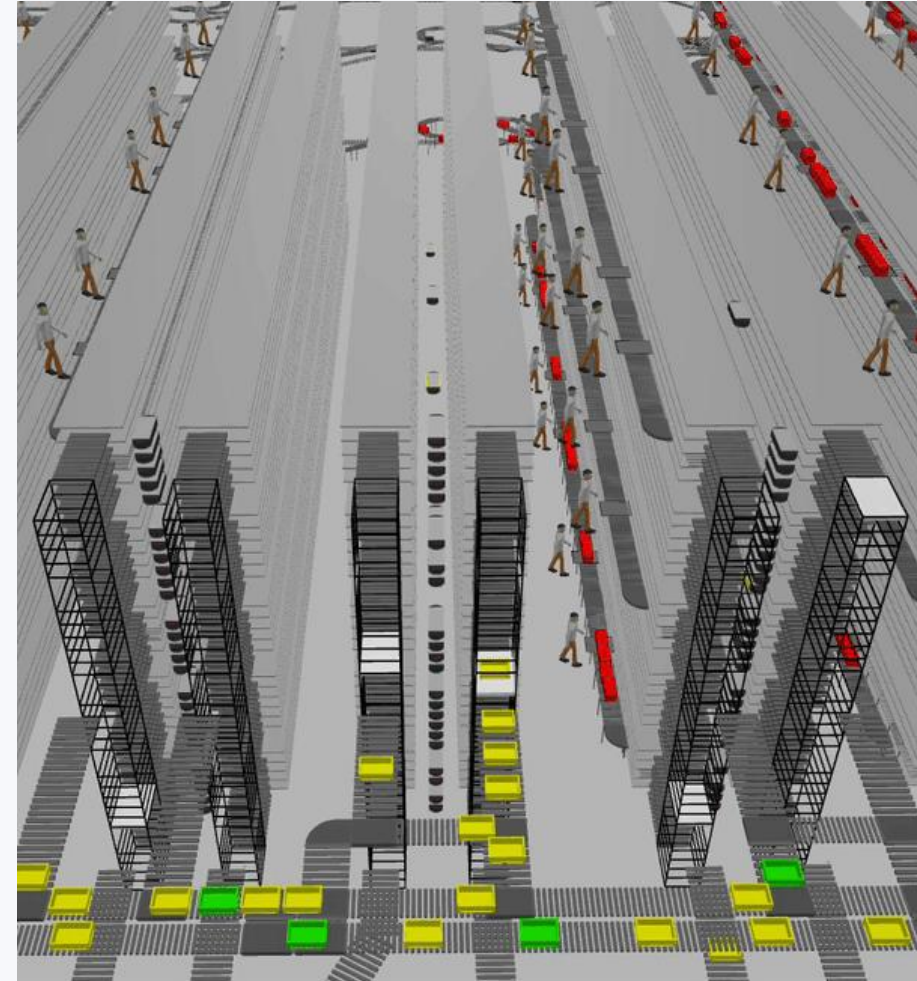
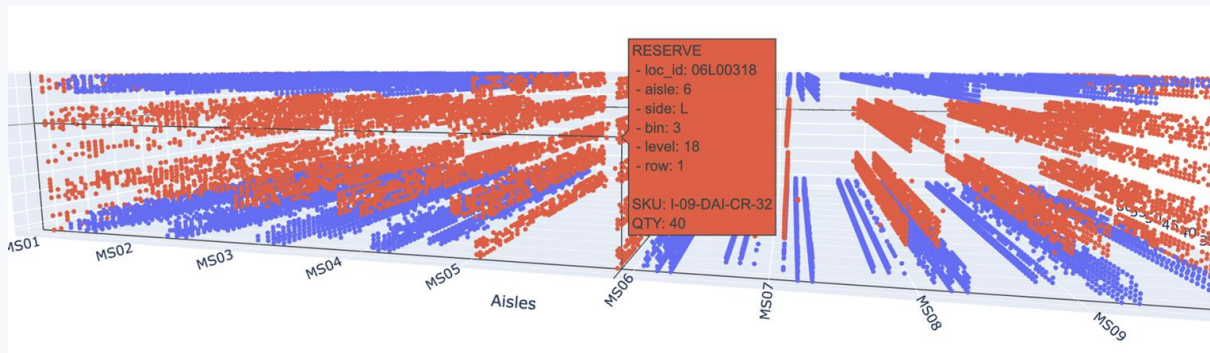
Bin

Spine

Replen.

Storage choices affect future travel distance, inventory availability and congestion

The model has to recreate both the map and the decisions taken on that map



Mechanical reliability is modelled asset by asset

Every lift and shuttle can replay real downtime or be stress-tested synthetically

20

vertical
lifts

270

automated
shuttles

HISTORIC

Replay actual downtime

SYNTHETIC

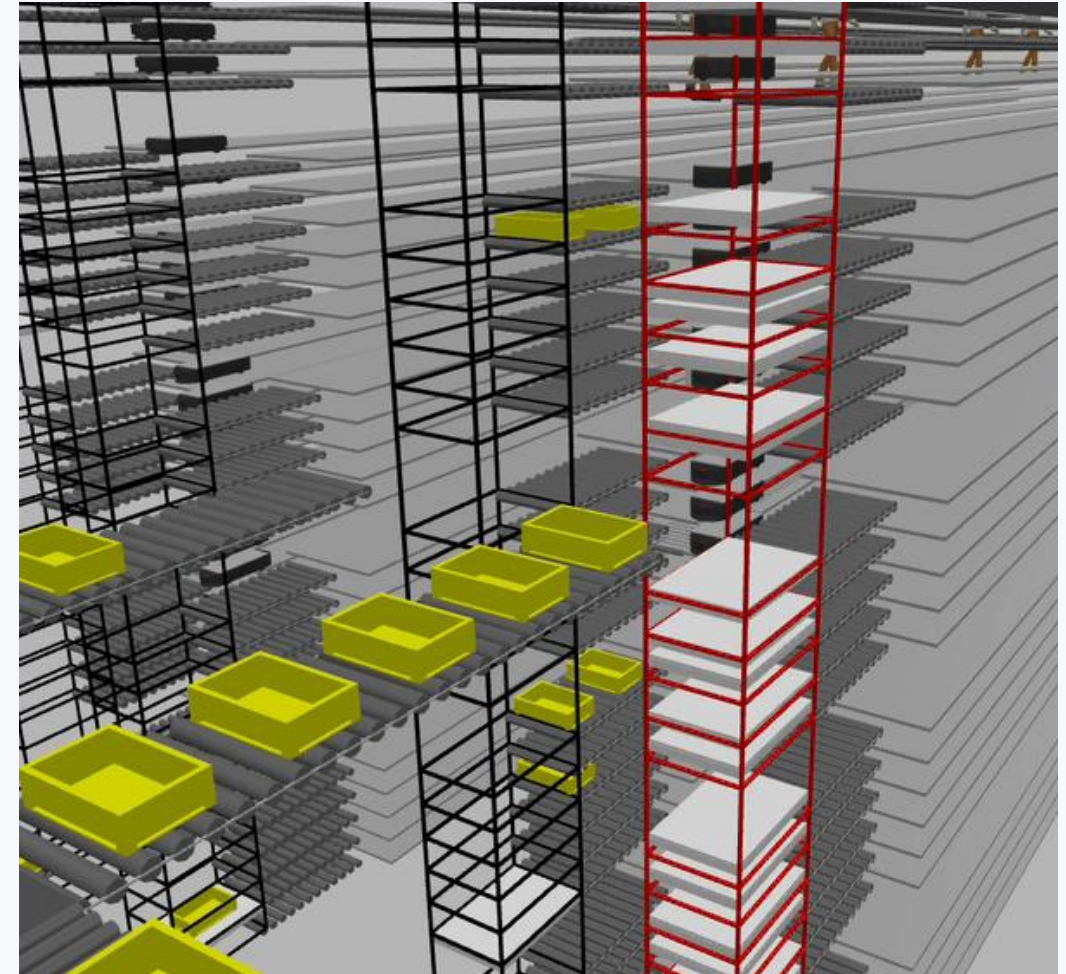
Force selected assets offline

STOCHASTIC

Sample failure behaviour

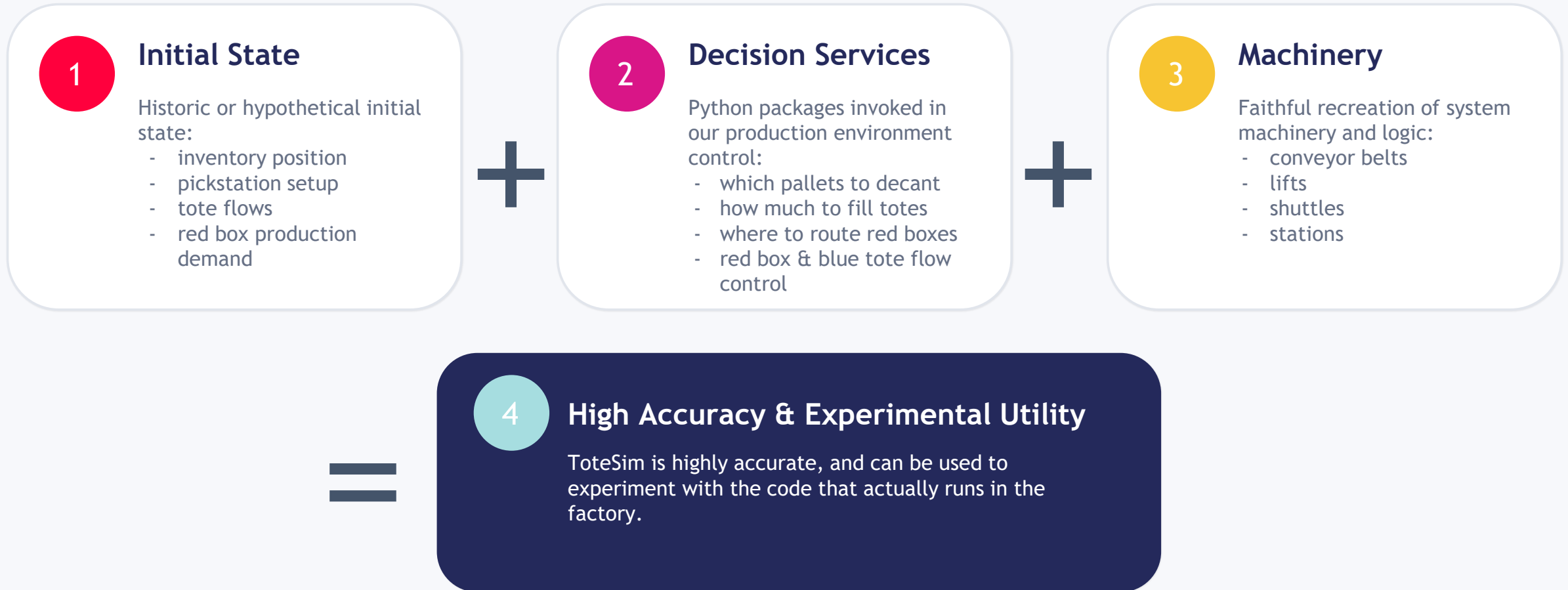
Reliability becomes an experimental variable
not uncontrolled noise.

Useful for replaying the past and stress-testing resilience before the factory sees the disruption



Integration of production decision-making code

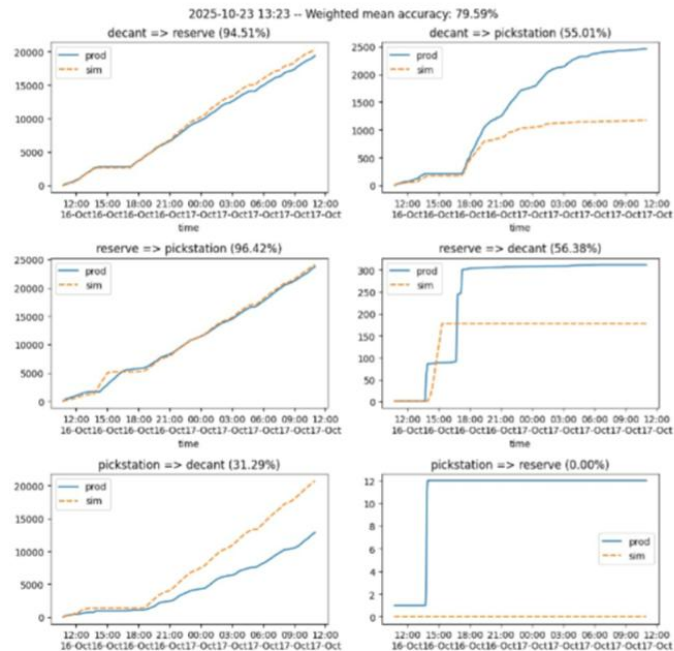
ToteSim invokes the same decision-making code that runs in our production environment, making it highly accurate and experimentally useful



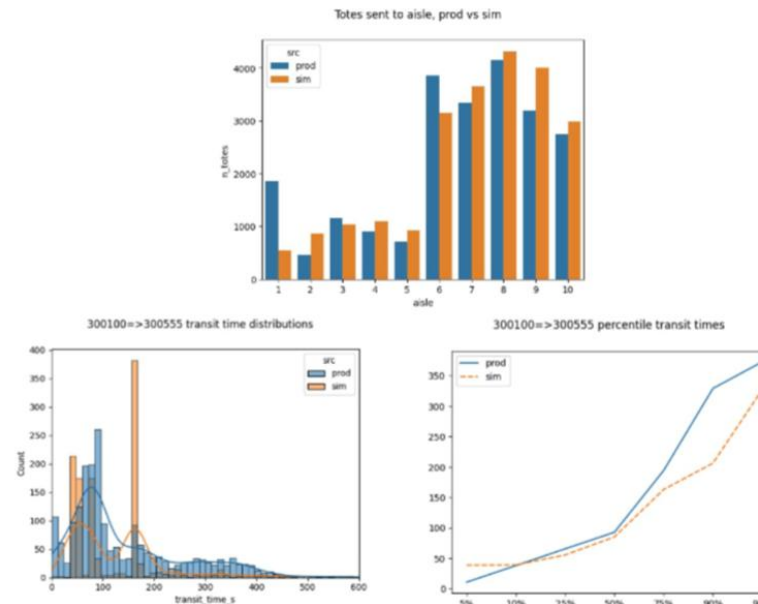
Trust was built for a decision – not claimed globally

Could *ToteSim* reproduce the operational behaviours needed to compare replenishment and control policies?

Headline Accuracy



Granular Accuracy



~80% accuracy

broad network measures

Weighted replay accuracy across multiple flows.

>95% accuracy

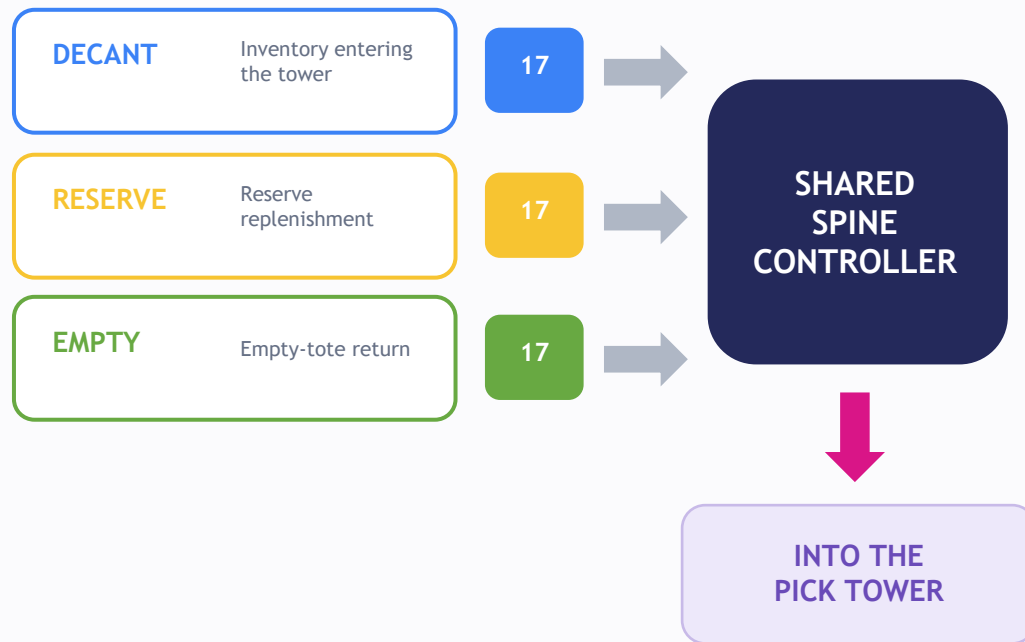
selected critical pathways

High-volume reserve-to-pick and decant-to-reserve flows.

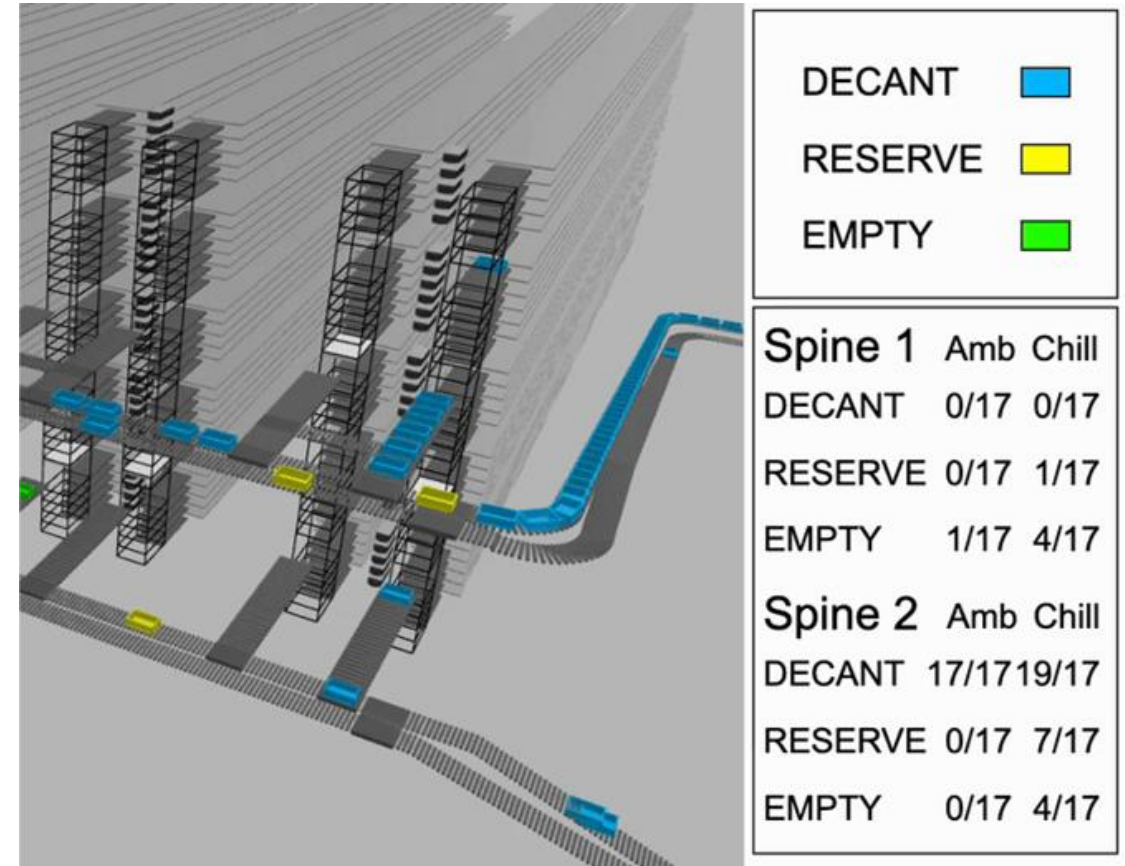
Could smarter control move more inventory into the pick tower?

The shared tote spine controller creates queues even when transport capacity is still available.

THE CONTROL PROBLEM



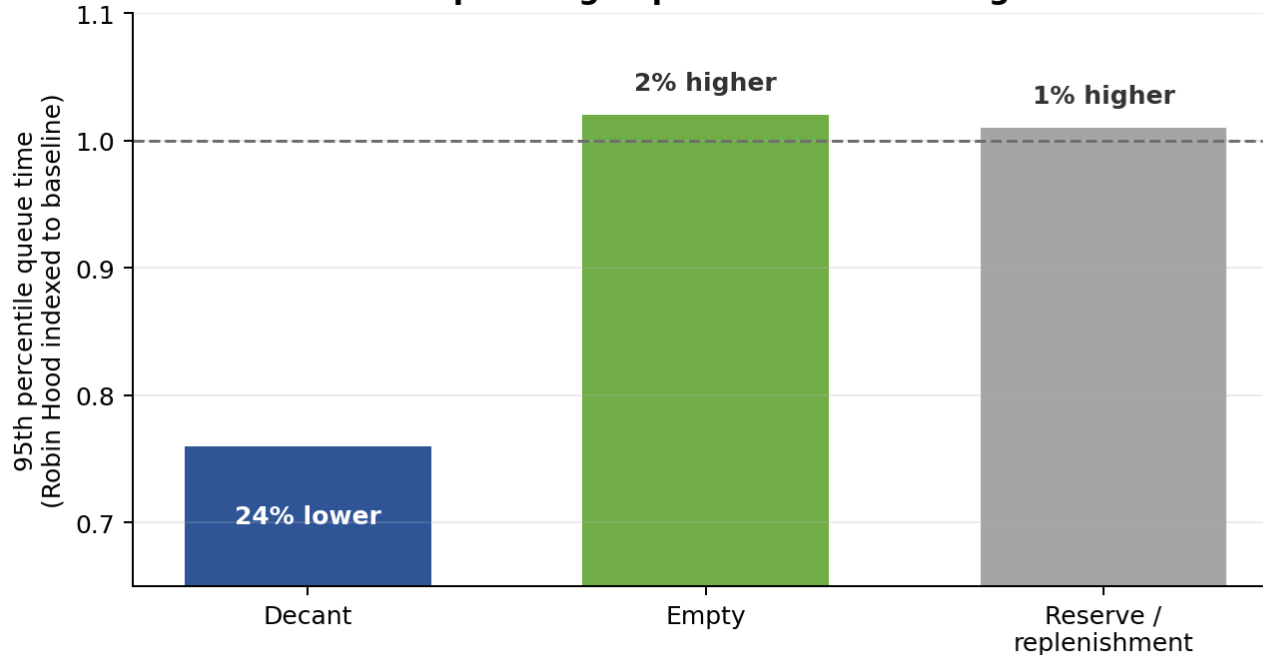
A tote stream could hit its category limit and queue while capacity remained elsewhere.



The new algorithm makes the difference during peak times

Normal movement stayed stable; the constrained tail improved where the controller was designed to act.

Production evaluation: tail queueing improves where the algorithm is intended to act



Source: six-week before/after evaluation; every tote using the spine included (>2 million journeys). Lower is better.

Lower is better. The decant tail improved while empty and reserve/replenishment remained broadly stable.

6-WEEK BEFORE / AFTER
NOT A CONTROLLED A/B TEST

>2M

live tote journeys

All spine journeys across six-week periods before and after deployment.

24% lower

decant p95 queue time

The largest effect appeared in the high-pressure tail.

THREE TAKEAWAYS

Simulation became an operational decision capability

Not a one-off model. A reusable path from operational question to production evidence.

1

Build on proven capability

ToteSim extended four years of red-box simulation — it did not start from a blank page.

2

Put real decisions inside the twin

Operational data, the AnyLogic model and Gousto decision services ran as one executable environment.

3

Carry evidence into production

The same control logic moved from experiment to deployment and was evaluated on live journeys.

5 months

to build *ToteSim*

2,000+

simulation runs

2M+

live tote journeys

24% lower

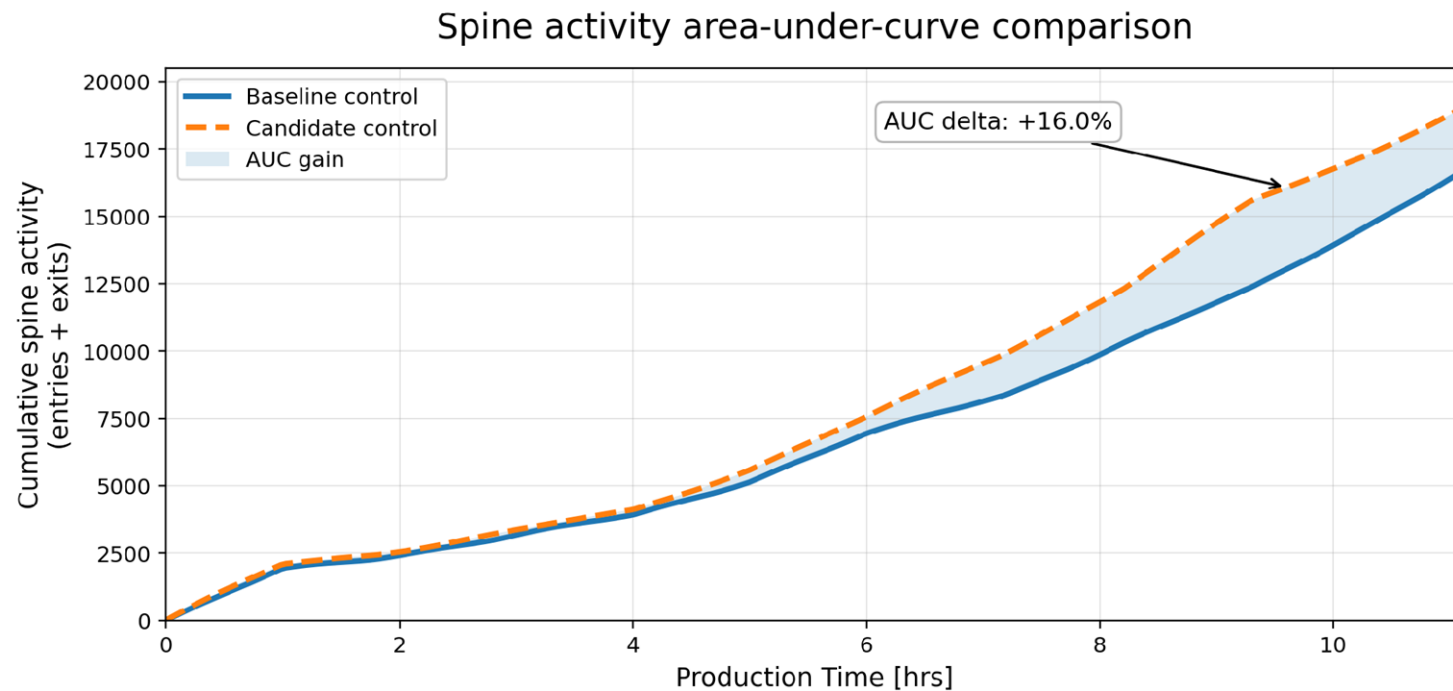
decant p95 queueing

Questions?

From Simulation Twinning to Factory Control

2,000+ experiments made a subtle effect measurable

The metric asked a practical question: did the same required work happen earlier, with less avoidable waiting?



Illustrative redraw of the Use Case 6 spine_activity metric: higher or earlier cumulative activity increases AUC.

A higher cumulative activity curve means the spine transported more totes and/or transported them earlier.

2,000+

simulation runs

3x

replications per scenario

$p < 0.001$

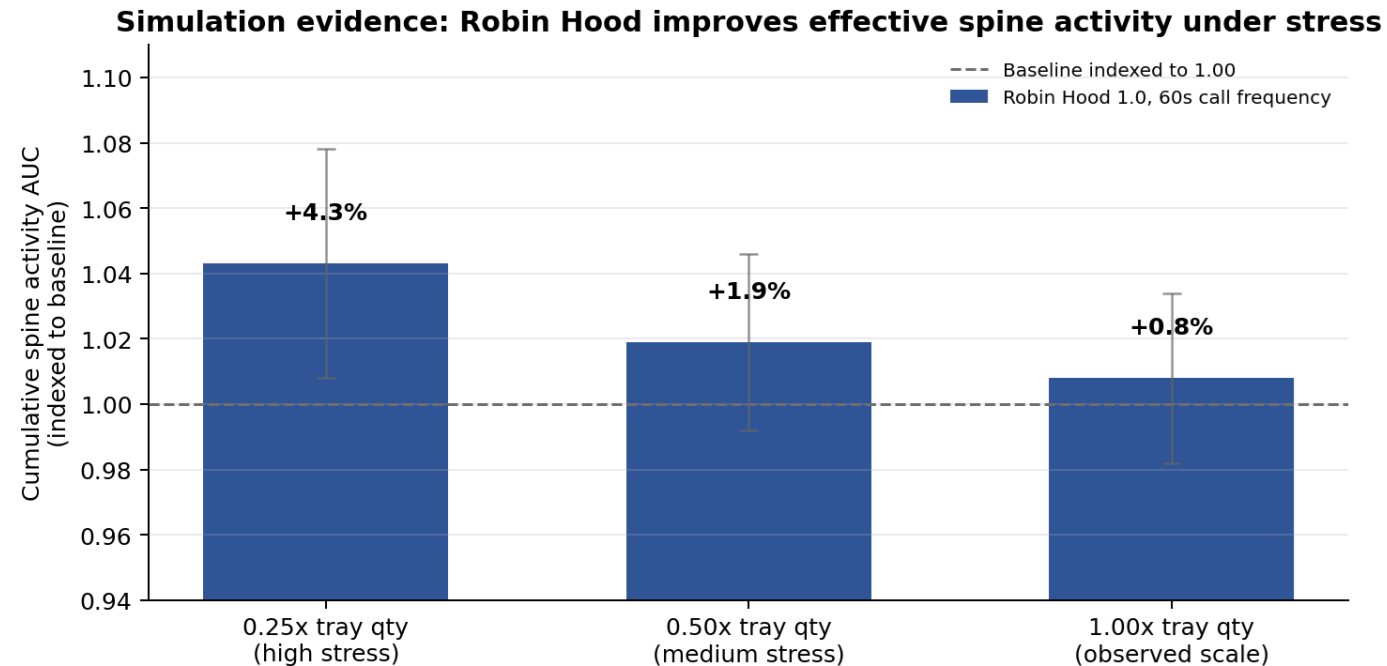
higher-stress comparisons

+4.3%

best mean AUC result

Our algorithm worked, and was a drop-in replacement

Headroom balancing redistributed spare category capacity while preserving the existing interface.



HEADROOM BALANCING

When total occupancy is below target, equalise the remaining headroom across tote categories.

DROP-IN DESIGN

Same inputs • same outputs • no extra sensors
• no new control-system interface.

TWIN



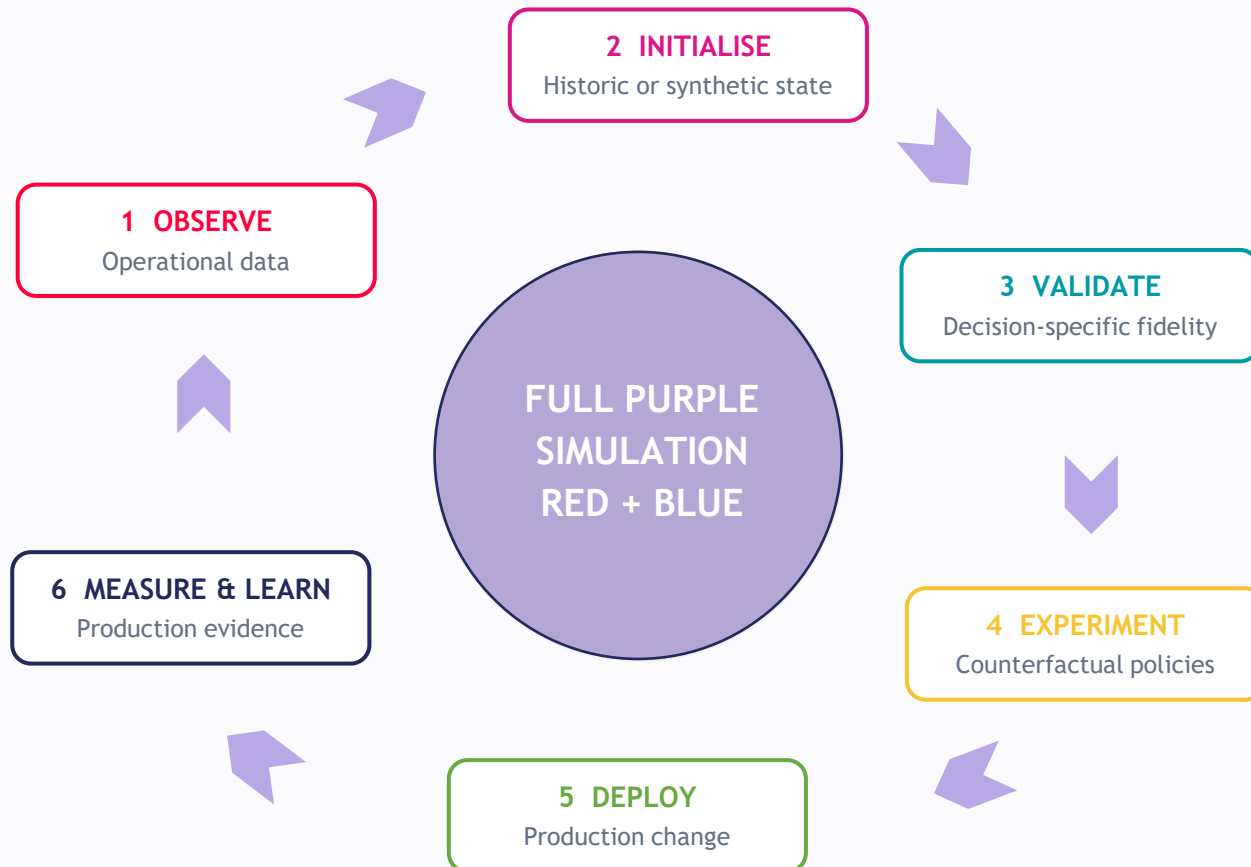
SAME PYTHON LOGIC



FACTORY

The full Purple model turns new questions into fast learning loops

Observe → initialise → validate → experiment → deploy → measure → learn.



CURRENT PURPLE-SIM INVESTIGATIONS

Prioritise pallet movement

Use downstream demand and current system state.

Allocate decanting capacity

Direct constrained induction capacity to the most valuable flows.

Fine-tune storing & spreading

Balance inventory across reserve and active pick faces.

NONE OF THESE END-TO-END QUESTIONS COULD BE TESTED BEFORE THE FULL PURPLE MODEL.

Plain-English glossary for the main system terms

Red box

A customer order moving through the picking line

Blue tote

A container carrying ingredient inventory

Decant

Inventory induction into the automated system

Reserve

High-density automated storage and retrieval

Pick station

Where ingredient inventory meets the customer box

Spine

The shared conveyor loop used by multiple streams

Headroom

Remaining allowance before a tote-category limit is hit

Simulation twin

A validated, data-integrated model used to replay and test operational decisions