Applications of Causal Machine Learning in Building a Unified Metric System

Wenjing Zheng

3rd Workshop on Causal Inference and Machine Learning in Practice

KDD 2025

ROBLOX

Collaborators

Wally Toh

João Dimas

Phil Hebda

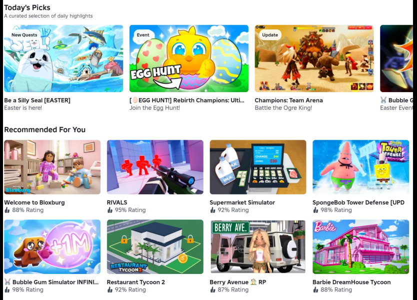
Yidi Wang

Zhenyu Zhao

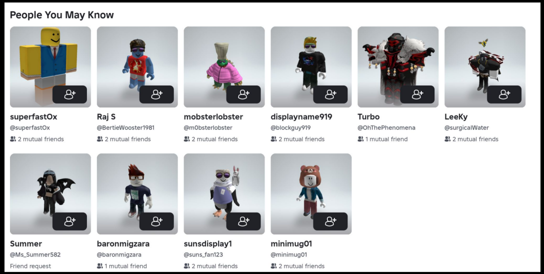
Roblox mission: connect a billion people every day with optimism and civility

Why a Unified Metric System?

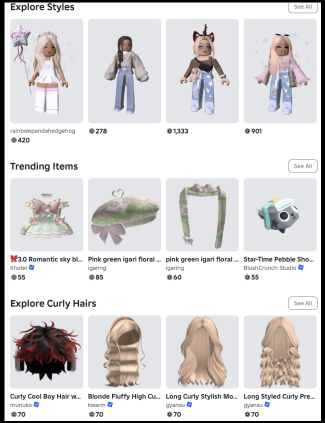
- Common language vs conflicting local signals
- Faster decision-alignment
- Moving all areas towards the same goal



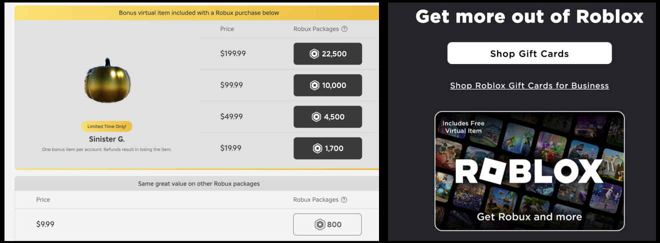
Experience Discovery



Social network



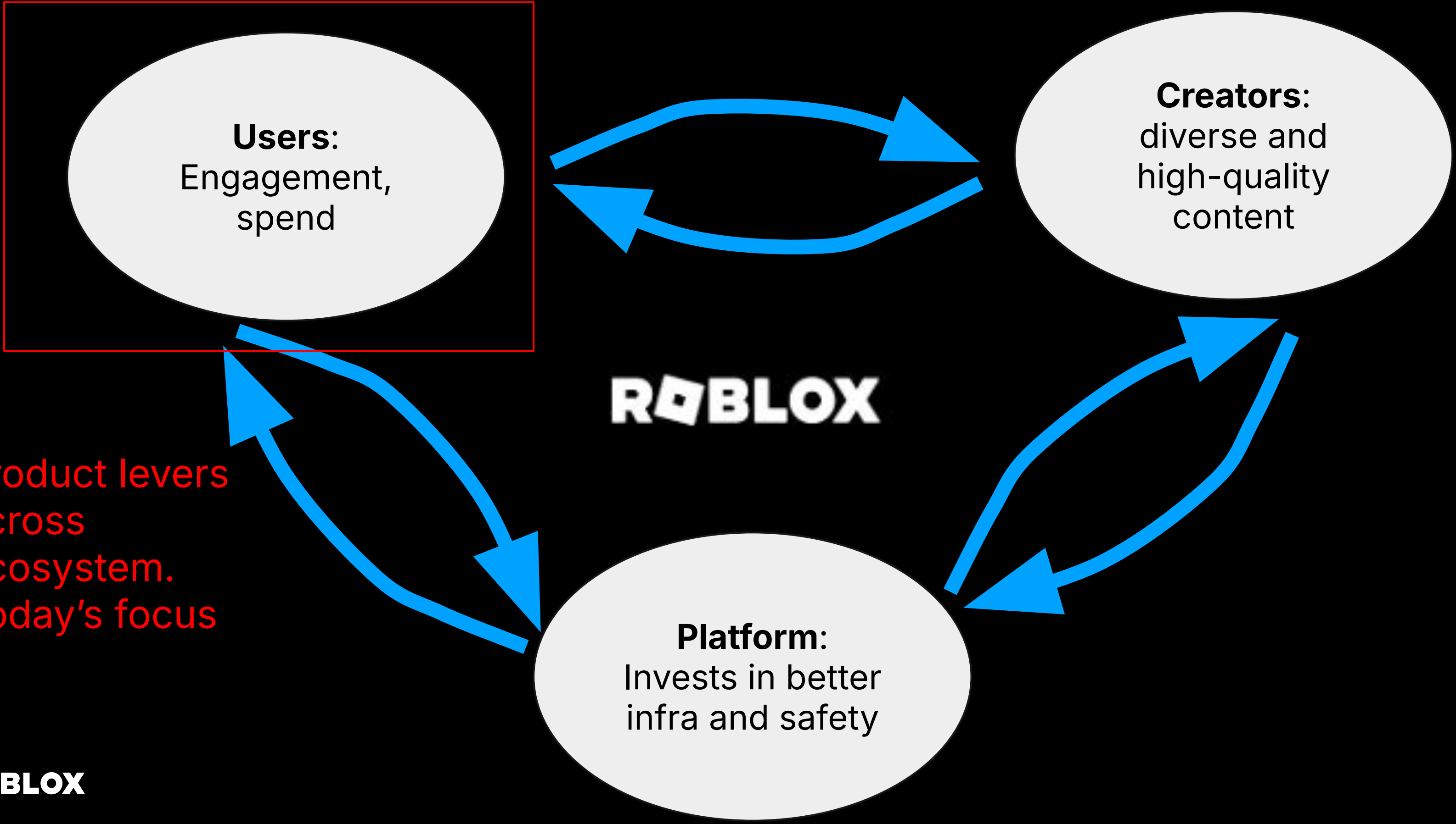
UGC avatar items



Virtual Money: Robux

Technical performance

Roblox Ecosystem



- Product levers across ecosystem.
- Today's focus

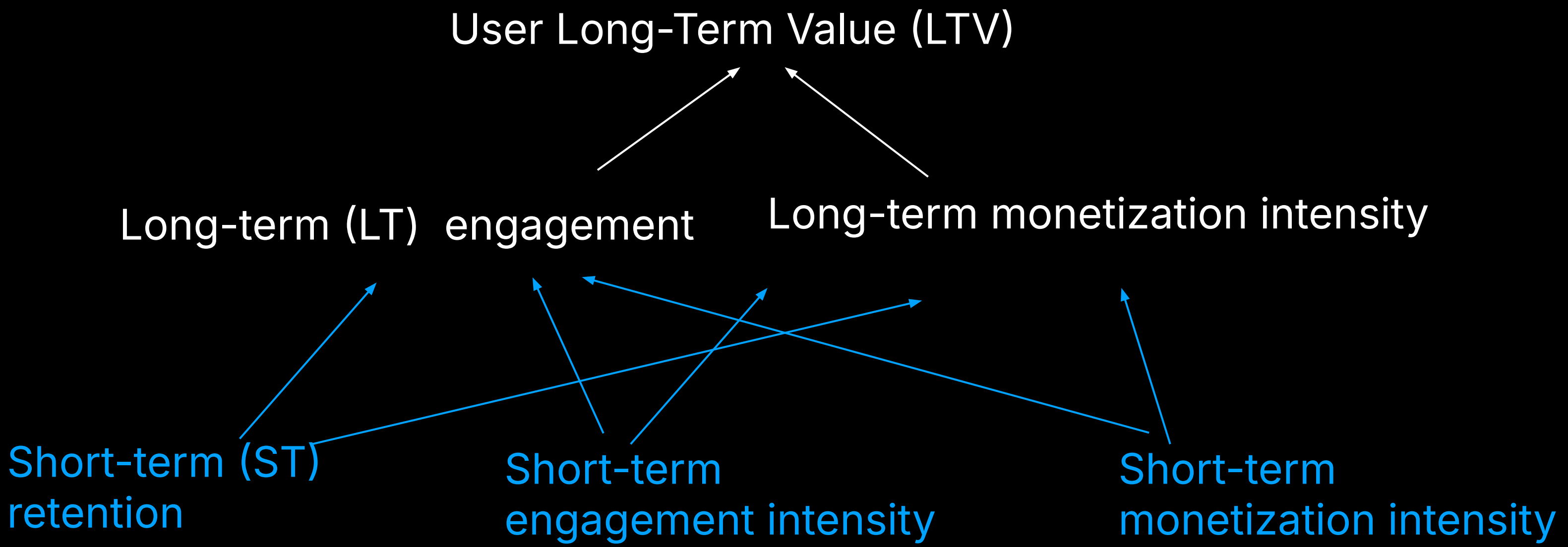
North Star Decomposition

User Long-Term Value (LTV)

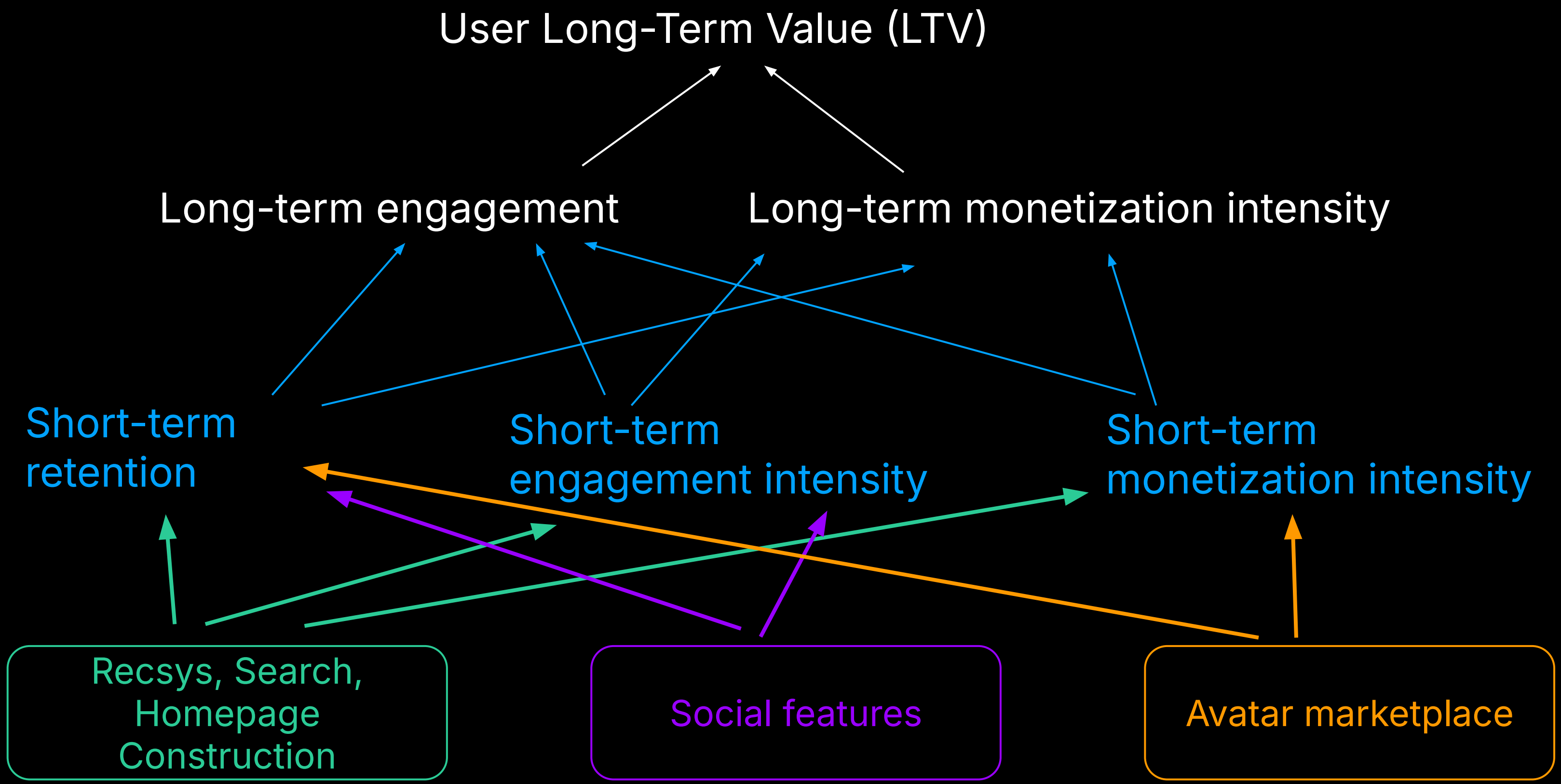
Long-term engagement

Long-term monetization intensity

Produce levers — timely short-term company-level metrics

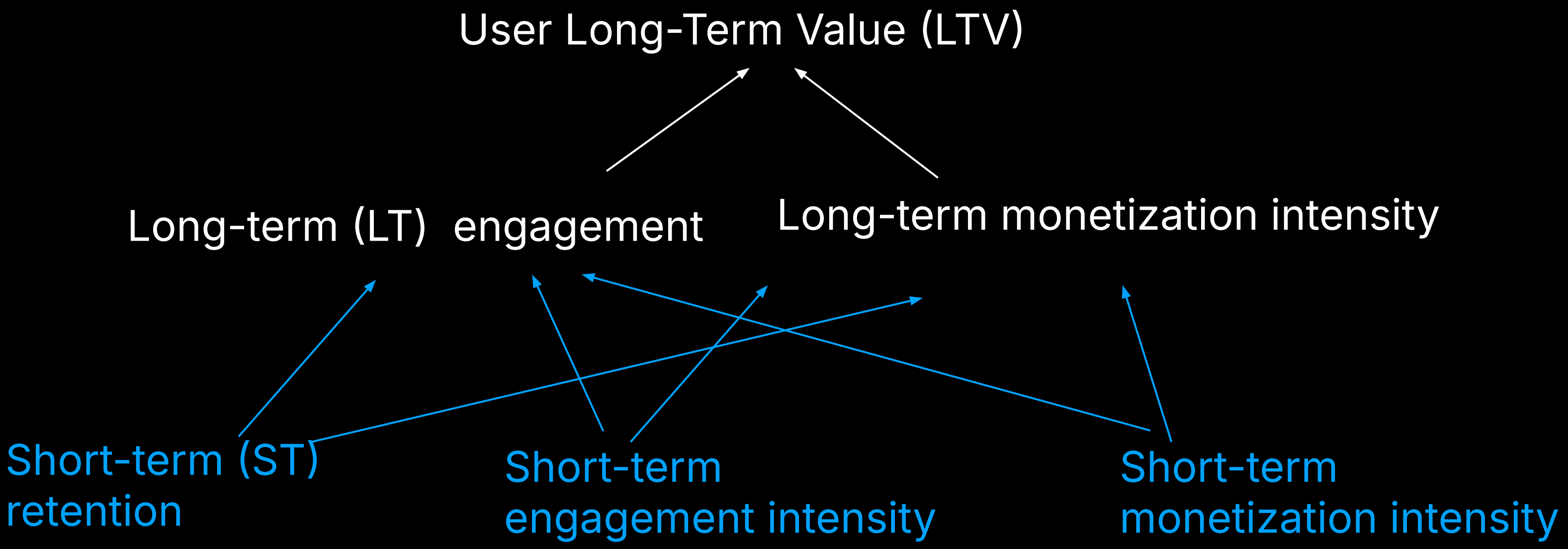


Product levers — can move multiple metrics



Product Decisions for Long-term Optimization

Produce levers — timely short-term company-level metrics



- **Experiments only measure the short-term metrics.**
- **How to make tradeoff decisions that optimize for long-term?**

Product decisions — optimize towards long-term

Ensemble approach with two models:

1. Joint relationship between movements of ST metrics and LTV.

2. Blackbox prediction of user's LTV trajectory.

→ Do they agree? What is highest confidence level?

ROBLOX

Collaborators: Wenjing Zheng, Wally Toh, Joao Dimas, Phil Hebda, Zhenyu Zhao, Yidi Wang

Produce decisions — optimize towards long-term

Ensemble Approach:

1. **Joint relationship between movements ST metrics and LTV.**

$$\textit{LT engagement} = \alpha_1\ \textit{ST retention} + \alpha_2\ \textit{ST engagement int.} + \alpha_3\ \textit{ST monetization int.} + f(C)$$
$$\textit{LT monetization} = \beta_1\ \textit{ST retention} + \beta_2\ \textit{ST engagement int.} + \beta_3\ \textit{ST monetization int.} + f(C)$$
$$\rightarrow \textit{LTV} = \gamma_1\ \textit{ST retention} + \gamma_2\ \textit{ST engagement int.} + \gamma_3\ \textit{ST monetization int.} + f(C)$$

Training

- **multi-exposure DML with user-level observational data**
  - ST metrics are exposures, LT metrics are outcomes.
  - Lots of user-level baseline confounders in DML nuisance models (GBM)
- **Simple partial linear models give interpretable exchange rates.**
- **Can be extended to group-level ST metrics**

ROBLOX

Collaborators: Wenjing Zheng, Wally Toh, Joao Dimas, Phil Hebda, Zhenyu Zhao, Yidi Wang

Produce decisions — optimize towards long-term

Ensemble Approach:

1. Joint relationship between ST metrics and LTV movements.

LT engagement = α_1 *ST retention* + α_2 *ST engagement int.* + α_3 *ST monetization int.*
LT monetization = β_1 *ST retention* + β_2 *ST engagement int.* + β_3 *ST monetization int.*
→ *LTV* = γ_1 *ST retention* + γ_2 *ST engagement int.* + γ_3 *ST monetization int.*

Application in experiment

- ST core metrics changes between treatment and control groups
- Plug into the formulas to get LT impact estimates and SEs.

PSA: check out Wally Toh's talk in the E2E Customer Journey workshop for more details

Collaborators: Wenjing Zheng, Wally Toh, Joao Dimas, Phil Hebda, Zhenyu Zhao, Yidi Wang

Produce decisions — optimize towards long-term

Ensemble Approach:

2. Blackbox prediction of user's LTV trajectory.

- Lots of user-level features, focuses on prediction accuracy
- Produce LTV prediction for every user.
- Compare average predicted LTV between treatment and control groups

Collaborators: Wenjing Zheng, Wally Toh, Joao Dimas, Phil Hebda, Zhenyu Zhao, Yidi Wang

Produce decisions — optimize towards long-term

Validation:

- Long-running experiments
  → ground truth = measured actuals of LTV (or mid-term version)

- Validation question:
  Would we have made the same decision if we had measured the ground truth vs using the ensemble approach?

Launch (stat sig pos) /No launch decisions:
- Trade-off formula: precision 50%, recall 80%
- predicted LTV: precision 100%, recall 60%.
(caveat: number of experiments limited)

PSA: check out Wally Toh's talk in the E2E Customer Journey workshop for more details

ROBLOX

Collaborators: Wenjing Zheng, Wally Toh, Joao Dimas, Phil Hebda, Zhenyu Zhao, Yidi Wang

Produce decisions — optimize towards long-term

Ongoing work:

- **Validation:**
    - Continuously invest in golden data sets
    - Simulations from experiment data.

- **Improve models for predicting long-term treatment effects. eg**
    - Trend stabilization
    - Confounding in observational data
    - Continuing treatment effect

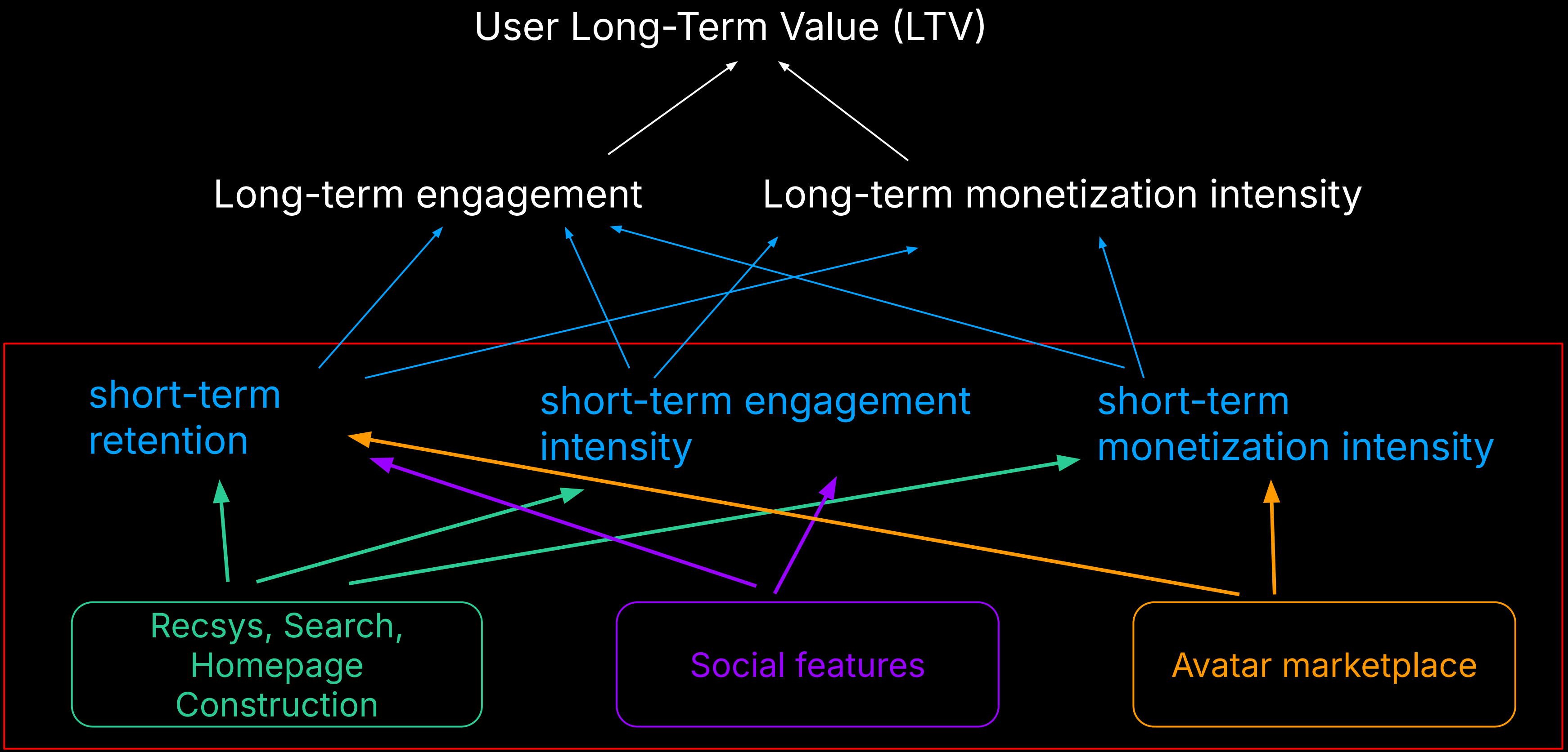
- **Variance reduction in core metrics in models and experiments**

ROBLOX

Collaborators: Wenjing Zheng, Wally Toh, Joao Dimas, Phil Hebda, Zhenyu Zhao, Yidi Wang

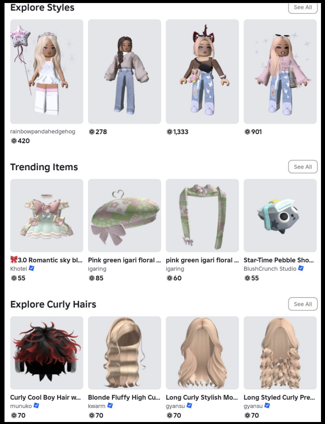
Informing Product area objective functions

Product levers — can move multiple metrics

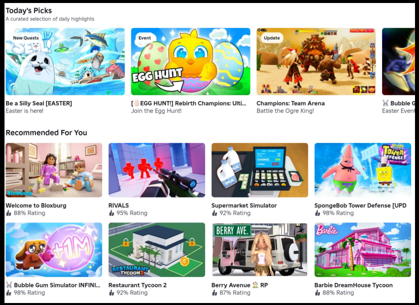


Product area objective functions

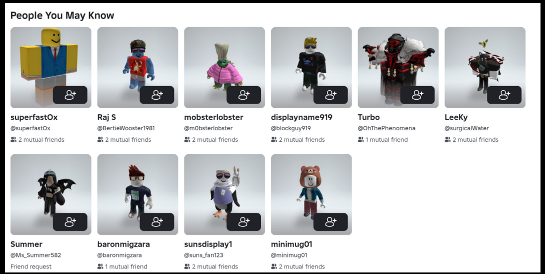
Product areas have different levers to nudge user action



Decorating avatars



Play different experiences

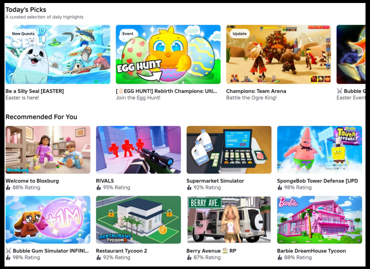


Making friends

Collaborators: Zhenyu Zhao, Yidi Wang

Product area objective functions

Product levers



Recsys
Homepage construction
Server-matching

ST area metrics


play


spend

Technical performance

ST core metrics

Retention
Engagement
Monetizations

→ LTV

Numerous, short-term, sensitive.

How to prioritize for long-term goals (less sensitive)

ROBLOX

Collaborators: Zhenyu Zhao, Yidi Wang

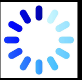
Product area objective functions - guiding for long-term goals

1. Which short-term metrics is this product area most leveraged to move?

    Example: Recsys ⟶ ST retention

2. Large set of ST area-specific metrics. Examples

 play  spend 

Technical performance 

3. Which ST metrics are most incremental to ST retention?

Collaborators: Zhenyu Zhao, Yidi Wang

Product area objective functions

3. Which ST area-specific metrics are most incremental to ST retention? (even when controlling for the effect of others?)

i. Candidates: Start with a large set to cover diverse dimensions.
ii. Narrow down: apply multi-exposure DML with user-level observational data.

Hypothetical example. Recsys:

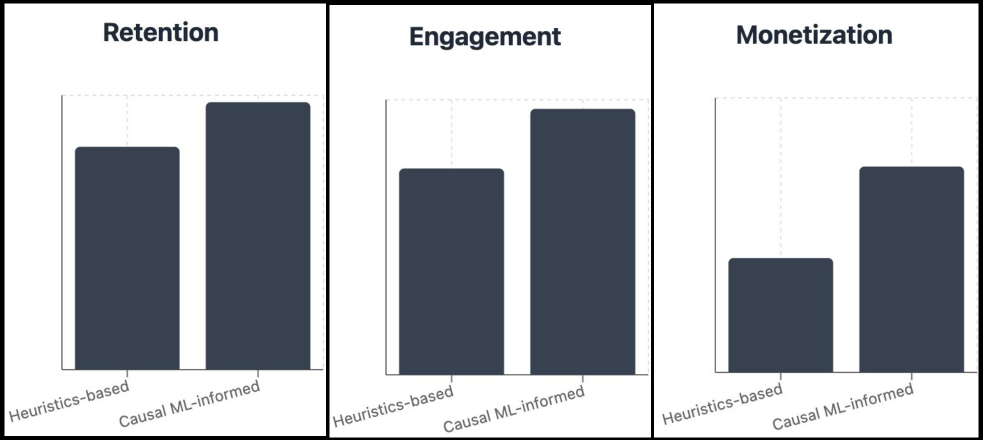
$$\text{ST retention} = \eta_1 \textit{ST playtime} + \eta_2 \textit{ ST in-experience spend} + \eta_3 \textit{ST playing with friends} + \eta_4 \textit{ST technical QoE} + \eta_5 \textit{ ST new games played} + f(C)$$

playtime, in-experience spend, and playing with friends are found most incremental. How to combine these to guide Recsys objective function?

Product area objective functions

4. Run experiments to confirm.

- We saw stat sig wins on short-term core metrics



- BUT if experiment had tradeoffs in core metrics, we use the previous framework to decide launch.

 Collaborators: Zhenyu Zhao, Yidi Wang

Actual ST metrics for Recsys obj fctn are disclosed to our creators [here](#)!

Lessons Learned So Far

Metric Frameworks

1. Unified Framework ~ accelerates shipping decisions while keeping the long-view front and center.

2. Objectives functions derived from the framework helps connect short-term sensitive product levers to long-term goals

3. Multi-altitude view:
    a. Top: LT goals
    b. Middle: ST to MT core metrics for shared guardrails and launch rubric at scale (built into our experimentation platform)
    c. Bottom: Group-owned ST area-specific metrics to operationalize.

4. Cross-team collaborative process

Building Trust in Causal ML (observational causal inference)

1. Technical Trustworthiness (can we believe the estimate?)

    a. Transparency: performance, methodology, assumptions. Reproducibility.
    b. Experiments and simulations to validate. Maintain a golden data set.
    c. Standardize metrics and methodology through tooling and infra
    d. Iterate on methods improvements

2. Organizational Trust and Alignment (will people use it?)

    a. Stakeholder education
    b. Interpretability (balance w/ model performance)
    c. Inclusive and iterative design
    d. Decision guidelines with escalation process

Thank you!