AI in Social Discovery

Blending Research and Production

Hyperconnect

Sungjoo Ha

September 1st, 2023

Today's Story

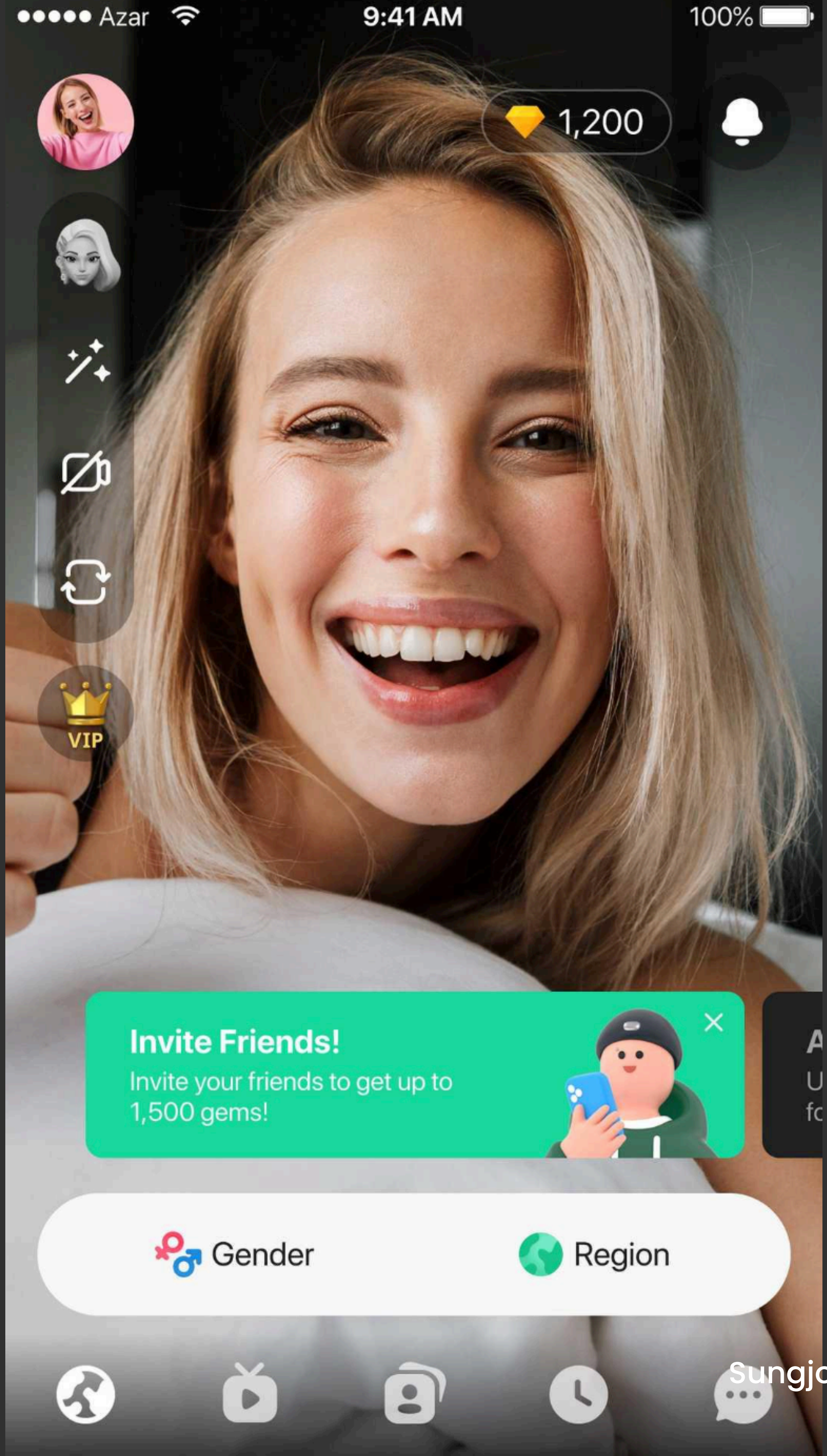
- Combining research and production
- How Hyperconnect AI navigated in this environment

Hyperconnect

- 2014 Azar
- 2019 Hakuna
- 2021 Match Group

azar

- Video messenger & social discovery service
- 115B matches
- 500M downloads
- 99% global user reach

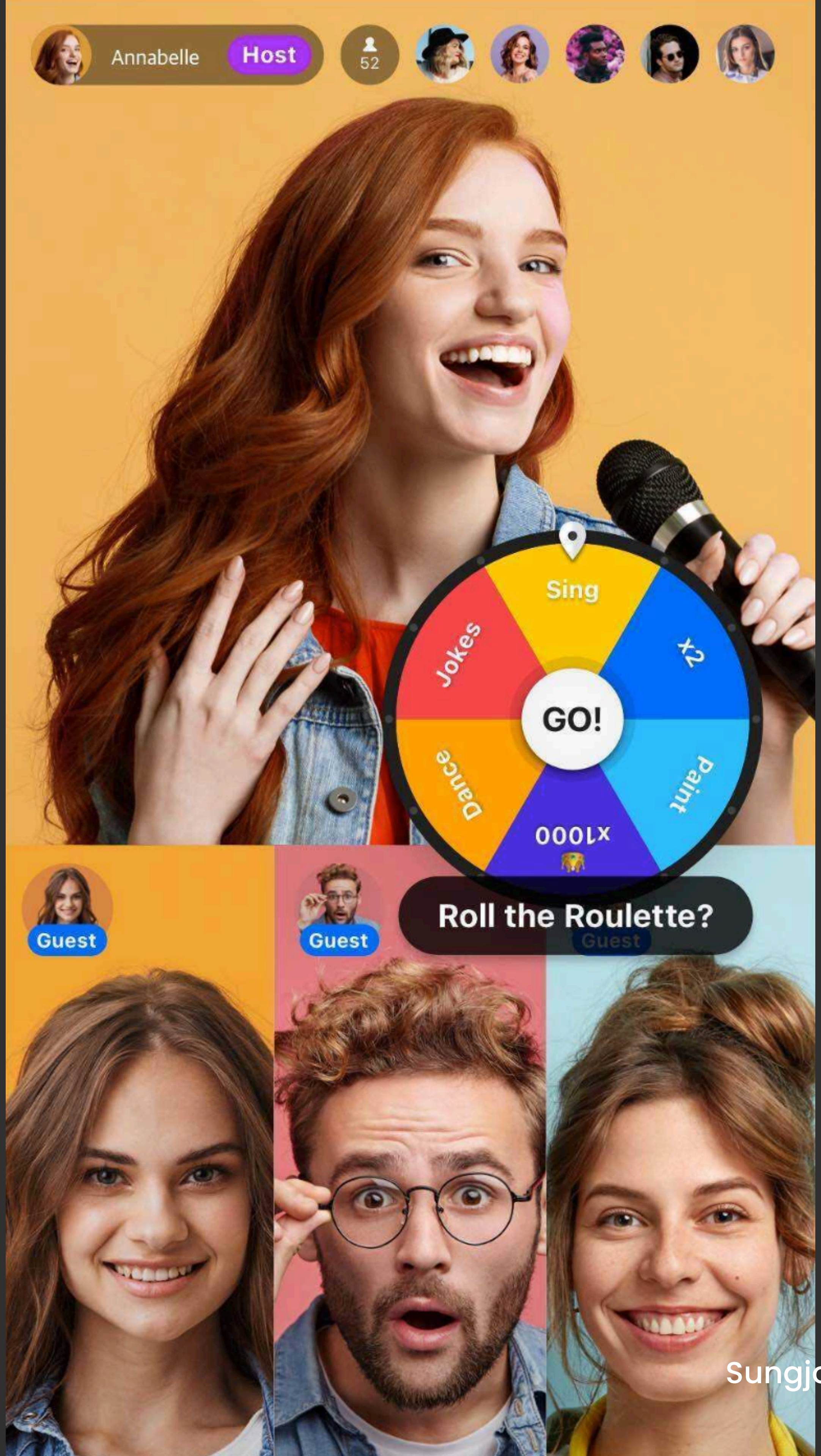


Invite Friends!

Invite your friends to get up to 1,500 gems!

Hakuna

- Social live streaming service
- Real-time multi-guest interaction via WebRTC



Spread the Joy of Live Conversation and Content Worldwide

- Hyperconnect's focus: social discovery
- Creating value through connecting people
  - Real-time communication and content
  - Utilizing AI

Hyperconnect AI Lab

- Handling all things ML/AI
  - Project selection
  - Project development
  - Data gathering
  - Model development
  - Experimentation
  - Paper writing
  - Data QA
  - Deployment
  - ...

Papers

- TiDAL: Learning Training Dynamics for Active Learning, ICCV 2023
- Reliable Decision from Multiple Subtasks through Threshold Optimization:Content Moderation in the Wild, WSDM 2023
- Measuring and Improving Semantic Diversity of Dialogue Generation, EMNLP 2022
- Learning with Noisy Labels by Efficient Transition Matrix Estimation to Combat Label Miscorrection, ECCV 2022
- Meet Your Favorite Character: Open-domain Chatbot Mimicking Fictional Characters with only a Few Utterances, NAACL 2022
- Understanding and Improving the Exemplar-based Generation for Open-domain Conversation, ACL 2022 Workshop
- Temporal Knowledge Distillation for On-device Audio Classification, ICASSP 2022
- Embedding Normalization: Significance Preserving Feature Normalization for Click-Through Rate Prediction, ICDM 2021 Workshop, Best Paper
- Efficient Click-Through Rate Prediction for Developing Countries via Tabular Learning, ICLR 2021 Workshop
- Distilling the Knowledge of Large-scale Generative Models into Retrieval Models for Efficient Open-domain Conversation, EMNLP 2021
- Disentangling Label Distribution for Long-tailed Visual Recognition, CVPR 2021
- Attentron: Few-shot Text-to-Speech Exploiting Attention-based Variable Length Embedding, INTERSPEECH 2020
- MarioNETte: Few-shot Face Reenactment Preserving Identity of Unseen Targets, AAAI 2020
- Temporal Convolution for Real-time Keyword Spotting on Mobile Devices, INTERSPEECH 2019

Research in a Company

- Industry research vs. academic research
- Defining research
  - Writing papers? Creating state-of-the-art models?
- Understanding production
  - Service with users?

Competition is for Losers[1]

To create a valuable company you have to basically both create something of value and capture some fraction of the value of what you've created.

You're the smartest physicist of the twentieth century, you come up with special relativity, you come up with general relativity, you don't get to be a billionaire, you don't even get to be a millionaire. It just somehow doesn't work that way.

[1] https://startupclass.samaltman.com/courses/lec05/

Value Creation & Value Capture

- Research: value creation
- Production: value capture
- Ultimately, all activities should contribute to company value
- Research labs in a company
    - Value creation alone is often insufficient
    - Aim to create value that is easily captured

Revisiting Social Discovery

- Creating value by connecting people
  - Obvious approach: recommendation via ML
  - Let's use ML to create better matches

Azar 1:1 Match

- Monetization through filters and pay-per-match
- Synchronous recommendation
    - Fully real-time -- supply & demand
    - Challenging to assume IID
        - Changes to the match algorithm inevitably affect others
        - Difficult to conduct A/B tests

Problem Definition

- What do we want to solve?
  - Use ML to provide users with better matches
- What defines a better match?
  - Unclear
  - Gauge via user feedback?
  - Maybe revenue is a signal that the users are having good experience?
  - Perhaps long matches?

Finding the Objective to Optimize

- Long-term user satisfaction
  - Don't even know how to measure exactly
- Cumulative revenue
  - However, delayed reward and not directly optimizable
- Chat duration maximization
  - Single/multiple matches, sessions?
  - Should we maximize the longest chat duration in a session?
  - Or the sum of chat durations within a session?

Pirate Metrics[2]

- Acquisition, activation, retention, revenue, referral
- Retention is king[3]
  - Whether a person returns to the service or not
  - Increasing retention is very difficult without improving the product
  - Also not directly optimizable

[2] https://500hats.typepad.com/500blogs/2007/06/internet-market.html, https://www.youtube.com/watch?v=irjgfW0BIrw

[3] https://andrewchen.com/retention-is-king/

Data Analysis

- Both exploratory & confirmatory data analysis are important
- Important to look at the data and get a feel for it
- So much cargo cult in data domain
- Know the correct tools, frame of mind, etc.

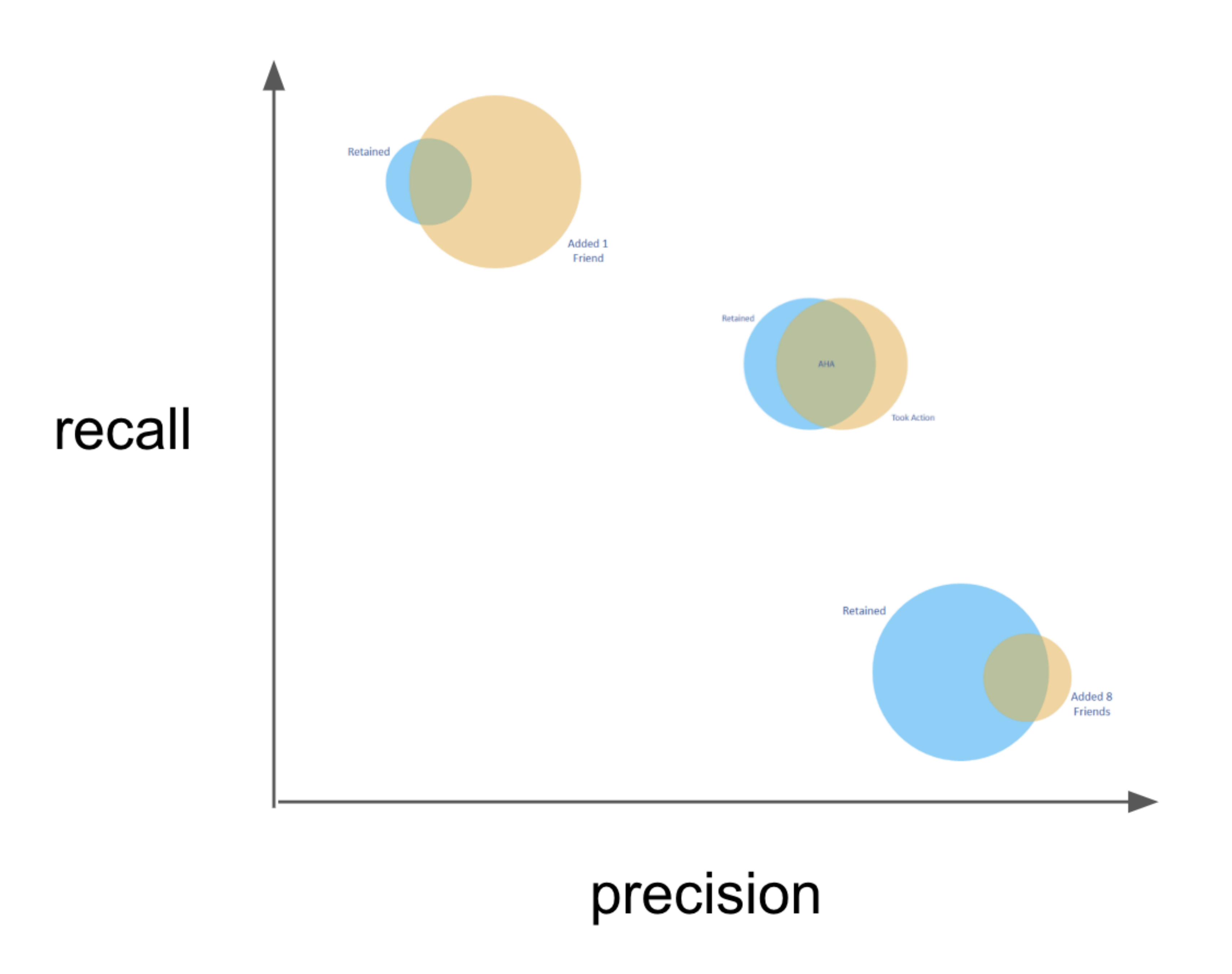
Aha Moment[4]

- Aha Moment: Perform Action Y, Z times within X days
  - The moment a user experiences the core value provided by the service
  - Users who experience the Aha Moment are retained, while those who don't are likely to churn
- Effective communication tool
  - Focus only on actions that lead to more Aha Moment experiences

[4] https://www.youtube.com/watch?v=raIUQP71SBU

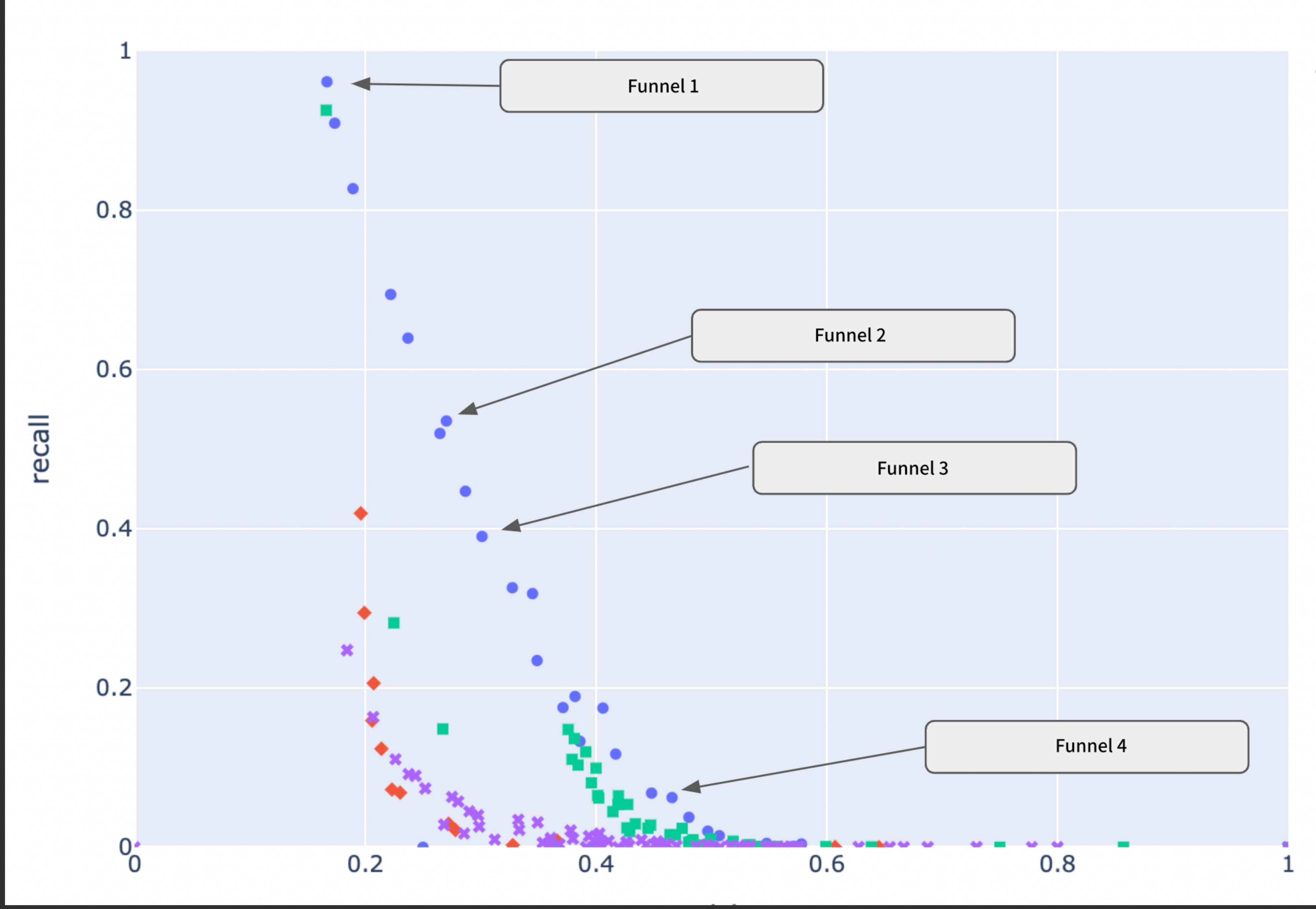
Aha Moment

- Perform Action Y, Z times within X days
  - Varying conditions X, Y, and Z result in different precision/recall values
- Identify all relevant actions
  - Develop complex conditions by logical operators
  - Calculate precision/recall for each condition



Funnel Analysis

- Consider this as a funnel
  - High recall & low precision → high precision & low recall
  - Provides insights on which funnel needs optimization

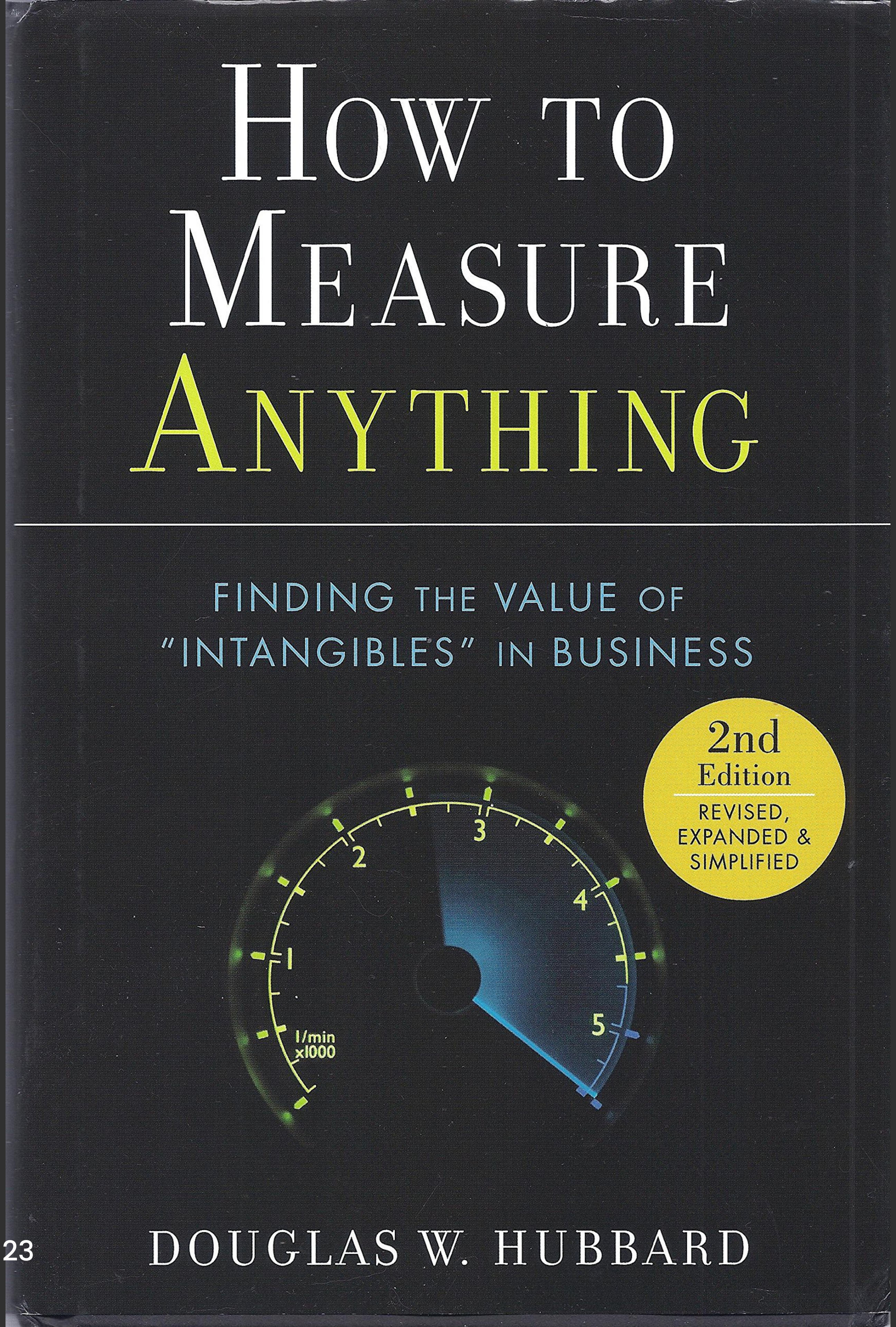


Problem Formulation

- Reduce your product problem into an AI problem
  - Your AI skills & product design skills count
  - Mathematical formulation, data strategy, AI/data flywheel
- Distinguish between exploration/exploitation projects
  - Most ML PoCs failed to deliver value to production
  - Know what works and doesn't work

Working with Legacy Systems

- Persuading stakeholders is an extremely important step
  - A working legacy system already exists
  - Why should it be replaced with an ML system?
- Engineering prowess alone is insufficient
  - Soft skills: communication, incentive design, sales


HOW TO MEASURE ANYTHING
FINDING THE VALUE OF "INTANGIBLES" IN BUSINESS
2nd Edition
REVISED, EXPANDED & SIMPLIFIED
1
2
3
4
5
l/min x1000
DOUGLAS W. HUBBARD

ROI Analysis

- Will the ML system result in better outcome?
  - Challenging to guarantee
  - Confidence increases with deeper understanding of the problem/system
  - Estimating the size of the upside is difficult
  - One heuristic: Is the problem sufficiently hard/complex?
- Adopt Bayesian decision theory framework when necessary

Working with Production Systems

- Think of the whole process as an anytime algorithm
- Create a well-designed interface & provide a baseline
  - Consider how the final model will integrate with the entire system and design an interface required for the final task
  - Begin by deploying the simplest model/heuristic
- Iteratively improve & continuously evaluate/monitor
  - Conduct small-scale experiments
  - Ensure your hypothesis aligns with reality

First Attempt

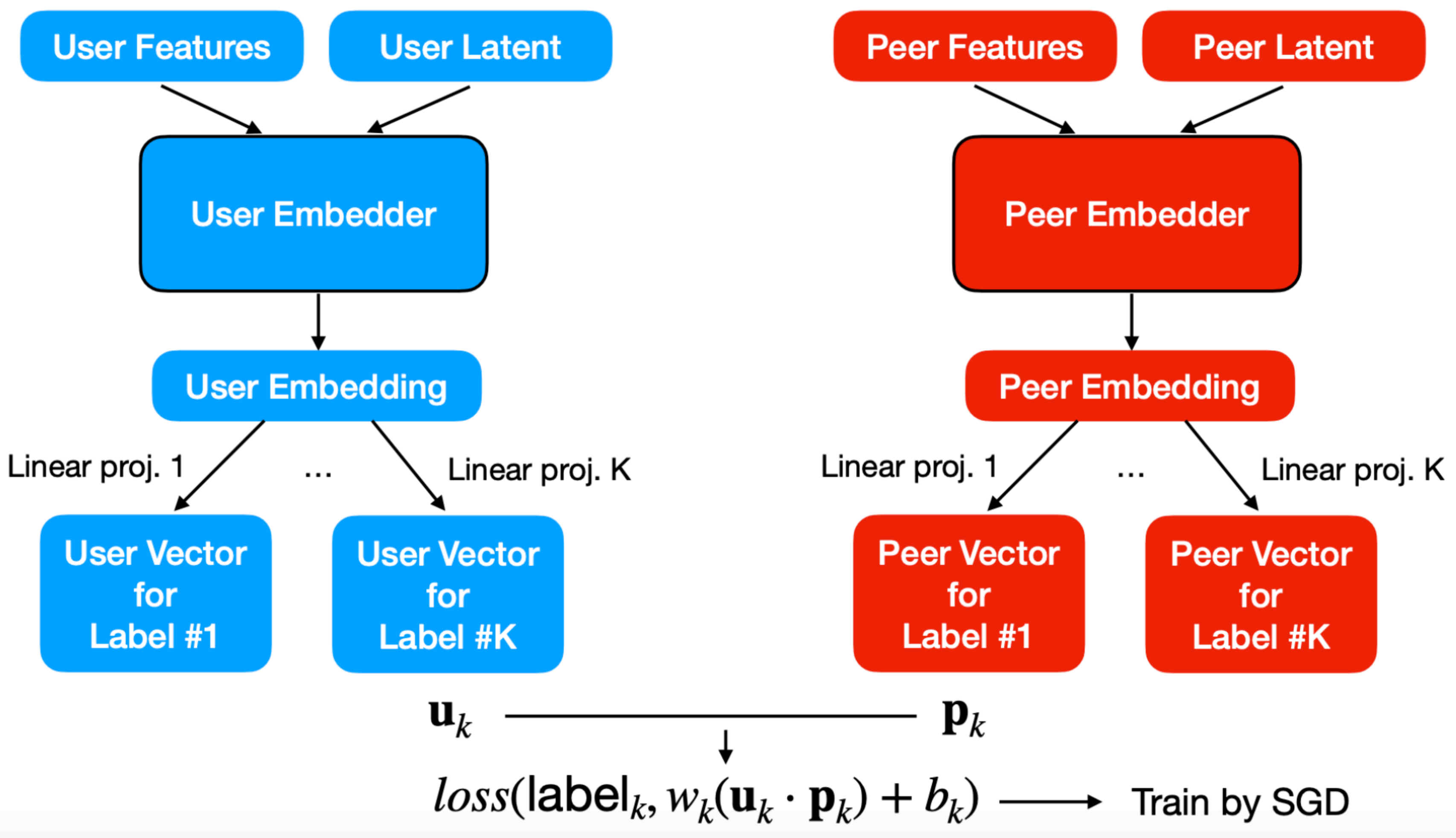
- Let's say we want to build a chat duration predictor
  - Pretend it generates more Aha Moments
  - Assumes IID, so can't address the supply-demand issue
  - However, tackling the most difficult problem from the start is not a good idea
- Even when addressing chat duration prediction
  - Consider how the model will be used and what the target metric should be
  - Example: AUROC & MSE
    - Low MSE indicates more accurate match duration predictions
    - High AUROC means better ordering

Problem Constraints

- Strict constraints
  - Low latency
    - A single tick is approximately half a second
    - ML can utilize around 100ms
  - Scalable
    - Need to reach more than 1500 TPS

Model Engineering

- $O(N^2)$ pairwise computation
  - Ensure the entire computation can be performed using a single dot product
- Cache the embedding layer, which can be computed asynchronously
- Knowing how each model differs in implementation level is essential



Parallelism

- Break down the problem into independent subproblems
- Enable parallel processing of user-peer pairs
- Simple in concept, difficult in practice
  - Distributed system causes all sorts of headache

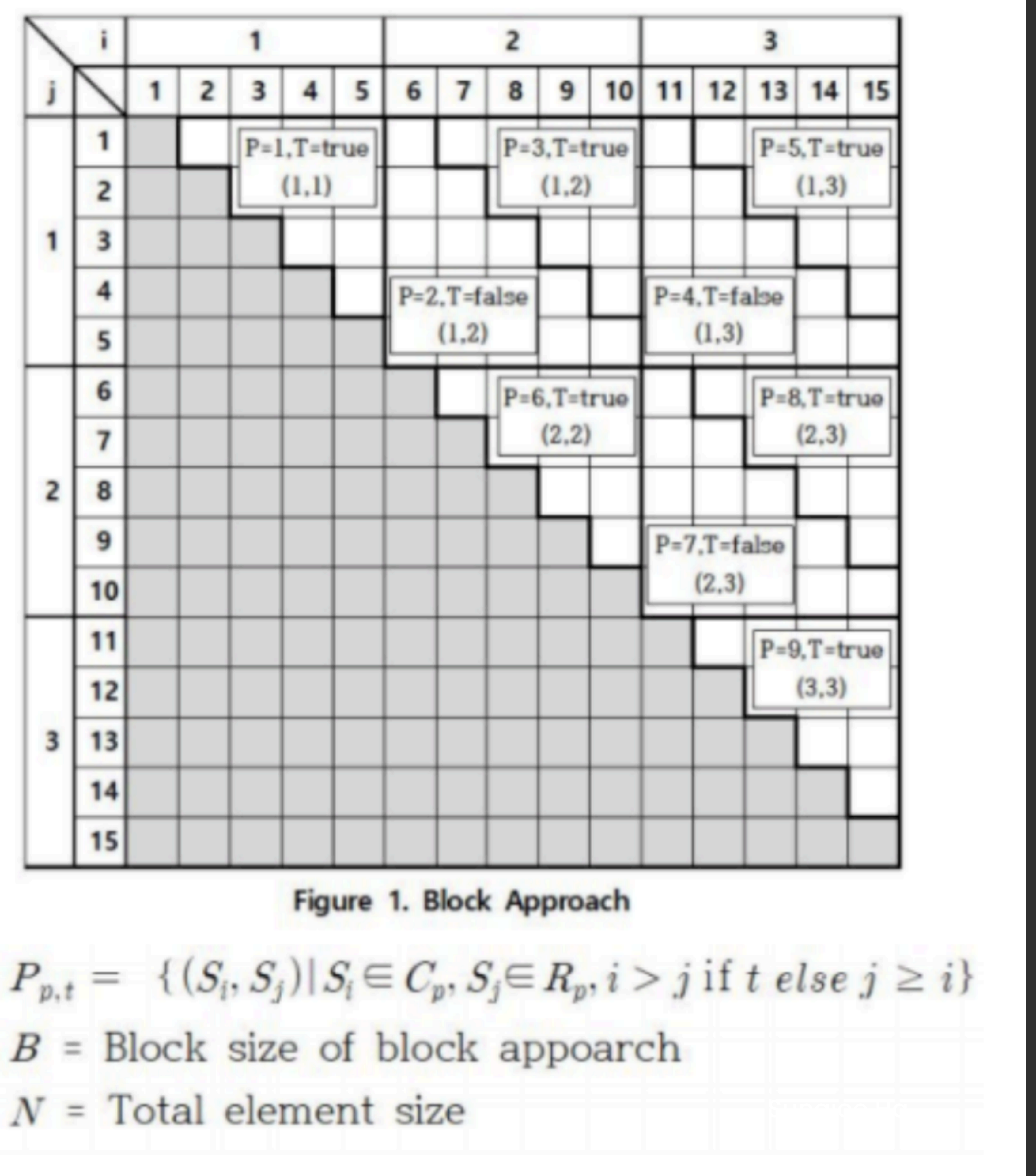


Figure 1. Block Approach

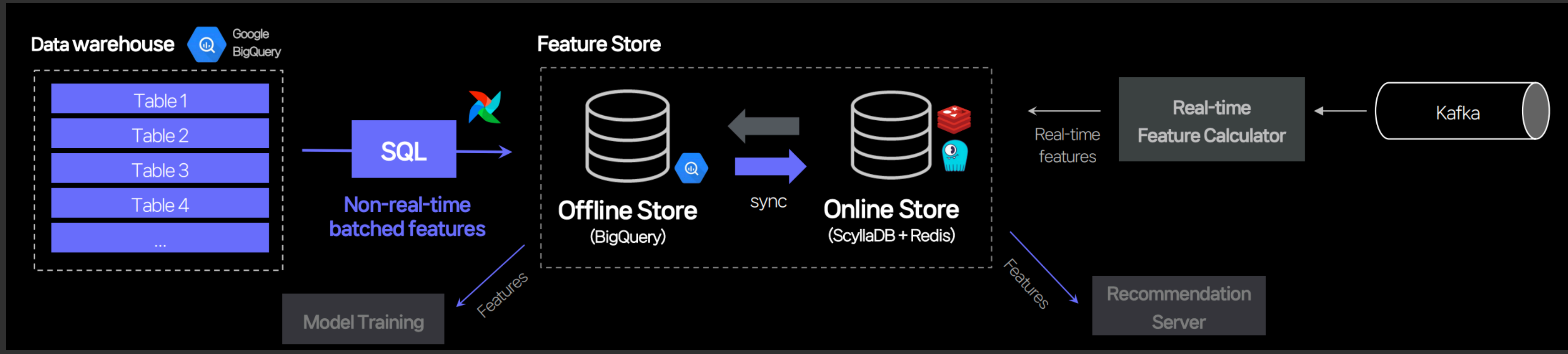
$$P_{p,t} = \{(S_i, S_j) | S_i \in C_p, S_j \in R_p, i > j \text{ if } t \text{ else } j \geq i\}$$

B = Block size of block appoarch

N = Total element size

Feature Store

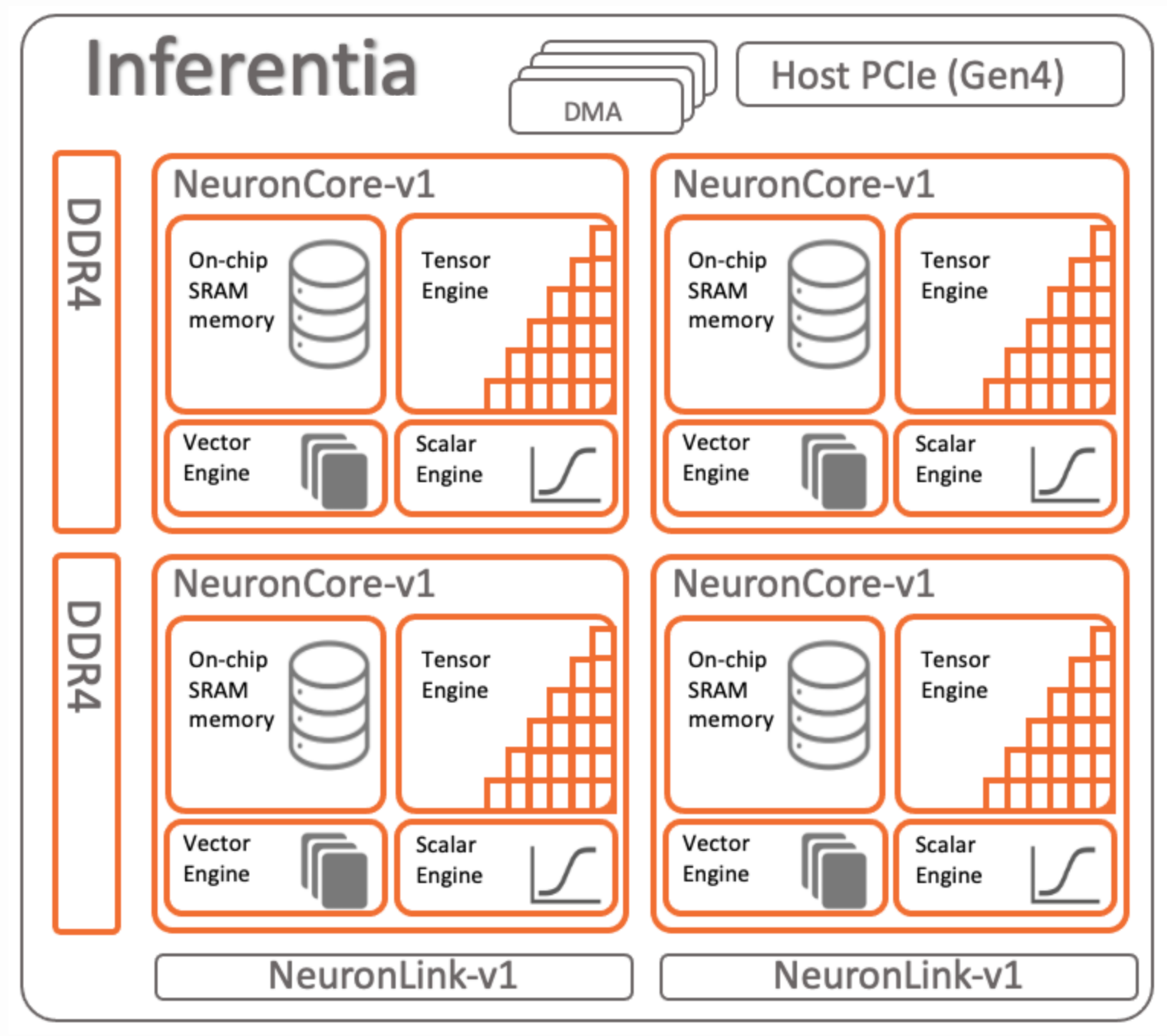
- Feature store[5] addresses the following issues:
  - Train/serving data discrepancies
  - High cost of adding features
  - Redundant components when deploying multiple ML applications
  - Difficulty sharing features when deploying multiple ML applications
  - Ensuring feature correctness



[5] https://deview.kr/2023/sessions/536

Inference Optimization

- AWS Inf1[6]
  - AI accelerator
- Improved TPS with consistent latency and lower cost
- Understanding how different parallelisms are exploited can help boost the performance
  - Dynamic batching, model pipelining

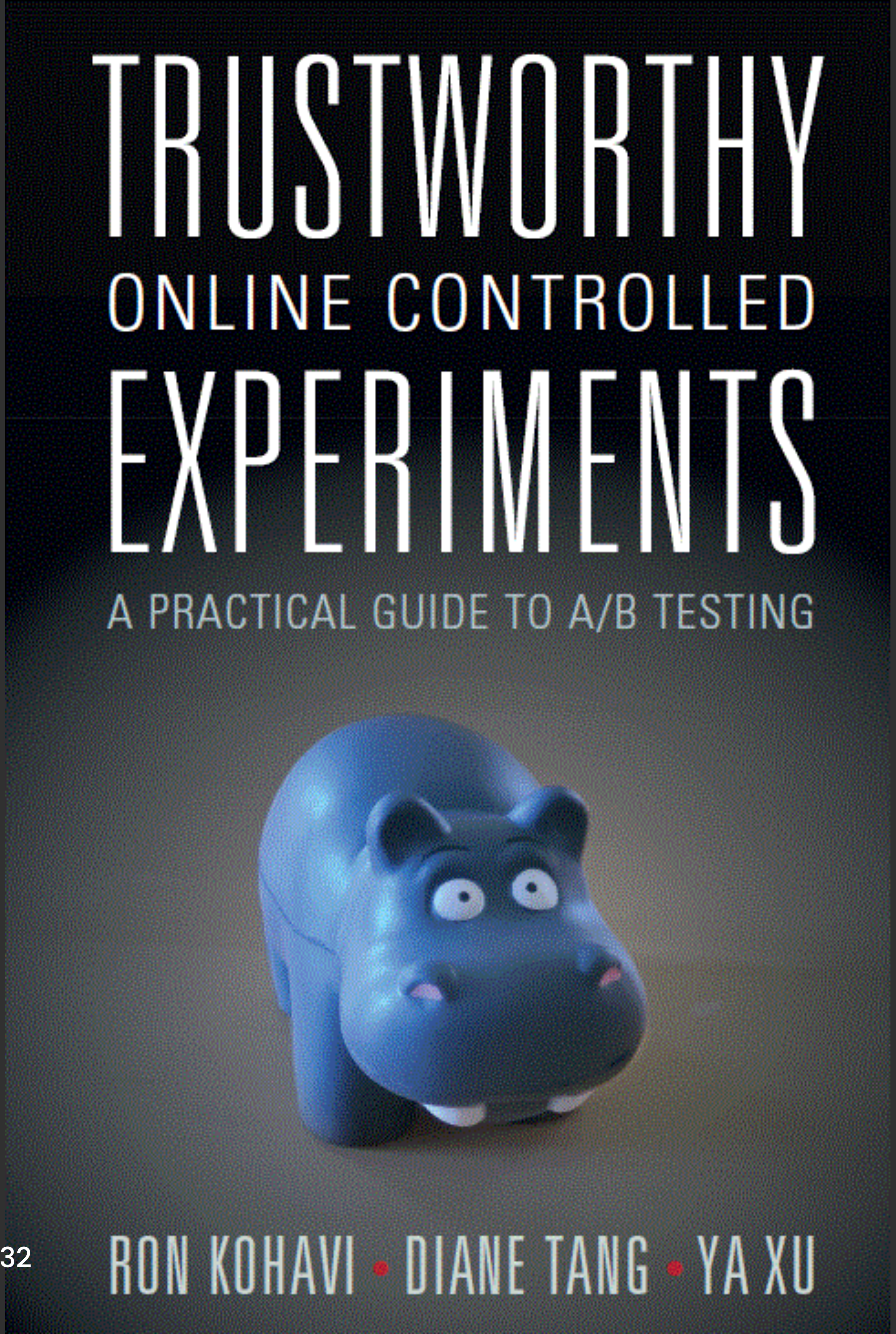


[6] https://hyperconnect.github.io/2022/12/13/infra-cost-optimization-with-aws-inferentia.html

Python Optimization[7]

- Optimize P99.9 latency
- Avoid using Python lists
  - Especially not Pandas
  - Use contiguous memory: array/numpy array
- Garbage collection optimization
  - Avoid stop-the-world
- Avoid context switching by optimizing the number of concurrent processes

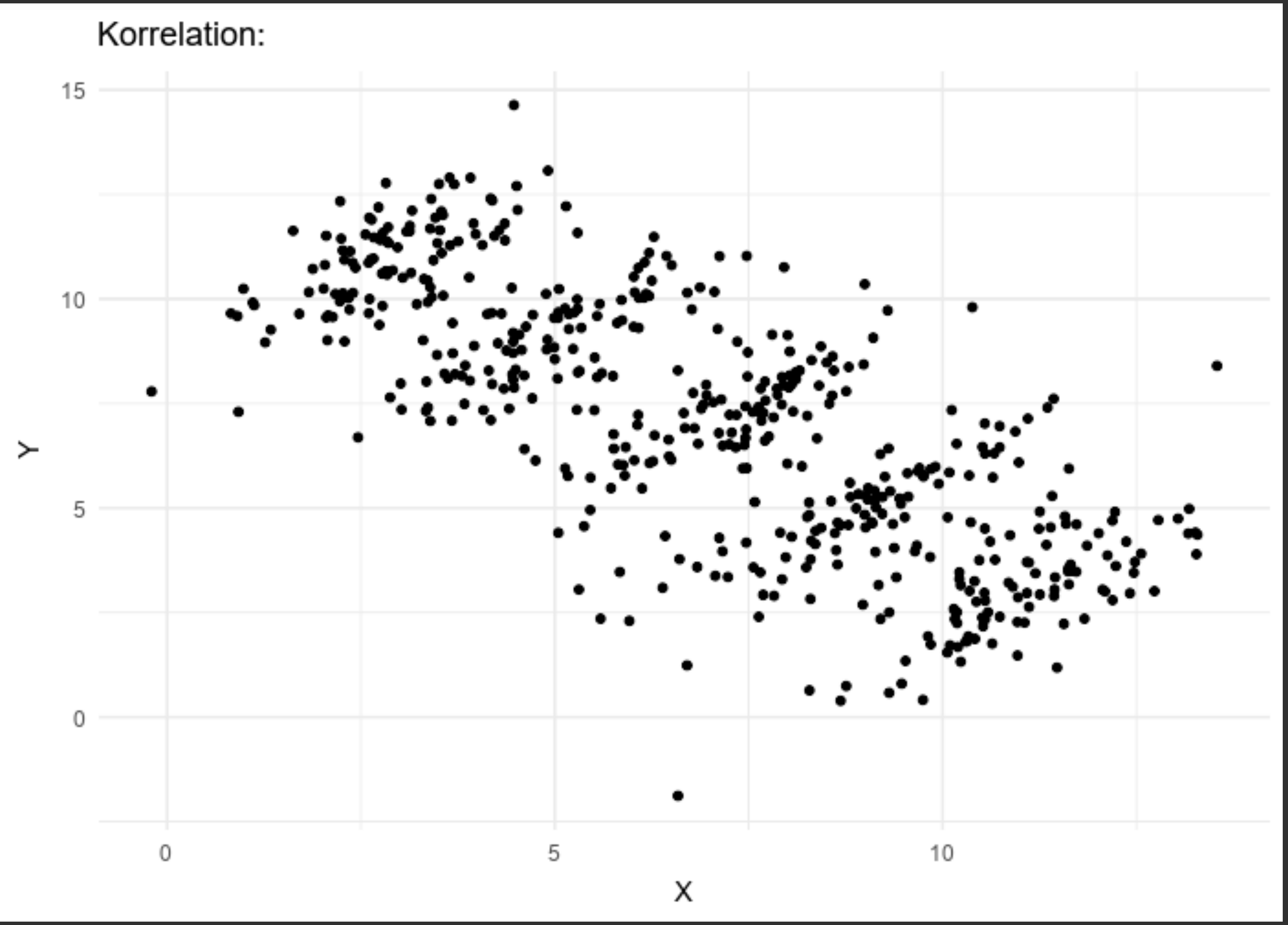
[7] https://hyperconnect.github.io/2023/05/30/Python-Performance-Tips.html


TRUSTWORTHY
ONLINE CONTROLLED
EXPERIMENTS
A PRACTICAL GUIDE TO A/B TESTING
RON KOHAVI • DIANE TANG • YA XU

Experiment Iteration

- Experiment a lot
- Conduct proper monitoring
- Perform A/B test[8] whenever possible
- Come up with concrete hypothesis if things go wrong for another analysis/experiment
- Get your hands dirty with data

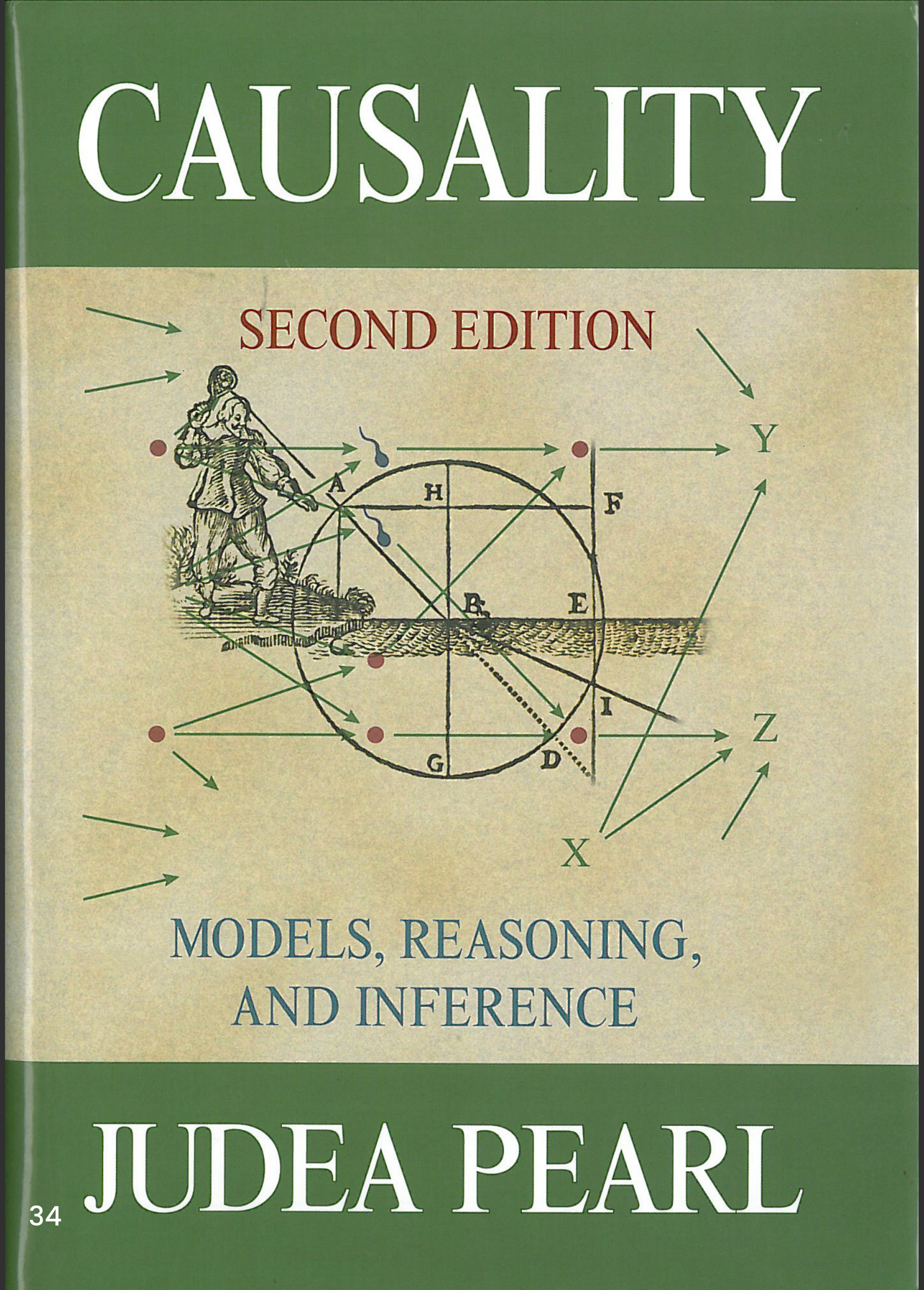
[8] https://exp-platform.com/talks/



Simpson's Paradox[9]

- Exactly the same data, different interpretation for different cases
- You encounter them once you start to replace your business logic with AI/ML models

[9] https://en.wikipedia.org/wiki/Simpson%27s_paradox



Causal Inference

- Gold standard to dealing with simpson's paradox
- Several methods available
  - Gold standard: randomized experiments
  - For observational data, use causal diagrams[10]

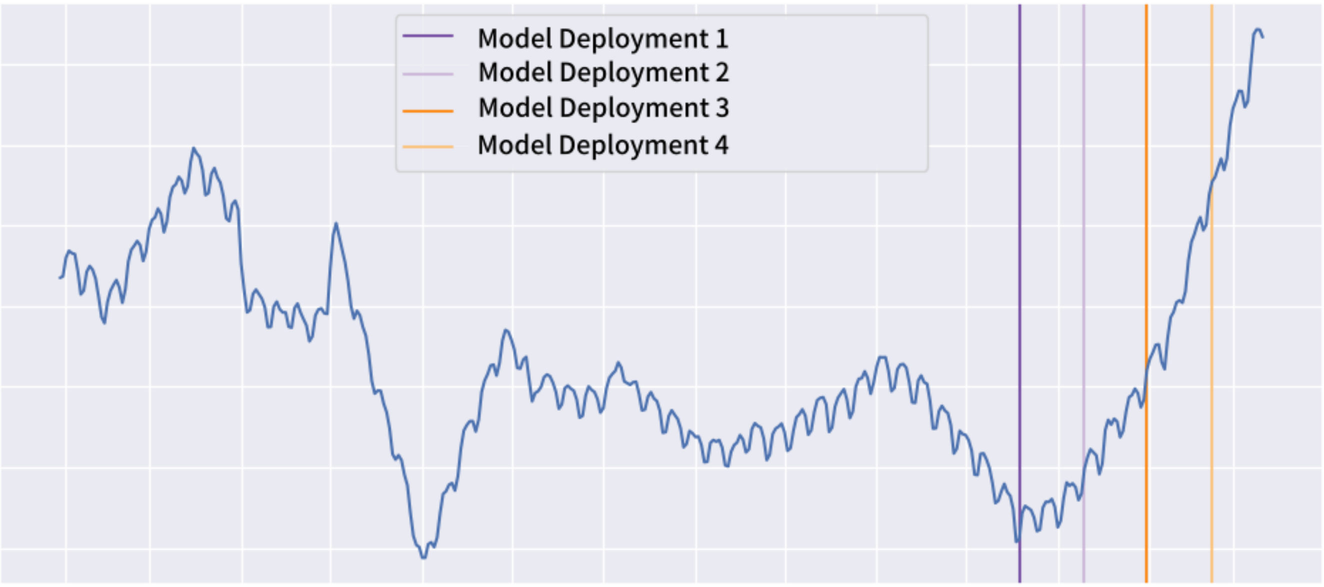
[10] https://pll.harvard.edu/course/causal-diagrams-draw-your-assumptions-your-conclusions

And Many More

- Better problem formulation
- Model improvements
- Overall MLOps ecosystem
- Stream processing
- Experiment design & management
- Monitoring and observability
- …

Result

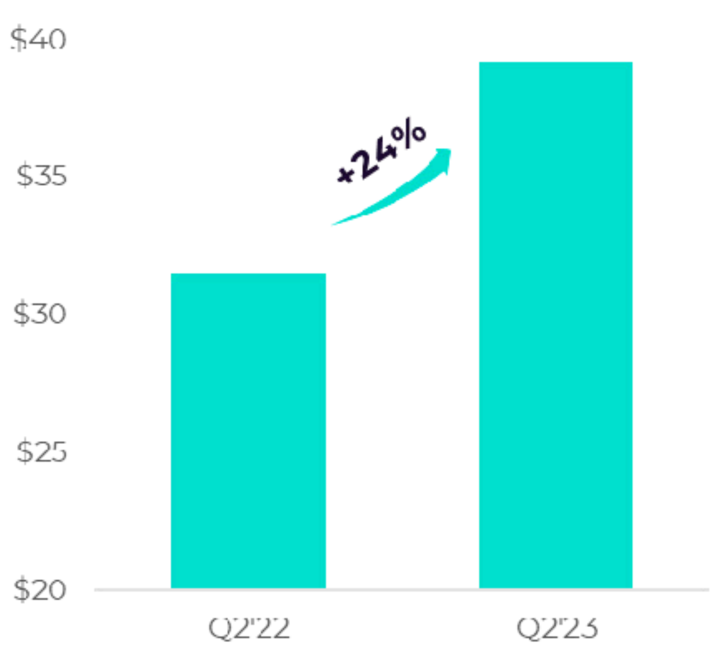
- Following numerous iterative improvements
- Deploying the recommendation model resulted in a dramatic increase in retention



MATCH GROUP ASIA

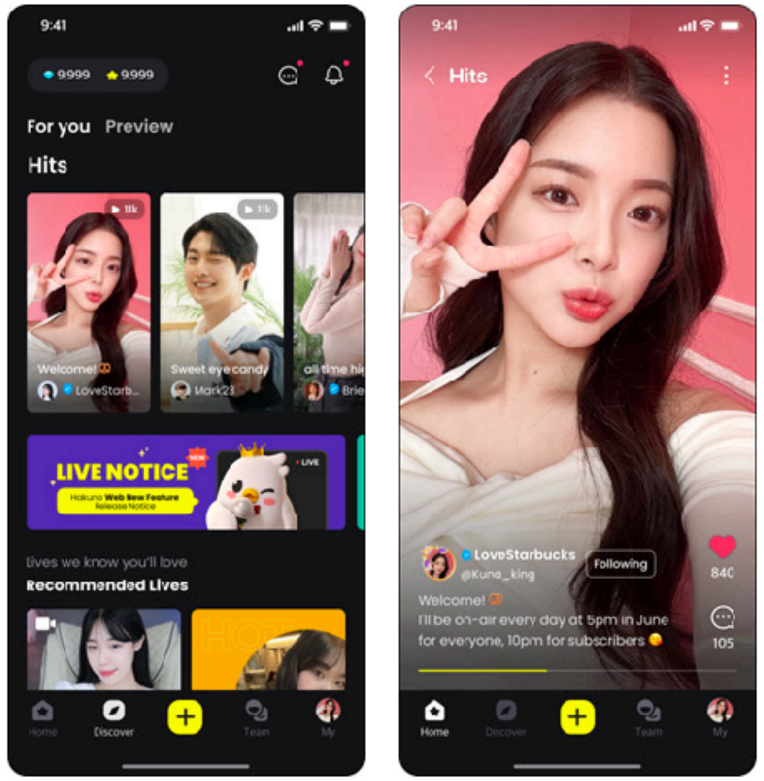
At Hyperconnect, Azar experienced very strong revenue growth in Q2, largely driven by improvements to its new AI-powered matching algorithm, which is increasing both user engagement and monetization, as well as by strong seasonal trends. Hakuna has reimagined its product to create a more personal connection between creators and their audience, targeted at key Asian markets. We're encouraged by the progress and direction at Hyperconnect, and the business's profitability trends.

Azar Direct Revenue ($M)



+30% Y/Y growth in KRW

Hakuna's Reimagined Experience



The acquisition of Hyperconnect brought Match Group a large team of talented AI engineers which we're leveraging to drive a number of important AI-related initiatives across the portfolio. Given Hyperconnect's strong reputation in Korea, we expect to be able to further grow this engineering team more quickly and effectively than we could in other markets.

How Did We Do This?

- Sane software engineering
- Sane machine learning & data science
- Other hard & soft skills
- Iterate & compound

Some Suggestions

- Striving for deep understanding
  - SWE, ML, DS, mental models
- Gaining deep dive experience is crucial
  - Problem finding, formulating, solving, and selling
  - Ability to navigate between abstraction layers
- Effective problem solving almost always involves other people
  - Alignment
  - Extreme ownership & high agency
  - Positive-sum game

Iterate & Compound

- There will be countless problems that you haven't thought of
- Solve/avoid one by one and make many small steps
- Compounding is a superpower

We Are Hiring!

- career-ai-recruit-2023.hpcnt.com