Bridging the Gap: AI Research and Real-World Deployment in AI Companies

Hyperconnect

Sungjoo Ha

March 29th, 2023

Today's Story

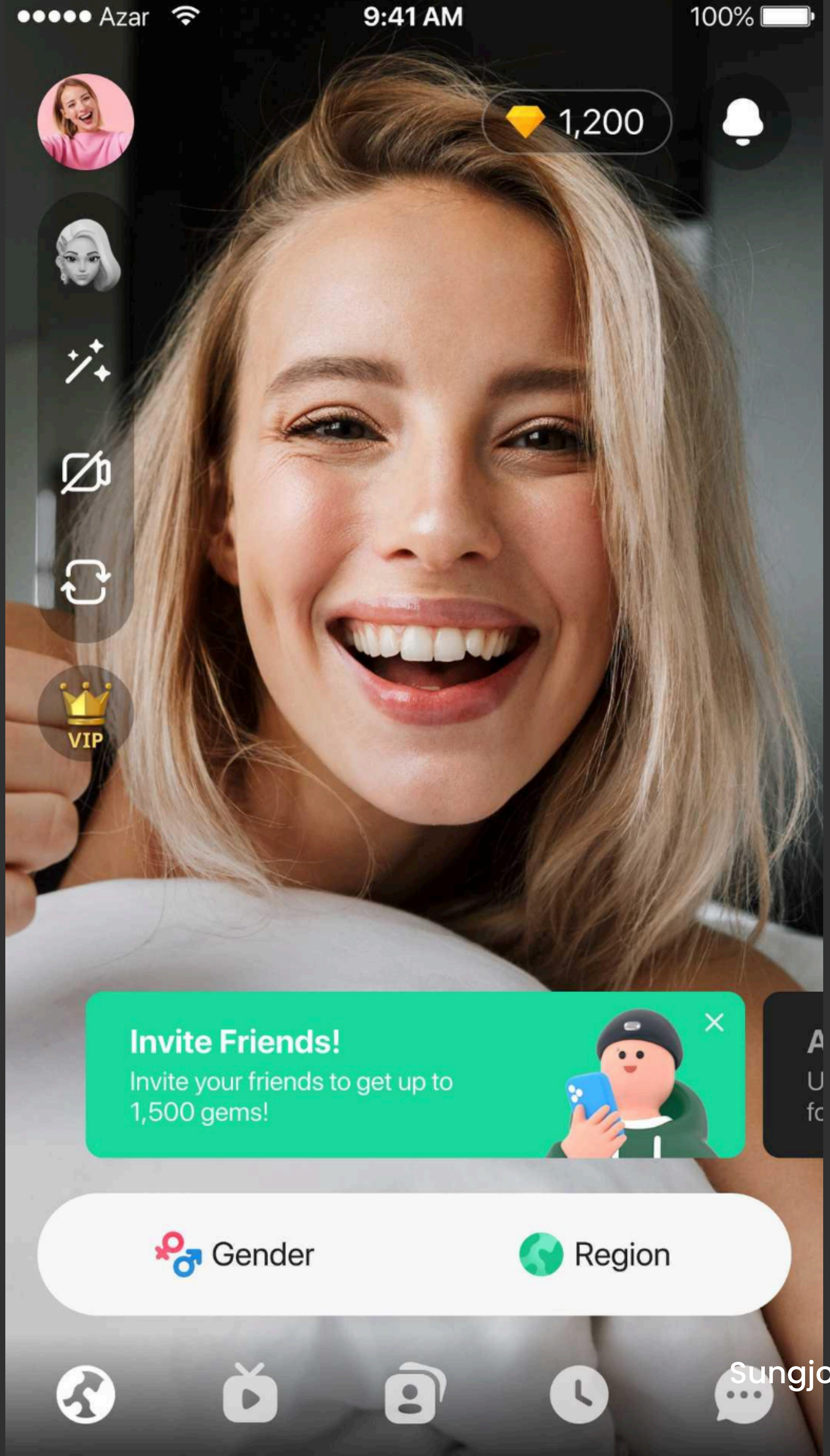
- Combining research and production
- The AI company & its implications
- Essential skills in this environment

Hyperconnect

- 2014 Azar
- 2019 Hakuna
- 2021 Match Group

azar

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

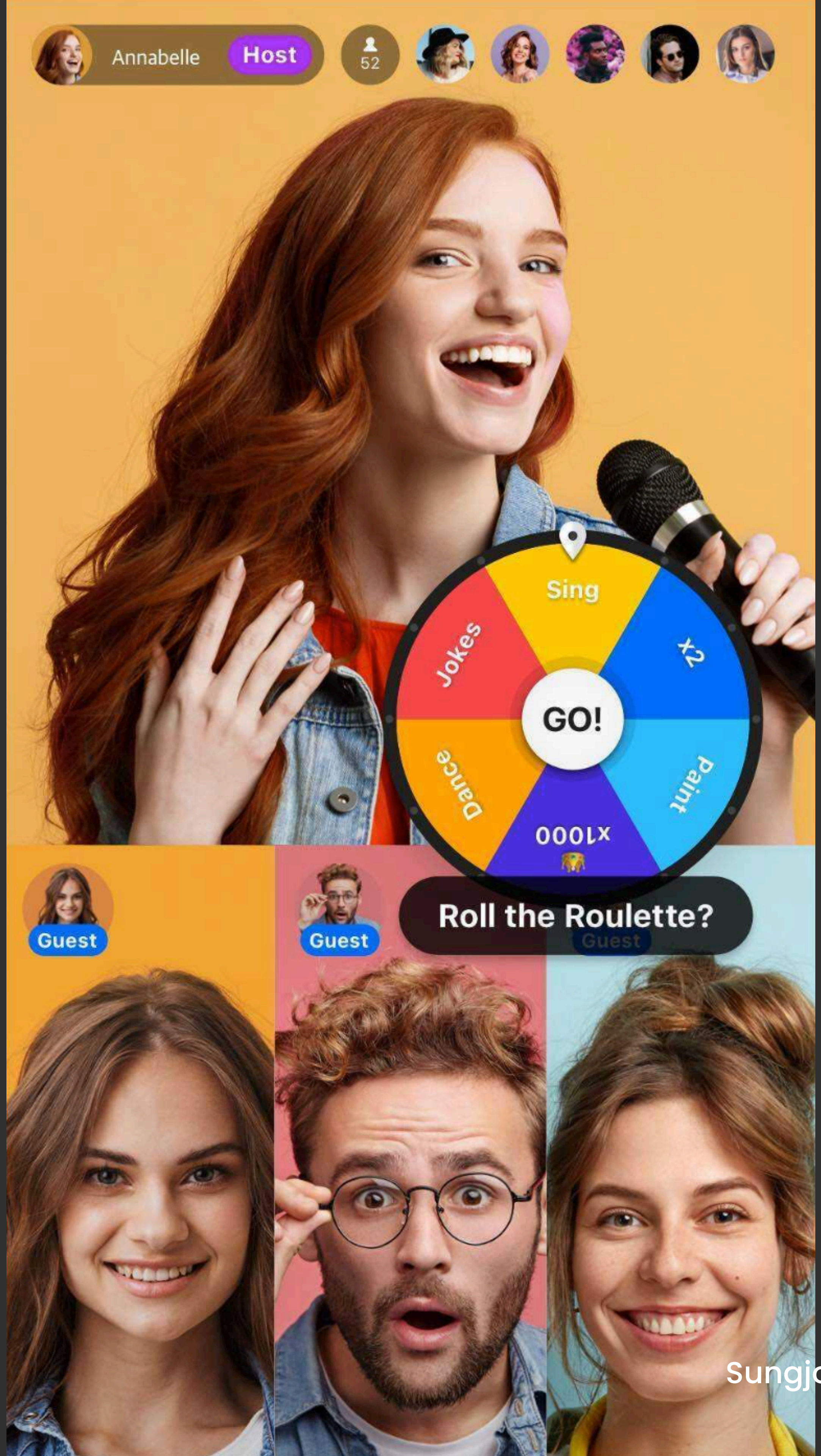


Invite Friends!

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
  - …

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

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.

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

AI Company

- Companies utilizing internet technology were called internet companies, and this trend continued into the mobile era
    - Amazon, Alphabet, Facebook, Alibaba, Tencent, etc.
- Defining an AI Company in the AI era

Shopping Mall + Web Page

≠ Internet Company

Jeff Bezos in 1997

In the book space, there are more than three million different books worldwide active and in print at any given time across all languages, so when you have that many items, you can literally build a store online that couldn't exist any other way. [1]

[1] https://youtu.be/rWRbTnE1PEM

Internet-Enabled Technology

- Technology of the Internet Era
  - Everyone had a web page during the internet era
  - Yet, companies fully utilizing internet-enabled technology were limited
  - Understanding users by collecting user behavior
    - Conducting A/B testing[2]
  - Transitioning from deploying once or twice per year
    - To continuous integration[3], continuous deployment, enabling daily deployment
  - Achieving an extremely short iteration cycle to explore product-market fit
    - An organizational structure that supports such exploration

[2] Google was already performing A/B test in 2000

[3] Martin Fowler wrote about CI in 2006

Learnings From The Past

- What can internet companies teach us about AI companies?
  - Businesses that cannot exist without AI
    - Achieving what was literally impossible before
  - Broadening the scope, companies utilizing AI-enabled technology

Any Company + AI/ML/DL

≠ AI Company

Aggregators

- Zero marginal cost
  - Selling additional copies of a digital item costs nothing
  - Distribution is free
  - Transactions are free
- Modern successful companies maximize this concept
  - Super-aggregators[4]
  - Merely existing on the internet is not a value proposition
  - Embrace what the internet offers and build a business that is impossible without the internet

[4] https://stratechery.com/concept/aggregation-theory/

Zero Marginal Content

- What businesses are impossible without AI?
- Some hints:
  - Zero marginal cost content creation
    - LLM, stable diffusion
  - Super-human decision-making
    - AlphaGo, AlphaFold

AI-Enabled Technology

- In the AI era, everyone will use AI models
- The crucial factor will be the ability to utilize the concepts, technologies, and culture stemming from this progress
    - Just as there are companies that use A/B testing and those that don't
    - Just as there are companies that use CI/CD and those that don't

Learned Business Logic

- Replace business logic with a model
  - Business logic: If A then do B
    - Most of what programmers create is business logic
- How does this differ? Wouldn't it be easier to write code rather than develop a complex model?
  - Models can outperform humans
    - If the condition A is too complex, humans are notoriously bad at it
    - Software 2.0

Software Rot

- Software, including business logic, rots
  - Environment changes
  - New features are deployed, product directions change, users change, …
  - How do we address this? Software engineers modify the code
    - If A then do B → If A then do C
  - However, if this was built using a model
    - The model processes the data and adapts itself
    - More data leads to better performance

Ideal

- All decision-making could be replaced by a model
  - Automate everything
- Particularly appealing if you can reduce the core business/product problem to an AI problem
  - Experience continuous improvement of your product

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
  - Perhaps long matches?
- What do we want to optimize?
  - Cumulative revenue
    - However, not directly optimizable
  - Chat duration maximization
    - Should we maximize the longest chat duration in a session?
    - Or the sum of chat durations within a session?
    - If we're paid per match, wouldn't this lead to lower overall revenue?

Objective

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

Exploratory Data Analysis

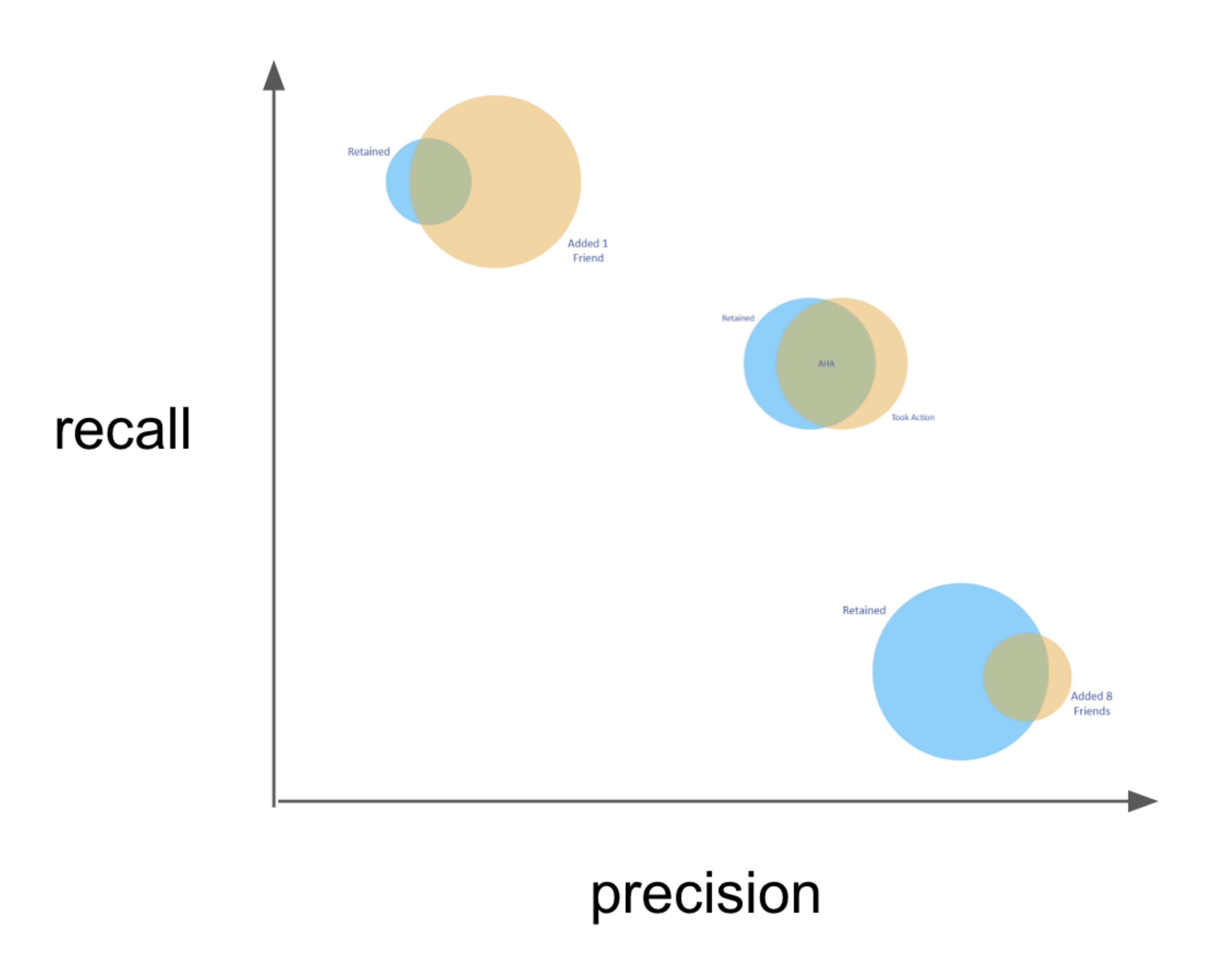
- Important to look at the data and get a feel for it
- So much cargo cult in data domain
- Know the correct tools, frame of mind, etc.

Aha Moment

- 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

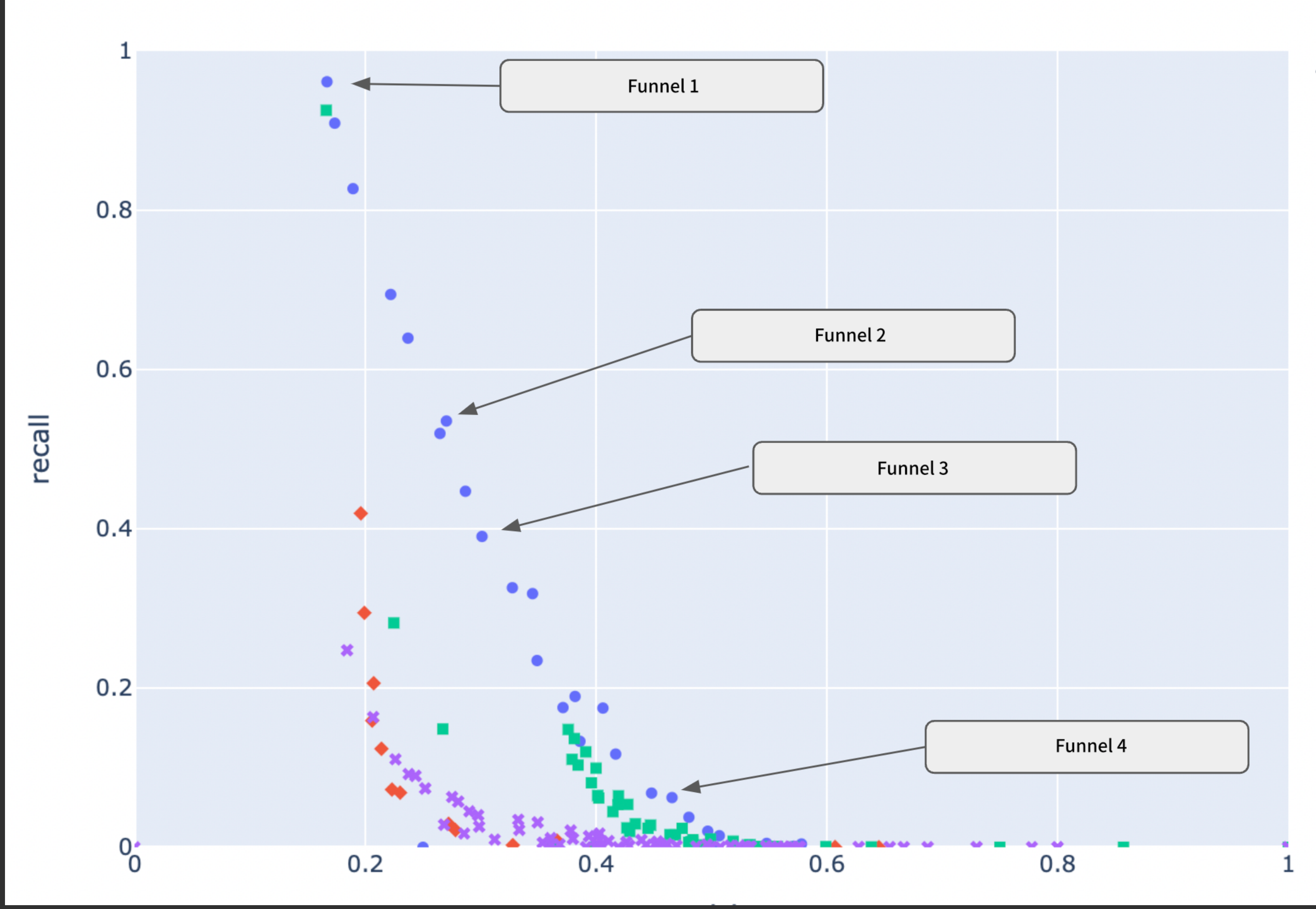
Aha Moment

- 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



Causal Inference

- Upon identifying a certain condition, conduct causal analysis
  - As correlation does not imply causation
- Several methods available
  - Gold standard: randomized experiments
  - For observational data, use causal diagrams

Legacy System

- 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
- Engineering considerations
    - Will the ML system result in better matches?
        - 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?

Working with Production System

- Interface
  - Consider how the final model will integrate with the entire system and design an interface required for the final task
- Baseline/heuristic
  - Begin by deploying the simplest model/heuristic
  - Start with a linear model or boosted tree, using features from the heuristics as inputs
- Iterative improvement
  - Conduct small-scale experiments
    - Target specific countries or segments
  - Perform A/B testing if possible; if not, use switch-back testing
- Evaluation & monitoring
  - Ensure your hypothesis aligns with reality
  - Identify and fix bugs

Chat Duration

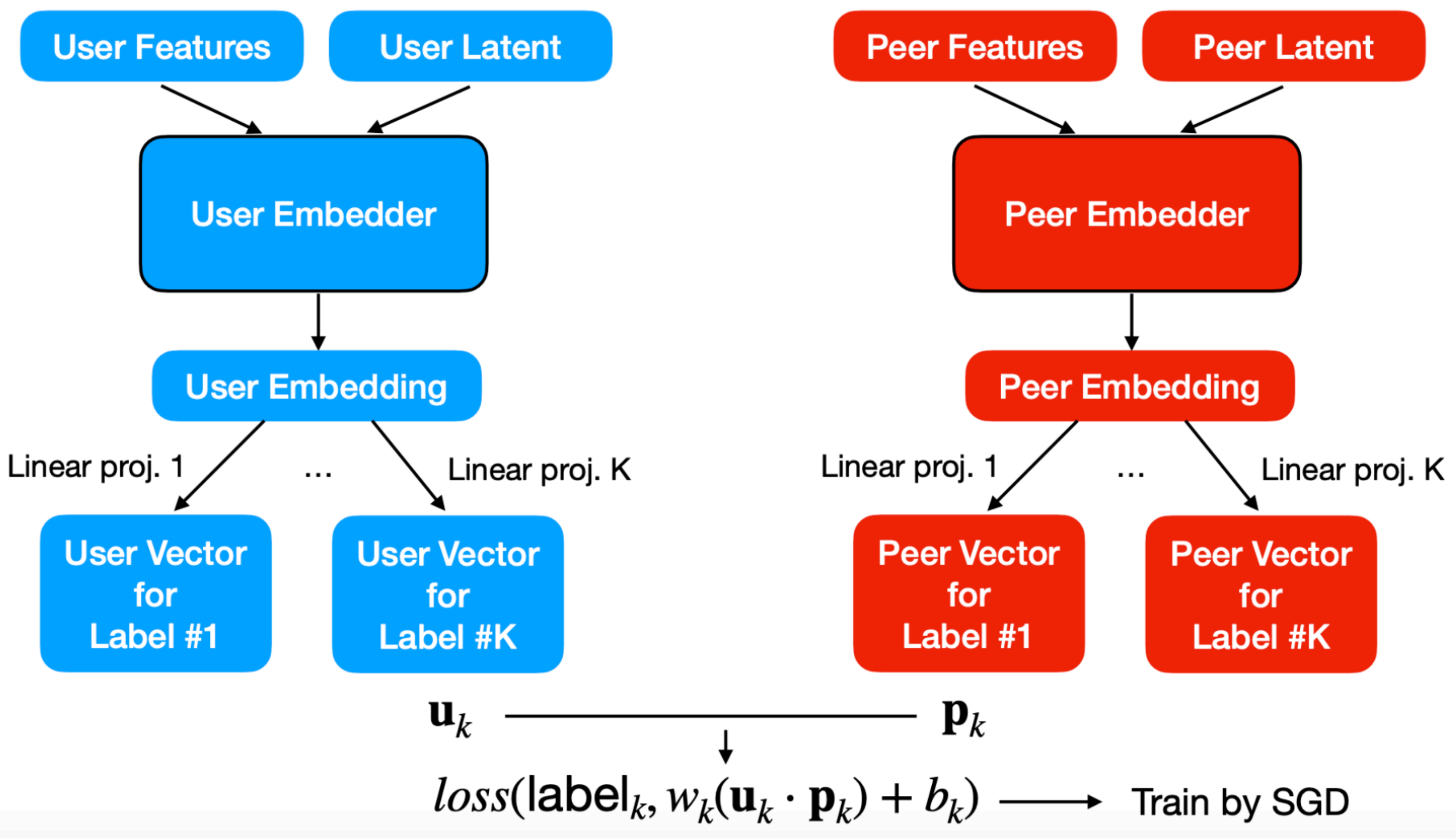
- First attempt
  - Develop a chat duration predictor and use it to generate more Aha Moments
  - Assumes IID, so can't address the supply-demand issue
  - However, tackling the most difficult problem from the start is not a good idea
    - Challenging to persuade stakeholders and iterate
- Even when 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

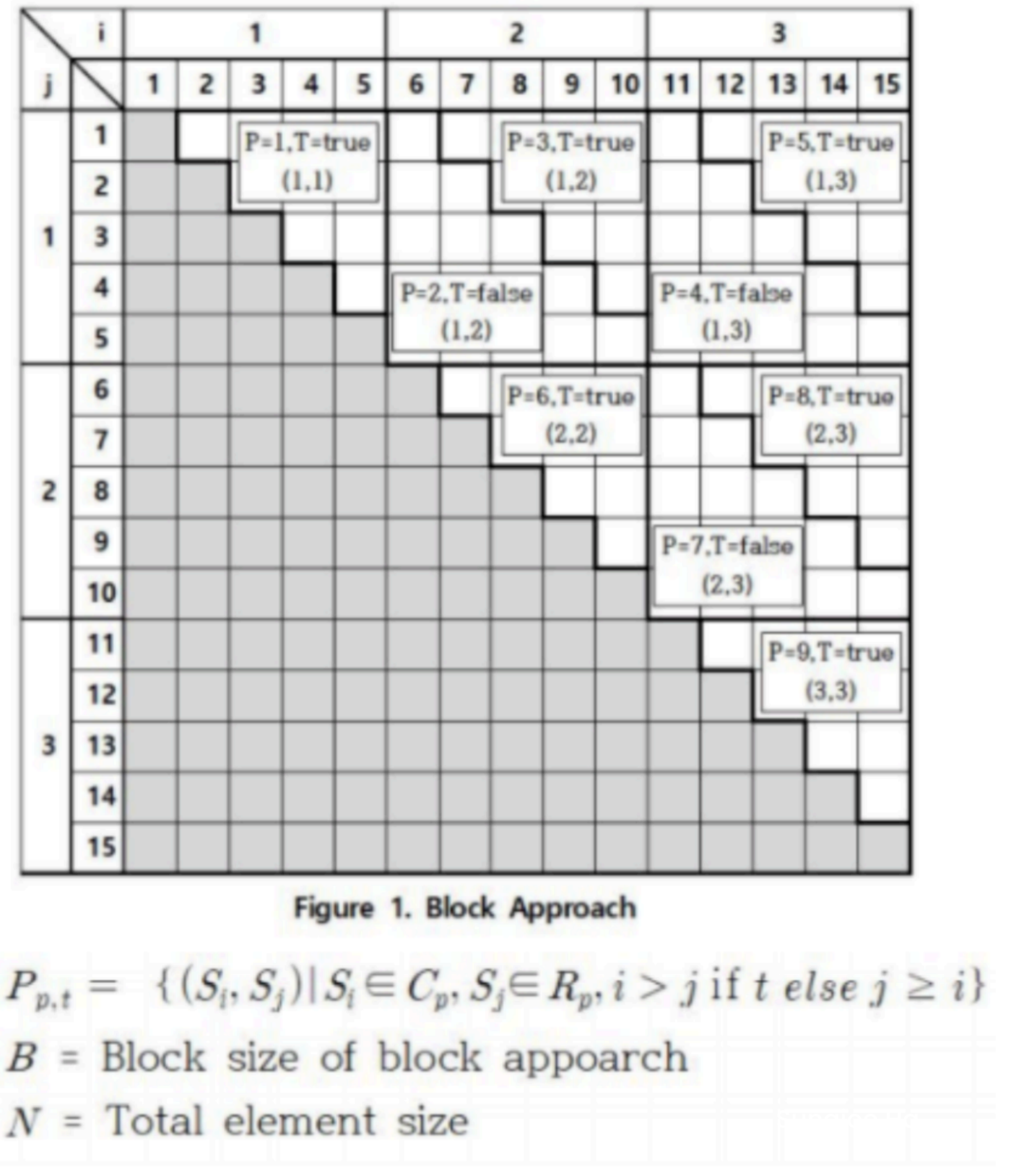


Figure 1. Block Approach

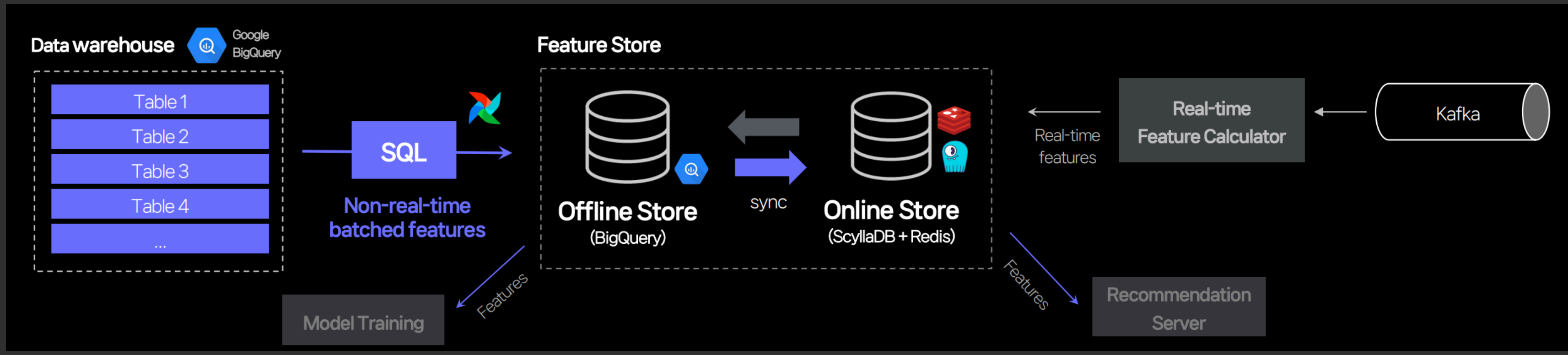
$$P_{p,t} = \{(S_i, S_j) | S_i \in C_p, S_j \in R_p, i > j \text{ if } t \text{ else } j \geq i\}$$

B = Block size of block appoarch

N = Total element size

Feature Store

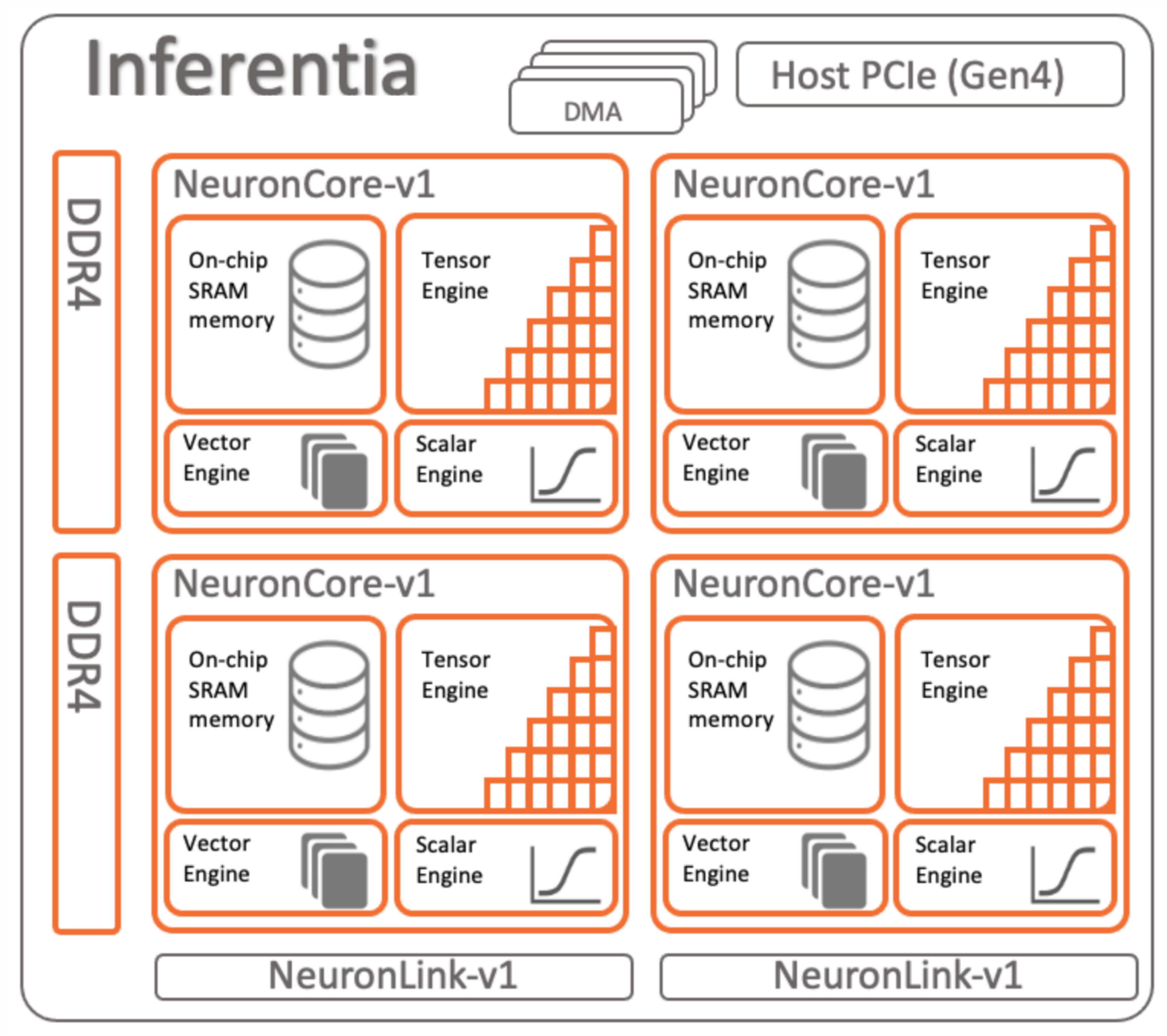
- Feature store[5] addresses the following issues:
  - Train/serving data discrepancies
  - High cost of adding features
  - Redundant components when deploying multiple ML applications
  - Difficulty sharing features when deploying multiple ML applications
  - Ensuring feature correctness



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

Inference Optimization

- AWS Inf1
  - 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

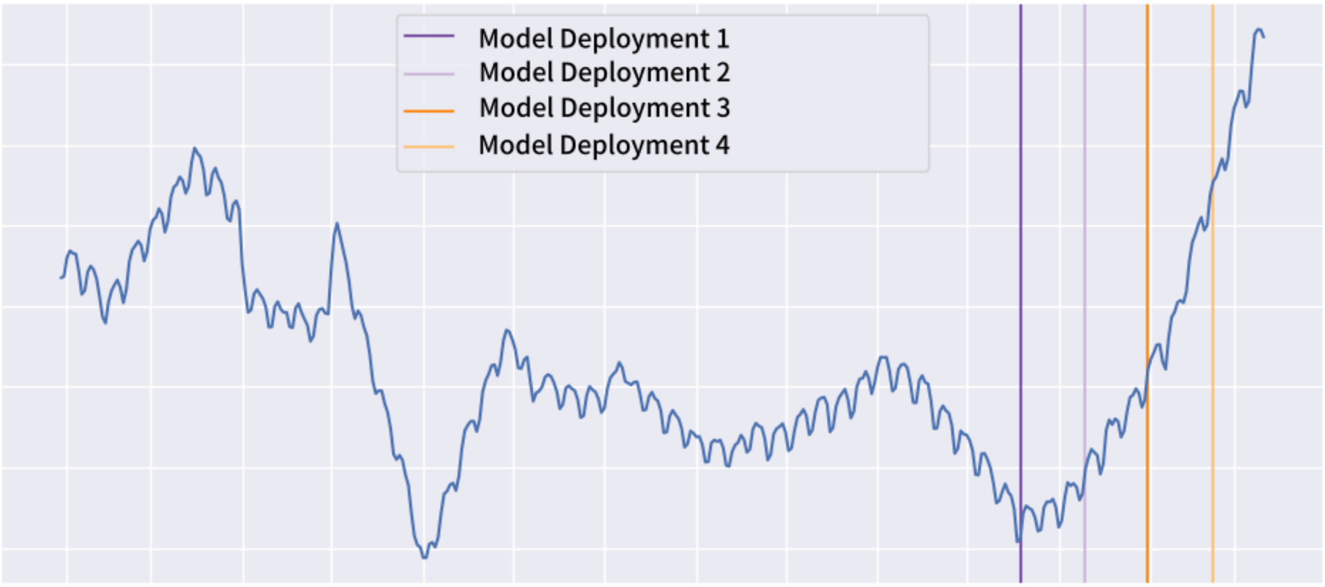


Engineering Optimization

- 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

Result

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



Recap

- Software engineering
  - Feature store
  - Parallelism
  - Python optimization
- Machine learning
  - Causal inference
  - Metrics
  - Inference optimization - batching & pipelining
- Broad view of the problem
  - AI/data flywheel
  - Learned business logic
  - Transforming core business problems into AI problems

Problem Formulation

- Problem finding, formulating, solving, and selling
  - Essential skills to acquire while in school
- Numerous problems exist in the world
  - Focus on finding suitable problems
    - Valuable and solvable
- Problem formulation
  - Various tools available
    - Ex: Using the language of mathematics to eliminate ambiguity
- Problem solving
  - The main focus of education
  - Strive for a deep understanding in whatever you do
- Selling
  - If no one buys what you're selling, you neither create nor capture value

Deep Dive

- Gaining deep dive experience is crucial
  - Ability to navigate between abstraction layers
  - A key quality sought during hiring
- As AI advances, this skill will become even more important
  - Superficial understanding will be replaced by AI
  - Developing your own perspective and deep understanding is difficult to replace
- Strive for a deep understanding of your work
  - Software engineering fundamentals
  - Machine learning foundations
  - Any other deep understanding