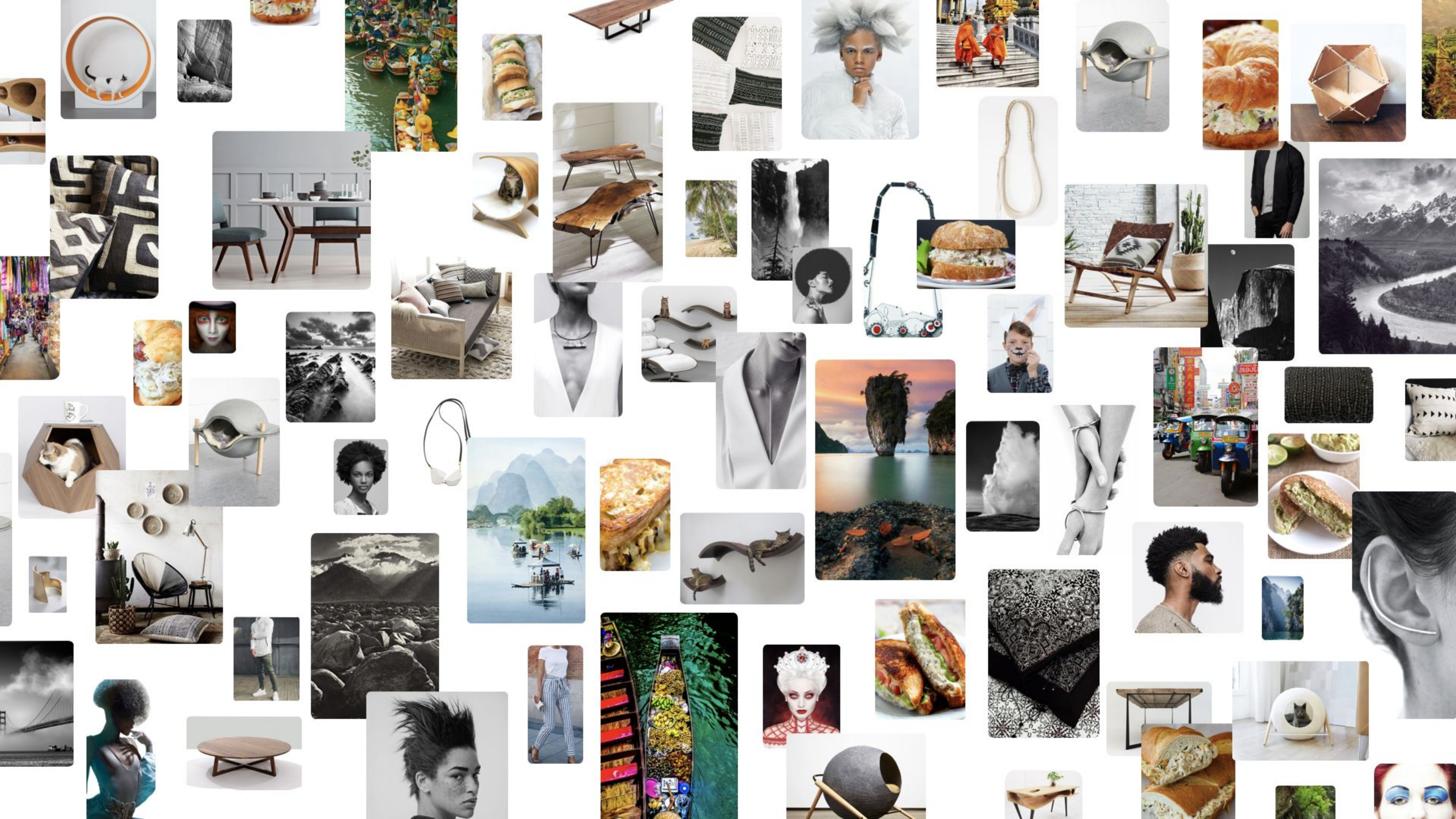
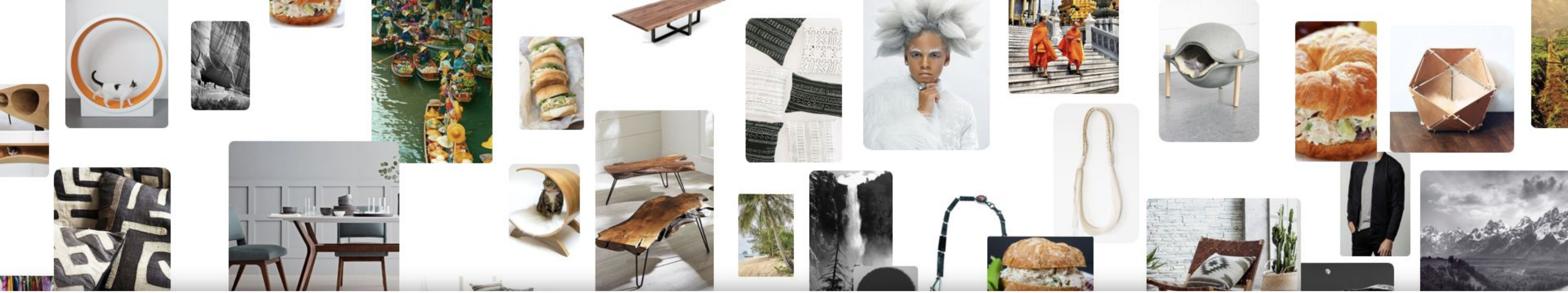
# *Powering the AI Inspiration Engine*

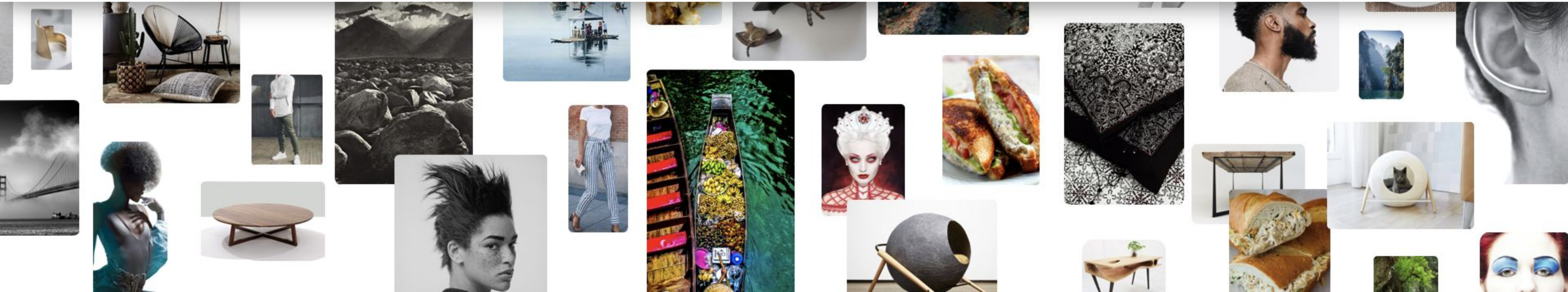
Andrew Zhai, Senior Staff Applied Scientist
Pinterest

April 25, 2022

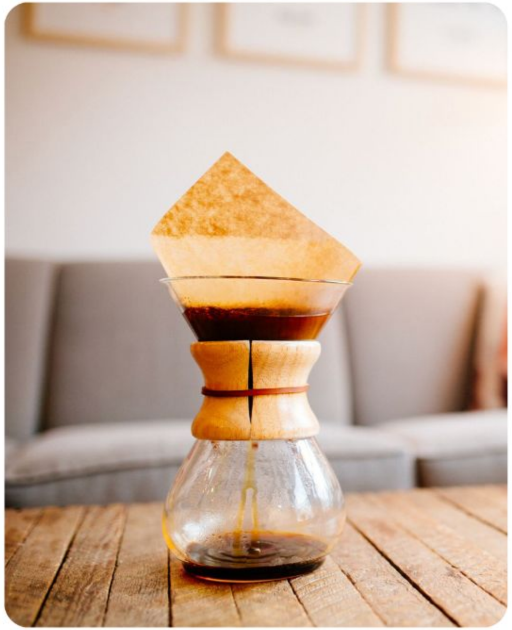




Bring ***everyone*** the ***inspiration*** to create a life they love



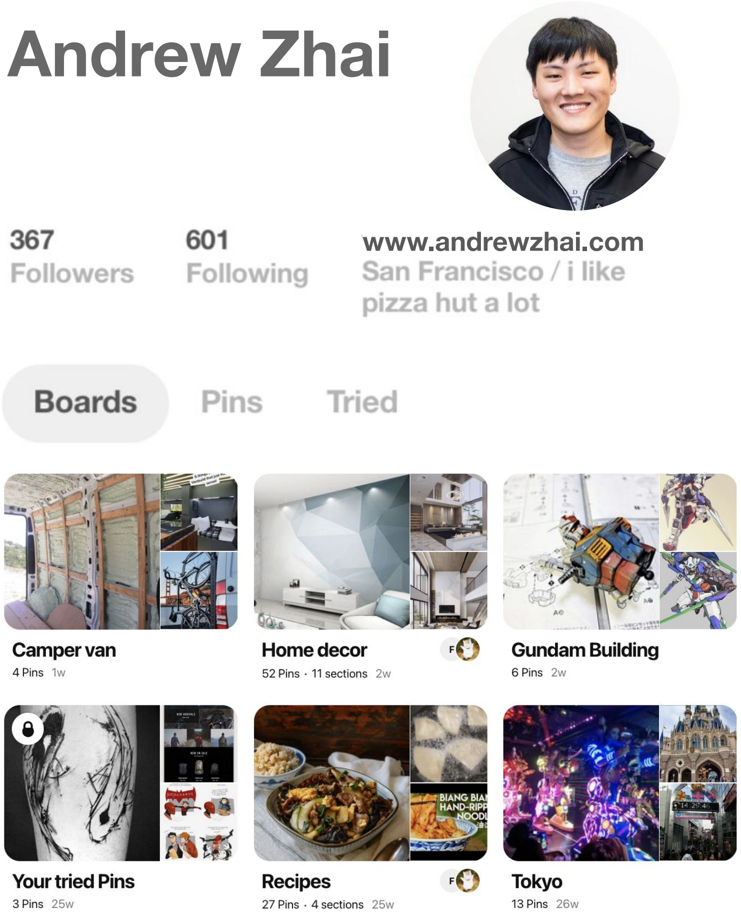
# Pin



**The perfect path to cold brew** 36

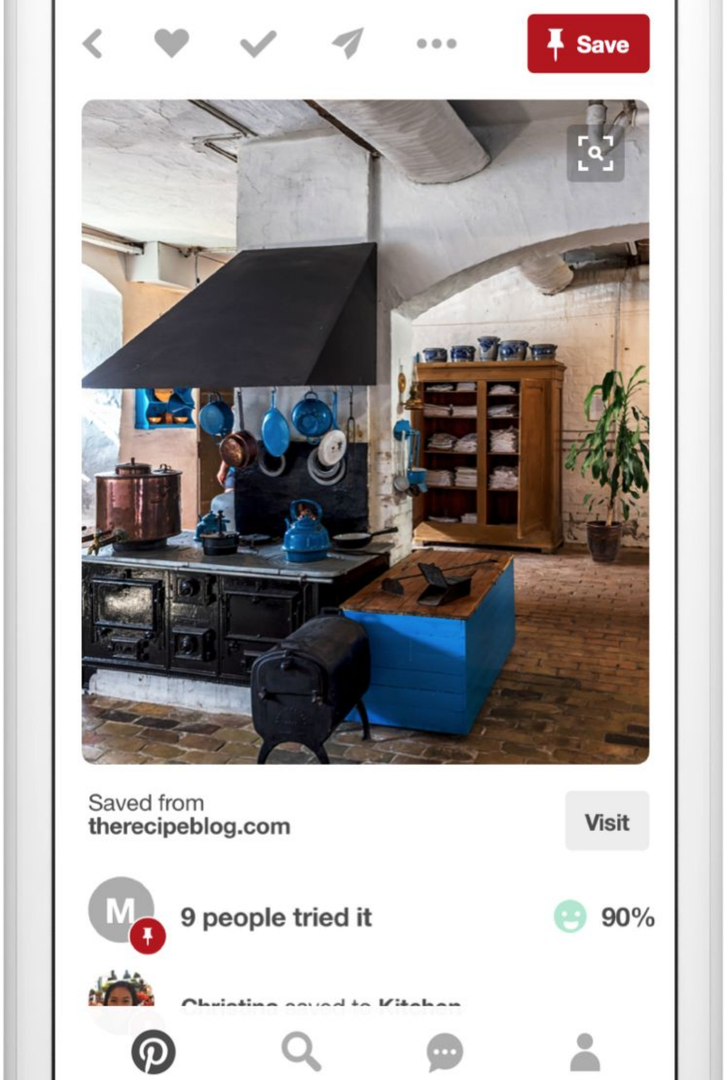
Caffeinated Inc.

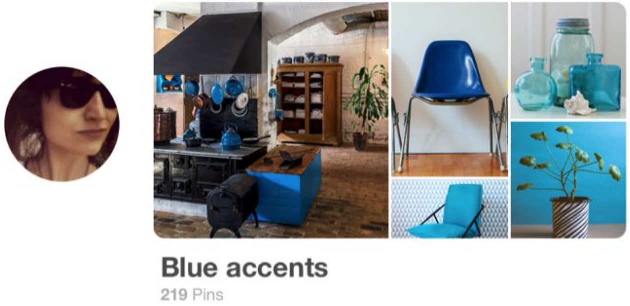
**Omar Seyal**
Cravings

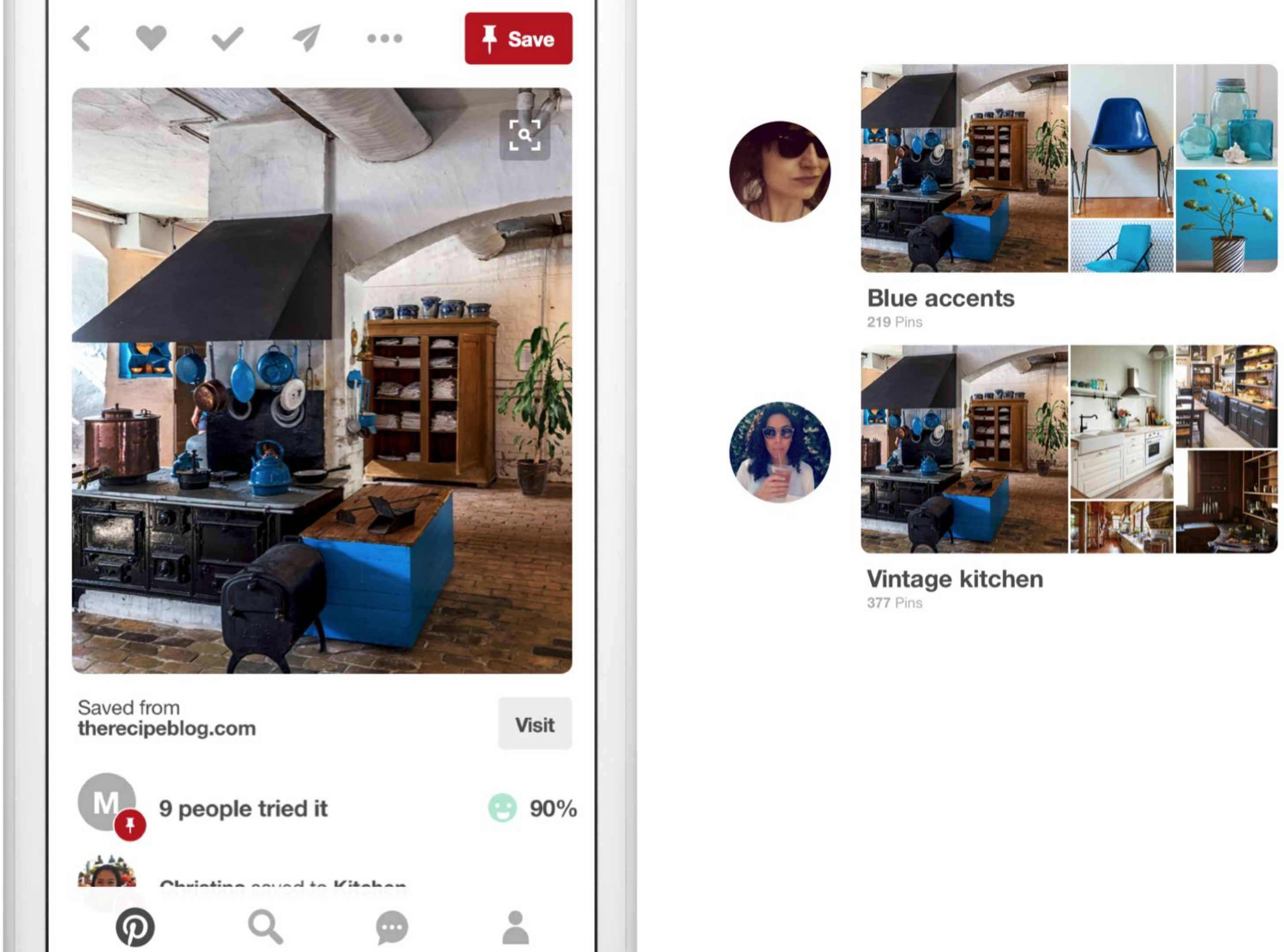


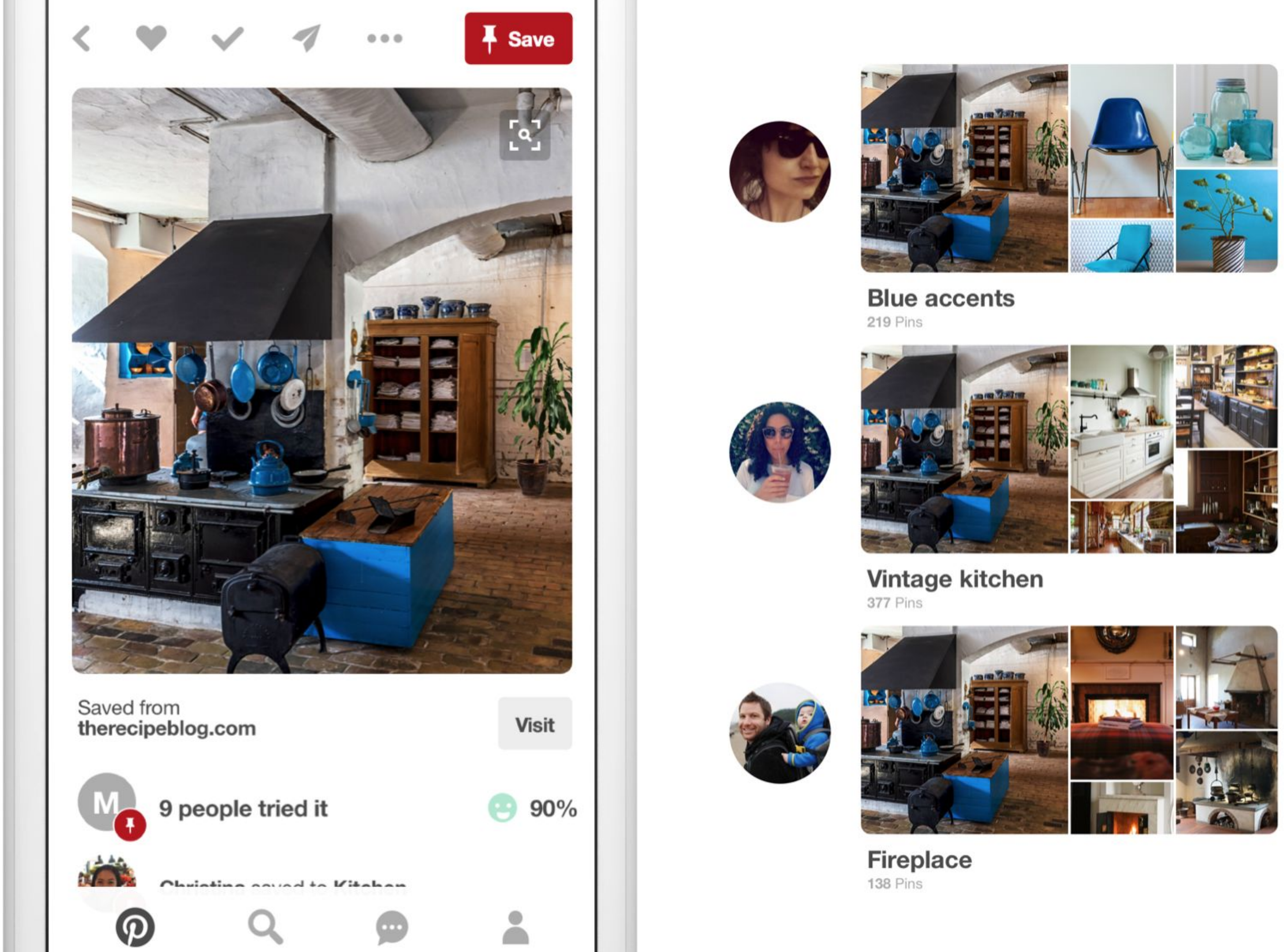

# Board

## A greater collection of ideas.


Save
Saved from
therecipeblog.com
Visit
M
9 people tried it
90%
Christina saved to Kitchen


Blue accents
219 Pins


Save
Saved from
therecipeblog.com
Visit
M
9 people tried it
90%
Blue accents
219 Pins
Vintage kitchen
377 Pins


Save
Saved from
therecipeblog.com
Visit
M
9 people tried it
90%
Blue accents
219 Pins
Vintage kitchen
377 Pins
Fireplace
138 Pins

People on Pinterest each month

# 400m+

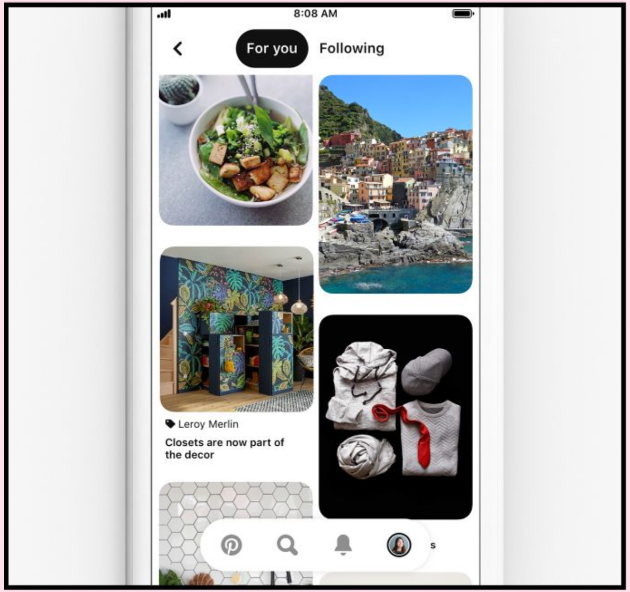
## 330b+
Pins

## 7b+
Boards

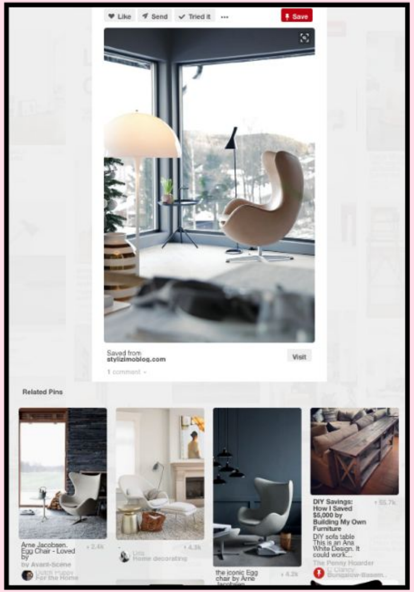
## 35+
Languages

## 91%
say Pinterest is a place filled with **Positivity**

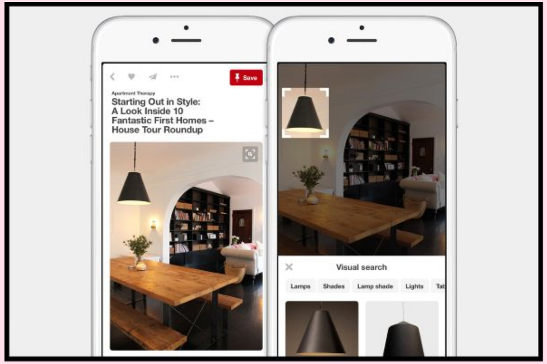
# The Inspiration Engine




Homefeed (User)




Related Pins (Pin)




Visual Search (Image)

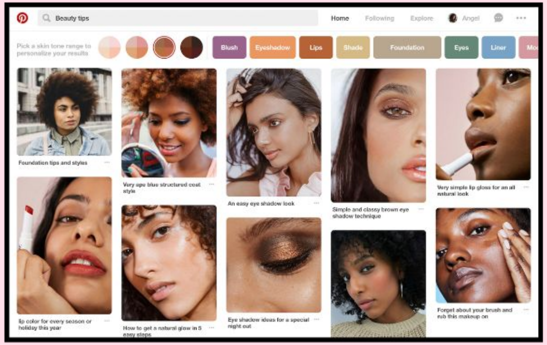
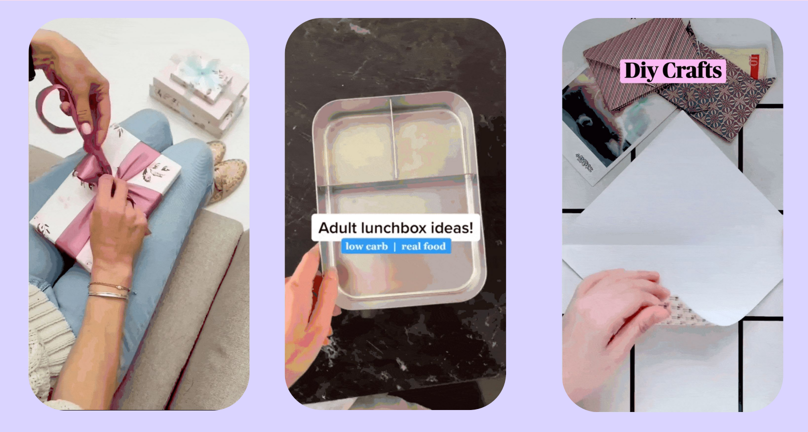



Image Search (Text)

# Inspirational Engagement


Top pins by view time

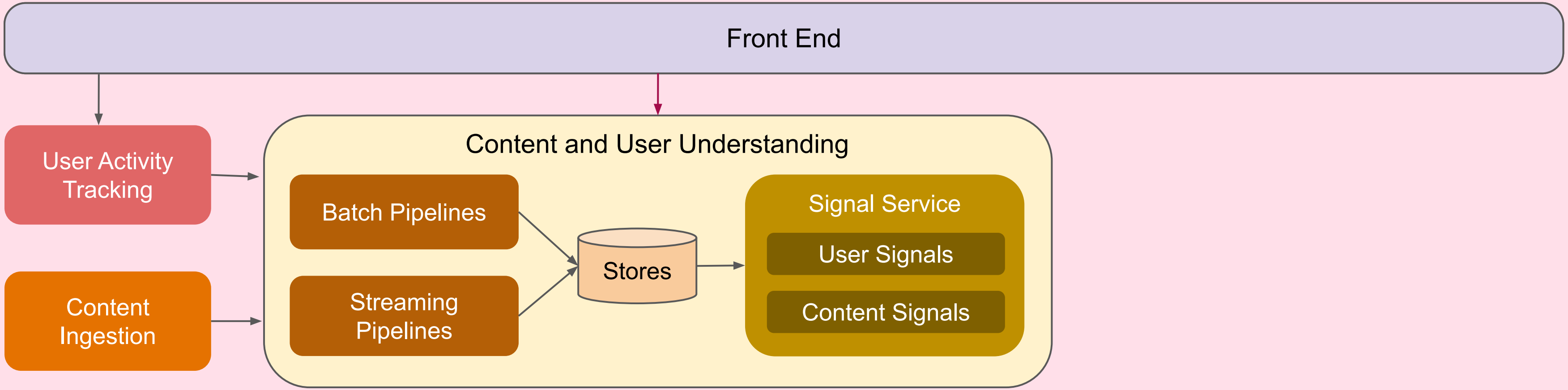



Top pins by Saves

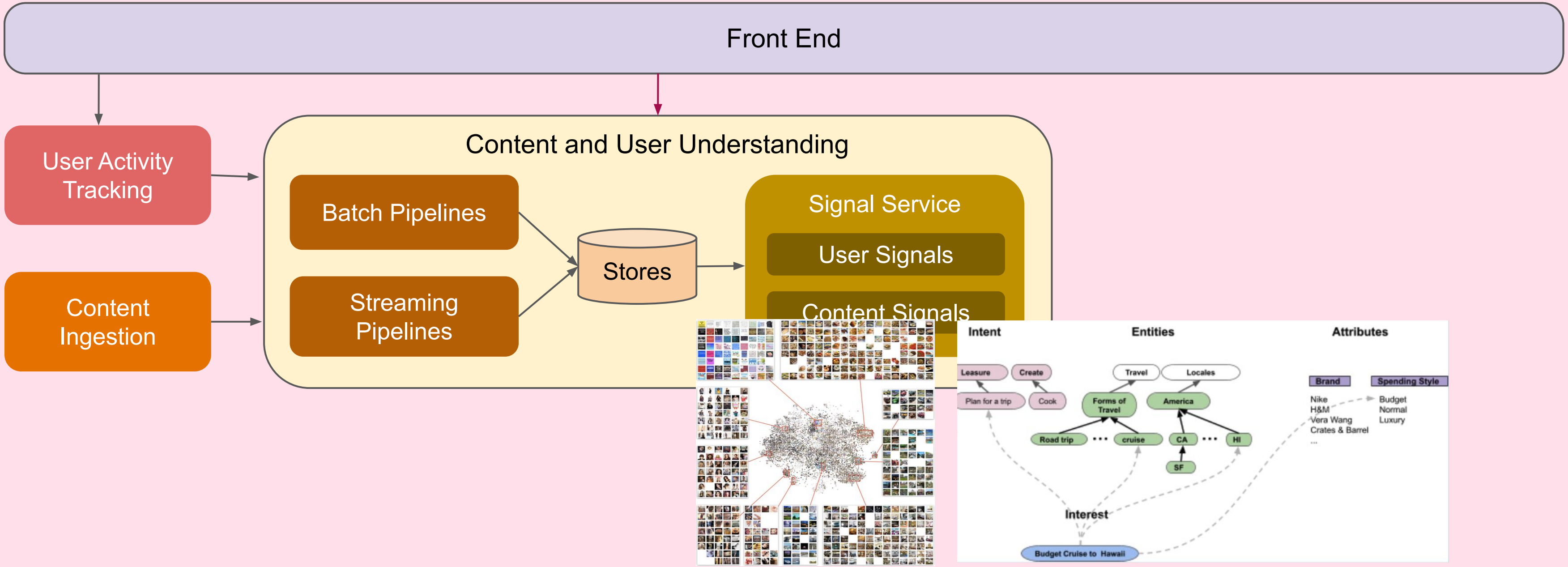
# System Architecture for scalability - 00s of thousands of users and few billion Pins (content)

Front End

# System Architecture for scalability - 00s of thousands of users and few billion Pins (content)

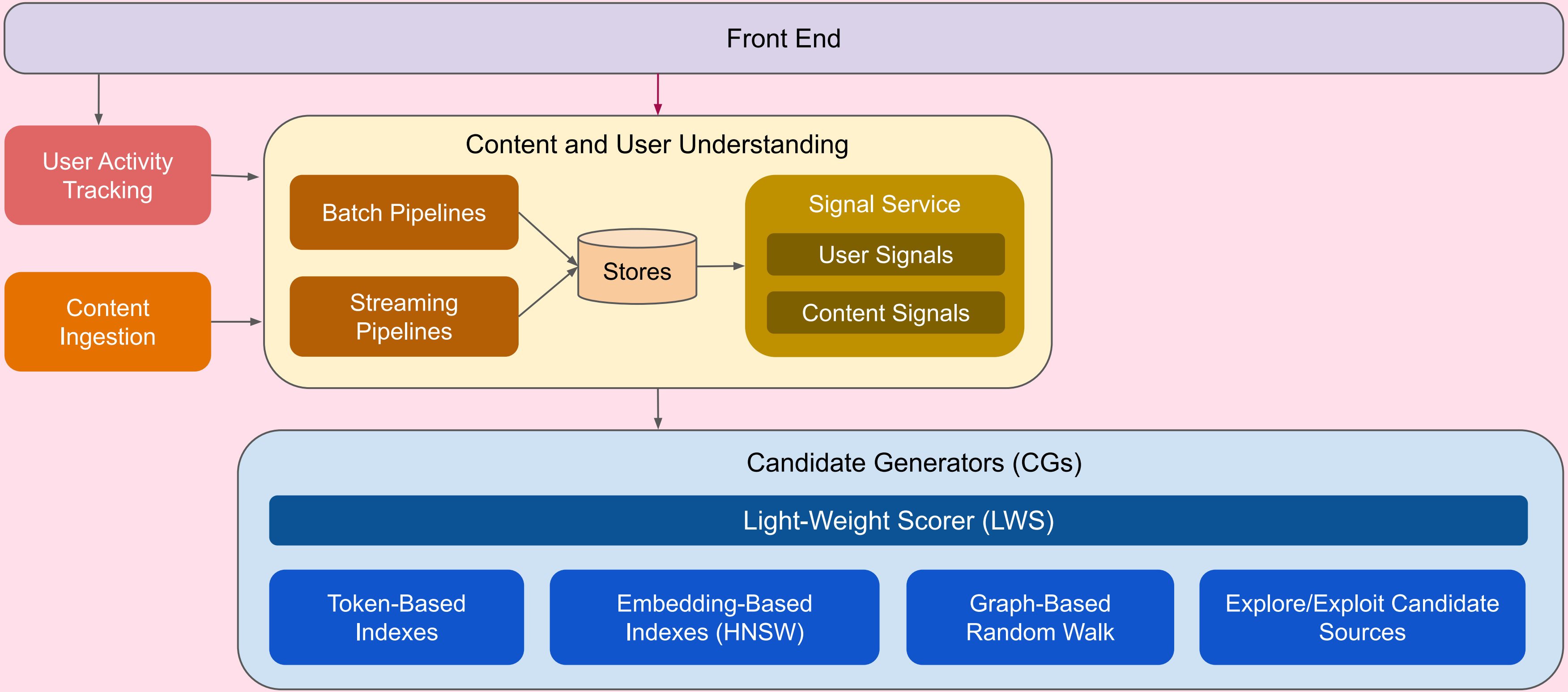


# System Architecture for scalability - 00s of thousands of users and few billion Pins (content)

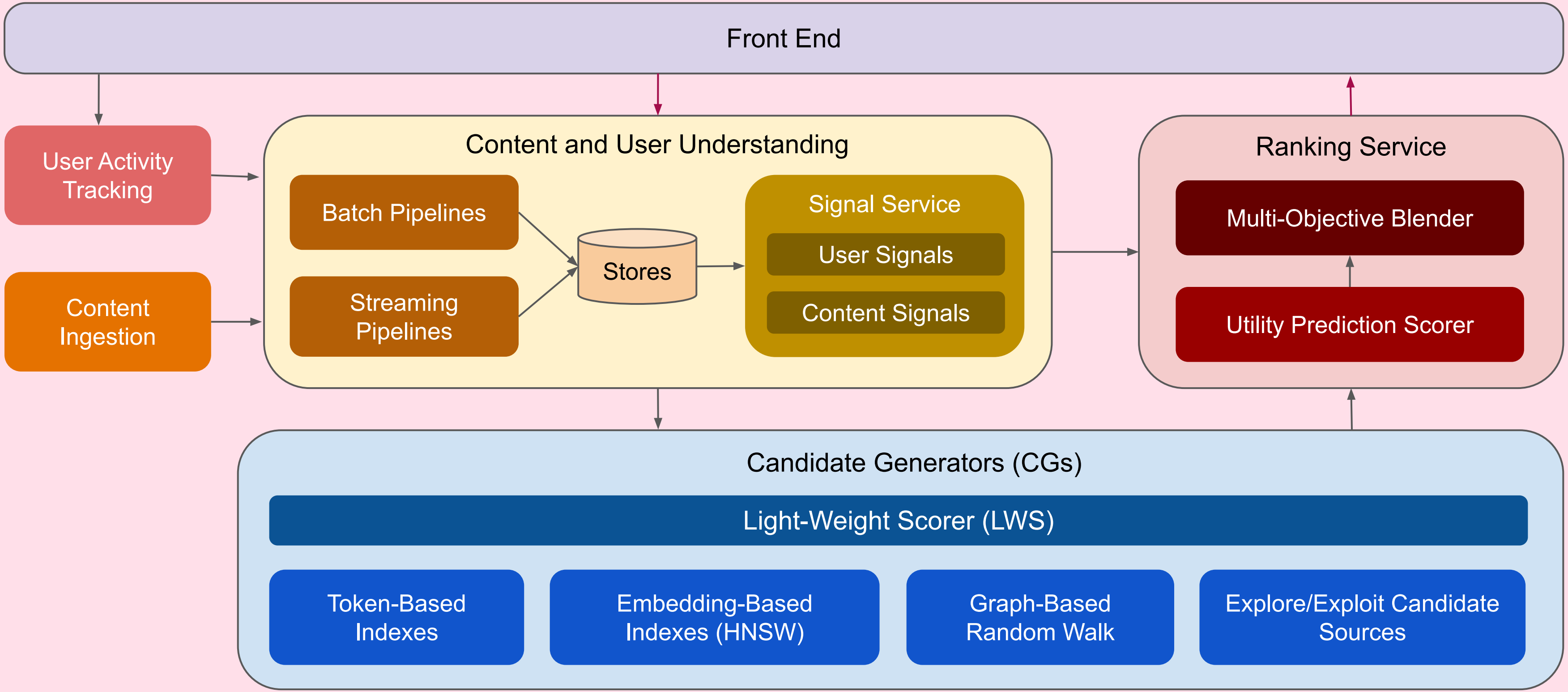



**Embeddings** **Knowledge Graph**

# System Architecture for scalability - 00s of thousands of users and few billion Pins (content)

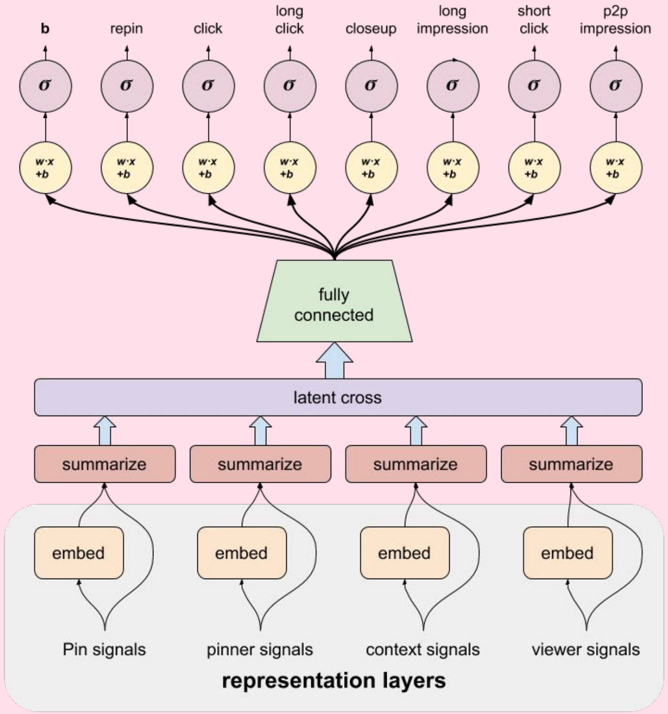


# System Architecture for scalability - 00s of thousands of users and few billion Pins (content)




Pinterest

# Ranking: User Action Prediction



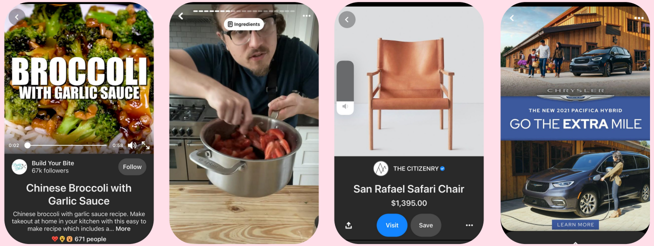

- Predict a wide variety of user actions for each (user, item) pair through multi-head deep neural network

# Multi Objective Optimization

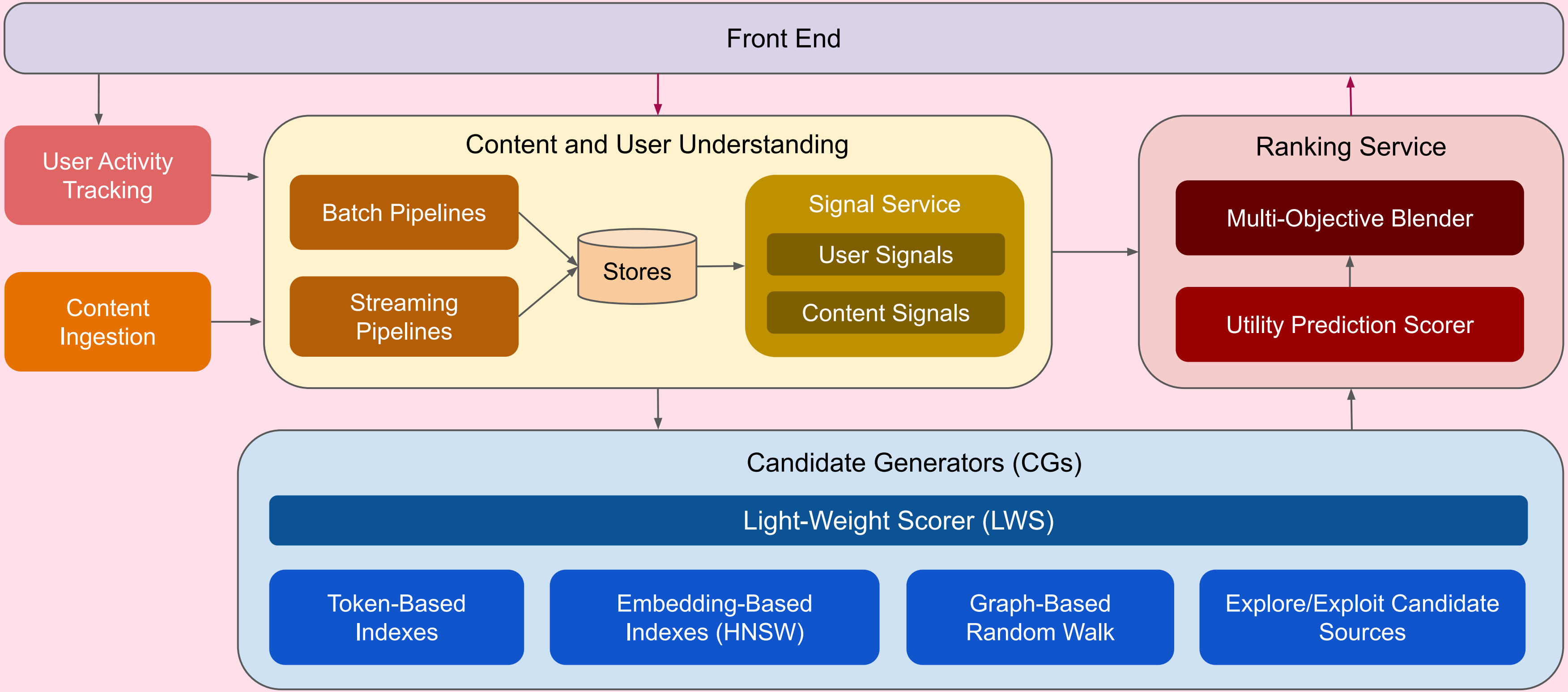
$$\max_{\boldsymbol{x}} \text{PinnerUtility}(\boldsymbol{x})$$
$$\text{s.t.} \quad \text{CreatorUtility}(\ ) \geq \theta_1$$
$$\text{MerchantUtility}(\boldsymbol{x}) \geq \theta_2$$
$$\text{AdUtility}(\boldsymbol{x}) \geq \theta_3$$

$$\max_{\boldsymbol{x}} \text{PinnerUtility}(\boldsymbol{x})$$
$$+ w_1 \text{ CreatorUtility}(\ )$$
$$+ w_2 \text{ MerchantUtility}(\boldsymbol{x})$$
$$+ w_3 \text{ AdUtility}(\boldsymbol{x})$$

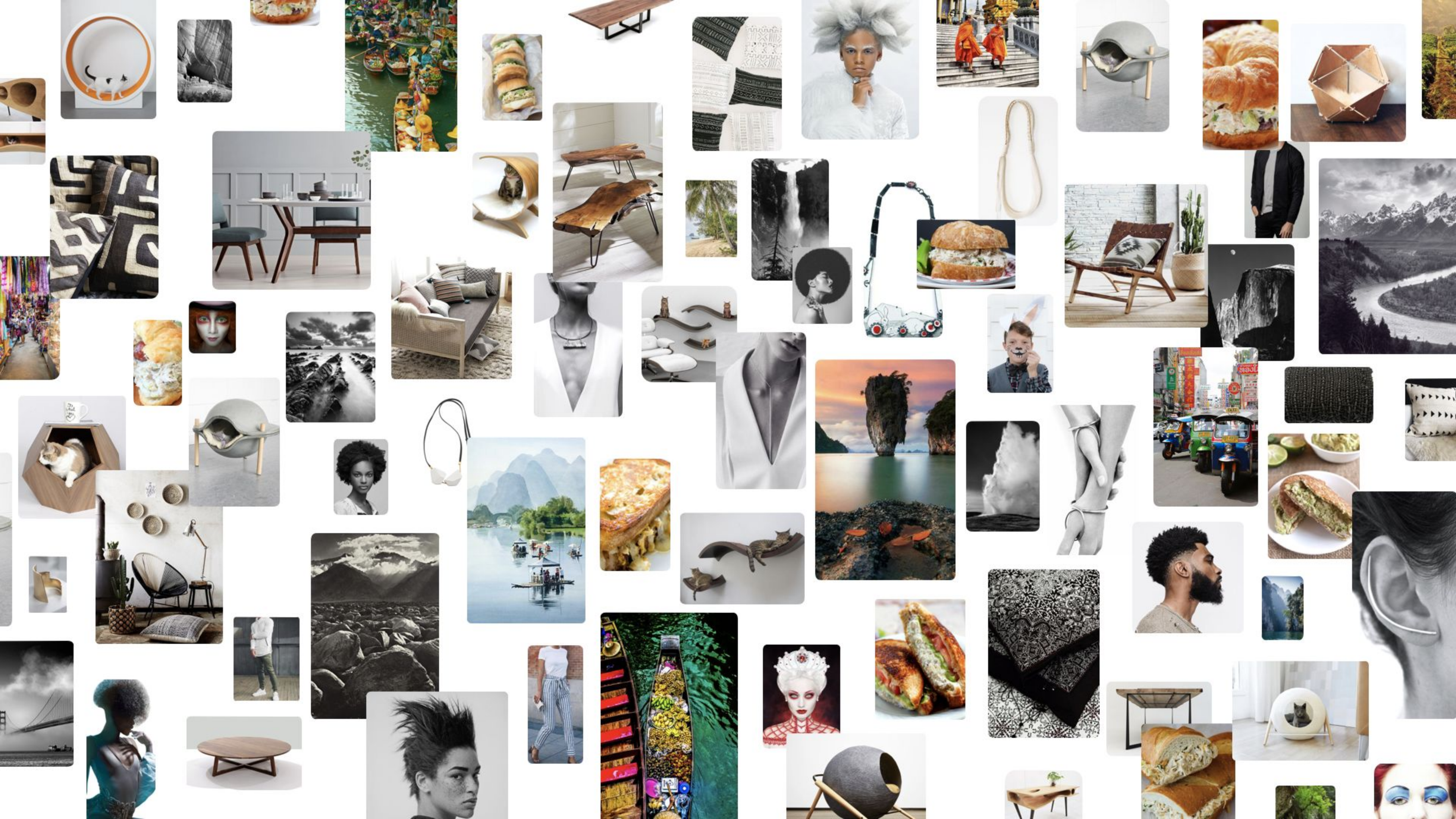
- Estimate utility values for different parties on Pinterest based on predicted action probabilities
- Tune the weights to achieve a desired tradeoff
- Real system - Functional form contains non-linearities are present



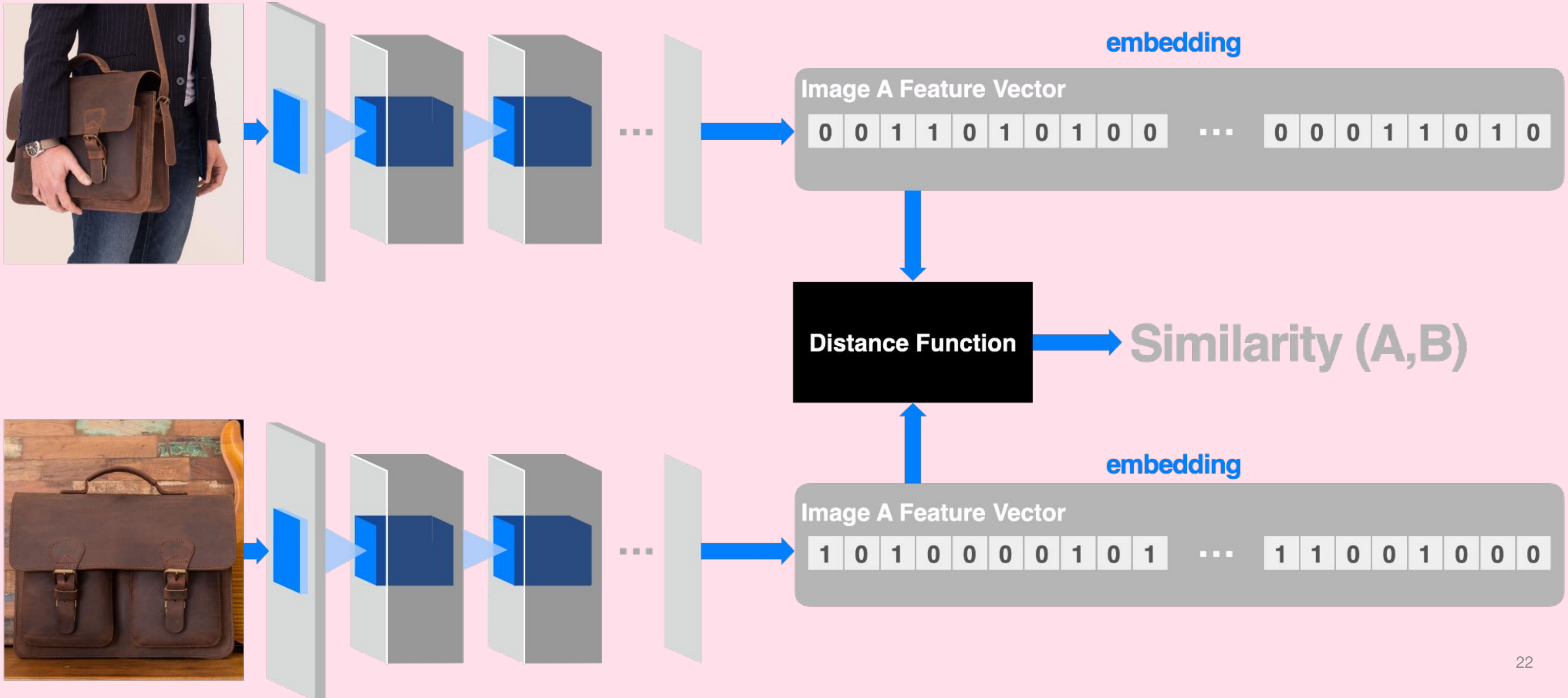
# System Architecture for scalability - 00s of thousands of users and few billion Pins (content)



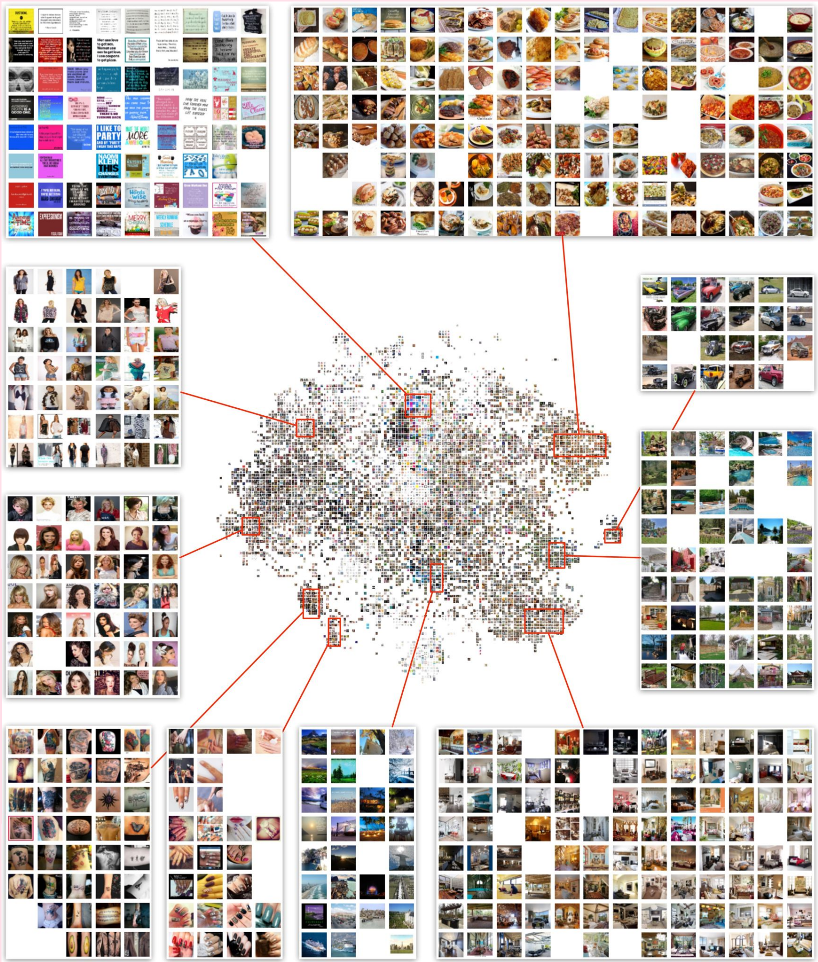
# Content Understanding



# Determining visual similarity



Zhai et al. “Visual Discovery at Pinterest”, WWW ‘17

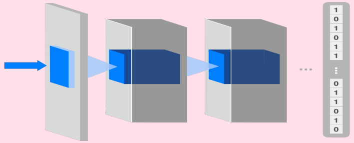


# Embeddings

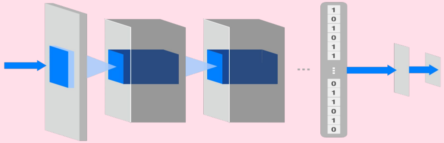
Encode very different types of data (images, pin, user)

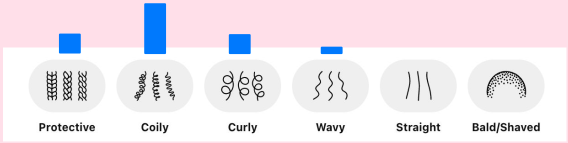
# Application 1

## Application 1



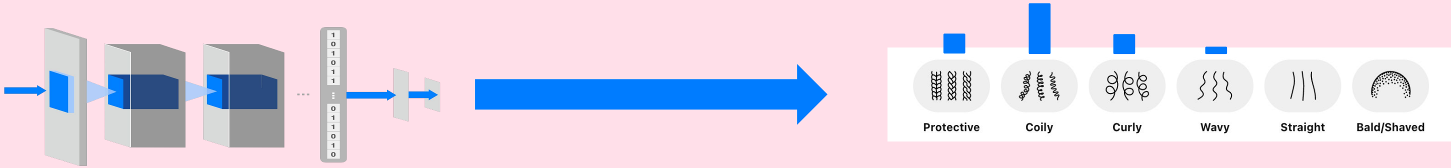
## Application 2





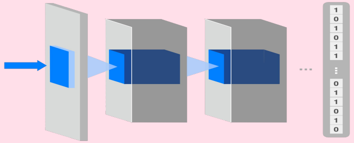
## Application 1

## Application 2

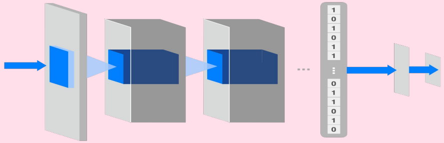


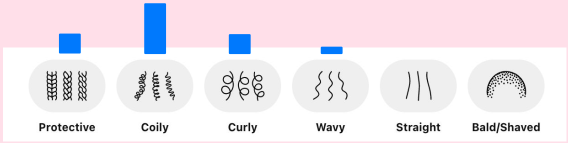

## Application 3

## Application 1

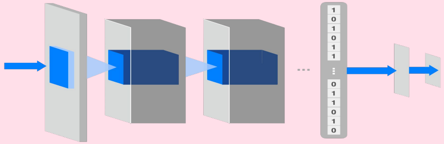


## Application 2



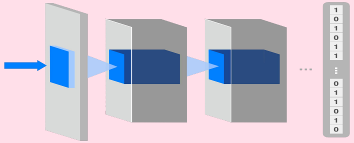



## Application 3



| Color | Fabric |
|---|---|
| **blue=38.78**, | **cotton=42.24**, |
| purple=16.46, | chiffon=14.61, |
| green=11.07, | silk=7.90, |
| black=8.26, | wool=5.61 |
| multi=7.52 | |

## Application N



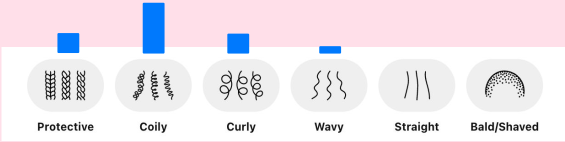
?

## Application 1

vgg16

## Application 2

resnet101




## Application 3

se-resnext101

| Color | Fabric |
|---|---|
| **blue=38.78**, | **cotton=42.24**, |
| purple=16.46, | chiffon=14.61, |
| green=11.07, | silk=7.90, |
| black=8.26, | wool=5.61 |
| multi=7.52 | |

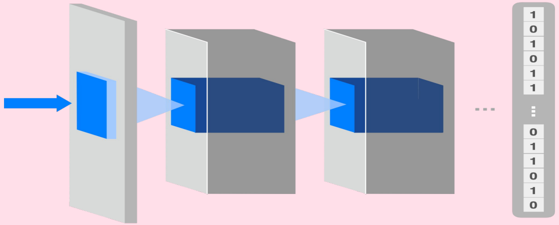
## Application N

?

?

# “Unified” Visual Backbone

Zhai et al. “Learning a Unified Embedding for Visual Search at Pinterest”, KDD’19





**Output: 20+ tasks** across exact product matching, neardup, skin tone classifier

**Benefits**

- Scalable maintainence (most important)
- Joint learning across dataset
- Share foundational improvements

Zhai et al. “Learning a Unified Embedding for Visual Search at Pinterest”, KDD’19

# “Unified” Visual Backbone

Zhai et al. “Learning a Unified Embedding for Visual Search at Pinterest”,

| Model | STL P@1 | Flashlight Avg P@20 | Lens Avg P@20 |
|---|---|---|---|
| Old Shop-the-Look | 33.0 | - | - |
| Old Flashlight | - | 53.4 | - |
| Old Lens | - | - | 17.8 |
| ImageNet | 5.6 | 33.1 | 15.0 |
| Ours | **52.8** | **60.2** | **18.4** |

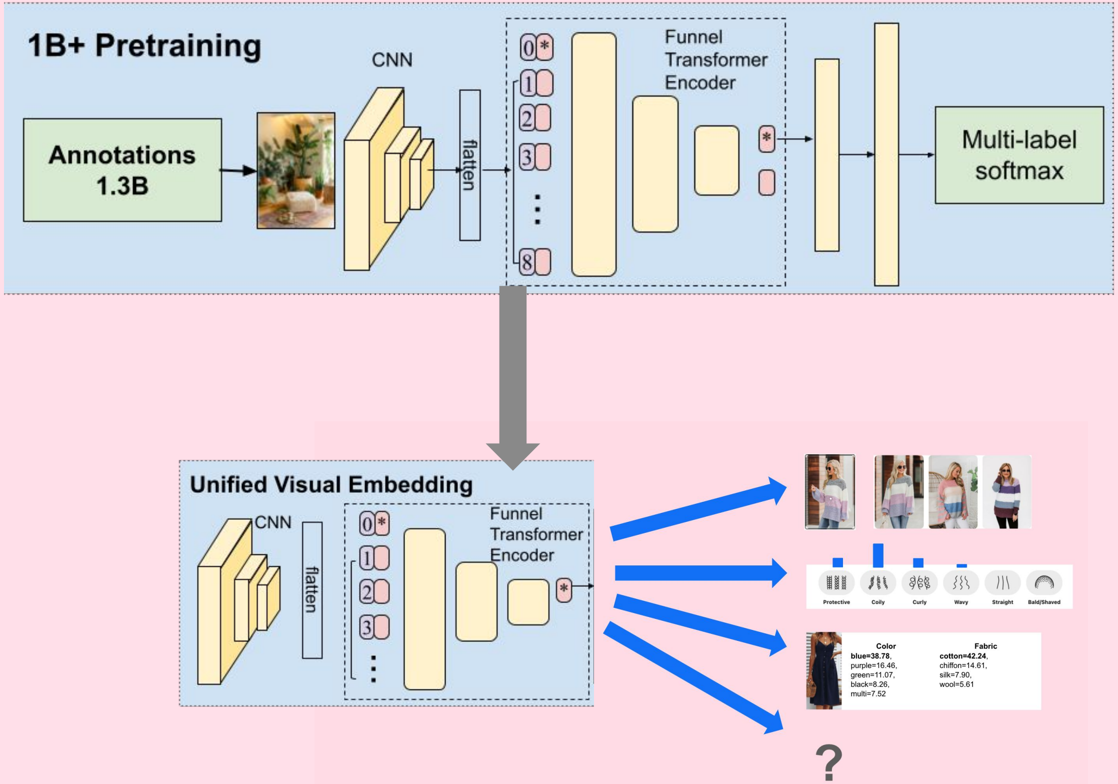
| Dataset | STL P@1 | Flashlight Avg P@20 | Lens Avg P@20 |
|---|---|---|---|
| Shop-the-Look (S) | 49.2 | 42.1 | 14.7 |
| Flashlight (F) | 11.0 | 53.4 | 16.1 |
| Lens (L) | 26.2 | 47.8 | 18.2 |
| All (S + F + L) | **52.8** | **60.2** | **18.4** |

Multi-Task Embedding > Single-Task Embedding

All Dataset > Single Dataset

Billion-Scale Pretraining with Vision Transformers for Multi-Task Visual Representations, Beal. et al, WACV 2022

# Billion-Scale Pretrain Lifts All




## Pretrain

- 1.3B image pretraining
- Funnel Hybrid ViT

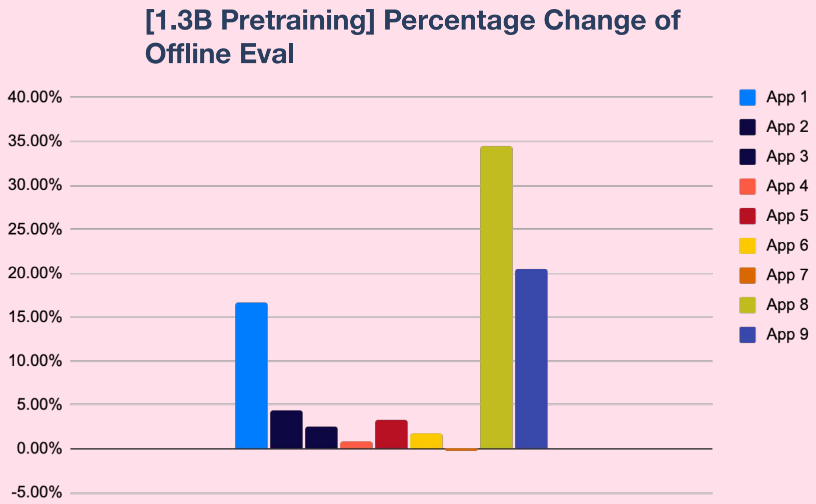
## Finetune

Pinterest

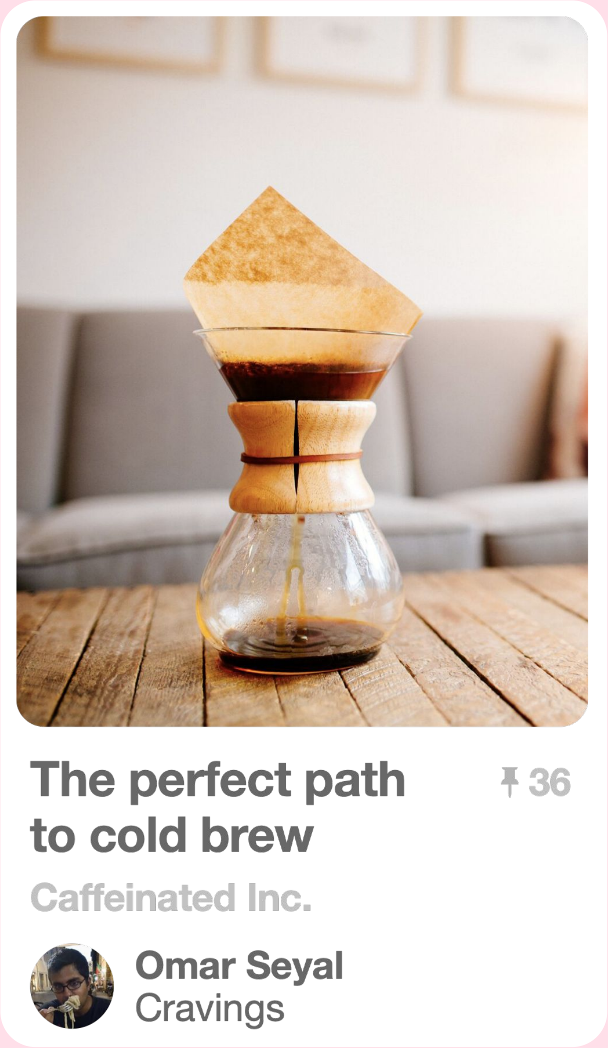
# Billion-Scale Pretrain Lifts All

Billion-Scale Pretraining with Vision Transformers for Multi-Task Visual Representations, Beal. et al, WACV 2022

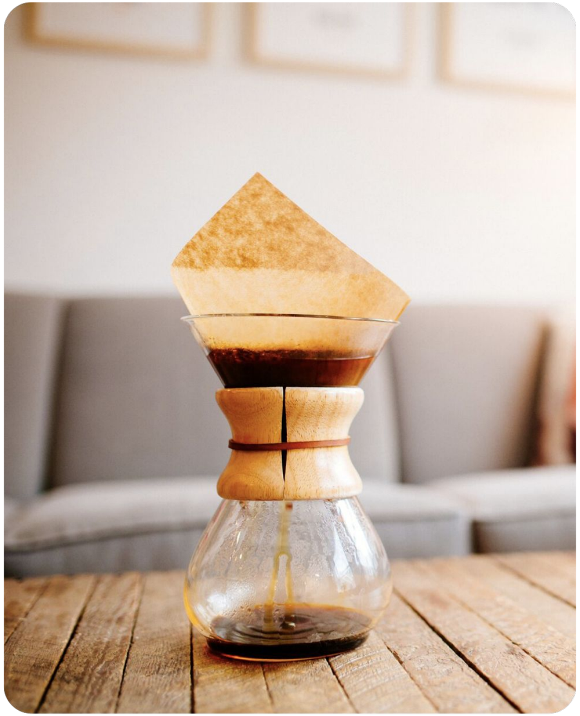
| Model | Pretraining | VS | F | L | C |
|---|---|---|---|---|---|
| RN-101 | IN-1k | 39.6 | 59.7 | 17.2 | 85.2 |
| RN-101 | IG-940M | 46.7 | 67.6 | 20.2 | 87.9 |
| RN-101 | ANN-1.3B | 52.4 | 70.8 | 22.7 | 88.8 |
| ViT-B/32 | IN-1k | 29.2 | 44.7 | 15.2 | 82.3 |
| ViT-B/32 | ANN-1.3B | 46.4 | 68.9 | 24.9 | 86.5 |
| ViT-B/16 | ANN-1.3B | **54.7** | **74.3** | **26.7** | **89.7** |




Billion-scale Pretraining Lifts majority of application performance




**Challenge:** How to represent all dimensions of our content?



**The perfect path to cold brew** 36

Caffeinated Inc.

**Omar Seyal**
Cravings

Image

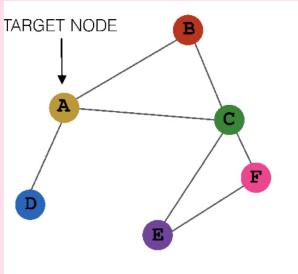
Title

**The perfect path to cold brew**

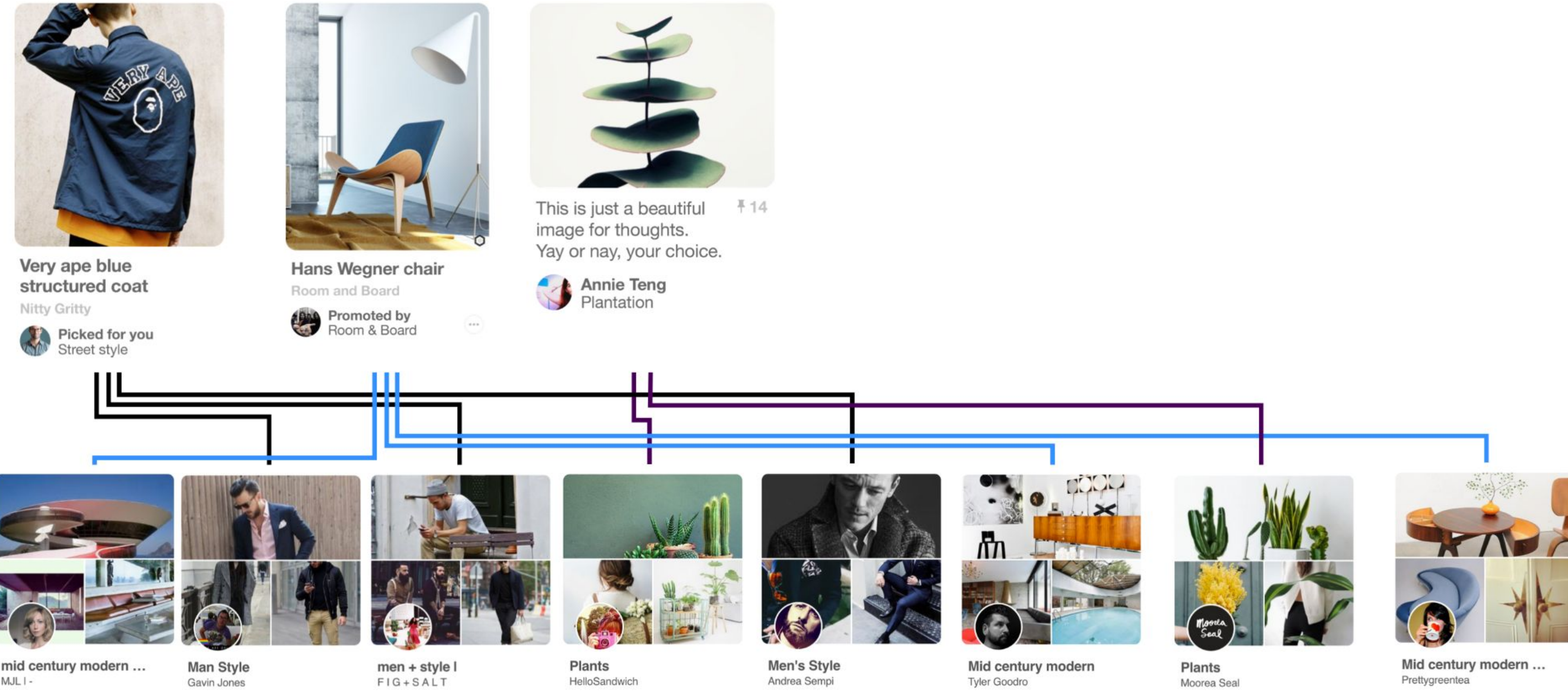
Creator

**Omar Seyal**

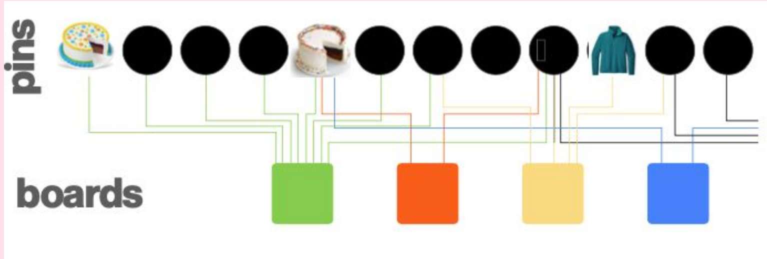
Pin-board Graph

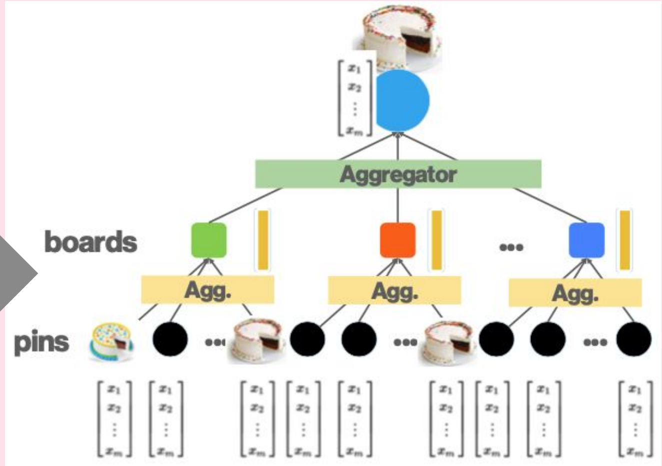


# Harnessing the Pin Board Graph



# PinSAGE: Graph Neural Network

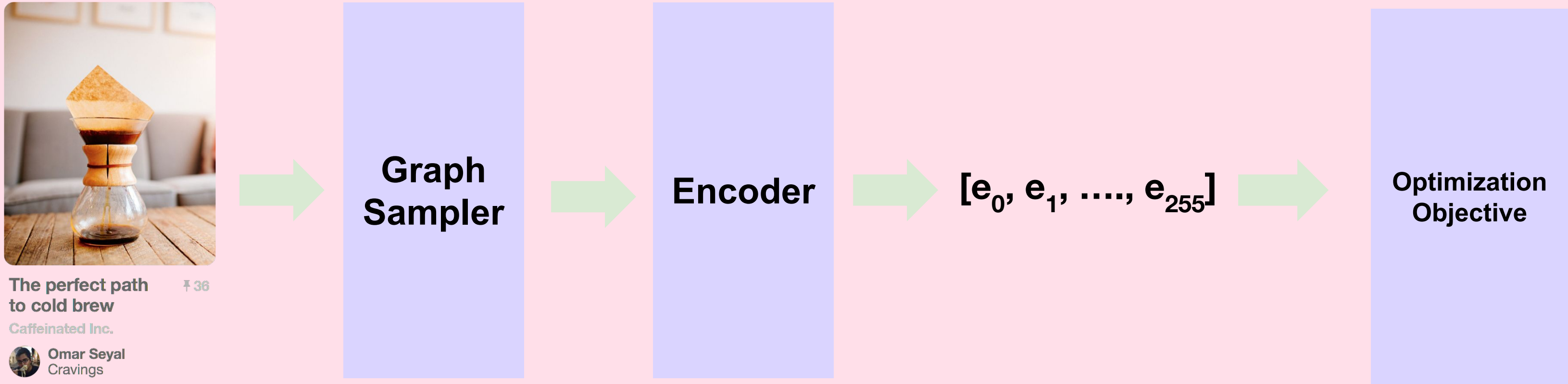



Graph with **3 billion** nodes and **18 billion** edges

Graph Convolutional Neural Networks for Web-Scale Recommender Systems, Ying et al., 2018

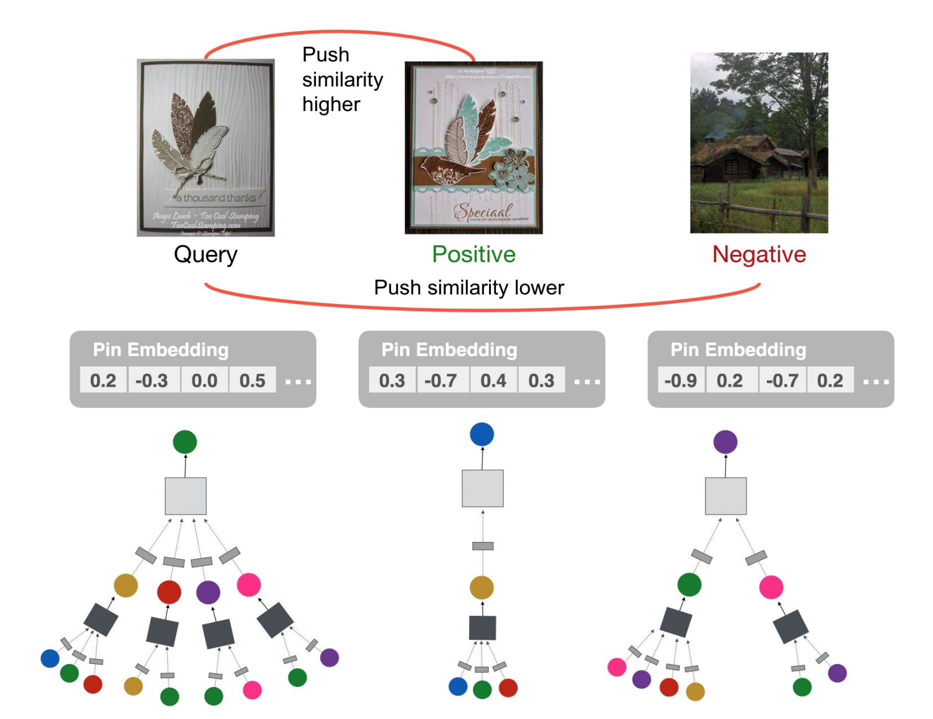
# PinSAGE: Graph Neural Network




From pin **features and graph**, encode into **embeddings** trained so pins that are "**related**" have **similar** embeddings

Graph Convolutional Neural Networks for Web-Scale Recommender Systems, Ying et al., 2018

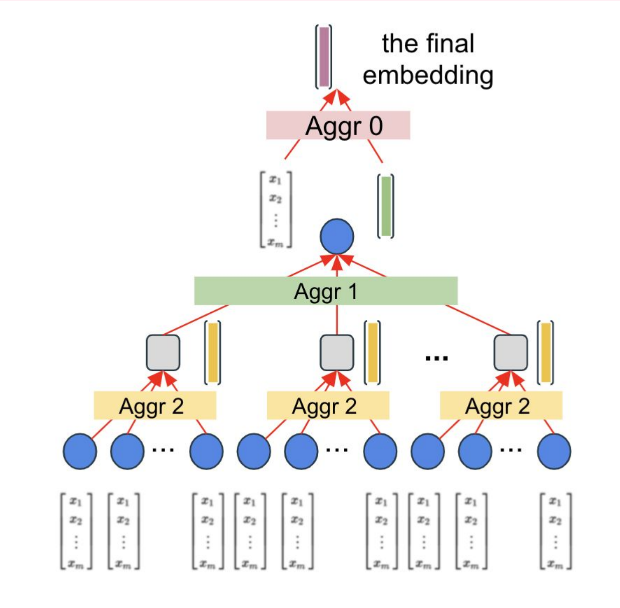
# PinSAGE: Optimization



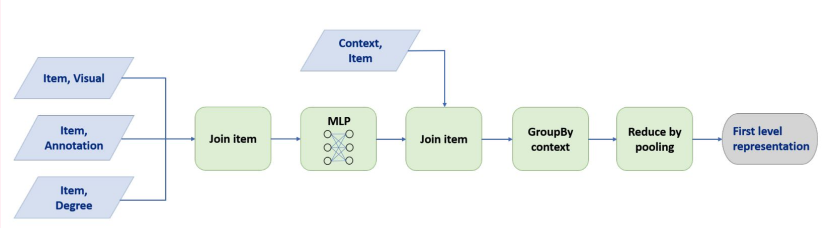

PinSage V1 (~triplet loss)

$$L = \frac{1}{|D|} \sum_{(q,p,n)\in D} max(0, e_q^T e_n - e_q^T e_p + m)$$

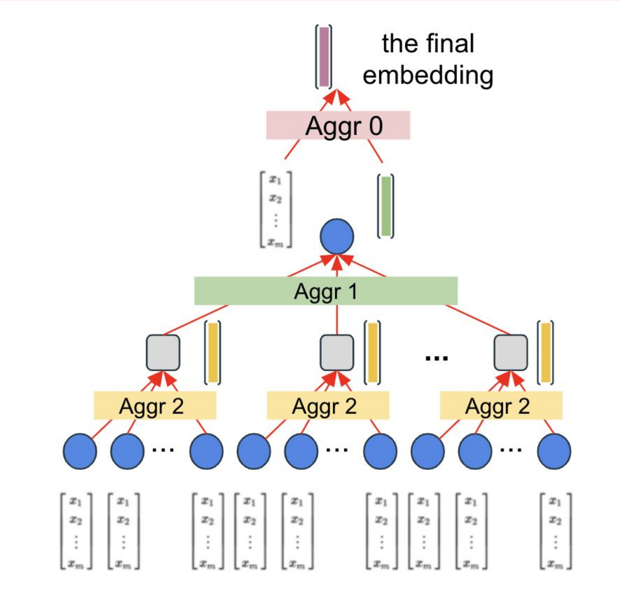
# V1: Graph Sampling on the Fly



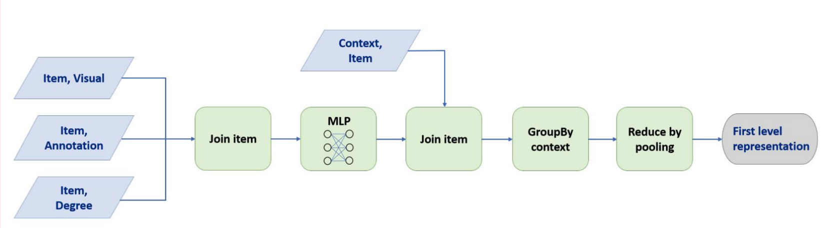

- **Sample Method**: K-hop neighborhood sampling
  - Pin -> board -> pin
- **Train Infra:** Graph sampling on the fly
  - 1.5TB RAM GPU machine (custom hardware)
  - **Only 2** available at Pinterest….
- **Inference Infra**: **Hardwire** architecture as Hadoop Jobs



# V1: Graph Sampling on the Fly




- **Sample Method**: K-hop neighborhood sampling
  - Pin -> board -> pin
- **Train Infra:** Graph sampling on the fly
  - 1.5TB RAM GPU machine (custom hardware)
  - **Only 2** available at Pinterest….
- **Inference Infra**: **Hardwire** architecture as Hadoop Jobs



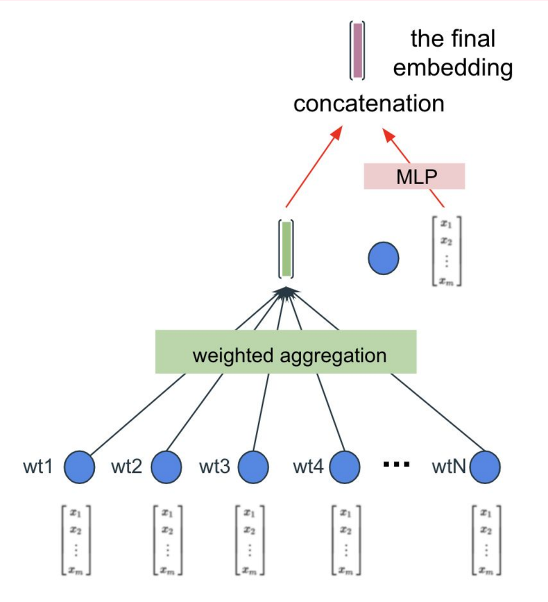

Pro:

- It works! Best performing content embedding at 3B nodes and 18B edges scale

Con:

- Not scalable to more developers nor flexible for iterations
- Train & serve completely separate stacks

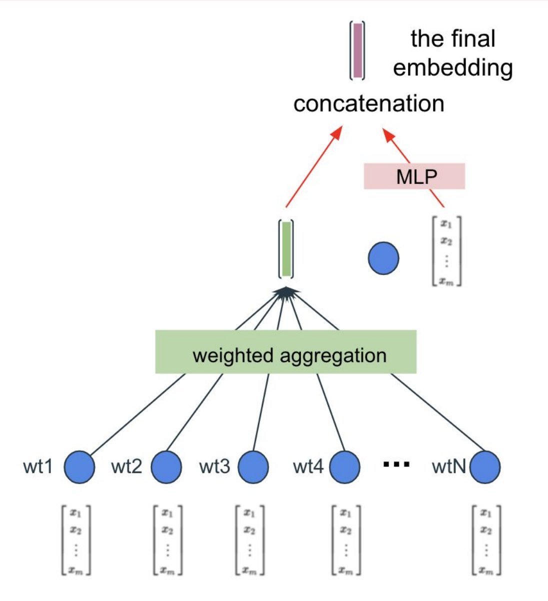
# V2: Offline Graph Sampling




- Scalability challenges due to graph sampling on the fly
  - **Solution**: Move sampling out of training / inference

- **Sample Method:** Random walks (50 neighbors)
- **Data Prep**:
  - Compute 3B * 50 random walk in a daily workflow
  - Materializes self + neighbor features for each pin example
- **Train & Inference Infra:**
  - Stream example through model

# V2: Offline Graph Sampling



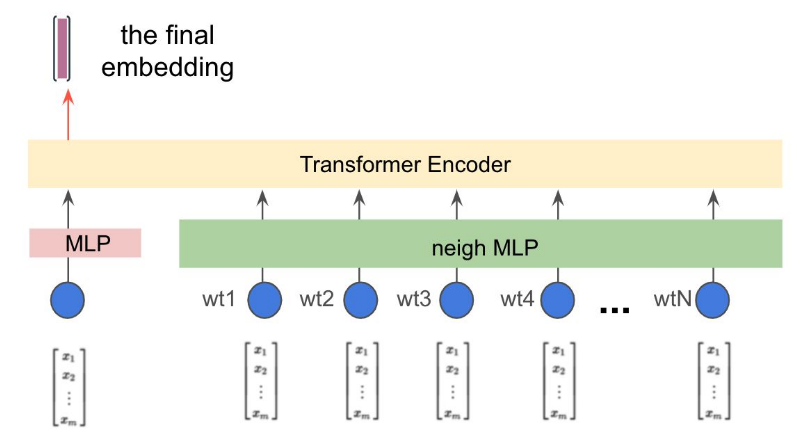

- Scalability challenges due to graph sampling on the fly
  - **Solution**: Move sampling out of training / inference

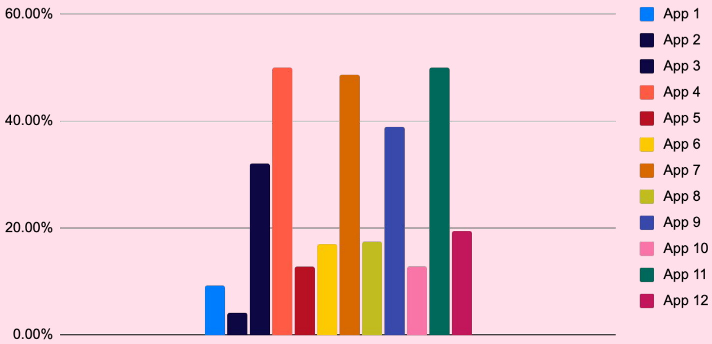
- **Sample Method:** Random walks (50 neighbors)
- **Data Prep**:
  - Compute 3B * 50 random walk in a daily workflow
  - Materializes <u>self + neighbor features</u> for each pin example
- **Train & Inference Infra:**
  - Stream example through model

Pro:
- Leverage commodity hardware
- +46% offline performance

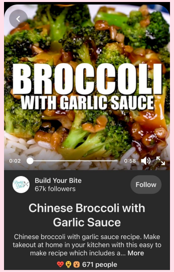
Con:
- Harder to iterate on graph sampling algorithm

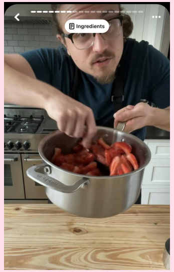
# V3: Multi-Task GNN Transformer

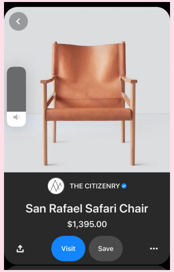



**PinSage V3** vs **V2**



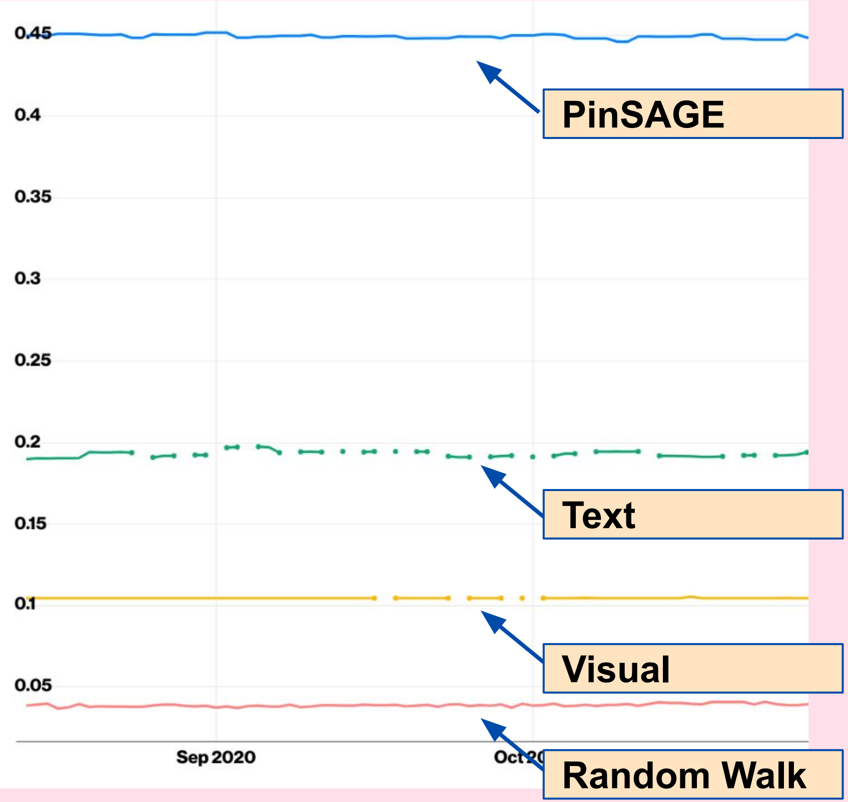





- **Multi-Task** - 16 objectives to optimize different content formats
- **TransformerEncoder** - why not early fuse neighbor and self features?

# GNNs produce the Best Content Representation



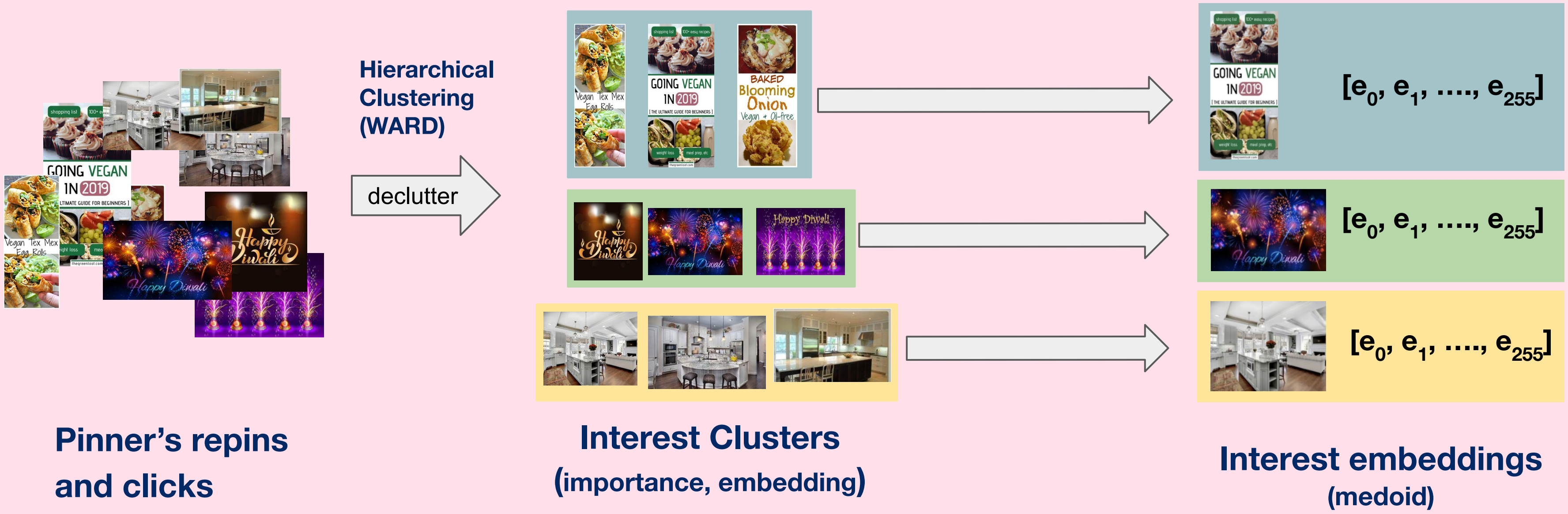

Pinterest

70+ launches

across recommendation systems, T&S, knowledge understanding, shopping, advertisement, ...

# User Modeling

# PinnerSage




PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest, Pal et al. KDD 2020

# PinnerSage

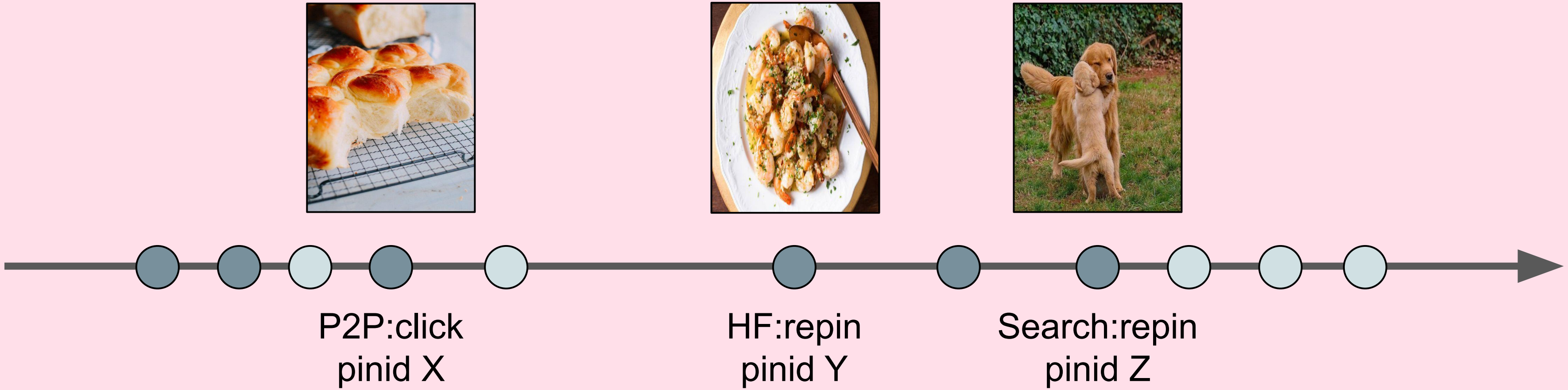

Pro:

- Simple and effective. 10+ launches (e.g. +3% HF repin/click volume)
- Interpretable, debuggable

Con:

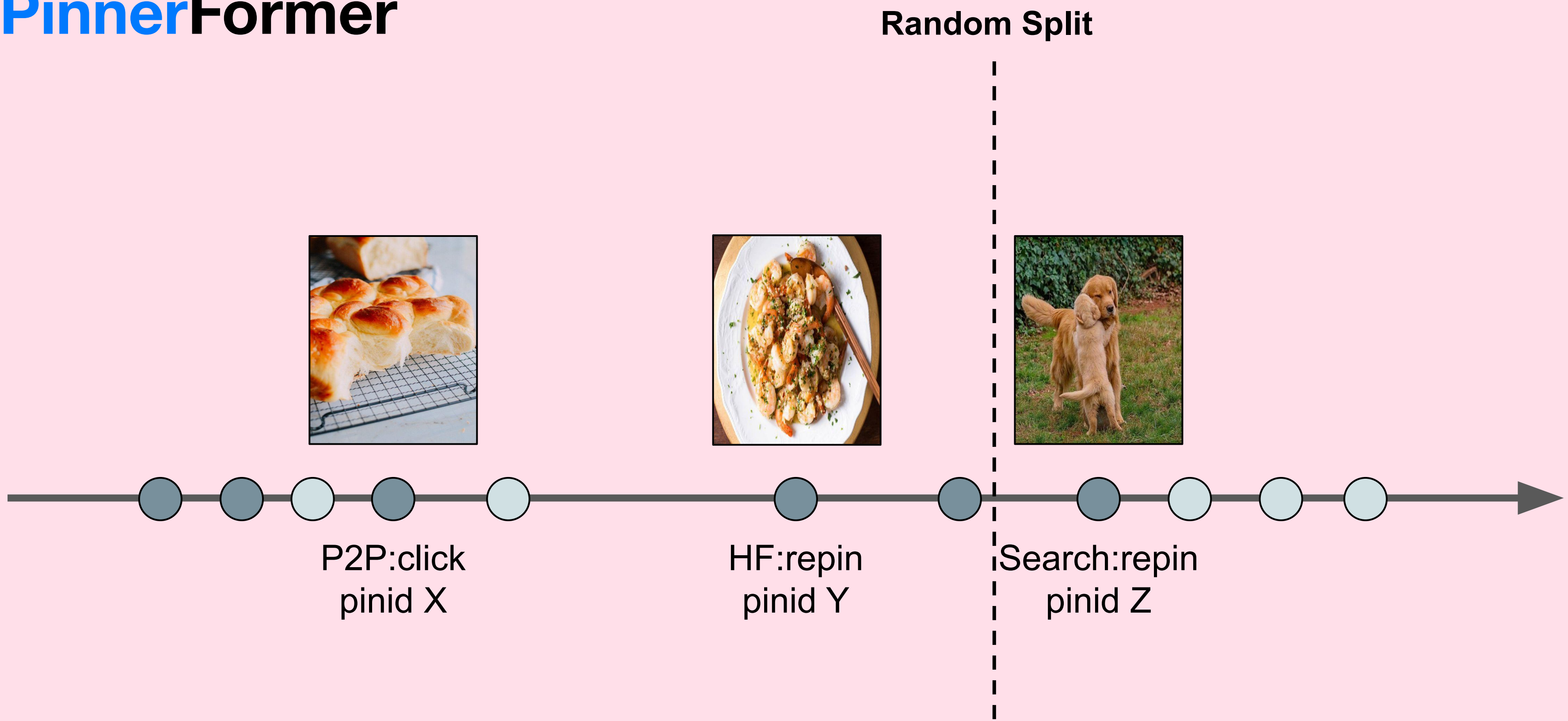
- Multiple embeddings challenging to use
- No parameter sharing across users
- No explicit objective learning

# PinnerFormer

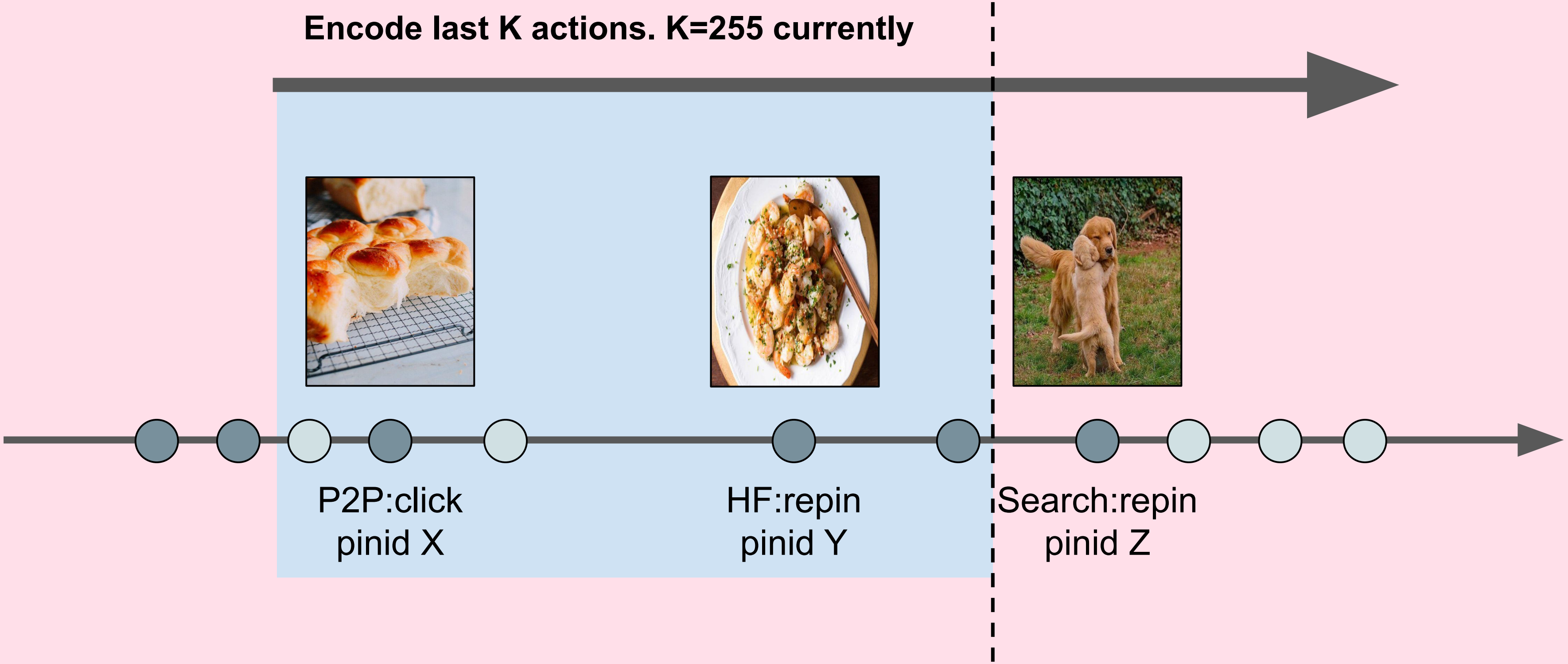



**User sequence activity for past year**

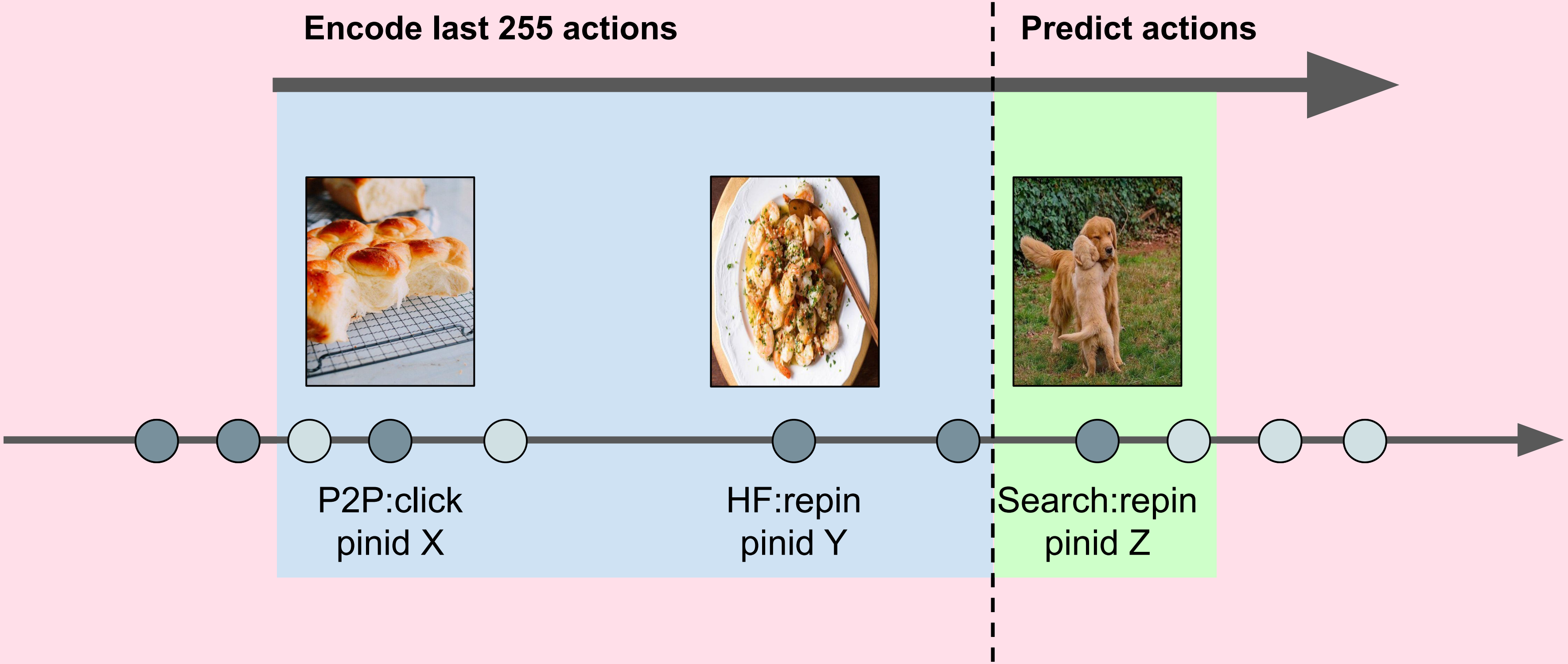
# PinnerFormer



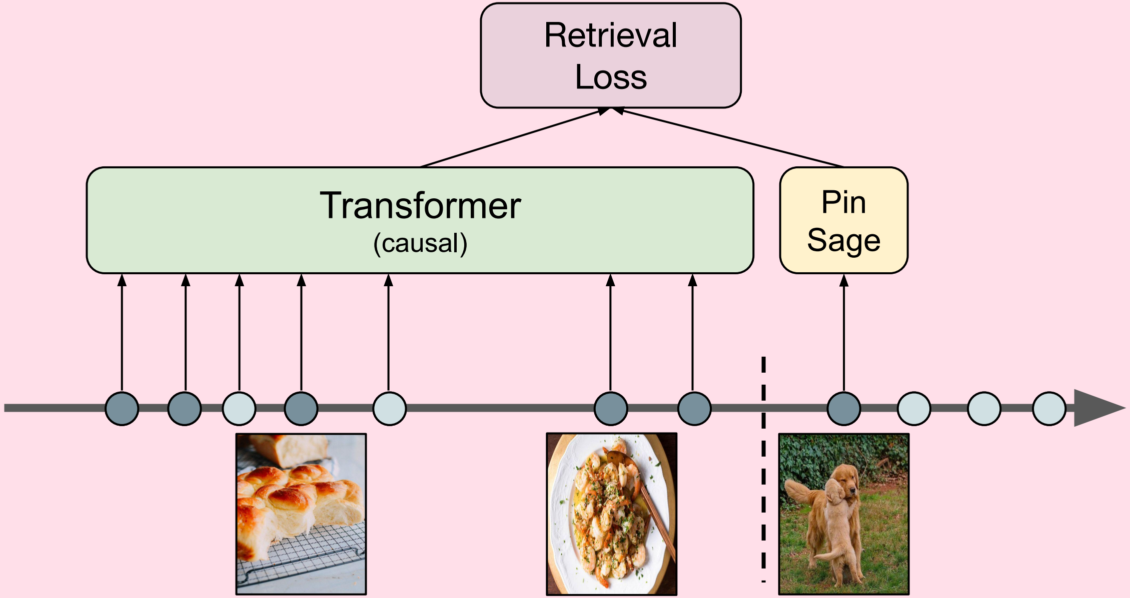
# PinnerFormer



# PinnerFormer

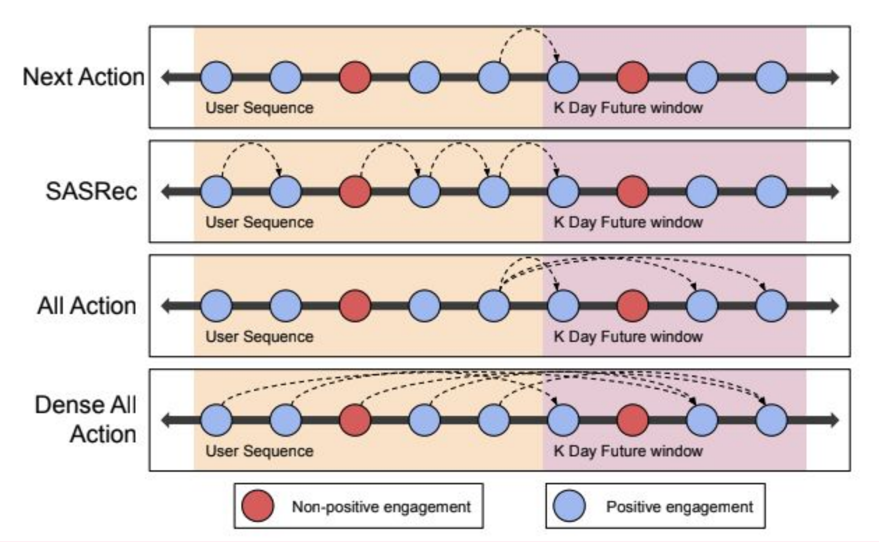


# PinnerFormer Architecture




- **Input**: Last K user activity sequence across all of Pinterest
- **Output**: one user embedding summarizing activity jointly for <u>short</u> and <u>long-term</u> activity prediction.

# PinnerFormer Optimization




| Training Objective | Recall@10 |
|---|---|
| Next Action | 0.186 |
| SASRec (Softmax) | 0.198 |
| All Action (28d) | 0.224 |
| Dense All Action (14d) | 0.223 |
| Dense All Action (28d) | 0.229 |

- **Dense All Action** leads to best performance
  - Optimize for all pos actions within 28d, densely across input seq to Transformer

Pinterest

# PinnerFormer Results

| | R@100 |
|---|---|
| (oracle) PinnerSAGE (5 clusters) | 0.125 |
| (oracle) PinnerSAGE (20 clusters) | 0.205 |
| **PinnerFormer (1 embedding)** | **0.255** |

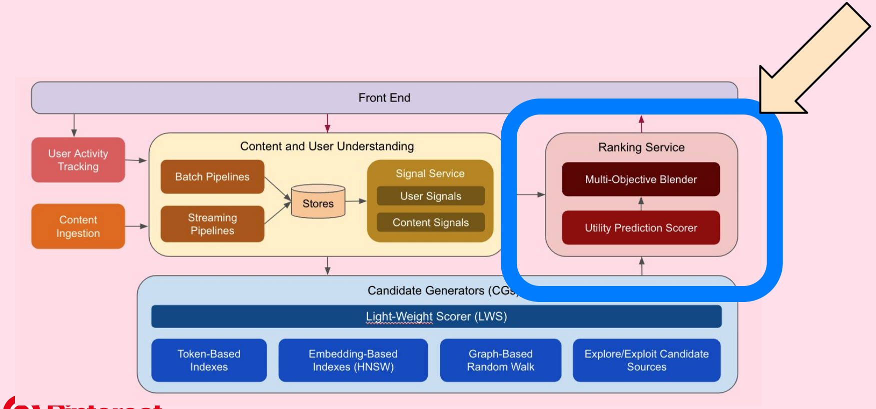
10+ launches

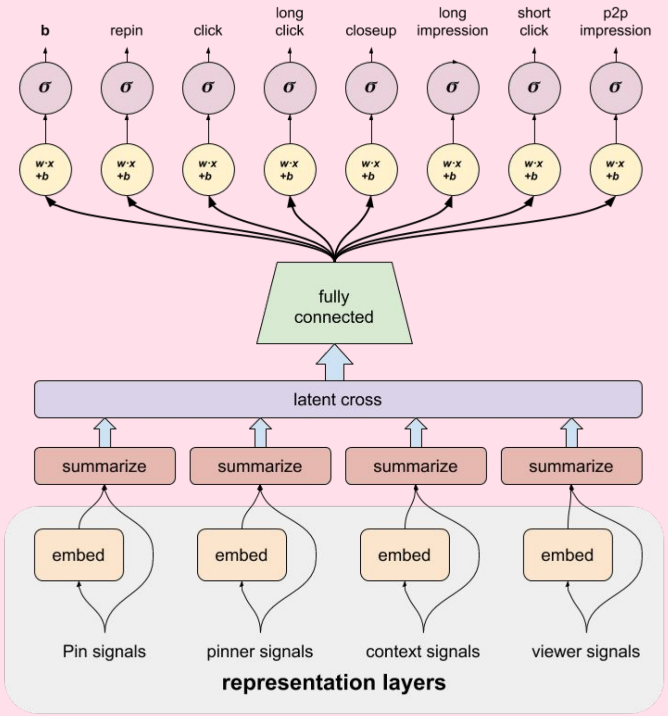
**Site-wide impact**
+1-2% timespent
+3-4% repins
-2.6% hides
+1.8% revenue

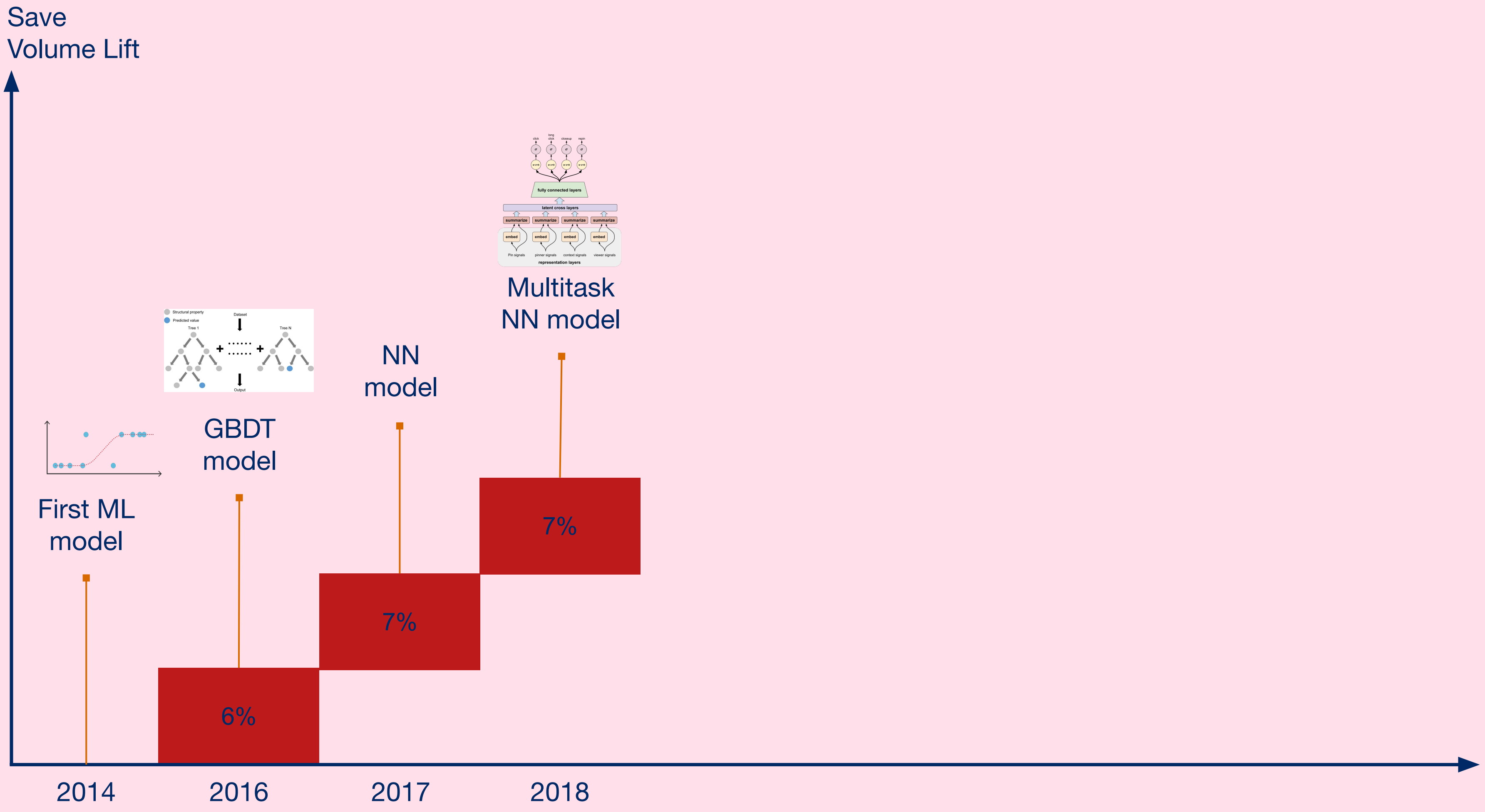
# Personalized Ranking

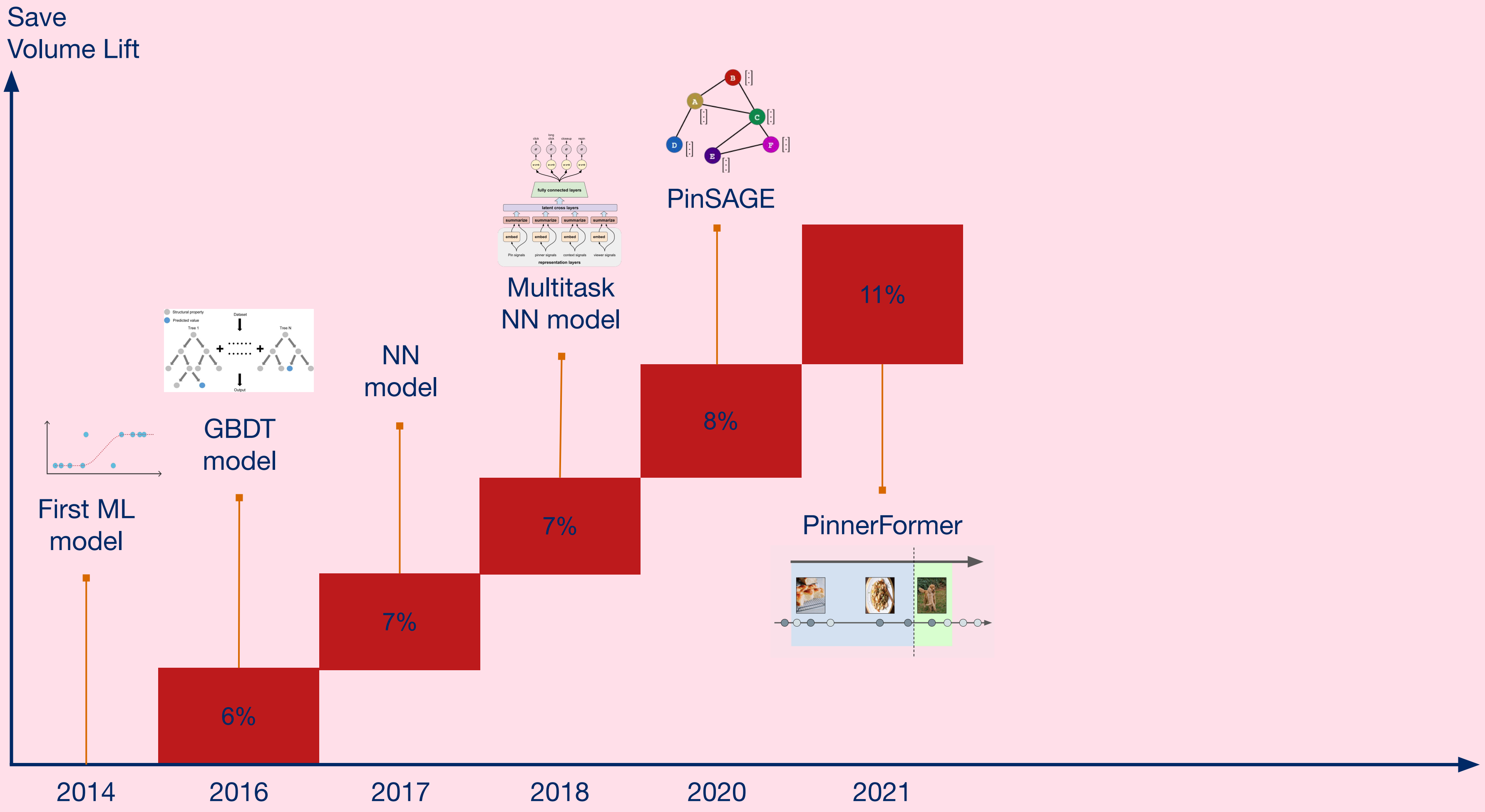
# **Ranking:** User Action Prediction

- Predict a wide variety of user actions for each (user, item) pair through multi-head deep neural network
- Combine 100s of features, served on CPU



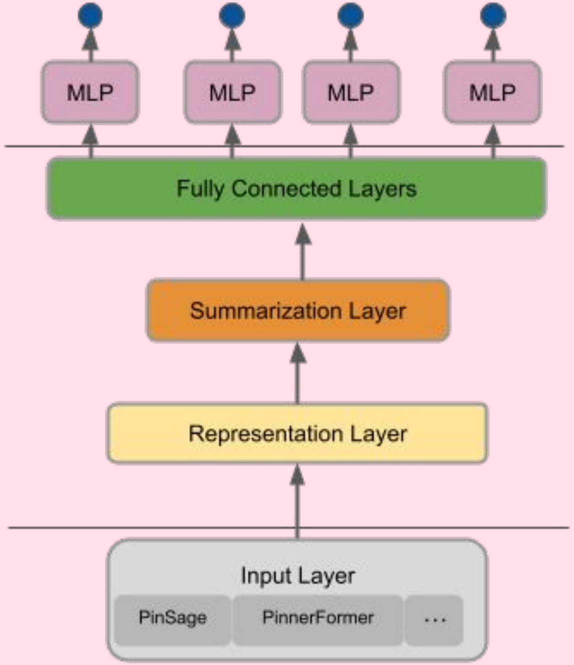

Save
Volume Lift
Multitask
NN model
NN
model
GBDT
model
First ML
model
7%
7%
6%
2014
2016
2017
2018



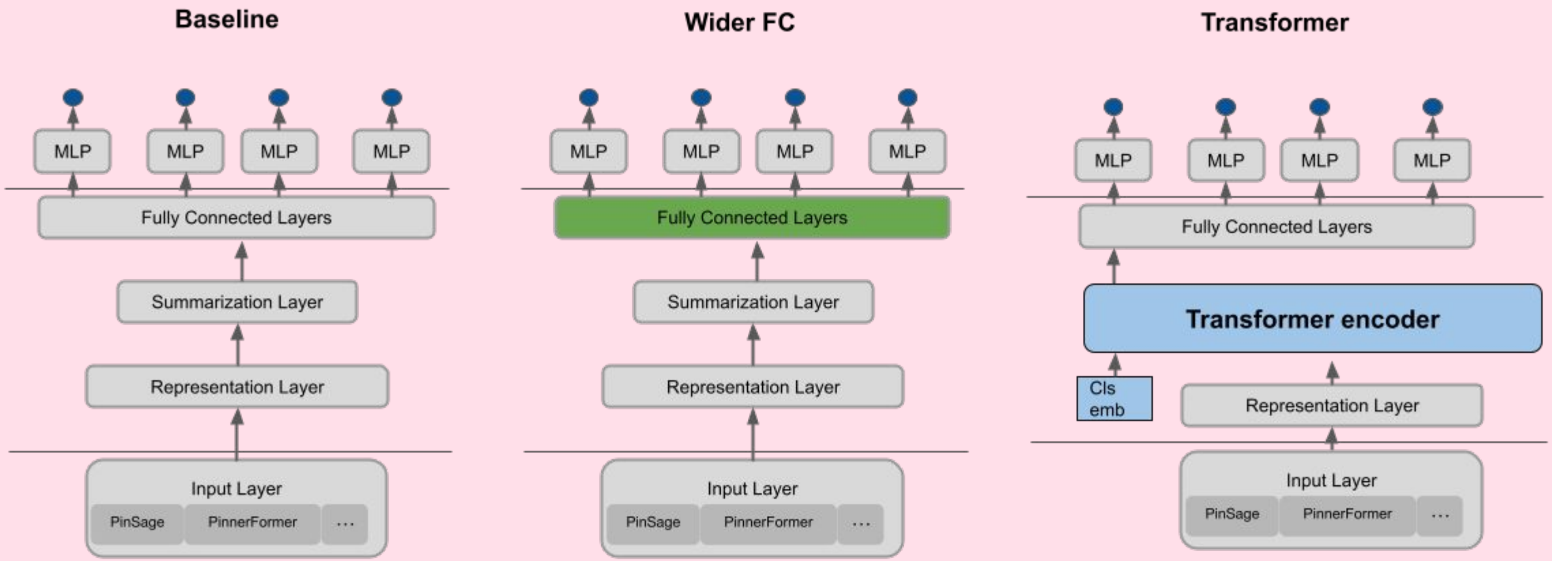

© 2022 Pinterest. All rights reserved.

# **Ranking**: User Action Prediction

- Two Trends for Performance:
  - **Increase parameters, complexity** for model expressivity



# **Ranking**: Scaling It Up

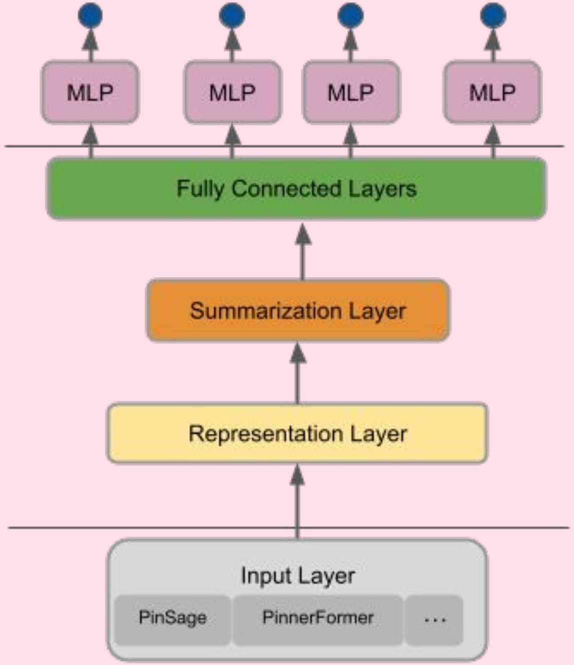



| Model | Expected Saves Gain | Latency Increase |
|---|---:|---:|
| 2x Wider Fully Connected | 5% | +10% |
| + Transformers | 4% | +300% |

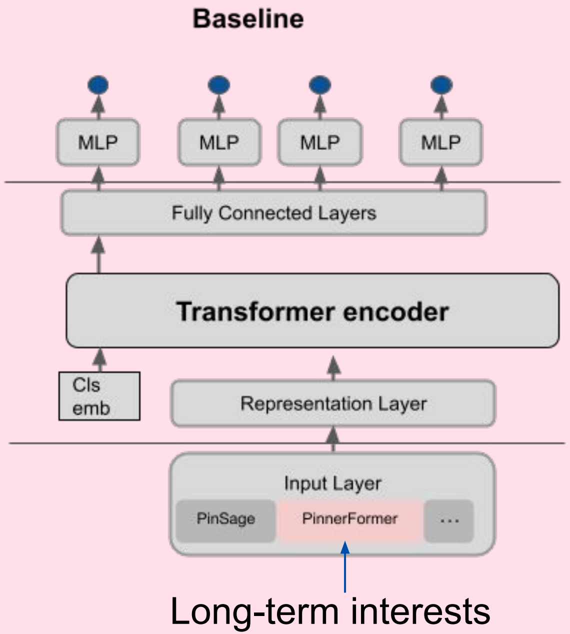
© 2022 Pinterest. All rights reserved.

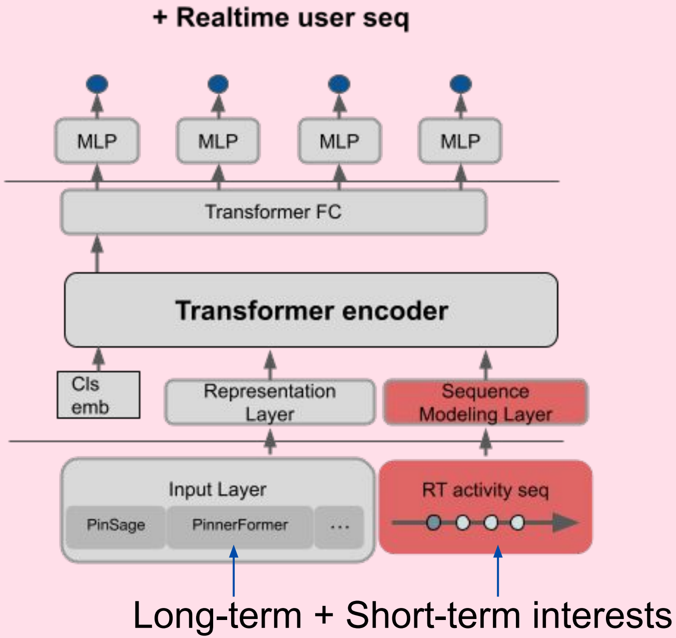
# **Ranking**: User Action Prediction

- Two Trends for Performance:
  - **Increase parameters, complexity** for model expressivity
  - **End-to-end learn** from raw (er) features



# **Ranking**: User journey modeling (E2E Learning)



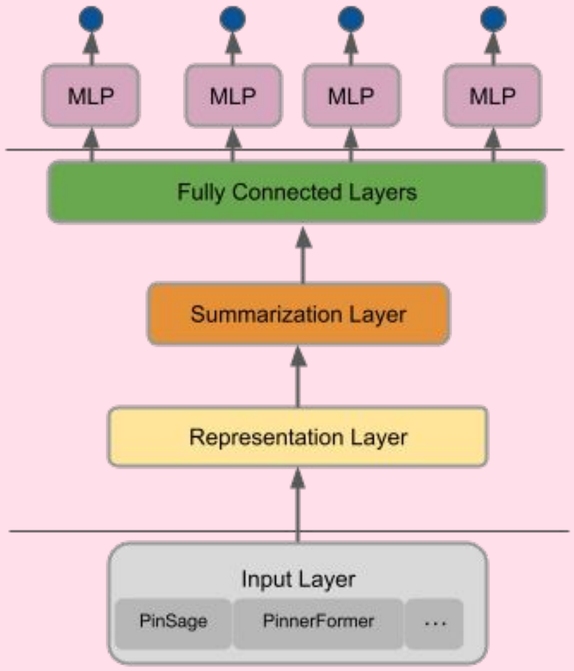

| Model | Expected Saves Gain | Latency Increase |
|---|---|---|
| + RT activity seq (early fuse) | 9% | +100% |

© 2022 Pinterest. All rights reserved.

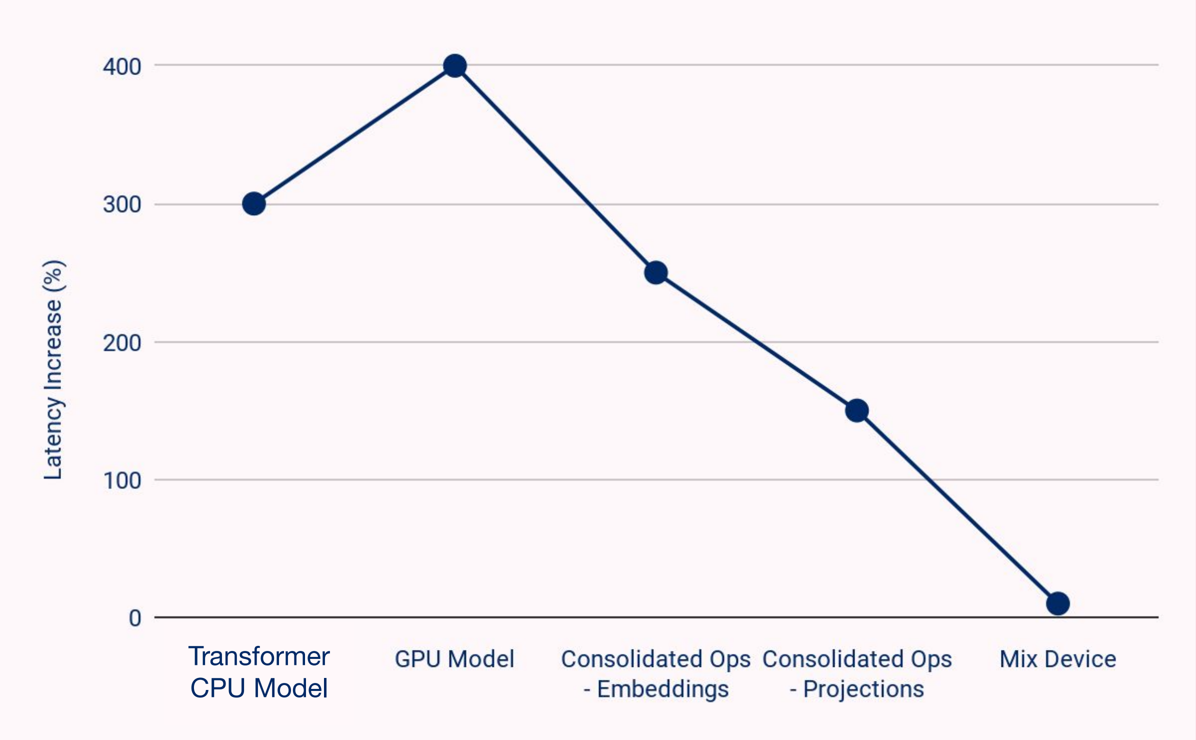
# **Ranking**: User Action Prediction

- Two Trends for Performance:
  - **Increase parameters, complexity** for model expressivity
  - **End-to-end learn** from raw (er) features

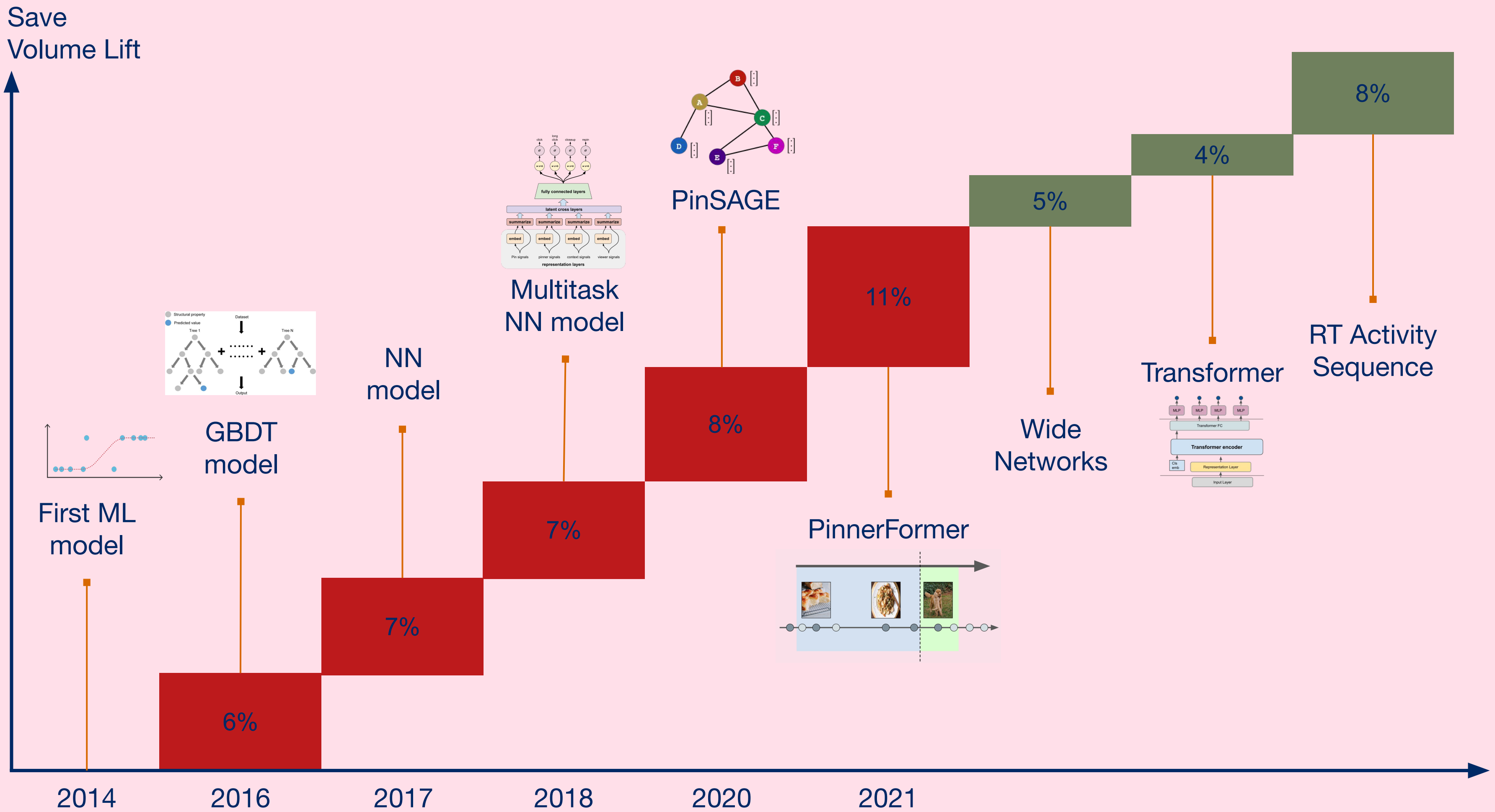
- **Challenge**:
  - **Latency** (~10ms P99)
  - **Throughput** (~10M inferences / sec)
  - **Cost** (+10% latency ~ $400k / year)



# **Ranking**: GPU serving




© 2022 Pinterest. All rights reserved.

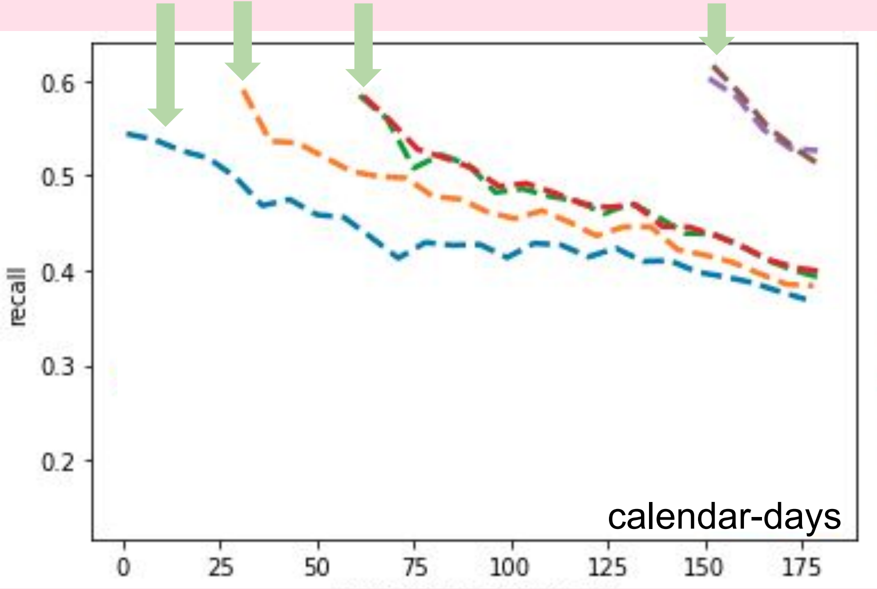



© 2022 Pinterest. All rights reserved.

# Challenges

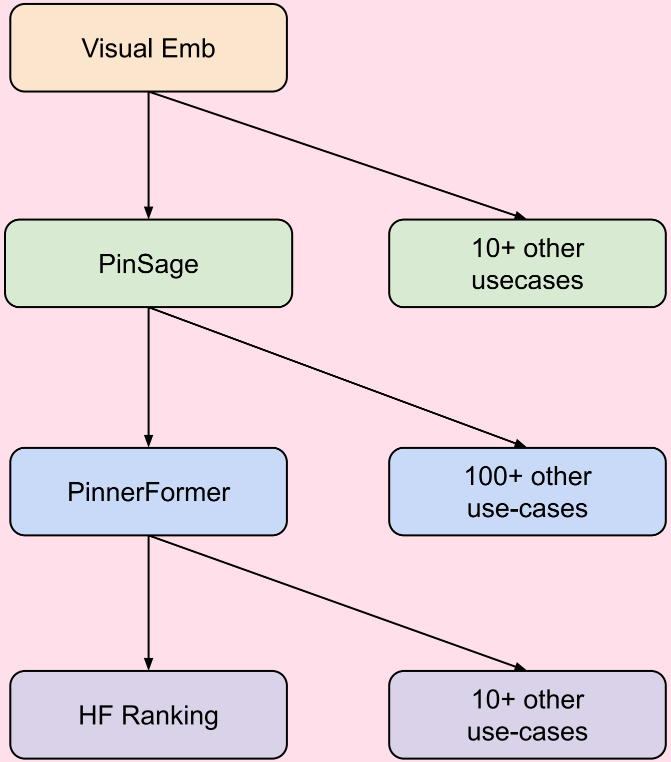
# ML Systems are Dynamic

Future calendar-day eval of same model setup
with different training windows




- Model degrades over time (e.g. Concept Drift)
- **Retraining** recovers performance
- Evaluating a “Good” model is at least 2-dimensional

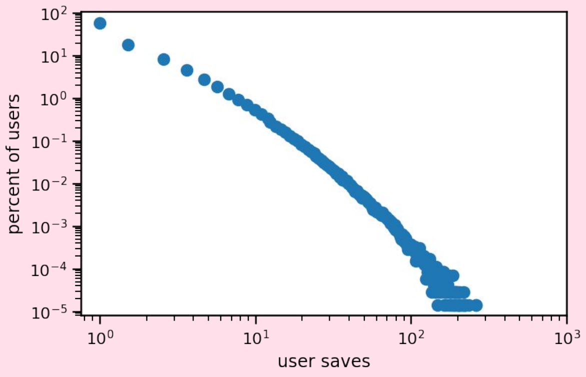
# ML Systems are Dynamic

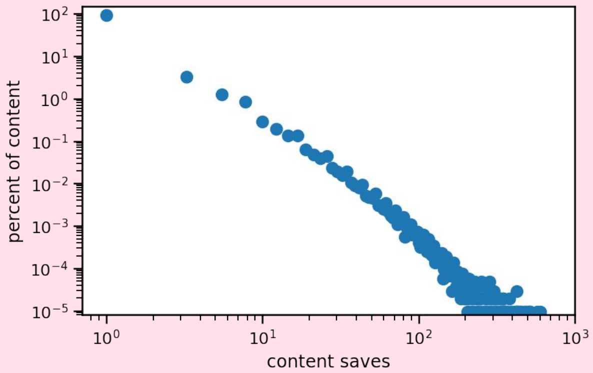



- In practice, long chains of model dependencies
- What is the ABI for ML models?

© 2022 Pinterest. All rights reserved.

# Curse of the Power Law Distribution

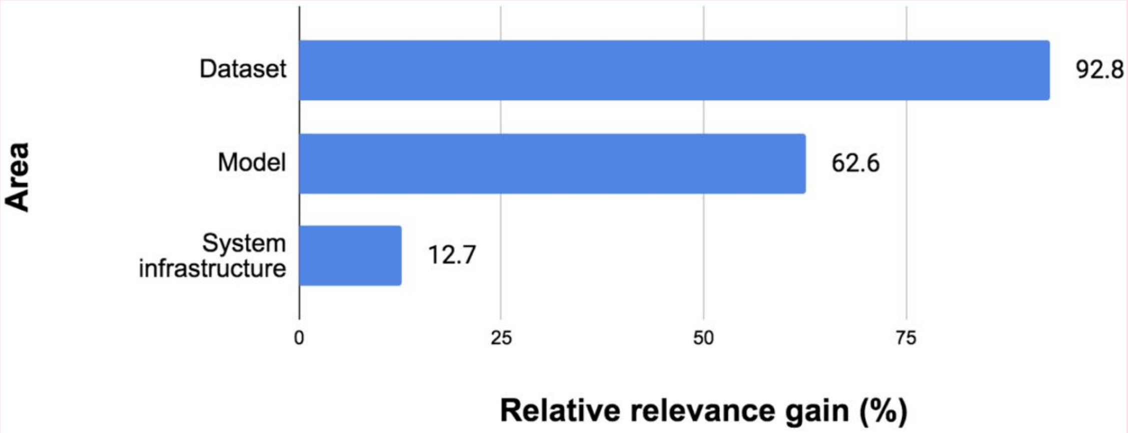



- Power law distributions exist for both users and content
  - Not much feedback for majority of content and users
- Methods
  - Dataset Sampling
  - Explore-Exploit
  - Counterfactual Learning
  - Content/User Embeddings
  - Self Supervision
  - ….

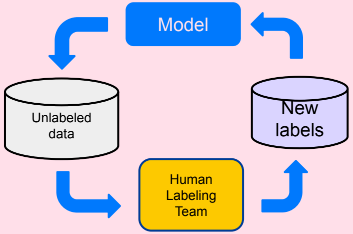
Pinterest

# Dataset is an Important Lever




Shop The Look: Building a Large Scale Visual Shopping System at Pinterest (KDD 2020)

- Research: model-centric Industry: data-centric
- Trends: Software 2.0, Data-centric ML
- How can we build systems and algorithms to iterate on datasets faster?




© 2022 Pinterest. All rights reserved.

# User Journey Optimization

To maximize long-term “reward”

| Aspiration | Inspiration | Consideration | Action |
|---|---|---|---|
| Off platform: I want to remodel my kitchen | Search for “kitchen remodel” in Pinterest | • Save lifestyle images to boards<br>• Explore products in shopping | Purchase products for my kitchen to complete my remodel |

- **User problem**: Want to find inspiration and complete project (e.g. summer vacation planning, cooking dinner). If Pinterest does well, plan more of life on Pinterest.
- **Today:** Utility function of immediate actions (e.g. save, click, closeup, hides).
  - Manual “gradient descent” (analysis, implement, ab experiment, feedback)

© 2022 Pinterest. All rights reserved.

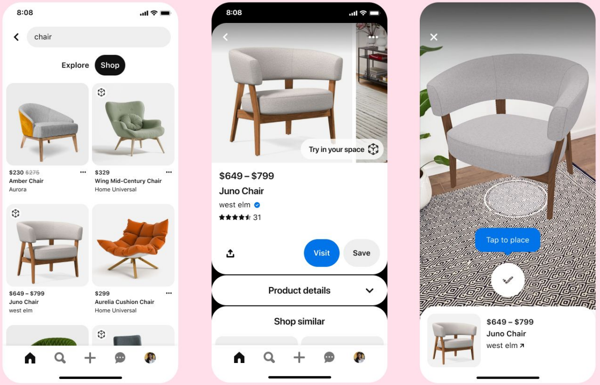
# User Journey Optimization

To maximize long-term “reward”

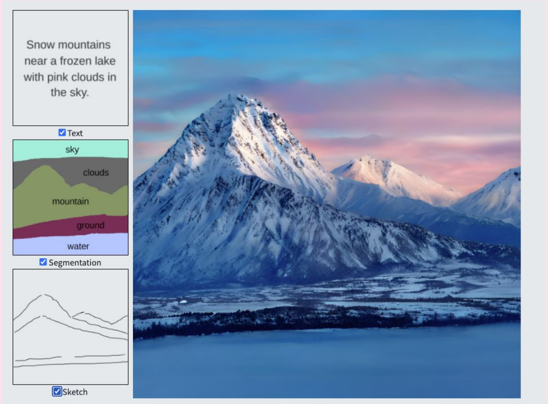
| Aspiration | Inspiration | Consideration | Action |
| --- | --- | --- | --- |
| Off platform: I want to remodel my kitchen | Search for “kitchen remodel” in Pinterest | • Save lifestyle images to boards<br>• Explore products in shopping | Purchase products for my kitchen to complete my remodel |

- **User problem**: Want to find inspiration and complete project (e.g. summer vacation planning, cooking dinner). If Pinterest does well, plan more of life on Pinterest.
- **Today:** Utility function of immediate actions (e.g. save, click, closeup, hides).
  - <u>Manual</u> “gradient descent” (analysis, implement, ab experiment, feedback)
- **Challenge**: Enable ML systems to optimize directly for “pinner satisfaction”
  - Causal inference for actions -> long-term satisfaction?
  - Off-policy Reinforcement Learning?
    - Reward function incredibly complex from multi-objective optimization

# Next Gen Inspirational AI Products




HD Virtual Try On AR




[Multimodal Conditional Image Synthesis with Product-of-Experts GANs](#)
2021

Fashion Virtual Try On AR

A bowl of soup that is a portal to another dimension as digital art

https://openai.com/dall-e-2/

© 2022 Pinterest. All rights reserved.

# A lot more going on…

- **Representation Learning** for videos, products, creators, search queries, notifications
- **Web Mining** through GNNs to extract attributes (e.g. recipe for food pins) from websites to create rich content at scale
- **Inspirational Knowledge Graph** to enable a vocabulary to communicate between ML and users to assist their journey
- **Learned Retrieval** to holistically learn candidate generation for recommendations and search
- **Notification Uplift Modeling** to learn the optimal intervention policy for share inspiration to Pinners outside of Pinterest

© 2022 Pinterest. All rights reserved.

# Takeaways

- **Pinterest** is a unique curated dataset of how people describe and organize things
- **ML** is leveraged throughout our inspiration funnel to enable us to bring *everyone* the *inspiration* to create a life they love
- **Deep Learning methods** (Transformers, GNN, Sequence) leading the way for performance
- **Scalability** of systems and ML algorithms are baked deeply into our culture and a continued trend for improvement
- A lot of technical **challenges** exist. Not even close to a solved problem

© 2022 Pinterest. All rights reserved.

# Thank you!

andrew@