# Amazon x UCLA MSBA Hackathon

## Team #4

(Ashwin Bharathwaj, Jungjin(JJ) Park, Karthik Maddipoti, Tharun Sriram, Yash Bhatia)

# Executive Summary

## What is the problem?

**Improve customer shopping experience**

- Identify a robust way to find right products among a wide selection at Amazon

## How can we fix this?

**Content-based Recommendation**

- Leverage product characteristics and description to identify similar products
- Build MVP that's scalable - word2vec embedding & content-based recommender

**Corner Case & Model Improvement**

- Preferences reflected in recommendations (i.e., brand, price)
- Experiment plans to improve the model

## Why does this benefit us?

**Amazon's Benefits**

- Capture opportunities for cross-selling and up-selling
- More satisfying purchases by finding better products easily, leading to low returns
- Stronger brand awareness with seamless shopping experience
- Better understanding of customers behaviors and tastes in products

# Modeling Approach

| | 1. Data Cleaning | 2. Feature Engineering | 3. Unsupervised Modeling |
|---|---|---|---|
| GOALS | Clean raw data to deal with data quality issues | Transform text-based data to numerical vectors for model ingestion | Train a recommendation model |
| SCOPE | Completed below data quality checks for selected set of important columns<br>• Missing values<br>• Outliers<br>• Special characters<br>• Duplicate items | Explore pre-trained word embedding methods and adjust feature weights<br><br>Validate performance on hand-picked samples | Use content-based filtering recommender to identify similar products<br><br>Validate model's findings based on hand-picked sample cases |
| FUTURE IMPROVEMENTS | Explore external data sources to enrich item-level data set<br>• Amazon Berkeley Objects (ABO) Dataset | Reduce feature space by re-training the text encoder on product data | Use an ensemble of models to make robust recommendation<br>• Hybrid recommendation models |

# Modeling Approach Details (1/3) - Data Cleaning

## Steps

1. **Excluded all non-alphanumeric characters** from all text columns
2. **Duplicates** – Description with n – 2 common words removed
3. **Outliers** - Imputed with 99th percentile value from same product category and type
4. **Missing values** - Imputed with average values from same product category and type

## Examples

*e.g. ′ ˆ •Ã § ˆ„ □µ ‘‹ £*

| Item Id | Item Name | Marketplace |
|---|---|---|
| B009SC4W3W | Organic Carrots, 2lb | AmazonFresh |
| B009SC4W3W | Organic Carrots, 2lb | AmazonGo |

| Item Name | Outlier Feature | Original Value | Imputed Value |
|---|---|---|---|
| Solimo Capri Fabric Recliner | Item Weight | 80000 | 176 |

| Item Name | Feature | Original Value | Imputed Value |
|---|---|---|---|
| Organic Light Cream 1 PT | Item Height | NULL | 6.2 |

# Modeling Approach Details (2/3) – Feature Engineering

To ingest data for training and prediction, textual data was converted to numerical features

## Steps

1. **One-hot encoding categorical features**
   a. Brand
   b. Product category
   c. Product Type
   d. Marketplace

2. **Word embedding methods** for product identifying texts
   a. TF-IDF vectorization
   b. Weighted Word2Vec – Weights corresponding to Idf of the word
   c. BERT inspired Sentence transformers

## Examples

Human-Readable → Machine-Readable

| Pet |
|---|
| Cat |
| Dog |
| Turtle |
| Fish |
| Cat |

| Cat | Dog | Turtle | Fish |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 |

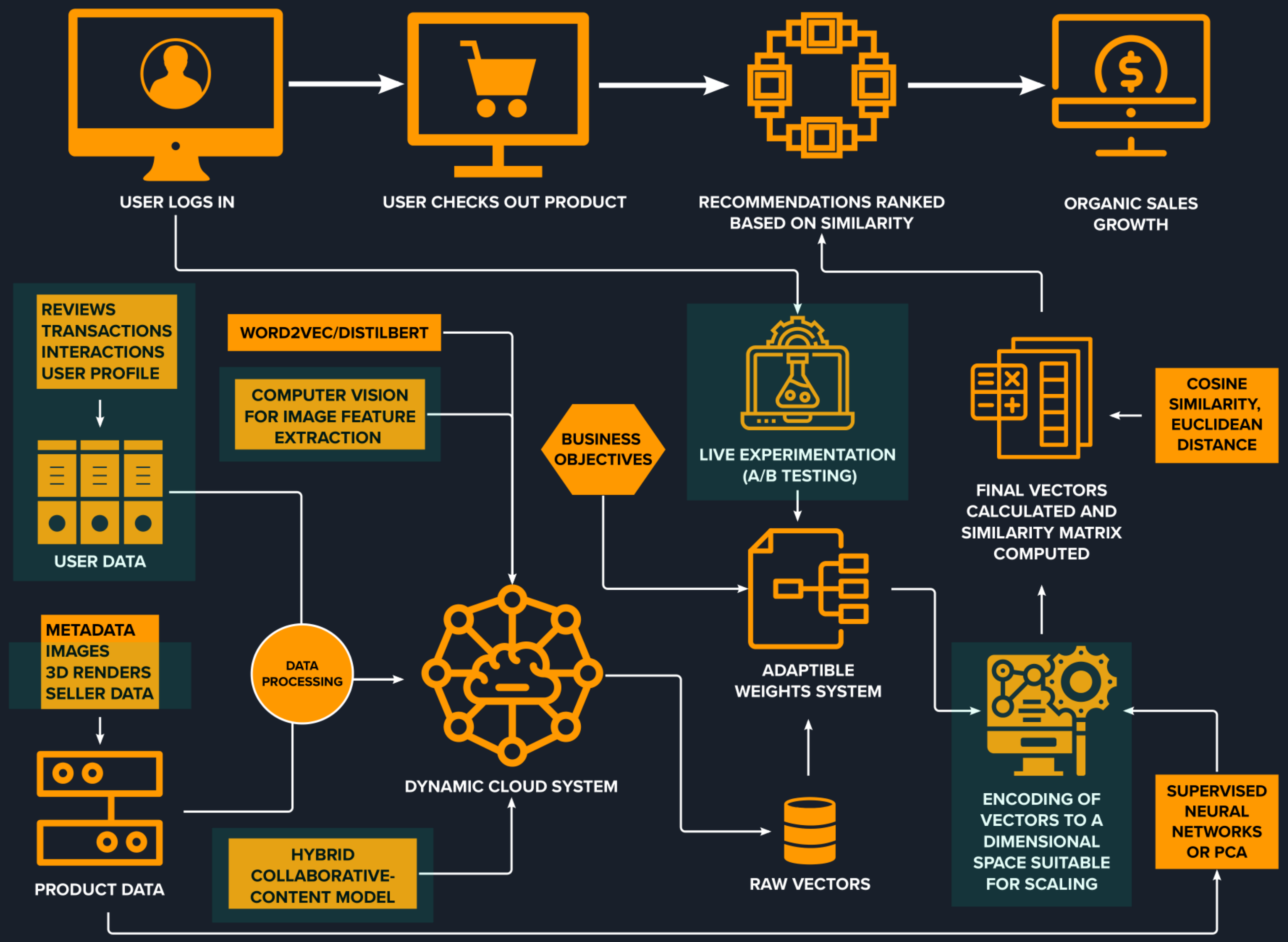
Word-embedding

$$\begin{pmatrix} the \\ cat \\ sat \\ on \\ the \\ mat \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

✓ *Weighted OHE vectors were concatenated to the text embeddings to create final raw vectors*

Used in the model

# Modeling Approach Details (3/3) - Methodology




*To be incorporated as we scale the system up*

# Baseline Recommendation Results

## *Item searched*

Amazon Basics Wireless Bluetooth Dual 3W Speaker with Built-in Microphone - Black

Amazon Brand – Rivet Industrial Coffee Table

Happy Belly – 2% Fat Milk One Gallon

## *Recommended Products*

Amazon Basics Portable Wireless Bluetooth Speaker

Amazon Basics Wireless Bluetooth Dual 3W Speaker with Built-in Microphone - Blue

Bose Sound-Link Wireless Bluetooth speaker with Mic

Amazon Brand – Stone & Beam Wood Coffee Table

Amazon Brand – Stone & Beam Industrial Coffee Table

Amazon Brand – Rivet Industrial Solid Wood Coffee Table

Happy Belly – 2% Fat Milk Half Gallon

Happy Belly – 1% Fat Milk One Gallon

Happy Belly – 1% Fat Chocolate Milk

# BUT WAIT!
# What about
# **Customer Obsession?**

Happy Belly – 2% Fat Milk One Gallon

Happy Belly – 2% Fat Milk Half Gallon

Happy Belly – 1% Fat Milk One Gallon

Happy Belly – 1% Fat Chocolate Milk

# Personalized Recommendations Ideas

Few examples of leveraging customer/product attributes for ranking/filtering/substitution

| | Price sensitive | Quality sensitive | Dietary preference | Brand loyalty |
|---|---|---|---|---|
| **Use cases** | • Recommendation based on customer's individual price preference | • Customers who strongly prefer high quality items | • Customized to personal dietary preference | • Positive/Negative propensity toward specific brands |
| **How to catch signals** | • Filtering on price range<br>• Tendency to buy products cheaper than median or 20% percentile | • Returns on low rating items<br>• High avg of ratings of items purchased<br>(cf. having considerable volumes) | • Product classification based on description<br>• Frequent purchases/ views/add to carts of certain grocery type | • Brand classification<br>• Including brand name when querying<br>• Frequent purchases/views/ add to carts of brand items |

*Item searched*


Whole Foods Market, Organic Mung Bean Gluten Free Fusilli, 8 oz

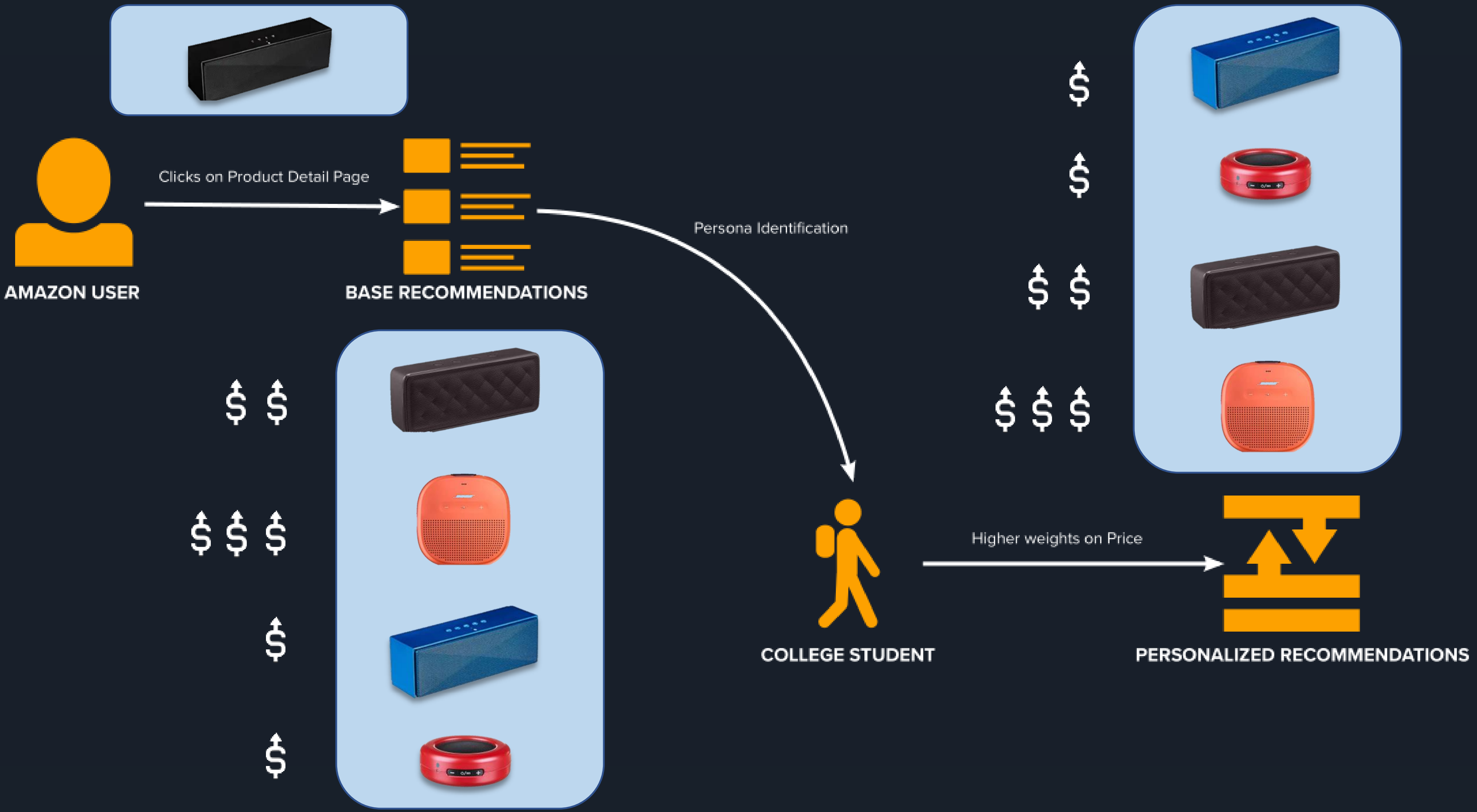
*Recommended products*


365 by Whole Foods Market, Pasta Corn & Rice Penne Rigate Gluten Free, 12 Ounce


Whole Foods Market, Organic Red Lentil Gluten Free Spaghetti, 8 oz

# Personalized Recommendations Journey





# Model Improvement Using Experimentation

## Improve personalization

**Feature Engineering**

- Word embedding: Word2Vec/ BERT/ TF-IDF
- Distance: Cosine similarity/ Euclidean distance
- Include additional data sources (ABO data, etc.)

## Improve UI/UX

**# of recommendations & ranking**

- Optimal # of products to display based on device type and screen orientation
- Displaying products to users based on what matters to them (price, brand, etc.)

**Recommendation widgets**

- Dynamic recommendations clusters (complimentary products, products from top sellers, etc.)

## Primary KPIs

1. Conversion rate
2. Profit per visit
3. AOV
4. Avg. CTR on recommended products

## Guardrail KPIs

Website-specific:

1. Bounce rate
2. Page load speed

Email-specific:

1. Unsubscribe rate
2. Open rate

## Model Performance Metrics

1. Hit rate
2. Mean reciprocal ranking

*…additionally, feedback can also be collected explicitly via built-in micro feedback system to get real-time actionable data*

# Thank you!