BRAND UNIFICATION FOR CRITEO

IMPROVING AD TARGETING, ANALYTICS, AND USER EXPERIENCES

ABOUT CRITEO

3,500+
Employees

100+
Countries

17,000+
Clients

1,100
R&D & Product Employees

AI

19
Years of Commerce-Focused AI

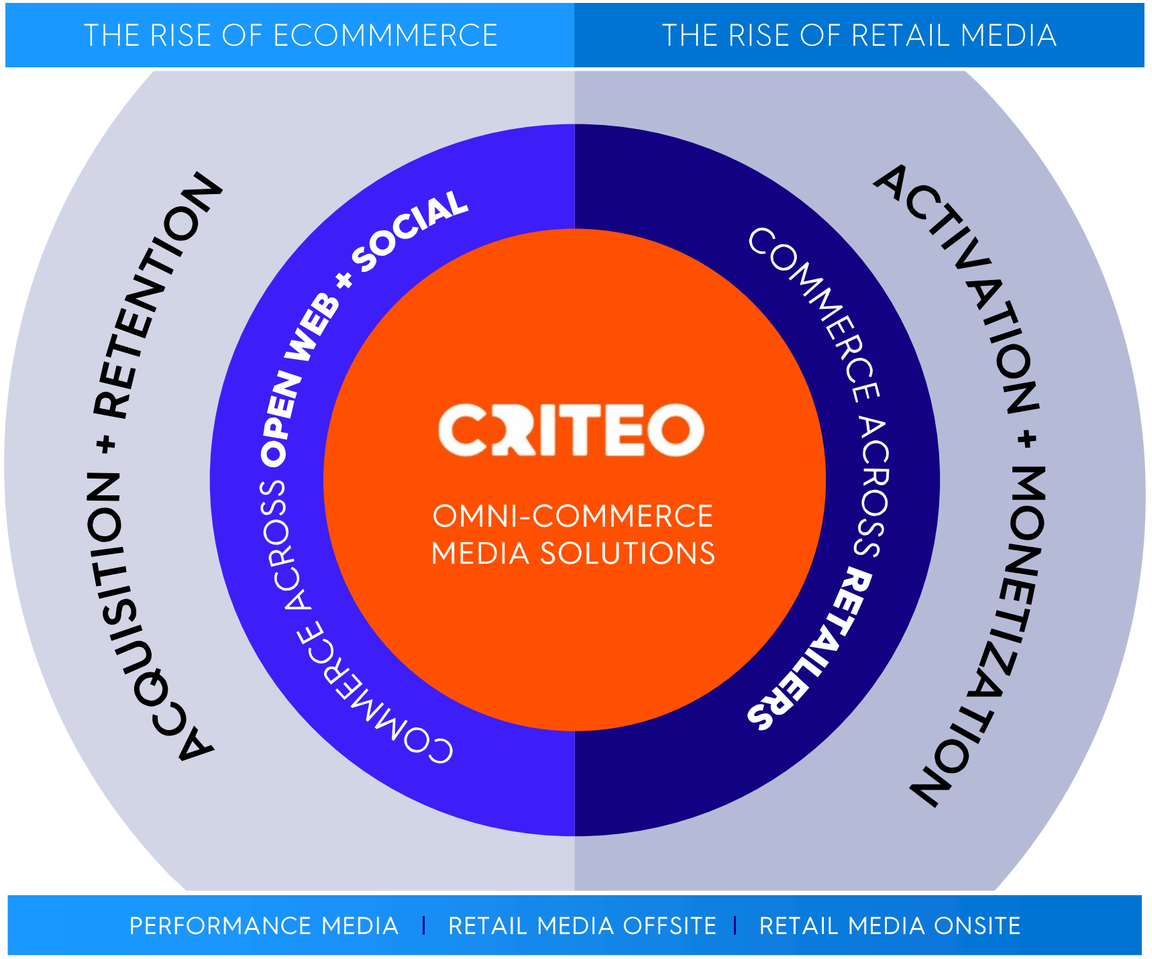
$4.3B
Activated Media Spend[1] LTM

$1 Trillion
Digital Transactions Observed

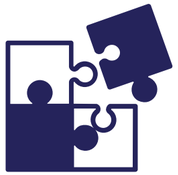
5 Billion
Ads Served a Day

700 Million
Daily Active Users

4 Billion
Product SKUs Observed



Case Study Objectives



Business Problem Statement

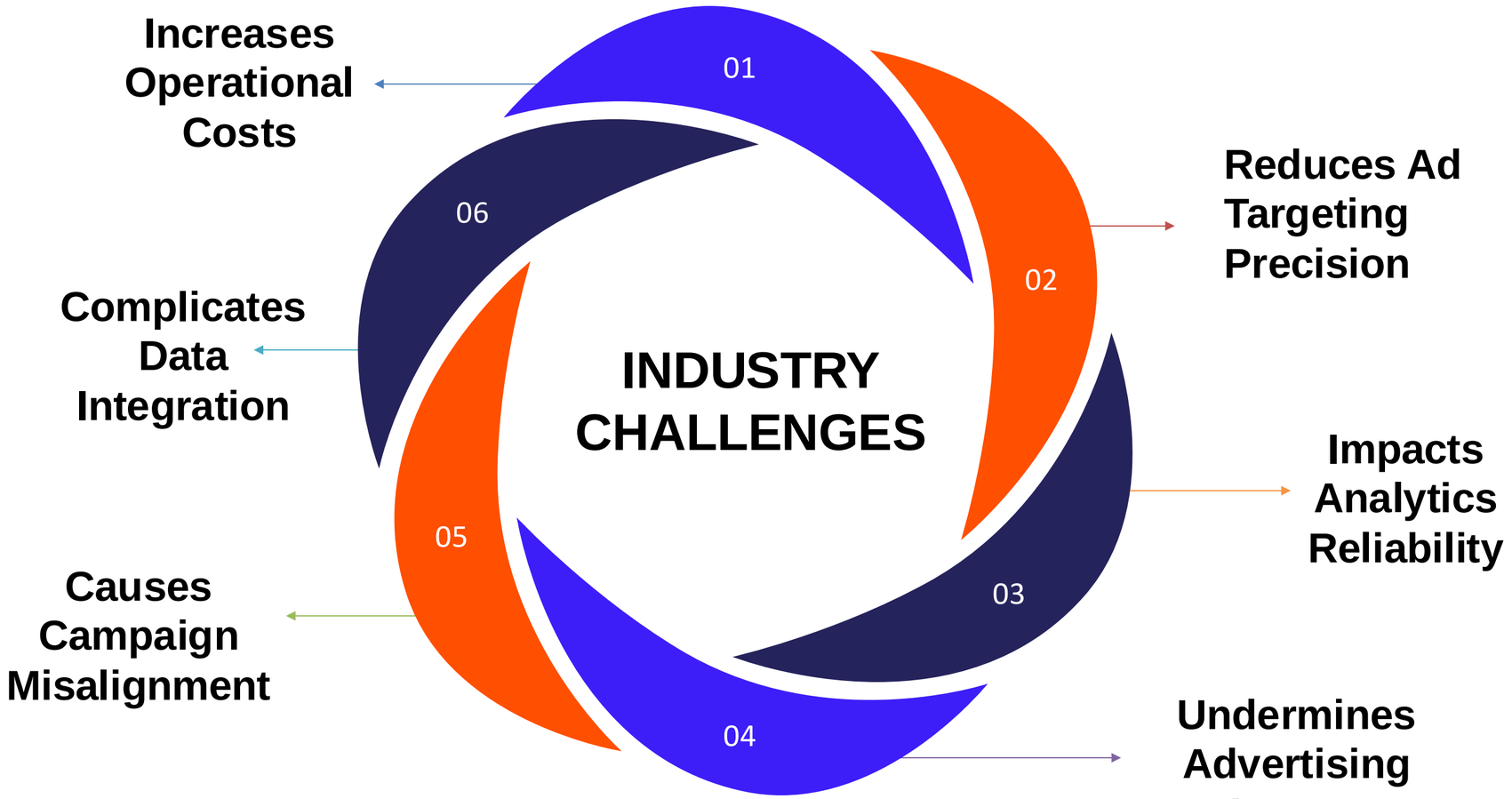
Criteo’s mission is to empower marketers with trusted and impactful advertising. However, inconsistent brand data creates inefficiencies in ad targeting and analytics, undermining this mission.

Our Solution

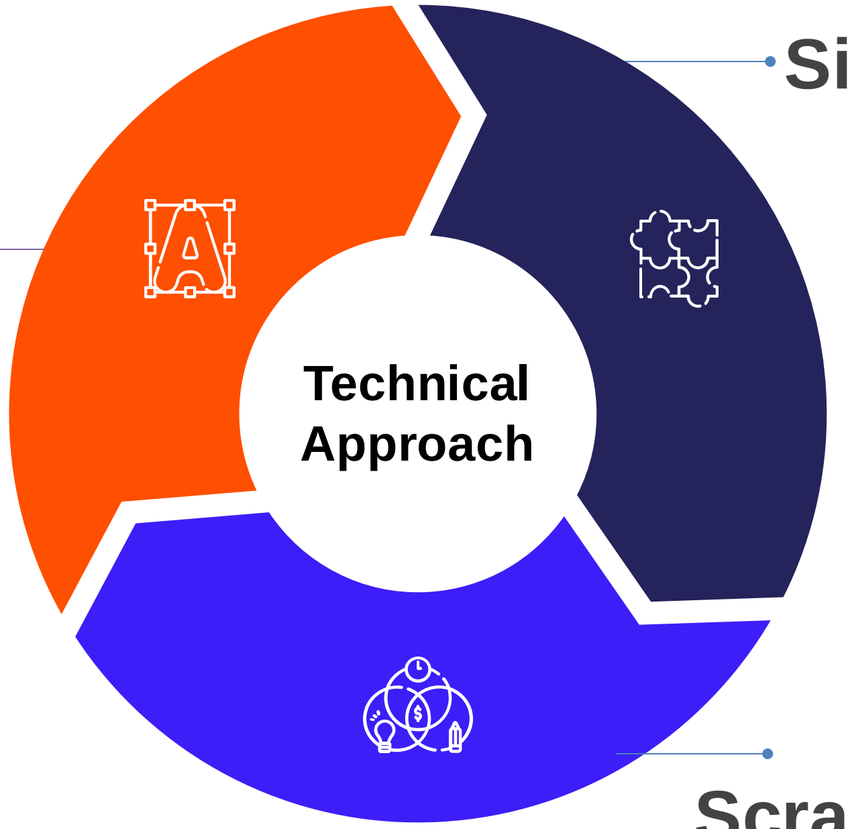
Dynamic Solutions for Brand Unification:

- Clustering Algorithms
- Fuzzy Matching
- External APIs

THE PROBLEM: INCONSISTENT BRAND DATA



THE SOLUTION: UNIFYING BRAND NAMES



Cleaning

Similarity

Scrapping and API

Cleaning

Data Preparation	• Loaded and normalized brand names. • Translated non-English names and replaced missing data.
Normalization and Translation	• Converted to lowercase, removed accents and special characters. • Korean names translated to English
Categorization	• Applied Google category IDs and taxonomy codes • Mapped brands to structured subcategories (Levels 1-7)
Metadata and Readiness	• Retained essential metadata for analysis • Dataset standardized and ready for modelling

Cleaning

Data Preparation	• Loaded and normalized brand names. • Translated non-English names and replaced missing data.
Normalization and Translation	• Converted to lowercase, removed accents and special characters. • Korean names translated to English
Categorization	• Applied Google category IDs and taxonomy codes • Mapped brands to structured subcategories (Levels 1-7)
Metadata and Readiness	• Retained essential metadata for analysis • Dataset standardized and ready for modelling

Cleaning

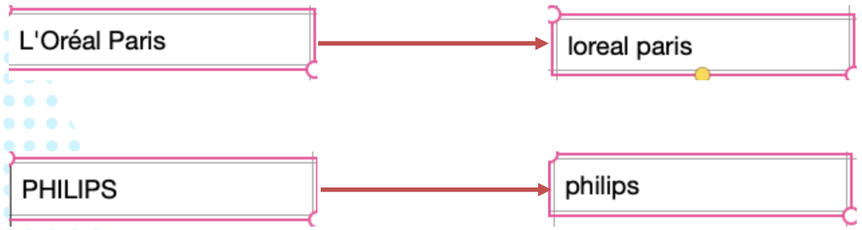
Data Preparation	• Loaded and normalized brand names. • Translated non-English names and replaced missing data.
Normalization and Translation	• Converted to lowercase, removed accents and special characters. • Korean names translated to English
Categorization	• Applied Google category IDs and taxonomy codes • Mapped brands to structured subcategories (Levels 1-7)
Metadata and Readiness	• Retained essential metadata for analysis • Dataset standardized and ready for modelling

Cleaning

Data Preparation	• Loaded and normalized brand names. • Translated non-English names and replaced missing data.
Normalization and Translation	• Converted to lowercase, removed accents and special characters. • Korean names translated to English
Categorization	• Applied Google category IDs and taxonomy codes • Mapped brands to structured subcategories (Levels 1-7)
Metadata and Readiness	• Retained essential metadata for analysis • Dataset standardized and ready for modelling

CLEANED DATA SAMPLE

968	Smoby	de	1251	0.3340042172	smoby	Toys & Games	Outdoor Play Equipment	Playhouses
969	Spam	en	4628	0.3196438862	spam	Food, Beverages & Tobacco	Food Items	Meat, Seafood & Eggs
970	TOMBOW	en	6065	0.3834294103	tombow	Office Supplies	Office Instruments	Writing & Drawing Instruments
971	TORY BURCH	en	3032	0.4782767253	tory burch	Apparel & Accessories	Handbags, Wallets & Cases	Handbags
972	Toteme	en	2271	0.4043098519	toteme	Apparel & Accessories	Clothing	Dresses
973	VEVOR	en	2549	0.2807043642	vevor	Home & Garden	Household Appliances	Laundry Appliances
974	VG Sports	en	3618	0.3926130958	vg sports	Sporting Goods	Outdoor Recreation	Cycling
975	Veilance	en	212	0.3106533353	veilance	Apparel & Accessories	Clothing	Shirts & Tops
976	Xiaomi	en	751	0.4188853113	xiaomi	Home & Garden	Kitchen & Dining	Kitchen Appliances
977	Yeele	en	2334	0.3446003723	yeele	Home & Garden	Decor	Wallpaper
978	gagaopt	en	2271	0.3049214426	gagaopt	Apparel & Accessories	Clothing	Dresses
979	jiadou	en	2334	0.5532031674	jiadou	Home & Garden	Decor	Wallpaper
980	키친아트	ko	751	0.4991738714	kitchen art	Home & Garden	Kitchen & Dining	Kitchen Appliances
981		de	3517	0.4765863021	-1	Hardware	Tools	Saws
982		de	6544	0.4871742719	-1	Electronics	GPS Tracking Devices	
983		en	1268	0.6956142346	-1	Toys & Games	Toys	Beach & Sand Toys
984		en	4295	0.7217419312	-1	Home & Garden	Decor	Piggy Banks & Money Jars



Similarity Calculation

```
dnm and dm: 80%
alanui and alanchi: 77%
rhone and hohner: 73%
rhone and zerone: 73%
naiveroo and native: 71%
burlington and burton: 75%
fenruien and fenrir: 71%
zeks and zest: 75%
marvis and tamaris: 77%
marvis and aries: 73%
marvis and marvo: 73%
advil and fanvil: 73%
garfield and rosefield: 71%
botter and trotters: 71%
botter and wotte: 73%
leviton and lito: 73%
boyy and oyy: 86%
aster and paste: 80%
```

```
Brand pairs with Weighted score above threshold:
nike and nike swim: 66%
rick owens and rick owens kids: 75%
calvin klein and calvin klein jeans: 75%
calvin klein and calvin klein performance: 75%
gucci and gucci eyewear: 66%
gucci and gucci kids: 66%
offwhite and offwhite kids: 66%
goya and stine goya: 66%
samsung and samsung electronics: 66%
disney and disney princess: 66%
versace and versace kids: 66%
versace and versace jeans: 66%
armani jeans and armani: 66%
valentino and red valentino: 66%
balenciaga and balenciaga kids: 66%
dkny and dkny sport: 66%
ralph lauren collection and ralph lauren: 75%
ralph lauren collection and lauren ralph lauren: 75%
karl lagerfeld and karl lagerfeld kids: 75%
marc jacobs and marc jacobs kids: 75%
stella mccartney kids and stella mccartney: 75%
michael kors and michael kors collection: 75%
michael kors and michael michael kors: 100%
missoni and missoni home: 66%
missoni and m missoni: 66%
missoni and missoni mare: 66%
agua bendita and agua by agua bendita: 75%
```

Overview of Brand Similarity Metrics

- Introduction to fuzzy and Jaccard similarity metrics.
- Emphasis on identifying small typographical and formatting differences.

Overview of brand pairs and their similarity scores highlighting the effectiveness of weighted metrics.

Similarity Calculation

```python
from nltk.corpus import words
english_vocab = set(w.lower() for w in words.words())

# add common words in brand names words to the english vocab
# so they're considered as common english words
english_vocab.update(["johnson", "co", "dr", "wd"])

def weighted_score(s1, s2, english_vocab, brand_weight=2.0, common_weight=1.0):

    # Tokenize strings into sets
    tokens1 = set(s1.split())
    tokens2 = set(s2.split())

    # Compute union and intersection
    union_tokens = tokens1.union(tokens2)
    intersection_tokens = tokens1.intersection(tokens2)

    # Handle edge case: if no tokens exist
    if not union_tokens:
        return 0

    # Function to determine token weight
    def token_weight(token):
        return brand_weight if token not in english_vocab else common_weight

    # Compute weighted intersection and weighted union
    weighted_intersection = sum(token_weight(t) for t in intersection_tokens)
    weighted_union = sum(token_weight(t) for t in union_tokens)

    # Compute weighted similarity
    weighted_score = weighted_intersection / weighted_union

    # Convert to percentage scale (0-100)
    return int(weighted_score * 100)
```

Tokenization and Weighting Techniques

Tokenization process: Splitting brand names into discrete elements.

Weight assignment:

- **Distinctive Tokens:** Higher weight for brand-specific terms.
- **Common Tokens:** Lower weight for frequently used English words.

Application of weights to evaluate the significance of each token in similarity computation.

Similarity Calculation

Key	Values
nike	nike, nike swim
rick owens	rick owens, rick owens kids
calvin klein	calvin klein, calvin klein jeans, calvin klein performance
gucci	gucci, gucci kids, gucci eyewear
offwhite	offwhite, offwhite kids
goya	goya, stine goya
samsung	samsung, samsung electronics
disney	disney, disney princess
versace	versace, versace jeans, versace kids
armani	armani jeans, armani, armani exchange
valentino	valentino, red valentino
balenciaga	balenciaga, balenciaga kids
dkny	dkny, dkny sport
ralph lauren	ralph lauren collection, lauren ralph lauren, ralph lauren, polo ralph lauren, polo ralph lauren kids
karl lagerfeld	karl lagerfeld, karl lagerfeld kids
marc jacobs	marc jacobs, marc jacobs kids
stella mccartney	stella mccartney kids, stella mccartney
michael kors	michael kors, michael kors collection, michael michael kors
missoni	missoni, m missoni, missoni mare, missoni home
agua bendita	agua bendita, agua by agua bendita
joseph joseph	joseph joseph, joseph
fendi	fendi, fendi kids
canada goose	canada goose kids, canada goose
moschino	moschino, love moschino, moschino kids, moschino jeans
pieces	pieces, filling pieces
balmain	balmain, balmain kids
philips	philips, philips hue
max mara	max mara, weekend max mara
victoria beckham	victoria beckham, victoria victoria beckham

Detailed Code Analysis

Steps Executed in Code:

- Tokenize both brand names into sets.
- Calculate the union and intersection of these sets.
- Assign weights and compute weighted intersection and union.
- Final similarity score calculation as a percentage.

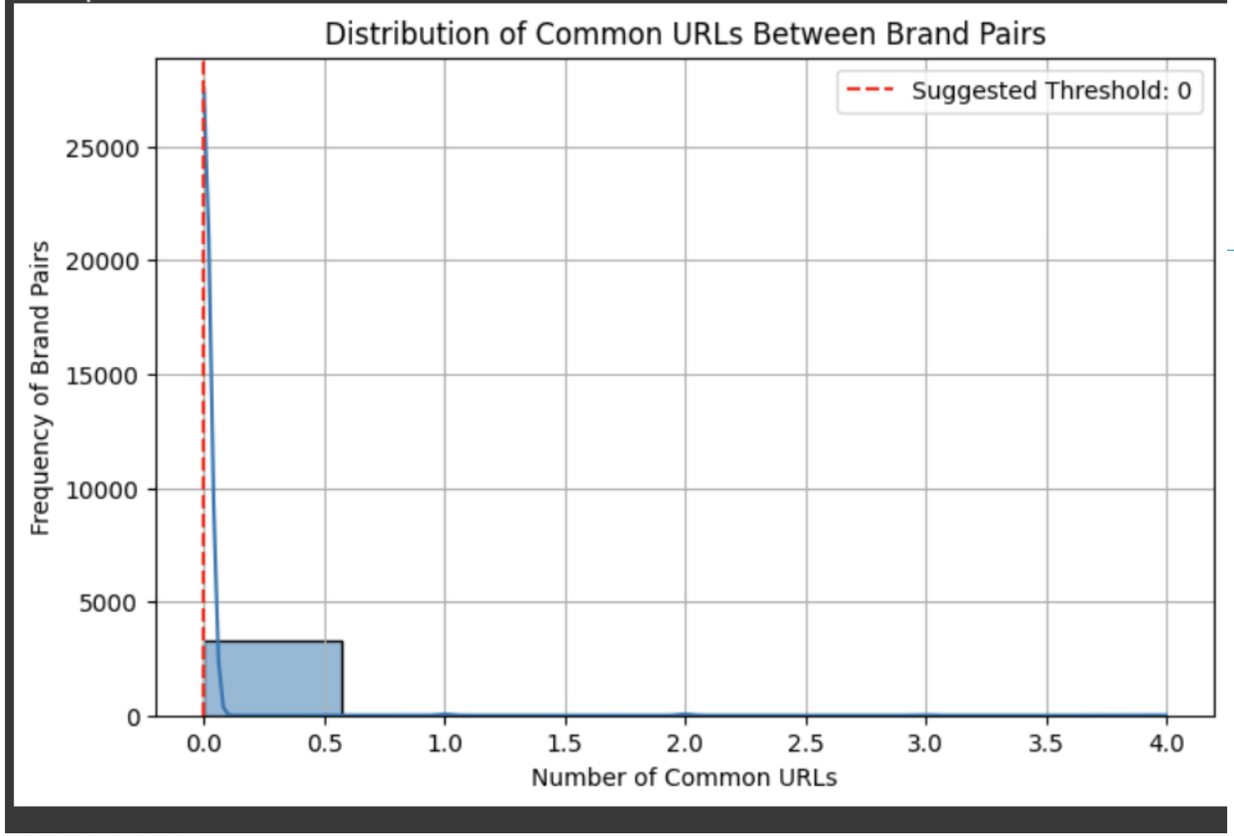
Similarity Calculation

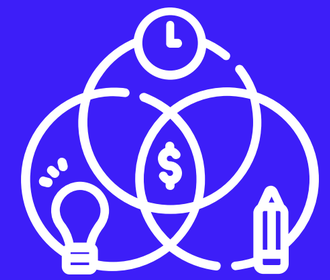
Computing Pairwise Weighted Similarity

	group_id	slug	row_ids	brand_names	categories	row_weights	languages	localization	n_brand_names	total_weight	n_languages	n_categories
0	-1	-1	[57307, 59378, 33579, 5946, 8662, 5943, 38581,...	[nan, nan, nan, nan, nan, nan, nan, ...	[504637, 5866, 499898, 5423, 1505, 1463, 525, ...	[0.5497975721, 0.5517016236, 0.5084593272, 0.4...	[de, it, de, fr, en, en, ko, it, en, es, pt, k...	{'de': nan, 'it': nan, 'fr': nan, 'en': nan, '...	67	1043.086962	10	1155
1	0	disney	[45686, 23448, 43300, 51506, 48882, 46640, 703...	[Disney, Disney, Disney, Disney, Disne...	[674, 1243, 6397, 211, 397, 3601, 5409, 499845...	[0.4732669698, 0.3633569839, 0.273265832, 0.41...	[es, en, en, en, en, en, en, en, en, en, e...	{'es': 'Disney', 'en': 'Disney'}	2	23.282313	2	58
2	1	vevor	[8038, 52528, 29662, 39966, 52968, 17904, 1119...	[VEVOR, VEVOR, VEVOR, VEVOR, VEVOR, VEV...	[952, 586, 3436, 3684, 616, 1194, 1184, 8236, ...	[0.2630395468, 0.2980161196, 0.349981633, 0.38...	[en, en, en, en, en, en, en, en, en, en, e...	{'en': 'VEVOR'}	1	14.008752	1	43
3	2	bosch	[57553, 58572, 17145, 14797, 13931, 23363, 130...	[Bosch, Bosch, Bosch, Bosch, Bosch, BOS...	[264, 503737, 2549, 618, 1203, 2727, 619, 8236...	[0.3479219774, 0.3704948612, 0.5129732222, 0.4...	[en, de, es, de, de, en, fr, en, es, en, en, p...	{'en': 'Bosch', 'de': 'Bosch', 'es': 'Bosch', ...	2	15.920517	6	31
4	3	calvin klein	[871, 32193, 52314, 41944, 2802, 21630, 39362,...	[Calvin Klein, Calvin Klein, Calvin Klein, Cal...	[5183, 2668, 5424, 212, 178, 1831, 479, 2580, ...	[0.3805605416, 0.4281684108, 0.3221414653, 0.5...	[en, de, en, nl, de, de, en, de, en, en, de, d...	{'en': 'Calvin Klein', 'de': 'Calvin Klein', '...	4	13.444201	5	26

Scrapping and API

```
50th percentile (Median): 0.0 URLs
75th percentile: 0.0 URLs
90th percentile (Suggested Threshold): 0.0 URLs
95th percentile: 0.0 URLs
99th percentile: 0.0 URLs
```





4 layers of added variables:

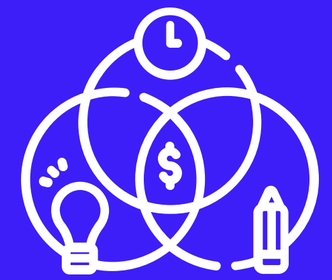
- 3 by using a **google search engine API** and **scraping URLs** from specific search terms in the bar search ("brand name", "brand name + logo"…)
- 1 by using **Gemini API** and prompting to get descriptions for given brand names

Scrapping and API

```
# Prompting in Gemini and extracting its answer in a specified format
response = client.models.generate_content(
    model="gemini-2.0-flash", contents=f"""\
Here is a list of brands: {brand_n}
Please analyze those brands and find out the short description of each (between 100 and 150 words).
Return a list with the following JSON format:
[{{
    "brand_name": "brand name 1",
    "brand_description": "short description of brand 1"
  }},
  {{
    "brand_name": "brand name 2",
    "brand_description": "short description of brand 2"
  }},
  #... more brands]
Please include all the brands in the original list and do not add any other comment
and do not use any punctuation of other than dots in the descriptions
"""
)

print(response.text)
```

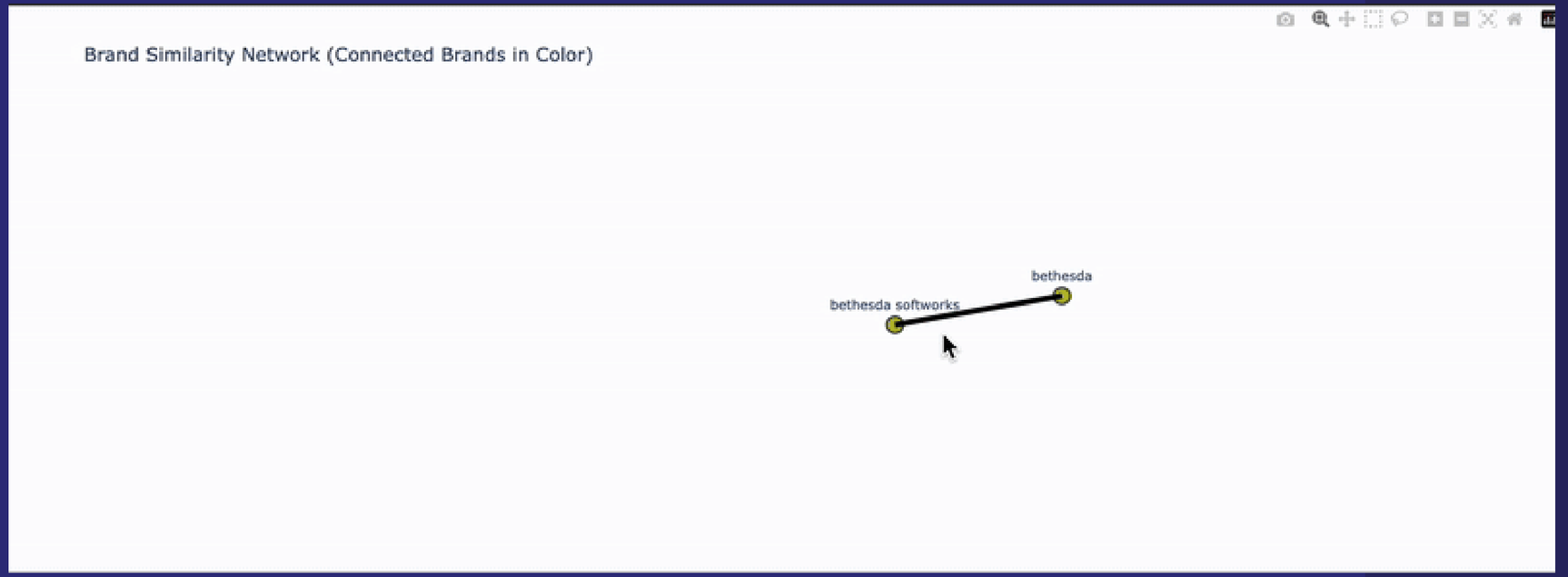
```
  {
    "brand_name": "microsoft",
    "brand_description": "Microsoft is a multinational technology corporation. It develops licenses supports and sells computer software consumer elec
  },
  {
    "brand_name": "adobe",
    "brand_description": "Adobe Inc is a multinational computer software company. It focuses on creativity design and digital document experiences. Its
  },
  {
    "brand_name": "corel",
    "brand_description": "Corel Corporation specializes in graphic design and productivity software. Its well known products include CorelDRAW Graphics
  },
  {
    "brand_name": "2k games",
    "brand_description": "2K Games is a video game publisher. It is a subsidiary of Take Two Interactive. 2K Games publishes a wide variety of games a
  },
  {
    "brand_name": "paladone",
    "brand_description": "Paladone is a supplier of gifts and merchandise. They specialize in creating innovative and trending products for various re
  },
  {
    "brand_name": "rockstar games",
    "brand_description": "Rockstar Games is a video game publisher and developer. It is a subsidiary of Take Two Interactive. Rockstar Games is best k
  },
  {
    "brand_name": "renegade game studios",
    "brand_description": "Renegade Game Studios is a board game publisher. It focuses on developing and publishing engaging and innovative tabletop ga
```



Then adding data to the dataframe and analyzing the similarity between these added variables and grouping brands when the similarity is above a chosen threshold

Analysis and Visualization: Bringing Brands Together

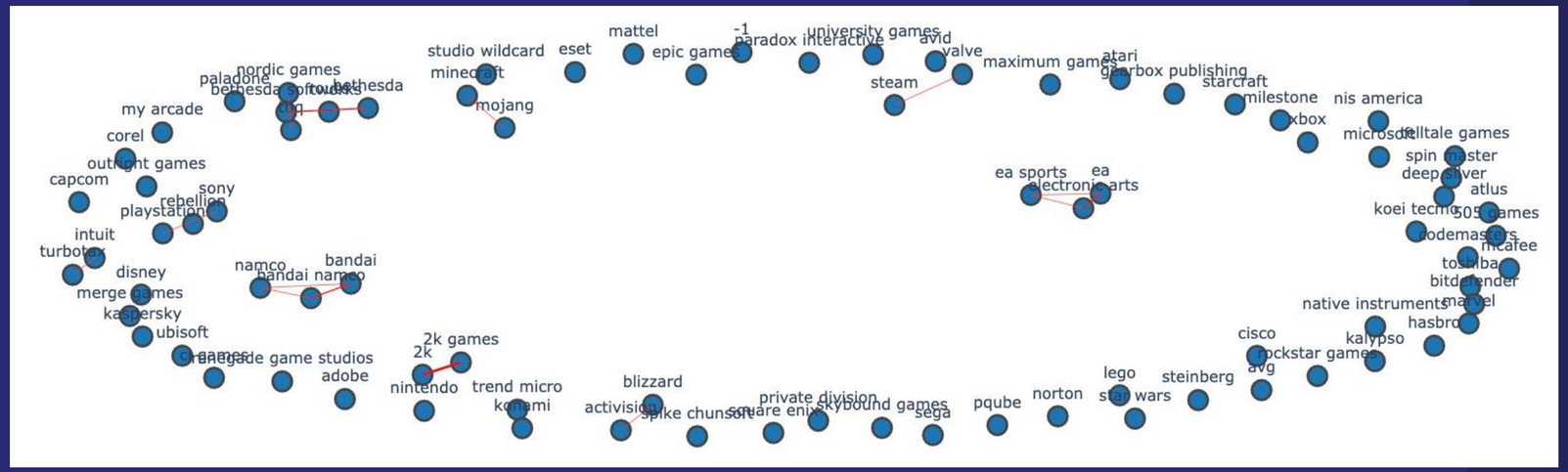
Similarity grouping with added data



Analysis and Visualization: Bringing Brands Together

Similarity grouping with added data

For the 3 layers of added URLs, a brand pairwise count of common URL is computed and the brands are added to same group if the count is above selected threshold



Key	Values
2k	2k games, 2k
bethesda	bethesda softworks, bethesda
ea	ea, ea sports, electronic arts

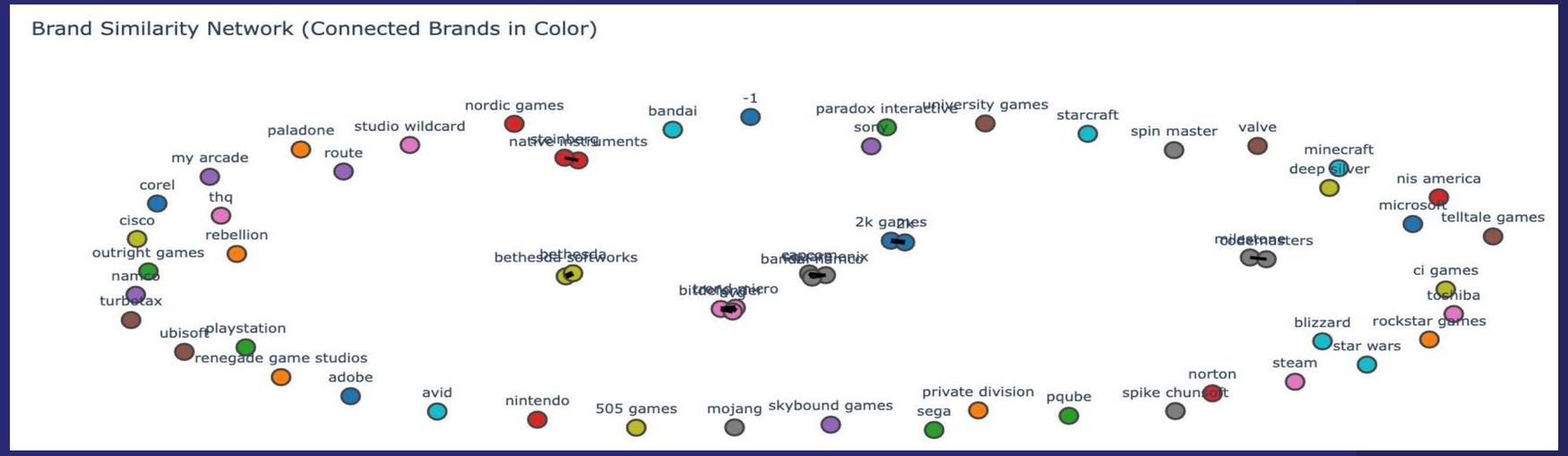
Key	Values
2k	2k games, 2k
playstation	playstation, sony
nordic games	nordic games, thq
namco	namco, bandai, bandai namco
valve	valve, steam
turbotax	turbotax, intuit
minecraft	minecraft, mojang
bethesda	bethesda softworks, bethesda
blizzard	blizzard, activision
ea	ea, ea sports, electronic arts

	slug	row_ids	brand_names	categories	row_weights	languages	localization	brand_keywords	n_brand_names	n_languages	n_categories
0	-1	[[985, 1350, 1585, 2449, 7209, 9442, 11018, 12...	[[nan, nan, nan, nan, nan, nan, nan,...	[[5304, 1279, 5303, 5300, 4952, 5303, 1279, 49...	[[0.7406684242, 0.7383853176, 0.4478219223, 0....	[['en', 'ja', 'es', 'en', 'en', 'fr', 'fr', 'e...	{['en', 'ja', 'es', 'en', 'en', 'fr', 'fr', '...	[Brand likely placeholder error company gaming...	1	9	15
1	2k	[[53487], [32875]]	[['2K'], ['2K Games']]	[[1279]]	[[0.4212443607], [0.4469868681]]	[['en']]	{['en']: []}	[K video game publisher subsidiary Interactive...	2	1	1
2	505 games	[[32184, 37481]]	[['505 Games', '505 Games']]	[[1279, 1279]]	[[0.3877262922, 0.4087040274]]	[['en', 'en']]	{['en', 'en']: []}	[global video game publisher diverse range pla...	1	1	1
3	activision	[[35287, 52248]]	[['Activision', 'Activision']]	[[1279, 1279]]	[[0.4176148247, 0.5366593163]]	[['en', 'en']]	{['en', 'en']: []}	[nan]	1	1	1
4	adobe	[[48680, 52310]]	[['Adobe', 'Adobe']]	[[5303, 4951]]	[[0.353601628, 0.3779333934]]	[['en', 'en']]	{['en', 'en']: []}	[Inc multinational computer software company c...	1	1	2
5	atari	[[62725]]	[['Atari']]	[[1279]]	[[0.497306481]]	[['en']]	{['en']: []}	[nan]	1	1	1
6	atlus	[[49482]]	[['Atlus']]	[[1279]]	[[0.4026980106]]	[['en']]	{['en']: []}	[nan]	1	1	1
7	avid	[[19440]]	[['Avid']]	[[5096]]	[[0.3831552337]]	[['en']]	{['en']: []}	[Technology technology company audio video edi...	1	1	1
8	bandai	[[29445, 40457]]	[['Bandai', 'Bandai']]	[[1279, 1279]]	[[0.4875174143, 0.408956381]]	[['en', 'en']]	{['en', 'en']: []}	[Japanese toy maker entertainment company popu...	1	1	1
9	bandai namco	[[21135]]	[['Bandai Namco']]	[[1279]]	[[0.4331794997]]	[['en']]	{['en']: []}	[Entertainment Japanese video game developer p...	1	1	1

Analysis and Visualization: Bringing Brands Together

Similarity grouping with added data

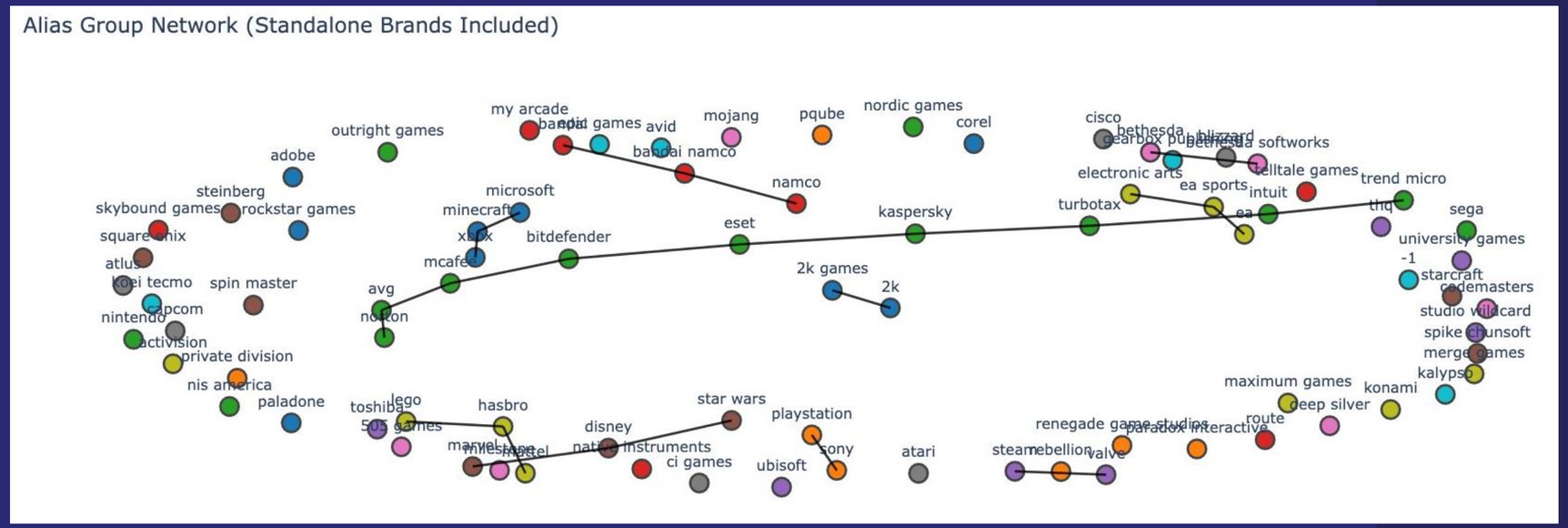
For the added Description, a cleaning is performed to keep only keyword and compute the brand pairwise similarity of their associated keywords vectors
the brands are then added to same group if the similarity is above selected threshold




```
Brand pairs with Weighted Jaccard above threshold:
2k games and 2k : 90%
trend micro and bitdefender : 92%
trend micro and avg : 84%
bethesda softworks and bethesda : 100%
bitdefender and avg : 81%
```

Evaluation metric's result

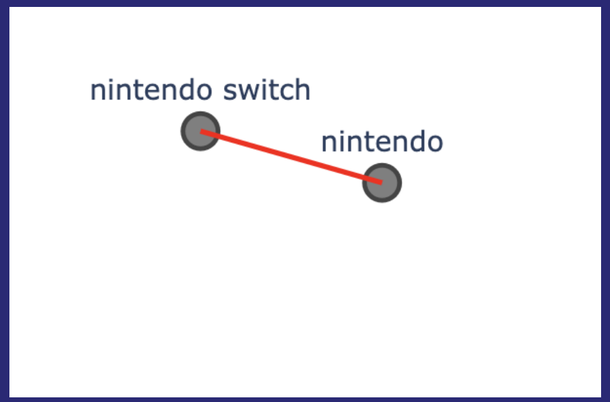
To evaluate the performance of the different methods, we compared the groups made after analyzing for similarity and the groups made by Gemini when prompted to group the brands together based on his knowledge. We get the following output

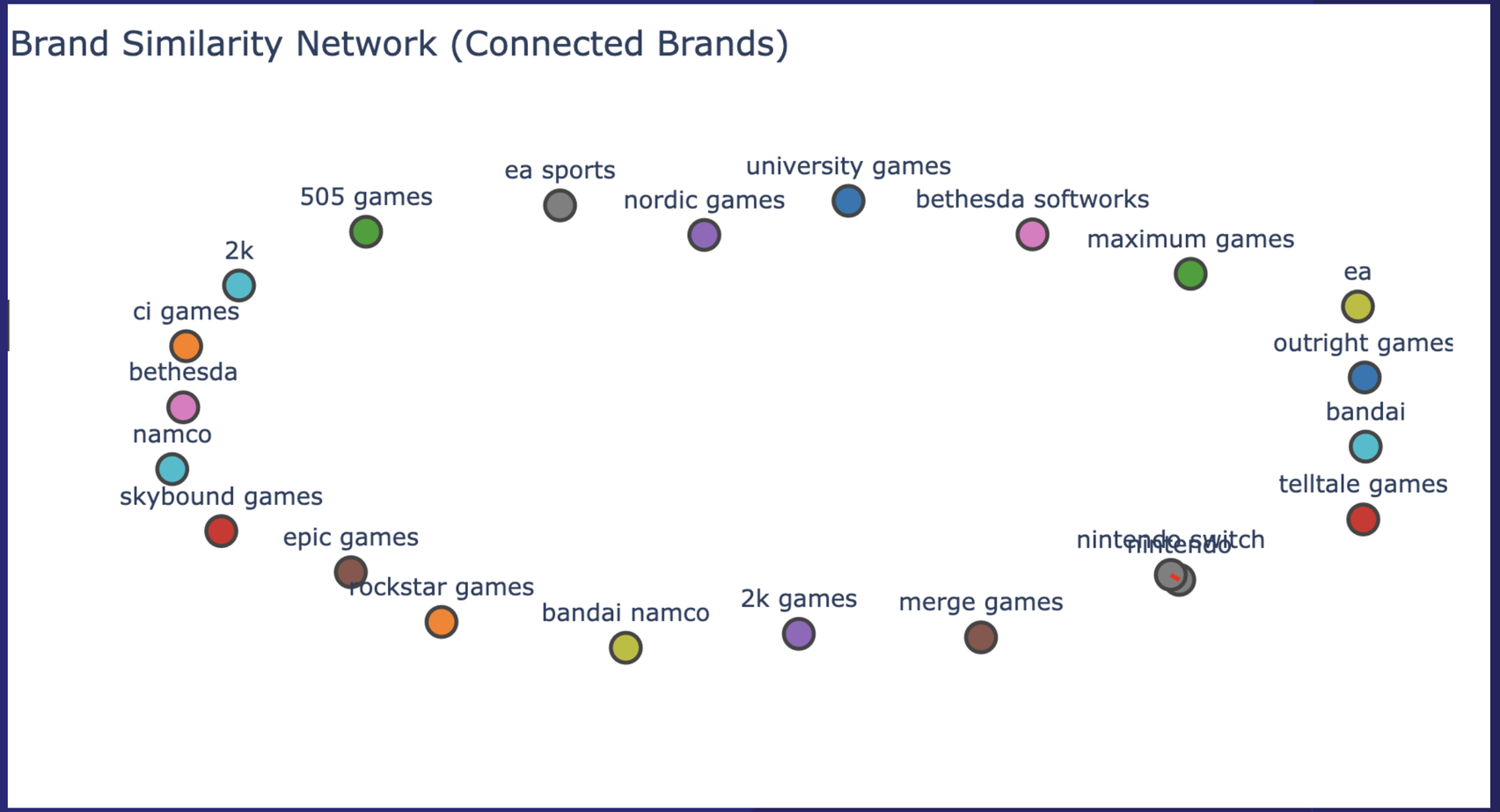



```
{"Microsoft": ["microsoft", "minecraft", "xbox"]},
{"Adobe": ["adobe"]},
{"Corel": ["corel"]},
{"2K Games": ["2k games", "2k"]},
{"Paladone": ["paladone"]},
{"Rockstar Games": ["rockstar games"]},
{"Renegade Game Studios": ["renegade game studios"]},
{"Rebellion": ["rebellion"]},
{"Private Division": ["private division"]},
{"PQube": ["pqube"]},
{"Playstation": ["playstation", "sony"]},
{"Paradox Interactive": ["paradox interactive"]},
{"Outright Games": ["outright games"]},
{"Sega": ["sega"]},
{"NortonLifeLock": ["norton", "avg", "mcafee", "bitdefender", "es
{"Nordic Games": ["nordic games"]},
{"NIS America": ["nis america"]},
{"Nintendo": ["nintendo"]},
{"Native Instruments": ["native instruments"]},
{"Namco Bandai": ["namco", "bandai namco", "bandai"]},
{"My Arcade": ["my arcade"]},
```

slug	row_ids	brand_names	categories	row_weights	languages	localization	n_brand_names	n_languages	n_categories
-1	[[985, 1350, 1585, 2449, 7209, 9442, 11018, 12...	[[nan, nan, nan, nan, nan, nan, nan, nan, nan,...	[[5304, 1279, 5303, 5300, 4952, 5303, 1279, 49...	[[0.7406684242, 0.7383853176, 0.4478219223, 0....	[['en', 'ja', 'es', 'en', 'en', 'fr', 'fr', 'e...	{['en', 'ja', 'es', 'en', 'en', 'fr', 'fr', '...	1	9	15
2k	[[53487], [32875]]	[['2K'], ['2K Games']]	[[1279]]	[[0.4212443607], [0.4469868681]]	[['en']]	{['en']: []}	2	1	1
505 games	[[32184, 37481]]	[['505 Games', '505 Games']]	[[1279, 1279]]	[[0.3877262922, 0.4087040274]]	[['en', 'en']]	{['en', 'en']: []}	1	1	1
adobe	[[48680, 52310]]	[['Adobe', 'Adobe']]	[[5303, 4951]]	[[0.353601628, 0.3779333934]]	[['en', 'en']]	{['en', 'en']: []}	1	1	2
atari	[[62725]]	[['Atari']]	[[1279]]	[[0.497306481]]	[['en']]	{['en']: []}	1	1	1
atlus	[[49482]]	[['Atlus']]	[[1279]]	[[0.4026980106]]	[['en']]	{['en']: []}	1	1	1
avg	[[62288]]	[['AVG']]	[[5299]]	[[0.2997889495]]	[['en']]	{['en']: []}	1	1	1
avid	[[19440]]	[['Avid']]	[[5096]]	[[0.3831552337]]	[['en']]	{['en']: []}	1	1	1
bethesda	[[6636], [37313]]	[['Bethesda Softworks'], ['Bethesda']]	[[1279]]	[[0.4672479376], [0.412768472]]	[['en']]	{['en']: []}	2	1	1
bitdefender	[[33653]]	[['Bitdefender']]	[[5299]]	[[0.3409507661]]	[['en']]	{['en']: []}	1	1	1

Evaluation metric's result





Use Cases

RETAIL MEDIA and UNIFICATION OF DATA

CASE	SOLUTION	EXAMPLES AND IMPACTS
Dynamic Pricing Campaigns	Implement real-time pricing strategies using unified data	Users see consistent prices across platforms, increasing trust and likelihood to buy.
Inventory-Driven Advertising	Adjust ad placements based on real-time stock levels	Prevent ad spend on out-of-stock items, enhancing ROI and customer satisfaction.
Seasonal Promotion Optimization	Harmonize seasonal campaigns across multiple retailers	Coordinated ads boost visibility and sales during peak seasons.
Enhanced Customer Segmentation	Improve accuracy in customer targeting with cleaner data	Targeted ads based on accurate user profiles increase conversion rates.
Compliance-Driven Targeting	Ensure regional compliance through consistent data use	Unified branding helps adhere to advertising standards, reducing legal risks

BUSINESS IMPACT

Sustainability in Advertising

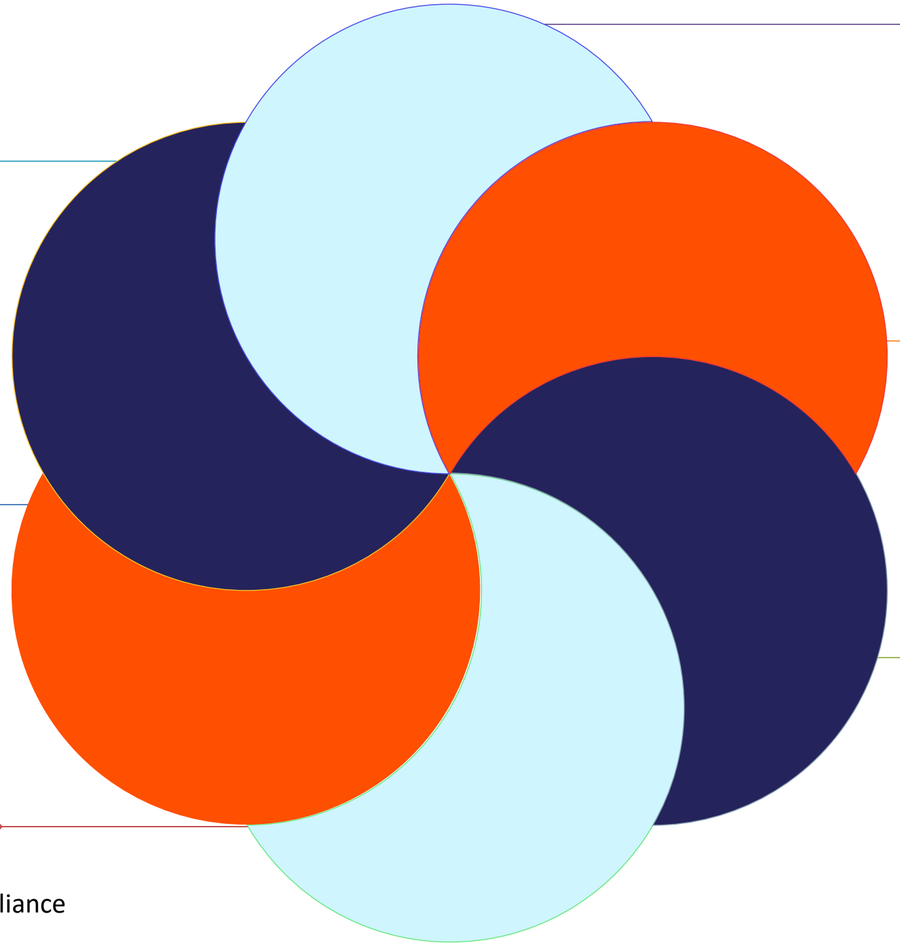
- Reduced digital waste through targeted advertising reduces carbon footprint.
- Alignment with consumer values around sustainability enhances brand reputation.

Innovative Advertising Solutions

- Advanced personalization capabilities from accurate data analysis.
- Development of cutting-edge ad technologies that resonate with target demographics.

Enhanced Compliance and Reduced Risk

- Unified data helps meet diverse regional compliance requirements.
- Minimized legal risks through adherence to international privacy laws.



Enhanced Customer Experience

- Consistency across all platforms increases brand loyalty and customer satisfaction.
- Seamless interactions elevate user experience, driving repeat business

Operational Efficiency

- Streamlined operations reduce costs and improve speed to market.
- Data centralization simplifies management and enhances productivity.

Data-Driven Decision Making

- Access to cleaner, consolidated data enables precise analytics.
- Real-time insights allow for dynamic marketing and optimized ad spend.

Thank you!

We are Group 17