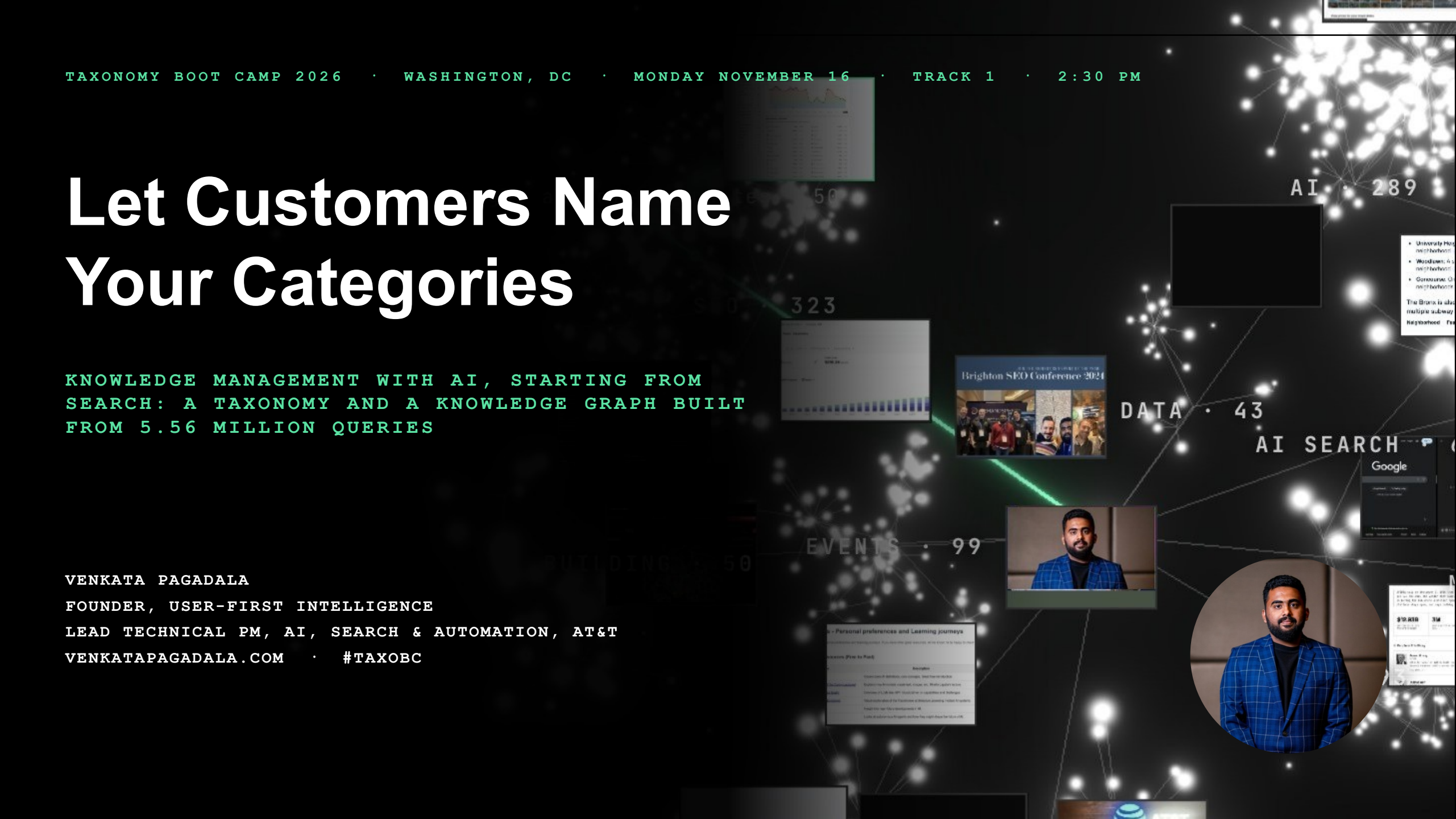
TAXONOMY BOOT CAMP 2026 · WASHINGTON, DC · MONDAY NOVEMBER 16 · TRACK 1 · 2:30 PM

Let Customers Name Your Categories

KNOWLEDGE MANAGEMENT WITH AI, STARTING FROM SEARCH: A TAXONOMY AND A KNOWLEDGE GRAPH BUILT FROM 5.56 MILLION QUERIES

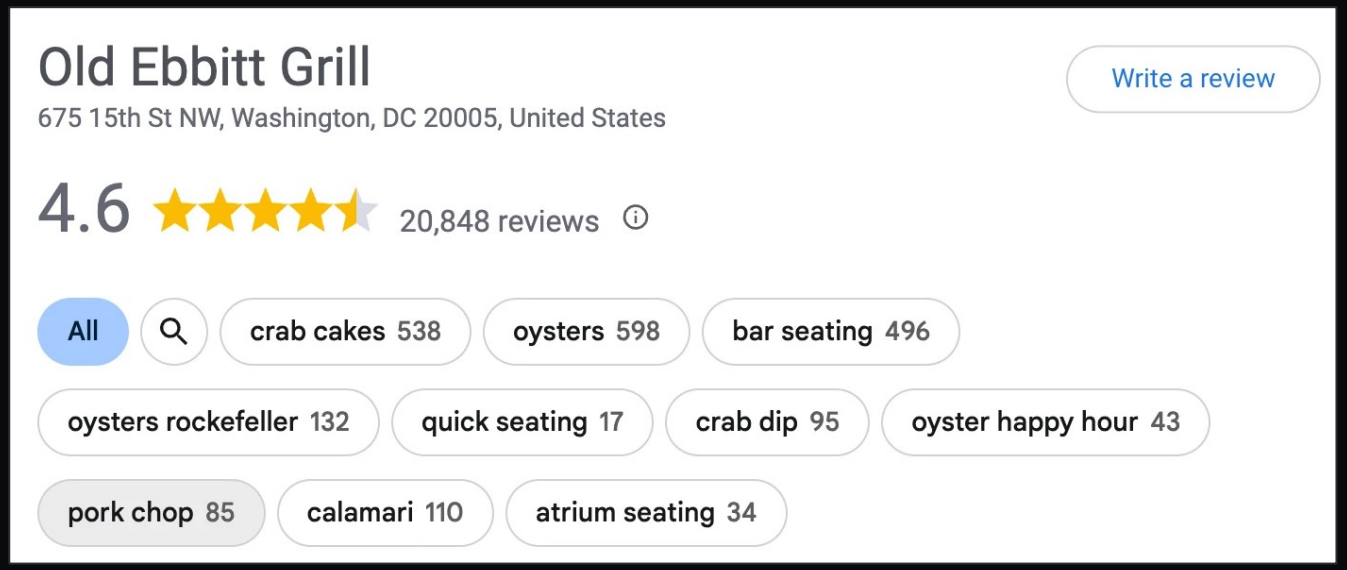
VENKATA PAGADALA
FOUNDER, USER-FIRST INTELLIGENCE
LEAD TECHNICAL PM, AI, SEARCH & AUTOMATION, AT&T
VENKATAPAGADALA.COM · #TAXOBC



LAST NIGHT · 9:30 PM · OLD EBBITT GRILL, WASHINGTON DC

Last night, I asked Google what to order.

WHAT GOOGLE SHOWED ME · THE WORDS REVIEWERS USE MOST



WHAT I ORDERED · MY OWN PHOTO



So I ordered the crab cakes.

Then I read the menu. It calls them something else.

SOURCE · GOOGLE REVIEWS FOR OLD EBBITT GRILL, NOVEMBER 2026

The restaurant named the dishes. Customers renamed them.

THE MENU SAYS		REVIEWERS SAY · WITH GOOGLE'S COUNT
Jumbo Lump Crab Cakes	→	**crab cakes** 538
Day Boat New England Calamari	→	**calamari** 110
Crab & Artichoke Dip	→	**crab dip** 95
Duroc Pork Chop	→	**pork chop** 85

THE ONE DISH NAME THAT MATCHED

Oysters Rockefeller = **oysters rockefeller** 132

When the menu already uses the customers' words, there is nothing to rename.

Nobody asked them to rename the dishes. They did it anyway.

SOURCES · MENU: EBBITT.COM/MENUS, READ 16 NOV 2026 · COUNTS: GOOGLE REVIEWS, NOVEMBER 2026

BEFORE WE START

No need to take notes.

Everything from this talk is on one page:

→ The checklist, all eight steps

→ Every example, with its source

→ The recording, with a full transcript

venkatapagadala.com/research-and-talks/taxonomy-boot-camp-2026

Listen now. Scan any time.



SCAN FOR THE NOTES

THE ONE IDEA

Your customers, across the whole industry, have already searched for it.

THE ONE IDEA

Your customers, across the whole industry, have already searched for it.

Existing, future and potential customers. That search is step one of your taxonomy.

THE ONE IDEA

Your customers, across the whole industry, have already searched for it.

Existing, future and potential customers. That search is step one of your taxonomy.

01 HOW MANY LEVELS A KNOWLEDGE STRUCTURE NEEDS

Four. Deeper only where the data forces it.

02 A HEALTH SCORE AND TWELVE GATES

55 to 91.5, and nothing is published until every gate is green.

03 WHERE AI STOPS AND PEOPLE DECIDE

Agents gather, classify, propose. People own the labels and the gates.

The biggest names in AI and tech are all building the same thing

Google Search Knowledge Graph behind AI Mode answers	**Google Cloud** Knowledge Catalog, BigQuery Graph 2026	**Microsoft** GraphRAG, Fabric IQ Ontology 2024 to 2026	**AWS** Graph and AI, Neptune connected data for agents	**Amazon** Product Knowledge Graph Amazon Science	**IBM** GraphRAG, watsonx.data 2026
Oracle GraphRAG, AI Database 26ai 2026	**SAP** SAP Knowledge Graph 2026	**Salesforce** Semantic layer, Data Cloud 2025	**ServiceNow** Knowledge Graph 2025	**Databricks** Genie Ontology June 2026	**Snowflake** Ontology for Cortex Agents May 2026
Palantir The Ontology the core of the platform	**Atlassian** Teamwork Graph opened up, May 2026	**OpenAI** Company knowledge in ChatGPT 2025	**Glean** Enterprise Graph 2025	**LinkedIn** Knowledge Graph since 2016	**Neo4j** The GraphRAG manifesto 2024

Seven names, one idea: knowledge graph, ontology, GraphRAG, catalog, semantic layer, enterprise graph, teamwork graph.

EACH FROM THE COMPANY'S OWN SITE, CHECKED NOVEMBER 2026 · SHOWN AS EVIDENCE OF THE SHIFT, NOT AS RECOMMENDATIONS

Venkata Pagadala

FOUNDER, USER-FIRST INTELLIGENCE · LEAD TECHNICAL PM, AI, SEARCH & AUTOMATION, AT&T · RESEARCHER · SPEAKER

I build User-First Intelligence systems with agentic AI: users, personas, platforms, content, business. AI agents gather and classify the signals at every step, and people make the calls.

WORKED WITH AND RESEARCHED

- AT&T
- Meta
- Etsy
- Coursera
- Sage
- Apartments.com
- ApartmentFinder.com
- Rehabs.com
- DrugAbuse.com
- American Addiction Centers
- Previsible
- Demand Engine
- Amazing Digital
- DYAD
- Sprout for Business

10+
YEARS IN SEARCH
first search role, April 2016

20,000+
LINKEDIN FOLLOWERS
profile, 22 Oct 2026

484,000+
POST IMPRESSIONS, 12 MONTHS
LinkedIn analytics

931
POSTS AND ARTICLES ARCHIVED
2016 to 2026

3
PUBLISHED PAPERS
2 peer-reviewed, 1 preprint

10
TALKS, PODCASTS, INTERVIEWS
each linked to its recording

Every number names where it comes from. venkatapagadala.com/about

01 HOW IT STARTED

the same question every week; what one query carries

02 FROM SOURCES TO A PRODUCT

22 kinds of source, 105K reviews, one map, one node, one product

03 FROM A TREE TO A GRAPH

four levels, facets, the graph, sorting at scale

04 THE SCORE, THE AUDIT AND THE GATES

a health score, 55 to 91.5; an audit, 30 out of 100; twelve mistakes, twelve gates

05 THE SAME METHOD, AND HOW TO START

education, healthcare and home services; the checklist

The same question, every week

> "Do we have a page for this?"

The same question, every week

"Do we have a page for this?"

"Which topic does this launch belong to?"

The same question, every week

"Do we have a page for this?"

"Which topic does this launch belong to?"

"Is there demand for this, or are we guessing?"

The same question, every week

"Do we have a page for this?"

"Which topic does this launch belong to?"

"Is there demand for this, or are we guessing?"

"Why do we have three pages that say the same thing?"

2016 ME

Chasing volumes and trends.

2026 ME

Chasing users and their pain points. The answer was a person. It should have been a knowledge base.

So I built the map once, and let the data maintain it.

From raw queries to a map

5.56M

RAW QUERIES, 22 KINDS OF SOURCE

From raw queries to a map

5.56M → **1.90M**

RAW QUERIES, 22 KINDS OF SOURCE

KEPT, UNCOUNTED QUESTIONS INCLUDED

From raw queries to a map

5.56M → **1.90M** → **37.5K**

RAW QUERIES, 22 KINDS OF SOURCE

KEPT, UNCOUNTED QUESTIONS INCLUDED

TOPICS, L1 TO L8

From raw queries to a map

5.56M RAW QUERIES, 22 KINDS OF SOURCE

➡ **1.90M** KEPT, UNCOUNTED QUESTIONS INCLUDED

➡ **37.5K** TOPICS, L1 TO L8

➡ **17** TOP-LEVEL CATEGORIES, NAMED BY QUERIES

3.7 MILLION QUERIES REMOVED AS NOISE BEFORE ANYTHING WAS CLASSIFIED · THE STRUCTURE IS MINE, THE QUERIES ARE THE INDUSTRY'S

One query, four signals

```
cheapest unlimited plan for 2 lines near me
```

ENTITY

unlimited plan

what they are asking about: a node in the taxonomy

One query, four signals

```
cheapest unlimited plan for 2 lines near me
```

ENTITY

unlimited plan

what they are asking about: a node in the taxonomy

NEED

cheapest

the reason behind the ask: it becomes a label, in their words

One query, four signals

`cheapest unlimited plan for 2 lines near me`

ENTITY

unlimited plan

what they are asking about: a node in the taxonomy

NEED

cheapest

the reason behind the ask: it becomes a label, in their words

INTENT

compare, then buy

research, compare, purchase, support, navigate

One query, four signals

`cheapest unlimited plan for 2 lines near me`

ENTITY

unlimited plan

what they are asking about: a node in the taxonomy

NEED

cheapest

the reason behind the ask: it becomes a label, in their words

INTENT

compare, then buy

research, compare, purchase, support, navigate

MODIFIERS

2 lines · near me

quantity, place, price, urgency: facets, never levels

Every query is a vote: for a category, for a label, and for where that label sits. Count the votes and the taxonomy starts to write itself.

Where to start: one hundred places your customers already talk

01 SEARCH AND DISCOVERY

- Google Search
- Google autocomplete
- People also ask
- Google Trends
- Bing
- DuckDuckGo
- YouTube search
- Amazon search
- App Store search
- Pinterest

02 AI ANSWER ENGINES

- ChatGPT
- Perplexity
- Gemini
- Claude
- Copilot
- Google AI Overviews
- Meta AI
- Grok
- You.com
- DeepSeek

03 VIDEO AND AUDIO

- YouTube videos
- YouTube comments
- YouTube live chat
- TikTok
- Instagram Reels
- Twitch
- Spotify podcasts
- Apple Podcasts
- Vimeo
- Webinar replays

04 SOCIAL

- X posts and replies
- LinkedIn posts
- LinkedIn comments
- Facebook groups
- Facebook pages
- Instagram comments
- Threads
- Bluesky
- Snapchat
- Nextdoor

05 COMMUNITIES AND Q&A

- Reddit
- Quora
- Stack Exchange
- Stack Overflow
- Discord servers
- Slack communities
- Hacker News
- Product Hunt
- Brand support forums
- XDA and niche forums

SEARCH IS ONE GROUP OF TEN. THE OTHER NINETY ARE WHERE THE INDUSTRY NAMES THINGS IN ITS OWN WORDS. · PAGE 1 OF 2

Where to start: one hundred places your customers already talk

06 EXPERT AND LONG-FORM

- Substack
- Medium
- Hugging Face
- GitHub issues
- arXiv
- Google Scholar
- Wikipedia
- Wikidata
- Conference notes
- Industry newsletters

07 COMPANY AND MARKET

- SEC filings
- Earnings calls
- Investor presentations
- Press releases
- Job postings
- Competitor sitemaps
- Competitor pricing pages
- Crunchbase
- Patents
- Analyst reports

08 MARKETPLACE REVIEWS

- Amazon reviews and Q&A
- Best Buy reviews
- Target reviews
- Costco reviews
- Walmart reviews
- Home Depot reviews
- eBay listings
- Etsy reviews
- Wayfair reviews
- Newegg reviews

09 RATINGS AND RESPONSES

- Google Business reviews
- Owner responses to reviews
- Yelp
- Trustpilot
- Better Business Bureau
- G2
- Capterra
- Glassdoor
- App Store reviews
- Google Play reviews

10 YOUR OWN AND PUBLIC DATA

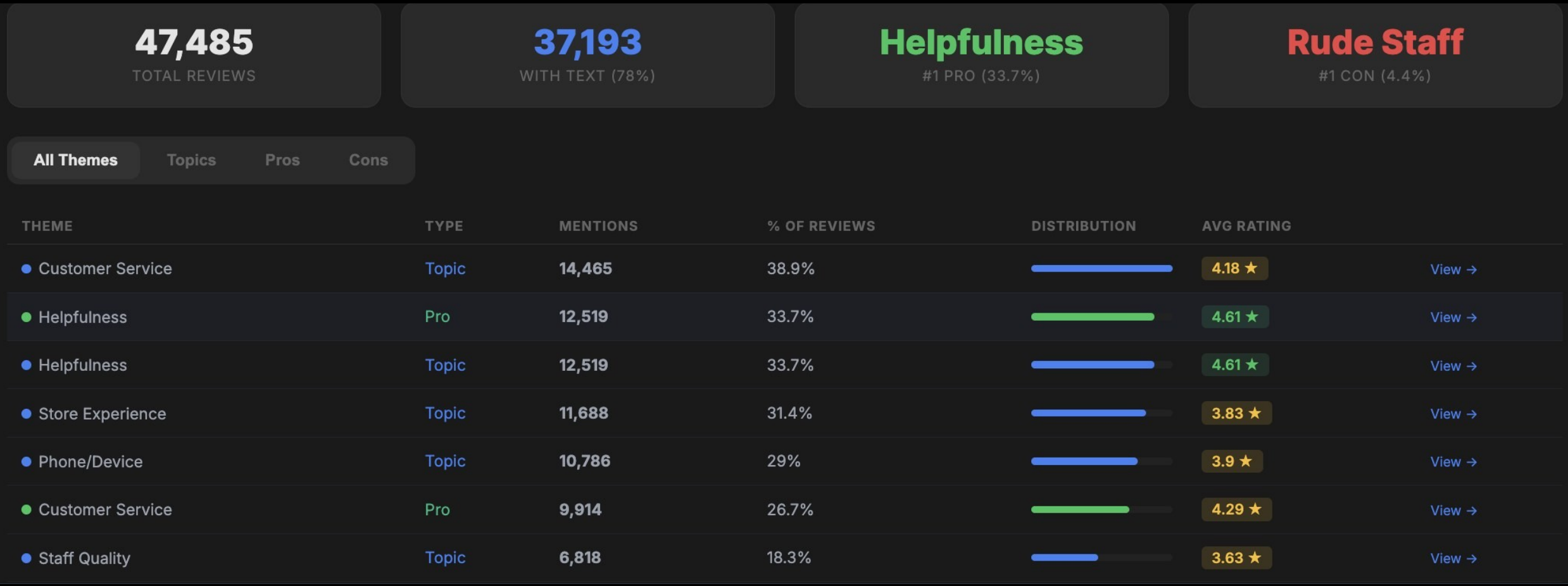
- Site search logs
- Support tickets
- Chat transcripts
- Call center logs
- Sales call notes
- Surveys and NPS verbatims
- Regulations.gov
- Census and data.gov
- Court records
- Common Crawl

YOU DO NOT NEED ALL HUNDRED ON DAY ONE. YOU NEED MORE THAN ONE. PICK THE TWO YOUR CUSTOMER ACTUALLY USES. · PAGE 2 OF 2

Let us follow one source all the way to a product

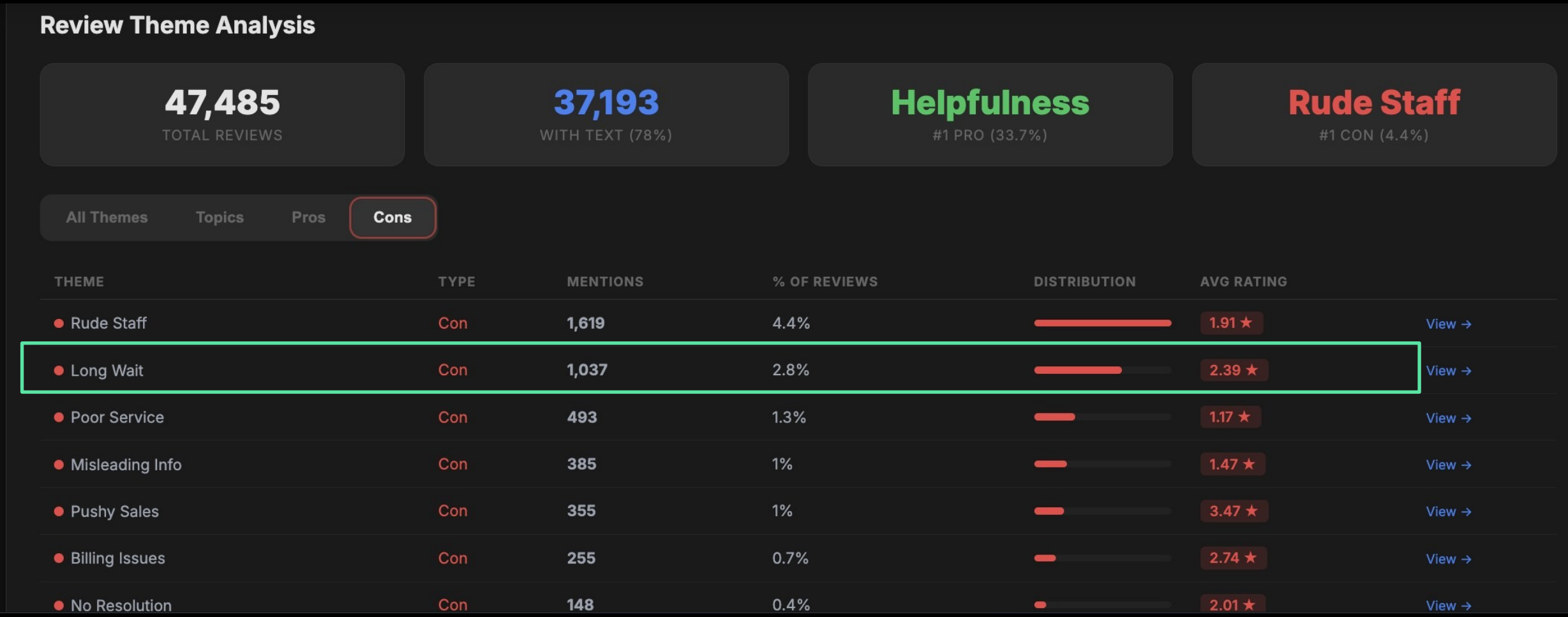
06 EXPERT AND LONG-FORM	07 COMPANY AND MARKET	08 MARKETPLACE REVIEWS	09 RATINGS AND RESPONSES	10 YOUR OWN AND PUBLIC DATA
Substack	SEC filings	Amazon reviews and Q&A	**Google Business reviews**	Site search logs
Medium	Earnings calls	Best Buy reviews	Owner responses to reviews	Support tickets
Hugging Face	Investor presentations	Target reviews	Yelp	Chat transcripts
GitHub issues	Press releases	Costco reviews	Trustpilot	Call center logs
arXiv	Job postings	Walmart reviews	Better Business Bureau	Sales call notes
Google Scholar	Competitor sitemaps	Home Depot reviews	G2	Surveys and NPS verbatims
Wikipedia	Competitor pricing pages	eBay listings	Capterra	Regulations.gov
Wikidata	Crunchbase	Etsy reviews	Glassdoor	Census and data.gov
Conference notes	Patents	Wayfair reviews	App Store reviews	Court records
Industry newsletters	Analyst reports	Newegg reviews	Google Play reviews	Common Crawl

TARGET LOCKED · GOOGLE BUSINESS REVIEWS · 465 TELECOM STORES ACROSS THREE CARRIERS · WHAT CUSTOMERS SAID AFTER THE PURCHASE, IN THEIR OWN WORDS



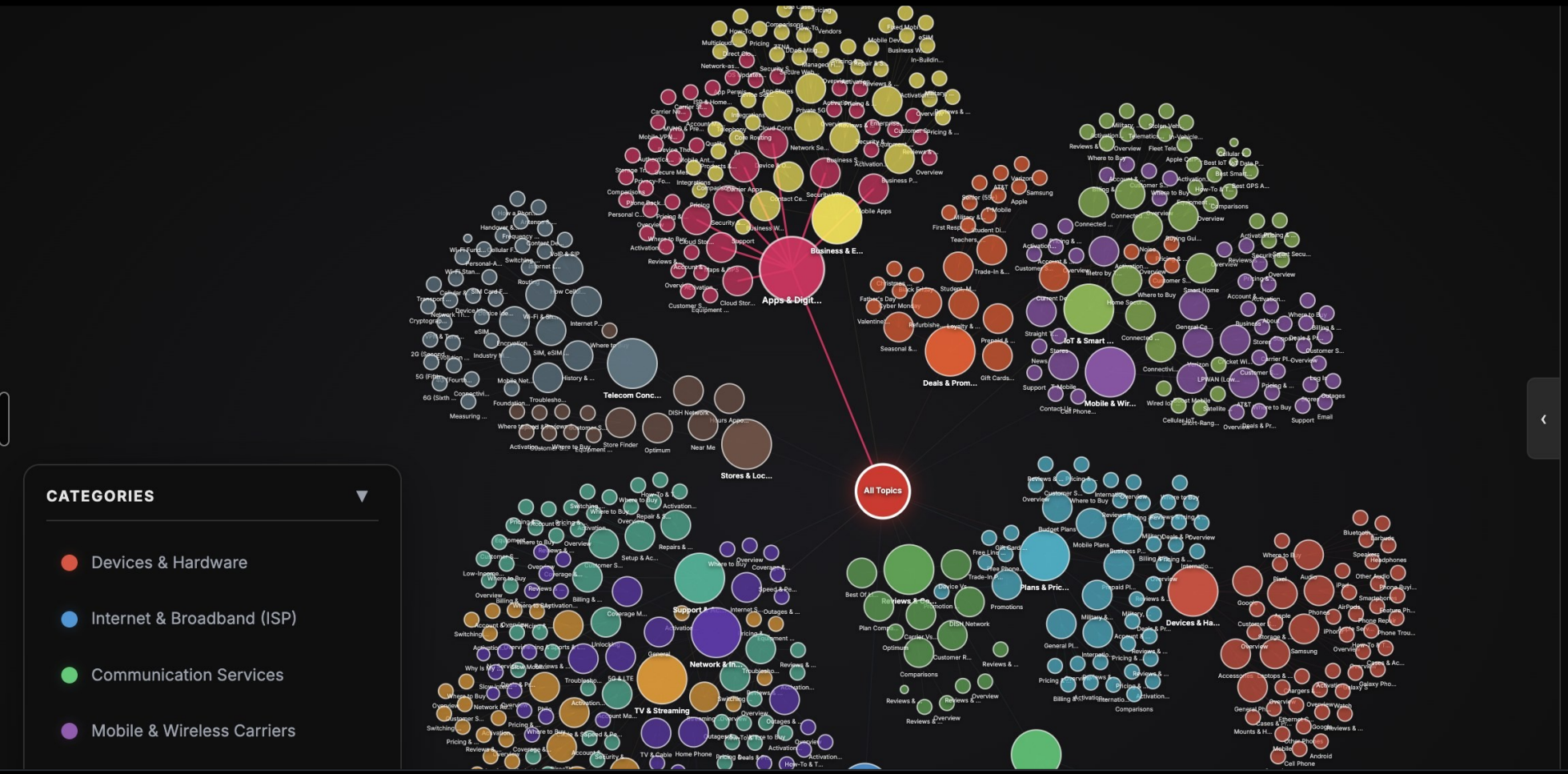
Step 2. One source no keyword tool sees: 105,400 store reviews.

465 TELECOM STORES ACROSS THREE CARRIERS · 47,485 REVIEWS IN THIS VIEW, 37,193 WITH TEXT · EVERY REVIEW BECOMES A THEME WITH A COUNT AND A RATING · MY OWN RESEARCH PLATFORM, SHOWN FOR THE METHOD



Step 3. One node: 'Long wait'. 1,037 mentions, 2.39 stars, never typed into a search box.

A PAIN POINT IN A REVIEW IS A QUERY NOBODY TYPED · NOBODY WAS WRITING CONTENT FOR IT · THE KNOWLEDGE BASE GOT THE NODE BEFORE THE CONTENT TEAM GOT THE BRIEF



57.1K KEYWORDS | 22.6M VOLUME | 35 TOPICS

MARKET INTELLIGENCE · B2C

$586M	$52M	$4.79
SEARCH VALUE/YR	CAPTURABLE	WTD CPC
99%	0.07	-4%
% COMMERCIAL	DIFFICULTY	MOMENTUM

Confidence: **Medium** · ad-equivalent value of search demand, not market revenue

telecom.apps_software

SUBTOPICS (35)

SUBTOPIC	KEYWORDS	VOLUME	TOPICS
Mobile Apps	25.9K	8.5M	20
Maps & GPS	6.6K	3.6M	19
Security & VPN Apps	2.1K	2.6M	8
Security VPN	7.3K	2.6M	20
Cloud Storage	7.8K	2.0M	18
Cloud Storage & Backup Apps	1.7K	1.4M	7
Device & OS App Ecosystems	429	487.7K	4

Step 4. The topics around it, as the data named them. Seventeen categories around one core.

APPS & DIGITAL SERVICES: 57.1K KEYWORDS · 35 SUBTOPICS · SUPPORT & ACCOUNT IS WHERE REVIEWS AND QUERIES MEET · EVERY RING IS A LEVEL

Step 5. A pain point became a node, and the node became a product.

PAIN POINT

"Long wait"

1,037 review mentions, 2.39 stars. Said in stores, never typed in a search box.

NODE

Support & Account › Store Experience › Wait Time

Sits beside 'appointment', 'check in', 'store hours' in the same branch of the map.

PRODUCT

Virtual appointments. Digital check-in. A virtual assistant that resolves it before the store.

The wait happens on the phone instead of in the store. Revenue up, pain down, and the knowledge base already has the node the content needs.

Knowledge management with AI, using search: the sources are the user research, the taxonomy is the knowledge structure, and the product is what the structure makes obvious.

Four levels of structure

THE DEFAULT. GO DEEPER ONLY WHERE THE DATA FORCES IT: TELECOM TO L8, EDUCATION AND ACCESSORIES TO L10.

L1 — CATEGORY
structure only, no keywords
Devices & Hardware

17 NODES AT L1

L2 — SUBCATEGORY
structure only, no keywords
Samsung Galaxy

885 NODES AT L2

L3 — TOPIC BUCKET
the classification layer
Galaxy S24 Ultra

10,717 NODES AT L3

L4 — KEYWORD NODE
one keyword, its own volume
galaxy s24 ultra case

14,886 NODES AT L4

Classify at L3. Everything above it is navigation. Everything below it is evidence.

One long query, classified: what earns a level, and what does not

~~best~~ clear magsafe case ~~for~~ iphone 17 pro max under $30 near me for my daughter

INCLUDED · L1 TO L10, DEPTH EARNED BY EVIDENCE

- L1 Devices & Hardware
  - L2 Accessories
    - L3 Phone Accessories
      - L4 Cases & Protection
        - L5 Phone Cases
          - L6 iPhone Cases
            - L7 iPhone 17 Cases
              - L8 iPhone 17 Pro Max Cases
                - L9 MagSafe Cases
                  - L10 Clear MagSafe Cases

FACETS · FILTERS, NOT LEVELS

under $30	price facet
near me	place facet
for my daughter	audience facet: a gift

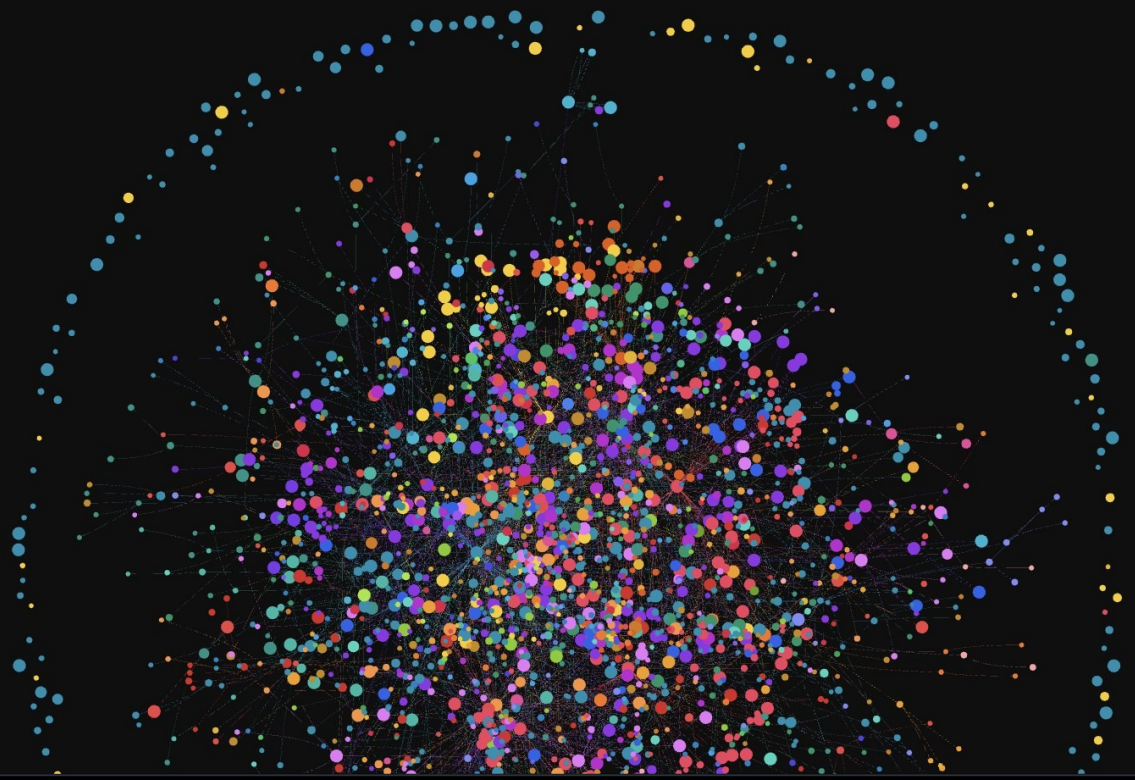
LINK · POINTS TO ANOTHER BRANCH

iphone 17 pro max	case levels borrow the name; the phone lives in Phones, linked by 'fits'

EXCLUDED · NO LEVEL, NO NODE

best, for	no meaning alone: kept as intent, never a node
Other, Misc, General	never created, at any level

Ten levels for the product, three facets for the person, one link to the phone, and two words that never become a node.



apple

- Apple
- Apple Services
- Apple TV
- Apple
- Apple Watch
- Apple Pencil

NODE INFO

Click a node to inspect it

COMMUNITIES

- [x] Select All
- [x] Mobile & Wireless Carriers 543
- [x] Devices & Hardware 192
- [x] T-Mobile 141
- [x] Communication Services 113
- [x] Network & Infrastructure 112
- [x] Plans & Pricing 102
- [x] Internet & Broadband (ISP) 99

Apps & Digital Services

TOPIC

57.1K	22.6M	35
KEYWORDS	VOLUME	TOPICS

MARKET INTELLIGENCE · B2C

$586M	$52M	$4.79
SEARCH VALUE/YR	CAPTURABLE	WTD CPC
99%	0.07	-4%
% COMMERCIAL	DIFFICULTY	MOMENTUM

Confidence: **Medium** · ad-equivalent value of search demand, not market revenue

telecom.apps_software

SUBTOPICS (35)

SUBTOPIC	KEYWORDS	VOLUME	TOPICS
Mobile Apps	25.9K	8.5M	20
Maps & GPS	6.6K	3.6M	19
Security & VPN Apps	2.1K	2.6M	8
Security VPN	7.3K	2.6M	20
Cloud Storage	7.8K	2.0M	18
Cloud Storage & Backup Apps	1.7K	1.4M	7

2,500 nodes, 2,449 edges, 40 communities: where a tree becomes a graph.

COMMUNITIES ARE DETECTED FROM SHARED KEYWORDS, NOT AUTHORED. 'T-MOBILE', 'VERIZON' AND 'SAMSUNG' SURFACE AS THEIR OWN COMMUNITIES BECAUSE CUSTOMERS TALK ABOUT THEM THAT WAY.

Four passes: simple rules first, AI last

PASS 1 · KNOWN NAMES

Look for names and words we already know, like 'iPhone 17'. A match goes straight to its topic. About 9 in 10 are settled here, each in under a millisecond.

Four passes: simple rules first, AI last

PASS 1 · KNOWN NAMES

Look for names and words we already know, like 'iPhone 17'. A match goes straight to its topic. About 9 in 10 are settled here, each in under a millisecond.

PASS 2 · WEB ADDRESS

When a carrier's own page shows up for a search, the address of that page tells us the topic. For example: /plans/unlimited.

Four passes: simple rules first, AI last

PASS 1 · KNOWN NAMES

Look for names and words we already know, like 'iPhone 17'. A match goes straight to its topic. About 9 in 10 are settled here, each in under a millisecond.

PASS 2 · WEB ADDRESS

When a carrier's own page shows up for a search, the address of that page tells us the topic. For example: /plans/unlimited.

PASS 3 · MEANING

Find the 8 topics closest in meaning. Pick one only if the match is strong. If not, the search waits. Waiting is better than guessing.

Four passes: simple rules first, AI last

PASS 1 · KNOWN NAMES

Look for names and words we already know, like 'iPhone 17'. A match goes straight to its topic. About 9 in 10 are settled here, each in under a millisecond.

PASS 2 · WEB ADDRESS

When a carrier's own page shows up for a search, the address of that page tells us the topic. For example: /plans/unlimited.

PASS 3 · MEANING

Find the 8 topics closest in meaning. Pick one only if the match is strong. If not, the search waits. Waiting is better than guessing.

PASS 4 · AI CHECKS

A few more rules catch most leftovers. Only the hard cases go to an AI model. It gives a one-sentence reason, and it may answer 'none of these'.

< 1 ms

TIME TO SORT ONE SEARCH

~90 / 10

SETTLED BY MATCHING / READ BY AI

10,600 h

HOURS TO SORT ALL 5.56 MILLION BY HAND

The label was written after reading the words, never before

WHAT PEOPLE TYPE	WHERE IT GOES
isp · internet provider	Internet & Broadband › Broadband
speed test	Internet & Broadband › Speed Tests
cost · price · how much	Plans & Pricing › Plan Comparisons
bill · payment	Plans & Pricing › Billing

The label was written after reading the words, never before

WHAT PEOPLE TYPE	WHERE IT GOES
isp · internet provider	Internet & Broadband › Broadband
speed test	Internet & Broadband › Speed Tests
cost · price · how much	Plans & Pricing › Plan Comparisons
bill · payment	Plans & Pricing › Billing
area code · phone code	Communication › Phone Numbers
sliding scale · free · afford	Insurance & Payment › Financial Assistance (healthcare)
therapist · therapy	Treatment Modalities › Evidence-Based Therapies (healthcare)

A RULE MOVES A NODE ONLY IF MORE THAN HALF ITS KEYWORDS MATCH · SPECIFIC BEFORE GENERIC · FIRST MATCH WINS · EVERY RULE IS BACKED UP BEFORE IT RUNS

What a rule-based first pass actually did

Nintendo Switch

404 keywords

→ Support › Switching

Ring camera

51 nodes

→ Communication › Ringtones

Santa tracker

204K searches a month

→ IoT › Connected Devices

tim mobile (Italian carrier)

11.1M searches a month

→ Devices, #1 keyword

What a rule-based first pass actually did

Nintendo Switch
404 keywords
→ Support › Switching

Ring camera
51 nodes
→ Communication › Ringtones

Santa tracker
204K searches a month
→ IoT › Connected Devices

tim mobile (Italian carrier)
11.1M searches a month
→ Devices, #1 keyword

family dollar near me
823K searches a month
→ Stores & Locations

republic services (waste)
673K searches a month
→ TV & Streaming

UPS package tracking
9.2M searches a month
→ IoT › Connected Devices

'skits a frenia'
673K searches a month
→ Substances › Abuse (healthcare)

Every taxonomy built from rules has a room like this. Most teams never open the door. The score opens it, and the queue tells you which door first.

One score: how much sits in 'Other'?

Score = 100 - the % with no real home

NO REAL HOME MEANS

- filed under 'Other' (counts twice)
- sitting in a 'General' bucket
- no subcategory at all

THE SCORE DOES NOT JUDGE THE NAMES. PEOPLE DO.

One score: how much sits in 'Other'?

Score = 100 − the % with no real home

NO REAL HOME MEANS

- filed under 'Other' (counts twice)
- sitting in a 'General' bucket
- no subcategory at all

THE SCORE DOES NOT JUDGE THE NAMES. PEOPLE DO.

THE LOOP · ONE RULE AT A TIME

1. **MEASURE** work out the score
2. **LOOK** find a word that repeats in a 'General' bucket
3. **PROPOSE** one rule: "internet provider" goes to Broadband
4. **TEST** apply the rule, then score again
5. **DECIDE** score up: keep the rule. score down: undo it.

One score: how much sits in 'Other'?

Score = 100 − the % with no real home

NO REAL HOME MEANS

- filed under 'Other' (counts twice)
- sitting in a 'General' bucket
- no subcategory at all

THE SCORE DOES NOT JUDGE THE NAMES. PEOPLE DO.

THREE 'GENERAL' BUCKETS · BEFORE AND AFTER THE LOOP

'General Internet'	1,072 topics → 40
'General Streaming'	1,072 topics → 23
'General Plans'	668 topics → 176

Rules are not the problem. Unchecked rules are.

THE LOOP · ONE RULE AT A TIME

1. **MEASURE** work out the score
2. **LOOK** find a word that repeats in a 'General' bucket
3. **PROPOSE** one rule: "internet provider" goes to Broadband
4. **TEST** apply the rule, then score again
5. **DECIDE** score up: keep the rule. score down: undo it.

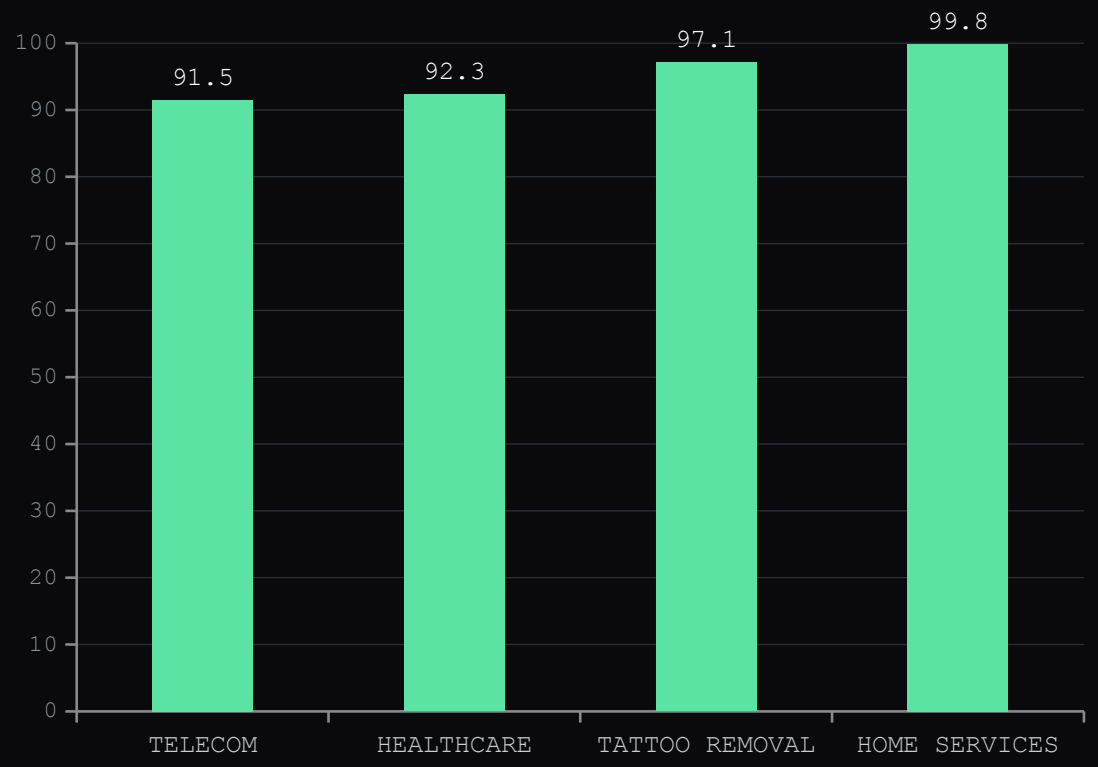
55

WHERE TELECOM STARTED: 36% SAT IN 'GENERAL' AND 'OTHER'

91.5

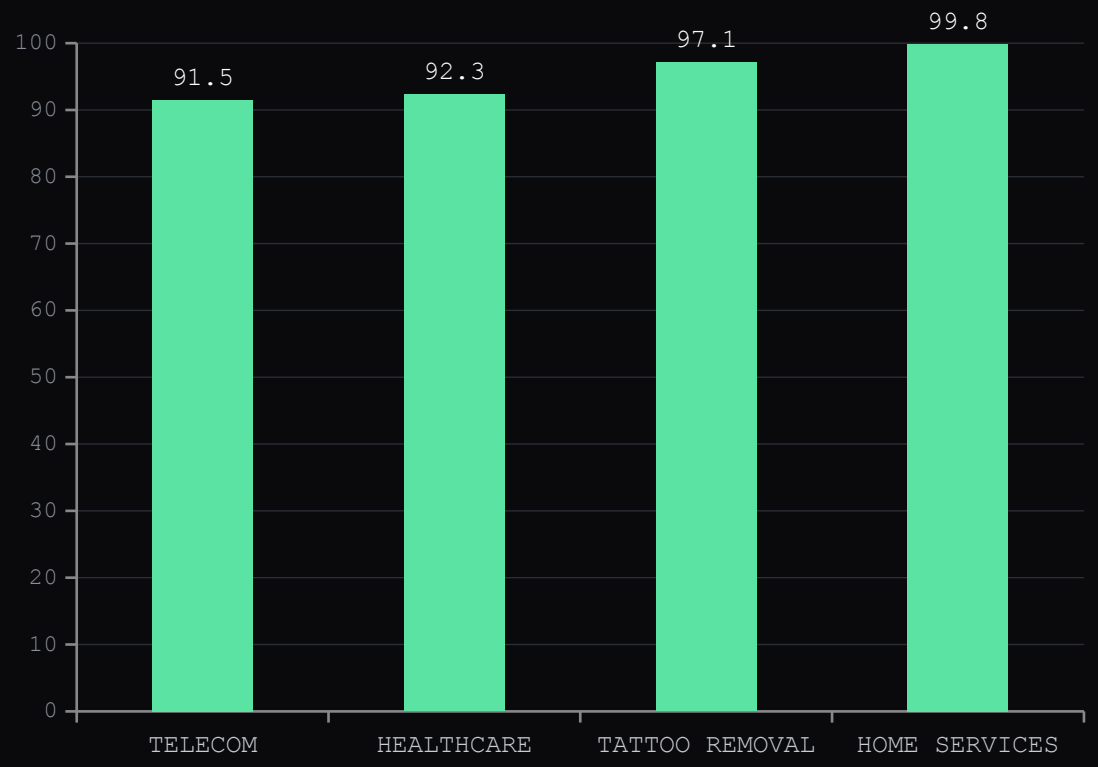
AFTER THE LOOP

55 to 91.5, and what the cleaning removed



SCORED CORPORA: TELECOM 2.9M KEYWORDS · HEALTHCARE 493K · HOME SERVICES 297K · TATTOO REMOVAL 38K. LARGER CORPORA CARRY MORE NOISE AND SCORE LOWER, AND THAT IS CORRECT.

55 to 91.5, and what the cleaning removed



SCORED CORPORA: TELECOM 2.9M KEYWORDS · HEALTHCARE 493K · HOME SERVICES 297K · TATTOO REMOVAL 38K. LARGER CORPORA CARRY MORE NOISE AND SCORE LOWER, AND THAT IS CORRECT.

~70M searches a month of phantom volume removed: an Italian carrier, a discount store, banks, hotels, utility outages

51.5K keywords re-attributed to the right topic

8,744 generic keywords pulled off carrier brand nodes, so brand counts stay honest

4,008 nodes moved out of generic buckets; 2,723 more corrected after the manual audit

22.4K predicted nodes with no demand yet: the content-gap map, on purpose

EVERY REMOVAL IS BACKED UP WITH A REASON AND RESTORABLE IN ONE COMMAND.

778,000

TELECOM KEYWORDS STILL IN 'OTHER UNCLASSIFIED'. THE LARGEST SINGLE NODE IN THE MAP.

It is on the map. It has a count. Everyone who opens the taxonomy sees it first. A score you cannot drive to 100 is a score you can trust; the day this bucket is empty, the score is lying.

ALSO STILL OPEN

- 'Other Samsung Galaxy' 55K
- 'Other Apple' 25K
- Labels like 'Mobile', 'Get', 'Plus' attract strangers
- Stores, Deals, Reviews re-homed under entities

30 / 100

AVERAGE SCORE OF 22 PARTS OF A 42,626-TOPIC TAXONOMY

70

CRITICAL PROBLEMS

104

SERIOUS PROBLEMS

1,959

DUPLICATE NAMES, 529 ACROSS CATEGORIES

8 to 46

LOWEST AND HIGHEST PART; EVEN THE BEST FAILED

The structure itself was rated near world-class. The mistakes were in the filling: searches in the wrong place, old and new versions mixed together, and things that did not belong.

Twelve kinds of mistakes. Each one became a gate.

F1 Matched on part of a word

'marriott code' was filed under 'VoIP codecs'

F2 Old and new versions mixed

the old layer still held 104,450 of 105,979 searches

F3 Same topics copied everywhere

21 topics copied onto 133 internet providers; 66% empty

F4 Levels out of order

443 level-4 topics sat directly under level 2

F5 Web pages copied in as topics

4,368 ZIP-code store pages under a topic named 'Bd'

F6 One giant catch-all bucket

one topic held 96% of its category

F7 Products that do not exist

90 Galaxy phones that were never made, S11 to S96

F8 Empty topics with no searches

88% of the Careers category was empty

F9 One word, several meanings

'Max' = CarMax, a Portland rail line, iPhone Pro Max

F10 Same brand in many categories

DISH appeared in 10 to 12 top categories

F11 Topic in the wrong category

consumer plans made up 65% of Regulatory

F12 Things outside the scope

2,200 China-only products with zero searches

All twelve can be prevented before they happen.

Twelve gates: pass them all, or nothing goes live

G1 — No duplicate names

at most 1% repeated; none under the same parent

G2 — One version only

never two structures for the same idea

G3 — Levels in order

every topic sits one level below its parent

G4 — Most sorted, none too big

at least 90% sorted; no topic over 10% of its category

G5 — Off-topic test

trap words (off-topic look-alikes) must stay unsorted

G6 — Only real things

every product and brand traces to a real source

Twelve gates: pass them all, or nothing goes live

G1 No duplicate names
at most 1% repeated; none under the same parent

G2 One version only
never two structures for the same idea

G3 Levels in order
every topic sits one level below its parent

G4 Most sorted, none too big
at least 90% sorted; no topic over 10% of its category

G5 Off-topic test
trap words (off-topic look-alikes) must stay unsorted

G6 Only real things
every product and brand traces to a real source

G7 No loose ends
no topic without a parent; leftovers at most 10%, truly off-topic

G8 No copy-paste topics
copied topics only where they apply; at most 30% empty

G9 No web pages as topics
no web addresses, ZIP pages, job posts or help IDs as topics

G10 Brands kept apart
at most 0.5% of one brand's searches under another

G11 The numbers add up
totals match exactly; at most 2% counted twice

G12 A person checks for gaps
a subject expert, 30 to 45 minutes per category

Governance is not a meeting. It is a gate the data has to pass through before anyone sees the change. Michele Jenkins takes this further at 3:15.

The score is layer one. Trust takes seven.

1	STRUCTURE	**Does every search have a real home?**	The score: 100 minus the % in 'Other' and 'General'.	TODAY
2	RULES	**Is the tree well formed?**	Twelve gates: no duplicate names, levels in order, no topic over 10%.	TODAY
3	MEANING	**Right meaning, or just the right word?**	A strong match in meaning, or the search waits.	TODAY
4	CONTEXT	**Same word, different meaning?**	'Max' = CarMax, a rail line, iPhone Pro Max. Trap words stay unsorted.	TODAY
5	PEOPLE	**Does an expert agree?**	A subject expert checks every category, 30 to 45 minutes each.	TODAY
6	WEIGHT	**How often is it right, where it matters?**	Precision and recall on an expert-labelled sample, weighted by search volume.	NEXT
7	OUTCOMES	**Does it work for real people?**	A/B tests: fewer dead-end searches, more clicks, pain points fixed.	NEXT

To improve it: change one thing, score before and after, keep what improves.

STANDARDS AND MEASURES · ISO 25964 · W3C SKOS · QSKOS (2012) · OOPS! (2014) · PRECISION, RECALL, F1 · COHEN'S KAPPA · A/B TESTS

One number to test and improve

NEXT NUMBER: SCORE V2 = % OF SEARCH DEMAND IN THE RIGHT TOPIC

Experts label a random sample. Each search counts by its volume. Sliced by month and by country.

RUNS TODAY · CASCADE: THE FIRST SURE ANSWER WINS

1. **RULE MATCH**
   Known names and words. About 9 in 10 searches settle here.
2. **SERP MATCH**
   The carrier page that ranks for the search names the topic.
3. **SEMANTIC SIMILARITY**
   Cosine similarity to each topic. A strong match, or it waits.
4. **AI CHECK**
   Hard cases only. Gives a reason. May answer 'none of these'.

NEXT · SIGNALS TO ADD

5. **DEMAND WEIGHT**
   Weigh every topic, and every error, by search volume.
6. **DEFINITIONS + SCHEMA.ORG**
   A definition per topic, mapped to a schema.org type. Score the label's fit (WiseOWL, 2026).
7. **TIME, REGION, POLICY**
   Season, country and local rules change what a word means.
8. **MODEL + RETRIEVAL TESTS**
   A decision model (Jev) rates each match. Retrieval tests: does AI search find the same topic?

Change one signal. Score before and after. Keep what improves.

WISEOWL: DALAL AND OTHERS, ARXIV 2604.12025 (APRIL 2026), A PRELIMINARY STUDY · SIMILARITY: OPEN EMBEDDING MODEL, COSINE, IN POSTGRES

Education, healthcare, home services

EDUCATION · ONLINE LEARNING

Built from 53 providers' live navigation: 13,059 observed labels, an 846-node proposal, then expanded.

61,429
NODES

35
TOP CATEGORIES

L10
DEEPEST, EVIDENCE-DRIVEN

10 of 10
QA CHECKS PASSED

Education, healthcare, home services

EDUCATION · ONLINE LEARNING

Built from 53 providers' live navigation: 13,059 observed labels, an 846-node proposal, then expanded.

61,429
NODES

35
TOP CATEGORIES

L10
DEEPEST, EVIDENCE-DRIVEN

10 of 10
QA CHECKS PASSED

HEALTHCARE · ADDICTION TREATMENT

Built from queries and ten years in this domain. Evidence-based labels, no 'rehab' catch-all.

44,714
NODES

15
TOP CATEGORIES

493K
KEYWORDS

92.3
QUALITY SCORE

Education, healthcare, home services

EDUCATION · ONLINE LEARNING

Built from 53 providers' live navigation: 13,059 observed labels, an 846-node proposal, then expanded.

61,429
NODES

35
TOP CATEGORIES

L10
DEEPEST, EVIDENCE-DRIVEN

10 of 10
QA CHECKS PASSED

HEALTHCARE · ADDICTION TREATMENT

Built from queries and ten years in this domain. Evidence-based labels, no 'rehab' catch-all.

44,714
NODES

15
TOP CATEGORIES

493K
KEYWORDS

92.3
QUALITY SCORE

HOME SERVICES · ROOFING

The first one. 316K raw keywords; 28 categories at first, cut to 12 once the data spoke.

14,427
NODES

12
TOP CATEGORIES

297K
KEYWORDS

99.8
QUALITY SCORE

THE ABSTRACT PROMISED TELECOM, EDUCATION AND HEALTHCARE. ALL THREE ARE HERE, BUILT THE SAME WAY.

Where it pays off: six use cases

1 NEW PAGES

A topic with demand and no page becomes a page type.

EXAMPLE

area code pages

2 PRODUCTS

A complaint that repeats becomes a product.

EXAMPLE

'long wait': 1,037 reviews → virtual appointments, digital check-in

3 SEGMENTATION

Every page gets a topic, so a huge site becomes segments you can measure.

EXAMPLE

40 million pages

4 LINKS + SITE SEARCH

Topics that sit together link together. Customers' other words become search synonyms.

EXAMPLE

"internet provider" finds Broadband

5 FEEDS

Readers pick topics and platforms, and the taxonomy builds their feed.

EXAMPLE

venkatapagadala.com/feeds

6 AUTOMATION

Rules first. AI for the hard cases. A loop keeps only what improves the score.

EXAMPLE

'General Internet': 1,072 topics → 40

One taxonomy. Six payoffs.

The checklist: one spreadsheet, eight steps

YOU NEED · A SPREADSHEET · YOUR OWN SEARCH EXPORTS · ONE SUBJECT EXPERT · OPTIONAL: ANY AI MODEL

1 Collect

Export the last 12 months of searches from two or more places: site search, Search Console, support tickets, reviews. One row per search, with a count and a source.

DONE WHEN every search is in one sheet, with a count and a source

EXAMPLE

```
"cheapest unlimited plan
 for 2 lines near me"
from: site search
```

2 Clean

Mark every search that is not about your business as noise. Count it and keep it. Do not delete it.

DONE WHEN every row says keep or noise

EXAMPLE

```
"nintendo switch" in a phone
company's searches = noise
```

3 Read

Sort by count and start at the top. For each row, write four things: the thing they ask about, the need behind it, what they want to do, and the extras.

DONE WHEN new rows stop showing you anything new

EXAMPLE

```
thing: unlimited plan
need: cheapest
intent: compare, then buy
extras: 2 lines, near me
```

4 Name

Group rows that ask about the same thing. Name each group with the words customers used most. AI can suggest groups. You choose the names.

DONE WHEN every row you read has a topic, named in customers' words

EXAMPLE

```
rows that say "speed test"
→ topic name: Speed Tests
```

The checklist: one spreadsheet, eight steps

SAME SHEET · STEPS 5 TO 8

5 Level

Arrange topics in four levels: category, subcategory, topic, search. Price, place and audience are filters. Give each its own column, never a level.

DONE WHEN each topic has a category and a subcategory; no 'Other'

EXAMPLE

Devices & Hardware › Samsung Galaxy
› Galaxy S24 Ultra › galaxy s24 ultra case
"near me" = place, a filter column

6 Sort

Sort the rest with simple rules: this word goes to that topic. Give AI only what no rule can place, and let it say "none". The rest stays unsorted.

DONE WHEN every row is sorted by a rule, by AI, or marked unsorted

EXAMPLE

rule: "internet provider" → Broadband
AI says "none" → stays unsorted

7 Score + check

Score it: 100 minus the % with no real home. Before you share, check: no duplicate names under one parent, no topic over 10% of its category, and an expert reads each category.

DONE WHEN the score is written down and every check passes

EXAMPLE

telecom score: 55 → 91.5
'General Internet': 1,072 topics → 40

8 Repeat

Add new searches on a schedule. Keep the unsorted pile visible, with its count. A word that keeps landing there is a missing topic: add it.

DONE WHEN the unsorted pile is smaller than last time

EXAMPLE

telecom's unsorted pile: 778,000,
shown first on the map

One ask: before you call a question dead, go find the person who asked it.

Today is step one. Here is the staircase.

1 TAXONOMY

A tree of categories.

```
Phones › Samsung
› Galaxy S24
```

WHERE YOU SEE IT

Online store navigation. This talk.

2 THESAURUS

One preferred name, its other words, and related terms.

```
"internet provider"
→ Broadband
```

STANDARDS

ISO 25964 (2011). On the web: W3C SKOS (2009).

3 ONTOLOGY

The kinds of things, and how they connect.

```
a case FITS a phone
a plan INCLUDES a phone
```

SHARED MODELS

schema.org: Google, Microsoft, Yahoo, Yandex (2011). Telecom: TM Forum SID.

4 KNOWLEDGE GRAPH

The real things, connected, at scale.

```
this case FITS
the Galaxy S24 Ultra
```

WHO RUNS ONE

Google (2012), Amazon, Apple, Microsoft, Facebook, eBay, IBM, LinkedIn

Each step stands on the one below. The bottom step is the names.

SOURCES · GOOGLE BLOG, 16 MAY 2012 · AMAZON AUTOKNOW, KDD 2020 · APPLE SAGA, SIGMOD 2022 · LINKEDIN ENGINEERING, OCT 2016
NOY ET AL., COMMUNICATIONS OF THE ACM, AUG 2019 (MICROSOFT, GOOGLE, FACEBOOK, EBAY, IBM) · ISO 25964-1:2011 · W3C SKOS · SCHEMA.ORG · TM FORUM GB922 (SID)
THE ORDER: THE LAST FOUR STEPS OF JESSICA TALISMAN'S ONTOLOGY PIPELINE (TECHNICS PUBLICATIONS, 2026)

Go deeper: start with these five.

TAXONOMIES AND THESAURI

The Accidental Taxonomist, 3rd edition

Heather Hedden · Information Today, 2022

Step by step, for beginners. From the publisher of this conference.

THE WHOLE STAIRCASE

Ontology Pipeline

Jessica Talisman · Technics Publications, 2026

Every step, from word lists to knowledge graphs.

ONTOLOGIES

Semantic Web for the Working Ontologist, 3rd edition

Dean Allemang, James Hendler, Fabien Gandon · ACM Books, 2020

How to model the kinds of things and how they link.

KNOWLEDGE GRAPHS AT WORK

Designing and Building Enterprise Knowledge Graphs

Juan Sequeda, Ora Lassila · Morgan & Claypool, 2021

How to build one inside a company. Juan gave this morning's keynote.

KNOWLEDGE GRAPHS AT SCALE

"Industry-scale Knowledge Graphs: Lessons and Challenges"

Natasha Noy and others · Communications of the ACM, August 2019

Lessons from the graphs at Google, Microsoft, Facebook, eBay and IBM.

KNOWLEDGE MANAGEMENT WITH AI, STARTING FROM SEARCH

Let customers name your categories.

They have already searched for it. Start there: sources, map, score, gates, people.

01 FOUR LEVELS

deeper only where the data forces it

02 A HEALTH SCORE AND TWELVE GATES

nothing published until every gate is green

03 AGENTS GATHER, PEOPLE DECIDE

every figure carries the sentence it came from

VENKATA PAGADALA · VENKATAPAGADALA.COM
LINKEDIN.COM/IN/VENKATA-PAGADALA · #TAXOBC

Figures are from Omniscite, the research platform I built, and from venkatapagadala.com. No vendor endorsements.

