

Connecting The Probable Dots

Disambiguating strategies for entity oriented search and natural language understanding

Dawn Anderson



@dawnieando

#FOS20

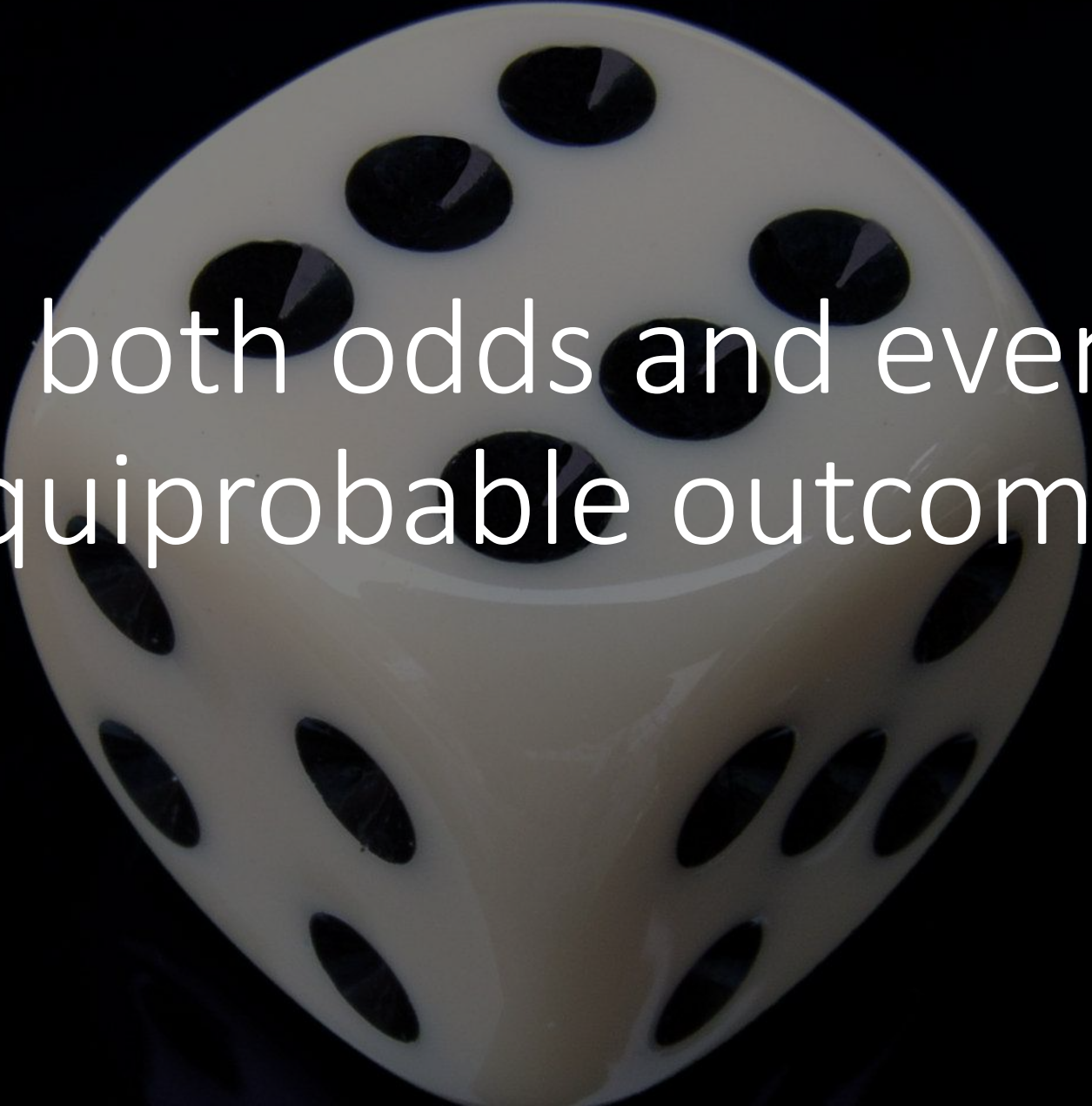




The probability
of rolling an even
number or an
odd number is
1:2



This is an
example of
'equiprobability'

A white, triangular die is centered in the image against a dark background. The die has black pips on its faces. The top face shows five pips, the front face shows four, and the right face shows three. The text is overlaid on the die.

Since both odds and evens
are equiprobable outcomes



“Assigns equal probabilities to outcomes when they are judged to be equipossible”
(Wikipedia)




In URLs, content, locations & entities equiprobability can be problematic



When one of a
number of pluralities
equally meets an
information need



A man with long, wavy brown hair and a serious expression is shown from the chest up. He is wearing a dark, fur-trimmed cloak over a dark tunic. He holds a sword with a silver blade and a dark hilt in his right hand. The background is a vast, hazy landscape with rolling green hills and distant mountains under a pale sky. The overall tone is dramatic and epic.

One must
beat the
other

THERE CAN BE ONLY ONE



This could be
duplicate content or
'ambiguity in
entities'





Equiprobability is like 'See saw SEO'

LaPlace's Principle of Indifference



la places principle of indifference



 All



Images



Videos



News



Shopping



More

Settings

Tools

About 8,260,000 results (0.53 seconds)

Principle of indifference philosophy

The **principle of indifference** states that in the absence of any relevant evidence, agents should distribute their credence (or 'degrees of belief') equally among all the possible outcomes under consideration. In Bayesian probability, this is the simplest non-informative prior.

en.wikipedia.org › wiki › Principle_of_indifference

[Principle of indifference - Wikipedia](#)

Search for: [Principle of indifference philosophy](#)



About Featured Snippets



Feedback

Where there is equi-
possibility there is a need for
further confirmations

Further 'confirmations'
beyond the same content or
surface form must be sought



Where two or more domain assets, surface forms, or entity determinations are considered equal a representative is 'the chosen one' (canonical)





Nearest Neighbours



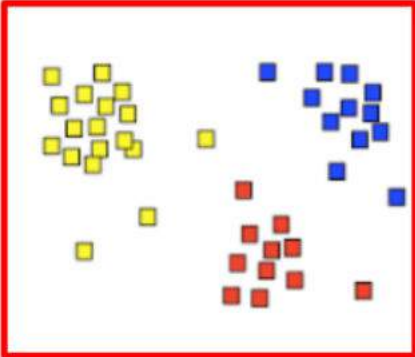
All the pluralities
are clustered
together and one
is picked

clusters

All Images News Shopping Videos More Settings Tools

About 492,000,000 results (0.99 seconds)

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (**clusters**).



[Cluster analysis - Wikipedia](#)
https://en.wikipedia.org/wiki/Cluster_analysis

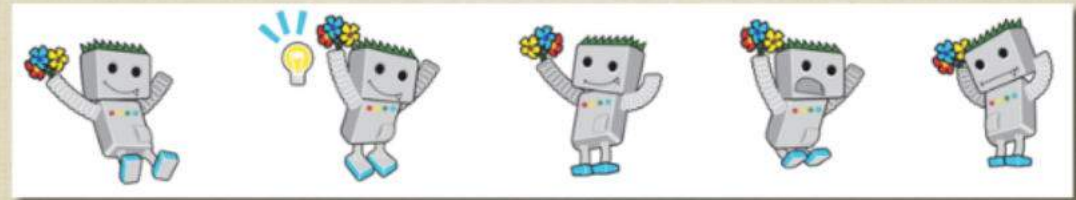
Sometimes 'bits'
are picked from
multiple
'equiprobable
outcomes'

Near Duplicate Content Can Cause Google to Choose Wrong Snippet

MARCH 14, 2017 AT 6:30 AM PST BY [JENNIFER SLEGG](#)

TheSEMPost

Near Duplicate Content Can Cause
Google to Choose Wrong Snippet





Today's Topic: Disambiguating
equiprobability for improved
search performance



WHO IS DAWN ANDERSON

 [linkedin.com/in/msdawnanderson](https://www.linkedin.com/in/msdawnanderson)

 @dawnieando

 dawn.anderson@[move-it-marketing.co.uk](mailto:dawn.anderson@move-it-marketing.co.uk)

11+ years SEO & Digital Marketing Consultant
& Pracademic:



Contributor:



Speaker & Trainer:



@dawnieando

#FOS20





WHO IS DAWN ANDERSON



[linkedin.com/in/msdawnanderson](https://www.linkedin.com/in/msdawnanderson)



@dawnieando



dawn.anderson@[@move-it-marketing.co.uk](mailto:dawn.anderson@move-it-marketing.co.uk)



move-it-marketing.co.uk



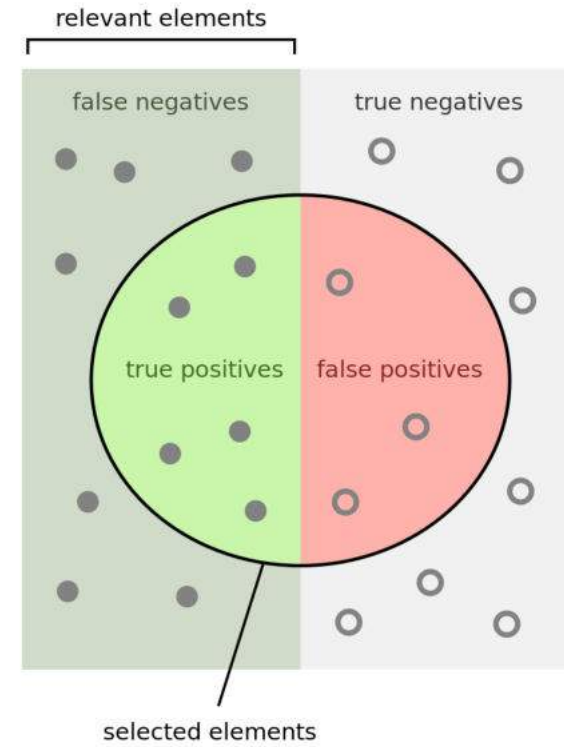
Meet Bert & Tedward

@dawnieando

#FOS20



In an ideal world
'precision' and
'recall' would
be perfect...
but... it's NOT



How many selected
items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant
items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

So search engines must
also work off probability
determination



Machine learning also
now powers some
probability
determination



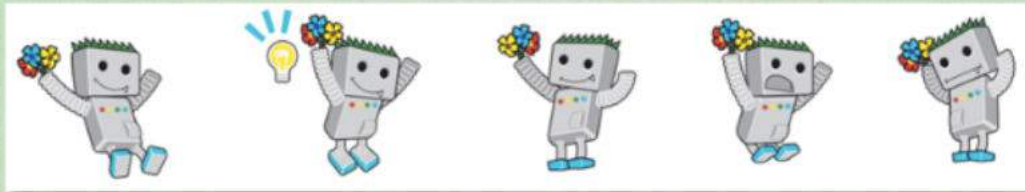
Google Uses Machine Learning for Crawling, Indexing & Ranking

Google Uses Machine Learning for Crawling, Indexing & Ranking

MAY 2, 2018 AT 7:20 AM PST BY JENNIFER SLEGG — [LEAVE A COMMENT](#)

TheSEMPost

Google Uses Machine Learning
for Crawling, Indexing & Ranking

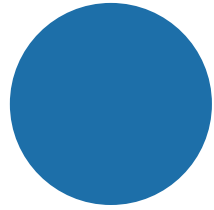
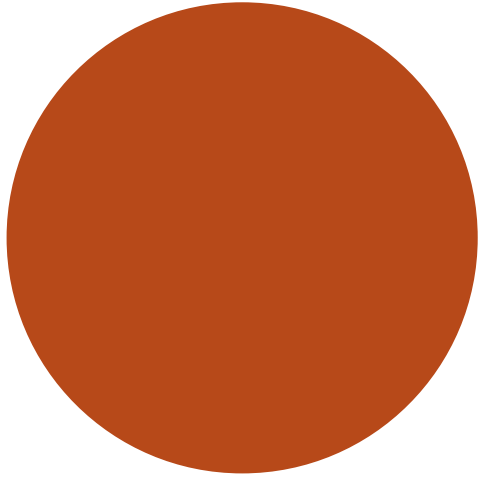




Using 'The
Law of Large
Numbers'



Progressive
learning &
increasingly
'educated'
guesses



Sometimes ‘probability’
predictions can be
unpredictable



On the 'long tail'
there's often
very little, or
nothing in it

Low to no page-rank pages
(but that is also now just
one of very many things

There will be lots of contributing factors now

1 - PAGE IMPORTANCE CONTRIBUTORS??

- Location in Site (e.g. home page more important than parameter 3 level output)
- PageRank
- Page type / file type ('about us' e.g. less important)
- Inclusion in XML sitemap (if others are excluded)
- Internal PageRank
- Internal Backlinks
- In-site Anchor Text Consistency
- Relevance (content, anchors and elements) to a topic (Similarity Importance)
- Directives from in-page robot and robots.txt management
- Parent quality brushes off on child page quality

**IMPORTANT PARENTS LIKELY
SEEN TO HAVE IMPORTANT CHILD
PAGES**

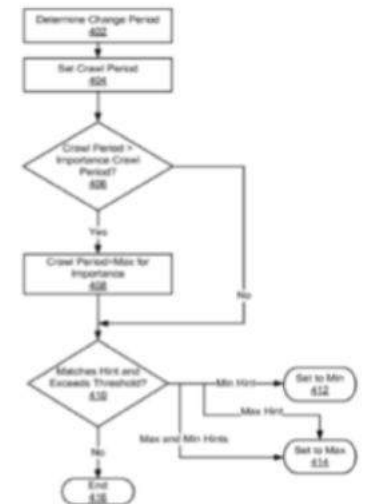


FIG. 4



See-saw SEO



Then the canonical
balance Tips From
Equiprobability
Slightly

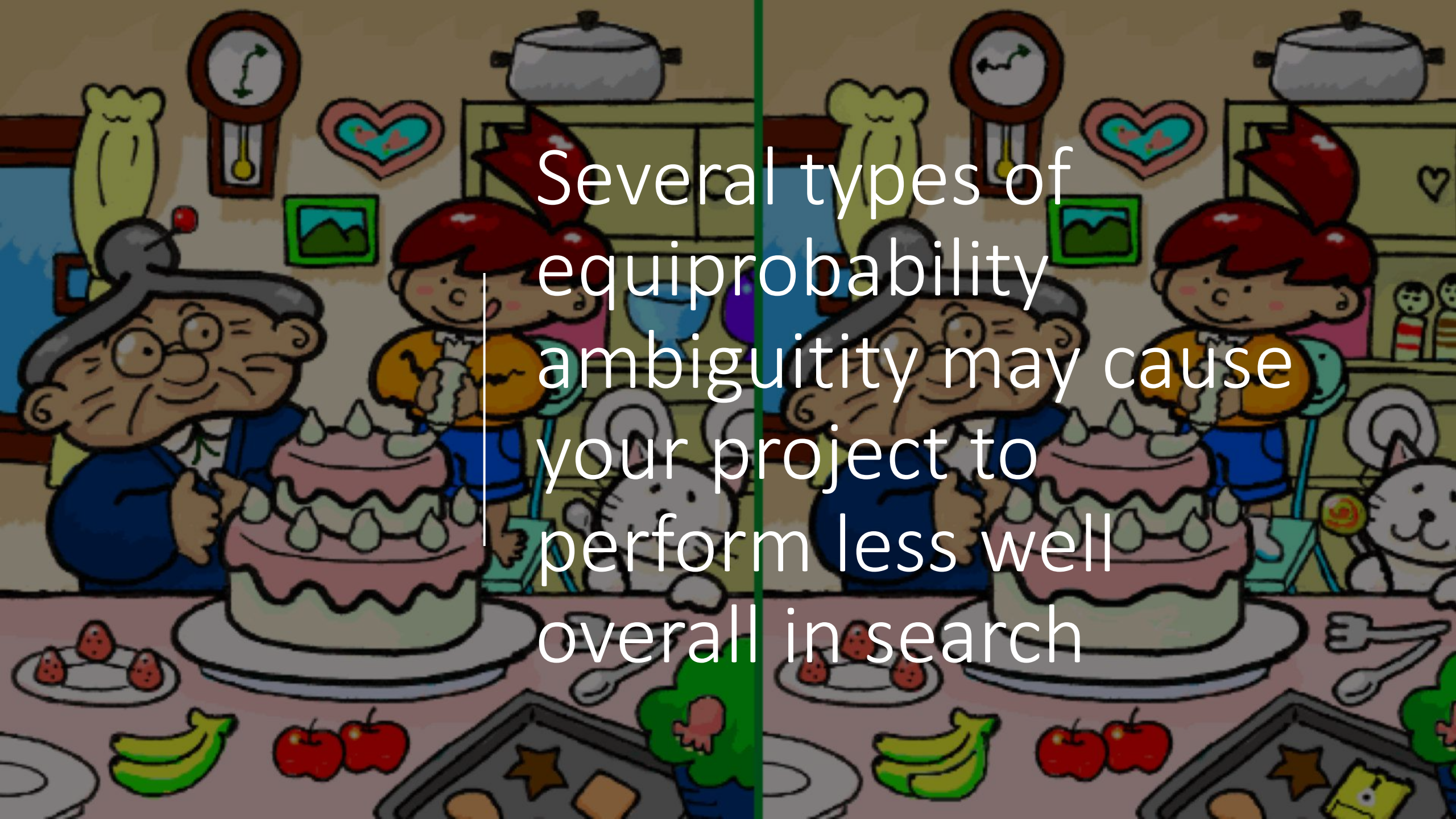


Where ambiguity
leads to
equiprobability
intermittently

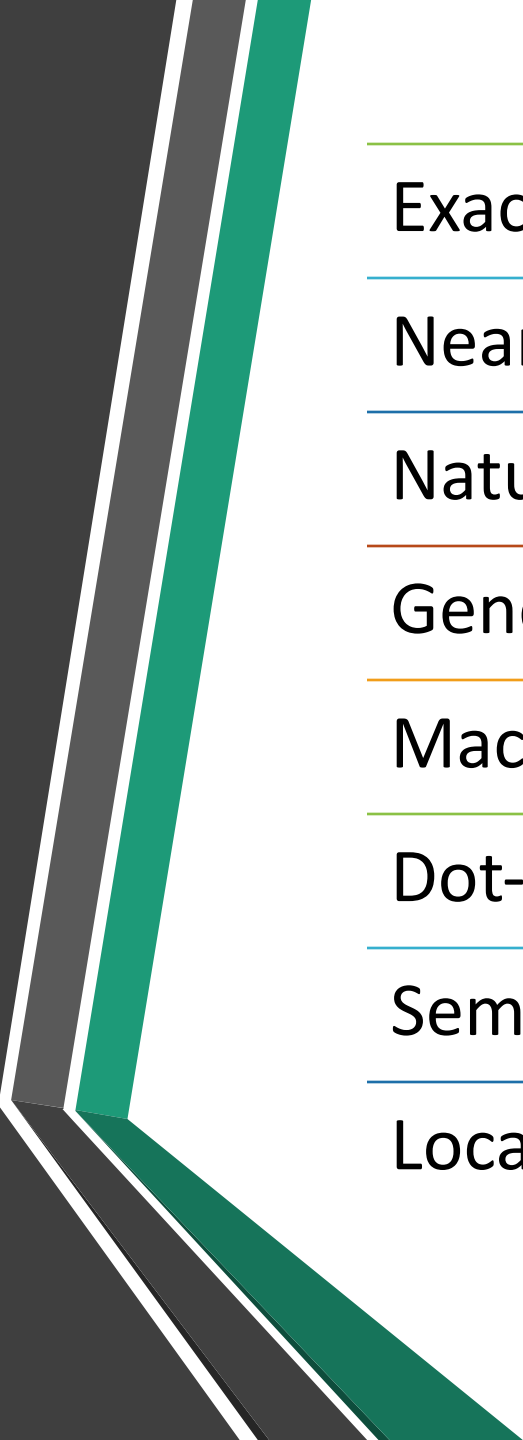




Your ranking flux might well be transferring equiprobability

A cartoon illustration of a kitchen scene. In the foreground, an elderly woman with grey hair and glasses is cutting a large, multi-layered pink cake with white frosting. A young girl with red hair is standing behind her, also looking at the cake. A white cat is sitting on the table to the right. The table is covered with various items: a plate of strawberries, a banana, two red apples, and a tray with star-shaped cookies. In the background, there is a kitchen counter with a pot, a clock, a heart-shaped picture, and a small framed picture. The scene is split vertically by a green line.

Several types of
equiprobability
ambiguity may cause
your project to
perform less well
overall in search



Different types of ambiguity can impact SEO

Exact duplicate content

Near duplicate content

Natural language ambiguity

Generational cruft based ambiguity

Machine learned ambiguity 'lag'

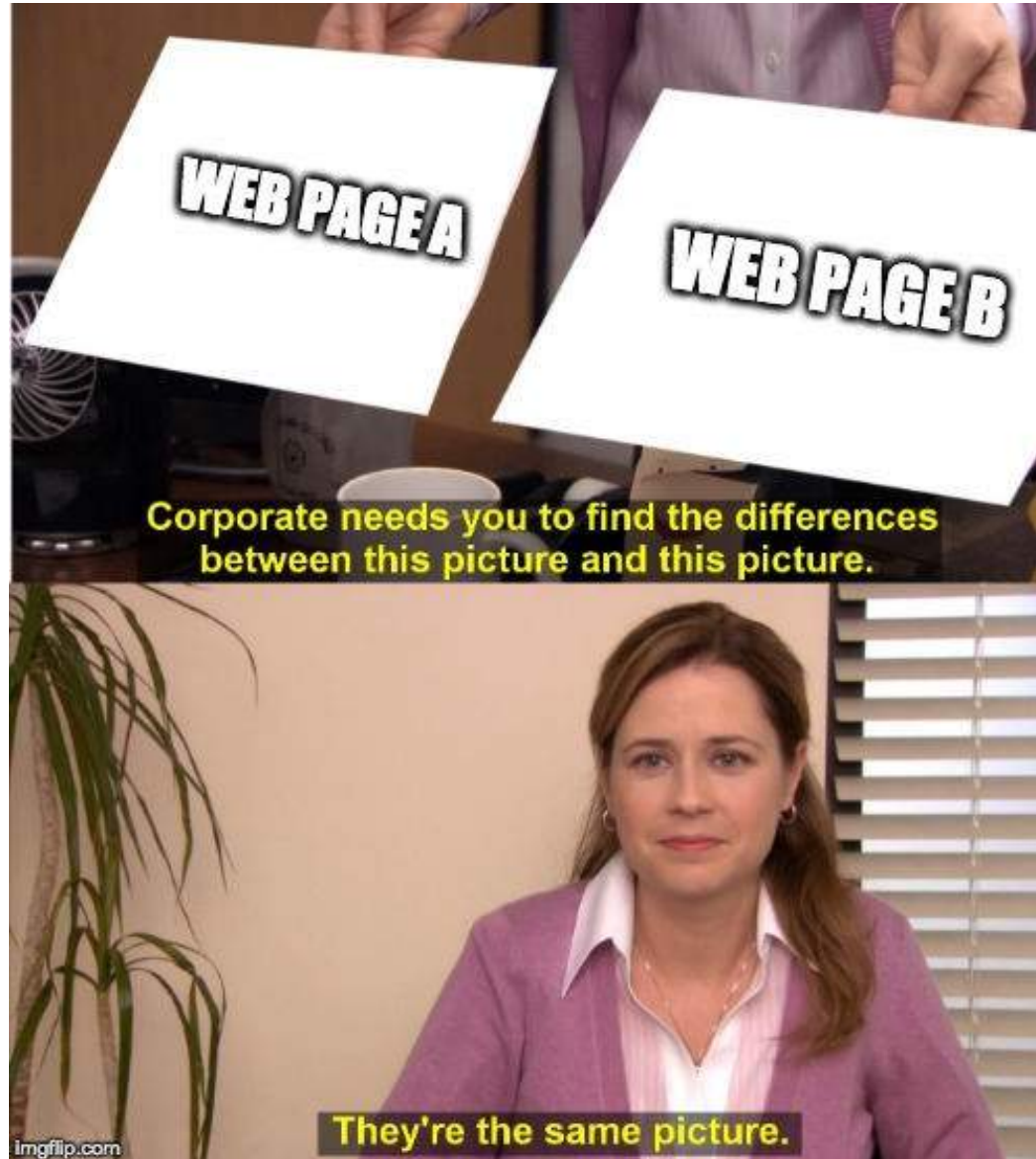
Dot-to-dot ambiguity

Semantic heterogeneity

Location based ambiguity

Exact duplicate
content

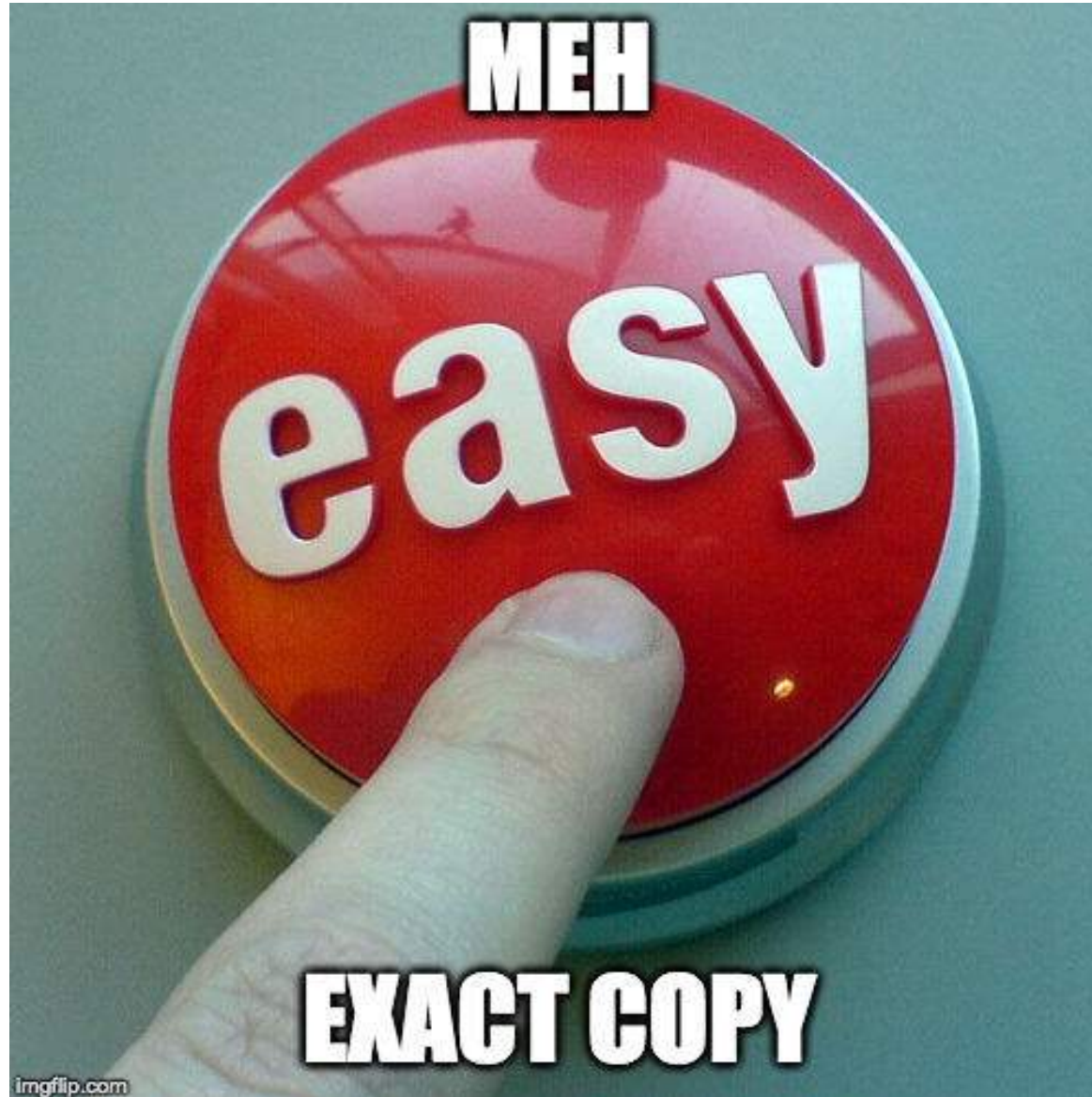




Corporate needs you to find the differences
between this picture and this picture.

They're the same picture.

imgflip.com



Google
probably
chooses the
one with
highest
'probability'

Most linked to internally

Most linked to externally

Included in XML sitemap

HTTPS

Prettier URLs

Maybe one or
several of many
'importance'
factors

1 - PAGE IMPORTANCE CONTRIBUTORS??

- Location in Site (e.g. home page more important than parameter 3 level output)
- PageRank
- Page type / file type ('about us' e.g. less important)
- Inclusion in XML sitemap (if others are excluded)
- Internal PageRank
- Internal Backlinks
- In-site Anchor Text Consistency
- Relevance (content, anchors and elements) to a topic (Similarity Importance)
- Directives from in-page robot and robots.txt management
- Parent quality brushes off on child page quality

**IMPORTANT PARENTS LIKELY
SEEN TO HAVE IMPORTANT CHILD
PAGES**

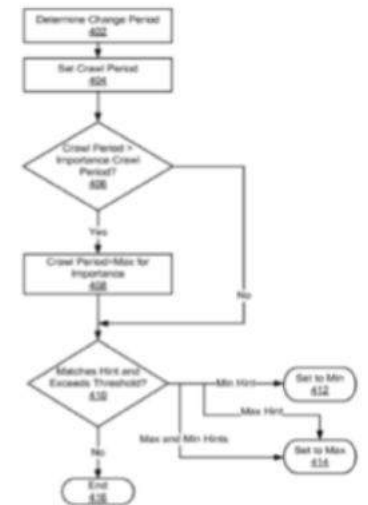


FIG. 4



Google works this
out VERY quickly



Near-duplicate
content







Easy(ish)



It starts off well
enough... but over
time



Google realizes
these pages are
'mostly' the
same as others

Some
symptoms

Many matching shingles

Quilting

‘Borrowed’ content at scale

Feeds

Big Boilerplate / little main value

Data driven sites with little value



Sometimes these are sites trying to make a URL footprint well beyond their contribution to a positive 'Network Effect'

Spoofing the 'Network Effect'

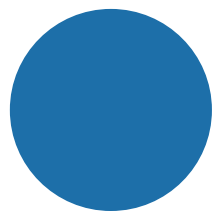
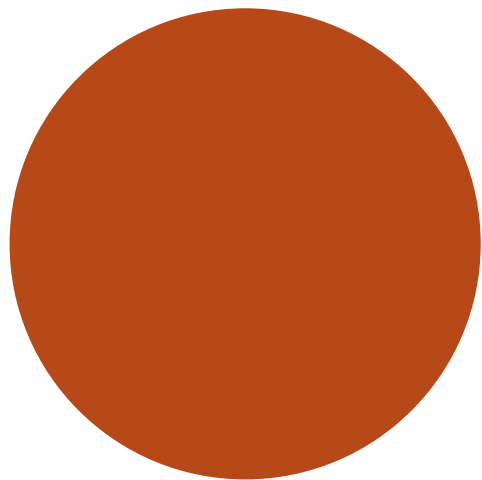




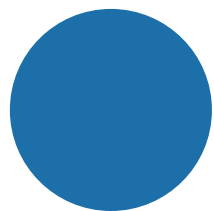
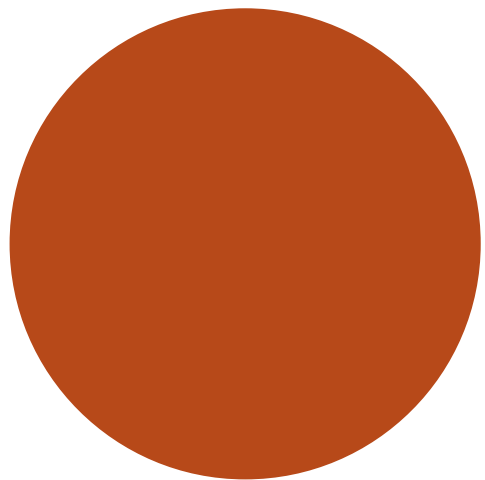
After some crawling
& sampling Google
puts the pieces of
the puzzle together



And begins to
exclude pages
from the index



Exclusion still
happens... A LOT



Particularly on
behemoth sites

I ran a little
Twitter
competition
with only
empathy as
the prize



Dawn Anderson
@dawnieando



Competition time:

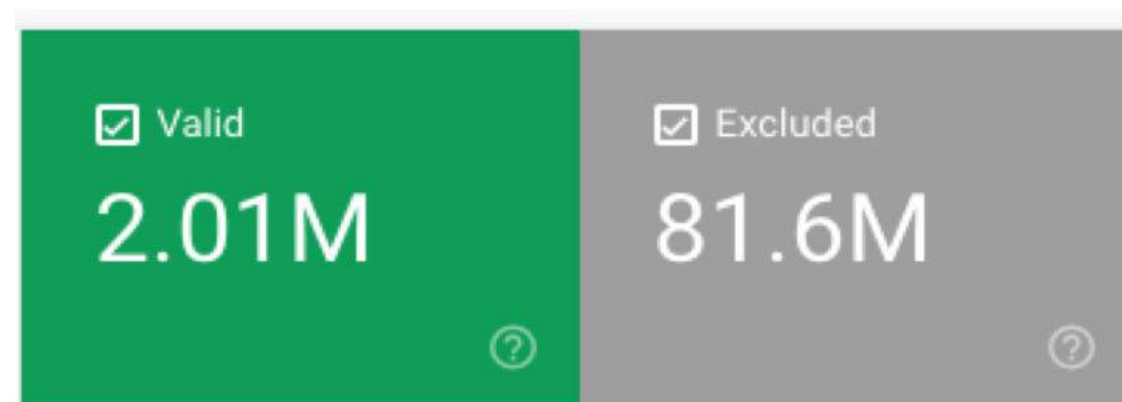
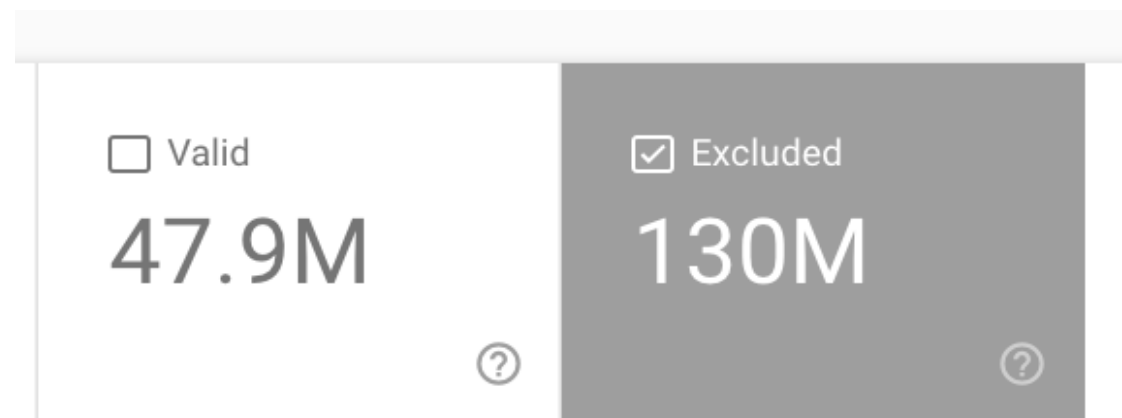
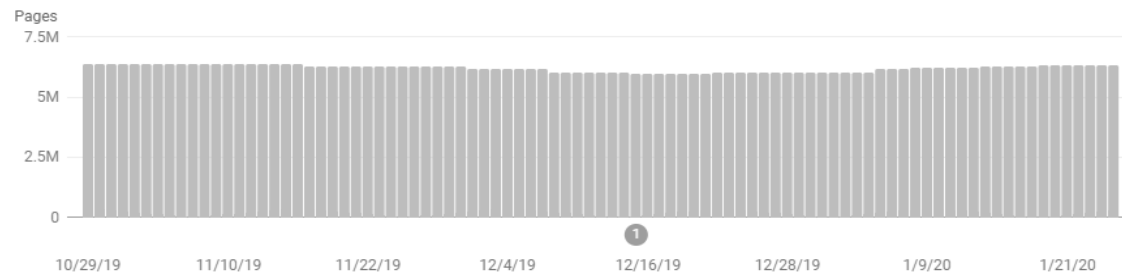
Who's got the highest number of 'Excluded' in any
GSC Coverage Report.

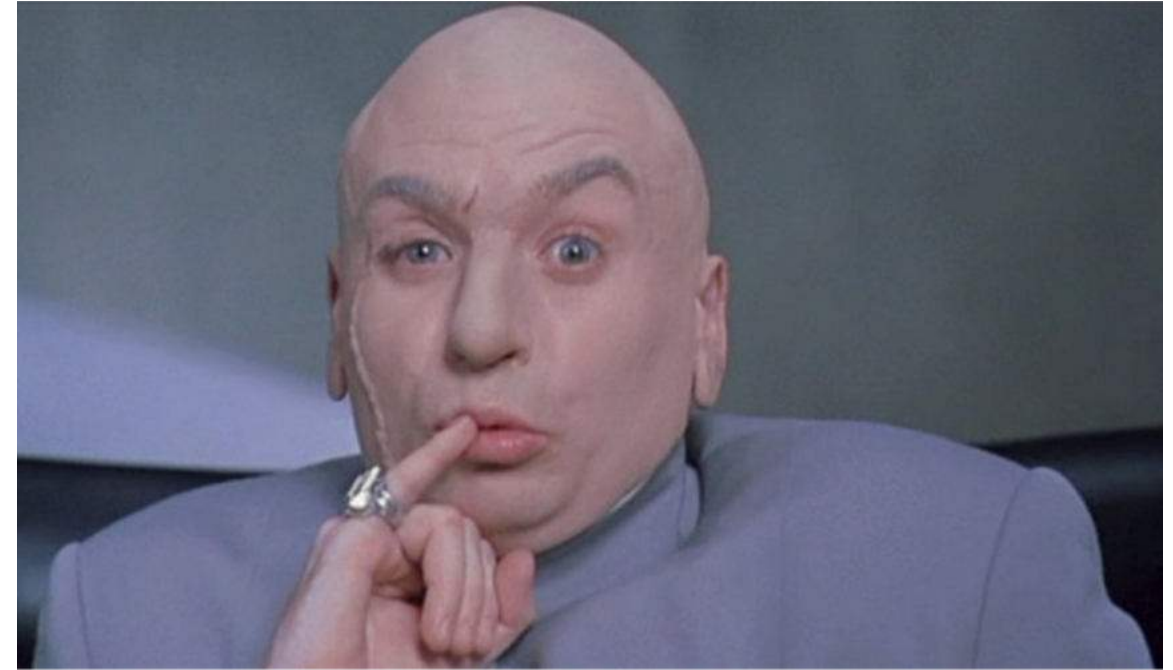
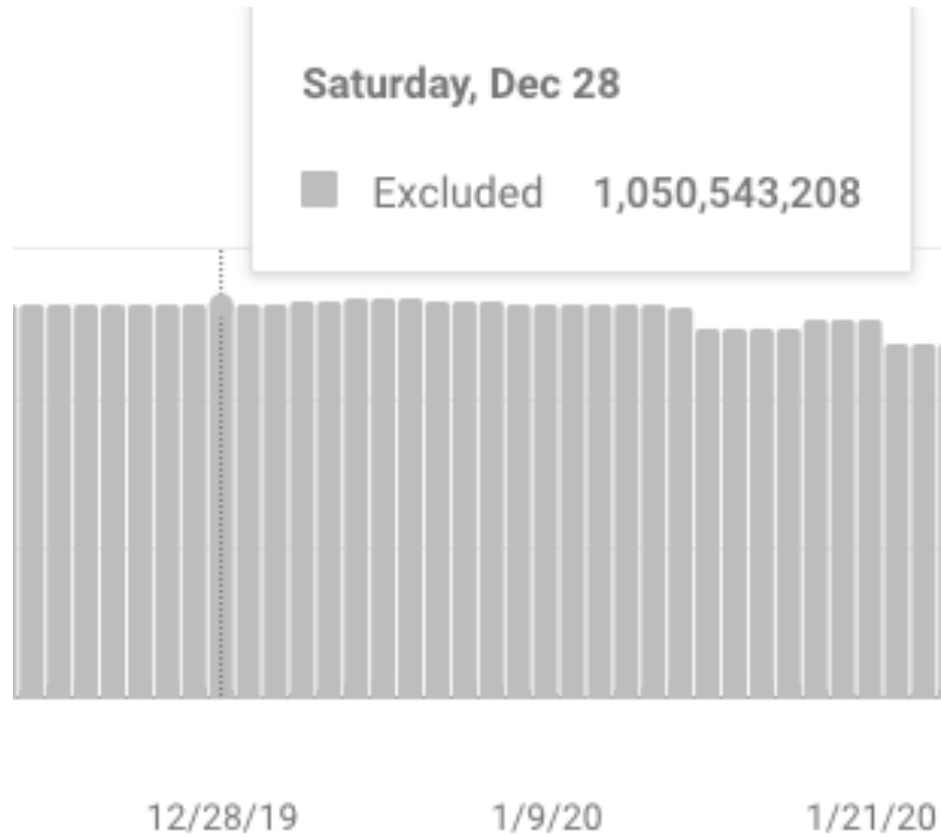
Show me the screenshots.

No prize for winning. Just the glory and empathy for
the work ahead ;P ;P

11:00 AM · Jan 27, 2020 · [Twitter Web App](#)

Some of the
GSC
'exclusion'
numbers
were huge...
millions





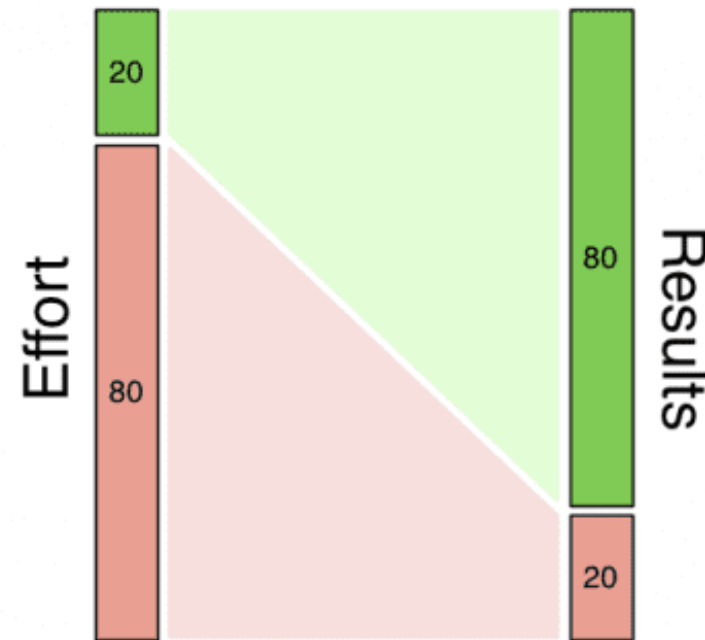
And even over... one billion

I would hazard a guess...The majority of the largest websites in the world have a LOT of near-duplicate 'transactional' pages

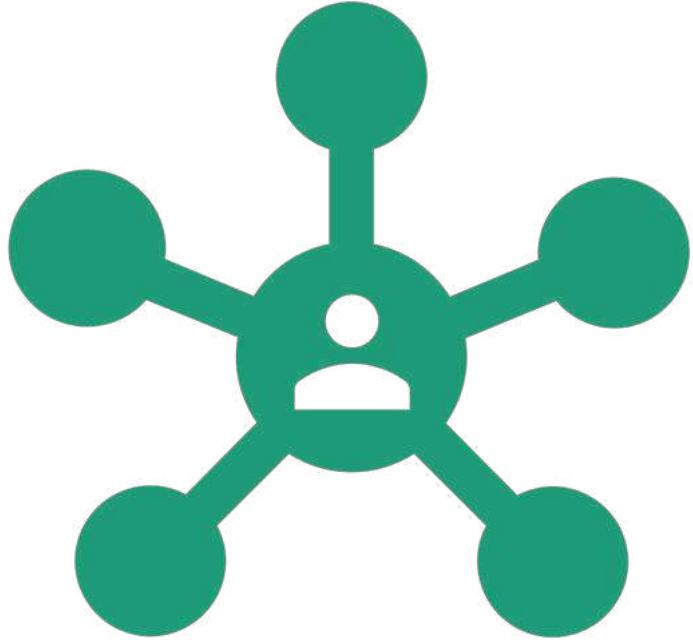
Crawl demand
probably
follows the
Pareto Principle

The 80-20 Rule

"For many events, roughly 80% of the effects come from 20% of the causes." - Pareto



Therefore 20% of the effort produces 80% of the results but the last 20% of the results consumes 80% of the effort.

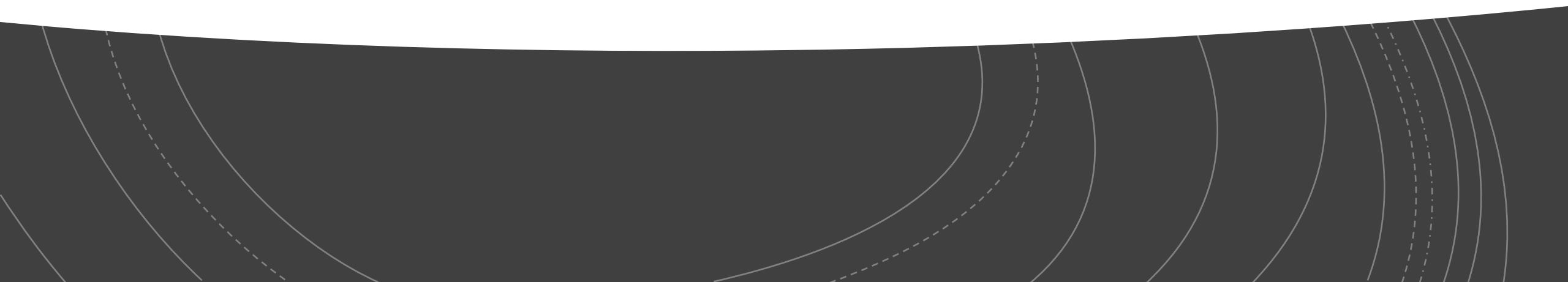


Probably 20% of the
URLs satisfy 80% of
the demand &
contribute well to
'The Network Effect'

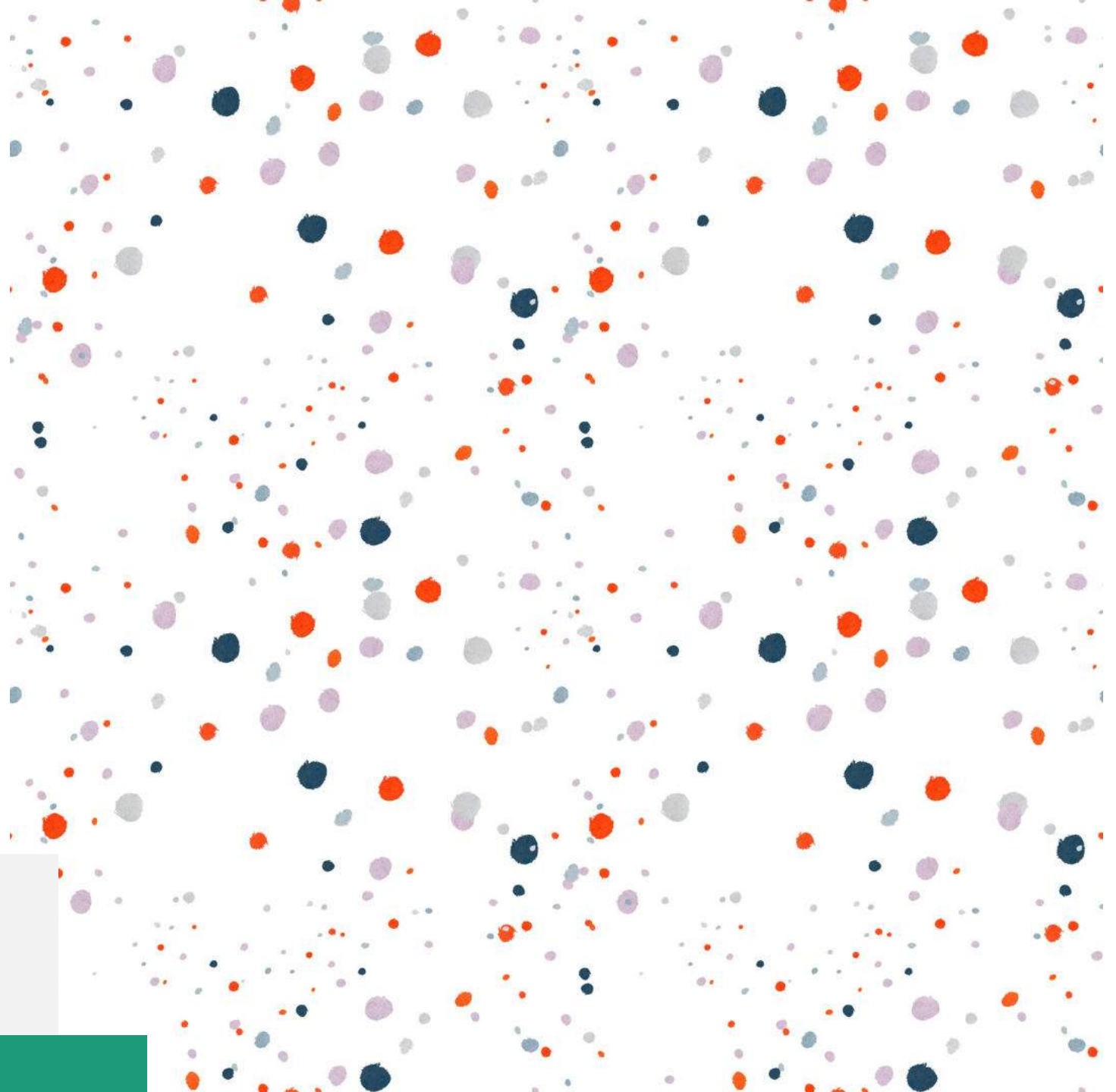
Since search engines
probably work with positive
'impact' value considered



Google will likely use
mostly 'Sampling' on those
heavy 'excluded URL' sites



Looking for small
samples of
content & URL
patterns... and...
‘Discovered, not
crawled’ will be
high



'Discovered,
not crawled'
in GSC is
Google
saying...



They know what's
down that site
section or URL
parameter path



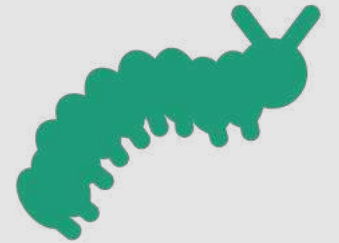
The verdict is
in

THEY KNOW IT'S A



LOAD OF POOP

Those URLs will be
'starved' of crawl



However many
times you submit
and inspect





You have your work cut out
once Google puts the pieces
of that puzzle together



Natural
language
ambiguity



For machines words are
problematic. Ambiguous...
polysemous... synonymous

In spoken word it is
even worse because of
homophones and
prosody

Like “four
candles” and
“fork handles”

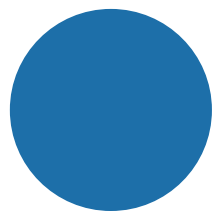
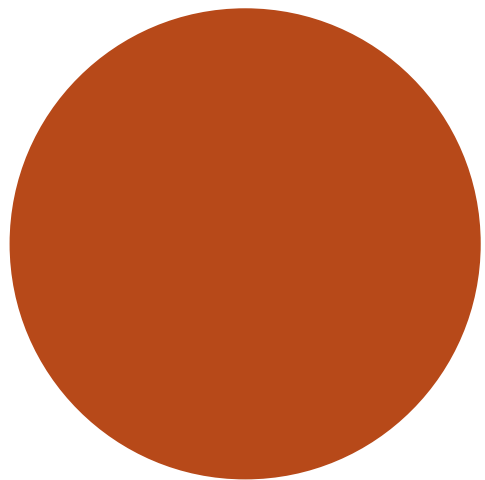


Many words have
multiple meanings.
Like “like” can be 5
possible parts of
speech (POS)



Word's context helps
enormously

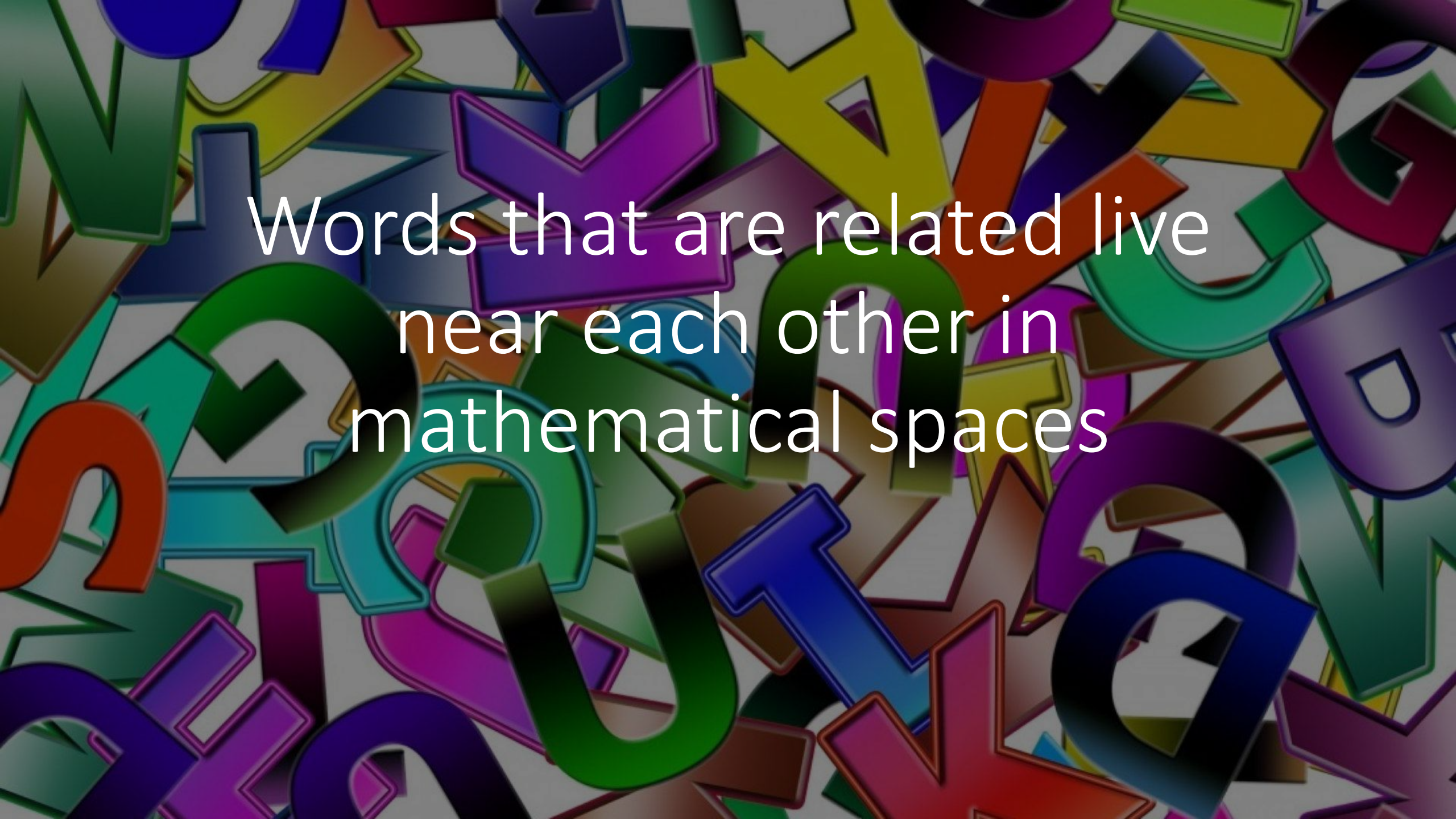




And disambiguating
words is getting easy(ier)



“You shall know a
word by the
company it keeps”
(Firth, 1957)

The background is a dense, chaotic arrangement of 3D letters and symbols in various colors including blue, green, yellow, orange, red, purple, and teal. The letters are of different sizes and are scattered across the entire frame, creating a textured, almost abstract look. The text is centered over this background.

Words that are related live
near each other in
mathematical spaces

There are HUGE leaps forward in
natural language understanding
now using machine learning

Accelerated mostly
by Google's BERT
(Bi-Directional
Encoder
Representations
from Transformers)



Co-
occurrence
in content
helps A LOT

In the page itself

In the interconnected pages

In the subcategorisation

In the site sections

In the external relationships

In the domain ontology as a whole

Relatedness



Informational content
adds to the contextual 2nd
level relatedness of
transactional content

The diagram features a green speech bubble on the left containing text. To its right are two dark blue rounded rectangular boxes, one labeled 'Informational content' and the other 'Transactional content'. These two boxes are connected by two curved lines, one above and one below them, suggesting a relationship or flow between the two types of content.

Informational
content

Transactional
content

If you can help
with
'informational
needs' you can
'probably' help
with
transactional
needs

Add value in rich
informational
needs



Contextual value
passes throughout
the whole site

Adding
Contextual
Value Does
Not Mean

Adding

Adding loads of content below
ecommerce pages

Adding

Adding LSI keywords in ecommerce
pages

High topical
relatedness
& context
vectors

Internal linking

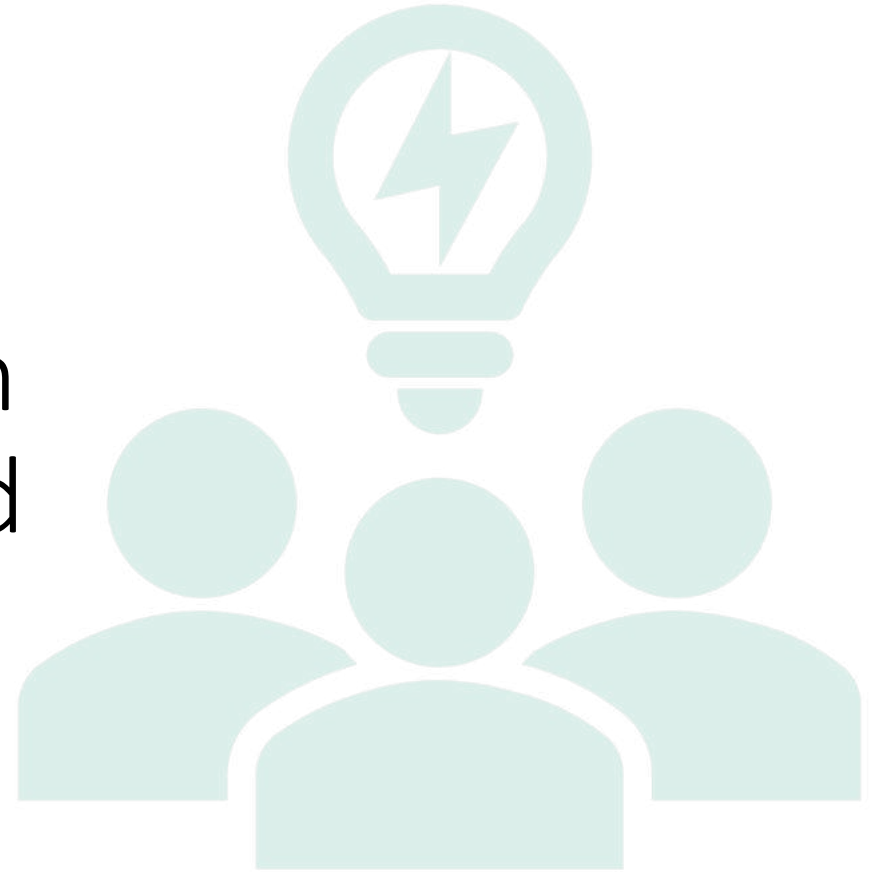
Not spam



Utilise conceptual
'nearest neighbours'
well



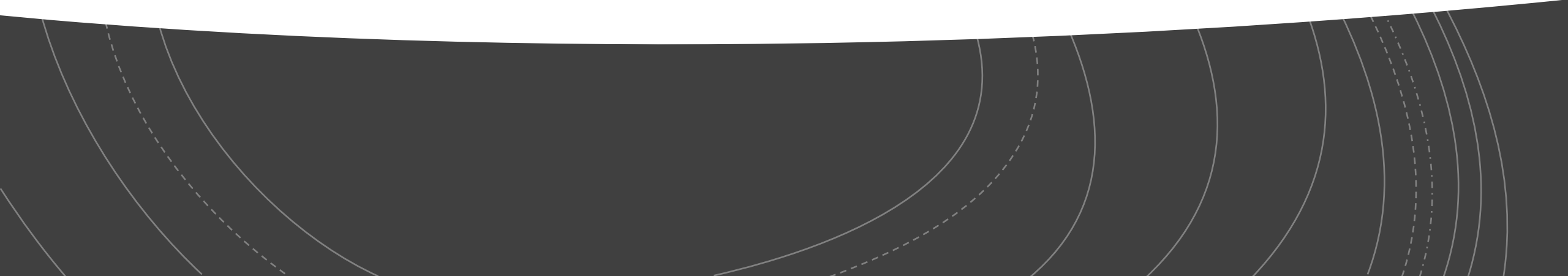
Identify content which
is 'very' closely related





Merge content
which is 'too'
conceptually
very similar with
no separate
demand

Utilise 'Overflow SEO' as
demand & content
naturally grows



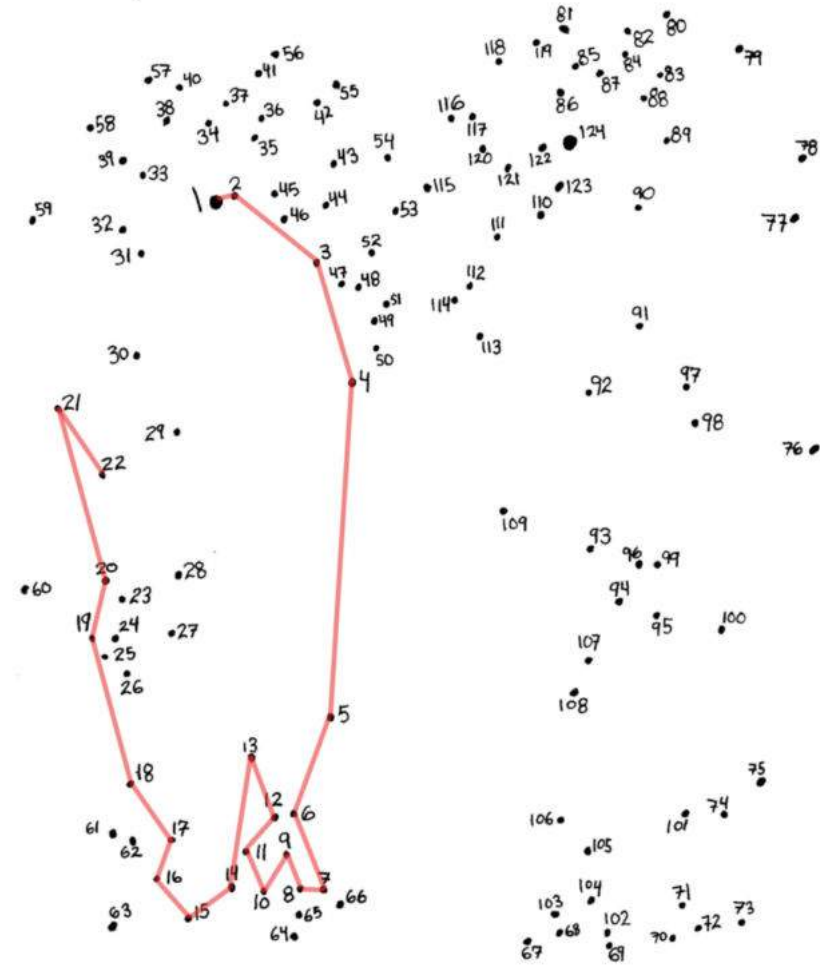


Google may realise you
can 'probably' help with
transactional queries

Dot-to-dot
ambiguity



Who knows what
this dot to dot
puzzle is creating?



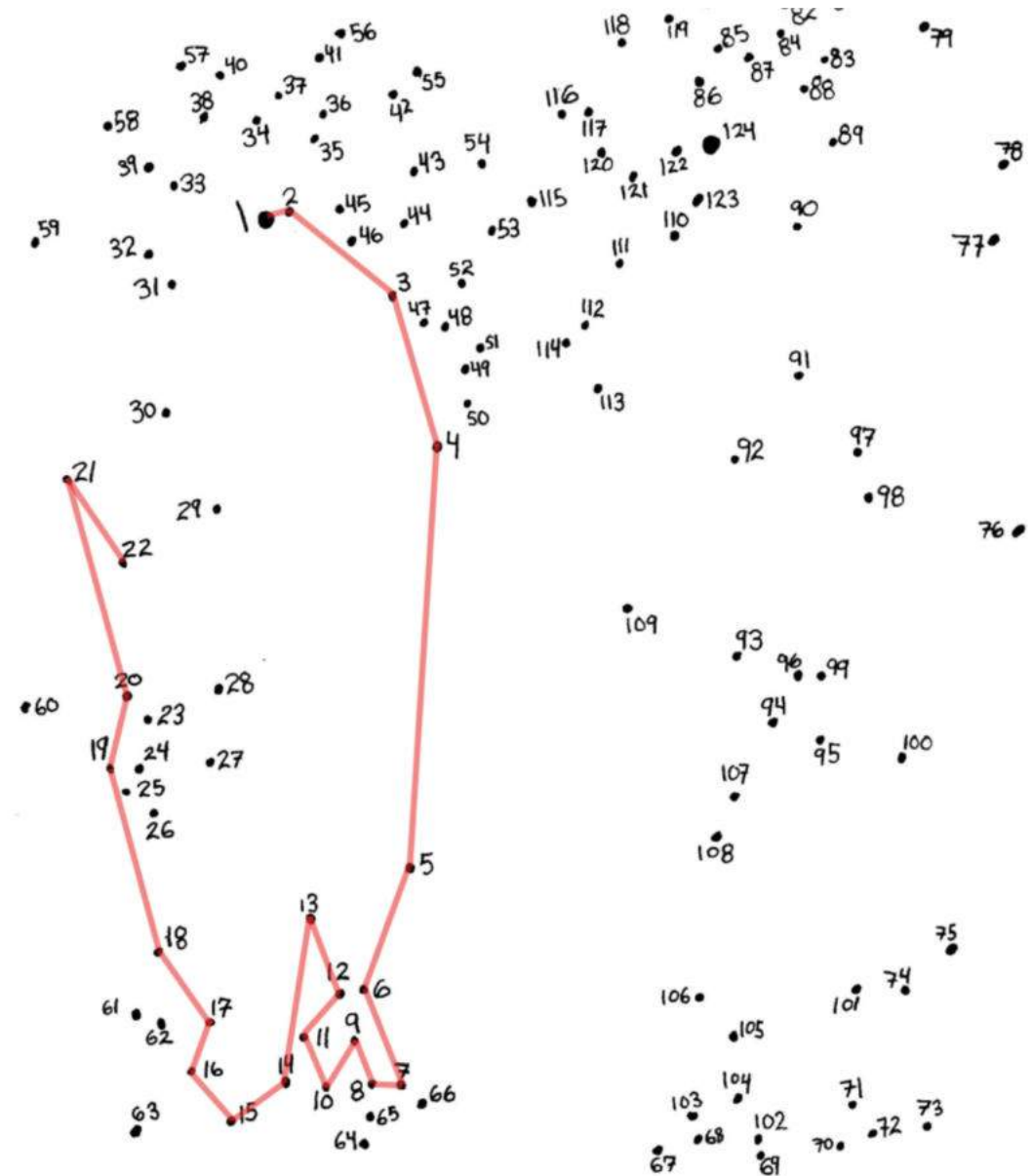
Absolutely
no-one



If you chop away or migrate
half of your website then
equiprobability will likely be a
problem



A half drawn website
is not the same as a
fully drawn previous
website?



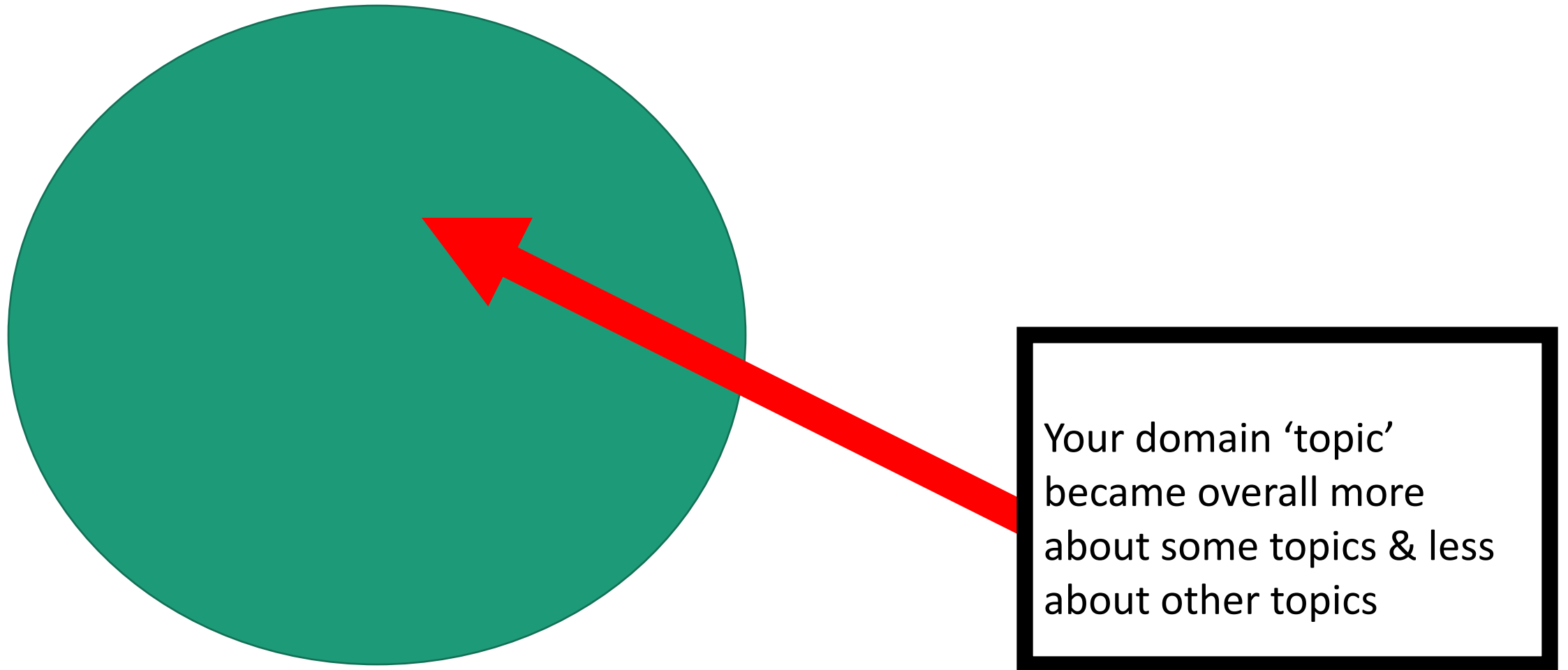
**Did you just
'prune away'
your corpus
'relatedness'?**



Informational
content adds
to the
contextual 2nd
level
relatedness of
transactional
content



Although you may rank 'marginally' better for fewer things



Prune, Merge, Improve, Archive

01

Prune – Only
ballast

02

Merge – Highly
semantically
related content

03

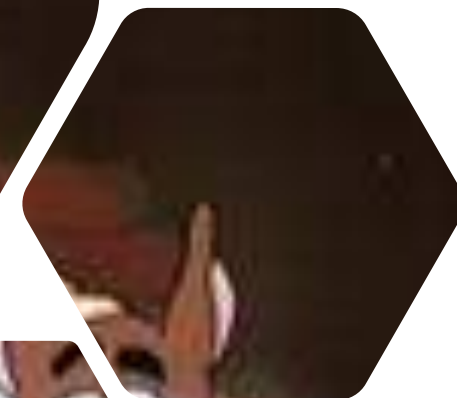
Improve –
Evergreen pages
worth the effort

04

Archive – Older
pieces of
temporal content
(like a library)

Avoid removing or
hiding user
generated content
(except spam)





Google puts the pieces of
a half-site puzzle together
(often with bad results)



Cruft based
ambiguity





Generational cruff
contributes to
equiprobability
issues

Just a few
examples...
legacy is a
problem

301 redirect chains

Inconsistent 301s

No 301s at all

Canonical points to a 301

Canonical points to a no-index

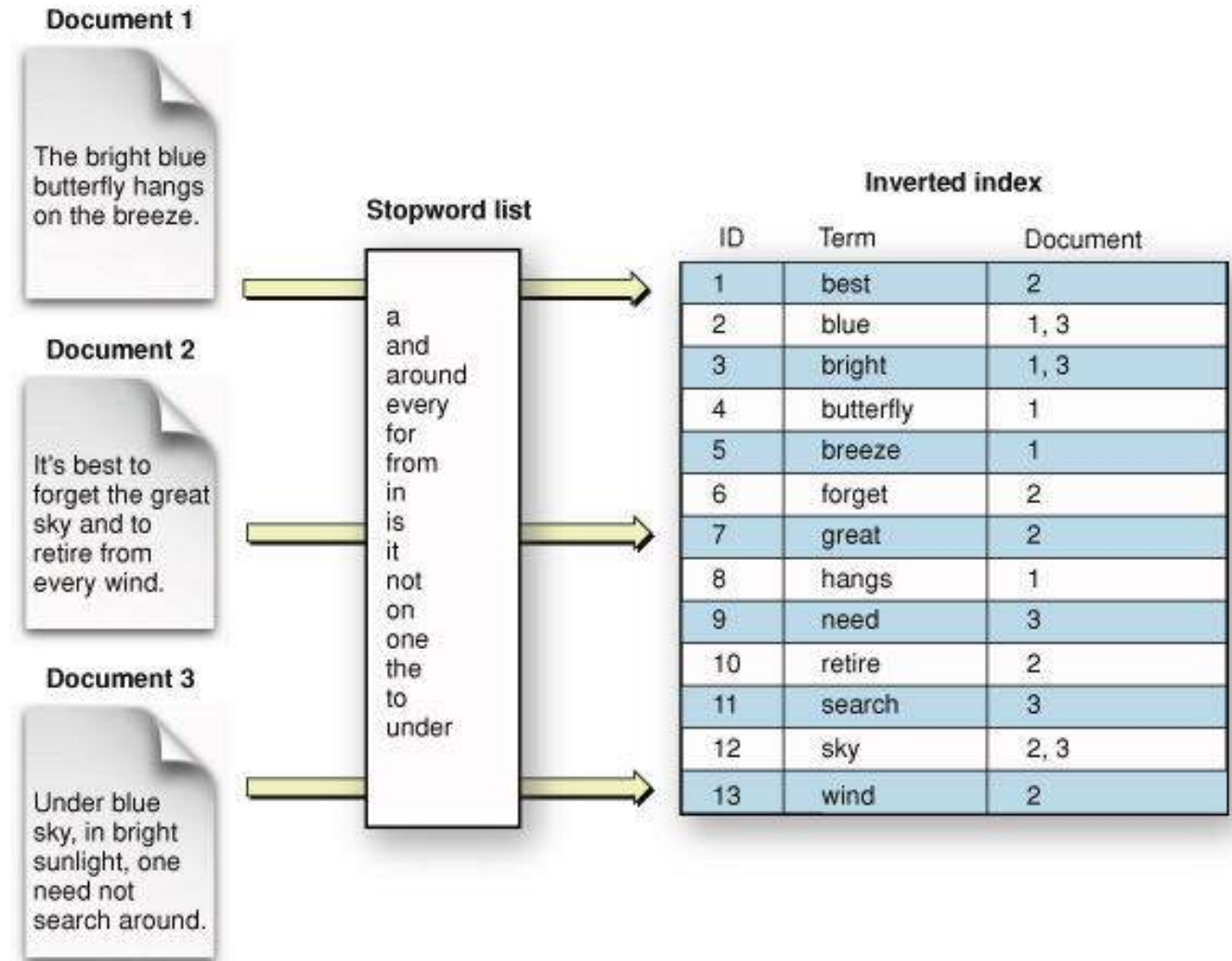
No-index points to a canonical

301 Should
Mean The
Resource
Moved



You're supposed to be
telling search engines
where the words
(tokens), topics,
concepts & entities went

You're saying
things moved
so they can
be re-filed



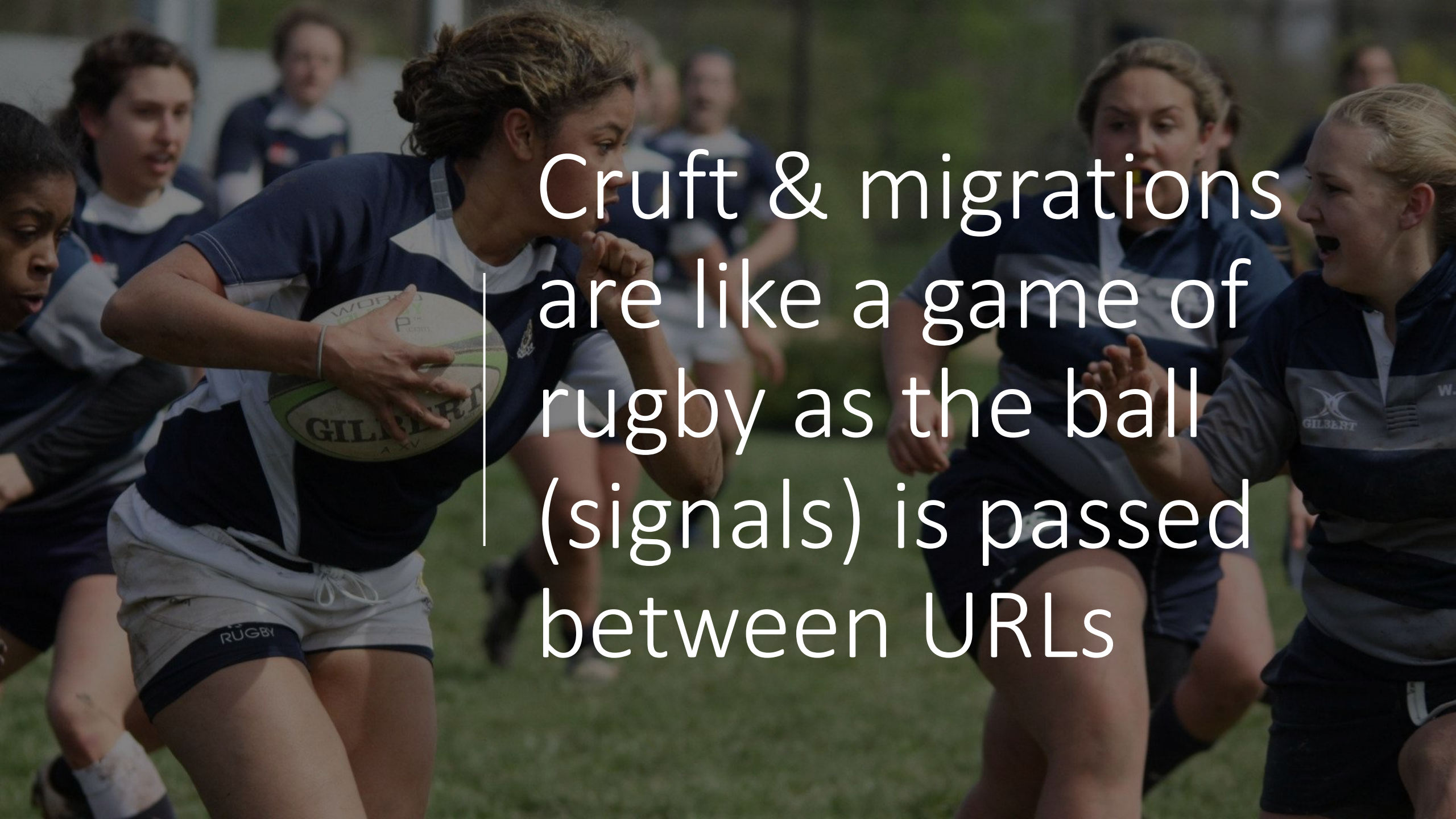
But often
there is little
or no match
at all



Borderline or
true 'soft 404'

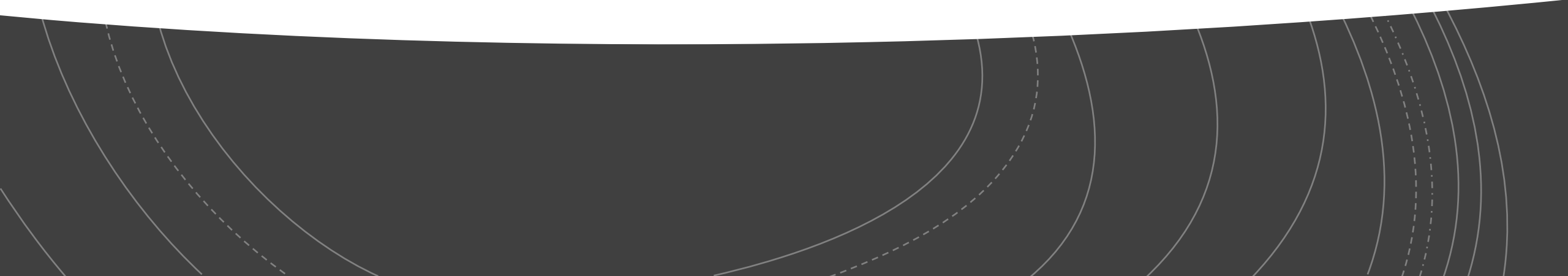


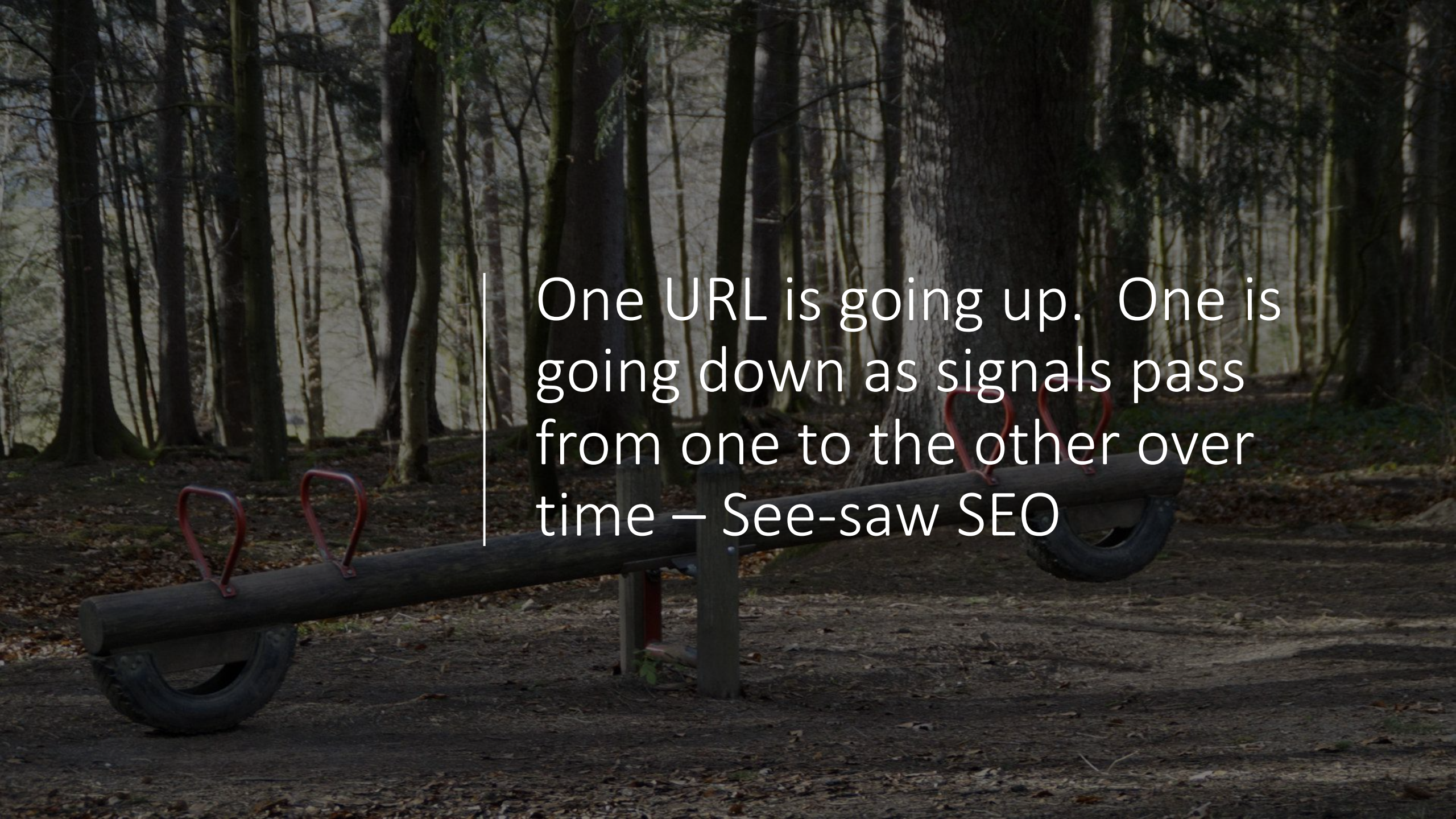
Also, any legacy canonicals
are now defunct by the
redirect status

A female rugby player in a dark blue and white jersey is running with the ball. She is looking down at the ball, which is a Gilbert A XV. Other players are visible in the background, some in similar jerseys and others in different colors. The scene is outdoors on a grassy field.

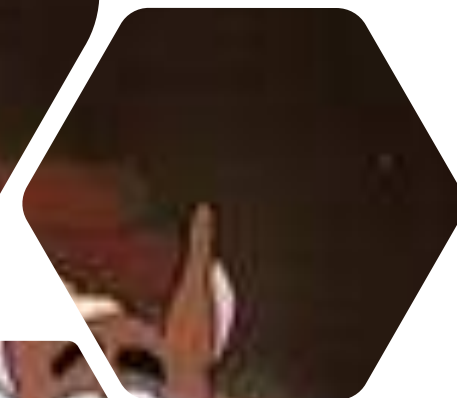
Cruft & migrations
are like a game of
rugby as the ball
(signals) is passed
between URLs

A sign of this is a
temporarily wrong target
ranking



A photograph of a wooden seesaw in a forest. The seesaw is made of a thick log with red metal handles and is supported by a central metal post. It is tilted slightly upwards on the left side. The background is a dense forest of tall, thin trees with bare branches, suggesting an autumn or winter setting. The ground is covered in dry leaves and dirt. Overlaid on the right side of the image is white text.

One URL is going up. One is going down as signals pass from one to the other over time – See-saw SEO



Google puts the pieces of
the puzzle together...
eventually (or never)

Machine
learned
ambiguity 'lag'



ALL of the above ==

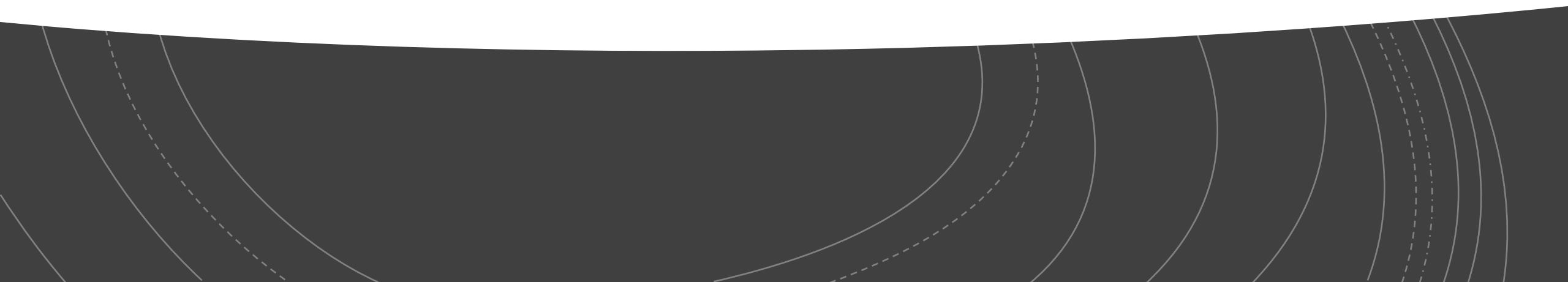


Crawl Budget
Woes / Learned
Quality Patterns

Your
reputation
precedes you
– You are
being judged
on the past



Your quality & crawl will be
'machine learned' by 'Large
Numbers' over time

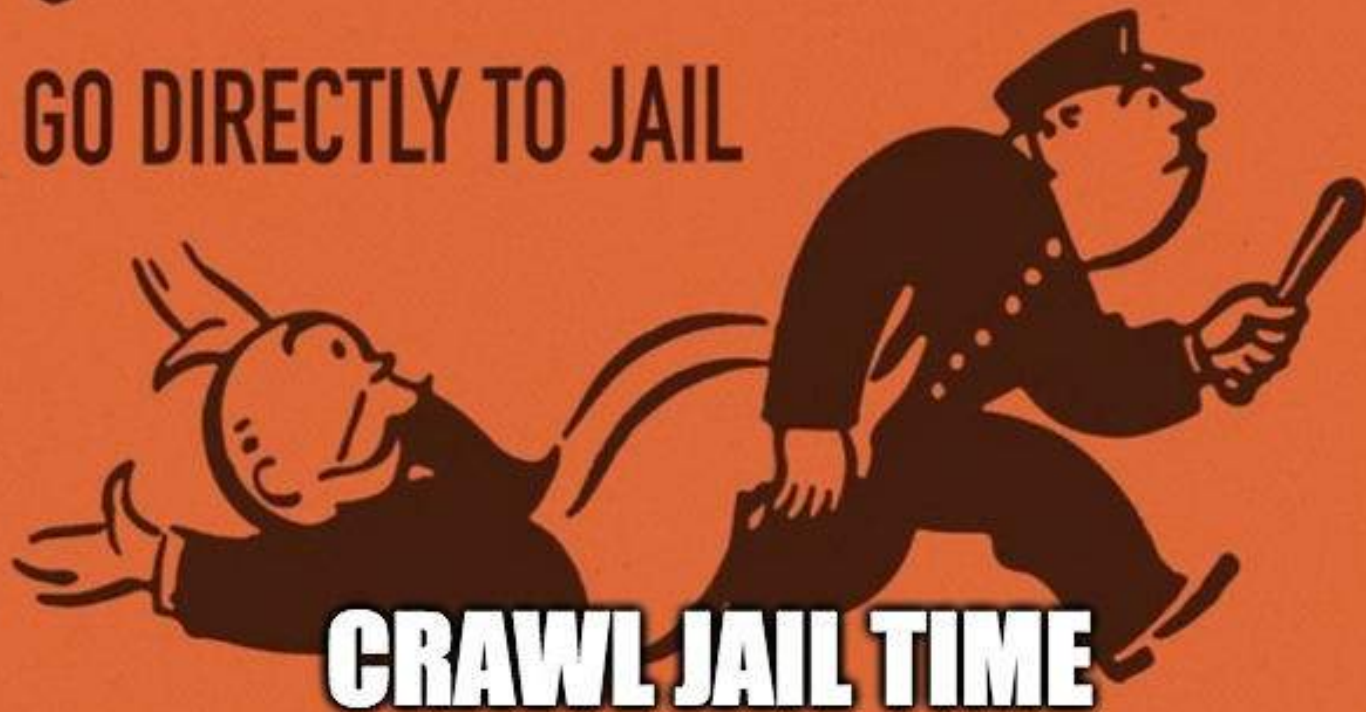


Probably you'll struggle
to get enough crawl for
Google to catch up

YOUR SITE IS DOING

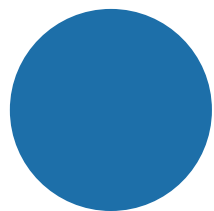
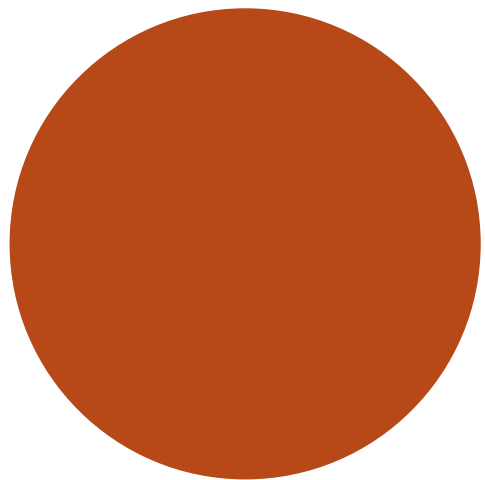
Chance

GO DIRECTLY TO JAIL



CRAWL JAIL TIME

DO NOT PASS GO, DO NOT COLLECT \$200



Search engines realise there is
no 'demand' for your pooppy
pages'

You've probably got
URLs which have not
been crawled for years

Your performance will be based on what is indexed



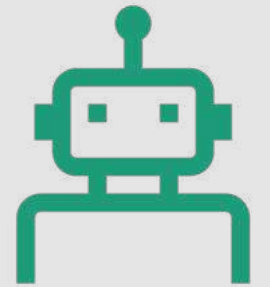
 **John**  
@JohnMu

Replying to [@kumarsinghdk](#), [@Marie_Haynes](#) and [@bill_slawski](#)

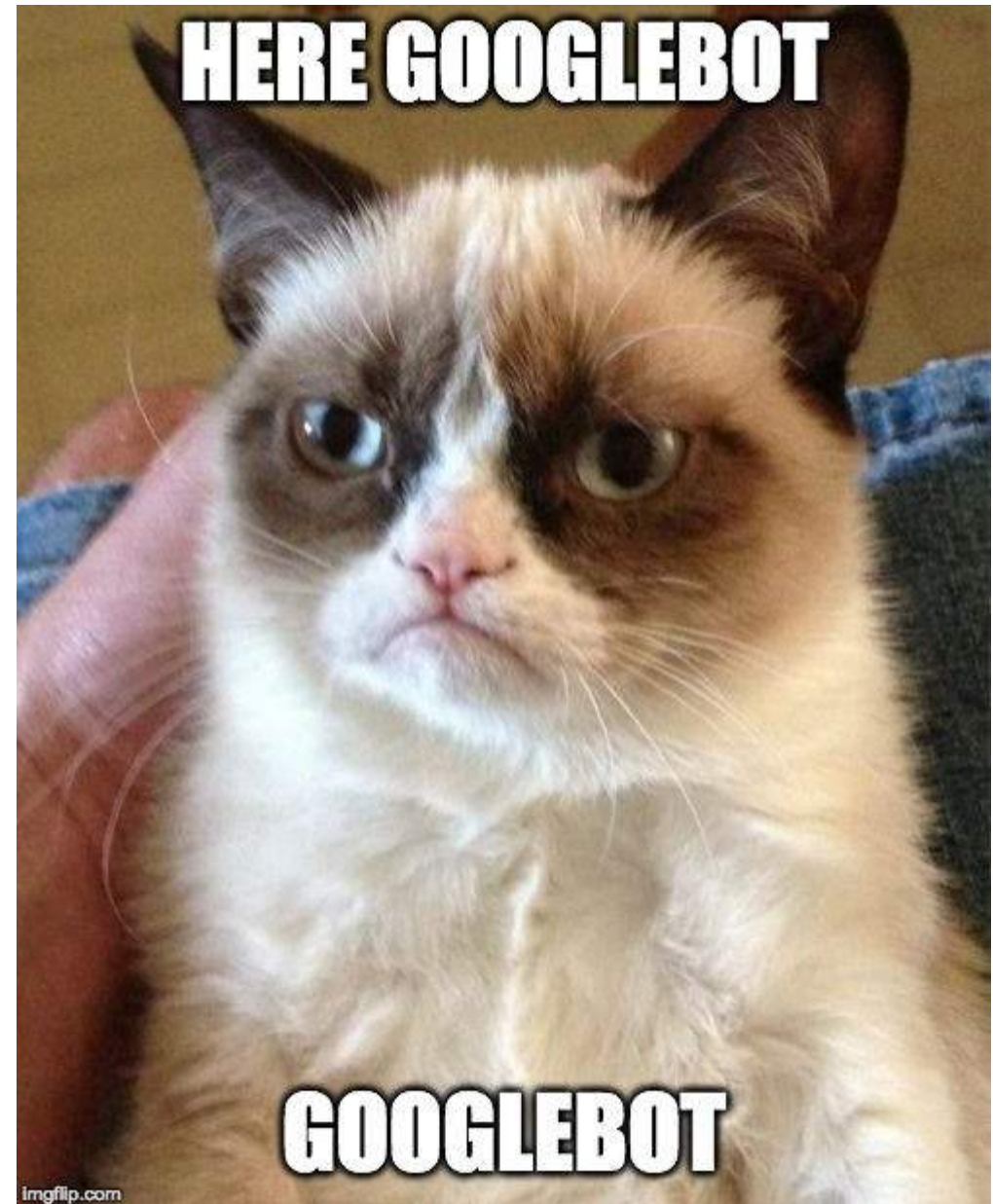
We do use the indexed content as a basis for what we show in search, so if you significantly change your page's content, I'd expect that to be reflected in search over time too.

10:04 am · 21 Jan 2020 · [TweetDeck](#)

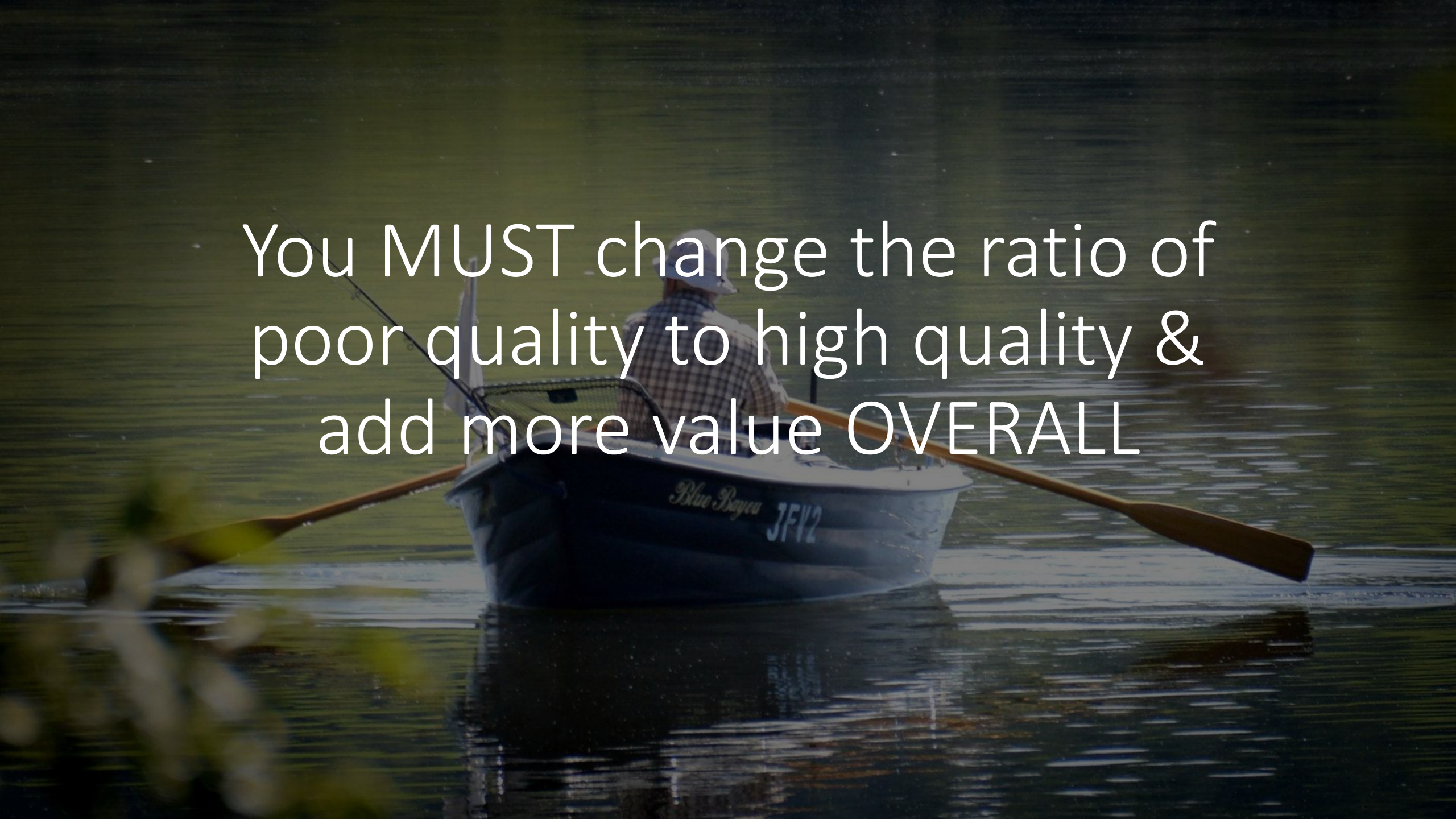
Fear not... Small
Wordpress site

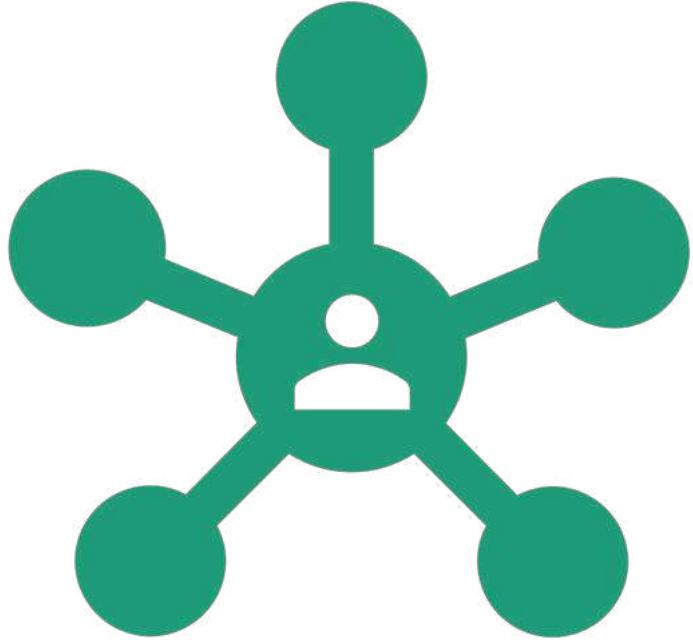


You need to lure
a Grumpy
Googlebot with
tasty quality
content morsels



You MUST change the ratio of
poor quality to high quality &
add more value OVERALL





And contribute
positively to 'The
Network Effect'

You need to substantively
improve the quality of your
pages, regain the crawl and get
more demand

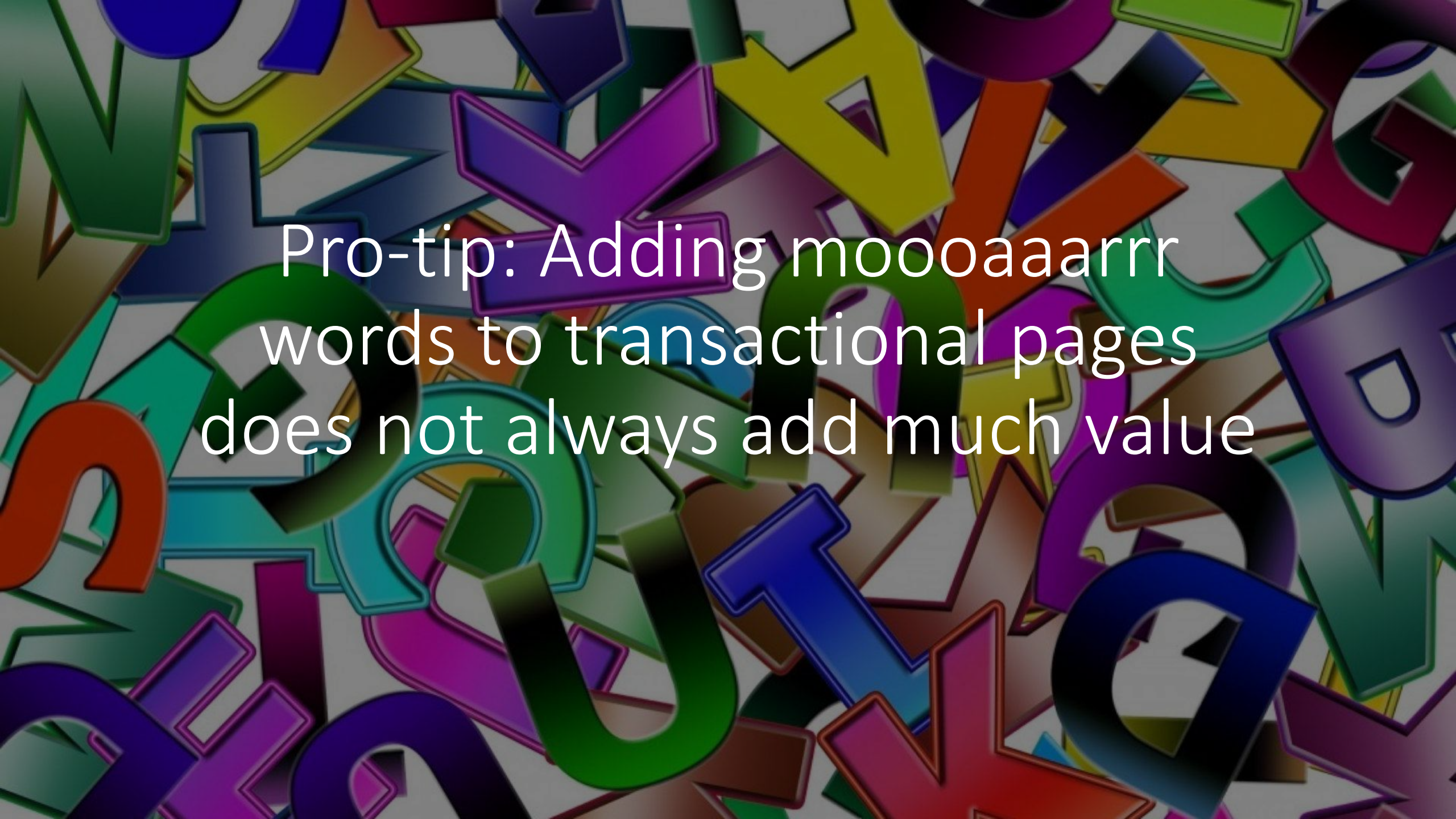




Watch for
patterns (clues) in
GSC Coverage &
take a demand-
driven approach

In transactional pages
identify valuable content,
features & attributes your
audience wants



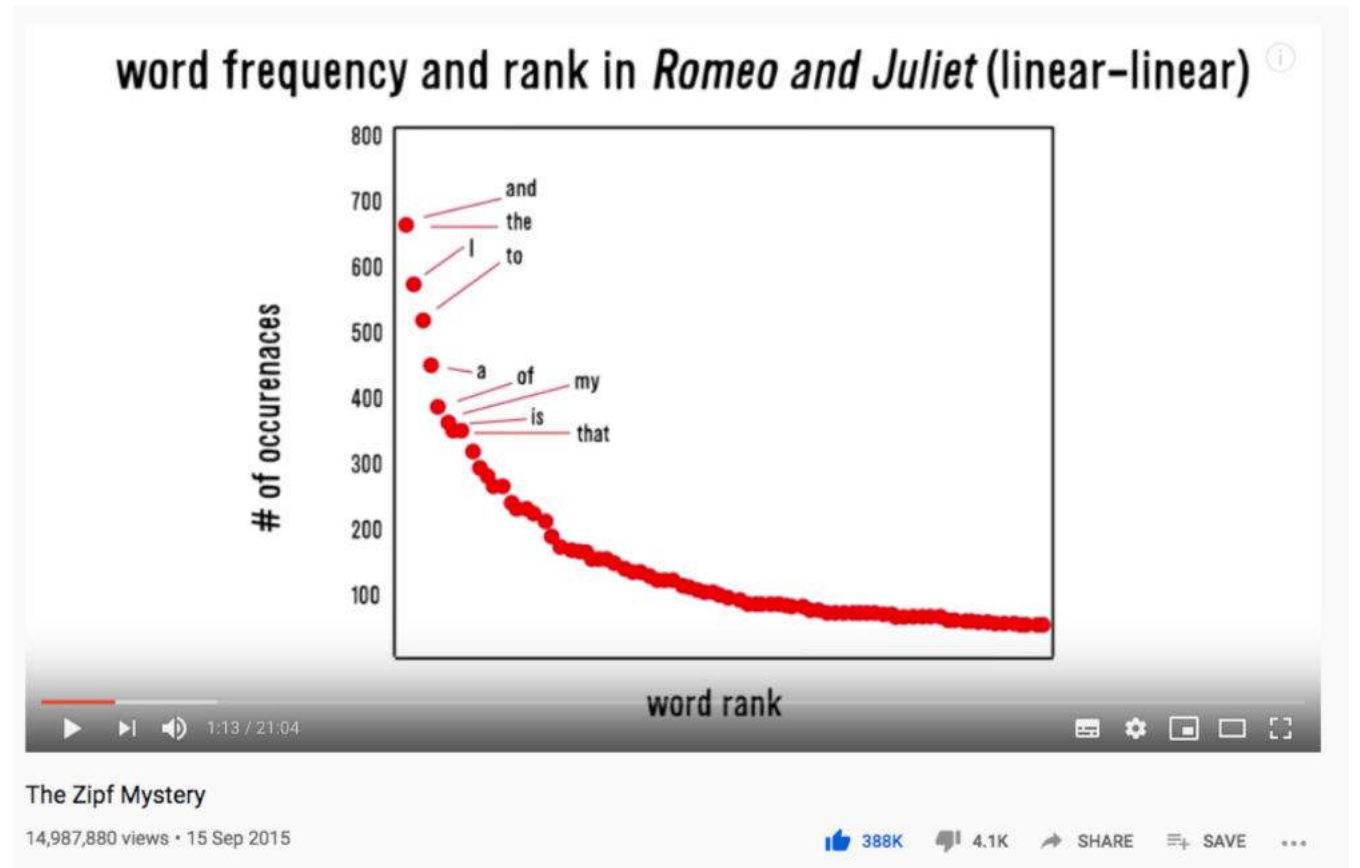


Pro-tip: Adding moooaaarr
words to transactional pages
does not always add much value



Plus... There is often a mathematically natural ordered pattern to things ranked by frequency

Like word frequencies
many types of
'ordered'
popularities will
have a Zipfian
Distribution



Where the frequency of
x is inversely
proportional to its
frequency table rank –
 $1/n$

Zipfian
Distribution
occurs in
many other
rankings
unrelated to
language

Population of cities in a country

Corporation sizes

Income rankings

No. of people watching same TV
channel

Nurture internal link graph
popularity that mirrors
'real life popularity'



Popularity & Zipfian Distributions

Is it really
popular in
'real life'?

Does it follow
'The Network
Effect'?

Or are you manipulating it for things you want to rank for (which aren't really popular at all)?



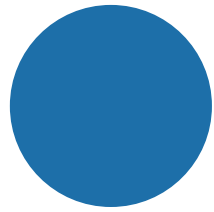
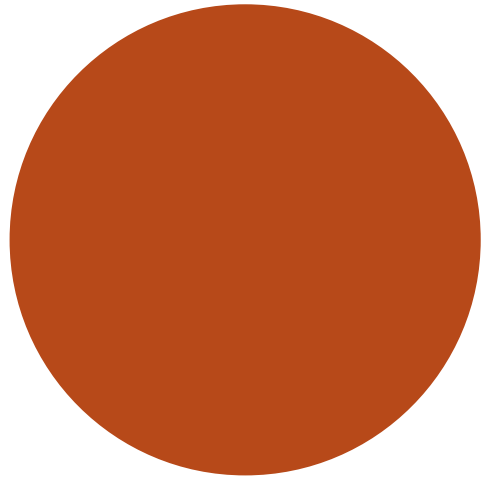
Bow tie of the web &
strongly connected
components



Building
genuine topical
hubs (Hub &
spoke)



Google puts the pieces of the puzzle together slowly but increases with quality improvements



Equipossibility goes well
beyond duplicate content too



Semantic
heterogeneity





The Web of Document Vs The Web of Data





Applies mostly to
entity-oriented
search

Mis-matching data types or
equivalencies in data tables
from same or other domains

Linked Data has been around for a long time

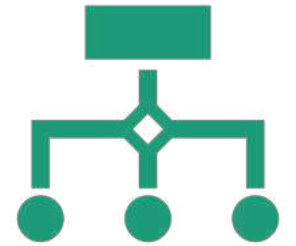
The **inventor** of the World Wide Web and the creator and advocate of the Semantic Web and **Linked Data**, Sir Tim Berners-Lee, laid down the four design principles of **Linked Data** as early as in 2006.

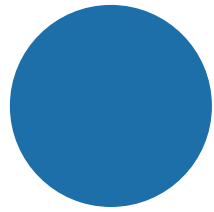
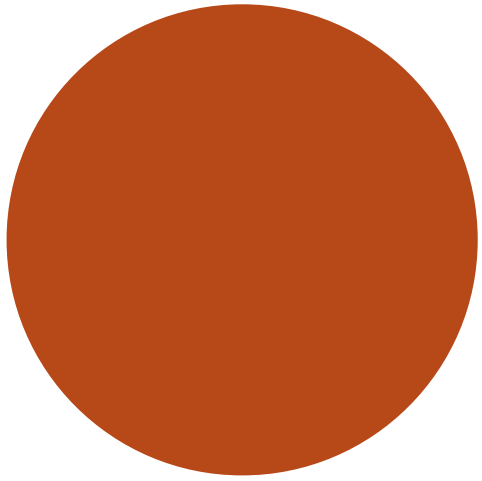


www.ontotext.com › [knowledgehub](#) › [fundamentals](#) › [linked-data-linke...](#)

What are Linked Data and Linked Open Data? - Ontotext

There have historically
been several ways to
implement linked data





Several linked-data types & web
sources come together causing
equipossibility

Some Types of Semantic Web Technologies & Their Markups

RDF

SPARQL

OWL

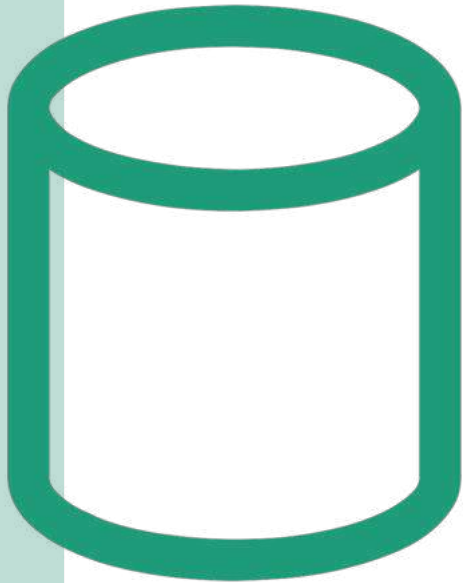
SKOS

RDFa

JSON-LD

Microdata

Implementation
Inconsistencies Prevail
(ed)



The Knowledge Graph & Its Data Sources

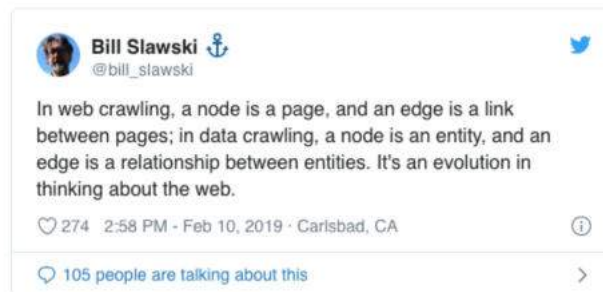
The majority of Google's entity types reference the Wikipedia URI or the Knowledge Graph MID

But Now Anything Structured Might Be Used

[ABOUT US](#)[SERVICES ▾](#)[CAREERS](#)[BLOG](#)

GOOGLE MAY USE ENTITY EXTRACTIONS, ENTITY CLASSES, ENTITY PROPERTIES, AND ASSOCIATION SCORES FROM PAGES TO BUILD KNOWLEDGE GRAPHS

When Google introduced the Knowledge Graph in 2012, it told us that it was going to start focusing upon [things and not strings](#). That process is maturing, and we have a chance to watch Google learn how to start crawling the Web to mine data and engage in entity extractions, instead of mining web information such as pages and links. As I wrote recently on Twitter about this:



A recently granted Google patent tells us about how the search engine may perform entity extractions from web pages, and store information about them. This goes beyond using knowledge bases as sources of information about entities, and moves on to finding more than what may be available in such sources, by looking at textual passages on web pages. The problem that this patent is intended to solve is described in this early line from the patent:

Conventional knowledge bases, however, can fail to provide up-to-date or reliable information regarding

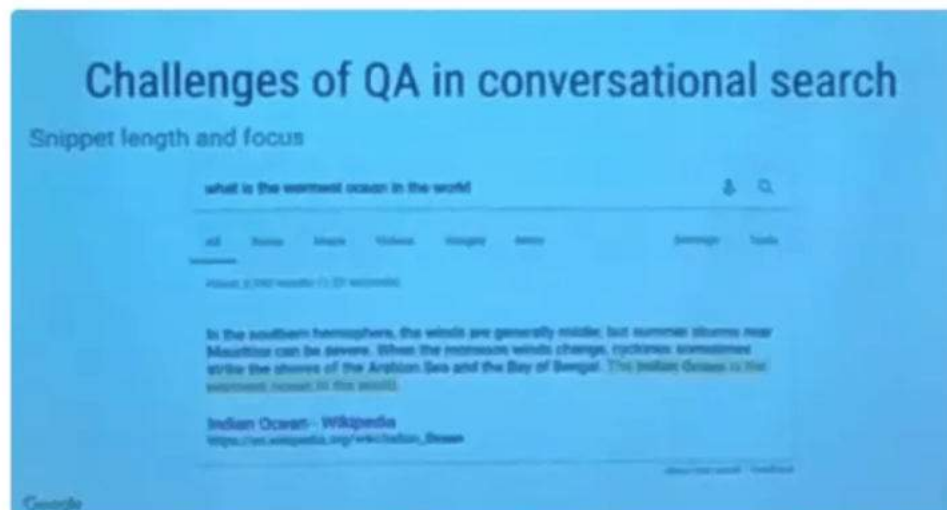
We already know conversational search fills gaps via web information



Dawn Anderson

@dawnieando

If there is no entity in schema conversational search next checks the web for information
[#essir2017](#)



9:17 AM - 6 Sep 2017

A dramatic landscape photograph at dusk or dawn. The sky is filled with dark, heavy clouds, with a warm orange and yellow glow from the setting or rising sun visible on the left side. In the background, dark, jagged mountain peaks rise above the horizon. A calm body of water, likely a lake or fjord, occupies the middle ground, reflecting the sky and the distant mountains. On the right side, a steep, grassy hillside slopes down towards the water. In the foreground on the right, there are large, dark, layered rock formations jutting out into the water. The overall mood is serene yet powerful, with a focus on natural beauty and vastness.

The web is a 'Data Lake'

Public Datasets (e.g. Kaggle) > 27k datasets

The screenshot shows the Kaggle Datasets page. At the top, there is a navigation bar with the Kaggle logo, a search bar, and links to Competitions, Datasets, Notebooks, Discussion, and Courses. Below this is a blue header with the word "Datasets" and buttons for "Documentation" and "New Dataset".

A light blue banner below the header contains the text "Help the community by creating and solving Tasks on datasets!" and a red box with the text "27,172 datasets". A red arrow points from this box to a search bar that contains the text "Search 27,172 datasets".

Below the search bar, there are tabs for "PUBLIC", "YOUR DATASETS", and "FAVORITES". The "PUBLIC" tab is selected. To the right of the tabs is a "Sort by:" dropdown menu set to "Hottest".

The main content area displays a list of datasets. The first four datasets are:

- Caselaw Dataset (Illinois)** by Caselaw Access Project, 929 MB, 8.1 rating, 2 Files (other), 49 votes.
- FIFA 20 complete player dataset** by Stefano Leone, 15 MB, 9.7 rating, 7 Files (CSV), 140 votes.
- eCommerce behavior data from multi category store** by Michael Kechinov, 4 GB, 10.0 rating, 2 Files (CSV), 1 Task, 52 votes.
- facenet pytorch vggface2** by timesler, 104 MB, 8.8 rating, 5 Files (other), 23 votes.

On the right side of the page, there is a section titled "Open Tasks" with a list of tasks:

- What to watch on Netflix ?** (2 Submissions - In Netflix Movies and TV Sh...)
- Visualize US Accidents Dataset** (4 Submissions - In US Accidents (3.0 million ...)
- Top 50 Spotify Songs - previous years** (1 Submission - In Top 50 Spotify Songs - 2019)
- Background analysis** (0 Submissions - In Credit Card Fraud Detect...)
- Data visualization** (5 Submissions - In Heart Disease UCI)

Enter Data Search by Google AI



The Keyword

Latest Stories

Product Updates

Company News

SEARCH

Discovering millions of datasets on the web

Natasha Noy

Research Scientist, Google Research

Published Jan 23, 2020

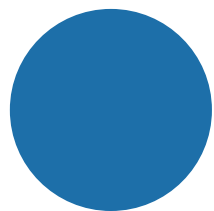
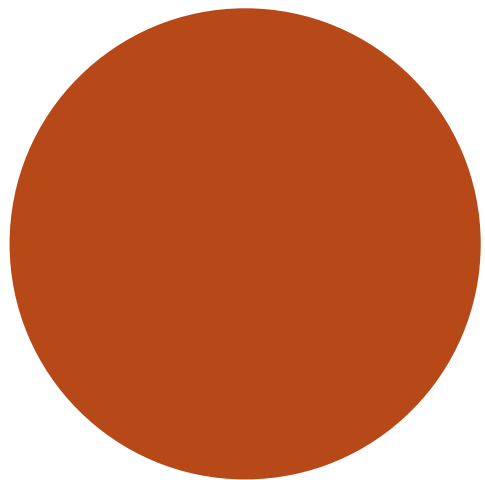
Across the web, there are millions of datasets about nearly any subject that interests you. If you're looking to buy a puppy, you could [find datasets](#) compiling complaints of puppy buyers or studies on puppy cognition. Or if you like skiing, you could [find data](#) on revenue of ski resorts or injury rates and participation numbers. [Dataset Search](#) has indexed almost 25 million of these datasets, giving you a single place to search for datasets and find links to where the data is. Over the past year, people have tried it out and provided feedback, and now Dataset Search is officially out of beta.



Skiing velocities and kinematic parameters for 36 world (fast)...
figshare.com
Updated Nov 9, 2017

Total Revenue for Skiing Facilities. All Establishments,
Employer Firms
[REVEF71392ALLEST](#)





Disambiguating Data
Sets (is hard)

Probability Predictions
Can Be VERY Wrong



Example: Philosophers Dates of Birth different in Wikidata & DBPedia



Ivo Velitchkov @kvistgaard · Jan 27

@dawnieando Here is a query w.wiki/GDq giving 50 philosophers for which @dbpedia and @wikidata state different date of birth.

You may change the DBpedia type for other occupations and the limit if you want different number of results.

#SPARQL

Hector Zagal	1966-06-06	1952-06-06
John Gardner	1965-03-25	1965-03-23
Paul Copan	1962-09-26	1962-09-20
Muhammad Tahir	1962-03-21	1962-01-06
Matthew Kramer	1959-06-09	1959-01-01
Mohsen Kadivar	1959-06-08	1959-06-07
Norbert Schmitt	1956-01-23	1956-01-01
Peter M. Haas	1955-01-23	1955-01-25
Michael Sandel	1953-03-05	1953-05-03
Javed Ahmad Ghamidi	1952-04-07	1951-04-18
Hitoshi Nagai	1951-11-10	1951-01-01

I asked for some examples on Twitter



Dawn Anderson


@dawnieando



Hey folks. Can anyone provide me with examples of inconsistencies in knowledge graphs (e.g. some data taken from one place and some taken from another place). e.g. photos from one place and content from another place). I know these are out there but please share examples

11:36 AM · Jan 26, 2020 · [Twitter Web App](#)

The community responded

 **Tom Folks** ✓ @tomtheseoguy · Jan 26
Replying to @dawnieando
Do you mean like this?


About 71,200,000 results (0.77 seconds)

1. Omega NC900HDC Juicer Extractor – **Best Juicer to Buy in 2020.** ...

2. Breville JE98XL Juice Fountain Plus – **Best Centrifugal Juicer.** ...

3. Tribest GSE-5050 Greenstar Elite Cold Press – **Best Twin-gear Juicer.** ...

4. SKG Slow Masticating Juicer Extractor – **Best Vertical Masticator.**


www.goodhousekeeping.com

[More items...](#) · Jan 2, 2020

healthykitchen101.com › [best-juicers](#)

[Top 5 Best Juicers of 2020 \(And Why They Are Worth Buying!\)](#)

for greens

vegetable

fruit

for citrus

for wheatgrass

for smoothies

ton

About Featured Snippets Feedback

'Mashed-up' Knowledge Graphs (Two Don Cherry's and his dog's name - 'Blue')



Aaron Bradley
@aaranged


Disambiguation is hard. The knowledge panel is a mashup of Don Cherry, the jazz trumpeter, and Don Cherry the (former) CBC hockey commentator. Keen eyes may also have correctly assessed that that's not a picture of the dog, Blue (it's Don Cherry, "big band singer and golfer").

don cherry's dog

About 943,000 results (0.70 seconds)

Don Cherry / Dogs

Blue



Blue (Don Cherry's dog) **Blue** was a white English Bull Terrier owned by hockey commentator Don Cherry. **Blue** was reportedly a gift from the members of the Boston Bruins when Cherry was their head coach from 1974 to 1979. The original **Blue**, who died in 1989, was a female.

[Blue \(Don Cherry's dog\) - Wikipedia](https://en.wikipedia.org/wiki/Blue_(Don_Cherry's_dog))
[https://en.wikipedia.org/wiki/Blue_\(Don_Cherry's_dog\)](https://en.wikipedia.org/wiki/Blue_(Don_Cherry's_dog))

Don Cherry

Jazz trumpeter

Born: November 18, 1936, Kingston
Died: October 19, 1995, Málaga, Spain
Spouse: Luba Cherry (m. 1995), Rosemarie Cherry (m. 1957–1997)
Teams coached: Colorado Rockies (Head coach, 1979–1980), Boston Bruins (Head coach, 1974–1979)
Children: Eagle-Eye Cherry, Christian Jon Cherry, David Ornette Cherry, Tim Cherry, Cindy Cherry

Movies and TV shows [View 10+ more](#)

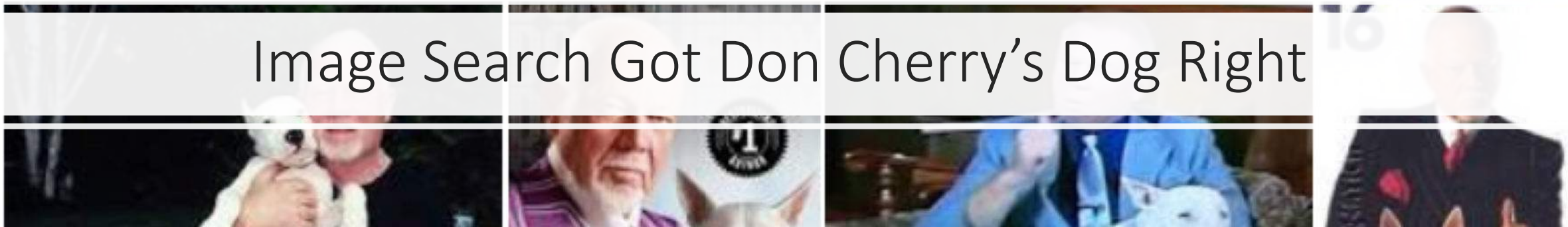
About 76,800,000 results (0.59 seconds)

Don Cherry's dog 'Blue'

Images for don cherry dog



Image Search Got Don Cherry's Dog Right



Wrong price included in images from other sites

Google search results for "ski insurance". The search bar shows "ski insurance" and the results indicate "About 143,000,000 results (0.44 seconds)".

Search results include:

- Ad · www.moneysupermarket.com/**
Travel Insurance for Skiing | Compare Cheap Ski Insurance
★★★★★ Rating for moneysupermarket.com: 4.8 - 20,061 reviews
Get **Travel Insurance** for Your **Ski Trip** Today at MoneySuperMarket. We're Here To Help You Find The Right **Travel Insurance** Policy. Compare Deals Online. PreExisting Medical Cover. Over 50s **Travel Insurance**. Compare 55+ Insurers. Default Star Ratings.
Winter Sports Cover · Single Trip Travel Cover · Cover For The Over 65s
Couples Travel Cover - from £5.45 - Cover for Couples · More
- Ad · www.statravel.co.uk/** 0333 321 0099
Cheap Skiing Insurance | Get Up To 24 Months Of Cover
Perfect Policies For Adventurous Travellers. Ideal For Backpackers, Students & More. Single...
- Ad · www.comparethemarket.com/**
Compare The Market™ Travel | Winter Holiday Insurance
Whether **Skiing** Or Taking In The Views, Get **Insured** Through CompareTheMarket Now!
- Ad · www.skicover.com/** 0333 252 9762
Save 50% on Ski Insurance | Designed for Wintersports
Best Prices Wintersports **Insurance**. Great Prices Including Off Piste Cover!

www.comparethemarket.com › travel-insurance › winter-sports › ski

The image gallery on the right shows various ski-related images. A red box highlights a price of £5.10 in one of the images, with a red arrow pointing to it from a text box that says "Price included in image from a different website".

Winter Sports Travel Insurance from just £5.10

Ski insurance

What is **ski travel insurance**? **Skiing** holiday **insurance** is cover specially designed for **ski** trips. It includes protection for your equipment, medical treatment in case you're injured, and specific costs like getting an injured skier off a mountain and safely home. 30 Oct 2019

www.moneysupermarket.com › travel-insurance › ski-insurance
Compare Ski Travel Insurance | MoneySuperMarket

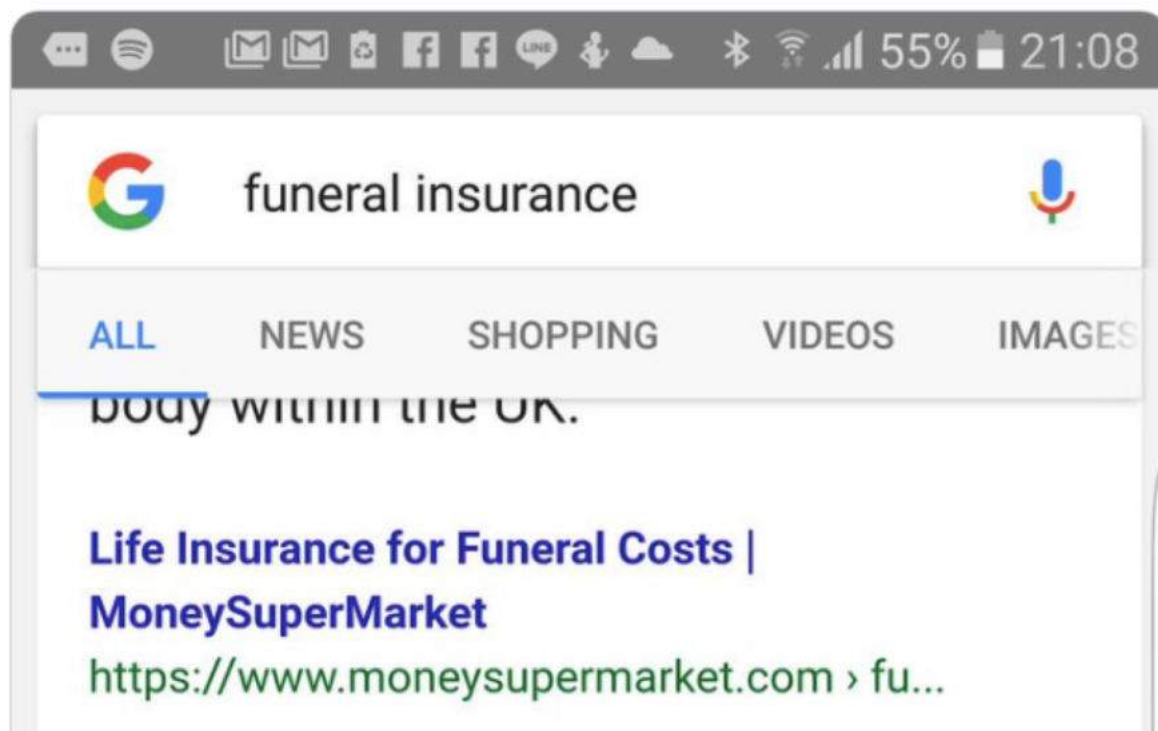
People also search for View 2+ more



Mr Jonathan Jones (謝強尼) 🇬🇧

@Jonny_J_

Do Google really think showing US websites for UK results is great user experience? #PeopleAlsoAsk #SEO @rustybrick @jenstar



Tweet

People also ask

US info in UK 'People also ask'

How much is a funeral insurance?

"The average coverage is **\$10,000-\$15,000** and most companies won't cover more than \$25,000 in funeral costs. It is usually purchased from a funeral home or mortuary. Funeral insurance premiums are generally paid as an initial lump sum (for the total amount being pre-paid) or with a three-, five- or 10-year payment plan."

Cost of Prepaid Funerals - Consumer Information and Prices ...

[personalfinance.costhelper.com > funeral...](https://personalfinance.costhelper.com/funeral...)

US Price Info Showing in UK 'People Also Ask'

Edgar's Were Comical



Edgar Meij
@edgarmeij

Replying to @dawnieando and @aaranged

See slideshare.net/edgar.meij/201... at around slide 23...



Web-scale semantic search

Most web search engine users are increasingly expecting direct and contextually relevant answers to their informatio...

slideshare.net

3:22 PM · Jan 29, 2020 · [Twitter for iPhone](#)

Van Diesel – Death – To Be Advised??



Vin Diesel , born as Mark Vincent, is an American actor, producer, director, and screenwriter. He came to prominence in the late 1990s, and first became known for appearing in Steven Spielbergs Saving Private Ryan... [wikipedia.org](https://en.wikipedia.org/wiki/Vin_Diesel)

Born: July 18, 1967, [New York City, New York, USA](#)

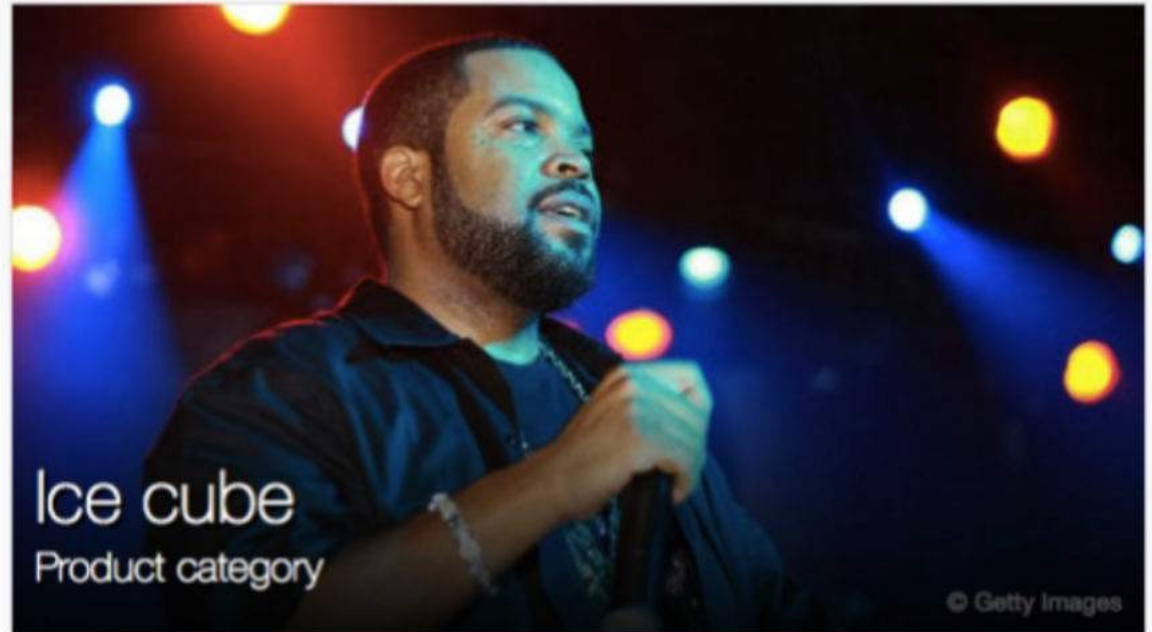
Died: January 30, 2014, [TBA](#)

Height: 5' 11" (1.82m)

Partner: [Paloma Jiménez \(2008-2014\)](#)

Parents: [Irving Vincent, Delora Vincent](#)

Does An Ice Cube Have Parents?



An **ice cube** is a small, roughly **cube**-shaped piece of **ice** (frozen water), conventionally used to cool beverages. **Ice** cubes are sometimes preferred over crushed **ice** because they melt more slowly; they are standard in... [wikipedia.org](https://en.wikipedia.org/wiki/Ice_cube)

Born: June 15, 1969 (age 44), Los Angeles, California, USA

Height: 5' 7" (1.73m)

Spouse: [Kimberly Woodruff](#) (m. 1992-present)

Parents: [Doris Benjamin](#), [Hosea Jackson](#)

Children: [Darrel Jackson](#), [O'Shea Jackson Jr.](#), [Shareef Jackson](#), [Karima Jackson](#), [Deja Jackson](#)

[Feedback](#)

Teenage Mutant Ninja Turtles paint works of art

Michelangelo

Artist



Michelangelo di Lodovico Buonarroti Simoni , commonly known as Michelangelo, was an Italian sculptor, painter, architect, poet, and engineer of the High Renaissance who exerted an unparalleled influence on the... [wikipedia.org](https://en.wikipedia.org/wiki/Michelangelo)

Born: March 6, 1475, [Caprese Michelangelo](#)

Died: February 18, 1564, [Rome](#)

Parents: [Ludovico di Leonardo di Buonarotto Simoni](#), [Francesca di Neri del Miniato di Siena](#)

[Feedback](#)

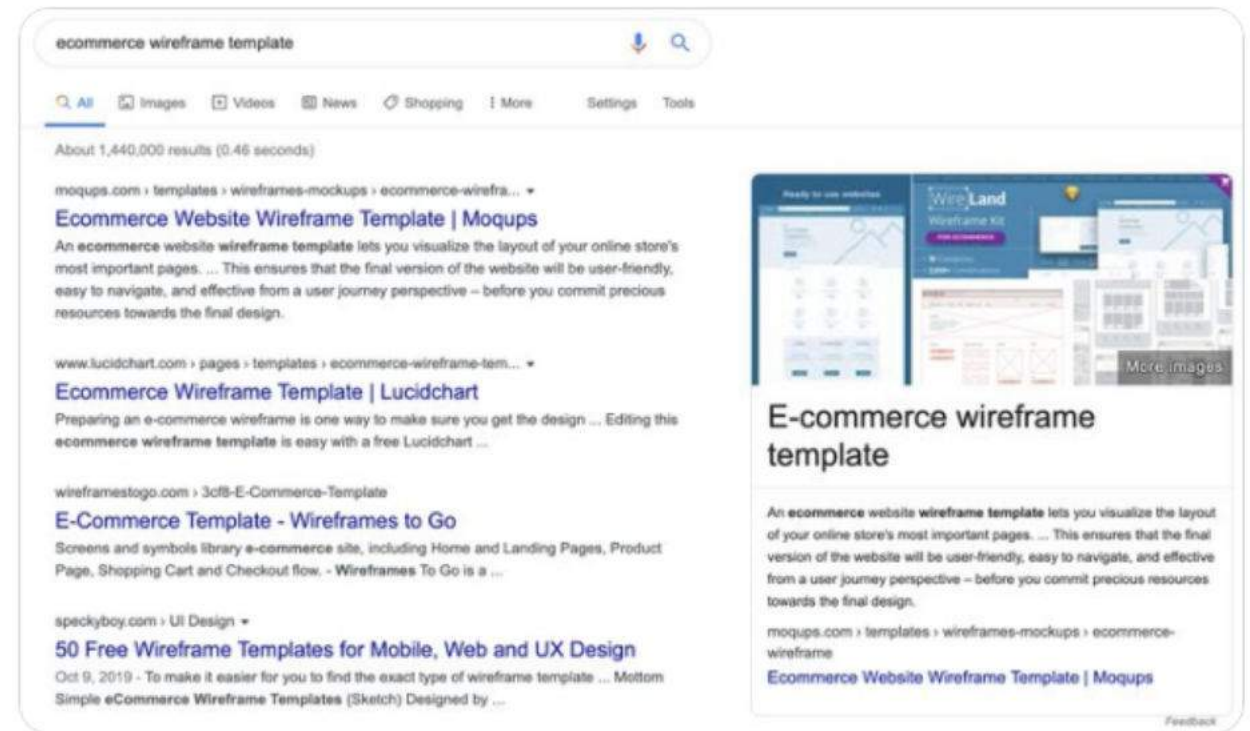
Images from Direct Competitors



Mihai Sterian
@MihaiSterian


Replying to @dawnieando

Link and content from Moqups, images from competitors.



3:29 PM • Jan 29, 2020 • Twitter Web App

Wikipedia Made Their Point Well



6,000,000 ARTICLES
WIKIPEDIA
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Wikipedia store

Interaction


- Help
- About Wikipedia
- Community portal
- Recent changes
- Contact page


Tools

- What links here
- Related changes
- Upload file
- Special pages
- Permanent link
- Page information
- Wikidata item

Print/export

- Download as PDF
- Printable version

Languages 


 Add links

Project page **Talk** Read Edit View history Search Wikipedia

Wikipedia:How many legs does a horse have?


From Wikipedia, the free encyclopedia

This page is about an explanation to WP:V. For the essay about flogging a dead horse, see WP:STICK.




This is an explanatory supplement to the Wikipedia:Verifiability page.
This page is intended to provide additional information about concepts in the page(s) it supplements. This page is not one of Wikipedia's policies or guidelines, as it has not been thoroughly vetted by the community.

Shortcut
WP:LEGS



This page **contains material that is kept because it is considered humorous.**
It is not meant to be taken seriously.



This page in a nutshell: Simply saying that a horse has five legs doesn't make it true – calling a horse's tail a leg does not make it one.

Jack and the Five Legged Horse

My uncle once said, "Boy, do you see that horse?"

Yessir.

"How many legs does that horse have?"

Four.

"Now what if I say that a horse's tail is a **leg**, too?"

Uh...

"If I call a horse's tail a leg, how many legs does that horse have?"


Five?

"Boy, will you never learn? It doesn't matter what you **call** something. That horse still has four legs."^[1]

Sometimes when confronted with a difficult challenge we reply, *That's not true!* We say, *That's not a (vanity page)/(rant)/(personal attack)/(tail); that's a (noteworthy article)/(discussion)/(polite comment)/(leg).* But this changes nothing.

It's never enough to reply to a comment by saying *You're wrong! That's not true!* If it really **is** a leg, then **show us how.**

[[WP:LEGS]]



It still has four legs.

“Simply saying that a horse has five legs doesn't make it true – calling a horse's tail a leg does not make it one.” (Wikipedia, 2019)

A six legged horse?



how many legs does a horse have



All

Images

News

Shopping

Videos

More

Settings

Tools

About 273,000,000 results (0.46 seconds)

six legs

All horses have **two legs** at the back. They also have fore legs at the front. That makes it a total of **six legs**. But six is an odd number of legs for a horse to have.



[Wikipedia:How many legs does a horse have? - Wikipedia](https://en.wikipedia.org/wiki/Wikipedia:How_many_legs_does_a_horse_have%3F)


https://en.wikipedia.org/wiki/Wikipedia:How_many_legs_does_a_horse_have%3F


? About this result

Feedback

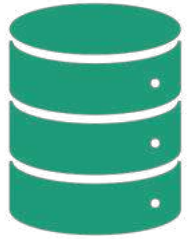
Somes Reputable
Knowledge
Repositories are
Just Plain 'Wrong'

Google Works To Fix How Many Legs Horses & Snakes Have

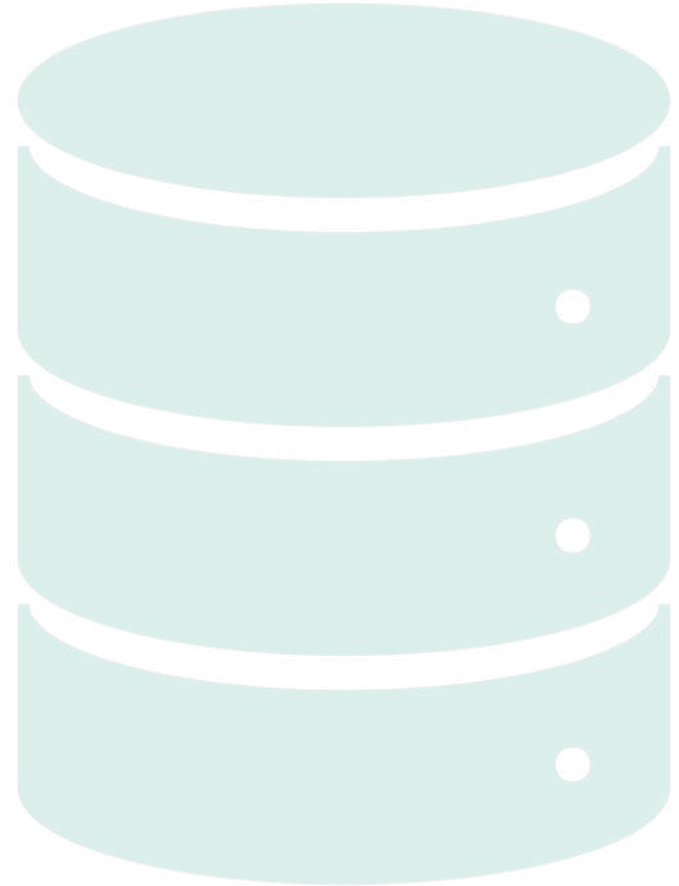
Apr 30, 2019 • 8:10 am |  (3)

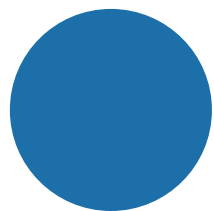
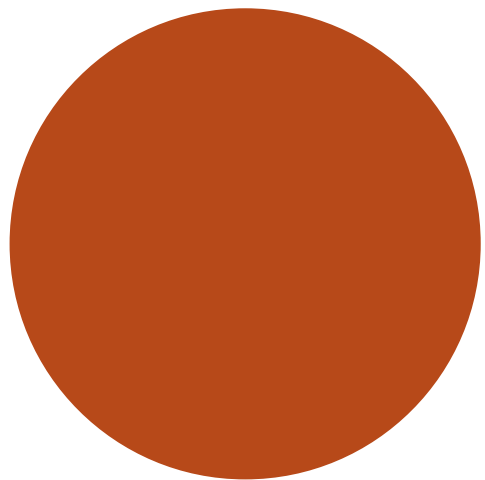
by [Barry Schwartz](#)  | Filed Under [Google Search Engine](#)





Equivalency
between data
sources is sought






Efforts Are Underway to Resolve

Since 2011 major search engines have been focusing on Schema commonly

Google Deprecating Data-vocabulary.org Schema

Google Sending data-vocabulary.org Schema Deprecation Notices

Jan 22, 2020 • 7:11 am |  (1)

by [Barry Schwartz](#)  | Filed Under [Google Search Engine Optimization](#)



Google announced yesterday that starting on April 6, 2020 it will [no longer support data-vocabulary.org schema](#). Hours later, Google began sending notifications via Google Search Console about this happening.

Your Breadcrumbs Markup May Be Impacted



Breadcrumbs issues detected on `https://www.`

To the owner of `https://www.` |/:

Search Console has identified that your site is affected by 1 Breadcrumbs issues:

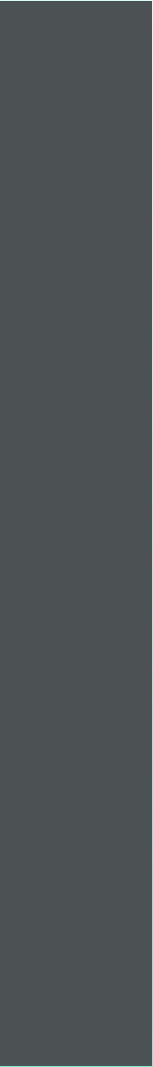
Top Warnings

Warnings are suggestions for improvement. Some warnings can affect your appearance on Search; some might be reclassified as errors in the future. The following warnings were found on your site:

`data-vocabulary.org` schema deprecated

We recommend that you fix these issues when possible to enable the best experience and coverage in Google Search.

[Fix Breadcrumbs issues](#)

A dark gray vertical bar is positioned on the left side of the slide, extending from the top to the bottom.

Entities often have
many surface forms

Schema:SameAs and OWL:SameAs

schema.org

Custom Search

HomeSchemasDocumentation

sameAs

[Thing](#) > [Property](#) > [sameAs](#)

URL of a reference Web page that unambiguously indicates the item's identity. E.g. the URL of the item's Wikipedia page, Wikidata entry, or official website.

[more...]


Values expected to be one of these types

[URL](#)


Used on these types

[Thing](#)

SameAs Seems To Be Effective



VRIJE
UNIVERSITEIT
AMSTERDAM




[Home](#) [Profiles](#) [Research Units](#) **[Research Output](#)** [Datasets](#) [Activities](#) [...](#)

On the Impact of sameAs on Schema Matching

[Joe Raad](#), [Erman Acar](#), [Stefan Schlobach](#)

[Artificial Intelligence \(section level\)](#), [Artificial intelligence](#), [Knowledge Representation and Reasoning](#)

Research output: Chapter in Book / Report / Conference proceeding > Conference contribution > Academic > peer-review



BE consistent in ALL
mentions of your brand
name & ALL THINGS
GENERALLY



Remember: Consistency is one of the Kings



Utilise 'known'
(popular) public
datasets & knowledge
repositories

Avoid conflict with
'known' (popular)
public datasets &
'knowledge
repositories'



Realise that well
known
knowledge
repositories
may be wrong





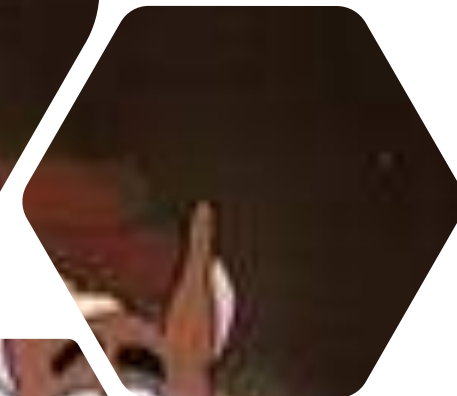
But they may be
conceptually still
be 'right'



They'll likely win in an
equipossibility face-off



Leave no room for
ambiguity where
possible



Google puts the pieces
of the puzzle together
often badly

Location based
ambiguity



Semantic Heterogeneity
is a problem locally

Entity
Address
(location) is
a VERY Big
Problem
with entity-
oriented
search

street number

Locality - city or town

street/route name, if detected

postal code, if detected

country, if detected

broad_region - administrative area, such as the state, if detected

Narrow_region - smaller administrative area, such as county, if detected

Narrow_Region
is VERY
Problematic

Inconsistent data from
knowledge repositories

Inconsistent data from websites

Inconsistent data from searchers

HUGELY SPLIT EQUIPROBABILITY

Which County is Bristol in?

which county is bristol in



All

Maps

Images

News

Shopping

More

Settings

Tools

About 128,000,000 results (0.64 seconds)

Bristol / Counties

Gloucestershire



The historic centre of **Bristol** and the sections of the city north of the River Avon (Lower, or **Bristol**, Avon) were part of the historic **county** of Gloucestershire, while the areas south of the Avon lay within the historic **county** of Somerset until the creation of the **county** of **Bristol** (1373–1974) and then of the **county** ...

[Bristol | History, Points of Interest, & County | Britannica.com](https://www.britannica.com/place/Bristol-England)

<https://www.britannica.com/place/Bristol-England>

Bristol is
BOTH a
county and a
city





In Scotland and Wales
it's worse

Which County is Cardiff in?

which county is cardiff in

All Maps Images Shopping News More Settings Tools

About 58,100,000 results (0.90 seconds)

Cardiff / Counties

Glamorgan



Cardiff is the county town of the historic **county of Glamorgan**, and in 1974, South **Glamorgan**. Cardiff is part of the Eurocities network of the largest European cities.

en.wikipedia.org › wiki › Cardiff

[Cardiff - Wikipedia](#)

Cardiff is in
South
Glamorgan
NOT
Glamorgan



But even
humans are
not so sure

which county is cardiff in



Cardiff, Welsh Caerdydd, city and capital of Wales. **Cardiff** exists as both a city and a **county** within the Welsh unitary authority system of local government. It is located within the historic **county** of Glamorgan (Morgannwg) on the Bristol Channel at the mouth of the River Taff, about 150 miles (240 km) west of London.

[Cardiff | History, Facts, & Attractions | Britannica.com](#)

<https://www.britannica.com/place/Cardiff-Wales>

Feedback

People also ask

Is Cardiff in the Vale of Glamorgan?



Is Glamorgan a county?



Is Cardiff South Glamorgan?



Does Mid Glamorgan still exist?



Feedback



Google Does Not
Understand the
Word ‘Historic’



What about
counties that
don't even exist
any more?


which county is twickenham in

Search icons: All, Maps, News, Images, Shopping, More, Settings, Tools

About 16,000,000 results (0.89 seconds)

Twickenham / Shire counties

Middlesex County



It doesn't even exist anymore

In 1965 **Middlesex County** Council was abolished and replaced with the **Greater London** Council and the boroughs of Twickenham, **Richmond** and Barnes were combined to form the **London Borough of Richmond upon Thames**.

en.wikipedia.org › wiki › Twickenham

[Twickenham - Wikipedia](#)

At least this
one mentions
London as well
as Middlesex

which county is hounslow in



All

Maps

News

Shopping

Images

More

Settings

Tools

About 2,660,000 results (0.71 seconds)

county of Middlesex

Hounslow. Hounslow, outer borough of **London**, England, on the western periphery of the metropolis. It is part of the historic county of Middlesex and lies in the valley of the River Thames.

[www.britannica.com](https://www.britannica.com/place/Hounslow) › place › Hounslow

[Hounslow | borough, London, United Kingdom | Britannica](https://www.britannica.com/place/Hounslow)

About Featured Snippets

Feedback

Everyone... Is confused

which county is hounslow in



People also ask

Is there still a county of Middlesex?



Is Hounslow in Middlesex or London?



When did Middlesex stop being a county?



Is the City of London a county?



Feedback

[en.wikipedia.org](#) › [wiki](#) › [Hounslow](#) ▼

Hounslow - Wikipedia

Hounslow is a large suburban town, and the principal town of the London Borough of ... OS grid reference · TQ140760. • Charing Cross, 10.7 mi (17.2 km) ENE · London borough · Hounslow · Ceremonial county · Greater London · Region.

Ceremonial county: [Greater London](#) **UK Parliament:** [Brentford & Isleworth, Feltham & ...](#)

London Assembly: [South West](#)

Sovereign state: [United Kingdom](#)

[London Borough of Hounslow](#) · [Hounslow Heath](#) · [Hounslow railway station](#)



The passage of time adds new meaning to
queries sometimes too

● blackberry
Search term

+ Compare

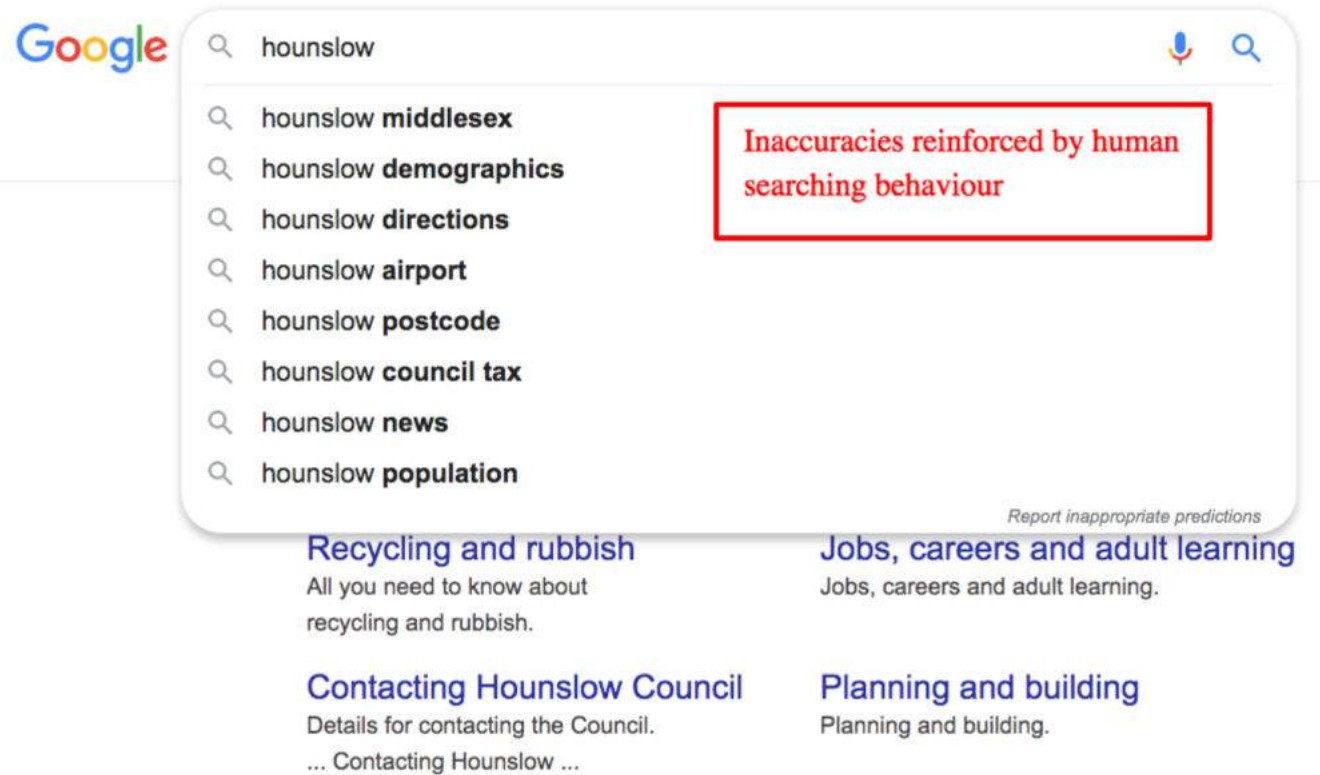
Worldwide ▼ 2004 – present ▼ All categories ▼ Web Search ▼

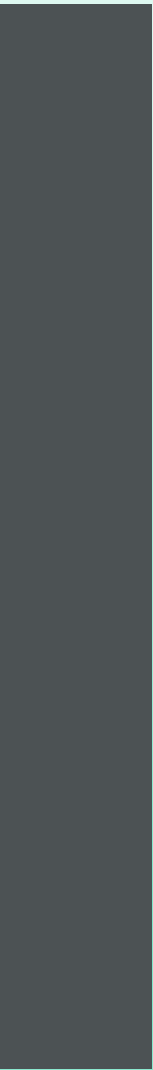
Interest over time ?



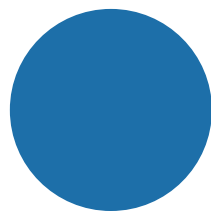
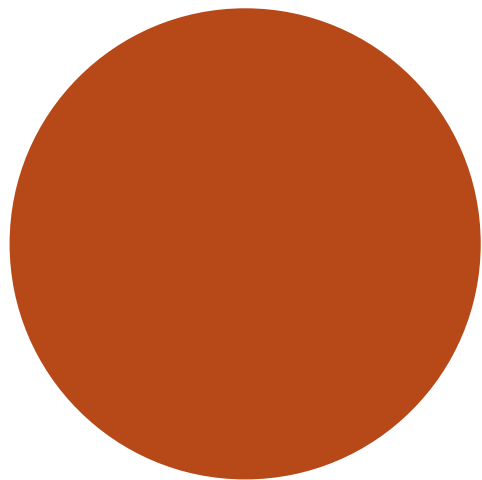
The rise and fall
of the
Blackberry?

Older humans
probably keep
ambiguous
county + town
combinations
alive



A dark grey vertical bar is positioned on the left side of the slide, extending from the top to the bottom.

Maximise the 'Local'
Opportunity

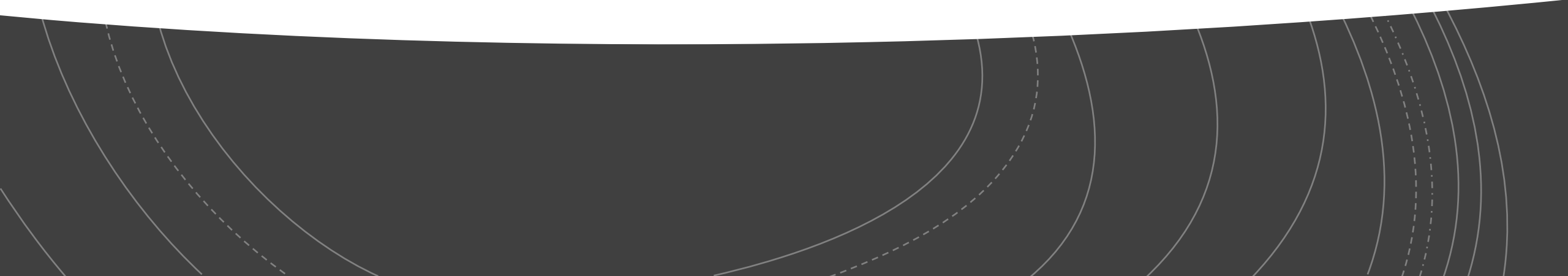


In location focused pages adding a
postcode could be much more valuable
than adding lots of words

Contextual relatedness
in local spaces is
proximity-based



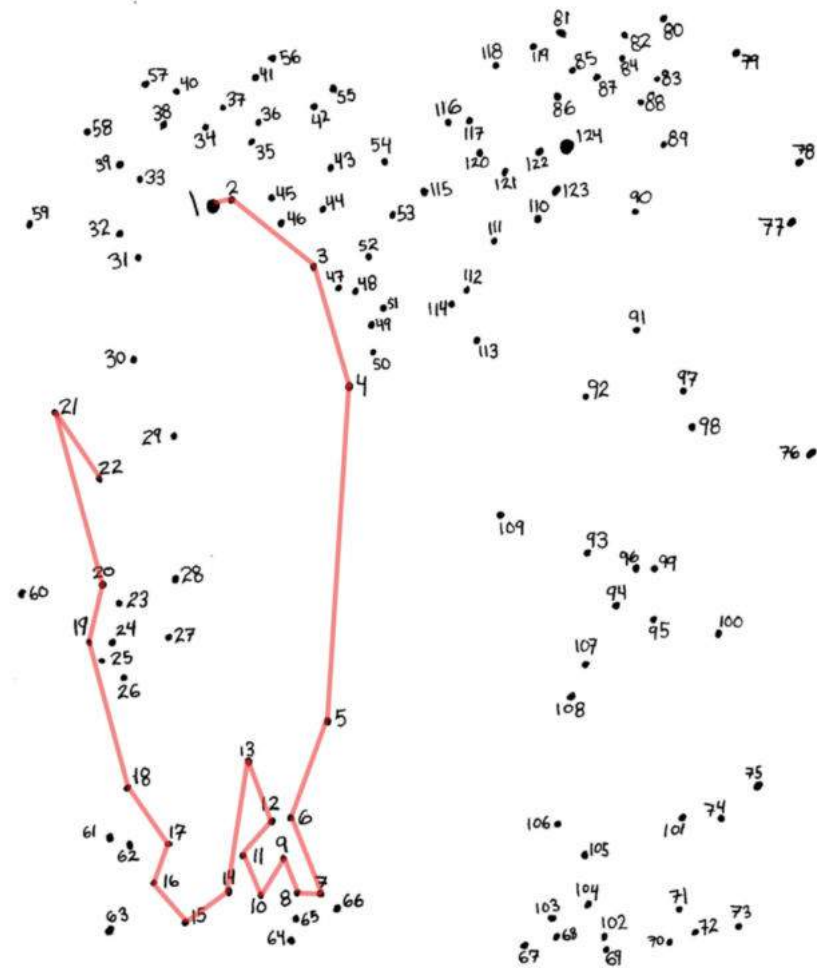
AreaServed Schema &
Internal linking by Proximity
(e.g. lat & long) is powerful

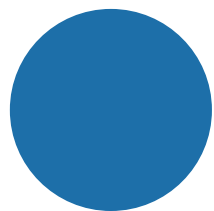
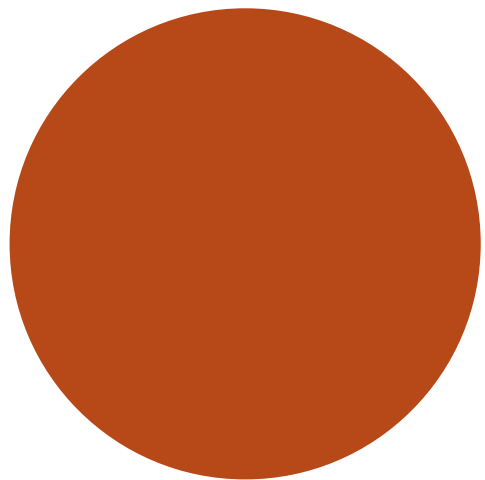




Google sometimes puts
the pieces of the puzzle
together sporadically

Align ALL the dots





Most Importantly -
Be Consistent

Keep in Touch

- @dawnieando
- @BeBertey





Some Take Aways



Go

- Watch for 'impact' patterns to maximise upon in GSC 'coverage' reports

Optimise

- Be careful you don't prune away second level relatedness signals

Map

- Consider Pareto's Law when identifying issues to fix in GSC coverage

Optimise

- Check for 'wrong page ranking' due to craft & lag passing signals

Get

- On location sites consider internal linking on proximity

Identify

- Add value to improve on 'The Network Effect' in attributes & informational content

Use

- Make it obvious you can 'probably' help with transactional needs through other content