How does the Hero do it?

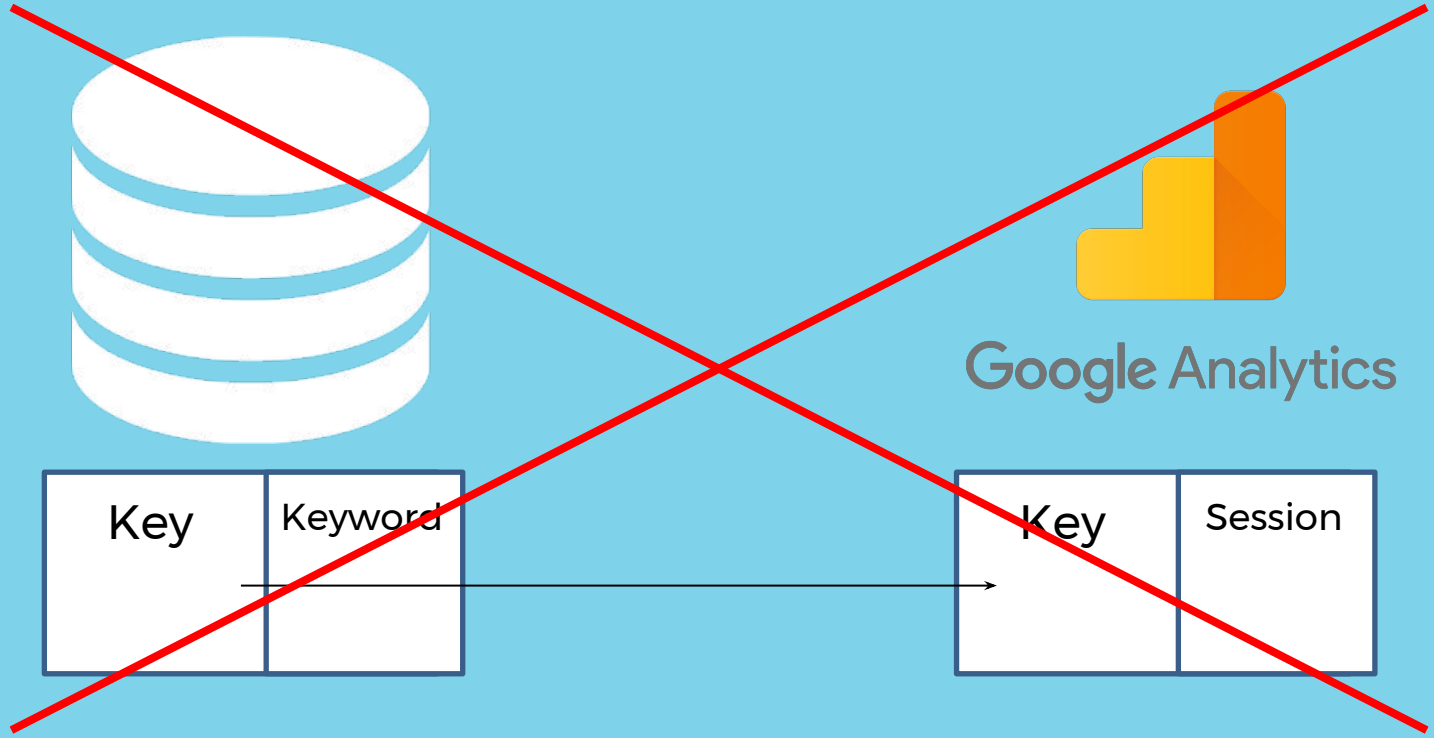
Extended Version

Table of Content

1. How it **does not** work

For Clarification Purposes

KEYWORD HERO



There are no unique keys that would allow you to map a keyword with a session. For this very reason we use several data sources to reduce the number of possible keywords for a specific session to the minimum.

2. Understanding the Domain

First Things First

KEYWORD HERO

Once a user signs up, the Hero analyzes the history of the GA account, looking for:

1. **patterns** e.g. site structure, URLs that pull traffic; clusters among organic traffic; device categories; time / date; locations; ...

2. **type and content** e.g. # of direct sessions attributed to organic; news or ecomm; one-pager or many product sites; semantical topics; ...

3. Generating Keywords

What could it be???

In the first step we only generate a large set of possible keywords per URL.

There is **no classification or attribution** in this step.

There are two types of data sources we use for this:

1. **3rd party data sources** e.g. rank monitoring services; browser extensions data; Bing search API; ....

2. **1st party user data** Search Console; remaining keywords in GA

YOURDOMAIN.COM/FOLDER

Keyword 1
Keyword 2
Keyword 3
Keyword 4

⋮

→ the result is one big bucket with possible keywords per URL

4. Clustering and Classifying Keywords

It’s getting tricky...

KEYWORD HERO

Using external APIs we look for anormal behaviour in a keyword's history

Google AdWords

were there changes in traffic in the last 12 months

(per country / device)

Google Trends

is this a new keyword?
did spikes occur?

is there a reference article on Wikipedia and how does it behave?

→ cool link

Semantics

After the Hero created a keyword set and checked for potential spikes and anormalities, he needs to “understand” the keyword.

is this a brand or a non-brand keyword?

is it a person / brand name? (important for ngram analysis)

is the keyword transactional / navigational or informational?

category clusters

Keyword	Traffic	Trending?	Brand / Non-Brand	Name	Type	Category	Category 2
iPhone 8	spike in last 3 months	new	yes (partly)	no	informational	electronics	Smartphone
shoes	constant	old/constant	no	no	informational	clothing	Shoe
how big is the eifeltower	constant	old/constant	no	no	informational - question phrase	travel	Eifeltower

→ the result is one big dataframe where certain parameters are attributed to the keywords

(looks sort of like this thing above)

In the next step we analyze differences in the performance of organic traffic due to keyword fluctuation in the SERPs.

suppose we have a site that ranked on position 13 - 15 in calendar week X for the term “shoes”. At this point we have the following metrics for this landing page:

Sessions	Bounce Rate	Conversions
500	50%	10

in calendar week **X + 1** the site now ranks on position 5 for shoes and metrics changed to:

Sessions	Bounce Rate	Conversions
1.500	50%	15

in this simple example we assume ceteris paribus, that nothing else has changed, so almost all new sessions can be attributed to “shoes”.

these differences tell us a couple of things:

1. “shoes” on position 5 brings in 1000 more daily sessions compared to position 13 - 15
2. the bounce rate of the keyword is identical to other keywords that rank for this site
3. the conversion rate of this keyword seems to be significantly lower than the CR of the others. It has brought only 5 more conversions with 1000 sessions, compared to the 10 conversions with the 500 sessions we had before.

→ when analyzing this, we take a couple of factors into consideration, most importantly:

- seasonal fluctuations
- overall site performance during this time frame
- keyword performance → spike due to trends?

Luckily the SERPs are changing quite a lot, otherwise this would not be possible.

5. Training Classifications

This is where his brains kick in...

A first step towards matching keywords and sessions

Using the aforementioned generated data, the Hero tries to calculate a first probability of a certain keyword matching a cluster of sessions.

Sessions that were captured thru extensions and those where the keyword is still visibly in GA (= “hard data”) can be matched with 100% certainty, they will be left out.

→ after the first probability is calculated, the data will again be compared against the “hard data”. This happens on the cluster level to adjust the classifications.

The second step towards matching

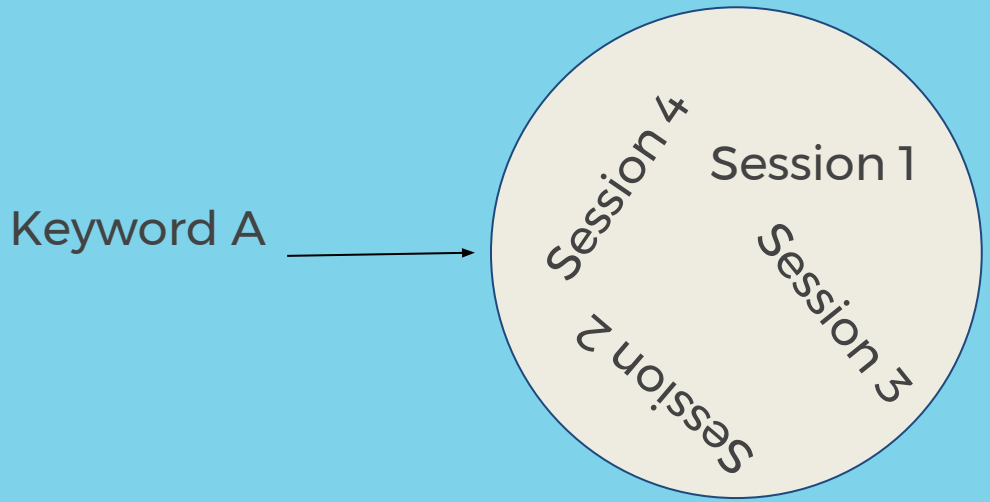
After the first iteration the Hero constructs strings from the GA data. These strings contain between 5 and 25 dimensions.

(It's interesting how few sessions are left if you query a big string - even if you have huge sites. You can check this for yourself using the Core Reporting API, where you get up to 7 dimensions in one string)

Now: the keywords have been clustered really well and you get only a handful of sessions back from one string, leaving very few potential keywords per cluster of session (the cluster has gotten smaller, too)

KEYWORD HERO

An unfair visual representation



This is a session cluster, containing four sessions.

This results from the Hero requesting all sessions with a certain **string**

Keyword A might now be matched with one or more of those sessions.

Training the classifications

What happened now: the Hero created an adjusted unique algorithm for your account, that tries to match sessions with the keywords.

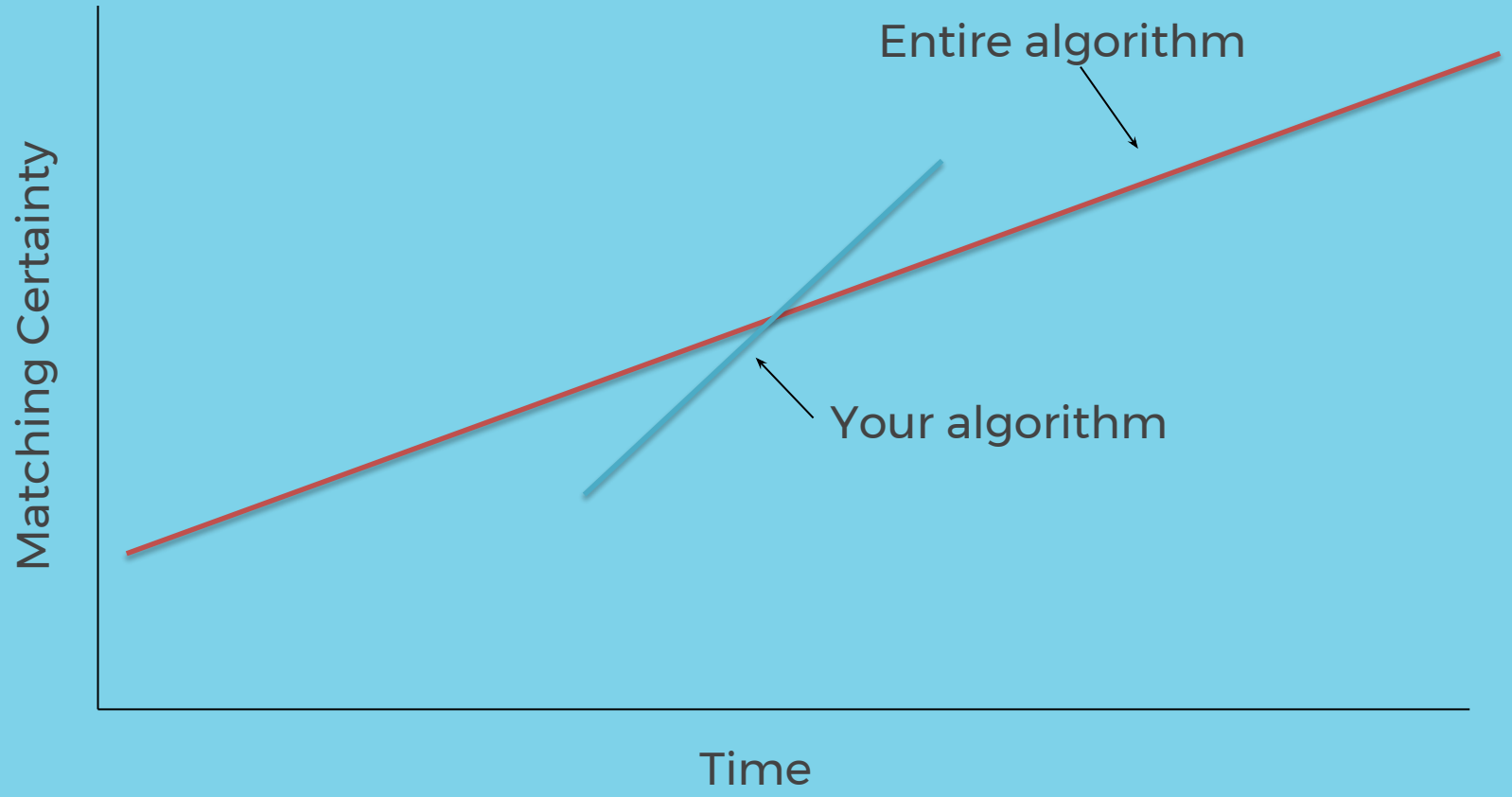
He applies this algorithm to the entire history of the GA account and then matches the results with the “hard data”, the remaining visible keywords and the keywords captured by browser extensions.

After this last step the algo is ready to go for each page.

In most cases the historic data is not enough as some data sources are not available retrospectively, which is why the Hero works far better after a month.

KEYWORD HERO

The learning curve



6. The First Computation

The magic happens...

After all that learning, the Hero can start

The process is essentially the same as the training process:

1. Computing possible keywords
2. Checking the results against the “hard data”
3. Adjust classifications based on results

Actually, quite straightforward!

7. The Final Output

What are you giving me here?

What data do you get?

Obviously the matching does not happen thru CIDs.

Looking in that mirrored account, you'll see that some dimensions are missing. This is because the matching happens based on dimension strings and **not sessions.**

KEYWORD HERO

Session vs. string-based matching

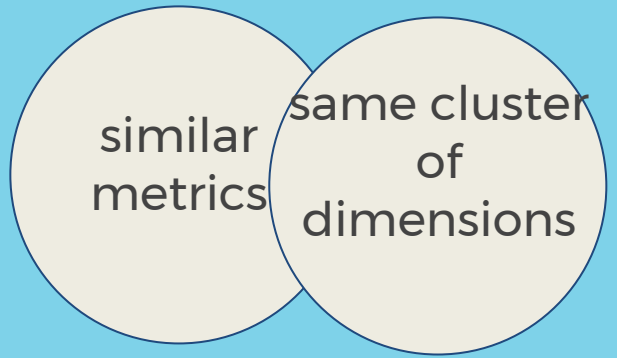
If the data was session based, you'd get all info:

{{Keyword}} ⟶ GA-CID = All Dimensions and Metrics for this CID

but what it looks like is this:

{{Keyword}} ⟶ GA-CID =

{{Keyword}} ⟶ GA-CID =



What do the 83% certainty mean?

This is just the average threshold but easier to communicate.
Really, the certainty threshold for a keyword to match with a session varies between 80% and 85%.

The certainty level is based on the assumption that 95% of all possible keywords have been found in the first bucket in page 9.

This keeps the Hero from top shape:

- you had a major redesign / relaunch / big change of your website resulting in very different metrics.
- the Hero can't recognize your brand or all of your (misspelled) brand terms. We adjust this manually.
- your GA account is in bad shape.

We hope that helped. We are certain about the validity of the mathematical concepts involved here.

We have to remain quite top level mostly, as the Hero tries to keep some secrets :-)

However, if you have any further questions about some of the concepts involved, we try our best to answer them. Just let us know @ tech-questions@keyword-hero.com

Appendix

Really???

How does the semantic clustering work?

After all possible keywords that rank were clustered and analyzed, the classification will be adjusted to what the Hero knows from analyzing the browser extension data and their users.

→

An example:

Search phrase: “How tall is the Eiffel Tower?”

Is the data set from the browser extensions large enough to attribute metrics to this keyword?

YES: how deep can you go? can you even attribute on country / device level?

NO: can we get an approximation of the performance from the columns in the dataframe?

Some stats about the browser extensions

We buy data from some browser extension conglomerates across the globe. Depending on vertical and country the ratio of desktop searches that we catch is between 0.5% and 8%.

For many URLs, however, we don't have a matching search phrase captured by a plugin. Still the data is very important as it trains the overall algorithm keyword related stats, such as:

"clicks on paid terms vs. organic"

"intention vs. vertical based click behaviour"