LOG FILES –
A DEEP DIVE
INTO YOUR DATA!

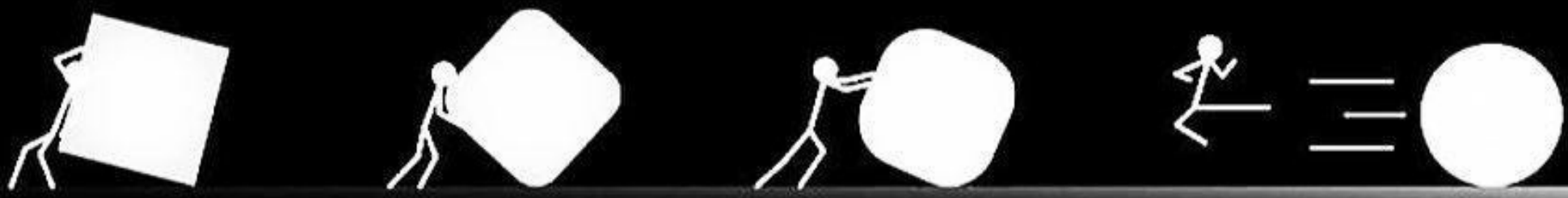
THE REAL WAY TO DETECT TECHNICAL ISSUES...

THAT NOBODY KNOWS ABOUT


SKIPPING THE BORING PART WITH CAT CONTENT


RESOURCES
BEGINNERS GUIDE
APACHE LOGS
NGINX LOGS
CLOUDFLARE LOGS
ELK STACK
USE CASES
CRAWLER
OVERVIEW

ALL DATA PROJECTS ARE DIFFICULT TO START



credits @addyosmani


ENHANCE YOUR VIEW!

DEPENDING ON ARCHITECTURE:

- TIMINGS
- CACHING INFOS
- EDGE / ORIGIN INFOS
- CUSTOM CONFIGS

TIMINGS

FOR WHICH RESOURCES WE HAD HIGHER TIME TO FIRST BYTE RESPONSES THAN THE AVERAGE?

THINK ABOUT: PARAMETERIZED URLS, RENDERING, LOAD BALANCING, DB QUERIES, API RESPONSES, CAMPAIGN URLS

CACHING

HOW IS THE CACHE HIT RATIO EVOLVING FOR EACH MIME TYPE?

THINK ABOUT: UNCACHED ELEMENTS, CACHING GUIDELINES, ARCHITECTURE PERFORMANCE

EDGE / ORIGIN SERVER

WHICH RESOURCES HAVE HIGH DNS RESOLVING TIMES ON THE ORIGIN SERVER?

THINK ABOUT: SERVER CONFIGS, EDGE WORKERS, SERVE RATIO FROM THE EDGE

CUSTOM CONFIGS

LOG THE ACCESS LOGS FOR GOOGLE-CRAWLER IN A CUSTOM FORMAT WITH CUSTOM METRICS

THINK ABOUT: CONDITIONAL LOGGING, LOG SEGMENTATION, STANDARDIZATION


CONNECT
THE DATA!

YOUR OWN DATA:

- GOOGLE SEARCH CONSOLE
- RANKING DATA, SEARCH VOLUME
- OWN CRAWLS
- BUSINESS METRICS

GSC

WHAT IS THE AVERAGE RANKING OF HIGHLY OR LESS CRAWLED URLS?

THINK ABOUT: CRAWL2IMPRESSION, RANK2CRAWLRATE

SEARCH VOLUME

WHICH URLS MAPPED TO HIGH SV KEYWORDS ARE NOT CRAWLED?

THINK ABOUT: POTENTIAL OF UNCRAWLED URLS, CRAWL BEHAVIOR OF MONEY URLS

OWN CRAWLS

HOW IS THE RELATIONSHIP BETWEEN INTERNAL LINKING SCORE AND CRAWL RATE?

THINK ABOUT: UNCRAWLED URLS, UNIMPORTANT CRAWLINGS, INDEXATION VS CRAWLING

BUSINESS KPIS

ARE ALL OF MY HIGH MARGIN PRODUCTS ON GOOGLEBOTS RADAR?

THINK ABOUT: STATUSCODE OF BESTSELLERS, RESPONSE TIMES OF BESTSELLERS


HOW TO PREPARE?

DATA RETENTION RATE

- LAST 3 MONTH NEARLINE
- LAST 12 MONTH IN COLDLINE
- LAST 24 MONTH IN ARCHIVE

EVERYTHING ELSE IS TOO MUCH!

CONDITIONAL LOGGING

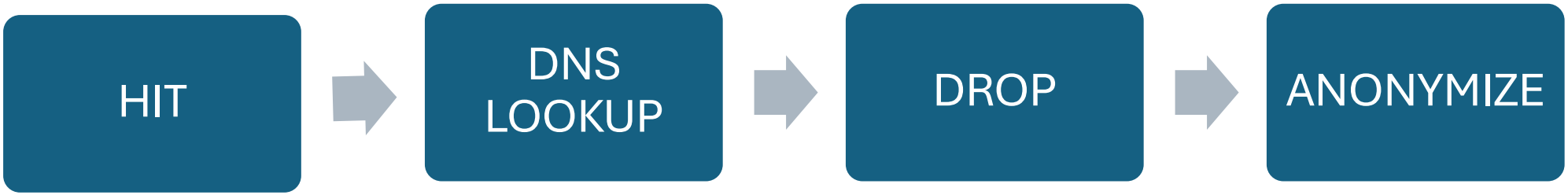
- PUSH ONLY IF REVERSE DNS IS FROM GOOGLE
- HITS FROM CUSTOMERS CONTAINS MORE FIELDS
- LOG KNOWN BOTS DIFFERENTLY

REDUCES STORAGE UP TO 90-95 PER CENT

IN GERMANY WE SAY:
AUFTRAGSDATENVERARBEITUNGS-VERTRAG

credits @liamcarpenter

ANONYMIZE ASAP!



DO NOT RELY ON THIRD PARTY TOOLS!

ANONYMIZING

66.249.66.1 → 66.249.0.0

66.249.66.1 → CC456FF2392XXZ 

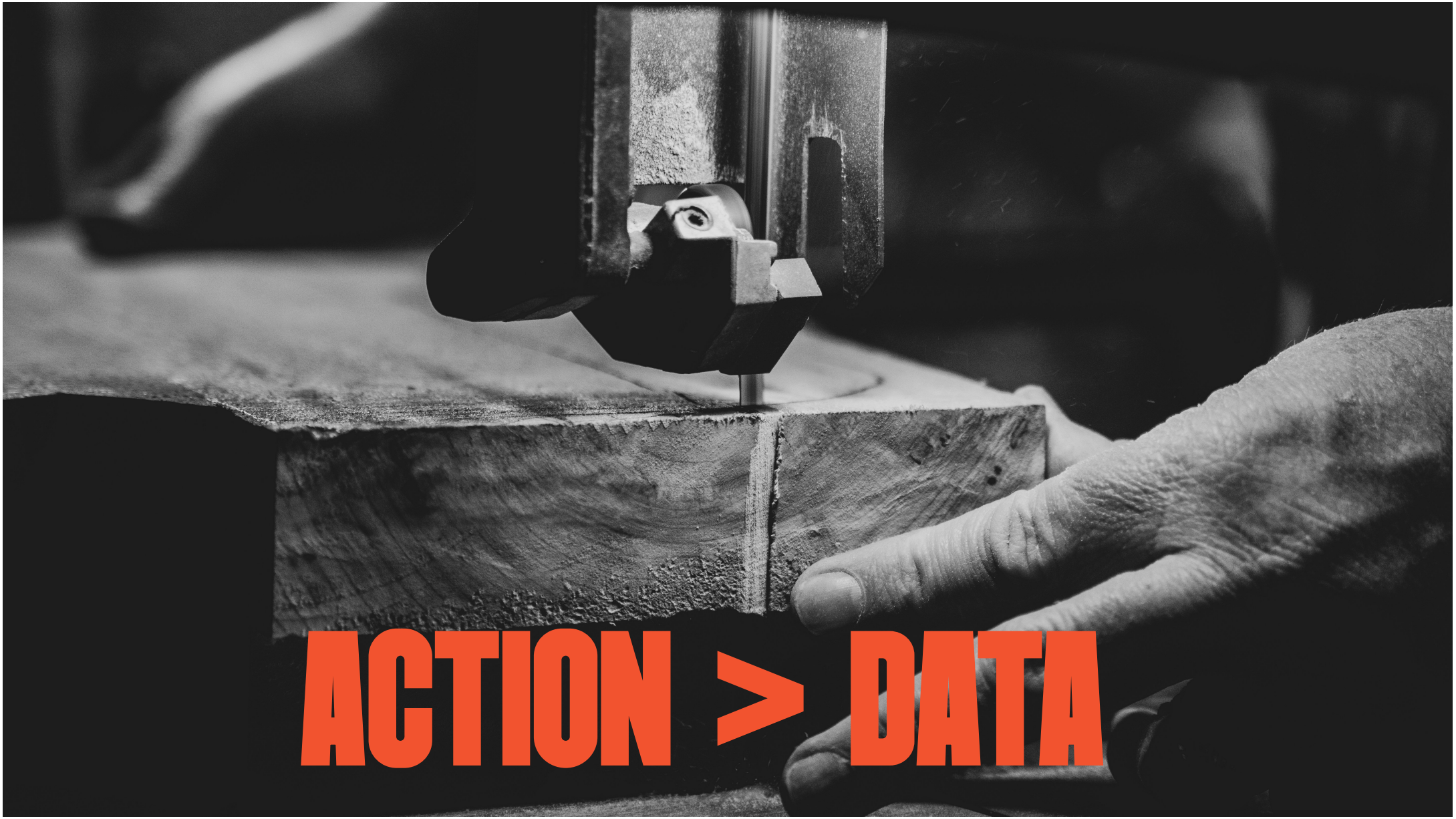
66.249.66.1 → GOOGLEBOT OR NOT!

NO DATA NO COMPLIANCE!

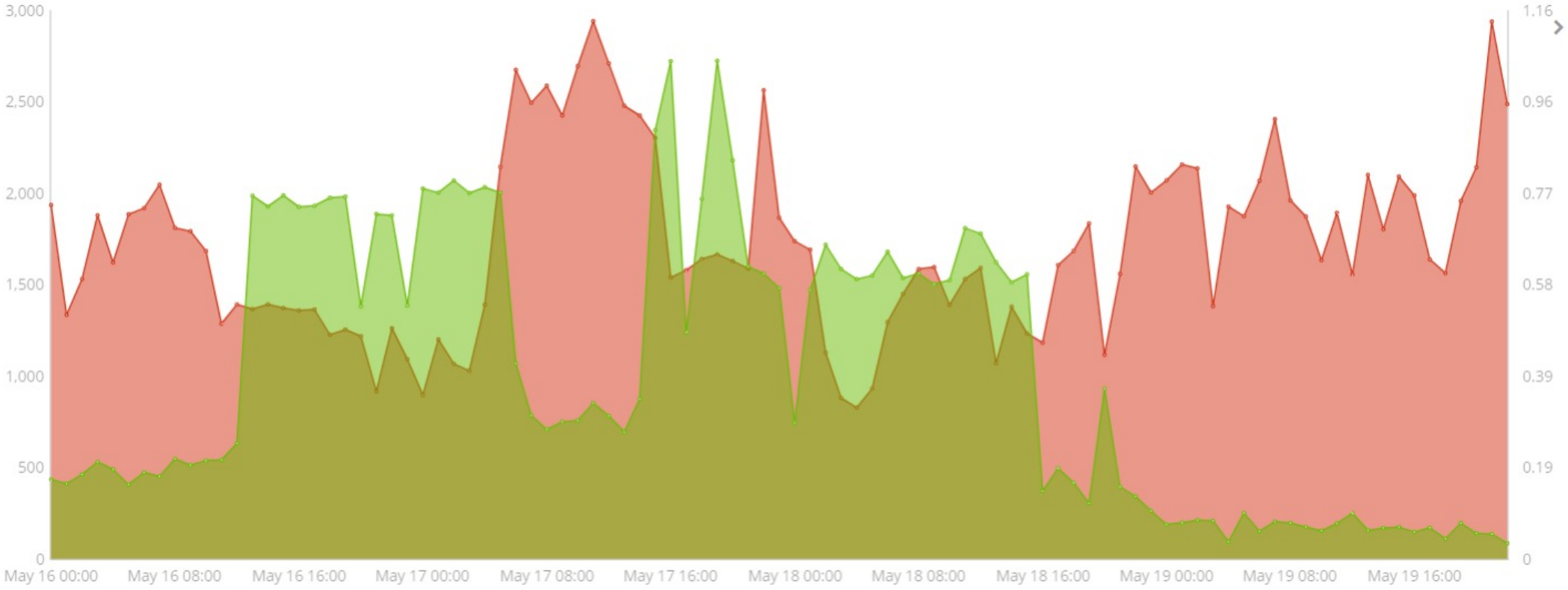
DUTIES

- CONTRACT FOR DATA PROCESSING
- USER RIGHTS
- PRIVACY STATEMENT
- OLD DATA MUST BE DELETED / MASKED

UP TO 20 MILLION EURO OR 4% REVENUE!


ACTION > DATA

SLOW REQUESTS — CRAWLRATE



SLOW REQUESTS — ON ORIGIN

Location	DNS	Connect	TLS	TTFB
Frankfurt	4.3ms	17ms	200ms	388ms
San Francisco	31ms	152ms	758ms	1300ms
Paris	15ms	18ms	250ms	528ms
Amsterdam	12ms	10ms	197ms	426ms

ANALYZE GRANULARLY - THINK GLOBAL

SLOW REQUESTS — PARAMETERS

/CAT/BRAKEDISCS → 800MS TTFB

/SEARCH?=BRAKEDISCS → 1800MS TTFB

PARAMETERS ARE EXPENSIVE!

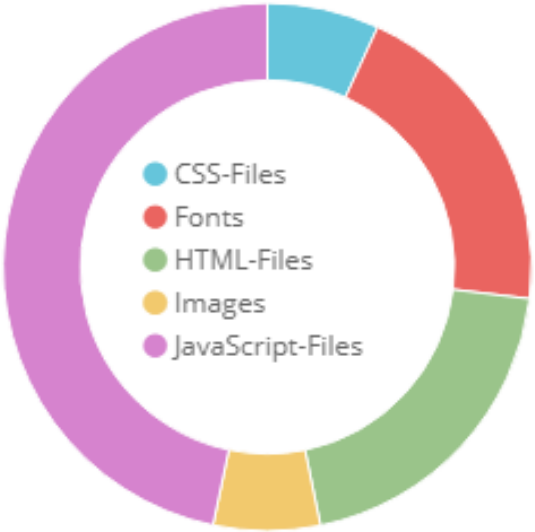
SLOW REQUESTS — CAMPAIGNS

/PRODUCT/DVC-1  NOINDEX, BUT HIGHLY REQUESTED

/TRACKING  NOT SEO RELEVANT

SWITCH YOUR MINDSET

SLOW REQUESTS — COMPRESSION



CSS-Files	3.018GB
Fonts	8.832GB
HTML-Files	8.738GB
Images	2.872GB
JavaScript-Files	20.553GB

DECIDE WHERE TO TWEAK COMPRESSION-RATE

CACHING — MISS HIT RATIO

PER RESOURCE

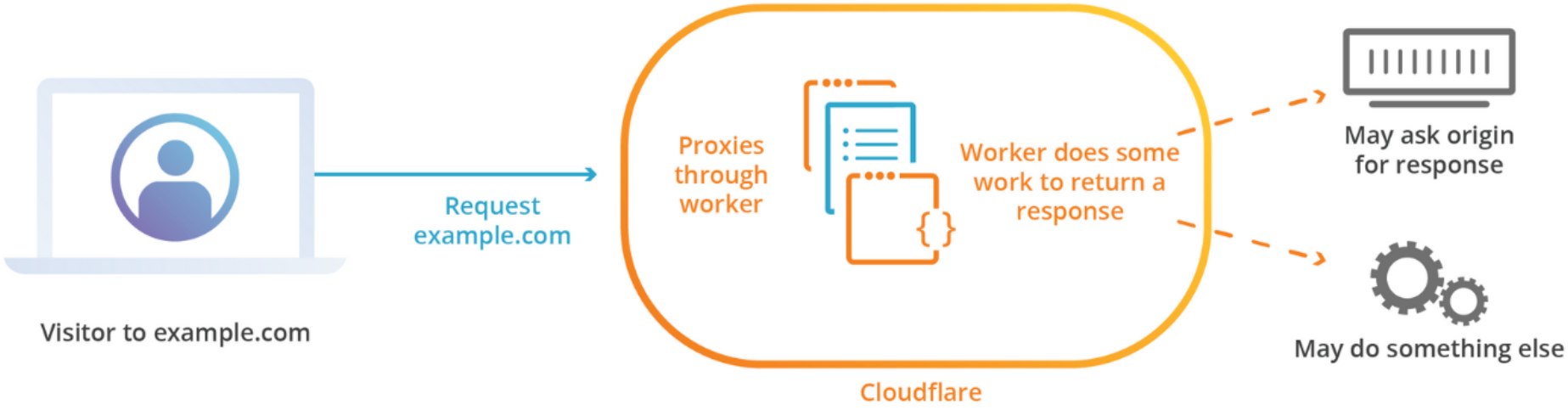
- IMAGES
- STATIC RESPONSES
- THIRD PARTIES

PER SEGMENT

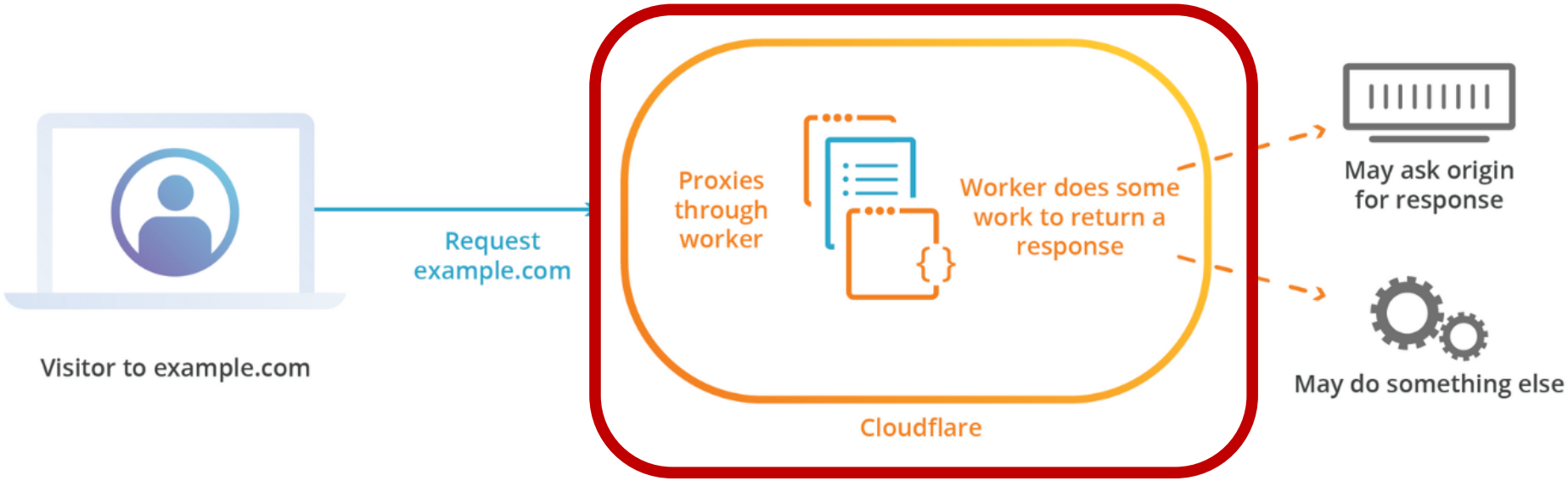
- PRODUCT PAGES
- TRACKING/RETURN PORTAL
- SEARCH

NO REQUEST IS THE BEST REQUEST

EDGE — PERFORMANCE

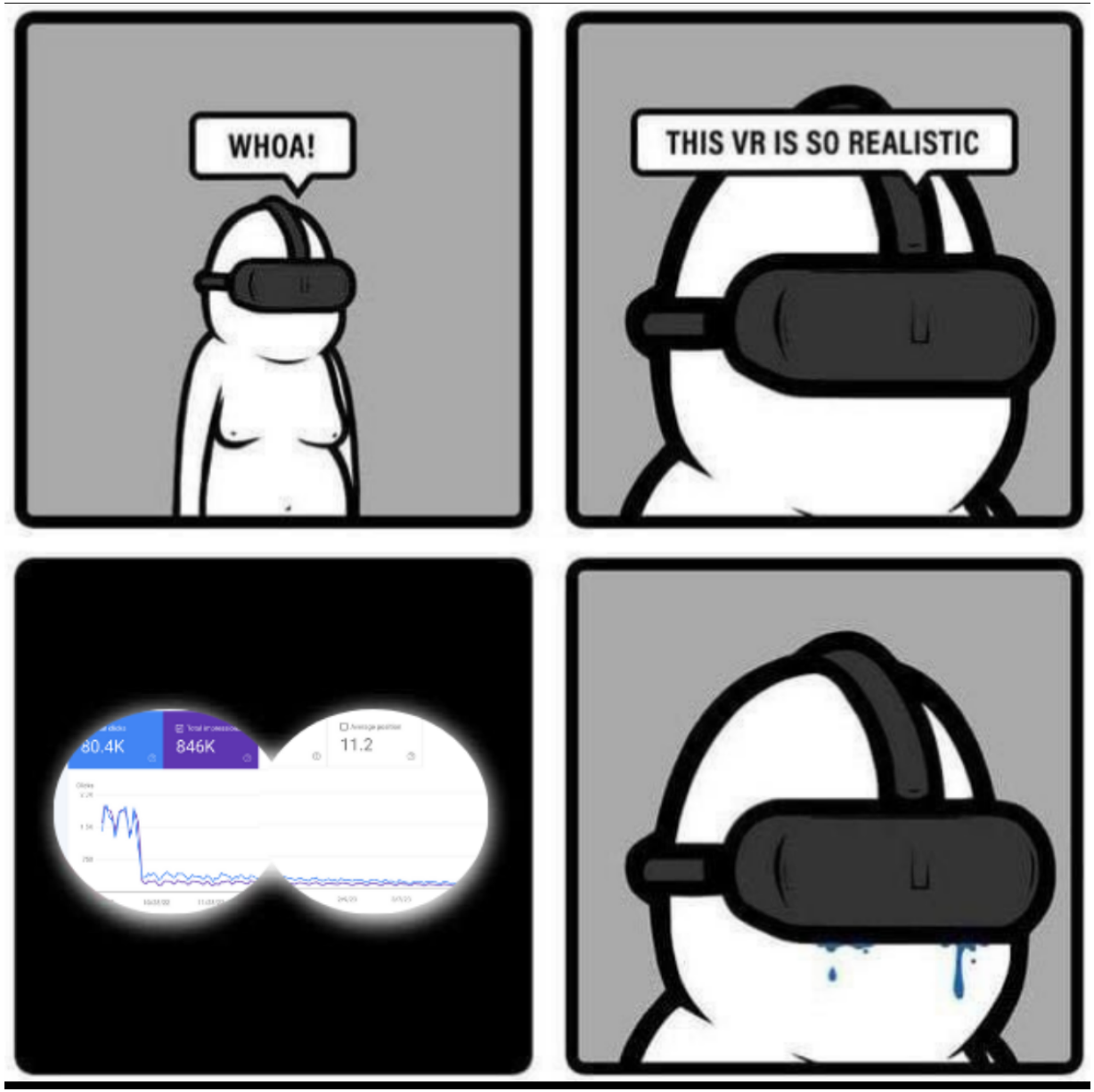


EDGE — PERFORMANCE



HOW DO YOU QUANTIFY THE IMPACT?

GIVE GSC DATA MORE CONTEXT!



GSC — COMBINE YOUR DATA

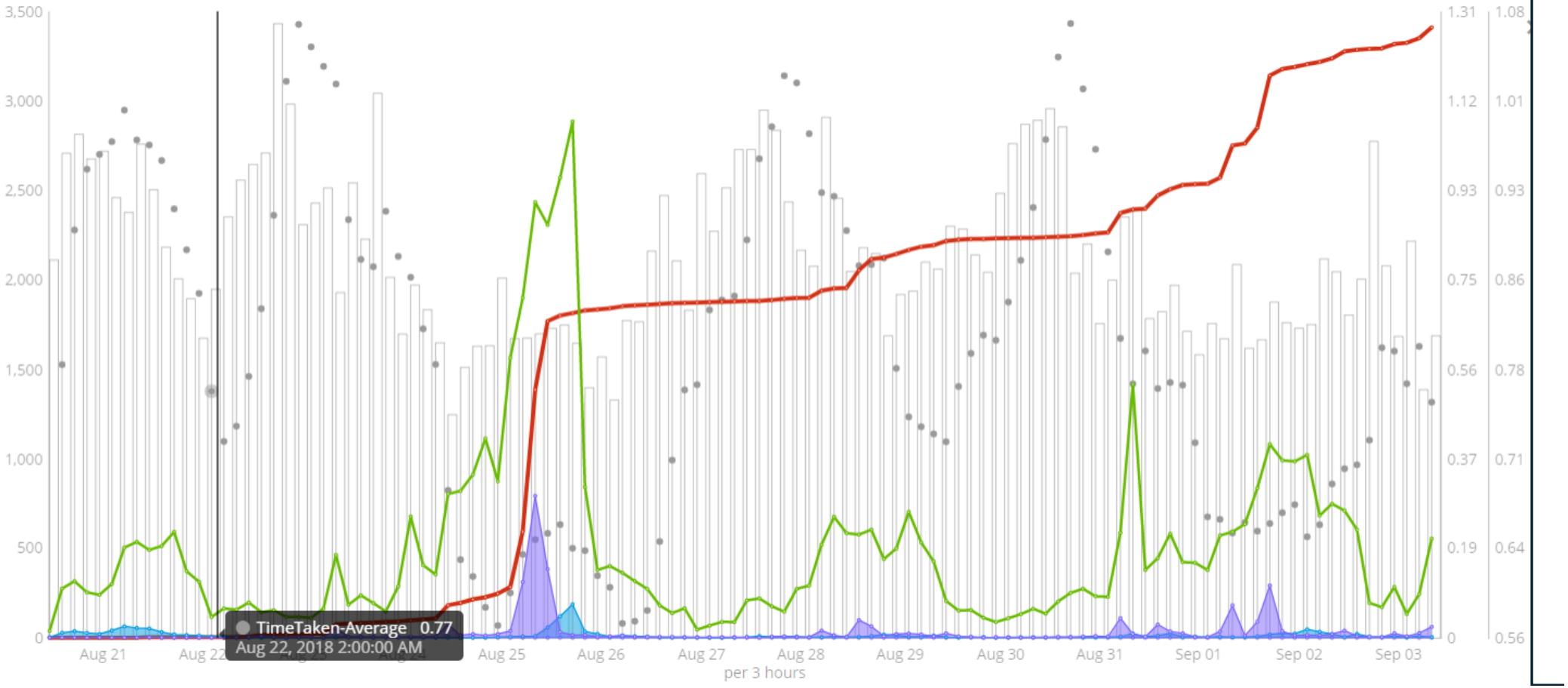
MAP PERFORMANCE DATA TO URLS

- ARE CRITICAL URLS CRAWLABLE?

- STATUS CODE ISSUES WITH HIGH-TRAFFIC URLS?

- CRAWLING TO IMPRESSION SPEED

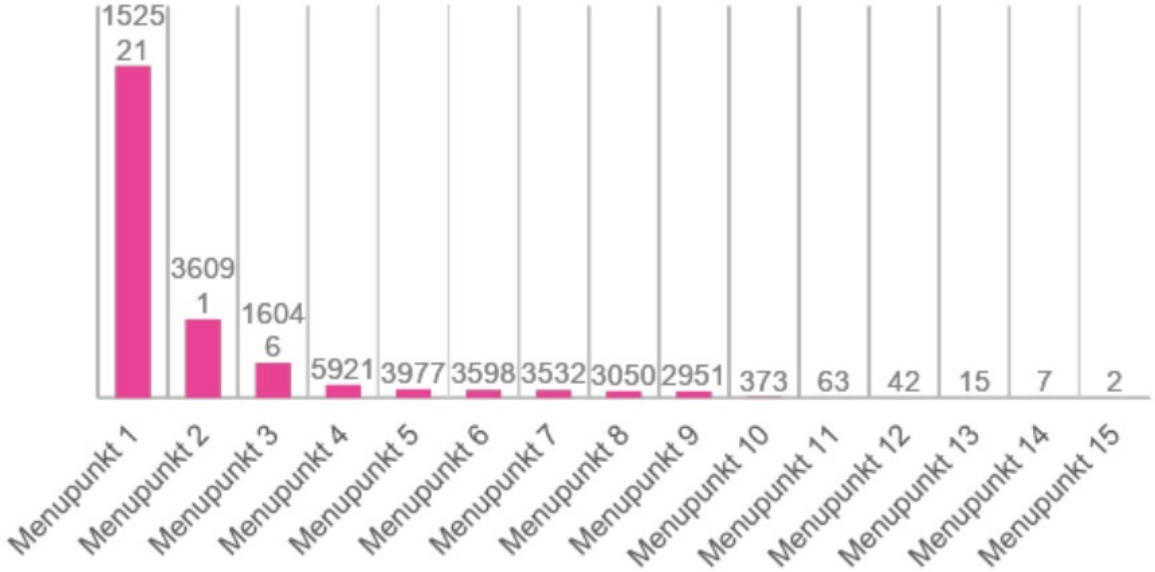
GSC — COMBINE YOUR DATA



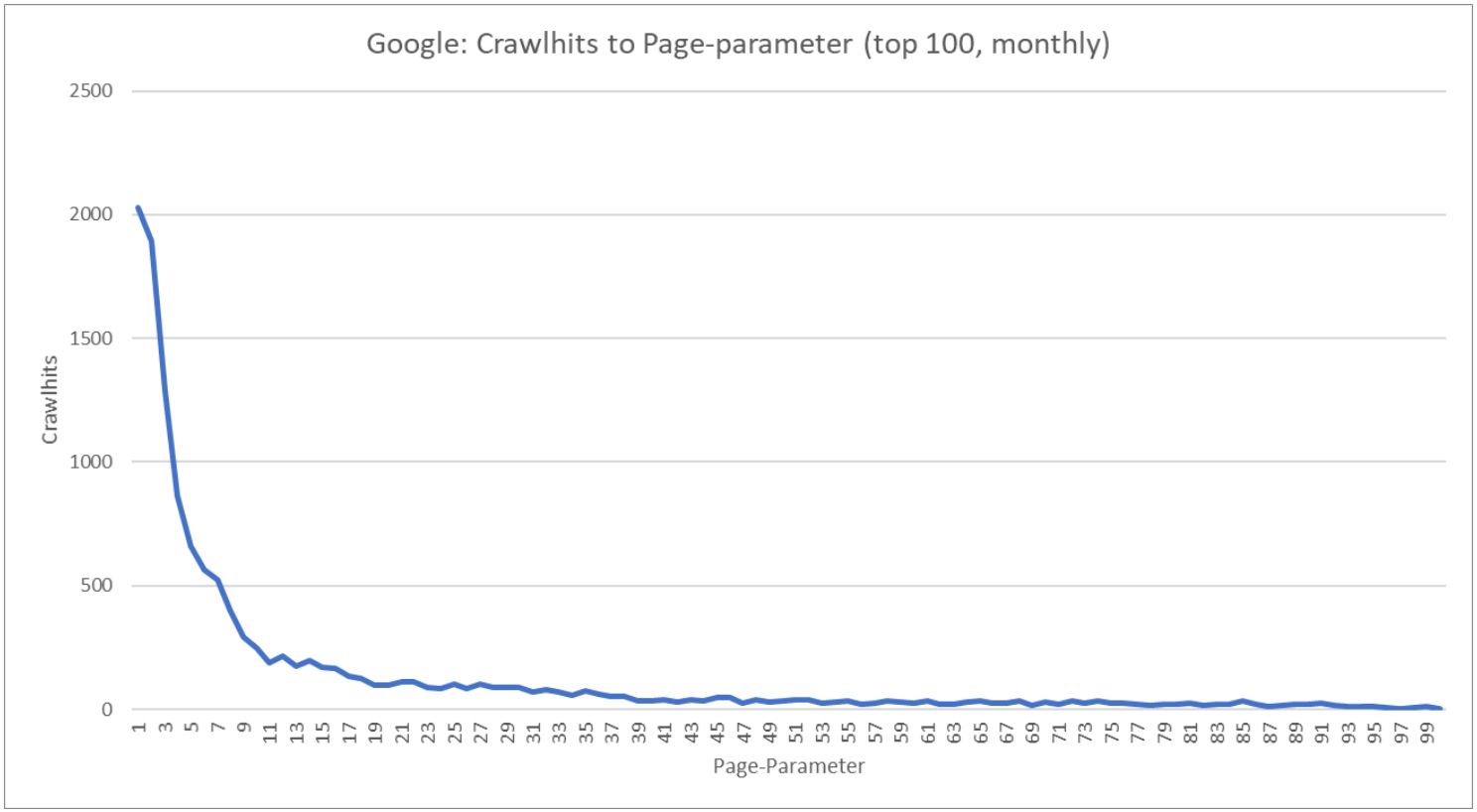
CRAWLS – THE MISSING PART

INTERNAL LINKING AUDITS

- ORDER IN NAVI IS KEY
- CRAWLING IS REFLECTING YOUR PRIORITIES



CRAWLS — DISCOVERY



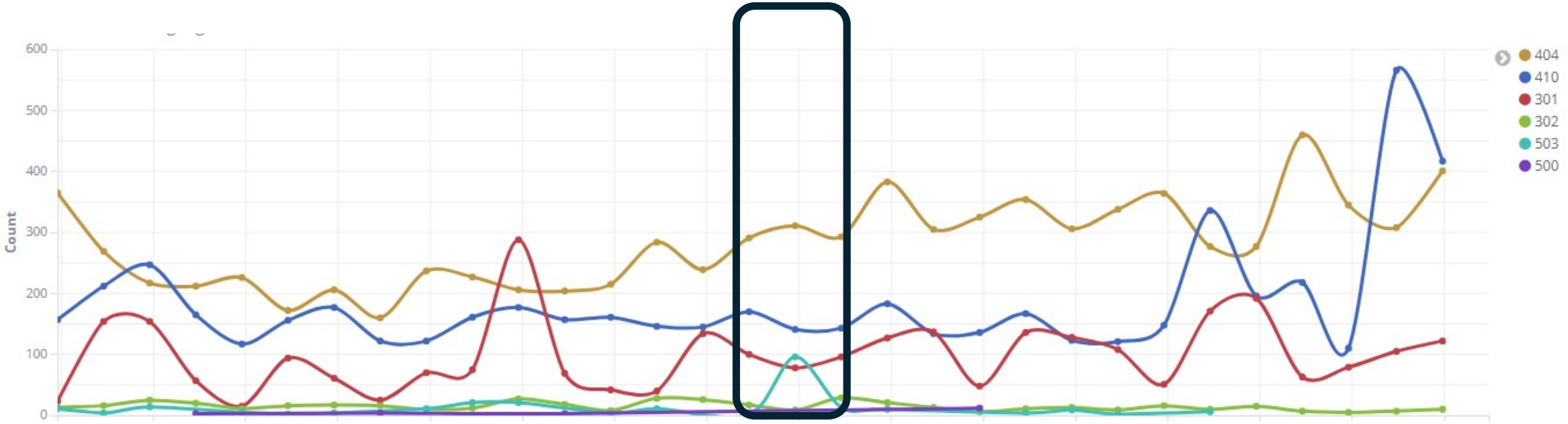
BUSINESS KPIS = VALUE

QUANTIFY CRAWLING EFFECTS

- BESTSELLERS ARE NOT CRAWLED!
- OUT OF STOCK IS KILLING CRAWLING -> RANKING
- SLOW RESPONDING CAMPAIGN URLS

BUSINESS DECISIONS

HOW WE FUCKED UP OUR FRONTEND SERVICE



RANDOMLY INCREASING 5XX RESPONSES?

ASK QUESTIONS FIRST INSTEAD FOR MORE DATA