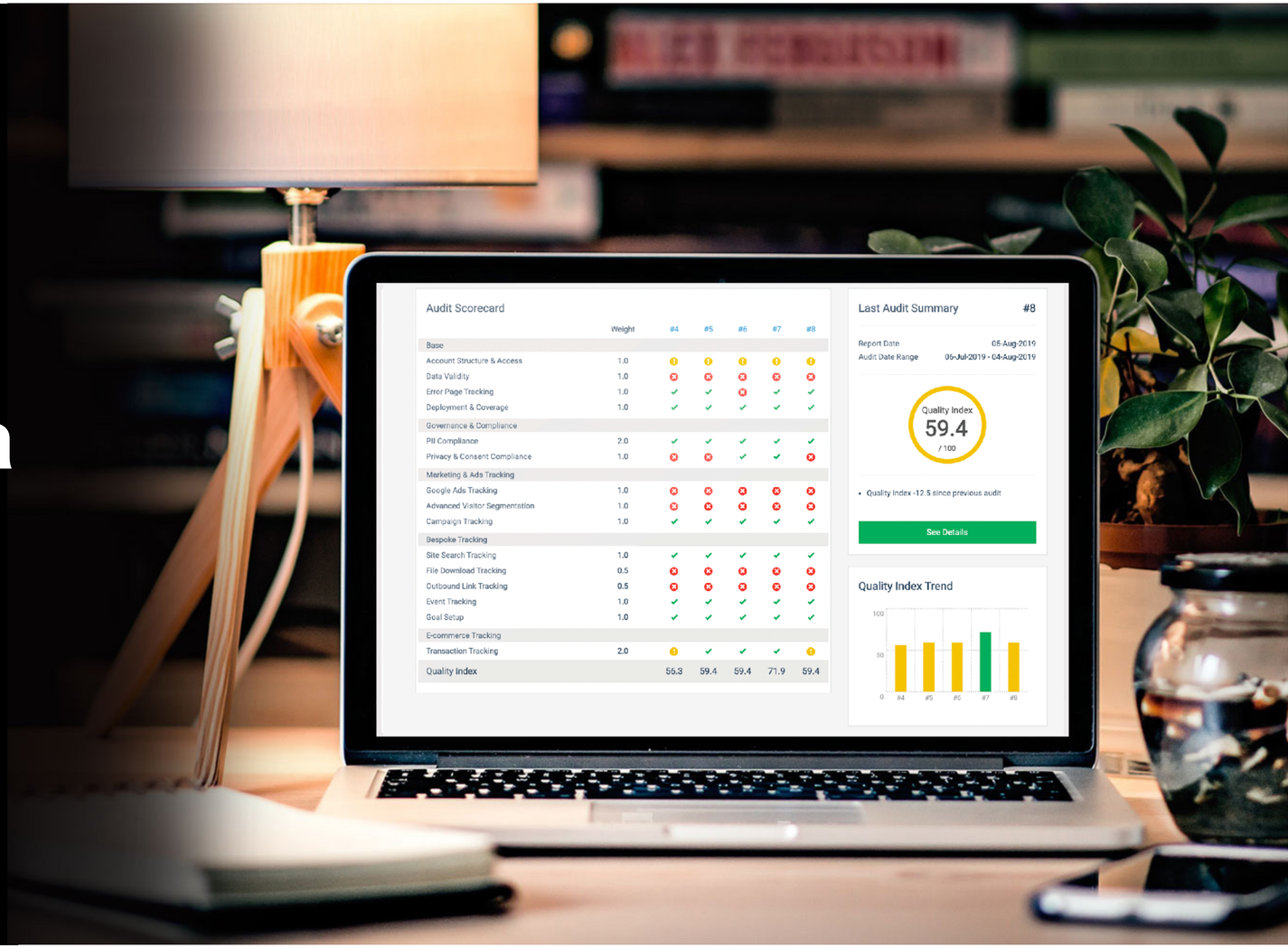
Verified Data

Google Analytics and How to Avoid Bad Data

An enterprise study of data quality



Brian Clifton - Author | co-founder Verified-Data.com

About Brian Clifton

Circa 1974...!

Trained as a scientist
(Chemistry)

Urchin

Google

Head of Web Analytics
EMEA 2005-8

100,000 copies sold

Automated audit tool

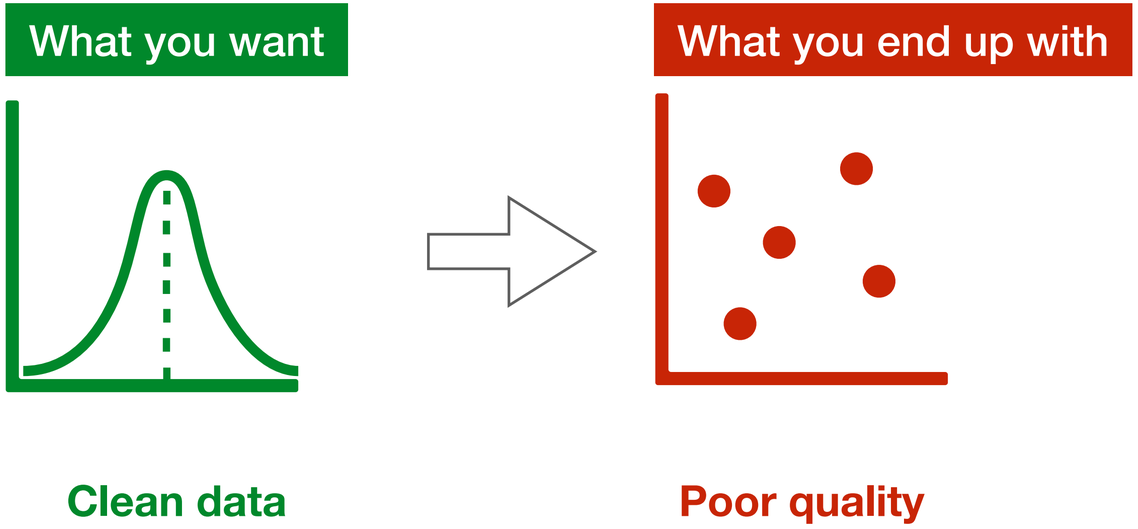
Verified Data

"Not really... Maybe just some of the numbers"

Why this lack of trust...?

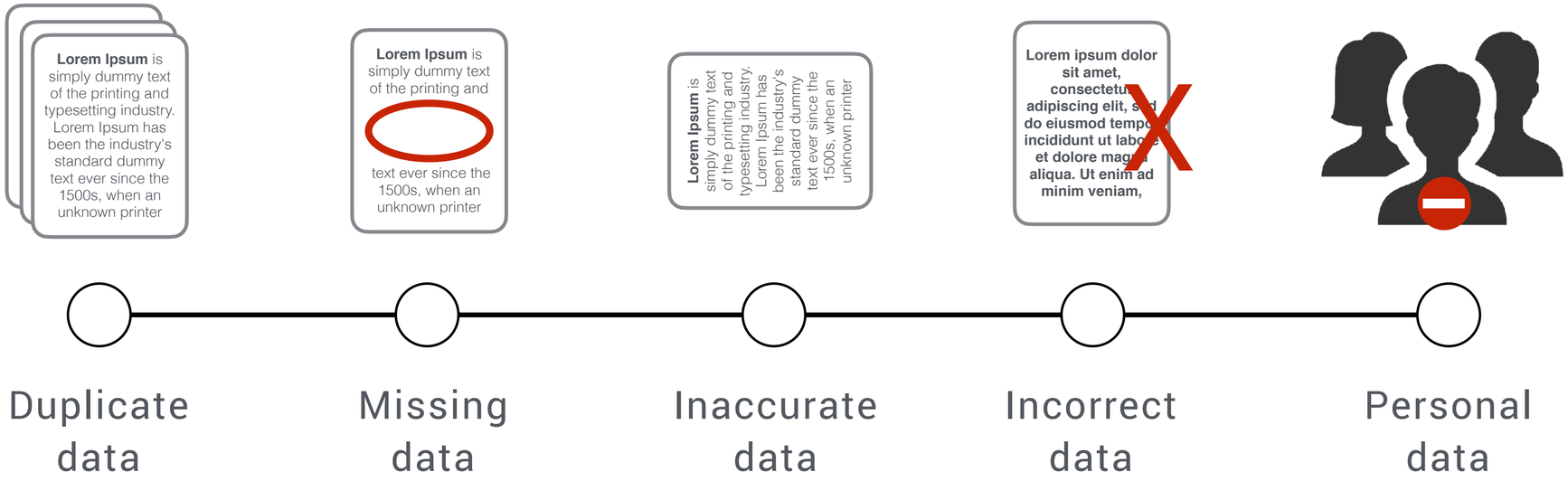
- **Setup** is often not understood
- Collection is **rarely verified** - often "*smells*"
- **Poor governance**

  - particularly Google Analytics



Bad data is easy to define...

BAD DATA IS...



Duplicate data | Missing data | Inaccurate data | Incorrect data | Personal data

But its **hard** to find because...

- Websites constantly change
- Changes usually performed by "*others*"
- Time pressures - campaign "*laziness*"
- **Bad data looks just like good data!**

  "Needles in impossible haystacks."

The Data Quality Study

(the method)

Verified Data

The how...

Scorecard

Weight scorecard technique

Governance

Accuracy

Summary of over 200 data quality tests

	Weight	Website 1	Website 2	Website 3	Website 4
Core Principals					
Account Structure	1.0	!	!	!	!
Data Validity	1.0	X	X	!	!
Error Page Tracking	1.0	X	✓	✓	✓
Deployment & Coverage	1.0	!	!	✓	X
Privacy & Compliance					
PII Compliance	2.0	X	✓	X	X
Privacy & Consent Compliance	1.0	✓	✓	✓	✓
Cookie Report	1.0	!	!	✓	!
Marketing & Engagement					
Google Ads Tracking	1.0	X	X	X	X
Advanced Visitor Segmentation	1.0	✓	X	✓	✓
Campaign Tracking	1.0	X	X	!	✓
Site Search Tracking	1.0	X	!	!	!
File Download Tracking	1.0	X	✓	!	X
Outbound Link Tracking	1.0	!	✓	✓	!
Event Tracking	1.0	!	X	!	!
Goal Setup	1.0	✓	X	X	X
E-commerce					
Transaction Tracking	2.0	X	X	!	✓
	Quality Index	**29.4**	**40.7**	**53.6**	**39.3**

The how...

Scorecard

	Weight	Website 1	Website 2	Website 3	Website 4
Core Principals					
Account Structure	1.0	!	!	!	!
Data Validity	1.0	X	X	!	!
Error Page Tracking	1.0	X	✓	✓	✓
Deployment & Coverage	1.0	!	!	✓	X
Privacy & Compliance					
PII Compliance	2.0	X	✓	X	X
Privacy & Consent Compliance	1.0	✓	✓	✓	✓
Cookie Report	1.0	!	!	✓	!
Marketing & Engagement					
Google Ads Tracking	1.0	X	X	X	X
Advanced Visitor Segmentation	1.0	✓	X	✓	✓
Campaign Tracking	1.0	X	X	!	✓
Site Search Tracking	1.0	X	!	!	!
File Download Tracking	1.0	X	✓	!	X
Outbound Link Tracking	1.0	!	✓	✓	!
Event Tracking	1.0	!	X	!	!
Goal Setup	1.0	✓	X	X	X
E-commerce					
Transaction Tracking	2.0	X	X		
		29.4			

Sum and normalise

Quality Index score out of 100

Section score

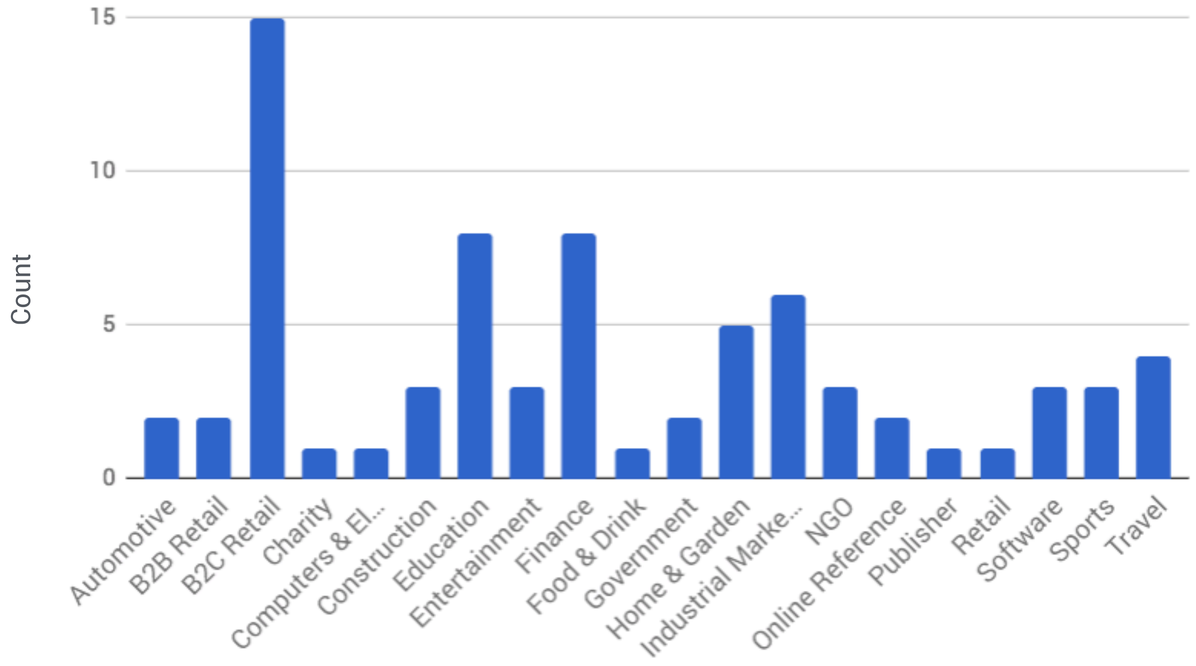
✓ = 10

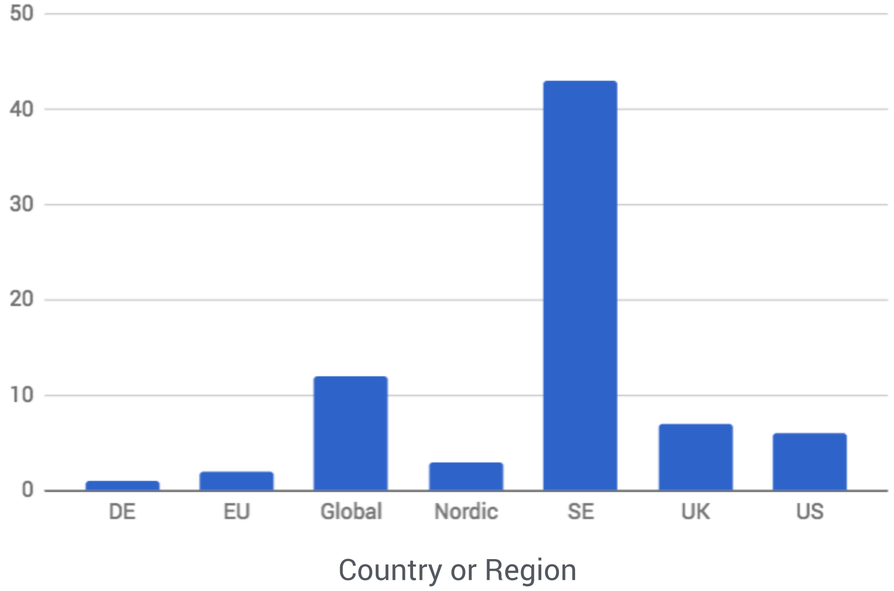
! = 5

X = 0

Study: 75 **enterprise** sites

- All **brand leaders**, 13 with a global presence.
- **41**% had an e-commerce facility.
- Monthly visits range from 100k - 100 million (three > 100M).





The Results

(a visualising data quality)

Verified Data

Scorecard

	Weight	Website 1	Website 2	Website 3	Website 4
Core Principals					
Account Structure	1.0	!	!	!	!
Data Validity	1.0	X	X	!	!
Error Page Tracking	1.0	X	✓	✓	✓
Deployment & Coverage	1.0	!	!	✓	X
Privacy & Compliance					
PII Compliance	2.0	X	✓	X	X
Privacy & Consent Compliance	1.0	✓	✓	✓	✓
Cookie Report	1.0	!	!	✓	!
Marketing & Engagement					
Google Ads Tracking	1.0	X	X	X	X
Advanced Visitor Segmentation	1.0	✓	X	✓	✓
Campaign Tracking	1.0	X	X	!	✓
Site Search Tracking	1.0	X	!	!	!
File Download Tracking	1.0	X	✓	!	X
Outbound Link Tracking	1.0	!	✓	✓	!
Event Tracking	1.0	!	X	!	!
Goal Setup	1.0	✓	X	X	X
E-commerce					
Transaction Tracking	2.0	X	X	!	✓
	Quality Index	**29.4**	**40.7**	**53.6**	**39.3**

75 enterprise websites

Data quality patterns - massive failures...!

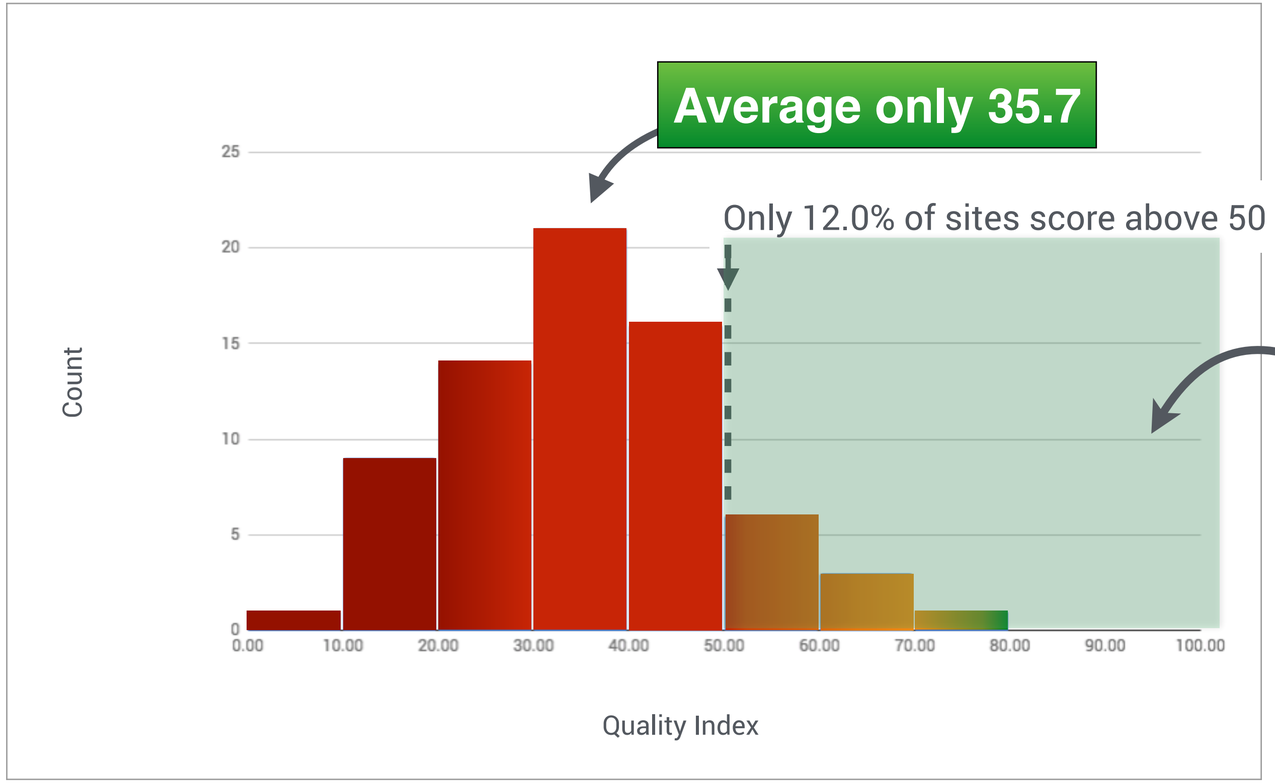
Website audit

75 websites

Account Setup & Governance	✓	!	✓	✓	!	x	!	✓	✓	x	✓	✓	x	x	!	!	x	!	✓	✓	✓	x	x	✓	!	!	✓	!	!	x
GATC Deployment	✓	x	✓	✓	!	✓	!	!	✓	x	!	x	✓	✓	x	!	!	✓	✓	✓	✓	✓	!	✓	✓	✓	✓	✓	✓	✓
AdWords	x	x	x	✓	!		!	!	✓	✓	!	✓	!	x	✓	!	!	x	!		!	✓	✓	✓	x	✓	✓	✓	✓	
Site Search Tracking	!	x	x	✓	x	x	x	✓	!	!	x	x	✓	x	x		!	x	!	✓	✓	!	!	x	x	✓		✓	x	
File Download Tracking	x				!	x	✓	x	✓	x	x	x	✓	✓	x	✓	x	x	x	x	x	✓	x	x	x		x	x	x	x
Outbound Link Tracking	x		!	x	x	x	x	x	✓	x	x	x	x	✓	x	✓	x	x	x	x	!	x	x	x	x	x	x	x	x	x
Form Completion Tracking	x	✓		x	✓	x	✓	x		x		✓	x		x	x	✓	x	!	!		x	x		!	x	x	x		x
Video Tracking			!				x				!	x	x		x			x				✓		x	x			x	x	x
Error Page Tracking	x	✓	x	x	x	x	✓	x	✓	x	!	✓	x	✓	✓	✓	x	x	✓	x	x	x	✓	x	✓	✓	x	x	x	✓
Transaction Tracking		!						✓	!	x		!		x	x			x	!	✓	x	✓	✓	x		!				
Event Tracking	!			✓	✓			x		!	!	x	✓		!	✓		x	✓	x	!	✓	x		x	!	x	x	x	x
Goal Setup	!	x	!	x	✓	x	x	x	!	!	!	!	x	!	x	!	x	x	!	✓	x	!	!	x	!	!	x	x	!	x
Funnel Setup	!	x	x	x		x		x	x	x	!	x	x	!	x	x	x	x	x	!	x	x	✓	x	x	!	x	x	!	
Visitor Segmentation	x	x	x	x	x	x	x	x	x	x	x	x	x	x	!	✓	x	x	x	!	x	x	x	x	!	!	!	x	x	x
CampaignTracking	!	!	✓	✓	!	!	x	x	✓	✓	✓	!	!	✓	!	!	x	!	!	x	!	!	✓	x	x	✓	x	!	!	x

Quality Index

The Quality Index score...



Only analyse your data if you are in this area

Lowest = 4.5

Highest = 73.1

Breakdown of results: verified-data.com/study

	Correct	Amber	Wrong or missing
Account Setup & Governance	24%	44%	32%
GATC Deployment	54%	32%	14%
AdWords	50%	25%	25%
Site Search Tracking	20%	30%	50%
File Download Tracking	34%	5%	61%
Outbound Link Tracking	19%	10%	71%
Form Completion Tracking	24%	16%	60%
Video Tracking	8%	23%	69%
Error Page Tracking	50%	3%	46%
Transaction Tracking	26%	19%	55%
Event Tracking	25%	36%	39%
Goal Setup	18%	39%	43%
Funnel Setup	8%	18%	73%
Visitor Segmentation	7%	8%	85%
CampaignTracking	18%	47%	35%

1 in 5 sites have PII issues

1/4 track conversions correct

82% track campaigns wrong!

Examples of data quality issues

Personally Identifiable Information

1 in 5

sites have

PII issues

PII most common in **URLs** and **page titles**...

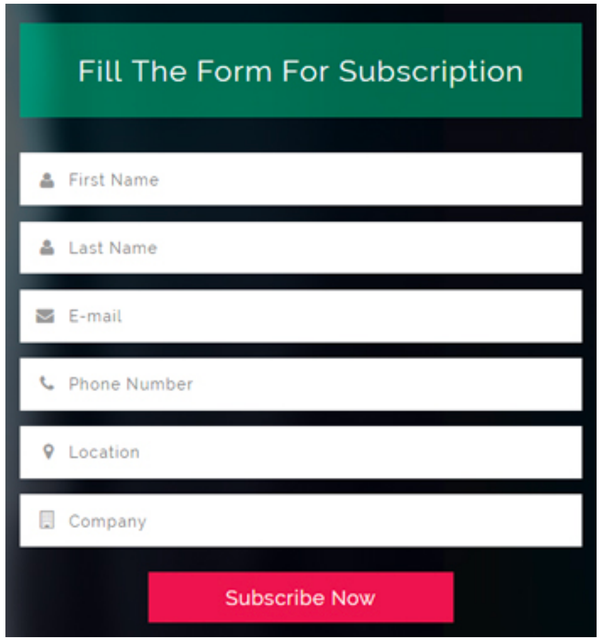
Usually by accident...

Page	Page Views
	13,382 % of Total: 0.10% (13,216,720)
1. /existing-appointment/log-in/?ins=wt16&email=ben_l[redacted]@yahoo.com&password=daxiMgjxG4	59 (0.44%)
2. /existing-appointment/log-in/?ins=wt16&email=[redacted]@aol.com&password=ngJ3UIifE7	45 (0.34%)
3. /existing-appointment/log-in/?ins=wt16&email=robbr[redacted]@hotmail.com&password=zMVMbBheN7	44 (0.33%)
4. /existing-appointment/log-in/?ins=wt16&email=ched[redacted]hotmail.co.uk&password=QF5HoNwzK8	40 (0.30%)
5. /existing-appointment/log-in/?ins=wt16&email=[redacted]ng@gmail.com&password=DxCe4pIcI5	36 (0.27%)
6. /existing-appointment/log-in/?ins=wt16&email=mcco[redacted]hotmail.co.uk&password=RLGt8tXnA9	33 (0.25%)
7. /existing-appointment/log-in/?ins=wt16&email=shini[redacted]hotmail.com&password=9vVnbQ7oZ8	33 (0.25%)
8. /existing-appointment/log-in/?ins=wt16&email=TON[redacted]YSECURITYGROUP.CO.UK&password=et7eNSylF4	33 (0.25%)

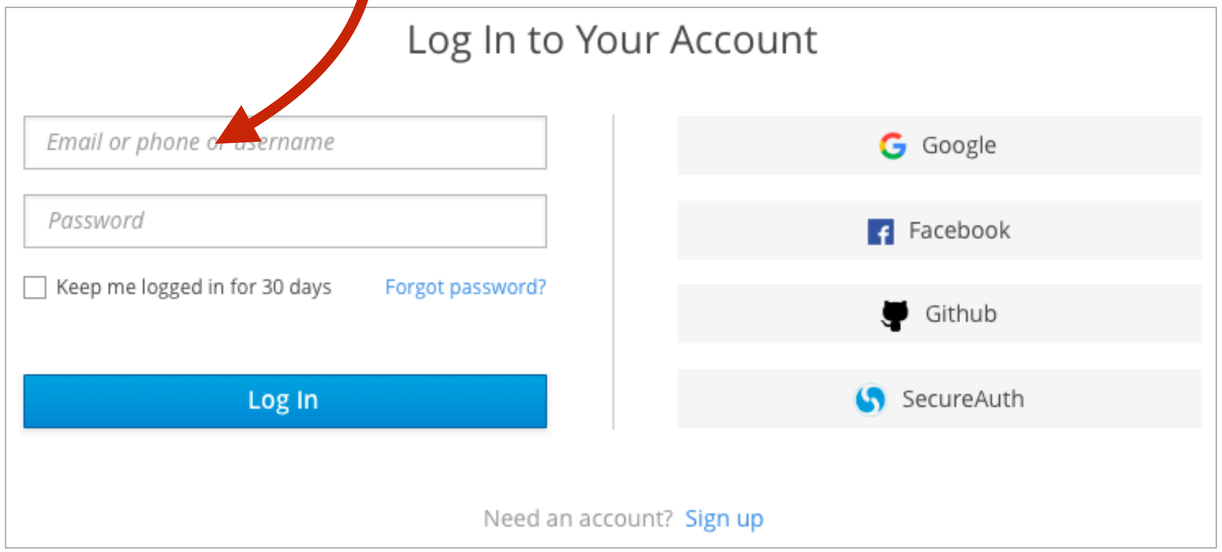
Show rows: 10 Go to: 1 1-10 of 4875

Where does "accidental" PII come from...?

Via forms submitted using GET method...



<title> tags: "*Welcome Brian@...*"



Can also be captured deliberately...

Property Settings

User Management

.

.

.

Dd Custom Definitions

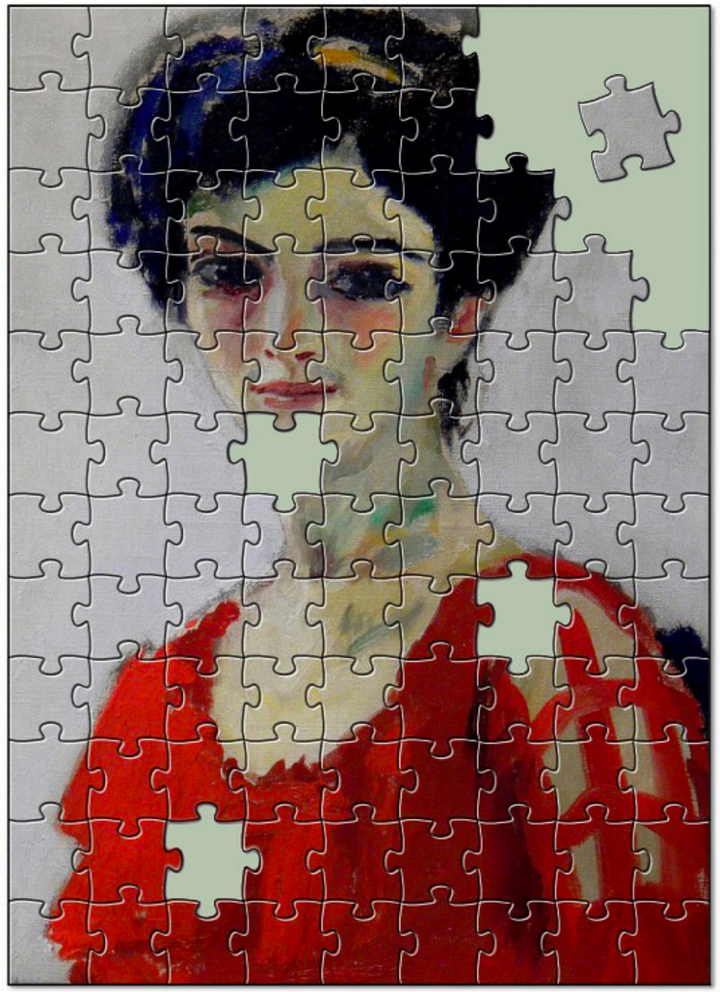
Custom Dimensions

Custom Metrics

Per field, this is not PII... BUT

- UI gender
- UI postal code
- UI city
- UI state
- UI birthday

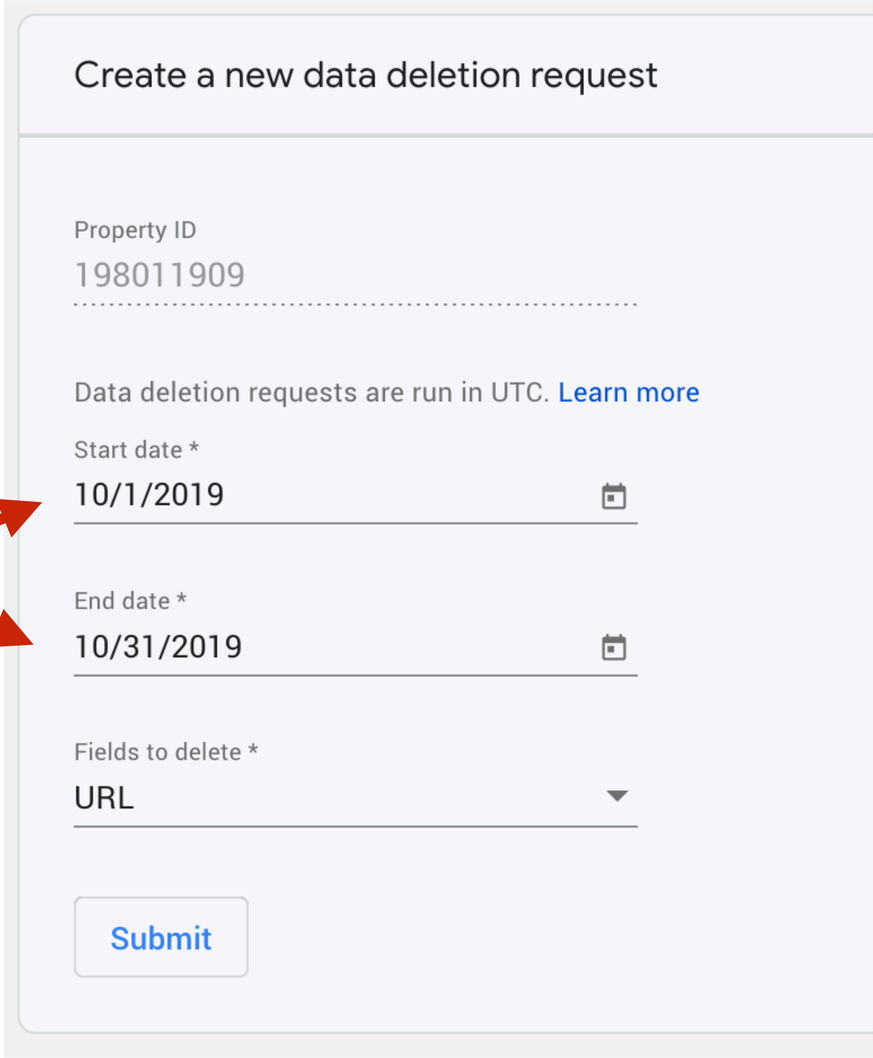
Jigsaw effect of data triangulation



Fixing historical data is **impossible**...

Only deletion is available for Google Analytics and its very blunt...

ALL URLs removed!



E-commerce examples

Only

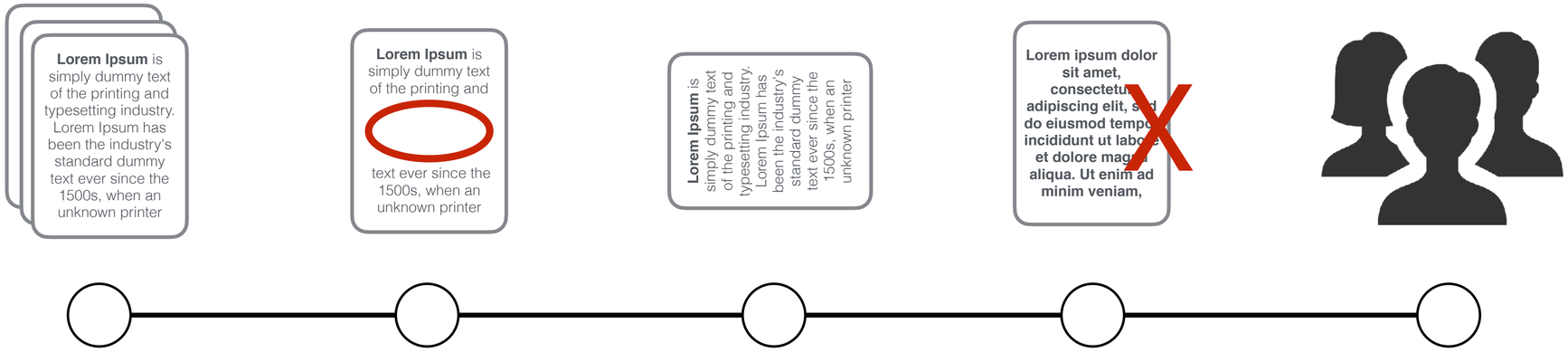
1/4

Track transactions correct!

Tran$action Tracking

Where E-commerce issues occur...

BAD DATA IS...



Duplicate data	Missing data	Inaccurate data	Incorrect data	Personal data
Duplicate transactions	**Product lists, Promotions**	**Inconsistent dataLayer**	**Timing issues**	**Name, address etc. (in the wrong place!)**

The most common problem - **Duplicate Transactions**

Overwrites the original referrer

If orders are not unique your attribution will be a mess...!

Primary Dimension: Transaction ID Date

Secondary dimension

73740657

Transaction ID	Revenue	Tax	Shipping
	SEK 3,580.00 % of Total: 0.00% (SEK 143,553,082.00)	SEK 716.00 % of Total: 0.00% (SEK 28,087,356.00)	SEK 0.00 % of Total: 0.00% (SEK 1,141,432.75)
1. 73740657	SEK 3,580.00 (100.00%)	SEK 716.00 (100.00%)	SEK 0.00 (0.00%)

If existing orders do need updating, **append** to the transID

73740657-M1

73740657-M2

Strange values...

???

Transaction ID	Revenue	Tax
	£1,024,352.00 % of Total: 0.71% (£143,553,082.00)	-£504,054.00 % of Total: -1.79% (£28,087,356.00)
500. 73740036	£98.00 (0.01%)	-£78.00 (0.02%)
501. 73739525	£118.00 (0.01%)	-£76.00 (0.02%)
502. 73764528	£99.00 (0.01%)	-£76.00 (0.02%)
503. 73703352	£5,890.00 (0.57%)	-£73.00 (0.01%)
504. 73788966	£1,890.00 (0.18%)	-£72.00 (0.01%)
505. 73742391	£89.00 (0.01%)	-£71.00 (0.01%)
506. 73700533	£399.00 (0.04%)	-£70.00 (0.01%)
507. 73720309	£648.00 (0.06%)	-£70.00 (0.01%)
508. 73788793	£149.00 (0.01%)	-£70.00 (0.01%)
509. 73780313	£669.00 (0.07%)	-£66.00 (0.01%)

Personal v B2B sales!

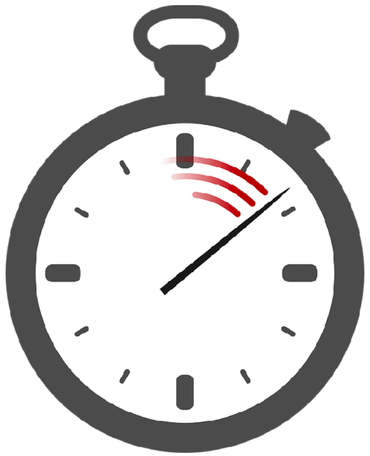
Did not match backend process.

Lessons learned:

KIS; **Don't make me think**

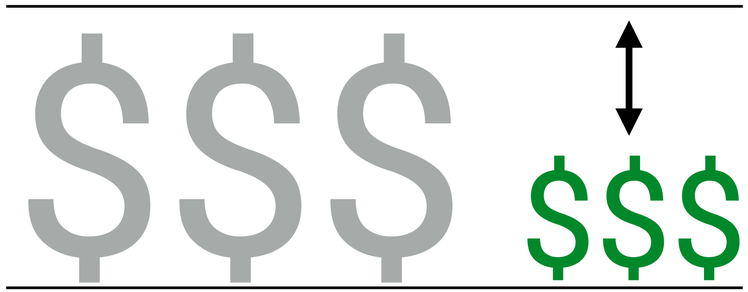
Data discrepancies

TIMING



- Google Analytics captures in **real time**.
- Backend processing is usually **batched**.

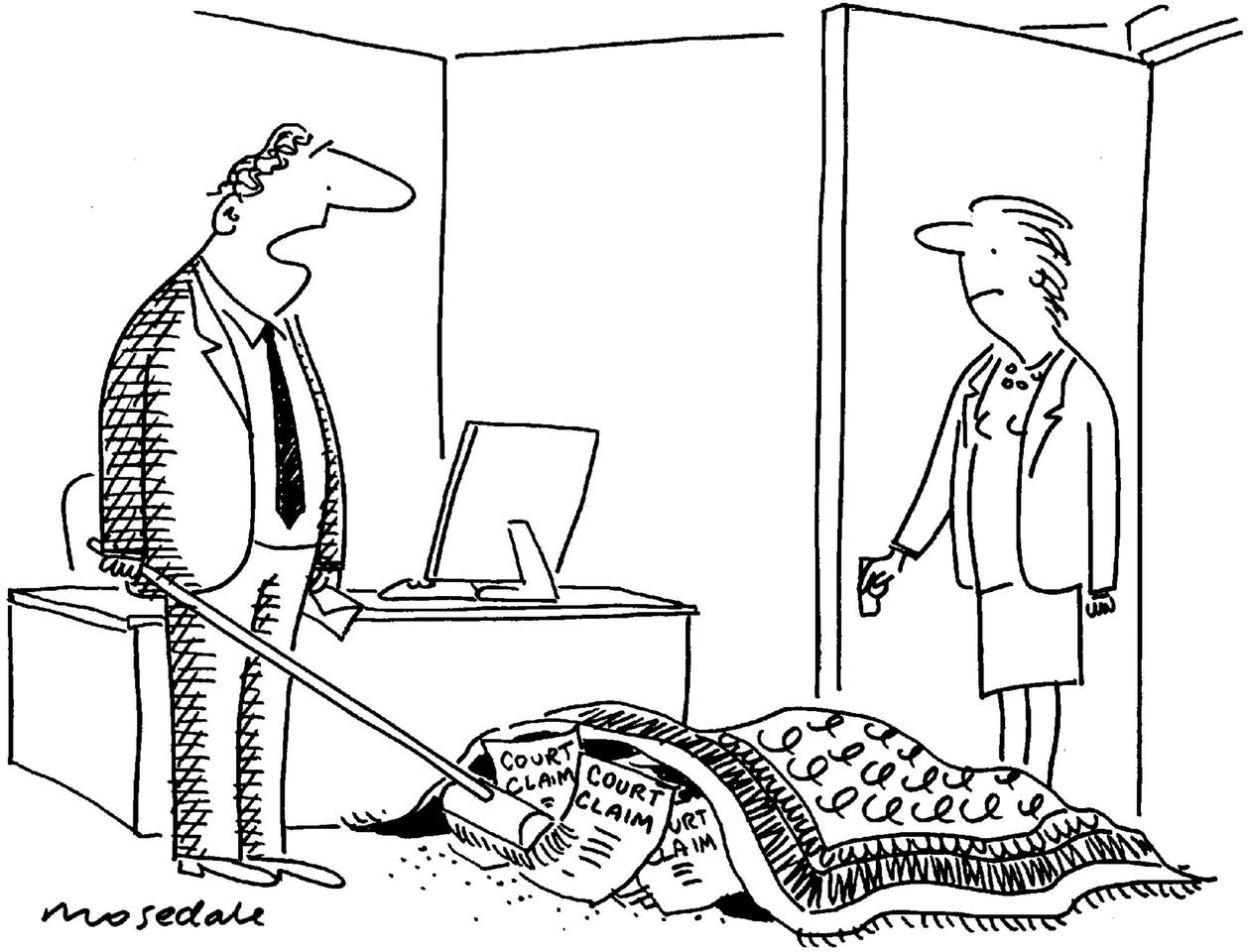
REFUNDS



- **Avoid refunds** in web analytics.
- Note them for backend reconciliation (or use a separate GA property).

Why are we in this position...?

Hard to find issues are **easy to not know about** (or even ignore)...



After all, these are needles in impossible haystacks...

This mess can be avoided...

TIPS to get on top of your data...

- **Stop** auditing manually
- **Automate** the heavy lifting
- **Fix** the priorities
- **Monitor** regularly

- Or you will lose good staff!

- Retain/recruit good staff

- Get your Quality Index score 50+

- Keep your QI > 80

The **1% rule** to follow my tips...

If you have $100 to "*make smart decisions using data*",

invest **$1** to monitor and verify its quality.