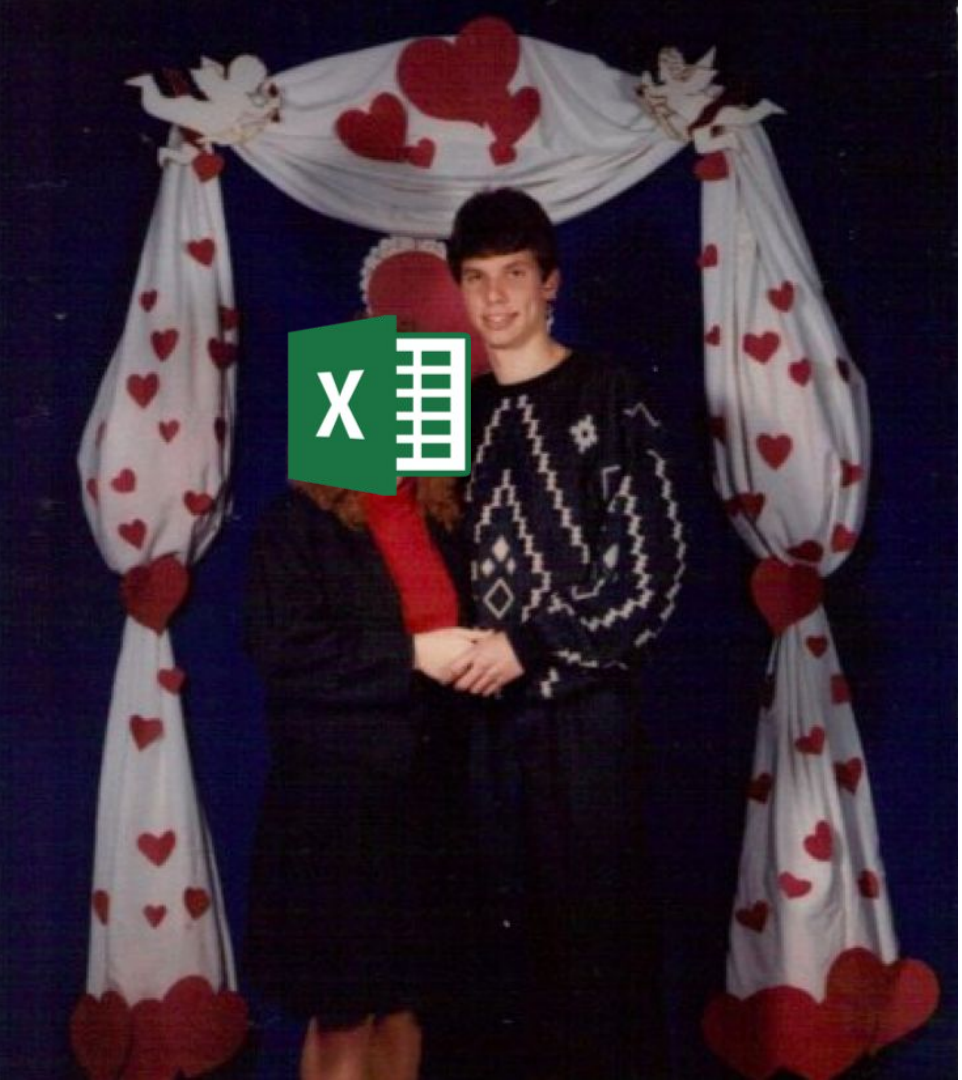
Moving Beyond Excel and Becoming More Data Science-y

Evolving How We Think About and Use Data

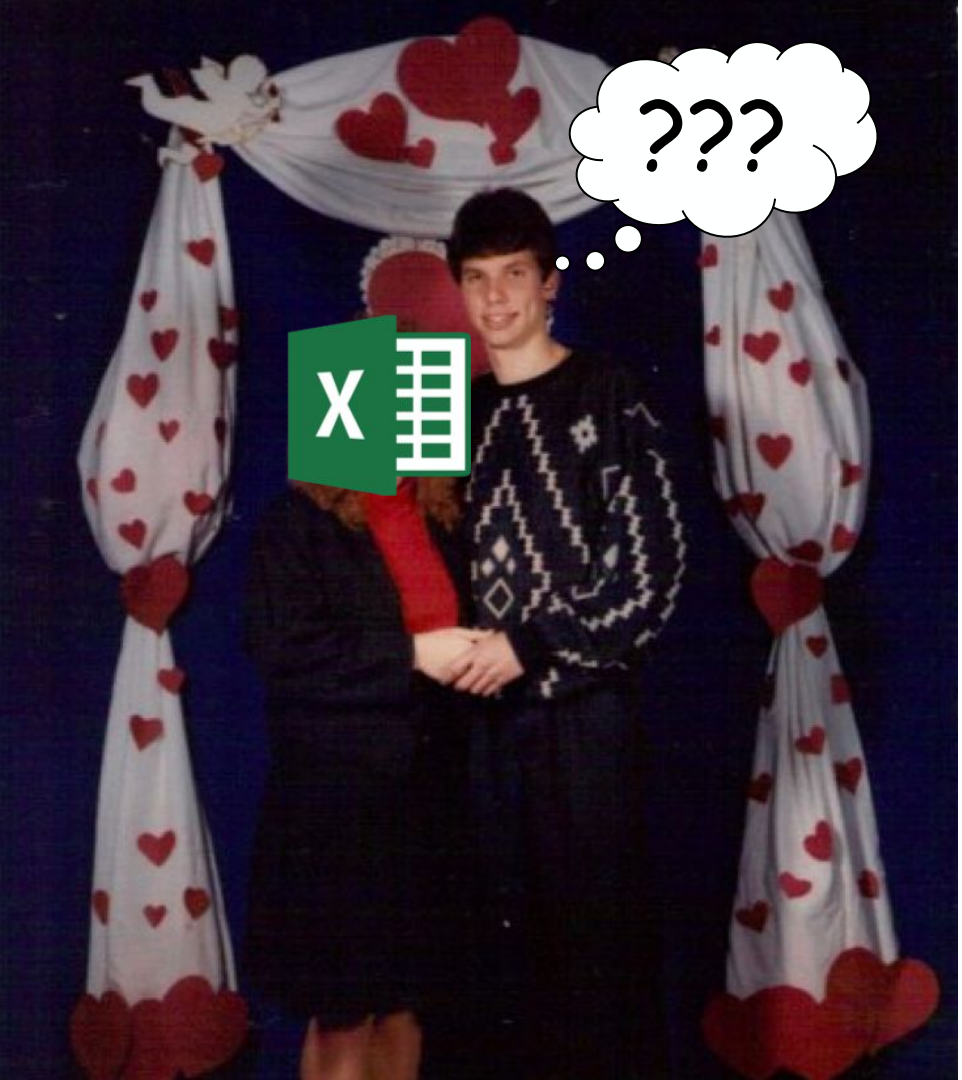

Once upon a time...


...there was an analyst.


And that analyst was me.



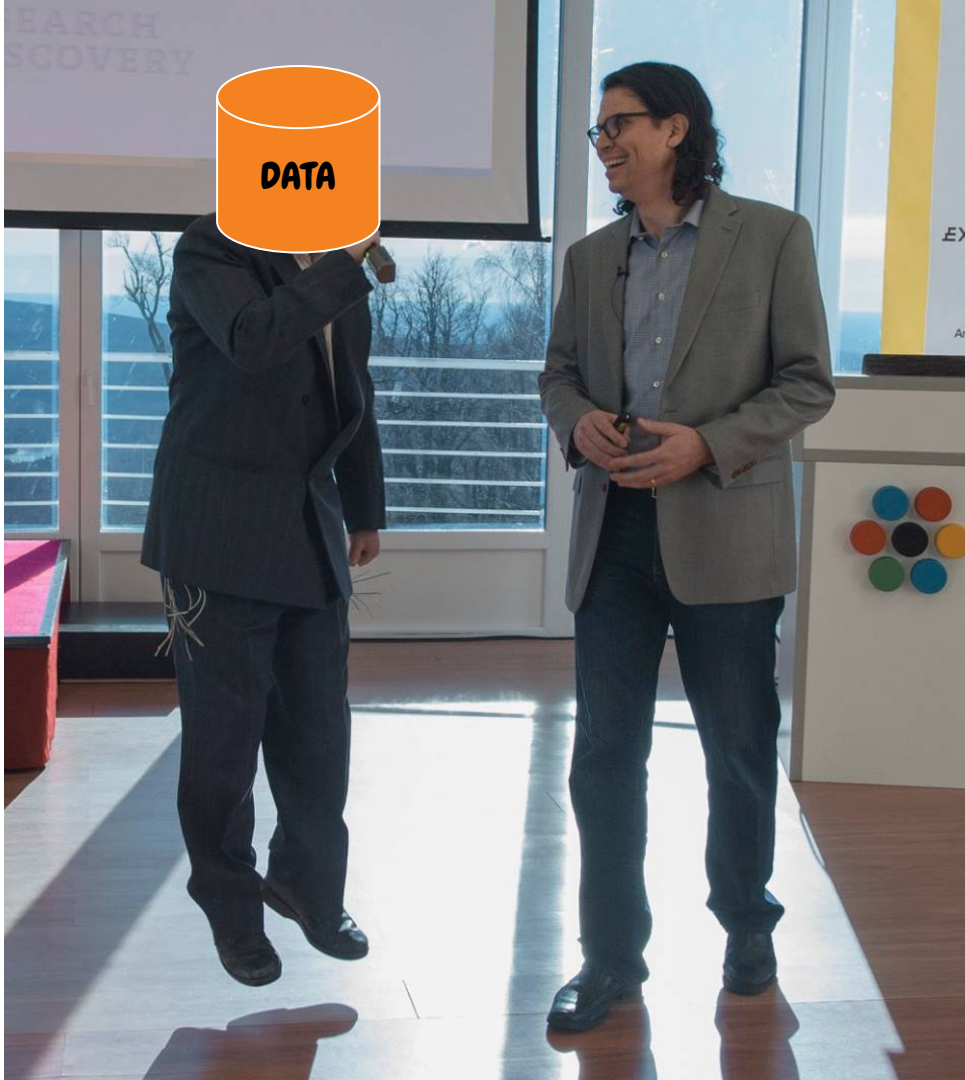
And I had a long and abiding relationship with Microsoft Excel.



But I started to question the merits of that relationship.

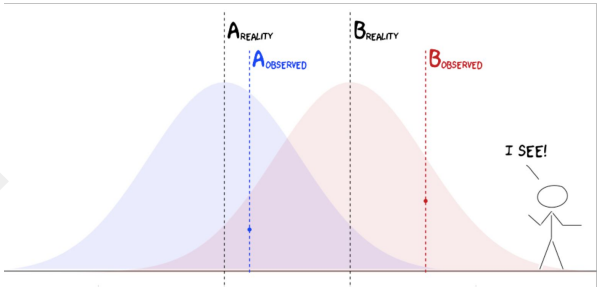
I decided I would become a **data scientist!**

I would do **magical things** with the data!



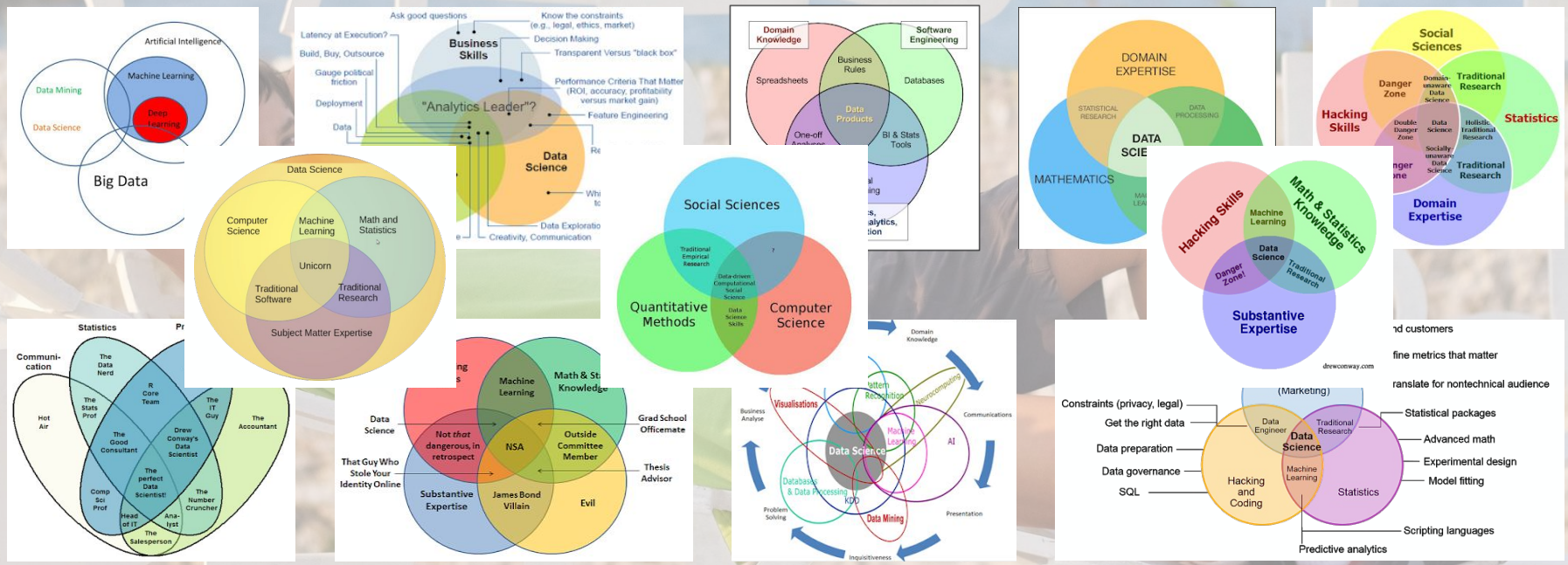
I quickly realized I would never be a **true** data scientist.

But I was intrigued...



What **is** data science?

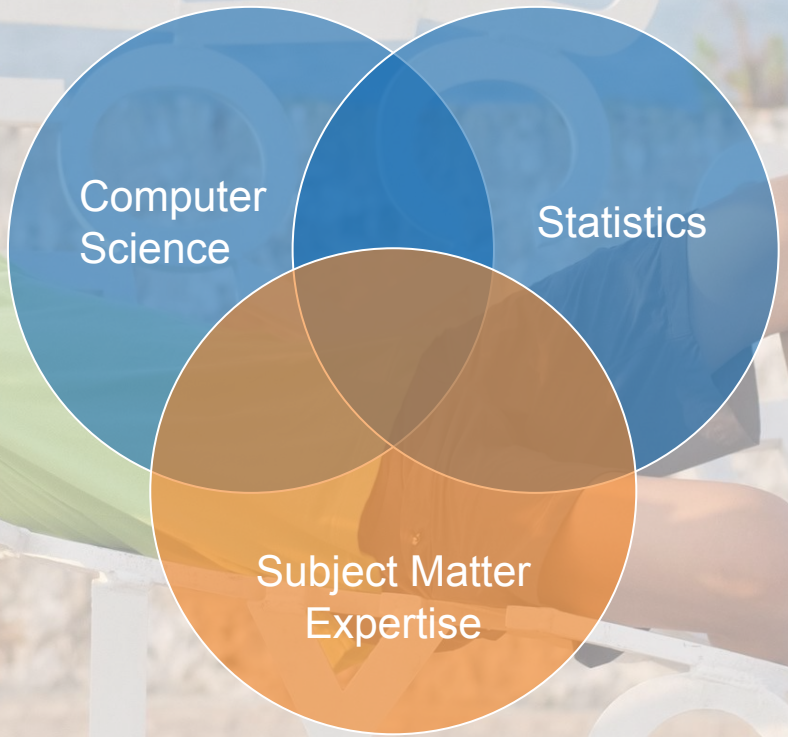
What **is** data science?



https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

An 85% Confidence Venn Diagram

Let’s Start with Subject Matter Expertise



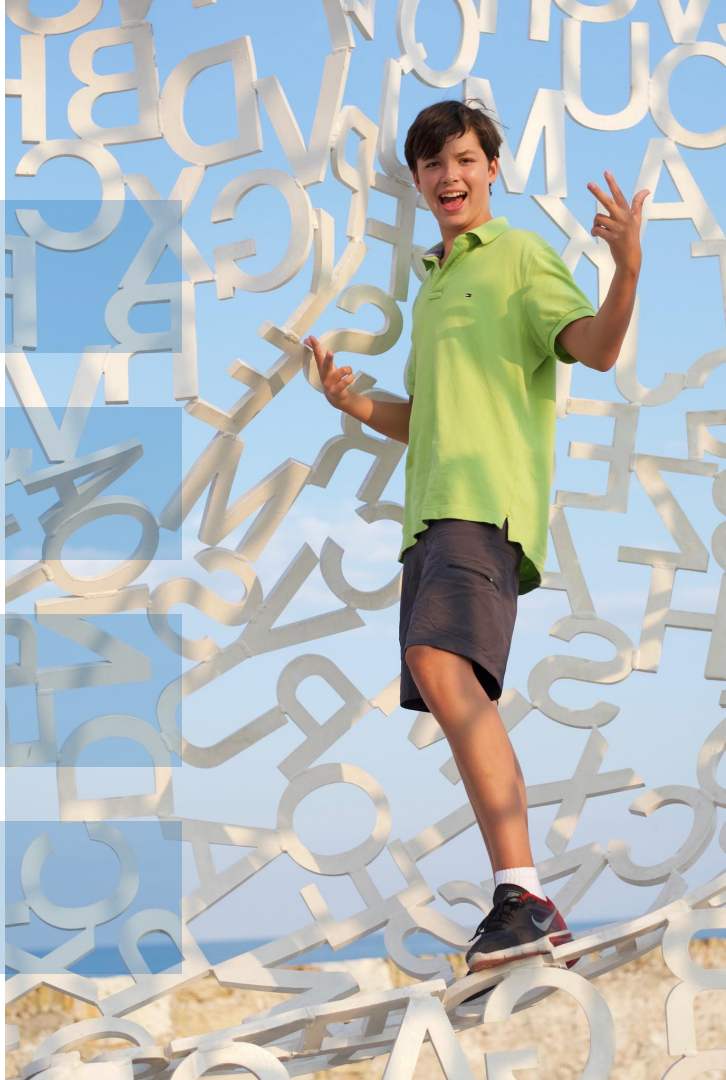
We've got this! We already...

...have deep knowledge of **how the data is collected**.

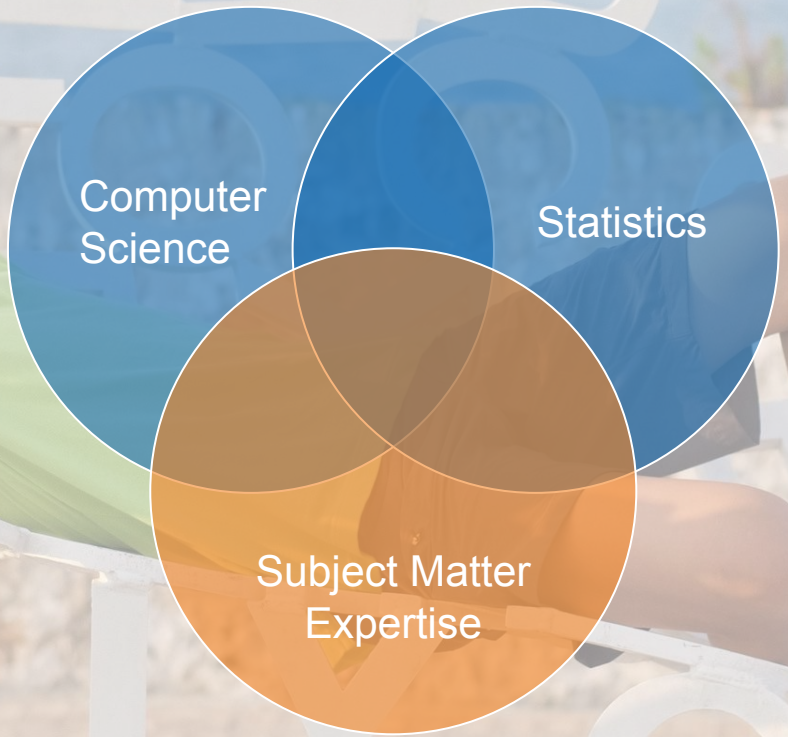
...have a deep understanding of **marketing**.

...are able to effectively articulate the **business problems** faced by our stakeholders.

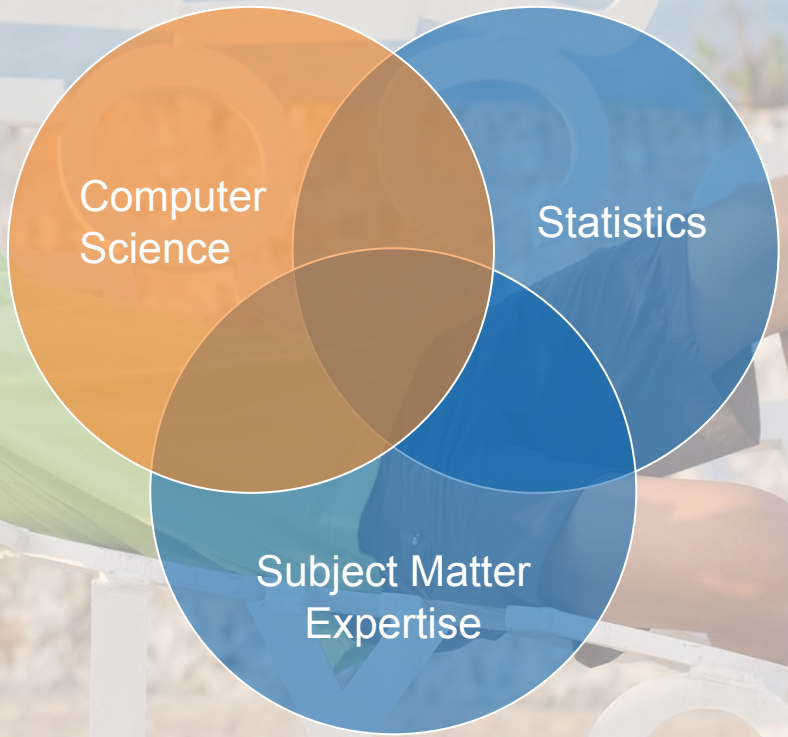
...are able to effectively communicate the **results of analyses** to stakeholders.



That’s It for Subject Matter Expertise



Let’s Talk...Computer Science!



Text-Based Programming

```
[ ] ###
    # Generate global vars from Predictor Configurations
    ##
    LOCAL_PREDICTOR_CONFIG = COMPANY + '-' + PROPERTY + '-' + PREDICTOR + '-
    GCS_BUCKET_LOCATION = 'astrologer-2/' + COMPANY + '/' + PROPERTY + '/' +
    GCS_PREDICTOR_CONFIG = GCS_BUCKET_LOCATION + 'predictorconfig.pckl'

    ###
    # Save configuration setti                    ile
    ##
    def saveConfig(PREDICTOR_C
        f = open(LOCAL_PRE
        pickle.dump(PRED
        f.close()

        # Copy updated
        !gsutil cp {LOCA                    GCS_PREDICTOR_CONFIG}

    ###
    # Check for .txt file, and                    does not exist
    ##
    def getConfig():
        # Copy file from GCS to local disk if it exists
        !gsutil cp gs://{GCS_PREDICTOR_CONFIG} {LOCAL_PREDICTOR_CONFIG}

        # Create new config if non exists already
        if not(os.path.isfile(LOCAL_PREDICTOR_CONFIG)):
            # Create new config
            PREDICTOR_CONFIG = [{
                'COMPANY': COMPANY,
                'PROPERTY': PROPERTY,
                'PREDICTOR': PREDICTOR,
                'SETTINGS': {
                    'TYPE': TYPE,
                                "CUSTOM_DATA_SOURCE_ID": "",
                                "WEB_PROPERTY_ID": "",
                                "ACCOUNT_ID": ""
                            }
                        }]
                    }
```

```
1  # Load the necessary libraries.
2  if (!require("pacman")) install.packages("pacman")
3  pacman::p_load(RSiteCatalyst, tidyverse)
4
5  # Load the username, shared secret, and report suite ID
6  username <- Sys.getenv("ADOBE_API_USERNAME")
7  secret <- Sys.getenv("ADOBE_API_SECRET")
8
9  # Authorize Adobe Analytics.
10 SCAuth(username, secret)
11
12 # Set the date range.
13 start_date <- Sys.Date(
14 end_date <- Sys.Date(
15
16 # Get the Report Sui
17 rsid_df <- GetReportS
18 numsuites <- nrow(rsi
19
20 # Set a counter just t
21 i <- 1
22
23 # Define a function to pull high
24 get_traffic <- function(rsid = "shell-d
25
26   # Output the RSID to the console so w
27   cat("Processing",i,"of",numsuites,":",rsid)
28
29   # Increment counter
30   i <<- i+1
31
32   # Pull the visits and pageviews data
33   traffic_df <- QueueSummary(rsid,
                                date = "",
                                ta for a
                                p_dfr(rs
                                e name b
                                <- left_
                                numeric(
46   arrange(desc(visits)) %>%
47   select(-url)
48
49 # Save the summary to a CSV
50 write.csv(traffic_summary_full, "rsid_traffic_summary.csv", row.names=FALSE)
51
```

CONNECTIVITY

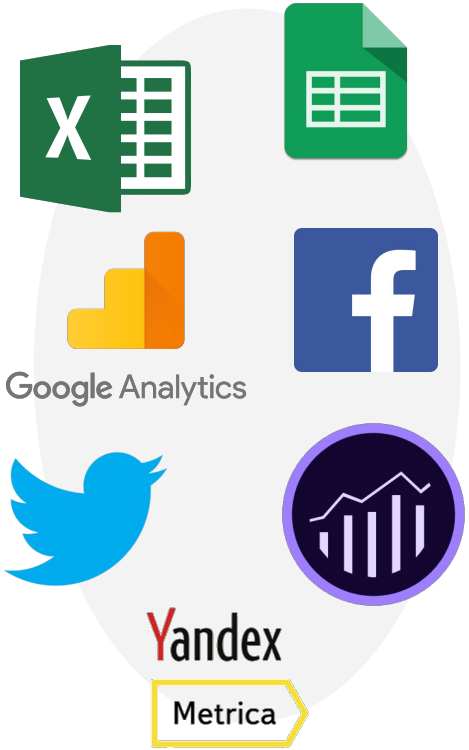
AUTOMATION

DATA VISUALIZATION

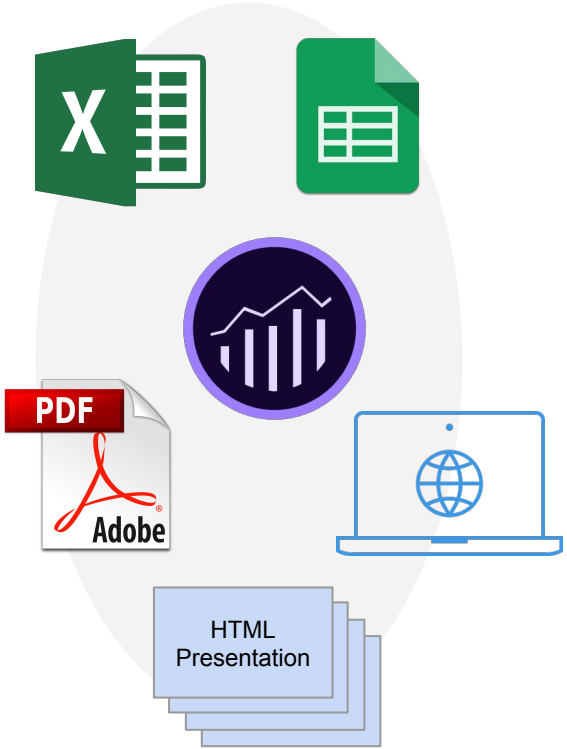
Connectivity

Inputs and Outputs

INPUTS*



OUTPUTS*



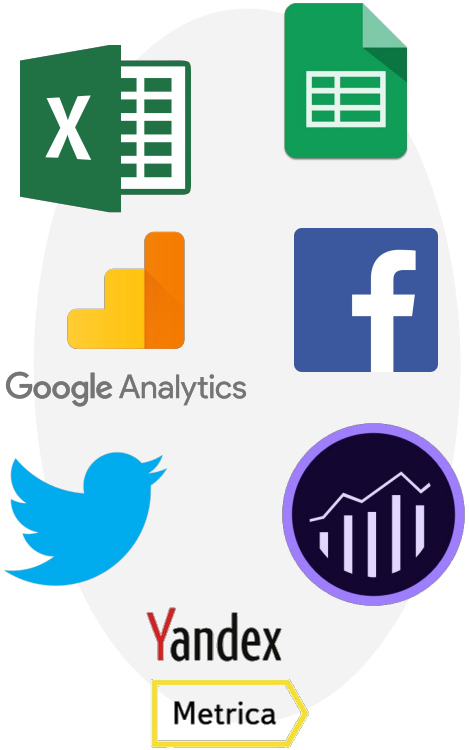
** These lists are essentially infinite, so they are limited here just to systems that Tim has pulled/pushed to using R.*

Automation

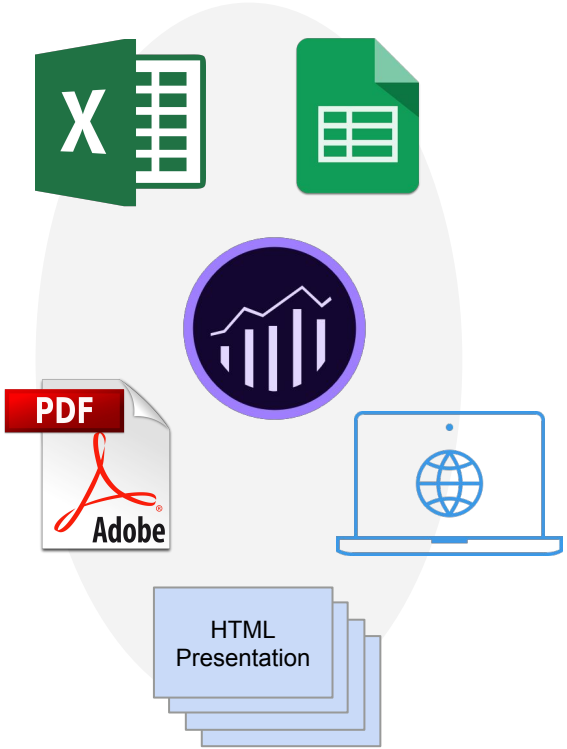
and Reusability

“How much traffic did we get to each of our sites?”

INPUTS*

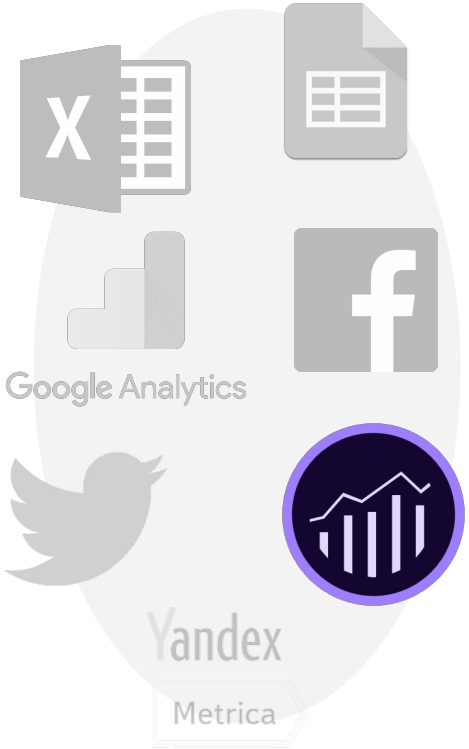


OUTPUTS*

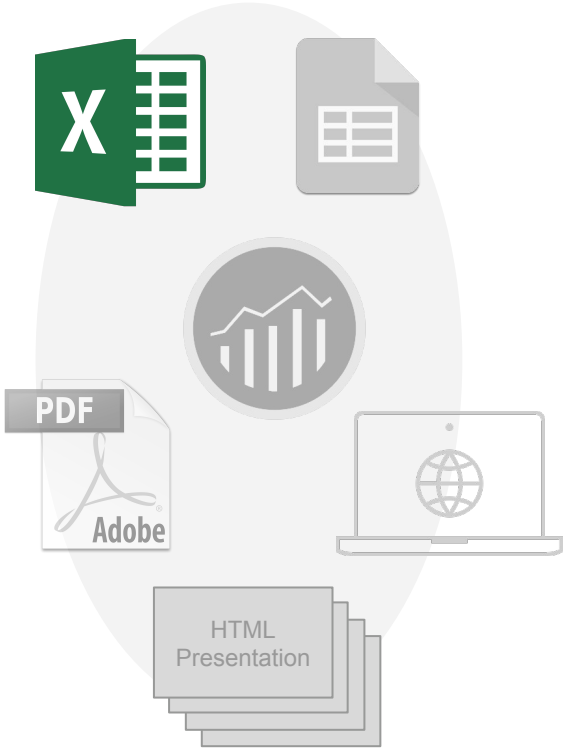


** These lists are essentially infinite, so they are limited here just to systems that Tim has pulled/pushed to using R.*

INPUTS*



OUTPUTS*

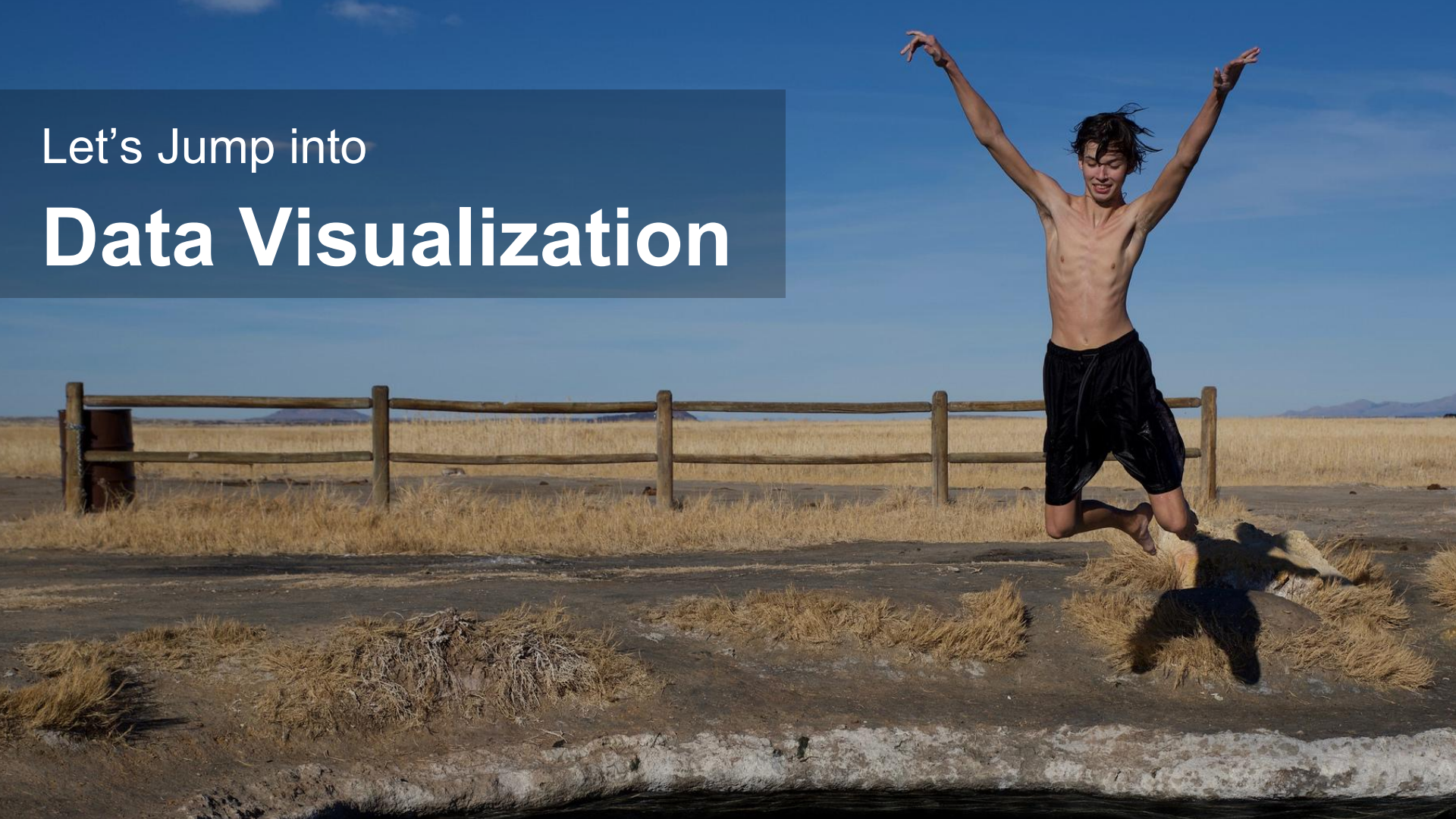


** These lists are essentially infinite, so they are limited here just to systems that Tim has pulled/pushed to using R.*

Visits and Page Views for 475 Report Suites!

```r
# Load the necessary libraries.
if (!require("pacman")) install.packages("pacman")
pacman::p_load(RSiteCatalyst, tidyverse)

# Load the username, shared secret, and report suite ID
username <- Sys.getenv("ADOBE_API_USERNAME")
secret <- Sys.getenv("ADOBE_API_SECRET")

# Authorize Adobe Analytics.
SCAuth(username, secret)

# Set the date range.
start_date <- Sys.Date() - 31        # 30 days back from yesterday
end_date <- Sys.Date() - 1           # Yesterday

# Get the Report Suites
rsid_df <- GetReportSuites()
numsuites <- nrow(rsid_df)

# Set a counter just to report out on the processing count
i <- 1

# Define a function to pull high-level data
get_traffic <- function(rsid = "shell-ac-site"){

  # Output the RSID to the console so we'll know what's being processed
  cat("Processing",i,"of",numsuites,":",rsid)

  # Increment counter
  i <<- i+1

  # Pull the visits and pageviews data
  traffic_df <- QueueSummary(rsid,
                             date = "",
                             metrics = c("visits","pageviews"),
                             date.from = start_date,
                             date.to = end_date)
}

# Pull the summary data for all report suites
traffic_summary <- map_dfr(rsid_df$rsid, get_traffic)

# Add the report suite name back on and sort by visits descending
traffic_summary_full <- left_join(traffic_summary, rsid_df, by = c("reportsuite" = "rsid"))  %>%
  mutate(visits = as.numeric(visits), pageviews = as.numeric(pageviews)) %>%
  arrange(desc(visits)) %>%
  select(-url)

# Save the summary to a CSV
write.csv(traffic_summary_full, "rsid_traffic_summary.csv", row.names=FALSE)
```

	A	B	C	D	E
1	reportsuite	visits	pageviews	site_title	virtual
2		6800684	14385490		NA
3		6298882	13477663		NA
4		3927516	6		NA
5		1044121	2582688		NA
6		808991	3803033		NA
7		679666	1011706		NA
8		563177	845258		NA
9		436928	794315		NA
10		362959	521924		NA
11		332231	532405		NA
12		329702	1463000		NA
13		299037	1604536		TRUE
14		285318	454644		NA
15		269896	795125		NA
16		247984	473322		NA
17		238520	396465		NA
18		237979	486079		NA
19		230299	609724		NA
20		224902	415389		NA
21		224055	244575		NA
22		222245	361846		NA
23		222093	1084912		NA
24		221132	587837		NA
25		160904	277098	s)	NA
26		118195	205674		NA
27		117493	236634		NA
28		99355	258287		NA
29		93757	164877		NA
30		91087	916501		TRUE
31		81557	150017		NA
32		76938	167569		NA
33		75003	157229		NA
34		71960	220363		NA
35		70105	143900		NA
36		69903	141548		NA
466		0	0		NA
467		0	0		NA
468		0	0		NA
469		0	0		NA
470		0	0		NA
471		0	0		NA
472		0	0		NA
473		0	0		NA
474		0	0		NA
475		0	0		NA
476		0	0		TRUE
477					



Let’s Jump into

Data Visualization

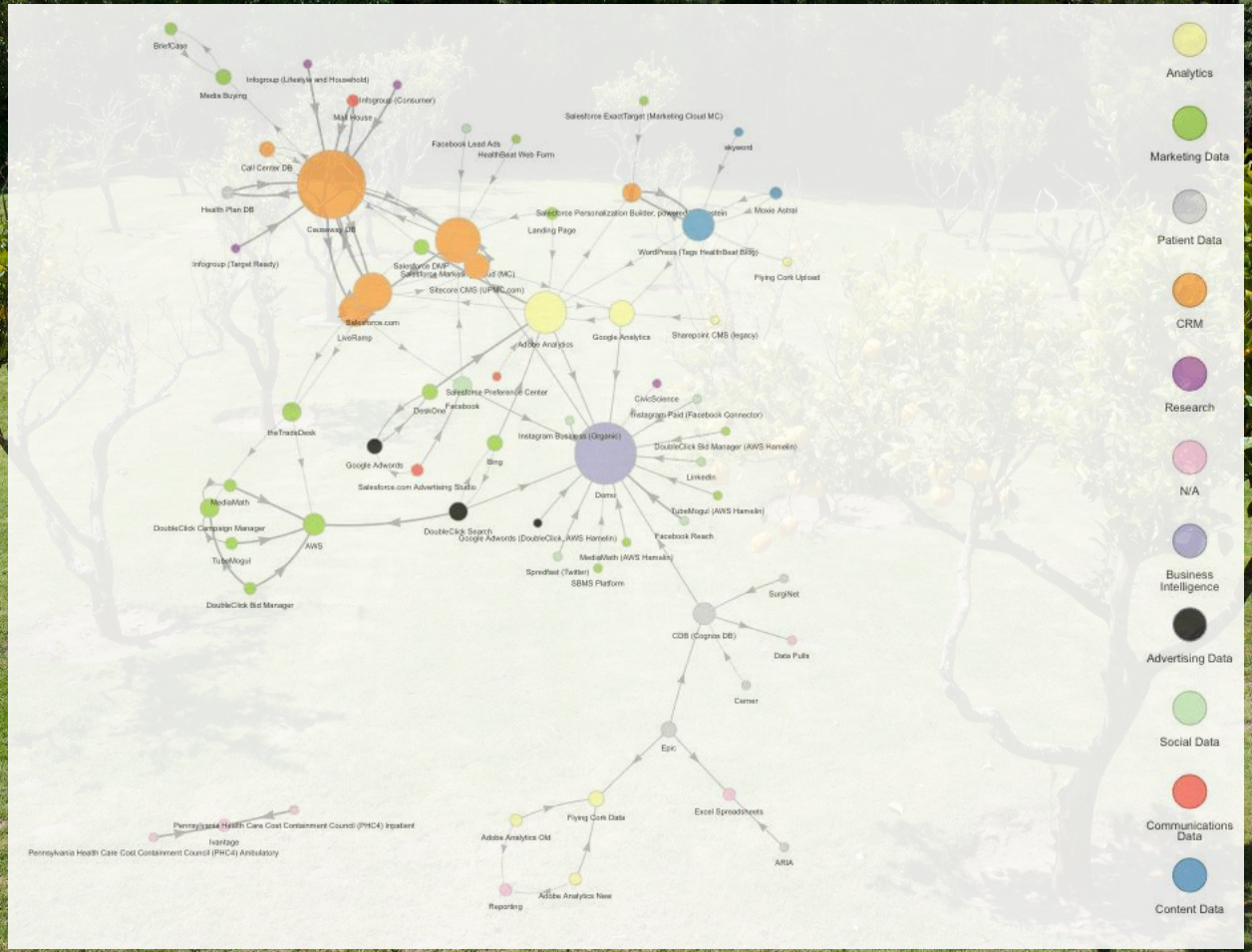
Mapping Out a Data Ecosystem

	A	B
1	**System**	
2	Adobe Analytics	
3	Adobe Audience Manager	
4	Domo	
5	Salesforce.com	
6	ExactTarget	
7	Google Analytics	
8	BigQuery	
9		
10		
11		
12		

Systems | Data Feeds

	A	B
1	**From**	**To**
2	Adobe Analytics	Adobe Audience Manager
3	Adobe Analytics	Domo
4	Adobe Analytics	Salesforce.com
5	Adobe Analytics	ExactTarget
6	Adobe Audience Manager	Adobe Analytics
7	Adobe Audience Manager	Domo
8	Salesforce.com	Domo
9	ExactTarget	Domo
10	BigQuery	Domo
11	Google Analytics	BigQuery
12	Google Analytics	Salesforce.com

Systems | Data Feeds

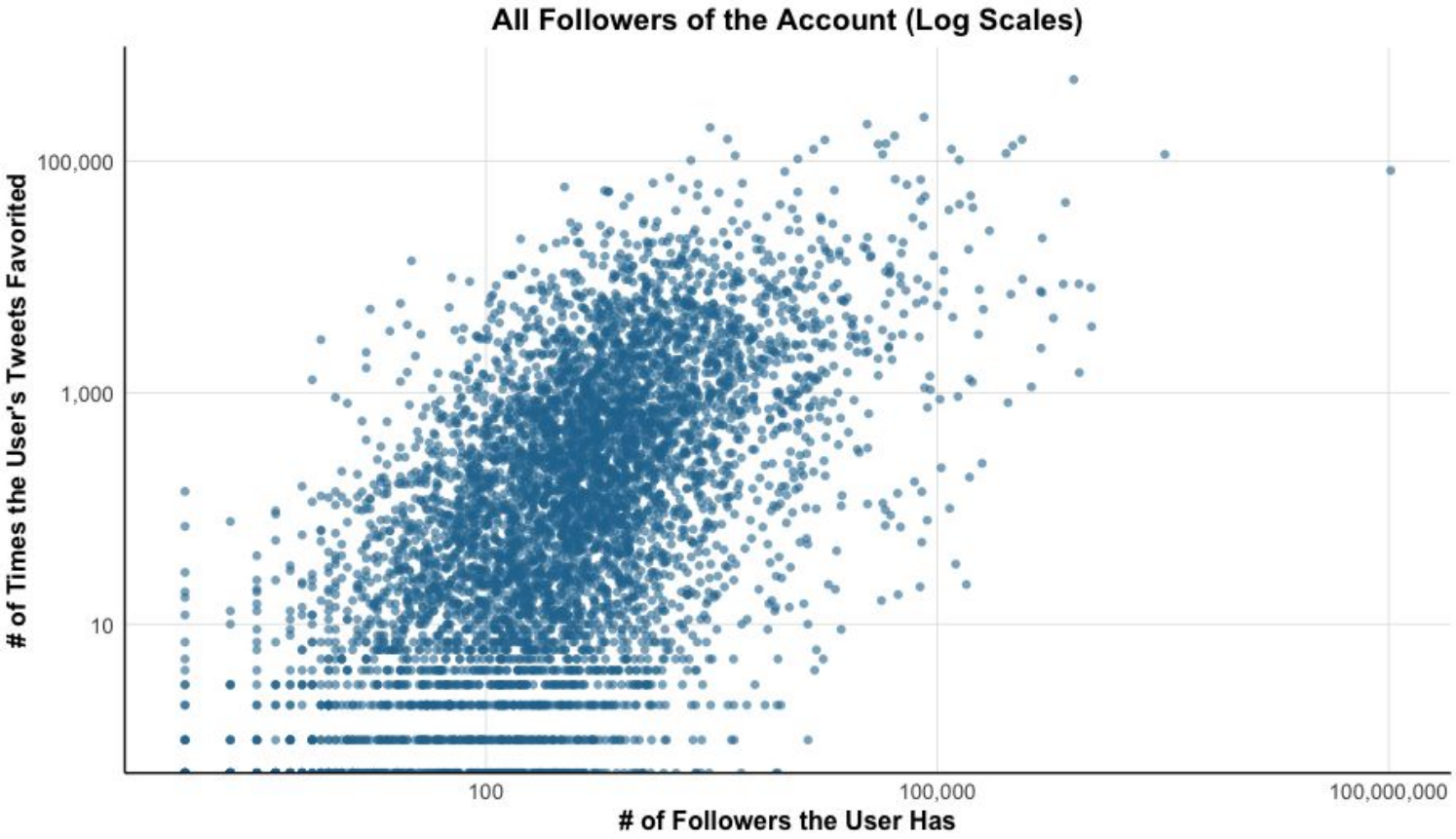

BriefCase
Infogroup (Lifestyle and Household)
Media Buying
Infogroup (Consumer)
Mail House
Salesforce ExactTarget (Marketing Cloud MC)
Facebook Lead Ads
HealthBeat Web Form
skyword
Call Center DB
Health Plan DB
Causeway DB
Salesforce Personalization Builder, powered
Landing Page
Movie Astral
Infogroup (Target Ready)
WordPress (Tags HealthBeat Blog)
Flying Cork Upload
Salesforce DMP
Salesforce Marketing Cloud (MC)
Sitecore CMS (UPMC.com)
Salesforce.com
LiveRamp
Adobe Analytics
Google Analytics
Sharepoint CMS (legacy)
Salesforce Preference Center
Facebook
DeskOne
CivicScience
Instagram Paid (Facebook Connector)
theTradeDesk
Instagram Business (Organic)
DoubleClick Bid Manager (AWS Hamelin)
Google Adwords
Bing
Salesforce.com Advertising Studio
LinkedIn
Demo
TubeMogul (AWS Hamelin)
MediaMath
DoubleClick Campaign Manager
DoubleClick Search
Google Adwords (DoubleClick, AWS Hamelin)
Facebook Reach
AWS
TubeMogul
MediaMath (AWS Hamelin)
Spredfast (Twitter)
SBMS Platform
DoubleClick Bid Manager
SurgiNet
CDB (Cognos DB)
Data Pulls
Cerner
Epic
Pennsylvania Health Care Cost Containment Council (PHC4) Inpatient
Ivantage
Pennsylvania Health Care Cost Containment Council (PHC4) Ambulatory
Flying Cork Data
Excel Spreadsheets
Adobe Analytics Old
ARIA
Adobe Analytics New
Reporting
Analytics
Marketing Data
Patient Data
CRM
Research
N/A
Business Intelligence
Advertising Data
Social Data
Communications Data
Content Data

Exploring Twitter Followers

Pull a List of All Followers of @tgwilson

Pull Basic Information About Each Follower

User	Tweets	Flwrs	Faves

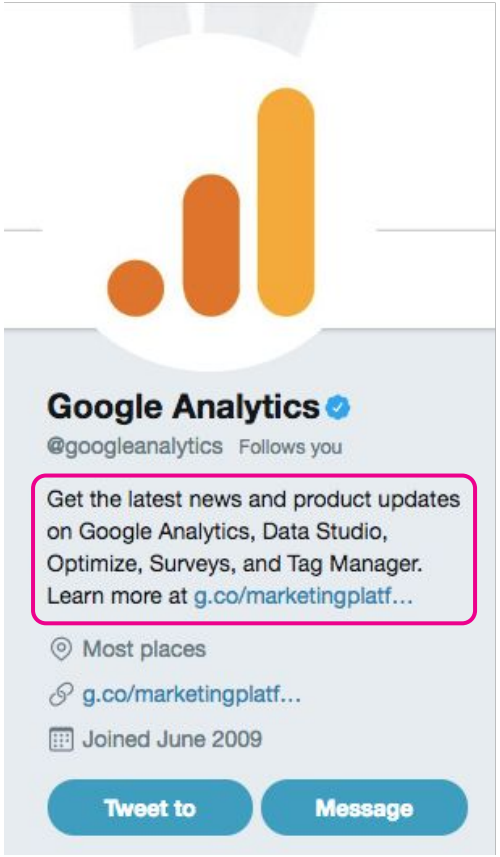


@tgwilson

SEARCH DISCOVERY


MIX.MASTA.KING
@MixMastaKing Follows you
Actor/Director For New Movie Makaveli #IMDB makavelithepacumentary.com
Los Angeles, CA
makavelithepacumentary.com
View broadcasts
Joined January 2009
Tweet to
Message


Larry Kim
@larrykim Follows you
CEO @MobileMonkey_, Founder @WordStream (acquired for $150M) Columnist @Inc, @Medium, @CNBC. Startups, AdWords, Chatbots. Popularized Unicorns in Marketing.
Boston, MA
bit.ly/LarryKim-Linke...
Joined December 2008
Tweet to
Message


Google Analytics
@googleanalytics Follows you
Get the latest news and product updates on Google Analytics, Data Studio, Optimize, Surveys, and Tag Manager. Learn more at g.co/marketingplatf...
Most places
g.co/marketingplatf...
Joined June 2009
Tweet to
Message


using industry
brand performance thoughts
traveler advocate conversion making use
addict better learning innovation beer enterprise
agency news interested
team optimization experience specialist fanatic
living like fun communications
years growth speaker strategy time entrepreneur scientist
just director will certified
account work father
cofounder founder consultant enthusiast former mother
global
great loves twitter tech food
people social data marketer writer create
building leader guy since leading
mobile music
blogger husband
stuff insights travel brands
community ppc fan dog junkie
lover creative
excel internet free
designer mom geek sports owner mba
home user
follow views help sem wife build
passion opinions customer
runner science website
software big online head search cro day startups
best one google manager good marketers
company
group family tweets now world services
everything also get new love sales adobe
lead design life web ceo media things real expert alum
mine student
based professional dad seo analyst live solutions senior avid
analysis tweet content business helping advertising
developer make product partner cloud
consulting strategist technology passionate can always football
bigdata intelligence ecommerce
health development author working businesses
columbus measure nerd coffee platform
clients management proud socialmedia player
adwords research ohio companies
digitalmarketing executive startup
photographer
marketing
analytics
digital


MIX.MASTA.KING
@MixMastaKing Follows you
Actor/Director For New Movie Makaveli
#IMDB makavelithepacumentary.com
Los Angeles, CA
makavelithepacumentary.com
View broadcasts
Joined January 2009
Tweet to
Message


Larry Kim
@larrykim Follows you
CEO @MobileMonkey_, Founder
@WordStream (acquired for $150M)
Columnist @Inc, @Medium, @CNBC.
Startups, AdWords, Chatbots.
Popularized Unicorns in Marketing.
Boston, MA
bit.ly/LarryKim-Linke...
Joined December 2008
Tweet to
Message

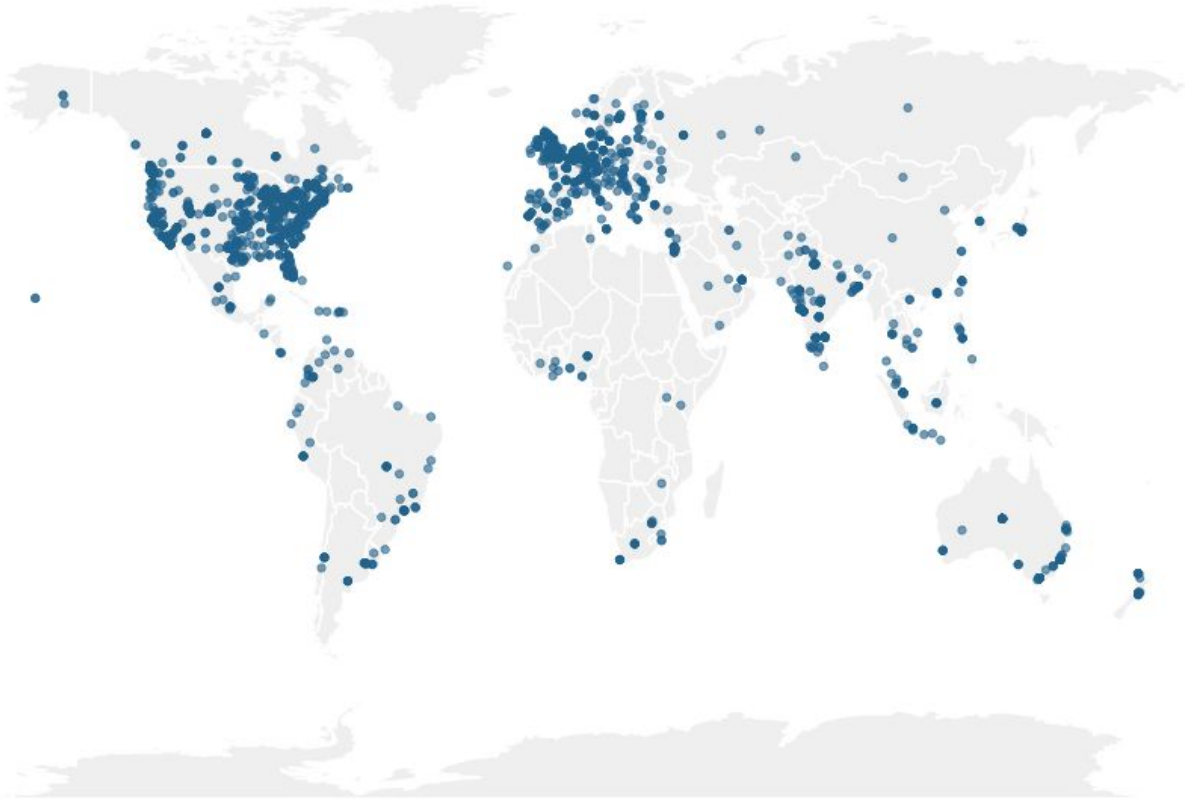

Google Analytics
@googleanalytics Follows you
Get the latest news and product updates
on Google Analytics, Data Studio,
Optimize, Surveys, and Tag Manager.
Learn more at g.co/marketingplatf...
Most places
g.co/marketingplatf...
Joined June 2009
Tweet to
Message

Pull a List of All Followers of @tgwilson

Pull Basic Information About Each Follower

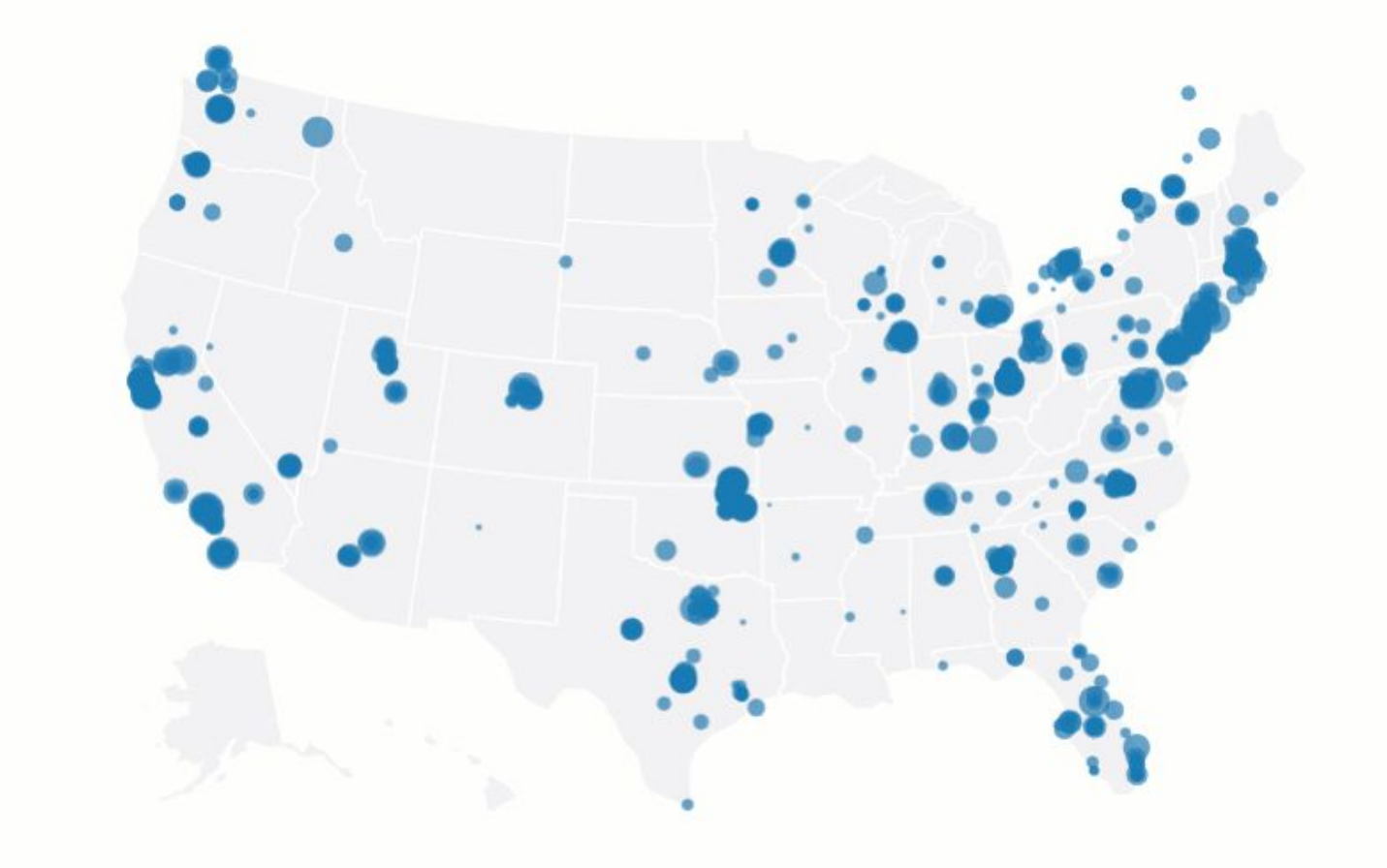
User	Tweets	Flwrs	Faves	Lat	Long

Top Followers of the Account (Based on Followers, When Location Available)



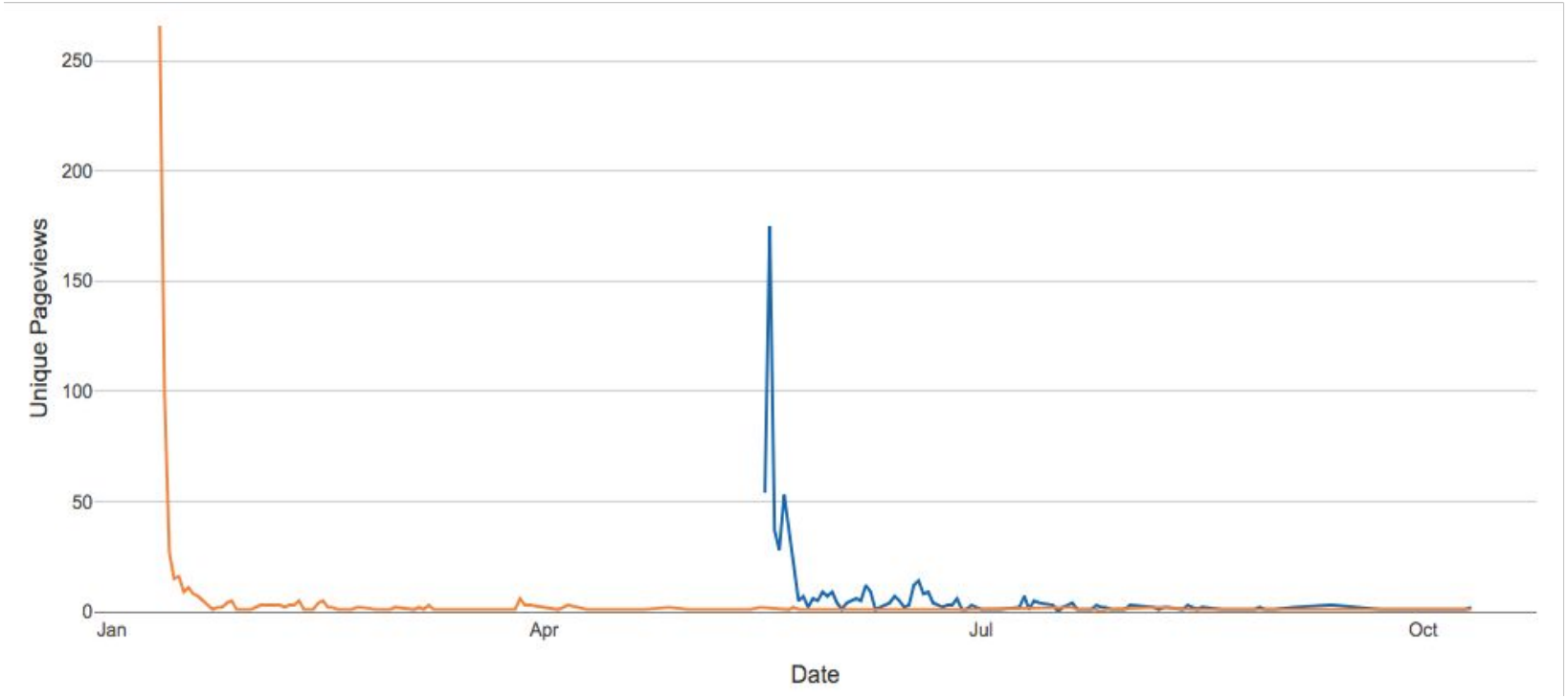
@tgwilson

SEARCH DISCOVERY

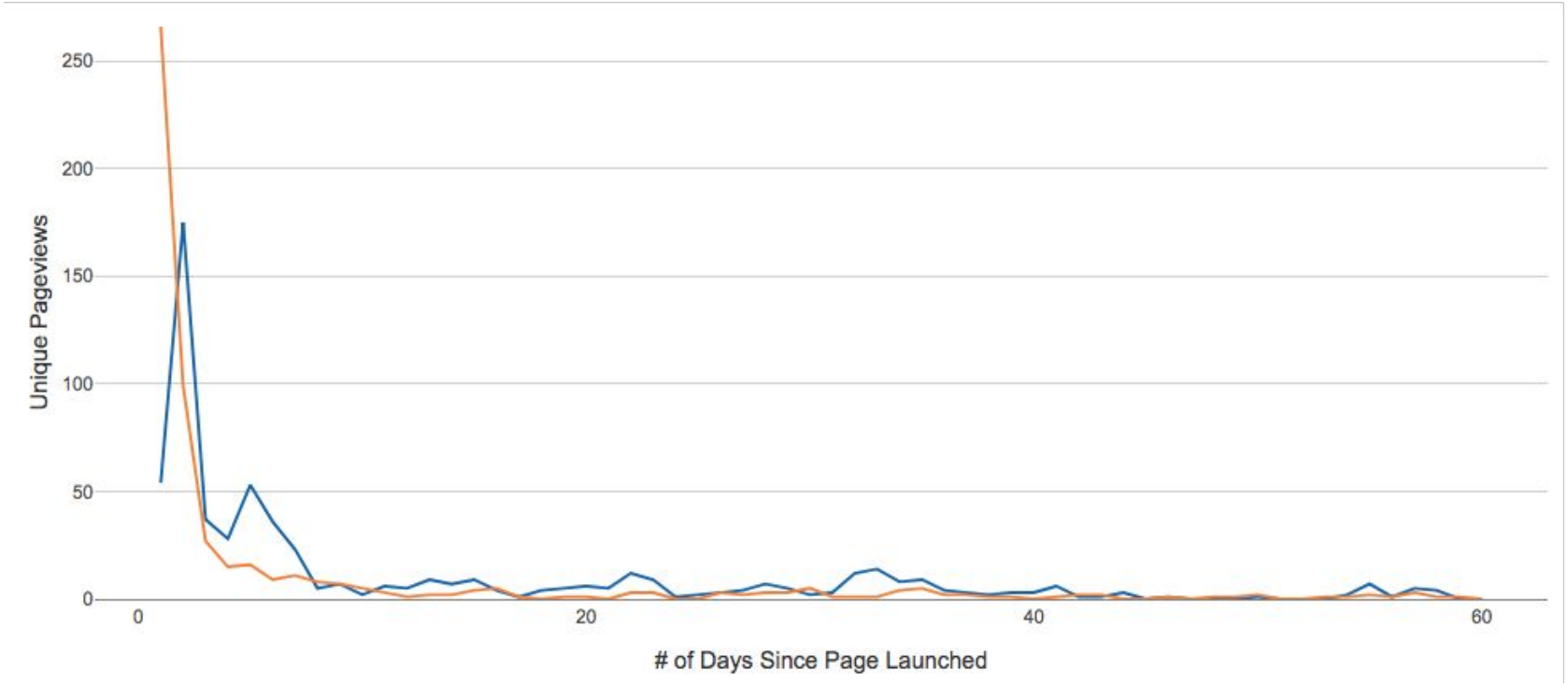


Page Traffic: Launch vs. Lifecycle

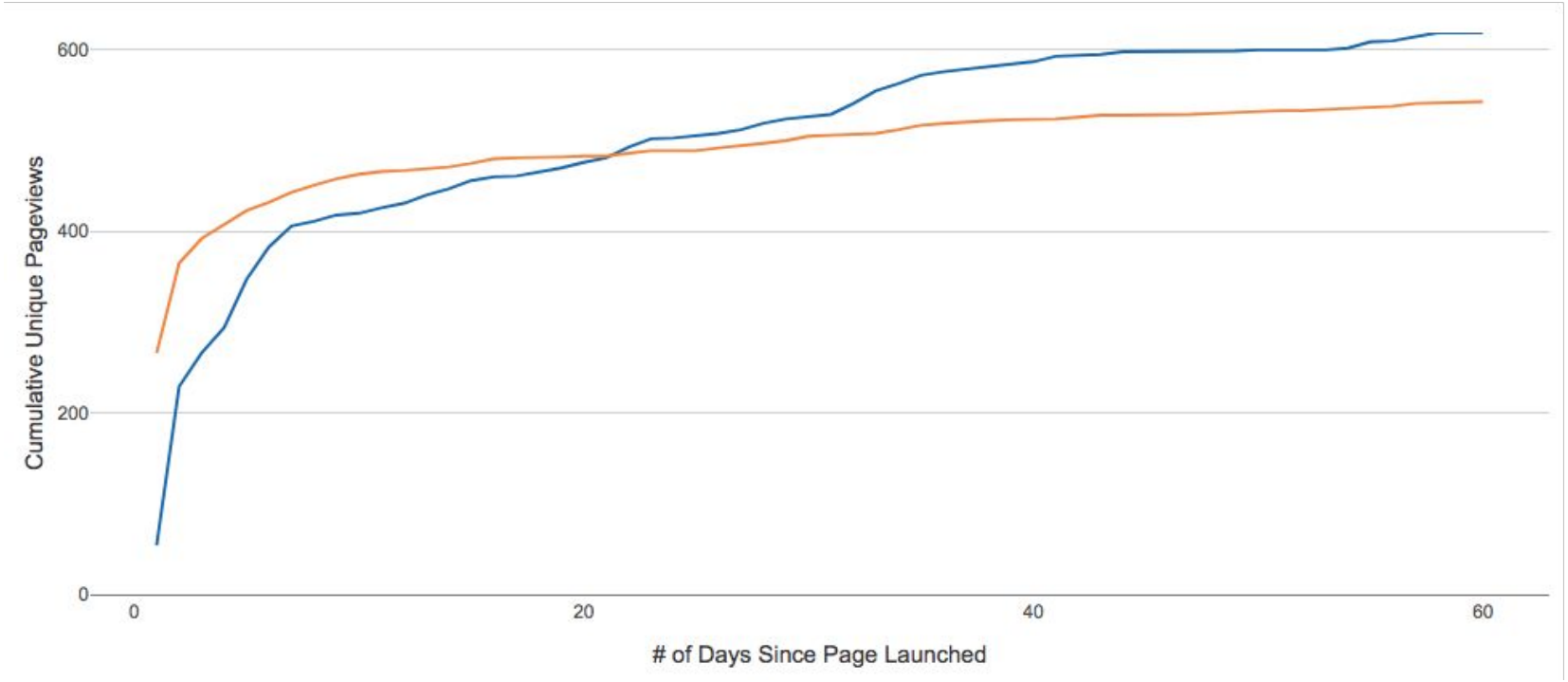
Consider Two Blog Posts



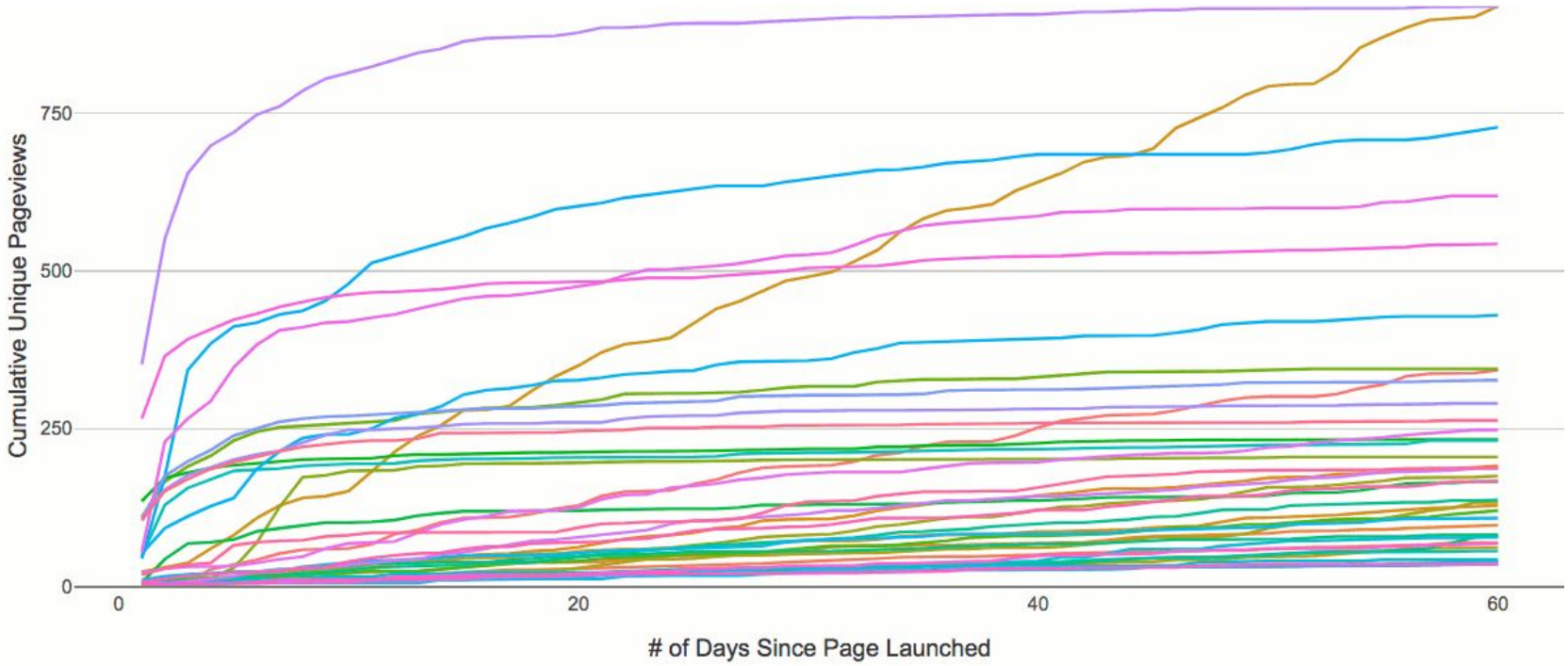
We Can “Time-Normalize” Them to Launch Date



We Can “Time-Normalize” Them to Launch Date



We Can “Time-Normalize” Them to Launch Date



So…should I learn

or

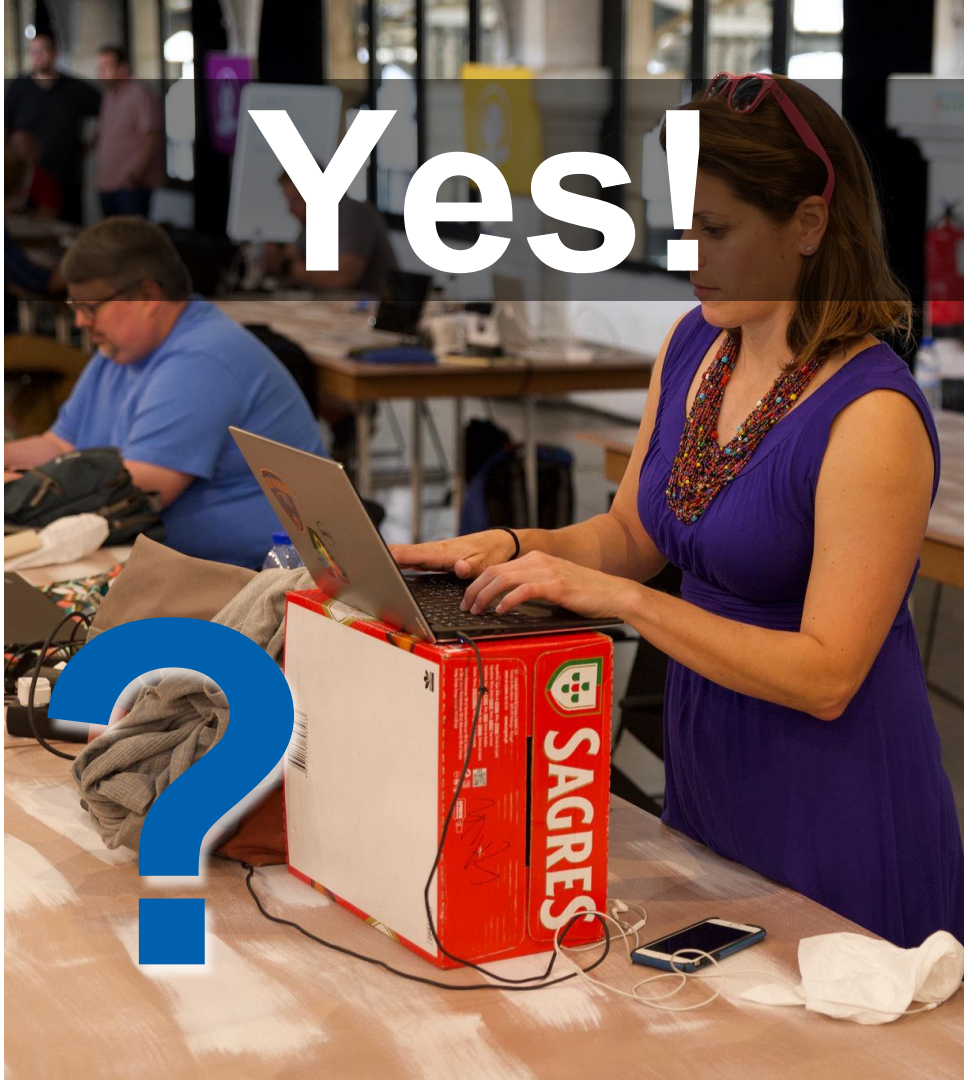
python™

So…should I learn

R

or

python™



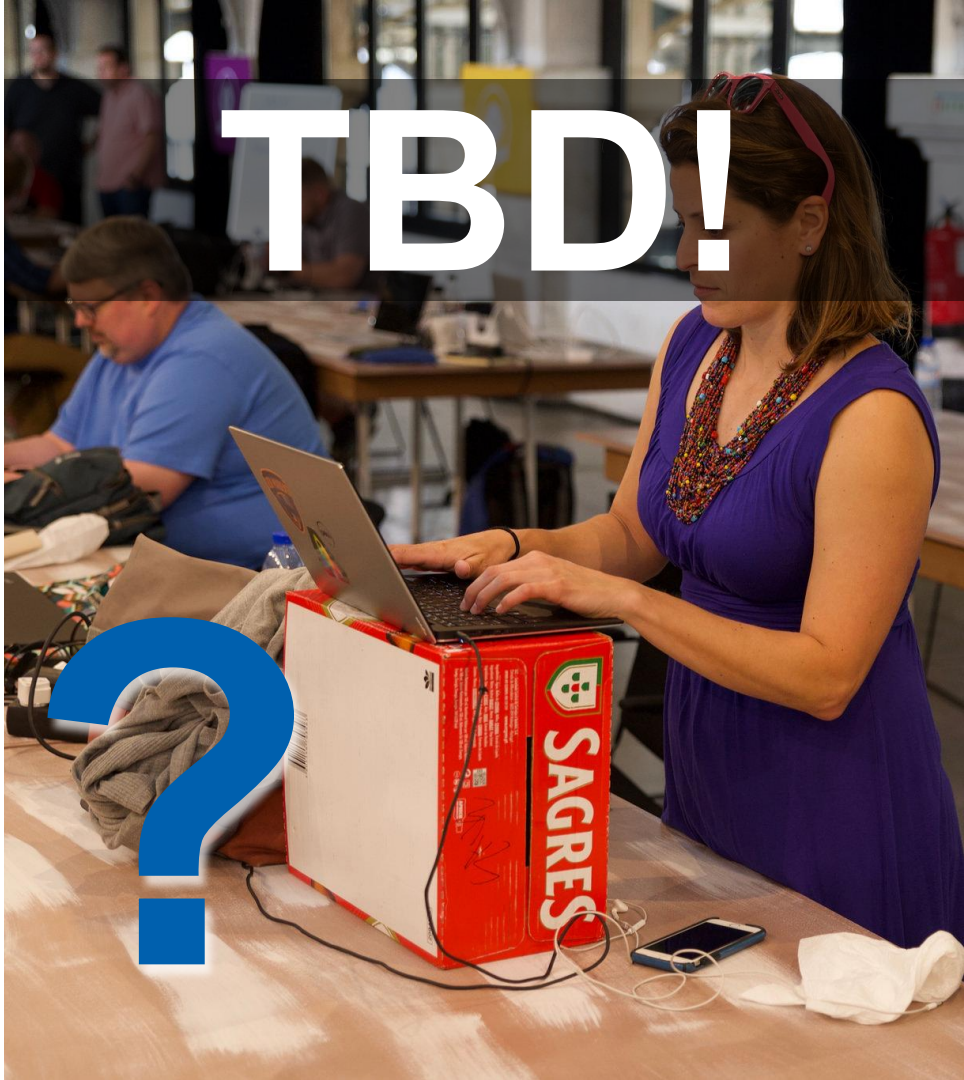
@tgwilson

So…should I learn

R

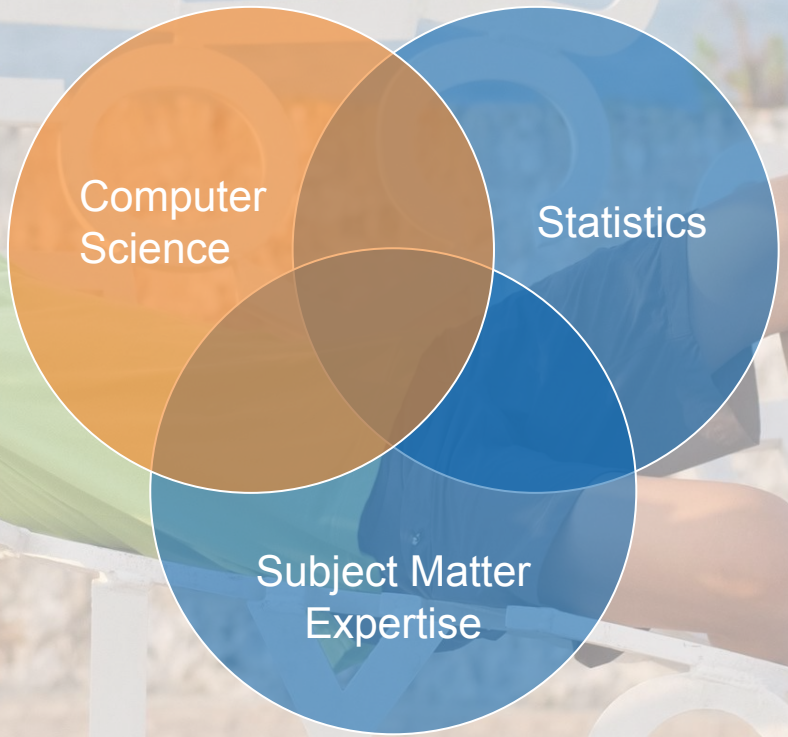
or

python™



@tgwilson

So...That Was Computer Science!



I’m 100% Confident We Should Talk Statistics

“We have a problem in America with thinking probabilistically.”

- Annie Duke

“We have a problem in [marketing] with thinking probabilistically.”

- Annie Duke

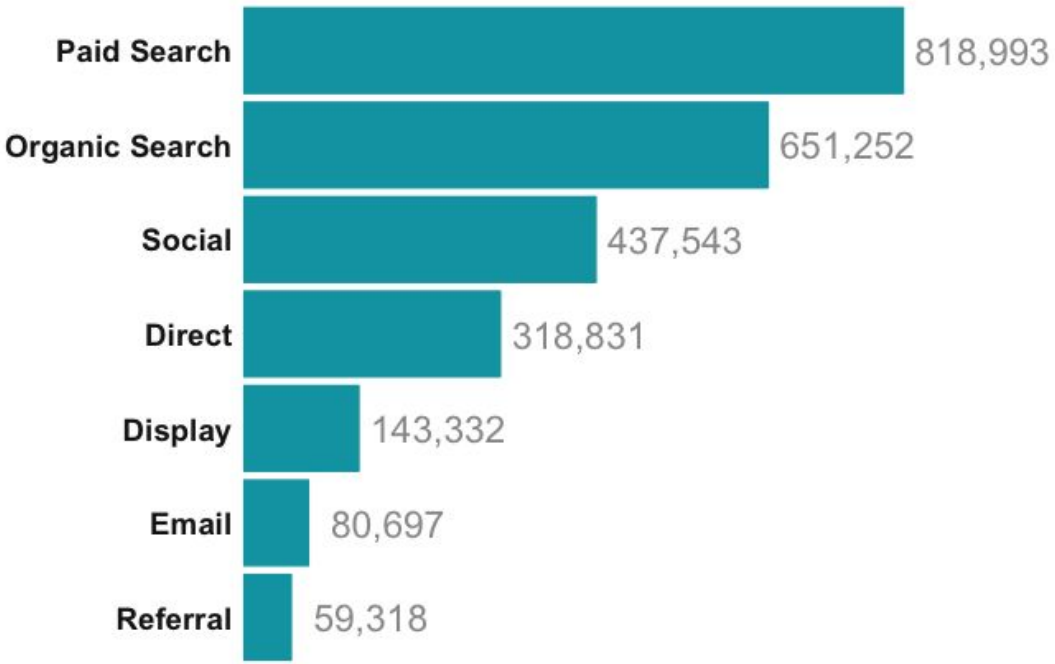
How much traffic came to the site last month?

2,509,966

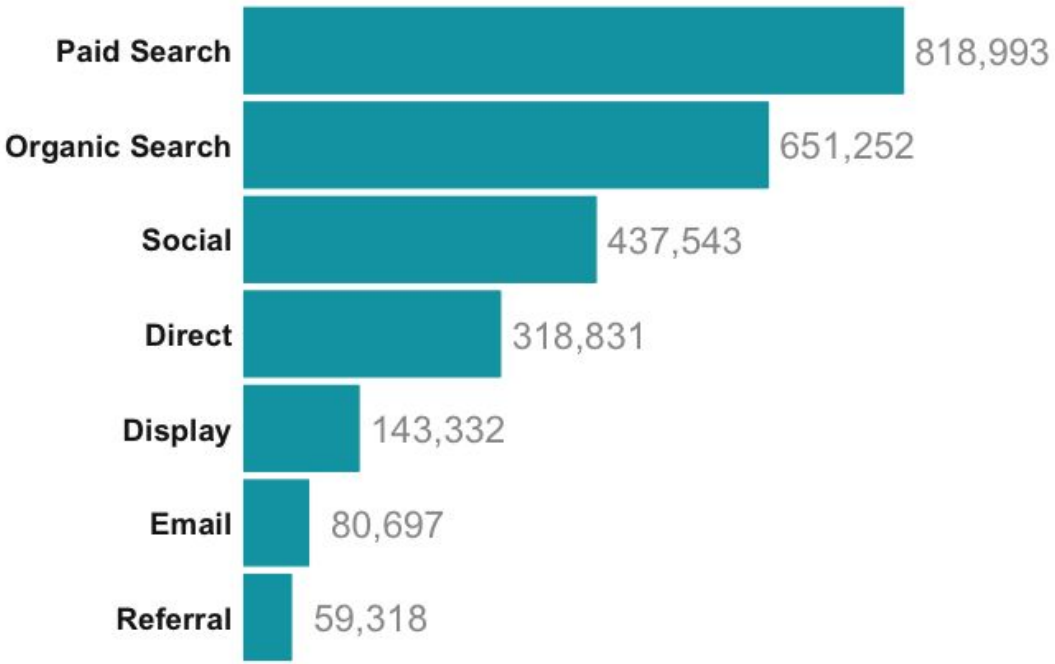
sessions

Trend it!

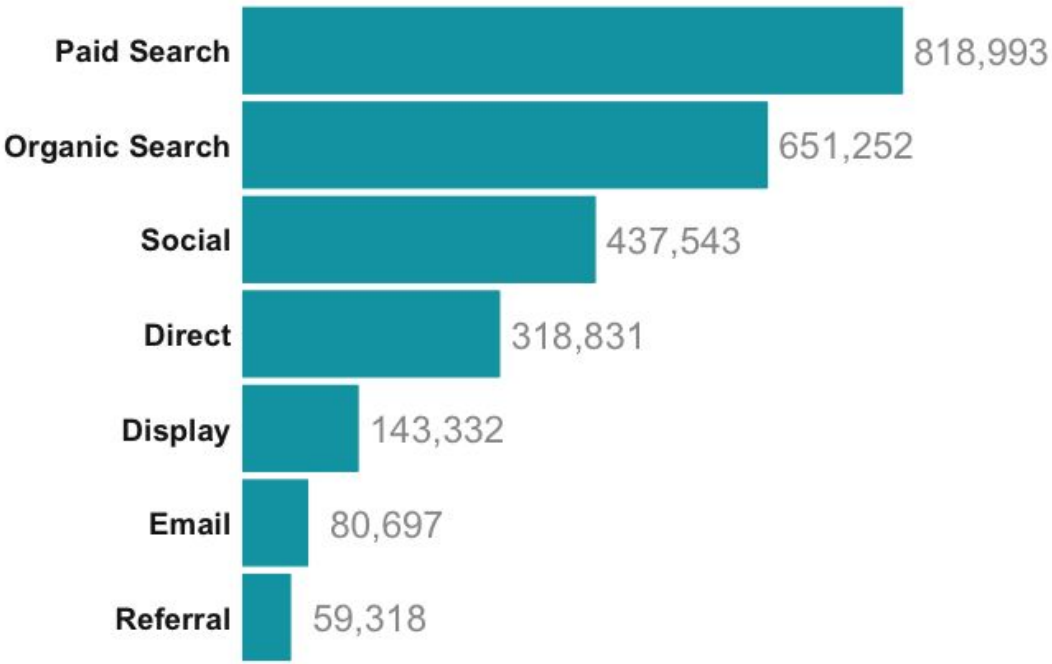
Or break it down by channel!



Let’s ponder this breakdown a bit.



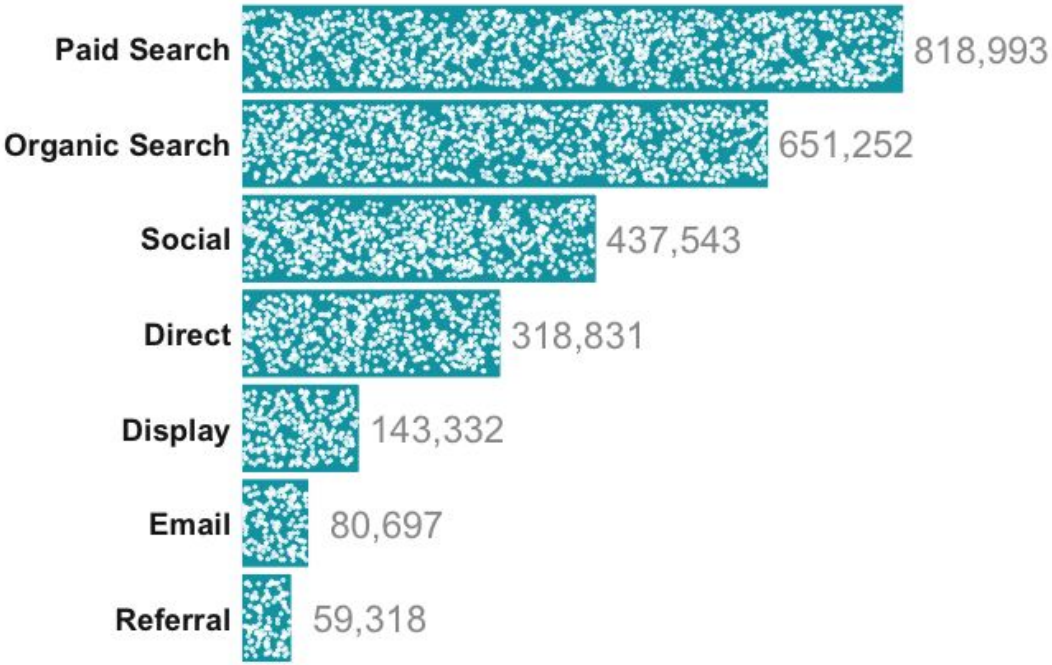
How many rows of data went into this table?



Channel	Sessions
Paid Search	818,993
Organic Search	651,252
Social	437,543
Direct	318,831
Display	143,332
Email	80,697
Referral	59,318

7!

How would a data scientist answer the question?



Visit ID	Channel
7521516054.5930117975	Organic Search
6639584132.5334606522	Organic Search
2072704531.5757008375	Organic Search
3700666076.4855492580	Social
3901278585.2803798674	Social
9493011014.8450893410	Referral
4257955221.2341693619	Organic Search
5653594021.2053548613	Paid Search
2103008782.3786630871	Organic Search
2156814867.9852847980	Direct
3969242874.8434979641	Paid Search
8975736469.8764719209	Social
3127960728.4142998058	Social
3863953834.9545306506	Organic Search
1769847183.6232068180	Organic Search

2.5 million

This shift also helps with some basic stats concepts

What is the variable?

What are the values of the variable?

Channel	Sessions
Paid Search	818,993
Organic Search	651,252
Social	437,543
Direct	318,831
Display	143,332
Email	80,697
Referral	59,318

?

Visit ID	Channel
7521516054.5930117975	Organic Search
6639584132.5334606522	Organic Search
2072704531.5757008375	Organic Search
3700666076.4855492580	Social
3901278585.2803798674	Social
9493011014.8450893410	Referral
4257955221.2341693619	Organic Search
5653594021.2053548613	Paid Search
2103008782.3786630871	Organic Search
2156814867.9852847980	Direct
3969242874.8434979641	Paid Search
8975736469.8764719209	Social
3127960728.4142998058	Social
3863953834.9545306506	Organic Search
1769847183.6232068180	Organic Search

This shift also helps with some basic stats concepts

What is the variable?

What are the values of the variable?

Channel	Sessions
Paid Search	818,993
Organic Search	651,252
Social	437,543
Direct	318,831
Display	143,332
Email	80,697
Referral	59,318

?

Visit ID	Channel
7521516054.5930117975	Organic Search
6639584132.5334606522	Organic Search
2072704531.5757008375	Organic Search
3700666076.4855492580	Social
3901278585.2803798674	Social
9493011014.8450893410	Referral
4257955221.2341693619	Organic Search
5653594021.2053548613	Paid Search
2103008782.3786630871	Organic Search
2156814867.9852847980	Direct
3969242874.8434979641	Paid Search
8975736469.8764719209	Social
3127960728.4142998058	Social
3863953834.9545306506	Organic Search
1769847183.6232068180	Organic Search

This shift also helps with some basic stats concepts

What is the variable?

What are the values of the variable?

Channel	Sessions
Paid Search	818,993
Organic Search	651,252
Social	437,543
Direct	318,831
Display	143,332
Email	80,697
Referral	59,318

?

Visit ID	Channel
7521516054.5930117975	Organic Search
6639584132.5334606522	Organic Search
2072704531.5757008375	Organic Search
3700666076.4855492580	Social
3901278585.2803798674	Social
9493011014.8450893410	Referral
4257955221.2341693619	Organic Search
5653594021.2053548613	Paid Search
2103008782.3786630871	Organic Search
2156814867.9852847980	Direct
3969242874.8434979641	Paid Search
8975736469.8764719209	Social
3127960728.4142998058	Social
3863953834.9545306506	Organic Search
1769847183.6232068180	Organic Search

This shift also helps with some basic stats concepts

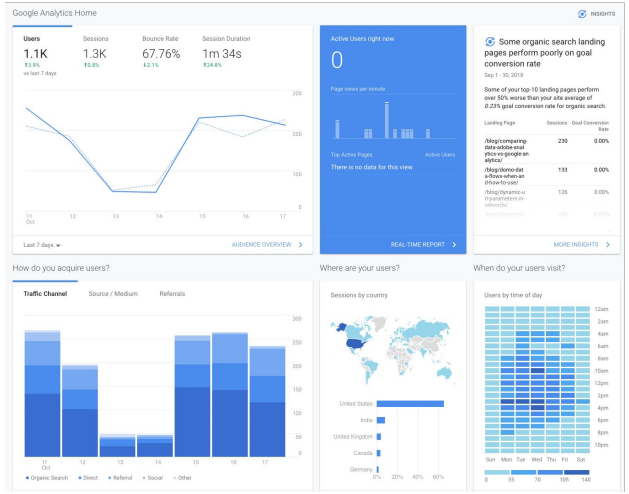
What is the variable?

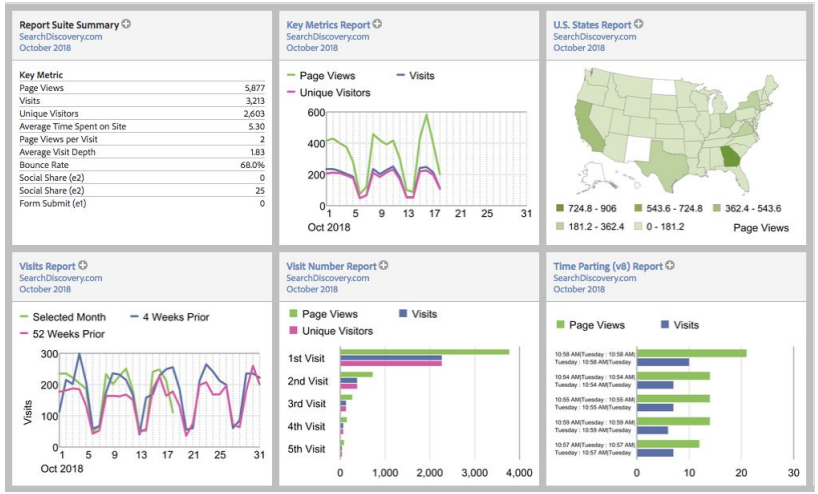
What are the values of the variable?

Channel	Sessions
Paid Search	818,993
Organic Search	651,252
Social	437,543
Direct	318,831
Display	143,332
Email	80,697
Referral	59,318

Visit ID	Channel
7521516054.5930117975	Organic Search
6639584132.5334606522	Organic Search
2072704531.5757008375	Organic Search
3700666076.4855492580	Social
3901278585.2803798674	Social
9493011014.8450893410	Referral
4257955221.2341693619	Organic Search
5653594021.2053548613	Paid Search
2103008782.3786630871	Organic Search
2156814867.9852847980	Direct
3969242874.8434979641	Paid Search
8975736469.8764719209	Social
3127960728.4142998058	Social
3863953834.9545306506	Organic Search
1769847183.6232068180	Organic Search

We are used to working with aggregated data!





+

Google Data Studio

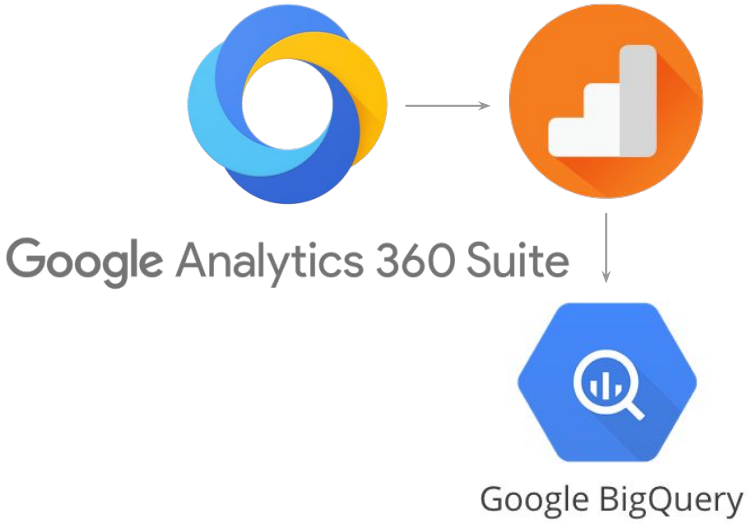
+

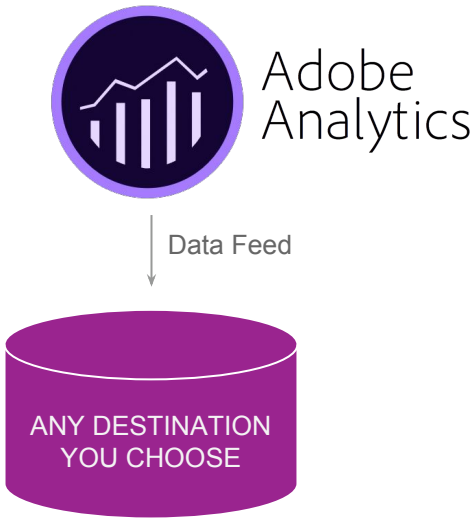
Analysis Workspace

+

Reporting APIs for Either Platform

Google and Adobe know this!





SNOWPLOW

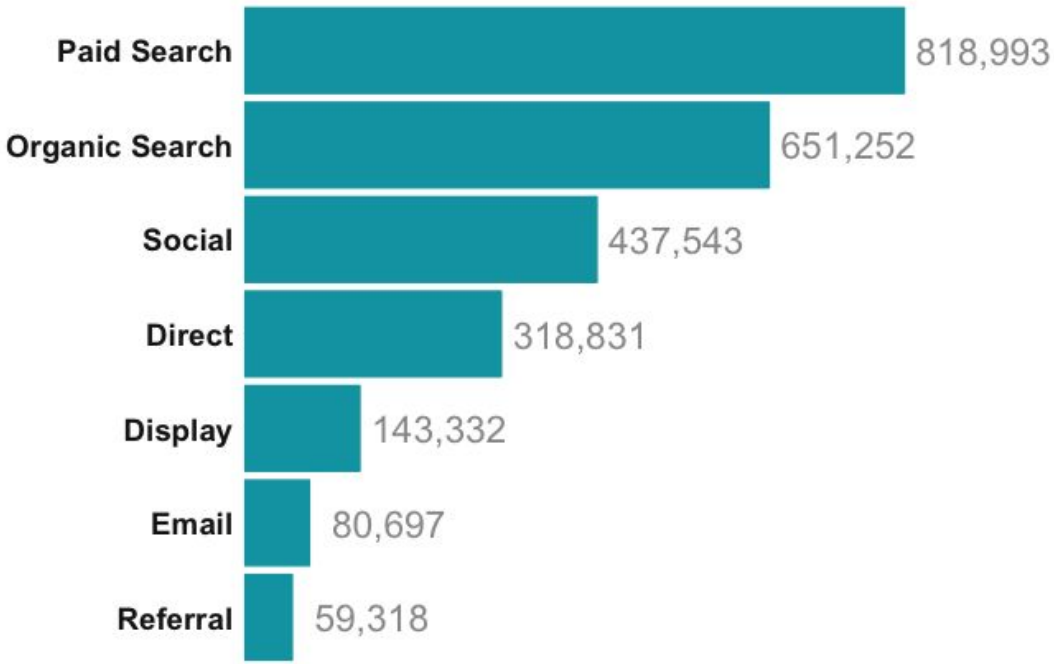
HEAP


Should we abandon all hope if we don’t have session-level detail?
No!
FEAR

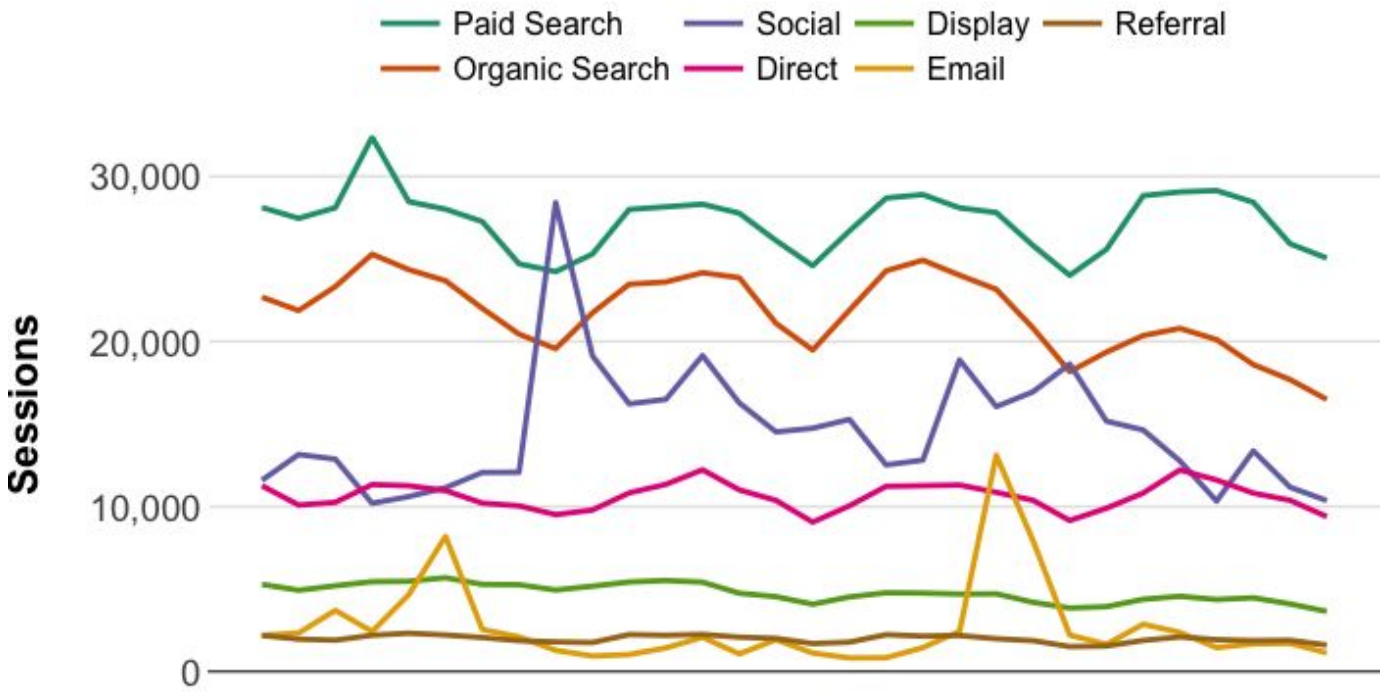
We just have to (carefully) cheat a little bit.

Let’s ask a different question.

Do sessions **really** differ by channel?



But if we eyeball it by day, things get murkier.



Deaggregation
(if not ideal detail)

30 Days

x 7 channels

---

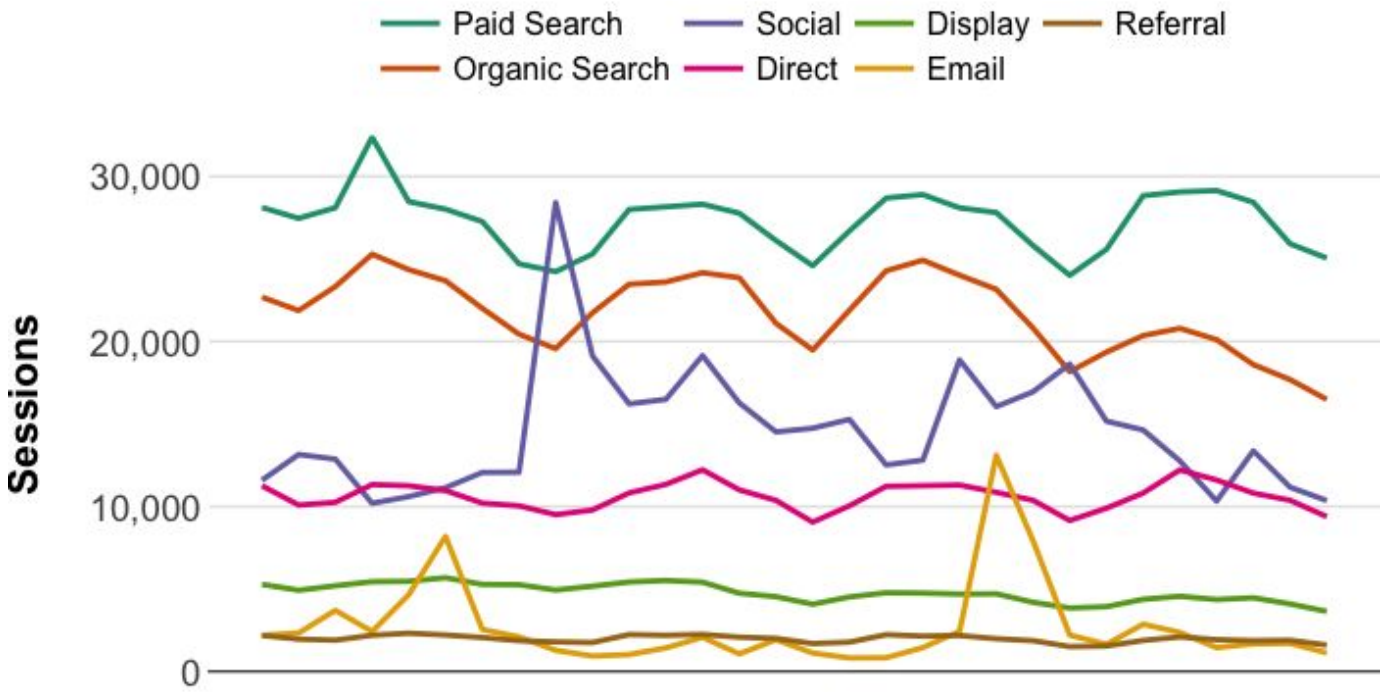
210 observations

Here’s where we’re going to cheat a little bit.

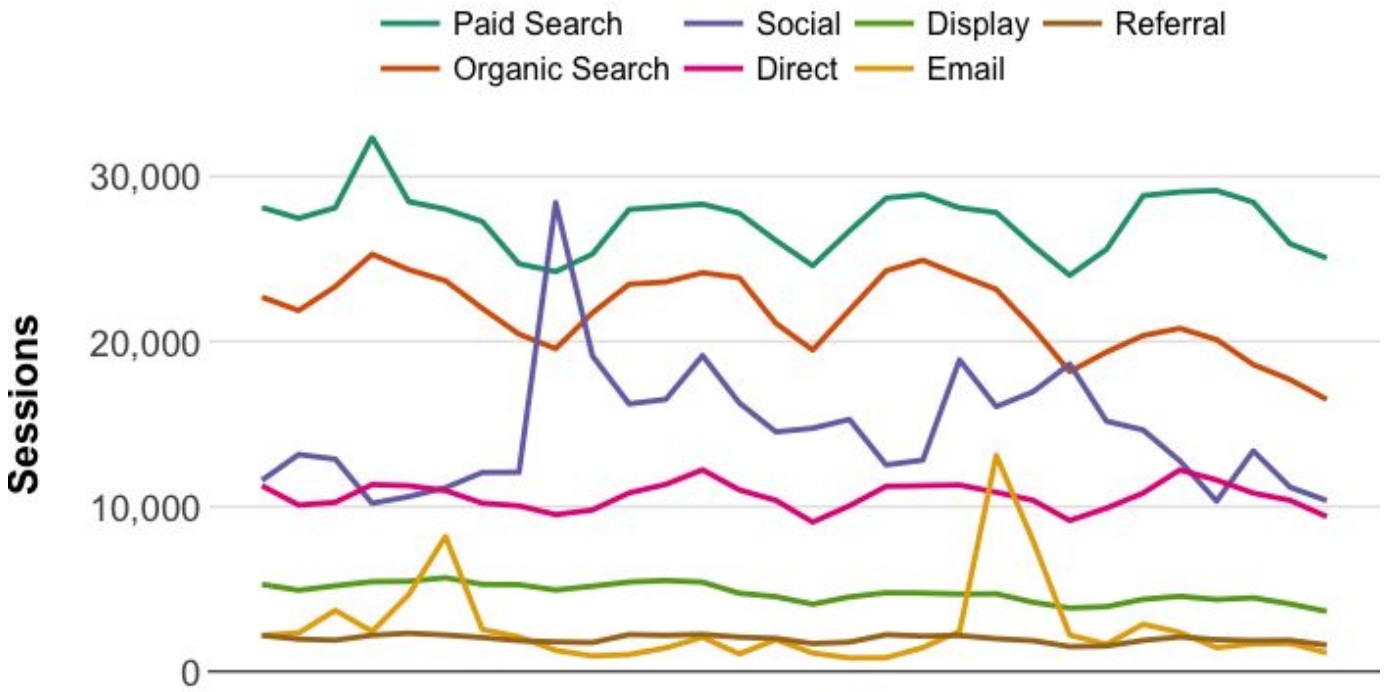
We're going to use **day**...

...as the **observational unit** for our analysis.

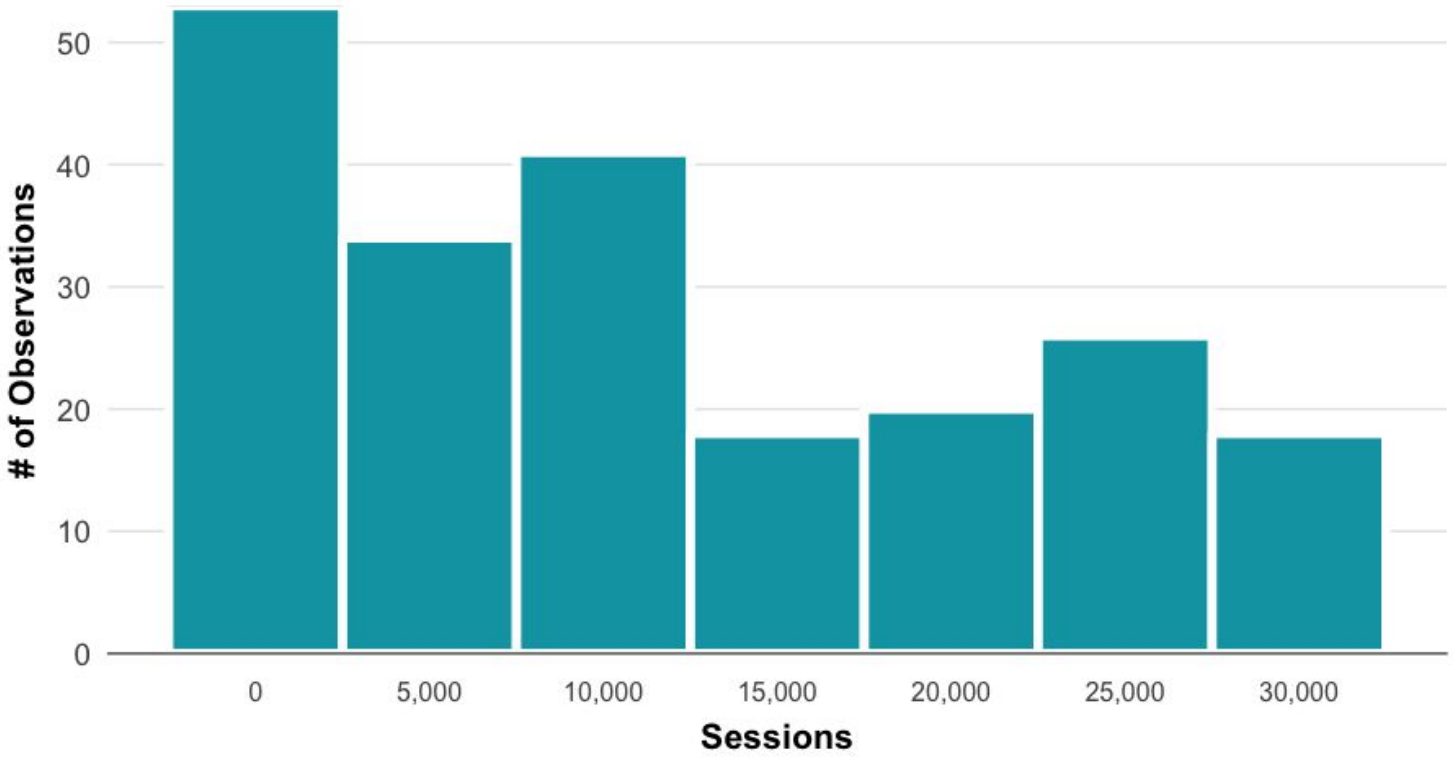
“Day” seems like an okay way to deaggregate.



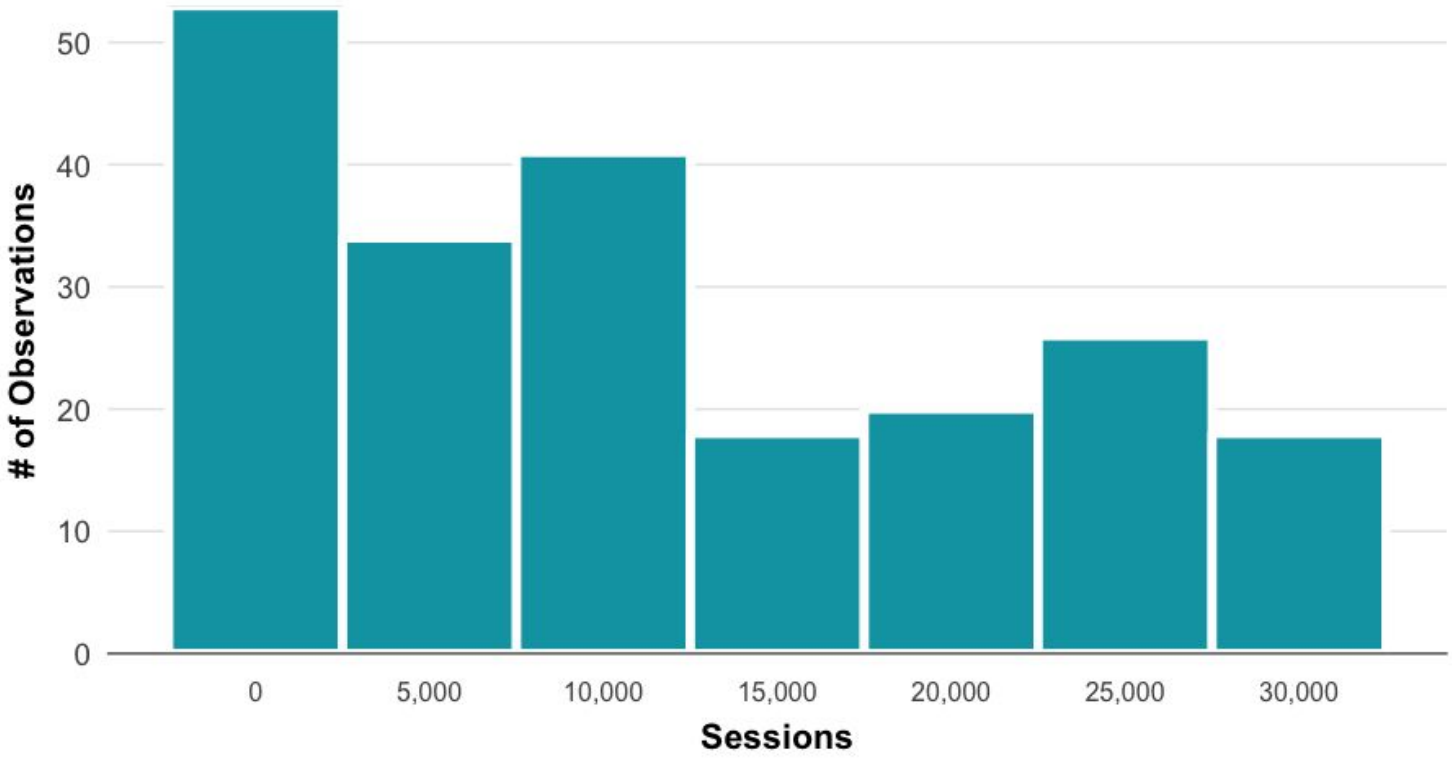
Let’s ignore the **time aspect of date**.



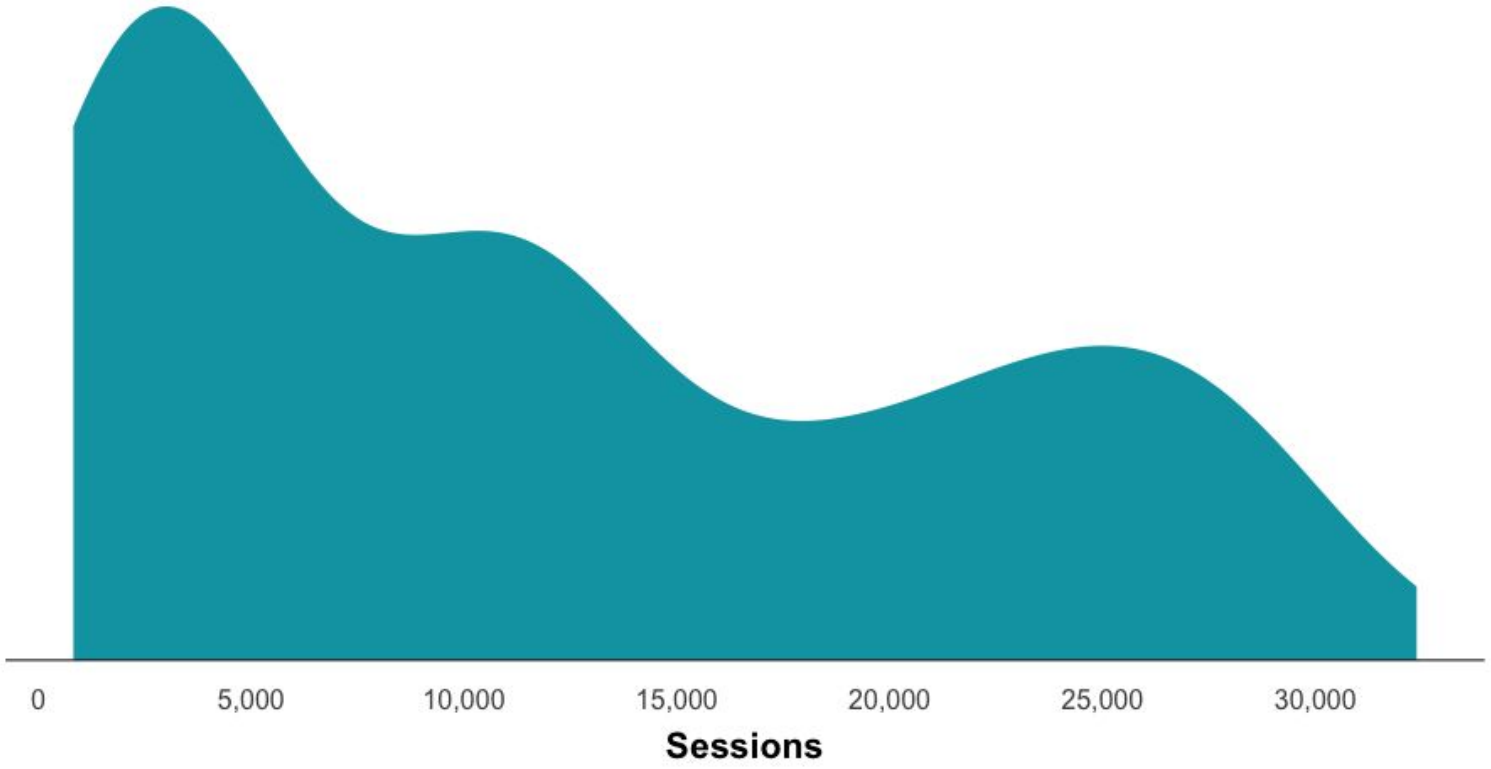
We can look at the data as a histogram (n = 210).



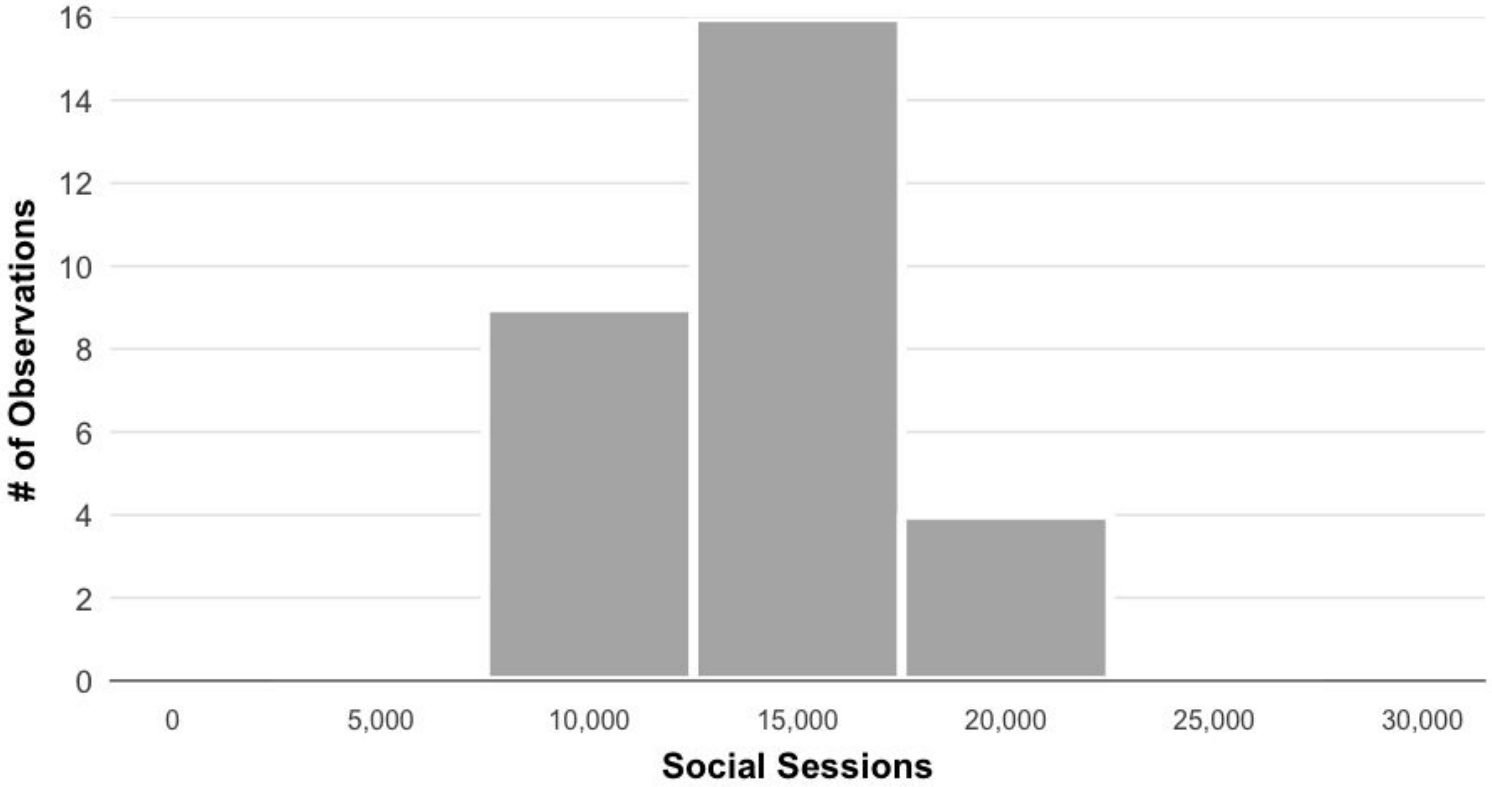
It’s the distribution of the observations.



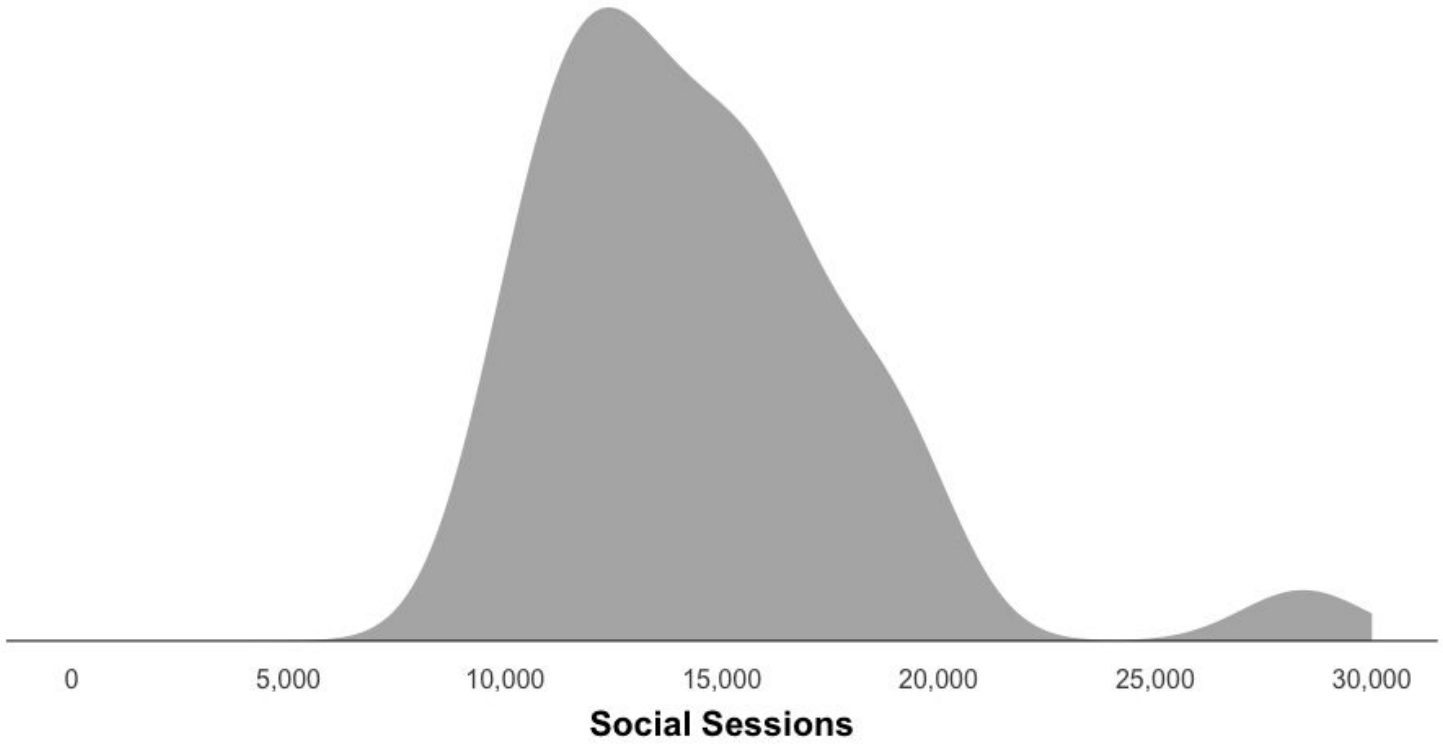
A **density plot** provides a smooth distribution.



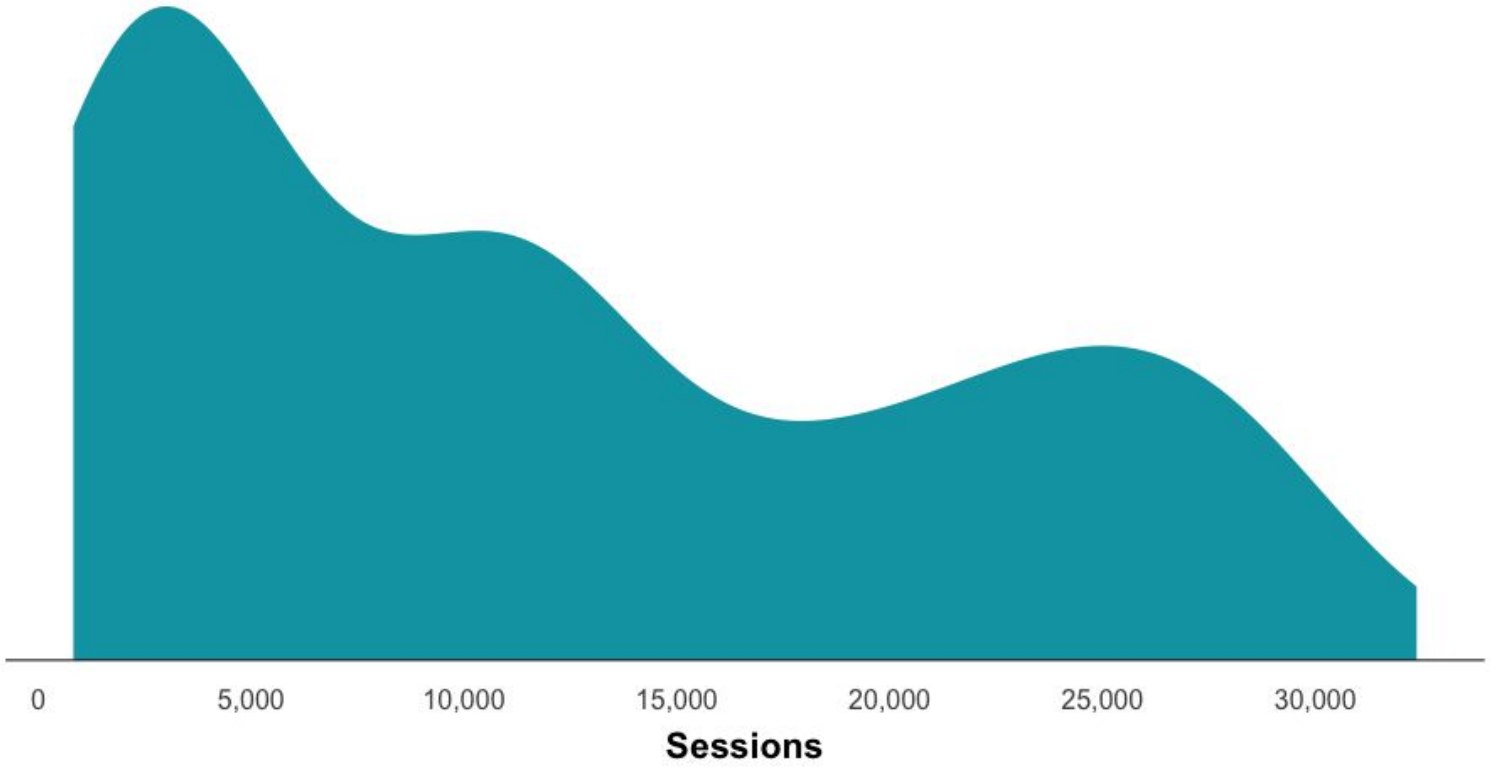
Let’s look at just one channel (n = 30)



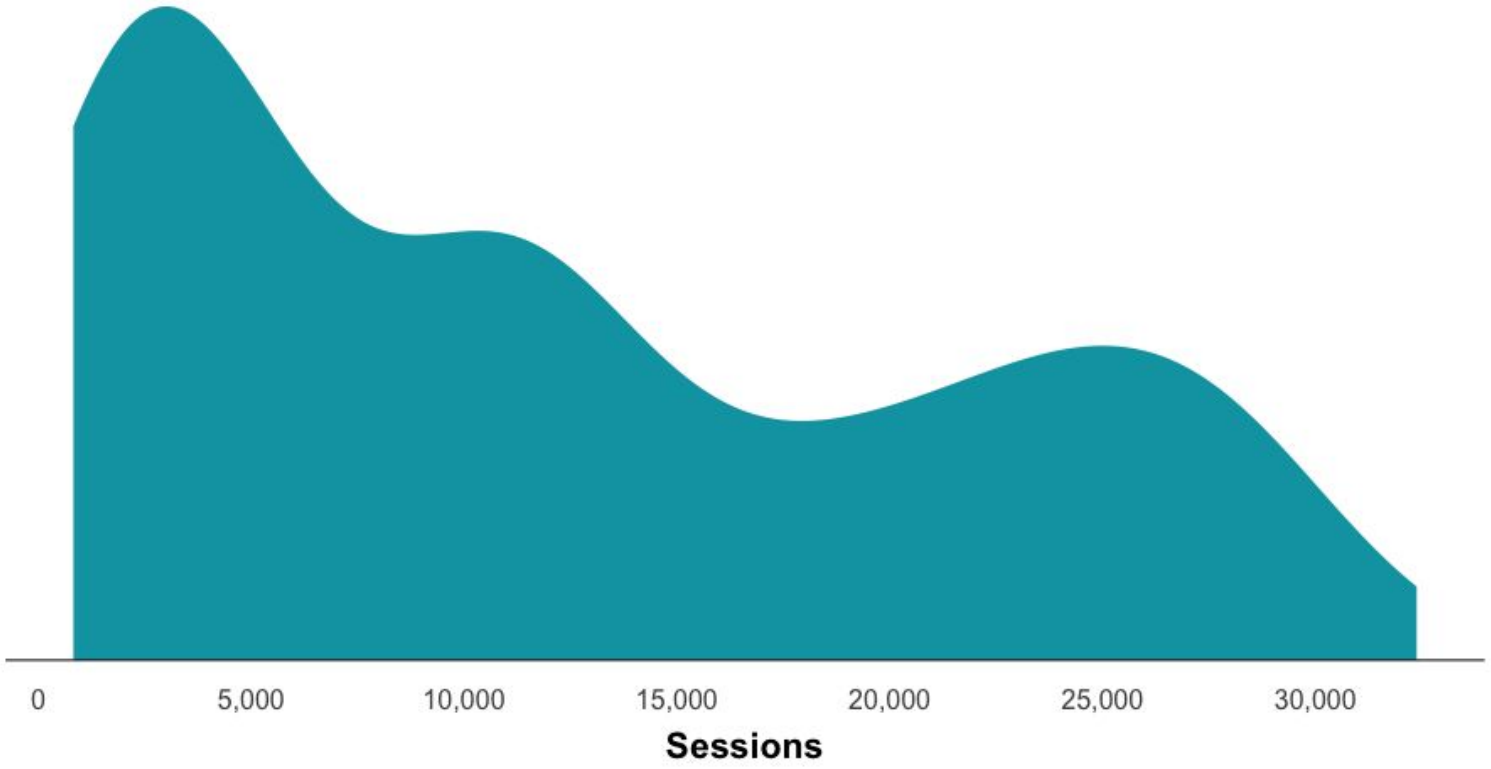
Let’s look at just one channel (n = 30)



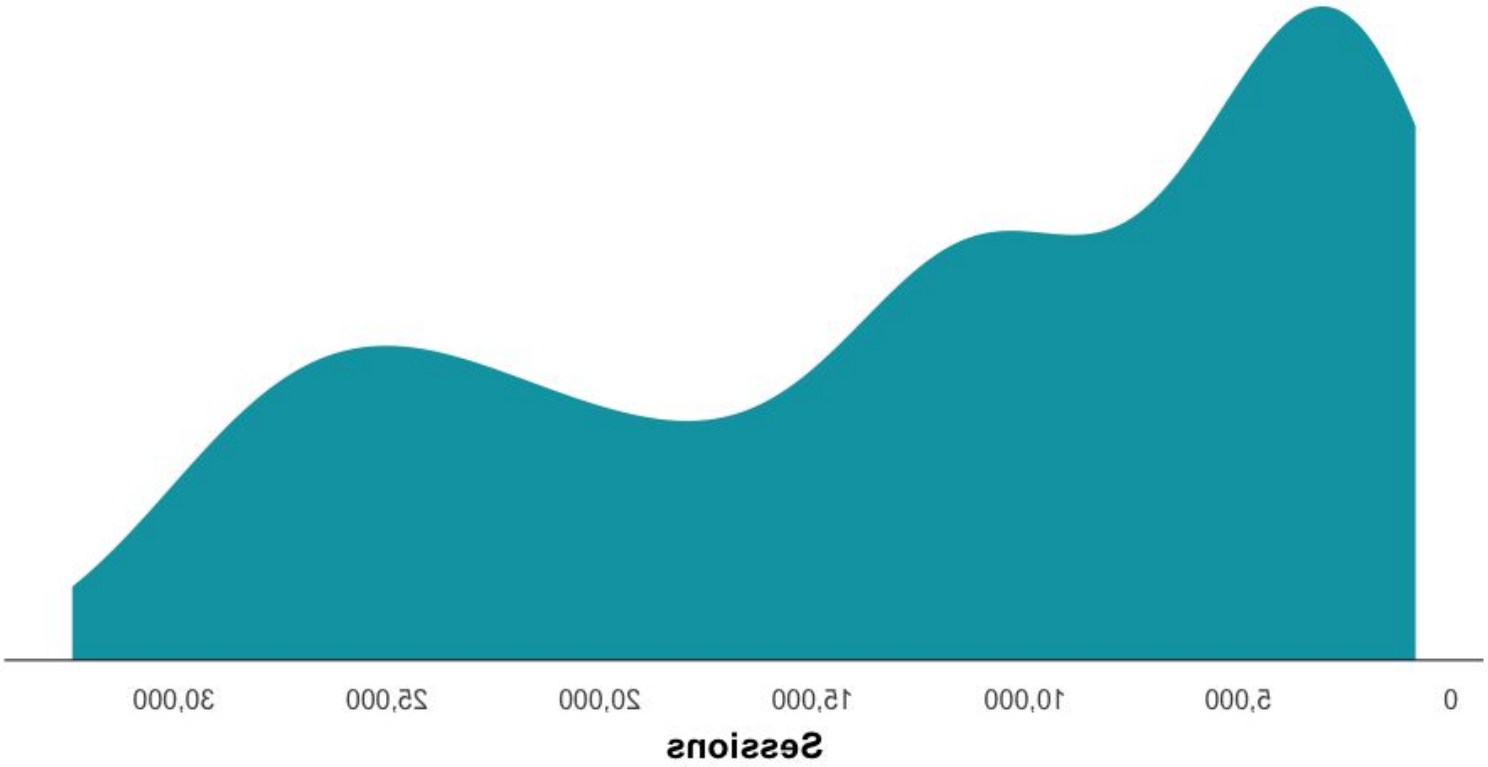
Start with our overall distribution (n = 210)



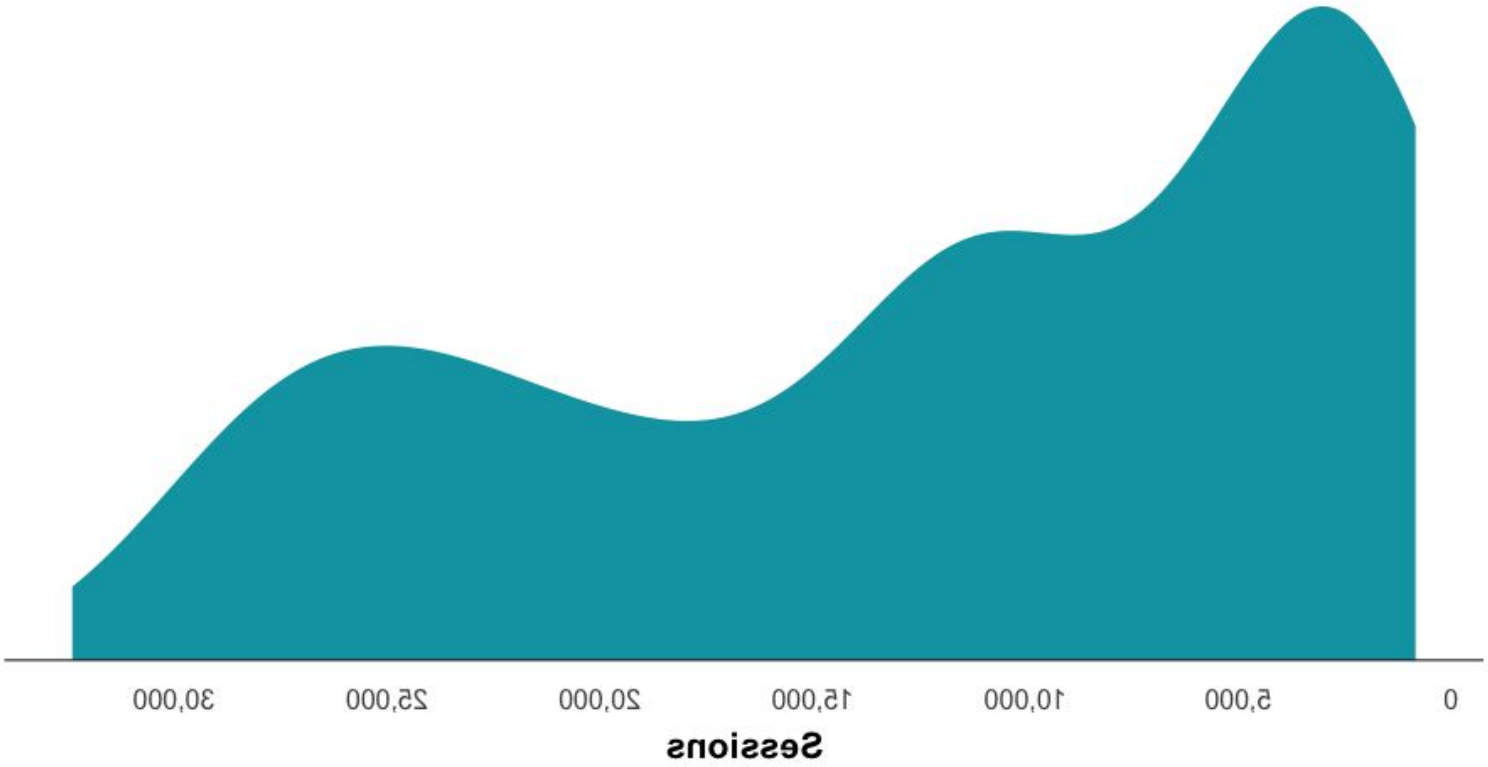
Flip it horizontally.



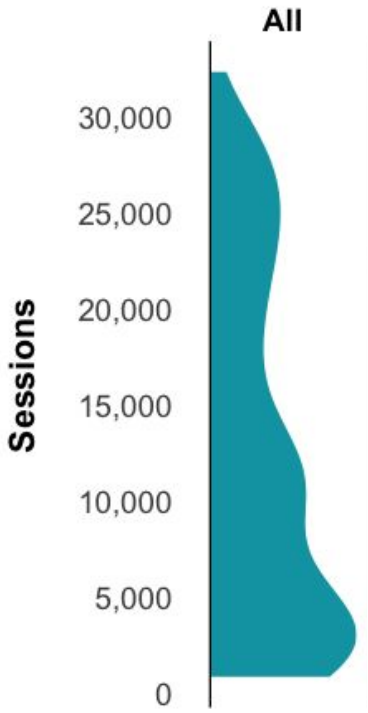
Flip it horizontally.



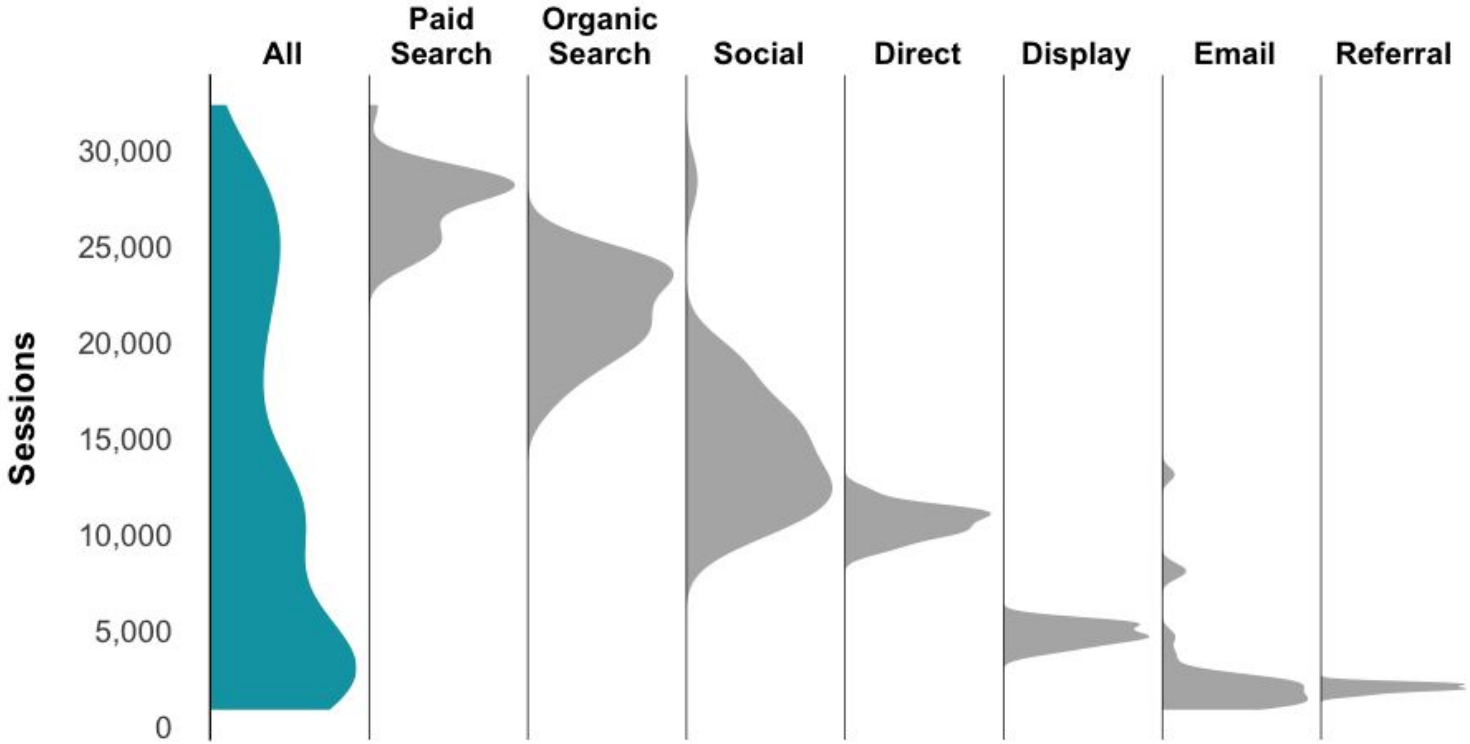
And rotate it and squish it.



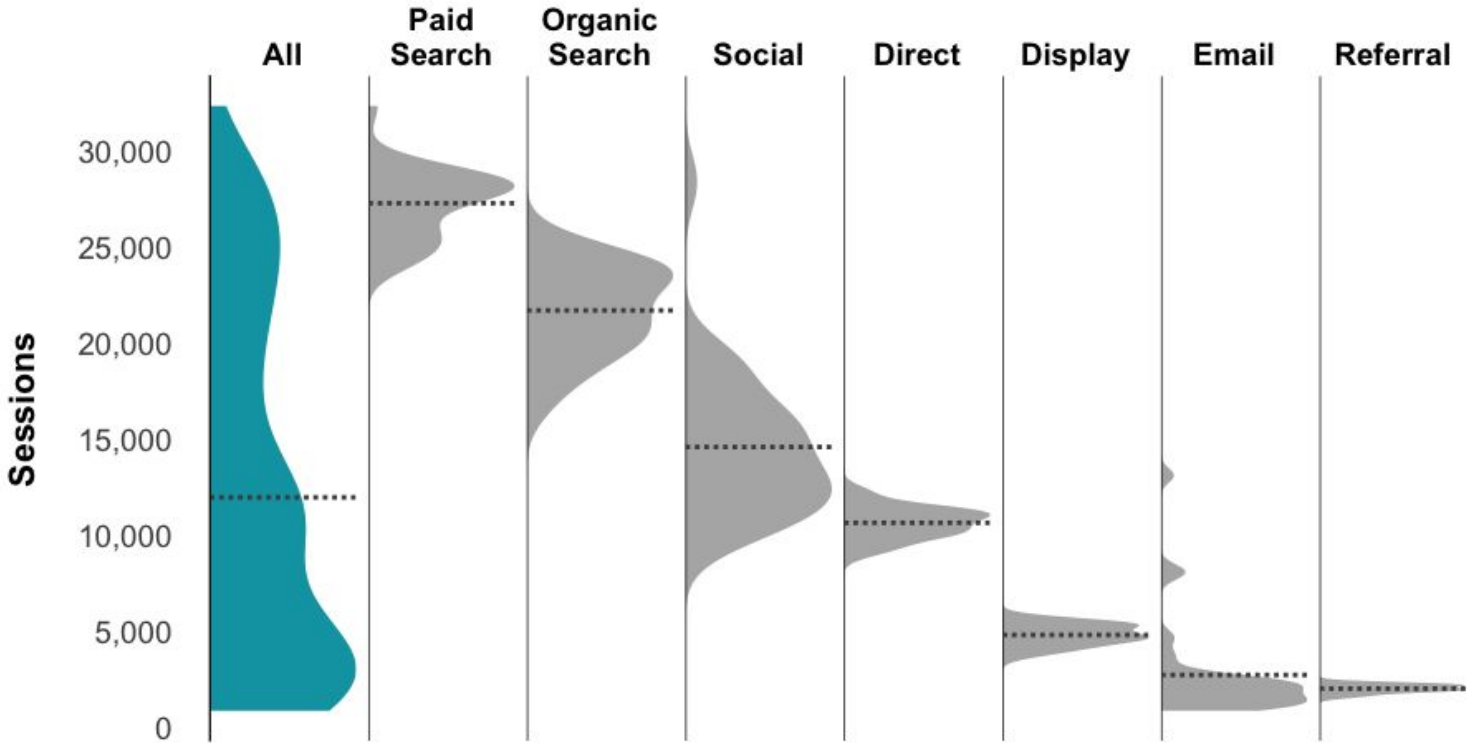
And rotate it and squish it.



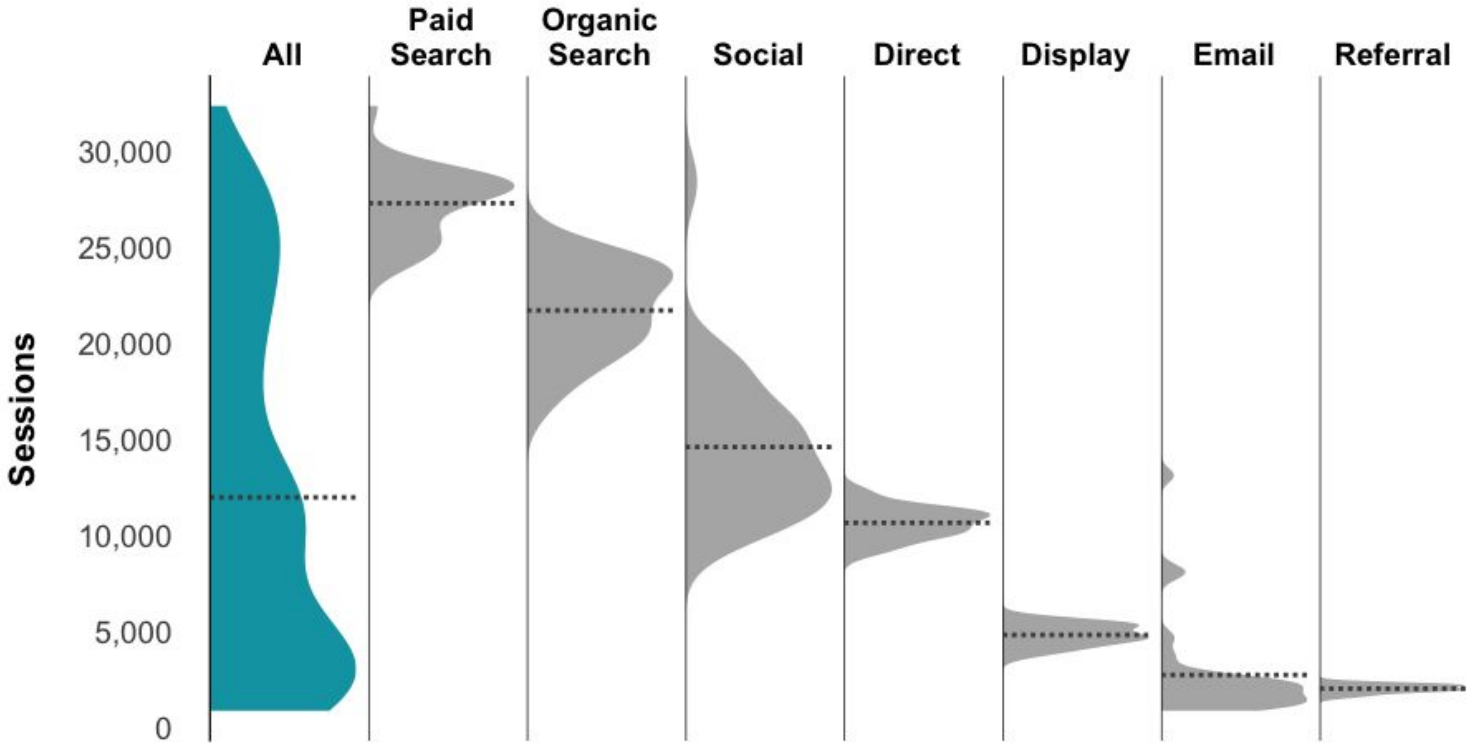
“All” is the sum of the channels.



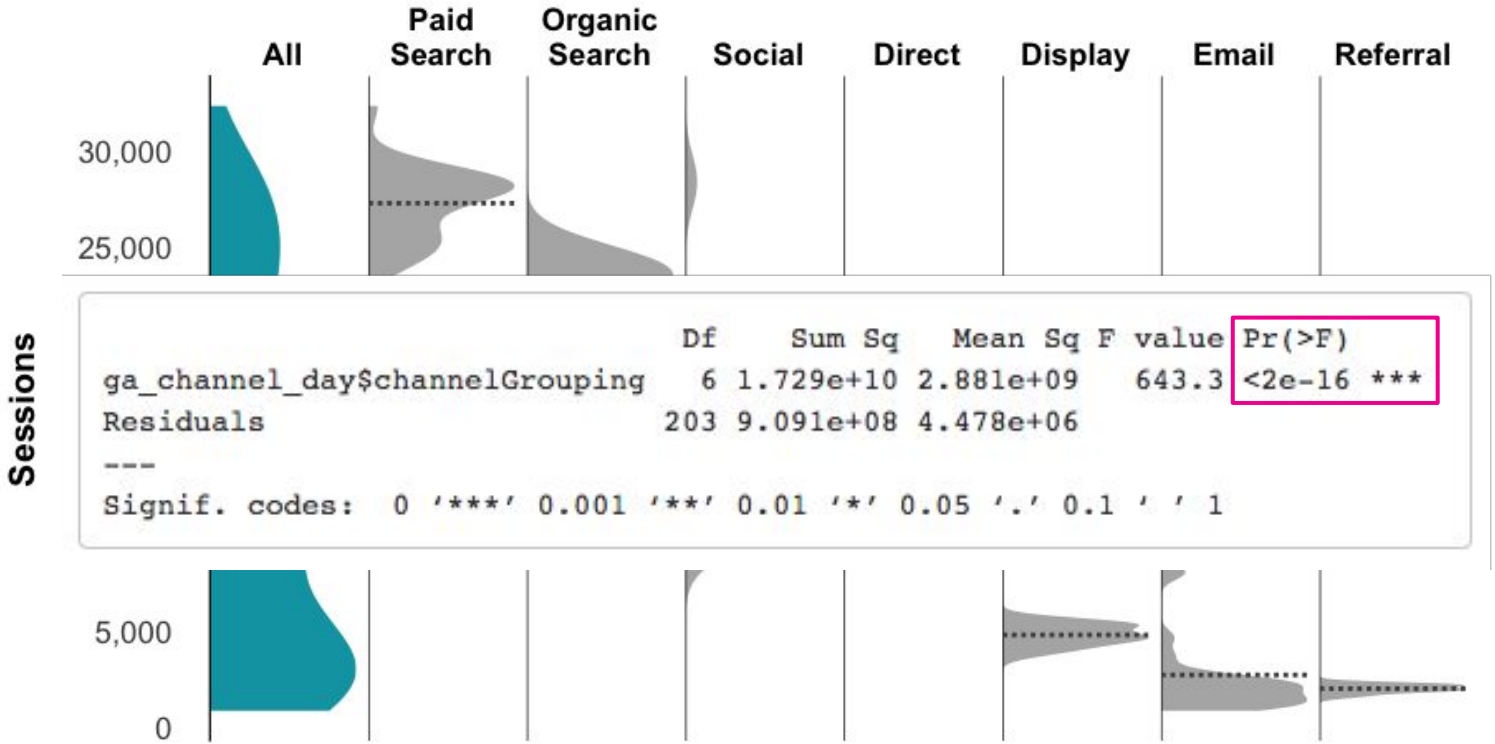
And we can add the **mean** for each distribution.



We're set to do an ANalysis Of VAriance

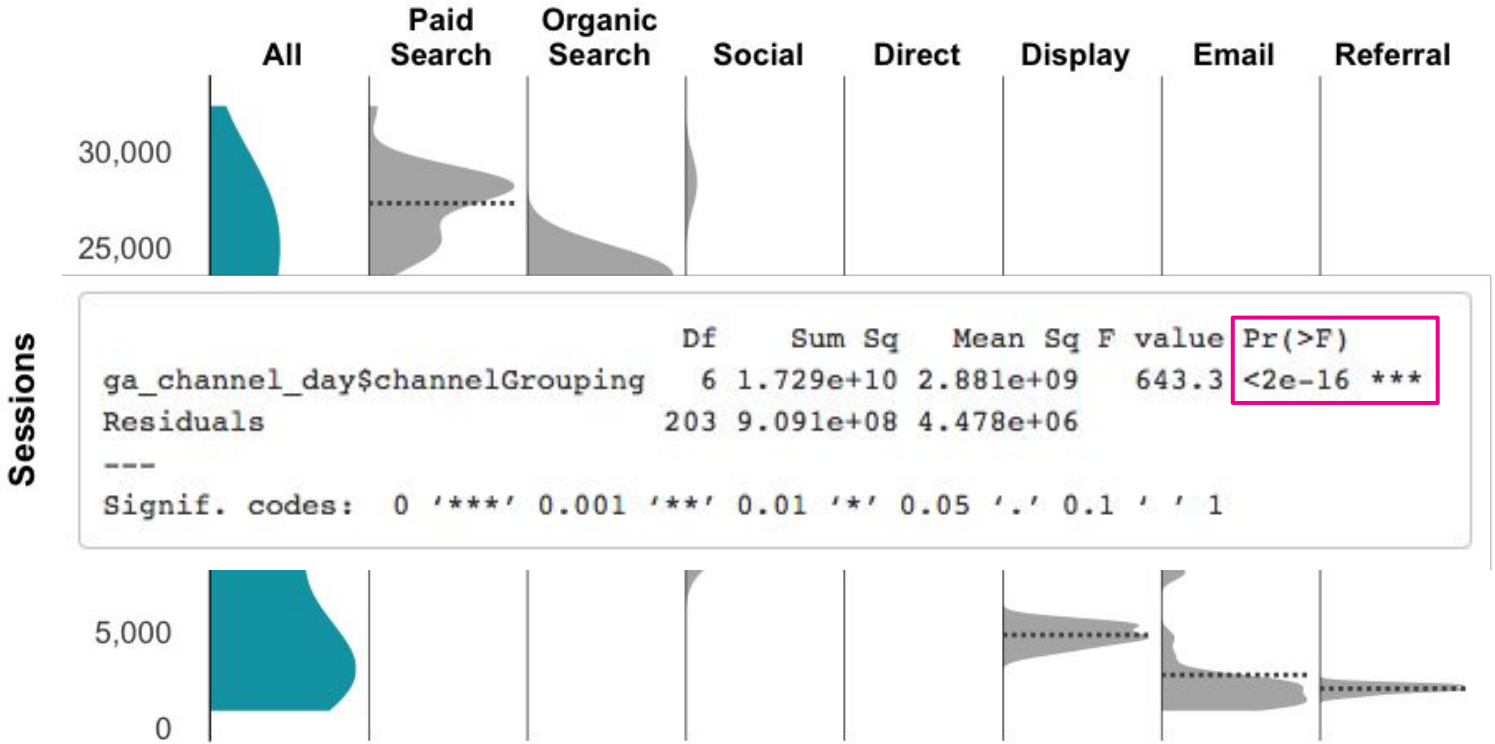


We’re set to perform a **1-Way ANOVA**.




```
                                 Df    Sum Sq   Mean Sq F value Pr(>F)
ga_channel_day$channelGrouping    6 1.729e+10 2.881e+09   643.3 <2e-16 ***
Residuals                       203 9.091e+08 4.478e+06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

WE can **reject the null hypothesis**.




```
                                 Df    Sum Sq   Mean Sq F value Pr(>F)
ga_channel_day$channelGrouping    6 1.729e+10 2.881e+09   643.3 <2e-16 ***
Residuals                       203 9.091e+08 4.478e+06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Null Hypothesis (H_0):

Sessions do not really ("significantly") differ by channel.

So...what we're saying:

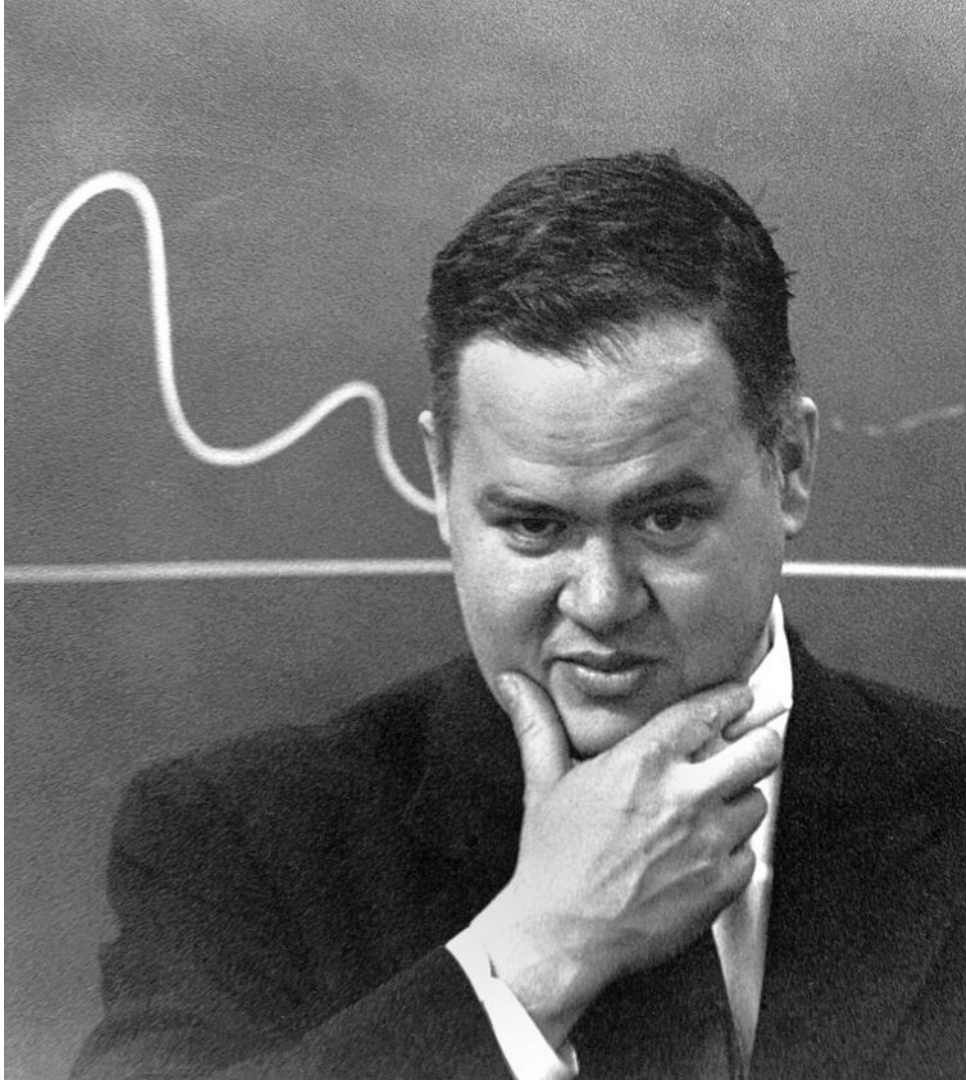
*Sessions **do** appear to differ by channel.*

An ANOVA does not tell us ***which*** *channels differ.*

*But a **post hoc analysis** can do that.*

John Tukey

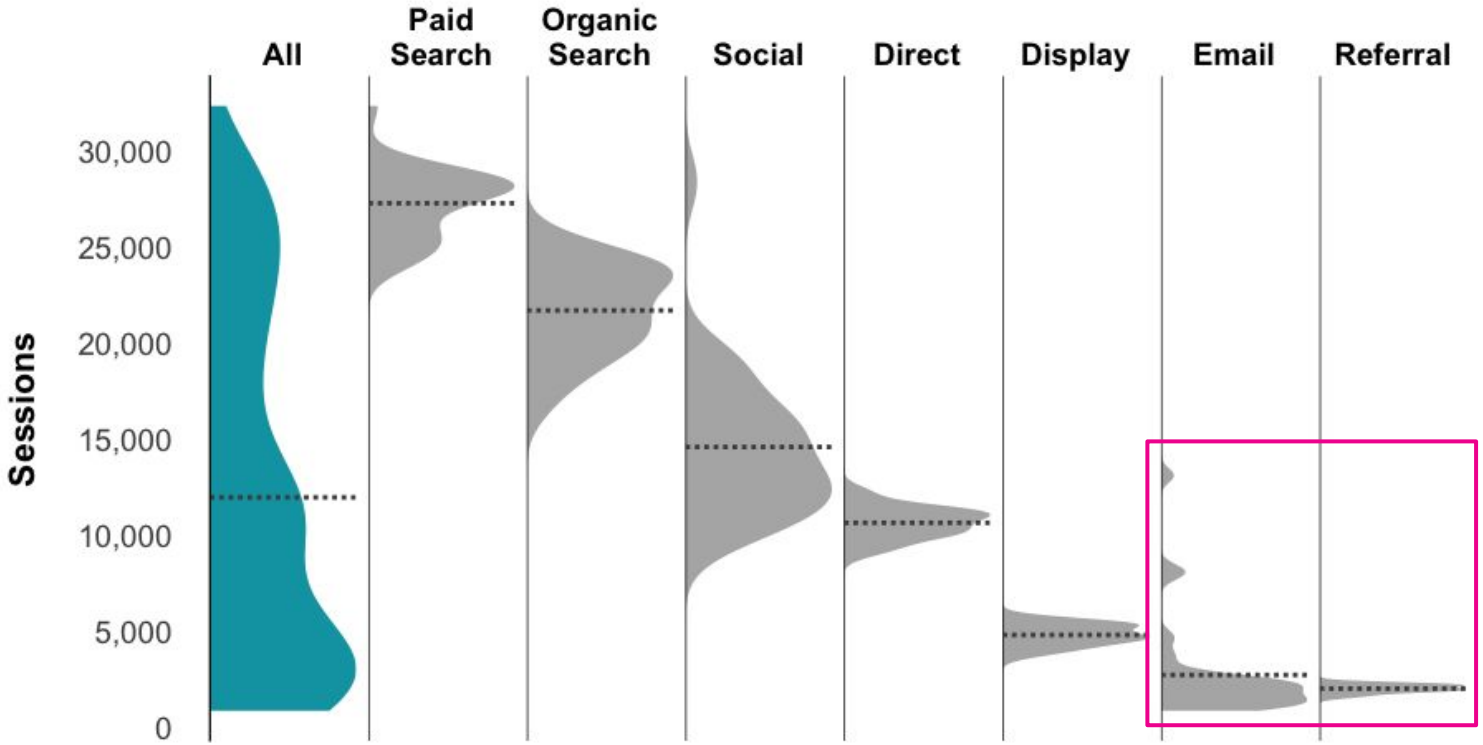
...brought us the
Tukey Post Hoc Test.

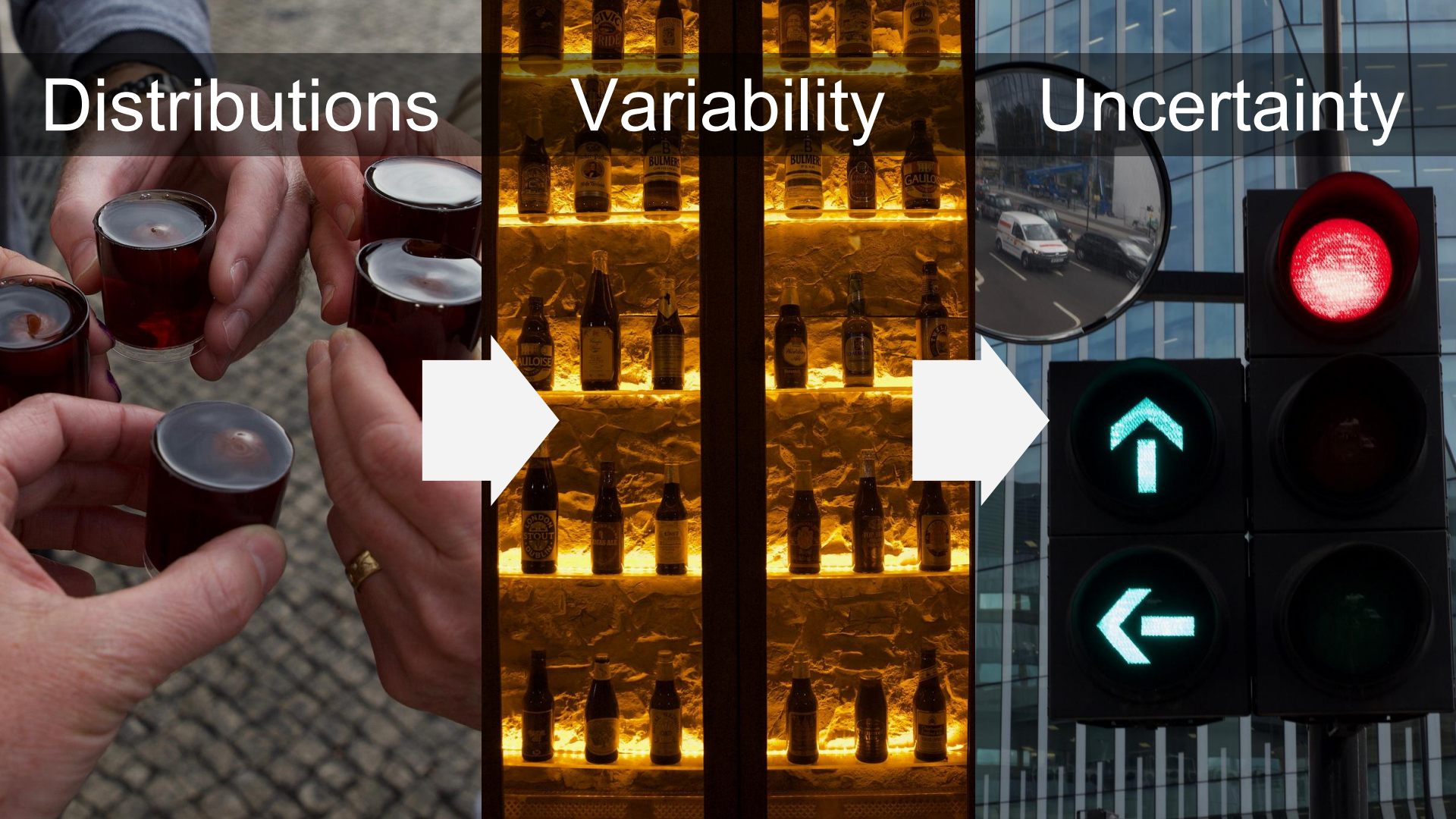


Tukey Sayeth: Most Channels Differ

	Referral	Email	Display	Direct	Social	Organic Search
Email	0.84938					
Display	0.00001	0.00330				
Direct	0.00000	0.00000	0.00000			
Social	0.00000	0.00000	0.00000	0.00000		
Organic Search	0.00000	0.00000	0.00000	0.00000	0.00000	
Paid Search	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

The result matches our intuition!




Distributions
Variability
Uncertainty

“We have a problem in [marketing] with thinking probabilistically.”

- Annie Duke

This is really important!

All marketers inherently operate under **conditions of uncertainty**.

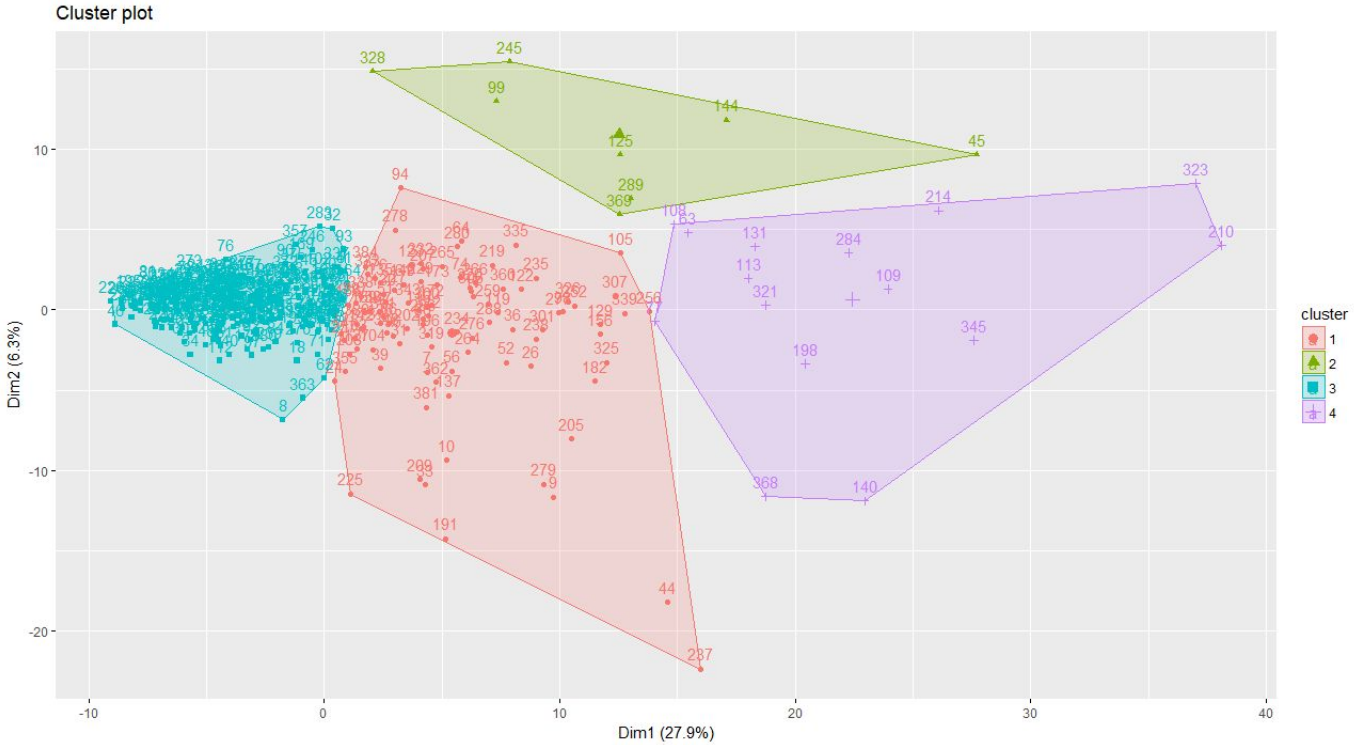
There is a **cost to reduce uncertainty**.

Uncertainty **can be eliminated**

Matt Gershoff
CEO, Conductrics

So...Data Science

k-means Clustering of Franchises



Propensity Modeling

Uncertainty!

Computer Science
(SQL + Python)

Web Analytics Data

CRM Data

POS Data

Integrate the Data

Model

Train

Predict

Analytics

Personalization

Retargeting

User-Level Data

Python + Statistics

So...where to begin?

bit.ly/data-science-y