Marketing & Data Analysis

Data sources

Santiago Alonso-Díaz

Tecnólogico de Monterrey
EGADE, Business School

Photo: Dalle2

Table of contents

External Data

Definition

Relevant data that is not produced by the company's activity

Examples

- Demographics
- Market shares
- Economic indicators
- Social media trends

Some sources

- Google Trends
- Statista
- INEGI
- Kaggle
- World Bank
- Passport Euromonitor
- Pew Research Center
- Mintel
- Web scrapping

Some samples

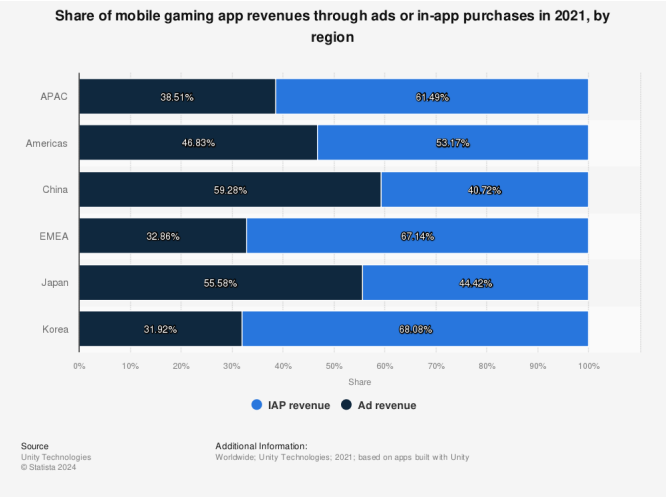


Figure: Statista

Some samples

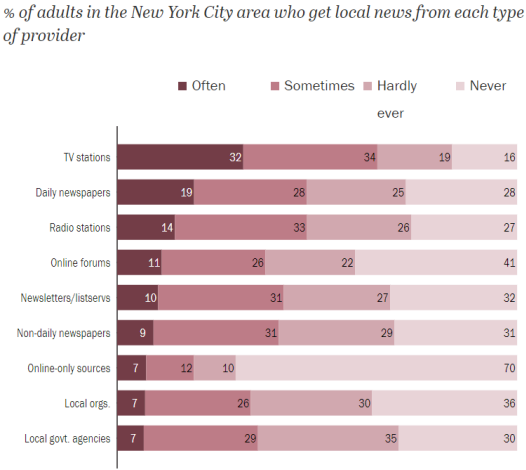


Figure: Pew Research Center

Some samples

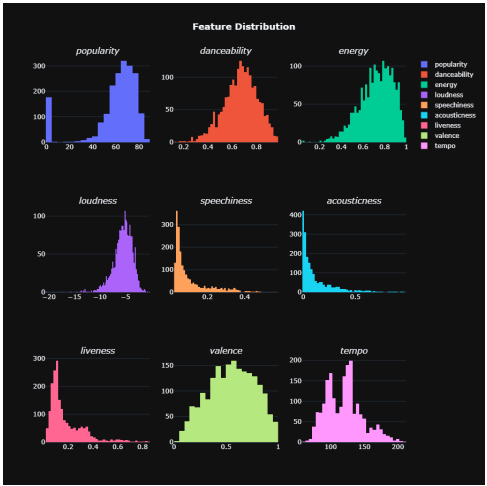


Figure: Spotify Song Features. Kaggle

Data Brokers

Where do firms get data? One source is online data brokers

In principle, with site cookies they can identify, say, age (e.g. click on retirement plans) or gender (e.g. click on tampons). Then, follow that user in their browsing behavior.

Data brokers then categorize individuals into audiences based on browsing behavior. This is useful. For instance, just show an ad to an specific audience.

Data Brokers

Can data brokers identify audience based on cookies? Let's test a campaign

Table 2. Study One: Campaign Criteria Given to Ad Platform

Criteria	Detail
Prespecified audience	Males between the age of 25 and 54
Campaign size	100,000 advertising impressions. Each time a display ad is shown on a website to a user, this counts as an impression.
Frequency	As many unique users as possible. Each user should see one impression, rather than one user seeing multiple impressions.

Figure: Neumann et al., 2019

Data Brokers

Table 3. Study One: Variable Definitions

Metric	Explanation
Accuracy	Percentage of impressions that were delivered to an audience that identified as male between 25 and 54 years old
Frequency	Average frequency of the campaign or how many impressions each viewer saw
Brand safe	Percentage of impressions that were served in a brand-safe environment
Nonhuman	Percentage of invalid (bot) impressions

Figure: Neumann et al., 2019

Data Brokers

Table 4. Study One: Campaign Performance Results

DSP	Accuracy, %	Frequency	Brand safe, %	Nonhuman, %
1	72	1.01	99.8	1.4
3	68	1.20	98.4	2.4
2	66	1.03	92.9	2.8
4	57	1.15	89.3	4.1
5	40	1.41	84.3	5.0
6	50	1.13	74.4	6.5
Average	59	1.15	89.9	3.7

Notes. Demand side platform (DSP) identities are anonymized. Accuracy refers to identifying males between the age of 25 and 54. See Table 3 for precise definitions.

Figure: On average, the campaign had a 59% accuracy. Compared to random (in Internet, 26.5% are males, 25-54), it is a 121% gain (59% / 26.5%)(Neumann et al., 2019)

Data Brokers

Table 9. Study Three: Data Broker Accuracy for Audience Interests

	Fitness interested		Sports interested		Travel interested	
Data broker	Accuracy, %	Sample size	Accuracy, %	Sample size	Accuracy, %	Sample size
Vendor A			86.2	571	64.7	697
Vendor B			91.0	1,428	64.0	2,564
Vendor C	81.2	611			74.0	704
Vendor D	78.6	117			83.5	127
Vendor E			89.6	4,371	87.8	1,753
Vendor F	82.1	196	86.0	285	67.5	243
Vendor G	83.2	393	86.3	729		
Vendor H	82.3	327				
Vendor I	82.4	307				
Vendor J			89.5	8,772	78.2	10,936
Vendor K			82.8	128	58.9	124
Vendor L			86.7	360	62.4	412
Vendor M	85.9	199	86.7	495	63.8	574
Vendor N			89.9	5,039	77.5	9,846
Vendor O	80.7	405	89.9	4,459	82.4	9,380
Vendor P			89.6	4,371	87.8	1,753
Vendor Q			86.9	604	67.5	499
Vendor R			82.1	168	78.2	10,904
Vendor S					65.9	857
Average	82.1	320	87.4	2,270	72.8	3,211

Figure: How about interest-based? It gets better, average accuracy can go up to 87.4% (Sports) (Neumann et al., 2019)

What did we learn?

Data from data brokers is useful. It is no necessarily perfect, but relative to random actions it is a good marketing strategy to consider that external data.

Internal Data

Definition

Relevant data that is produced by the company's activity and decisions

Examples

- User satisfaction
- Product revenue
- Segment growth
- Sale force efficiency
- Lifetime costumer value
- Costumer acquisition costs

Some sources

- Ads managers (e.g. Google ads, Facebook ad manager)
- Web managers (e.g. Google Search Console, Google Analytics)
- Focus groups (e.g. with customers)
- Interviews (e.g. with sale force)
- Experiments (e.g. A/B testing, Conjoint)

A/B test

Students: What A/B tests can you think of? In your jobs, business, life?

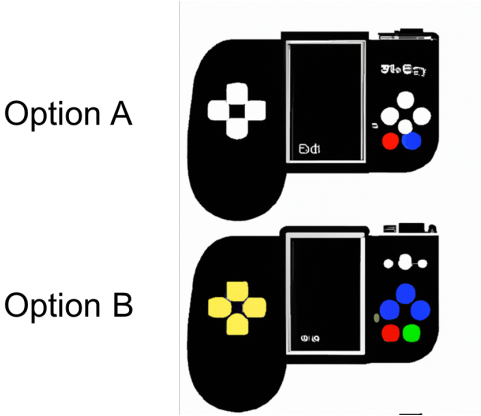


Figure: A/B testing

Startups and experimentation (Koning et al., 2022)

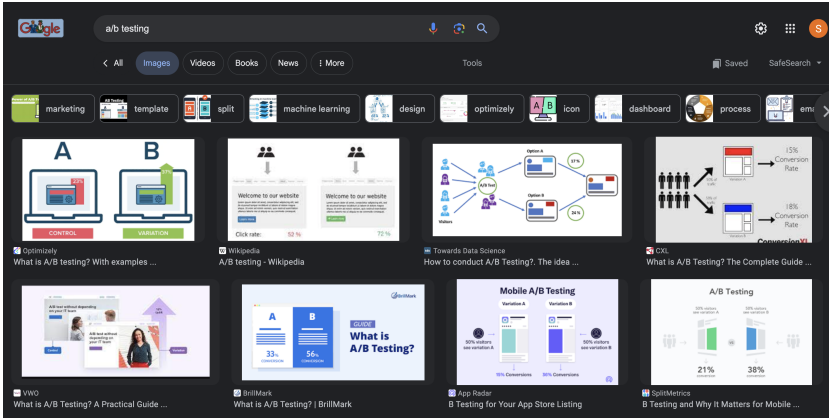


Figure: What is the role (or not) of experimentation in business (A/B testing)?

Startups and experimentation

Koning et al., 2022 evaluated 35,000 global startups over a 4-year period. Main results:

- A/B testing improves startup performance (e.g. introduce new products at higher rates).
- Venture capital (VC) startups do more A/B testing than non-financed startups.
- Silicon Valley startups do more A/B testing
- Regardless of the previous two results, A/B testing is beneficial to all.

Takeaway

"... experimentation helps drive both valuable incremental changes and the development of significant product improvements." (Koning et al., 2022, pp. 6436)

Exogenous shock (IV)

March 2017: Google launches Optimize and Optimize 360. Tools for A/B testing tools.

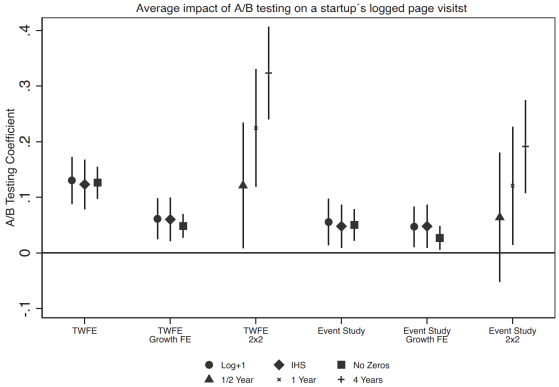
Results

Panel B: Start-up level		
	Number of start-ups	Percentage A/B testing
Not angel/VC funded	22,250	12.9
Angel/VC funded	13,012	25.2
Founded 2012–13	14,569	15.3
Founded 2010–11	11,966	16.5
Founded 2008–09	8,727	15.2
Outside United States	14,645	16.1
In United States, outside Bay Area	12,493	18.9
Bay Area	4,187	25.4
1-10 employees	15,393	13.0
11+ employees	19,840	20.7
Fewer than 1,500 weekly visits	17,189	8.1
More than 1,500 weekly visits	18,073	26.3
Commerce and shopping	4,517	24.1
Advertising	2,445	14.8
Internet services	2,079	17.2
Software	2,047	16.1
Data and analytics	1,940	21.6
Apps	1,746	17.1
Content and publishing	1,579	14.8
Financial services	1,547	23.6
Education	1,386	19.3
Information technology	1,233	20.0
Healthcare	1,042	19.2
Hardware	1,030	16.5
Other	12,671	14.2

Notes. Panel A provides summary statistics at the startup-week level. Panel B shows the number of startups of each type and the percent that use an A/B testing tool for at least one week during our panel.

Figure: Heterogeneity in A/B testing (Koning et al., 2022)

Results



Notes. "TWFE" indicates standard two-way fixed effects models, "Growth FE" indicates the model includes firm growth fixed effects, and "2 × 2" indicates that the estimate is from is a simplified difference-in-differences model that includes only data from the first week in our panel and a single observation either a half-year, year, or four years later. "Event study" indicates that only A/B switchers are included in the data. "IHS" indicates that we use the inverse hyperbolic sine instead of logged-plus-one visits. "No Zeros" indicates that all weeks in which page views are zero have been excluded from the data. All models include start-up fixed effects, week fixed effects, and a control for the size of the start-up's technology stack. Bars are 95% confidence intervals.

Figure: Robust effect on website visits (Koning et al., 2022)

Results

Figure 2. Event Study Plot Showing the Effect of A/B Testing over Time

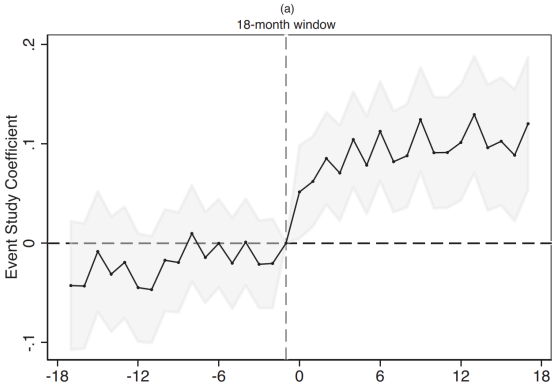
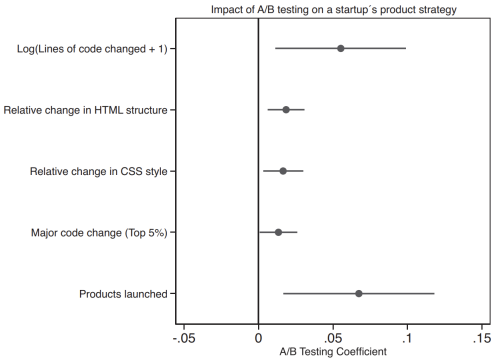


Figure: Persists over time (Koning et al., 2022)

Results

Figure 4. We Find That A/B Testing Does Not Lead to Incrementalism in Product and Website Development for the Nearly 10,000 Start-ups for Which We Have Website and Product Launch Data



Notes. Instead, these firms make larger changes to their website code, the structure of their homepage's HTML, and website style and are more likely to deploy major code changes. A/B testing firms are also more likely to launch a new product in a given week than those that do not.

Figure: A/B on other variables (Koning et al., 2022)

Conclusion

- Experimentation as a strategy (a single experiment may not work)
- Design the future with experiments
- Solving tension between routines and experimentation is critical
- Experimentation should aid innovation

Synthetic Data

Definition

Examples

Some Samples

- Structural models
- Synthetic controls
- Agent based models
- Generative AI (e.g. LLM labels)

Synthetic Controls

Sometimes we don't have an appropriate control for comparison.

One approach is to create synthetic controls: a weighted combination of other units so that it is as similar to treated units

NYT Synt. Controls (Pattabhiramaiah et al., 2019)

Is a paywall good or bad for newspapers? There are at least two externalities:

- Engagement of its online reader base
- Spillover effect on the print version

How to measure them for the New York Times? Create synthetic controls using other similar newspapers

NYT Synt. Controls (Pattabhiramaiah et al., 2019)

Table 1. Top Newspapers in the United States by Circulation: Comparison of *NYT*, *LAT*, and Control Newspaper Readership (Print + Online).

Newspaper	Rank in 2010	Rank in 2011	Rank in 2013	Circulation in 2010	Circulation in 2013
WSJ	1	1	1	1,752,693	2,378,827
USAT	2	2	3	1,671,539	1,674,306
NYT	3	3	2	1,086,293	1,865,318
LAT	4	6	4	1,078,186	653,868
WP	5	5	8	763,305	474,767
CT	7	7	10	657,690	414,930
NYDN	6	6	6	701,831	516,165

Notes: Source: AAM's annual Newspaper Audit Reports and http://www.thepaperboy.com/usa-top-100-newspapers.cfm.

Figure: Newspapers to build a synthetic control

NYT Synt. Controls (Pattabhiramaiah et al., 2019)

Table 4. Effect of the Paywall on *NYT* Online Visitation, Aggregate Data, and Generalized Synthetic Control.

	ln(Unique Visitors)		ln(Pages)		ln(Visits per Visitor)		ln(Pages per Visitor)		ln(Duration per Visitor)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
NYT × Paywall	−.184**	.029	−.428**	.073	.010	.125	−.104	.127	−.112	.148
# Observations: treated						1,025				
# Observations: control						5,125				

**$p < .01$.

Figure: Negative online effect of paywall for the NYT (relative to the synthetic control)

NYT Synt. Controls (Pattabhiramaiah et al., 2019)

Table 11. Effect of Paywall on Print Readership, Generalized Synthetic Control.

	All DMAs			
	Weekday Circulation Share (%)		Weekend Circulation Share (%)	
DV =	Est.	SE	Est.	SE
NYT × Paywall	.35**	.02	.34**	.03
N (treated)	202			
N (control)	404			

Figure: Positive print effect of paywall for the NYT (relative to the synthetic control)

NYT Synt. Controls (Pattabhiramaiah et al., 2019)

Business implications

- Paywalls do affect the NYT business
- Managers need cost-benefits analyses because paywalls in the NYT had different effects for online and print versions.

Python: loading, cleaning, exploring

Python

Let’s go to Python to find, load, clean, and explore some data

References

Koning, R., Hasan, S., & Chatterji, A. (2022). Experimentation and start-up performance: Evidence from a/b testing. *Management Science*, *68*(9), 6434–6453.

Neumann, N., Tucker, C. E., & Whitfield, T. (2019). Frontiers: How effective is third-party consumer profiling? evidence from field studies. *Marketing Science*, *38*(6), 918–926.

Pattabhiramaiah, A., Sriram, S., & Manchanda, P. (2019). Paywalls: Monetizing online content. *Journal of marketing*, *83*(2), 19–36.