CS341: Project in Mining Massive Datasets

Advanced ML in Google Cloud (2)

Abhay Agarwal (MS Design ‘19)

Agenda

- ‘Productizing’ analytics
- Data wrangling
- Data fundamentals
- Data studio vs datalab vs colab

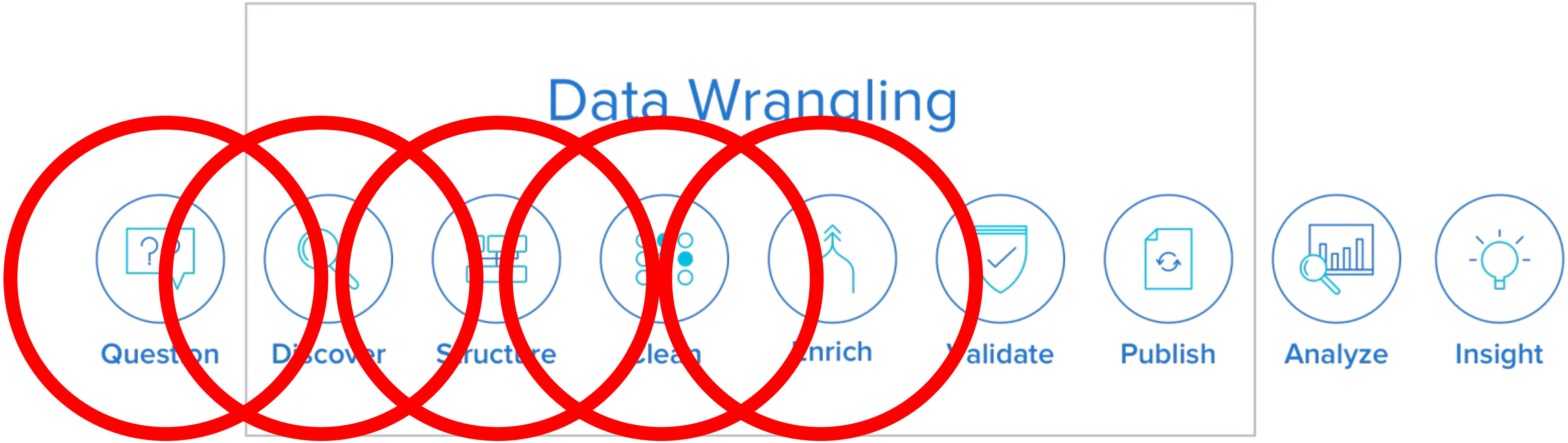
‘Productizing’

- What does it mean to ‘productize’ your ML?

Pitfalls in Productizing

- My algorithm has a 95% accuracy -- is it ready for production?
- My algorithm has a 95% accuracy and 95% precision -- is it ready for production?
- My algorithm has a 95% accuracy, 95% precision, and my training data is roughly sampled from real examples -- is it ready for production?
- My algorithm has a 95% accuracy, 95% precision, training data sampled from real examples, and my algorithm tests hypotheses that match the use cases -- is it ready for production?

Data wrangling



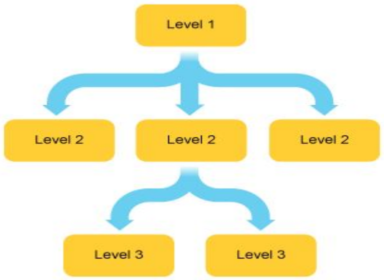
DATA COLLECTION FUNDAMENTALS

Kev Concepts

Quantity

Cost

Quality



Level 1

Level 2

Level 2

Level 2

Level 3

Level 3

Structure

Freshness

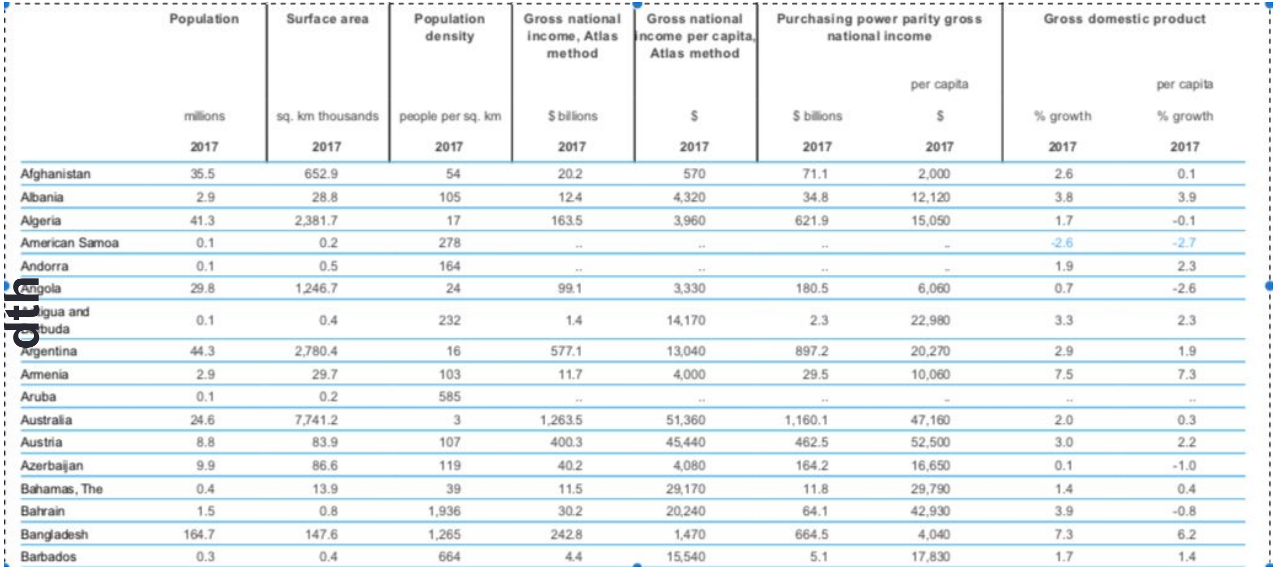
Quantity

- Breadth
  - Number of entities or observations
  - E.g., People, companies, stars, shopping trips,…
  - Ideally: comprehensive
- Depth
  - Data gathered on each entity or observation

Breadth and Depth

Depth

Brea
dth

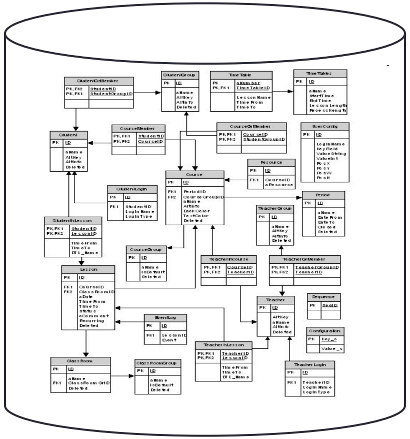


	Population	Surface area	Population density	Gross national income, Atlas method	Gross national income per capita, Atlas method	Purchasing power parity gross national income		Gross domestic product	
							per capita		per capita
	millions	sq. km thousands	people per sq. km	$ billions	$	$ billions	$	% growth	% growth
	2017	2017	2017	2017	2017	2017	2017	2017	2017
Afghanistan	35.5	652.9	54	20.2	570	71.1	2,000	2.6	0.1
Albania	2.9	28.8	105	12.4	4,320	34.8	12,120	3.8	3.9
Algeria	41.3	2,381.7	17	163.5	3,960	621.9	15,050	1.7	-0.1
American Samoa	0.1	0.2	278	..	..	..	..	-2.6	-2.7
Andorra	0.1	0.5	164	..	..	..	..	1.9	2.3
Angola	29.8	1,246.7	24	99.1	3,330	180.5	6,060	0.7	-2.6
Antigua and Barbuda	0.1	0.4	232	1.4	14,170	2.3	22,980	3.3	2.3
Argentina	44.3	2,780.4	16	577.1	13,040	897.2	20,270	2.9	1.9
Armenia	2.9	29.7	103	11.7	4,000	29.5	10,060	7.5	7.3
Aruba	0.1	0.2	585	..	..	..	..	..	..
Australia	24.6	7,741.2	3	1,263.5	51,360	1,160.1	47,160	2.0	0.3
Austria	8.8	83.9	107	400.3	45,440	462.5	52,500	3.0	2.2
Azerbaijan	9.9	86.6	119	40.2	4,080	164.2	16,650	0.1	-1.0
Bahamas, The	0.4	13.9	39	11.5	29,170	11.8	29,790	1.4	0.4
Bahrain	1.5	0.8	1,936	30.2	20,240	64.1	42,930	3.9	-0.8
Bangladesh	164.7	147.6	1,265	242.8	1,470	664.5	4,040	7.3	6.2
Barbados	0.3	0.4	664	4.4	15,540	5.1	17,830	1.7	1.4

World Bank Development Indicators

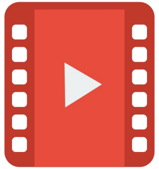
Structure

Structured



Semi-structured

Unstructured



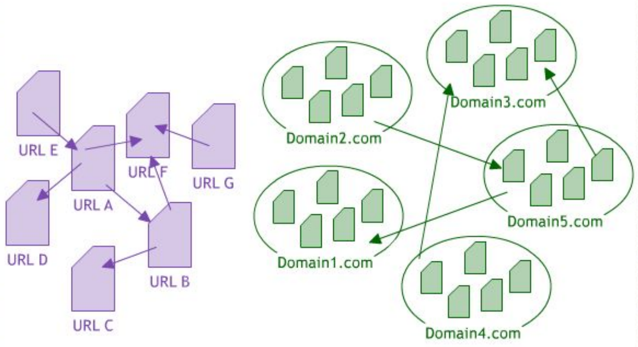
Graph Data

Graphs arise naturally in many settings

Many interesting techniques

e.g., Page Rank, community detection

Page vs. Domain-Level Link Graph



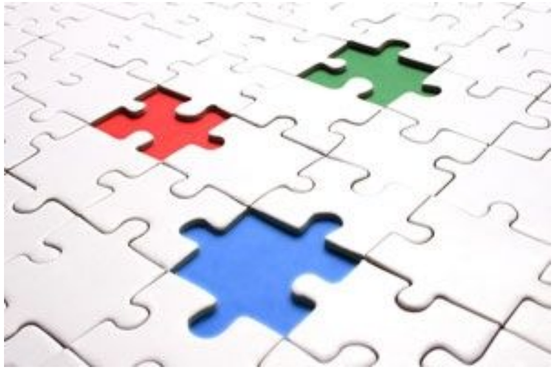
The page-level link graph counts links between individual URLs and values pages based on these links.

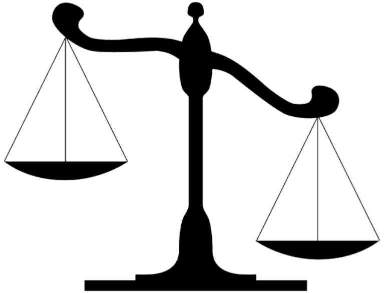
The domain-level link graph aggregates the links that exist on every page of a domain and considers only links that pass between unique sites to create domain-wide metrics.

Data Quality

- Errors
  - E.g., human labeling mistakes
- Missing data
  - E.g., missing addresses in customer records
- Bias
  - Sample bias, measurement bias, prejudice/stereotype

amazon mechanicalturk





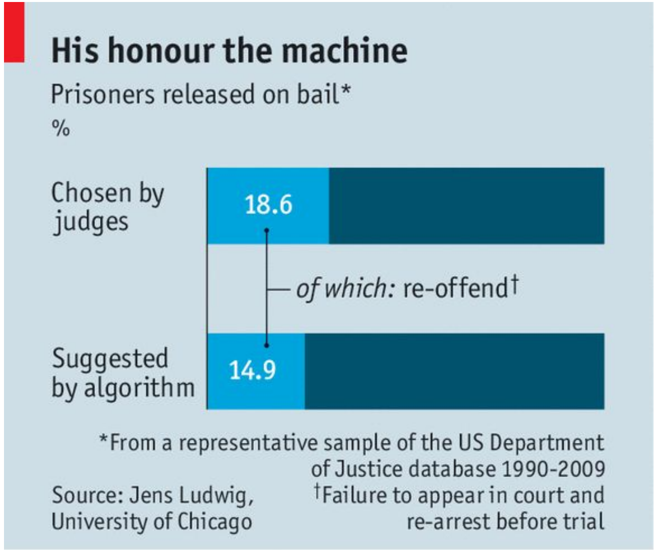
Data Quality: Sample Bias

Day Driving vs Night Driving

Tank recognition

Data Quality: Prejudice/Stereotype Bias Algorithmic Law Enforcement

Satoshi Kambayashi



His honour the machine

Prisoners released on bail*

%

*From a representative sample of the US Department of Justice database 1990-2009

†Failure to appear in court and re-arrest before trial

Source: Jens Ludwig, University of Chicago

Economist.com

But what about perpetuating bias against minorities?

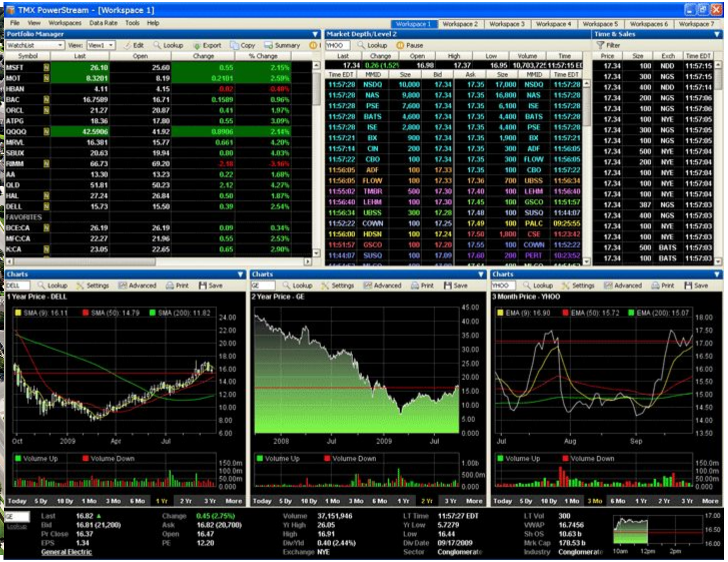
The Economist, August 20, 2016

Data Quality: Measurement Bias

HELLO, DO YOU HAVE ANY OPINIONS THAT FIT INTO OUR PRECONCEIVED QUESTIONS?
YES AND NO...
THANK YOU!
www.DOUBT.IT

Data Freshness

Rate of data collection must match rate of change of underlying phenomenon



Symbol	Last	Open	Change	% Change
MSFT	26.10	25.60	0.55	2.15%
MOT	8.3201	8.19	0.2101	2.59%
HBAN	4.11	4.15	-0.02	-0.48%
BAC	16.7589	16.71	0.1589	0.96%
ORCL	21.27	20.87	0.41	1.97%
ATPG	18.36	17.80	0.55	3.09%
QQQQ	42.5906	41.92	0.8906	2.14%
MRVL	16.381	15.77	0.661	4.20%
SBUX	20.63	19.94	0.80	4.03%
RIMM	66.73	69.20	-2.18	-3.16%
AA	13.30	13.23	0.22	1.68%
QLD	51.81	50.23	2.12	4.27%
HAL	27.24	26.84	0.50	1.87%
DELL	15.73	15.50	0.39	2.54%
FAVORITES				
BCE:CA	26.19	26.19	0.09	0.34%
MFC:CA	22.27	21.96	0.55	2.53%
K:CA	23.05	22.65	0.65	2.90%

Last	Change	Open	High	Low	Volume	Time
17.34	0.26 (1.52%	16.98	17.37	16.95	10,703,72	11:57:15 ED

Time EDT	MMID	Size	Bid	Ask	Size	MMID	Time EDT
11:57:28	NSDQ	10,000	17.34	17.35	17,000	NSDQ	11:57:28
11:57:28	NAS	9,800	17.34	17.35	16,800	NAS	11:57:28
11:57:28	PSE	7,600	17.34	17.35	6,100	ISE	11:57:28
11:57:28	BATS	4,600	17.34	17.35	4,400	BATS	11:57:28
11:57:28	ISE	2,800	17.34	17.35	4,400	PSE	11:57:28
11:57:21	BX	900	17.34	17.35	1,900	BX	11:57:21
11:57:14	CIN	200	17.34	17.35	300	ADF	11:56:05
11:57:22	CBO	100	17.34	17.35	300	FLOW	11:56:05
11:56:05	ADF	100	17.33	17.35	100	CBO	11:57:22
11:56:05	FLOW	100	17.33	17.36	700	UBSS	11:56:34
11:55:02	TMBR	500	17.30	17.40	100	LEHM	11:56:40
11:56:40	LEHM	100	17.30	17.45	100	GSCO	11:51:57
11:56:34	UBSS	300	17.28	17.48	100	SUSQ	11:44:07
11:52:22	COWN	100	17.25	17.49	100	PALC	09:25:55
11:56:00	HDSN	100	17.24	17.50	1,800	CSE	11:23:42
11:51:57	GSCO	100	17.20	17.55	100	COWN	11:52:22
11:44:07	SUSQ	100	17.09	17.60	200	PERT	10:23:52

Price	Size	Exch	Time EDT
17.34	100	NDD	11:57:15
17.34	300	NGS	11:57:15
17.34	400	NDD	11:57:14
17.34	200	NGS	11:57:06
17.34	100	NGS	11:57:06
17.34	100	NYE	11:57:05
17.34	300	NGS	11:57:05
17.34	100	NGS	11:57:05
17.34	500	NYE	11:57:04
17.34	200	NYE	11:57:04
17.34	100	NYE	11:57:04
17.34	100	NYE	11:57:04
17.34	100	NYE	11:57:04
17.34	387	NGS	11:57:03
17.34	400	NGS	11:57:03
17.34	100	NYE	11:57:03
17.34	100	NYE	11:57:03
17.34	500	BATS	11:57:03
17.34	100	BATS	11:57:03

GE

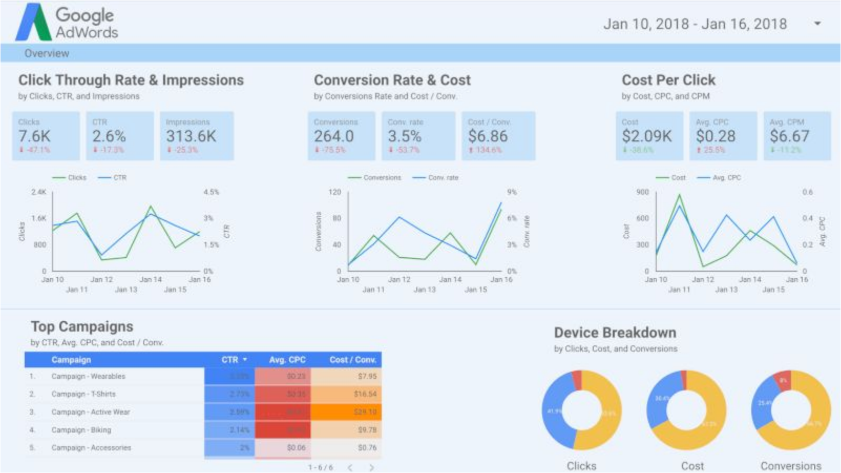
Last	16.82 ▲	Change	0.45 (2.75%)	Volume	37,151,946	LT Time	11:57:27 EDT	LT Vol	300
Bid	16.81 (21,200)	Ask	16.82 (20,700)	Yr High	26.05	Yr Low	5.7279	VWAP	16.7456
Pr Close	16.37	Open	16.47	High	16.91	Low	16.44	Sh OS	10.63 b
EPS	1.34	PE	12.20	Div/Yld	0.40 (2.44%)	Div Date	09/17/2009	Mrk Cap	178.53 b
General Electric				Exchange	NYE	Sector	Conglomerate	Industry	Conglomerate

Data manipulation in Google Cloud

- Data Studio
- Datalab
- Colab
- (offline!)

Data Studio

- Data Studio - glorified spreadsheets with a few integrations to Google Cloud to pull data
- Use cases: excel-like functions, simple visualizations (e.g. geographic)



Google AdWords

Jan 10, 2018 - Jan 16, 2018

Overview

Click Through Rate & Impressions

by Clicks, CTR, and Impressions

Clicks 7.6K -47.1%

CTR 2.6% -17.3%

Impressions 313.6K -25.3%

Conversion Rate & Cost

by Conversions Rate and Cost / Conv.

Conversions 264.0 -75.5%

Conv. rate 3.5% -53.7%

Cost / Conv. $6.86 134.6%

Cost Per Click

by Cost, CPC, and CPM

Cost $2.09K -38.6%

Avg. CPC $0.28 25.5%

Avg. CPM $6.67 -11.2%

Top Campaigns

by CTR, Avg. CPC, and Cost / Conv.

	Campaign	CTR ▾	Avg. CPC	Cost / Conv.
1.	Campaign - Wearables	[illegible]	$0.23	$7.95
2.	Campaign - T-Shirts	2.73%	$0.35	$16.54
3.	Campaign - Active Wear	2.59%	[illegible]	$29.10
4.	Campaign - Biking	2.14%	[illegible]	$9.78
5.	Campaign - Accessories	2%	$0.06	$0.76

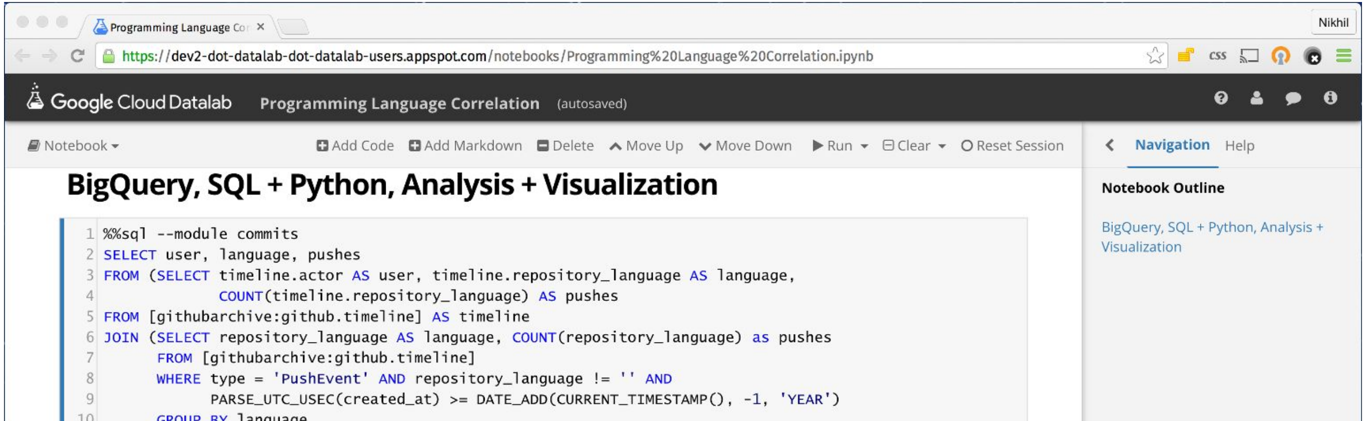
1 - 6 / 6

Device Breakdown

by Clicks, Cost, and Conversions

Datalab

- Datalab - hosted Jupyter instance with preset libraries
- Use cases: python scripting, visualization, ML pipelining, some long-running scripting, versioned scripts and models



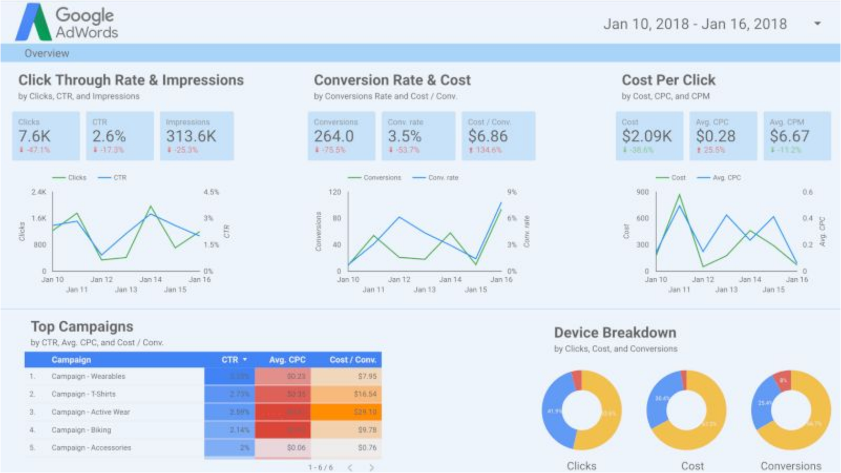
Programming Language Co ×
Nikhil
https://dev2-dot-datalab-dot-datalab-users.appspot.com/notebooks/Programming%20Language%20Correlation.ipynb
Google Cloud Datalab
Programming Language Correlation
(autosaved)
Notebook
Add Code
Add Markdown
Delete
Move Up
Move Down
Run
Clear
Reset Session
Navigation
Help
Notebook Outline
BigQuery, SQL + Python, Analysis + Visualization

BigQuery, SQL + Python, Analysis + Visualization

```
%%sql --module commits
SELECT user, language, pushes
FROM (SELECT timeline.actor AS user, timeline.repository_language AS language,
             COUNT(timeline.repository_language) AS pushes
FROM [githubarchive:github.timeline] AS timeline
JOIN (SELECT repository_language AS language, COUNT(repository_language) as pushes
      FROM [githubarchive:github.timeline]
      WHERE type = 'PushEvent' AND repository_language != '' AND
            PARSE_UTC_USEC(created_at) >= DATE_ADD(CURRENT_TIMESTAMP(), -1, 'YEAR')
      GROUP BY language
```

Colab

- Colab - Shared, no-setup version of Datalab that is designed around sharing
- Use cases: creating publicly accessible work, collaboration, but no long-running scripting



Clicks 7.6K -47.1%
CTR 2.6% -17.3%
Impressions 313.6K -25.3%

Conversions 264.0 -75.5%
Conv. rate 3.5% -53.7%
Cost / Conv. $6.86 134.6%

Cost $2.09K -38.6%
Avg. CPC $0.28 25.5%
Avg. CPM $6.67 -11.2%

Top Campaigns

by CTR, Avg. CPC, and Cost / Conv.

	Campaign	CTR ▾	Avg. CPC	Cost / Conv.
1.	Campaign - Wearables	3.05%	$0.23	$7.95
2.	Campaign - T-Shirts	2.73%	$0.35	$16.54
3.	Campaign - Active Wear	2.59%	[illegible]	$29.10
4.	Campaign - Biking	2.14%	[illegible]	$9.78
5.	Campaign - Accessories	2%	$0.06	$0.76

1 - 6 / 6