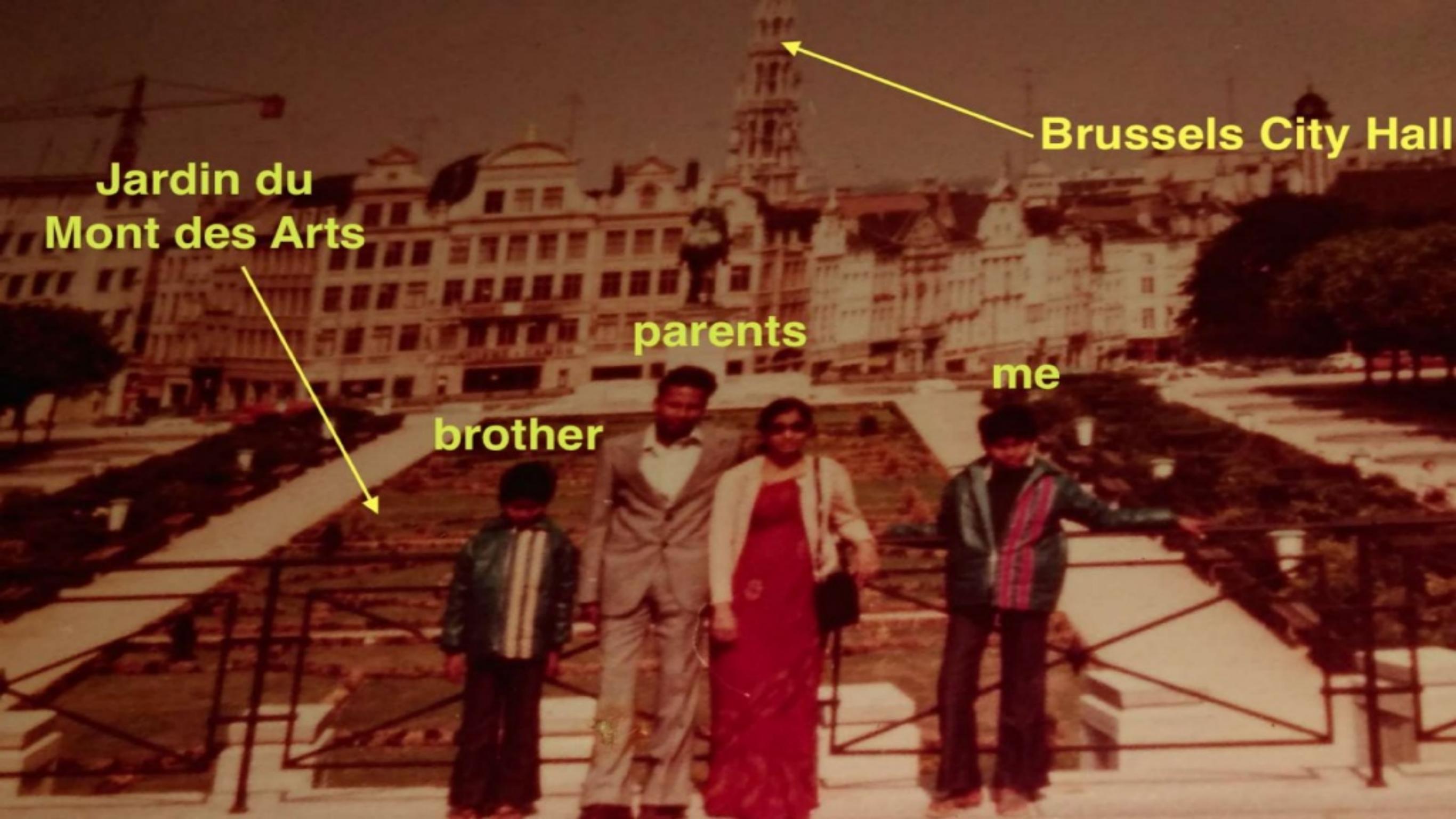


# extreme-scale ad-tech at MediaMath with Spark and Databricks

Prasad Chalasani, SVP Data Science, MediaMath

SPARK SUMMIT  
EUROPE 2016





Jardin du  
Mont des Arts

Brussels City Hall

parents

brother

me



digital display advertising  
is a \$25B/year business



a software platform for digital display advertising

we bid on the ad-exchanges  
on behalf of ~10K campaigns  
set up on our platform



40+ ad-exchanges send us  
200B ad-opportunities daily  
(2-3 M/s)



# our bidders respond within 10 ms with:

- optimal campaign
- optimal bid price

# powered by massive data

- 200 B daily ad-opportunities
- 1 B users (cookies)
- impressions, clicks, conversions, page-visits
- millions of sparse features
- terabytes of daily data





# spark

large-scale data applications with Spark @ MediaMath



# 1. machine learning for ad response prediction

what is  $\text{Prob}(\text{conversion} \mid \text{ad impression})$ ?



optimal bid price depends on  
 $\text{Prob}(\text{conversion} \mid \text{ad impression})$

- model for each campaign
- logistic regression
- stochastic gradient descent
- online learning (incremental)
- implemented in Scala + Spark

# Databricks visualization:

- many campaigns
- model variations
- calibration methods
- ML metrics (AUC, LogLoss, ...)

Model Results

	AUC				RIG				avErr				absAvErr			
	Prod	C	CPx	CPxImp	Prod	C	CPx	CPxImp	Prod	C	CPx	CPxImp	Prod	C	CPx	CPxImp
157904-	69	70	73	74	-2	4	5	6	94	6	3	4	292	195	190	190
199145-	54	58	65	66	-24	-4	0	0	274	71	47	47	469	266	233	233
171681-	64	71	74	76	0	5	6	7	20	-17	-16	-14	219	176	173	175
183010-	59	70	89	89	-11	2	18	18	101	36	4	5	300	228	138	137
170530-	63	76	79	80	0	8	10	10	-2	-1	-4	23	197	168	163	193
204737-	62	70	71	73	0	3	3	4	-4	-35	-33	-37	196	149	150	144
191286-	71	70	72	73	1	1	3	3	-55	-50	-44	-45	143	140	145	146
165354-	63	76	79	80	0	7	9	10	57	-13	-10	-8	256	163	165	166
197076-	61	79	90	92	-20	6	19	20	-90	-20	-7	-8	110	160	103	96
187201-	68	73	77	80	4	5	7	8	16	-33	-29	-27	215	161	159	160
219099-		56	61	62		0	0	0		7	23	27		200	210	214
193943-	73	83	85	86	6	8	10	11	-17	-34	-27	-27	182	152	146	148
129751-	62	73	81	84	-12	2	10	12	216	70	35	33	415	257	205	199
222838-	58	65	65	67	-39	1	-2	1	293	-4	3	-2	488	186	196	191
171935-	60	64	68	68	-27	1	4	3	291	26	40	55	490	219	210	224
203964-	61	80	81	81	-1	8	9	8	9	12	9	14	207	206	204	210
214899-	73	62	64	65	15	2	3	3	3	56	51	55	200	225	219	223
242996-	50	58	72	70	-2	-4	7	5	-10	-24	-16	-20	182	171	158	158
185849-	57	70	77	77	-1	3	6	6	19	2	9	11	219	194	183	188
199537-	60	84	83	83	1	12	11	12	11	-52	-51	-50	208	139	145	140
161586-	61	73	71	72	0	3	3	2	0	-5	-6	-16	202	183	181	173
244603-	54	63	61	60	0	-9	-8	-9	19	-78	-76	-77	19	116	116	123
243016-	59	57	61	59	-4	-6	-4	-16	-38	-60	-55	-59	163	139	135	136
123976-	49	75	70	70	-1	2	2	1	-35	42	30	40	152	233	221	234
239409-	57	69	69	69	-22	1	2	0	228	37	56	53	427	218	216	229
237023-	57	69	65	66	-16	0	0	0	286	68	45	41	508	270	247	245
221143-	76	79	74	79	-29	5	4	5	482	-1	1	10	681	195	157	206
212152-	50	83	66	79	-2	5	0	4	-63	99	51	51	143	306	258	258

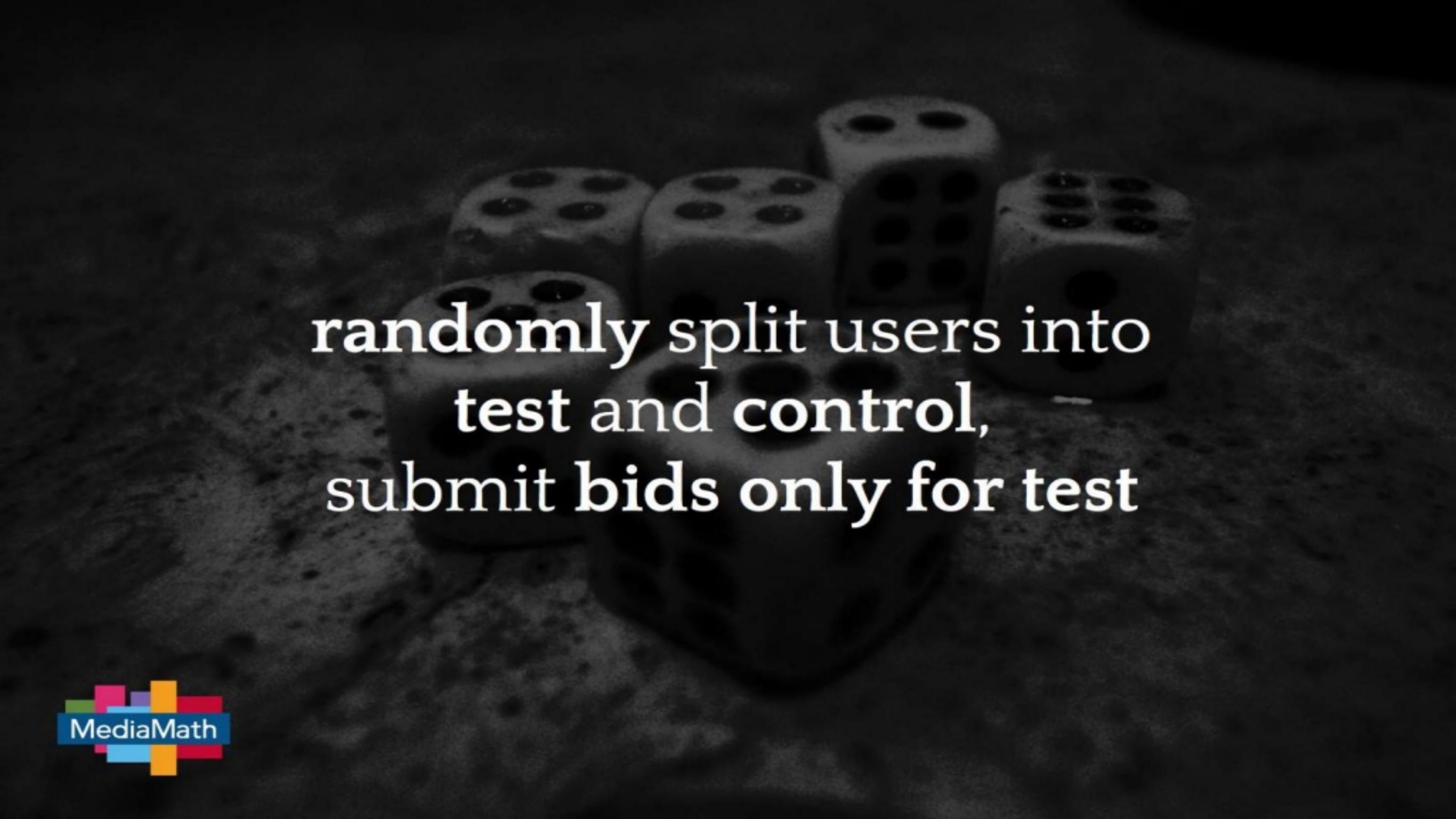
Campaigns (small --&gt; large)



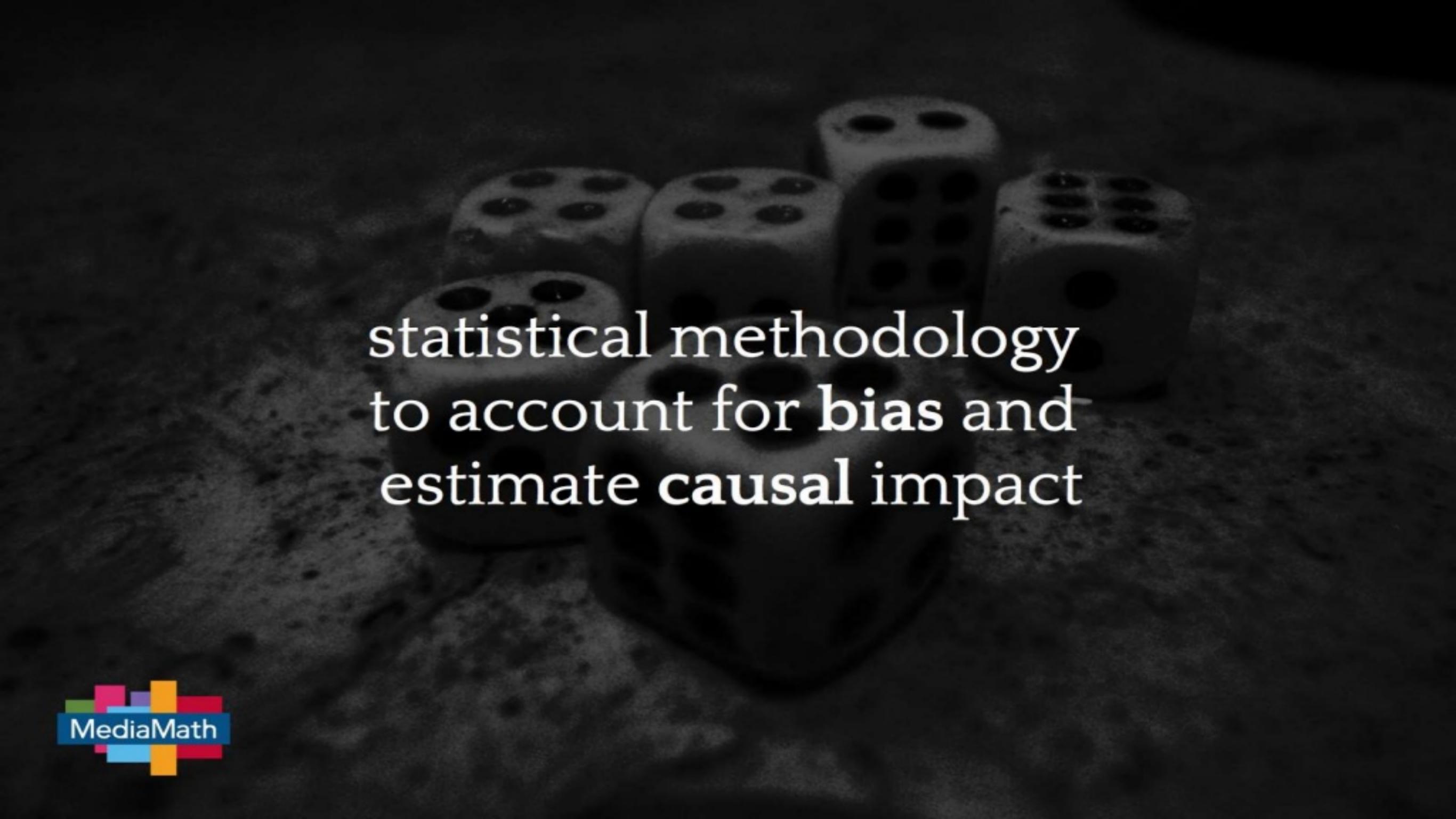
## 2. causal ad impact measurement



how much of response is driven by ad exposure?



**randomly split users into  
test and control,  
submit bids only for test**



statistical methodology  
to account for **bias** and  
estimate **causal** impact



# Gibbs Sampling for confidence intervals



runs daily in production  
using Scala + Spark

A wide-angle photograph of a large auditorium or theater. The seating consists of numerous rows of brown wooden chairs, arranged in a semi-circular pattern facing towards the left side of the frame. In the foreground, many people are seated, appearing as small figures due to the perspective. The upper levels of the auditorium are visible, showing more rows of empty chairs. The lighting is warm and focused on the seating area.

### 3. audience reach estimation

how many unique users satisfy targeting criteria?

# example targeting spec

- segment = sports-enthusiast
- country = Belgium
- dayOfWeek = Sunday
- deviceType = iPad
- channel = Video
- site = economist.com

# set intersection cardinality estimation

- source: incoming bid-ops
- 1M possible feature-values (sets)
- 1B users
- need answer in a minute
- cannot pre-compute all possible queries



offline:

- k-minHash sketch in Redis/AWS for each set
- Spark to update k-minHashes from bid-opps

# online (query-time):

- retrieve k-minHash sketches for sets in query
- compute intersection estimate fast



# other Spark applications at MediaMath

- attribution of response to ad-impressions
- predict win-rate as function of bid-price
- cross-device user identity



# Databricks @ MediaMath



indispensable for productive data science



Home



Workspace



Recent



Tables



Clusters

## Find nodes where production p percentile is < B

```
> val exch = 4  
display(getStats("leafP2B", 4).filter("rolP2B < 100 and pct = 80").orderBy("leaf"))
```

▶ (3) Spark Jobs

leaf	par	path	wins	losses
6610	6605	(ROOT_NODE=-998)(CHANNEL_TYPE=4)(WEEKPART=0)(SITE=0) (COUNTRY=60033)(CHANNEL=-998)	1412851	9361309
5264	5262	(ROOT_NODE=-998)(CHANNEL_TYPE=1)(WEEKPART=1)(DAY=7)(SITE=0) (CHANNEL=0)(CREATIVE_SIZE=10486360)(COUNTRY=60141)	89760	353957

quick ad-hoc analyses in scala, python, R, sql



Workspace	Users	pchalasani+db@gm...	brain
</> Application Examples	arsalan@databricks.c...	alloc	bigml
? Databricks Guide	cdizon@mediamath....	bidstats	brain-tests
Training	db-guest@mediamat...	bof	braineval
Shared	dwissinger@mediam...	brain	braineval-cmv
Users	ewinston@mediamat...	cid	braineval-cmv-analyze
GSW Passing Analys...	hkushary@mediamat...	Copy of Classificatio...	braineval-cmv-analyz...
	jmarshall@mediamat...	CreateHiveTables	braineval-cmv-multi
	jpollack@mediamath....	Decision Trees in Spark	braineval-creative-viz
	jstouli@mediamath.c...	deep	braineval-multi
	jthiagarajan@mediam...	Demo	cla

hundreds of notebooks:  
elegant, responsive, multi-language



Search Tables

Home

Workspace

Recent

Tables

activeplacebotests

activeplacebotests\_pqt

activeplacebotests\_str

amnet\_segment\_perf...

amnet\_segment\_perf...

amnet\_segment\_perf...

amnet\_segment\_perf...

atp\_cookieless

awsrdih

bidder\_logs\_raw

Refresh

Schema:

col_name	data_type
batch_id	bigint
auction_id	bigint
ts	timestamp
advertiser_id	int
creative_id	int
strategy_id	int

# hundreds of tables



 databricks

## Create Cluster

### New Cluster

Cancel **Create Cluster**

8 Workers, 240 GB Memory, 32 Cores, 8 DBU  and 1 Driver, 30 GB Memory, 4 Cores, 1 DBU 

Cluster Name  
MySpark2.0\_Cluster

Apache Spark Version   
**Spark 2.0.0 (Scala 2.11)**

Instance

Type  
On-Demand and Spot   Fall back to On-Demand 

On Demand Driver (30 GB Memory, 4 Cores, 1 DBU) 

---

Workers (240 GB Memory, 32 Cores, 8 DBU) 

On Demand      Spot

0  8   Enable Autoscaling (beta) 

easily spin-up clusters

run-jar-main-class (Scala)

Attached: pc-bof-10 ▾ File ▾ View: Code ▾ Permissions Run All Clear Results

```
> import com.mediamath.braineval.MyClass  
  
val args = Array(  
    "input", "/mnt/dspc/impressions",  
    "output", "/mnt/dspc/results",  
    "numDays", "3"  
)  
  
MyClass.main(args)
```

trivially run jars,  
no need for shell scripts





Clusters



Jobs



Search

run-daily-placebo

Prasad  
Chalasani

Notebook at [/Users/pchalasani+db@gmail.com/placebo/Gen-daily-placebo](http://Users/pchalasani+db@gmail.com/placebo/Gen-daily-placebo)

# run notebooks as jobs

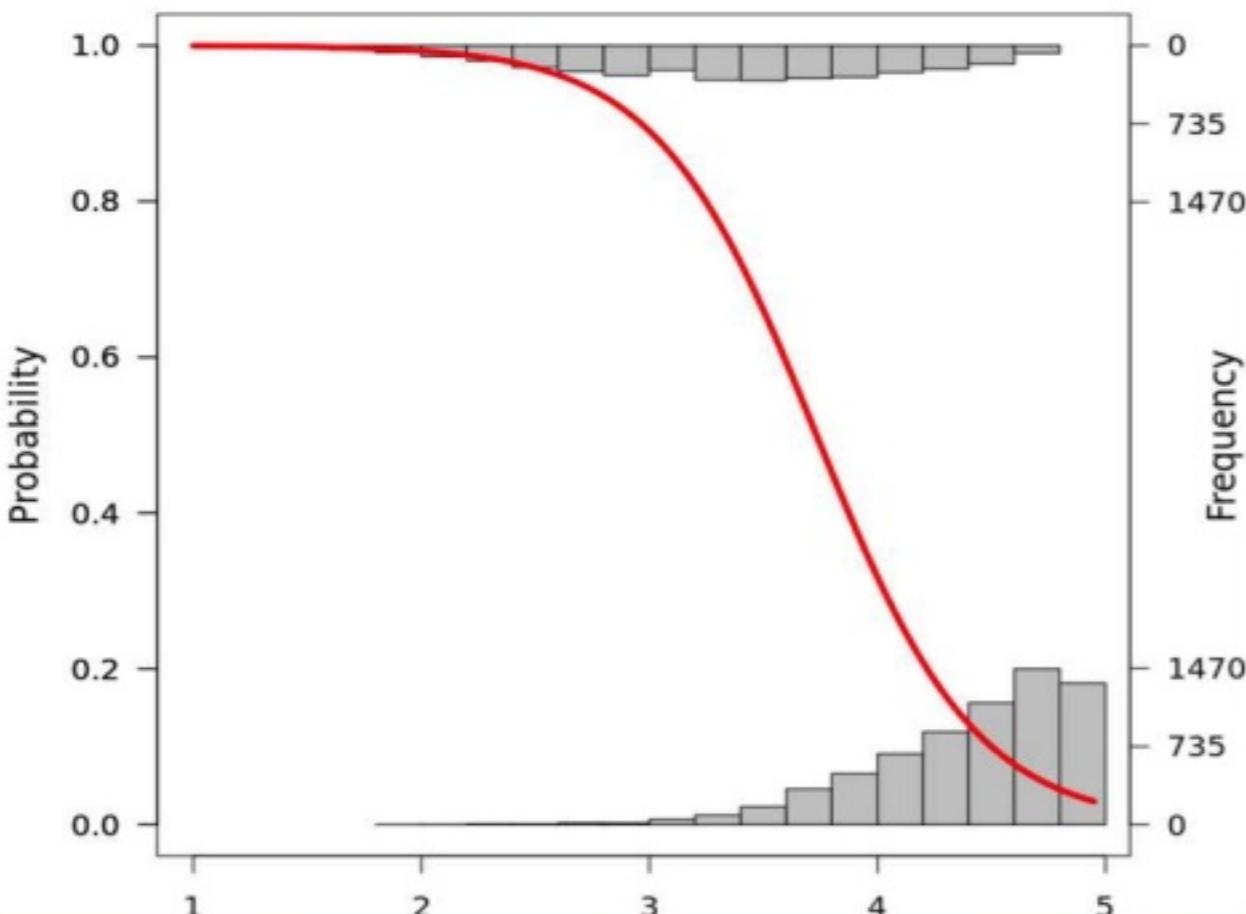


id	type	lm	incr	conf	lb	lbi	rtrcConf	ATT	gConf	glm5	glm50	glm95	rt	rc	rtw	w
796497	all	35	26	100	16	14	100	8	100	27	35	44	6	5	30	11
834256	all	411	80	100	17	15	100	3	100	238	408	935	3	3	4	13
857408	all	30	23	66	10	9	99	0	99	7	30	63	1	1	2	24
888637	all	26	20	100	20	17	100	517	100	25	26	26	2289	1911	2533	73
1041560	all	37	27	100	28	22	100	107	100	10	35	73	378	296	396	76
1056221	all	175	64	100	40	29	100	166	100	166	175	185	207	147	261	36

create, share dashboards



## P(Acceptable Sketch Estimate | Jaccard, k=5000)



## Model Results

Campaigns (small => large)	AUC				RIG				avErr				absAvErr			
	Prod	C	CPx	CPxImp	Prod	C	CPx	CPxImp	Prod	C	CPx	CPxImp	Prod	C	CPx	CPxImp
157904-	69	70	73	74	-2	4	5	6	94	6	3	4	292	195	190	190
199145-	54	58	65	66	-24	-4	0	0	274	71	47	47	469	266	233	233
171681-	64	71	74	76	0	5	6	7	20	-17	-16	-14	219	176	173	175
183010-	59	70	89	89	-11	2	18	18	101	36	4	5	300	228	138	137
170530-	63	76	79	80	0	8	10	10	-2	-1	-4	23	197	168	163	193
204737-	62	70	71	73	0	3	3	4	-4	-35	-33	-37	196	149	150	144
191286-	71	70	72	73	1	1	3	3	-55	-50	-44	-45	143	140	145	146
165354-	63	76	79	80	0	7	9	10	57	-13	-10	-8	256	163	165	166
197076-	61	79	90	92	-20	6	19	20	-90	-20	-7	-8	110	160	103	96
187201-	68	73	77	80	4	5	7	8	16	-33	-29	-27	215	161	159	160
219099-		56	61	62		0	0	0	7	23	27		200	210	214	
193943-	73	83	85	86	6	8	10	11	-17	-34	-27	-27	182	152	146	148
129751-	62	73	81	84	-12	2	10	12	216	70	35	33	415	257	205	199
222838-	58	65	65	67	-39	1	-2	1	293	-4	3	-2	488	186	196	191
171935-	60	64	68	68	-27	1	4	3	291	26	40	55	490	219	210	224
203964-	61	80	81	81	-1	8	9	8	9	12	9	14	207	206	204	210
214899-	73	62	64	65	15	2	3	3	3	56	51	55	200	225	219	223
242996-	50	58	72	70	-2	-4	7	5	-10	-24	-16	-20	182	171	158	158
185849-	57	70	77	77	-1	3	6	6	19	2	9	11	219	194	183	188
199537-	60	84	83	83	1	12	11	12	11	-52	-51	-50	208	139	145	140
161586-	61	73	71	72	0	3	3	2	0	-5	-6	-16	202	183	181	173
244603-	54	63	61	60	0	-9	-8	-9	19	-78	-76	-77	19	116	116	123
243016-	59	57	61	59	-4	-6	-4	-16	-38	-60	-55	-59	163	139	135	136
123976-	49	75	70	70	-1	2	2	1	-35	42	30	40	152	233	221	234
239409-	57	69	69	69	-22	1	2	0	228	37	56	53	427	218	216	229
237023-	57	69	65	66	-16	0	0	0	286	68	45	41	508	270	247	245
221143-	76	79	74	79	-29	5	4	5	462	-1	1	10	681	195	157	206
212152-	50	83	66	79	-2	5	0	4	-63	98	51	51	143	306	258	258

# data visualizations

The screenshot shows a Databricks interface. On the left, there's a sidebar with various icons and links: Home, Workspace, Recent, Tables, Clusters, and Jobs. The 'Training' link is currently selected. The main content area has a header 'Import Notebook' and a sidebar with 'Module 2: Word Count Lab Solutions', 'Module 3: Apache Log Lab Solutions', 'Module 4: Text Analysis and Entity Resolution Lab Solutions', and 'Module 5: Machine Learning Lab Solutions'. Below this, there's a section titled 'Scalable Machine Learning (CS190-1x) (new)' with links to 'Introduction (README)', 'Module 1: Lecture', 'Module 1: Math Review Lab', and 'Module 2: Lecture'. The main content area starts with the heading 'Part 3: Predictions for Yourself'. It contains text about the goal of predicting movie recommendations and adding ratings for yourself. It also includes a sub-section '(3a) Your Movie Ratings' and a code snippet for printing the most rated movies from a RDD.

## Part 3: Predictions for Yourself

The ultimate goal of this lab exercise is to predict what movies to recommend. To help you provide ratings for yourself, we have included the following code to print the most rated movies from `movieLimitedAndSortedByRatingRDD` which we created in part 1.

**(3a) Your Movie Ratings**

```
> print 'Most rated movies:'  
      print '(average rating, movie name, number of reviews)'  
      for ratingsTuple in movieLimitedAndSortedByRatingRDD.take(50):  
          print ratingsTuple
```

data science onboarding



# Data Science Institute

ABOUT

CENTERS

ACADEMICS

RESEARCH

ENTREPRENEURSHIP

INDUSTRY

## ACADEMICS

 **Master of Science in Data Science**

Certification of Professional Achievement in Data Sciences

Graduate Curriculum

Online Courses (ColumbiaX)

Frequently Asked Questions

## Master of Science in Data Science

The Master of Science in Data Science allows students to apply data science techniques to their field of interest, building on four foundational courses offered in our [Certification of Professional Achievement in Data Sciences program](#). Our students have the opportunity to conduct original research, included in a [capstone project](#), and interact with our industry partners and faculty. Students may also choose an elective track focused on entrepreneurship or a subject area covered by



# thank you!

[pchallasani@mediamath.com](mailto:pchallasani@mediamath.com)

