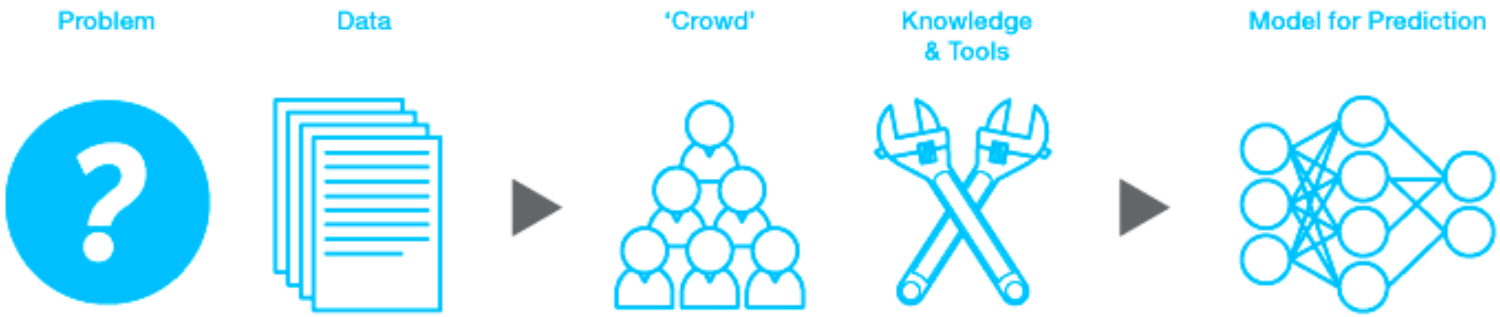
Dataiku

Play with kaggle

Matthieu Scordia

Kaggle?

Kaggle is a platform for predictive modeling competitions.



"We're making data science into a sport."

Let's enter a challenge!

Dataiku

Yandex

Personalized Web Search Challenge

31 days to go

Friday, October 11, 2013 — $9,000 • 99 teams — Friday, January 10, 2014

Dashboard

- Home
  - Data
  - Make a submission
- Information
  - Description
  - Evaluation
  - Rules
  - Prizes
  - Logs format
  - Organizers

Competition Details » Get the Data » Make a submission

Re-rank web documents using personal preferences

The Personalized Web Search Challenge provides a unique opportunity to consolidate and scrutinize the work from industrial labs on personalizing web search using user-logged search behavior context. It provides a fully anonymized dataset shared by Yandex, which has anonymized user ids, queries, query terms, urls, url domains and clicks.

The Data

Dataiku

Noteworthy characteristics of the dataset:

- **Unique queries:** 21,073,569
- **Unique urls:** 703,484,26
- **Unique users:** 5,736,333
- **Training sessions:** 34,573,630
- **Test sessions:** 797,867
- **Clicks in the training data:** 64,693,054

Total records in the log: 167,413,039 (=15Go!)

Let's shake the data

col_0 Number	col_1 Text	col_2 Text	col_3 Number	col_4 Number	col_5 Text	col_6 Number	col_7 Number	col_8 Number	col_9 Number
0	M	4	0						
0	0	Q	0	10047345	3080290,4098689	50504886,4217515	9848058,1084315	50534229,4217515	505916
0	108	C	0	50628761					
0	1080	C	0	50628761					
1	M	4	0						
1	0	Q	0	2057953	1093007	12695453,1284095	20124473,2056277	60660113,4693531	434896
2	M	27	0						
2	0	Q	0	2113437	1148783	33204613,3248226	2053036,303607	5878776,770558	346608
3	M	13	1						
3	0	Q	0	5239394	2365113,2856206,2491775	16457319,1712204	35513272,3344594	50566741,4217515	869993
3	76	Q	1	5239405	2365113,2856206,2491775,4219157,2383044	16457319,1712204	35513272,3344594	19587095,2024884	546383
3	162	C	1	35513272					
4	M	15	1						
4	0	Q	0	17143989	4219157,1841361,853923,1720163,1912374,1755325,4454730	65548702,4975721	197782,39086	54375043,4396765	315896
4	27	C	0	31589696					
4	86	C	0	6876504					
4	121	C	0	3374640					

Logs format.

col_0 Number	col_1 Text	col_2 Text	col_3 Number	col_4 Number	col_5 Text	col_6 Number	col_7 Number	col_8 Number	col_9 Number
0	M	4	0						
0	0	Q	0	10047345	3080290,4098689	50504886,4217515	9848058,1084315	50534229,4217515	505916
0	108	C	0	50628761					
0	1080	C	0	50628761					
1	M	4	0						
1	0	Q	0	2057953	1093007	12695453,1284095	20124473,2056277	60660113,4693531	434896
2	M	27	0						
2	0	Q	0	2113437	1148783	33204613,3248226	2053036,303607	5878776,770558	346608
3	M	13	1						
3	0	Q	0	5239394	2365113,2856206,2491775	16457319,1712204	35513272,3344594	50566741,4217515	869993
3	76	Q	1	5239405	2365113,2856206,2491775,4219157,2383044	16457319,1712204	35513272,3344594	19587095,2024884	546383
3	162	C	1	35513272					
4	M	15	1						
4	0	Q	0	17143989	4219157,1841361,853923,1720163,1912374,1755325,4454730	65548702,4975721	197782,39086	54375043,4396765	315896
4	27	C	0	31589696					
4	86	C	0	6876504					
4	121	C	0	3374640					

Session 0

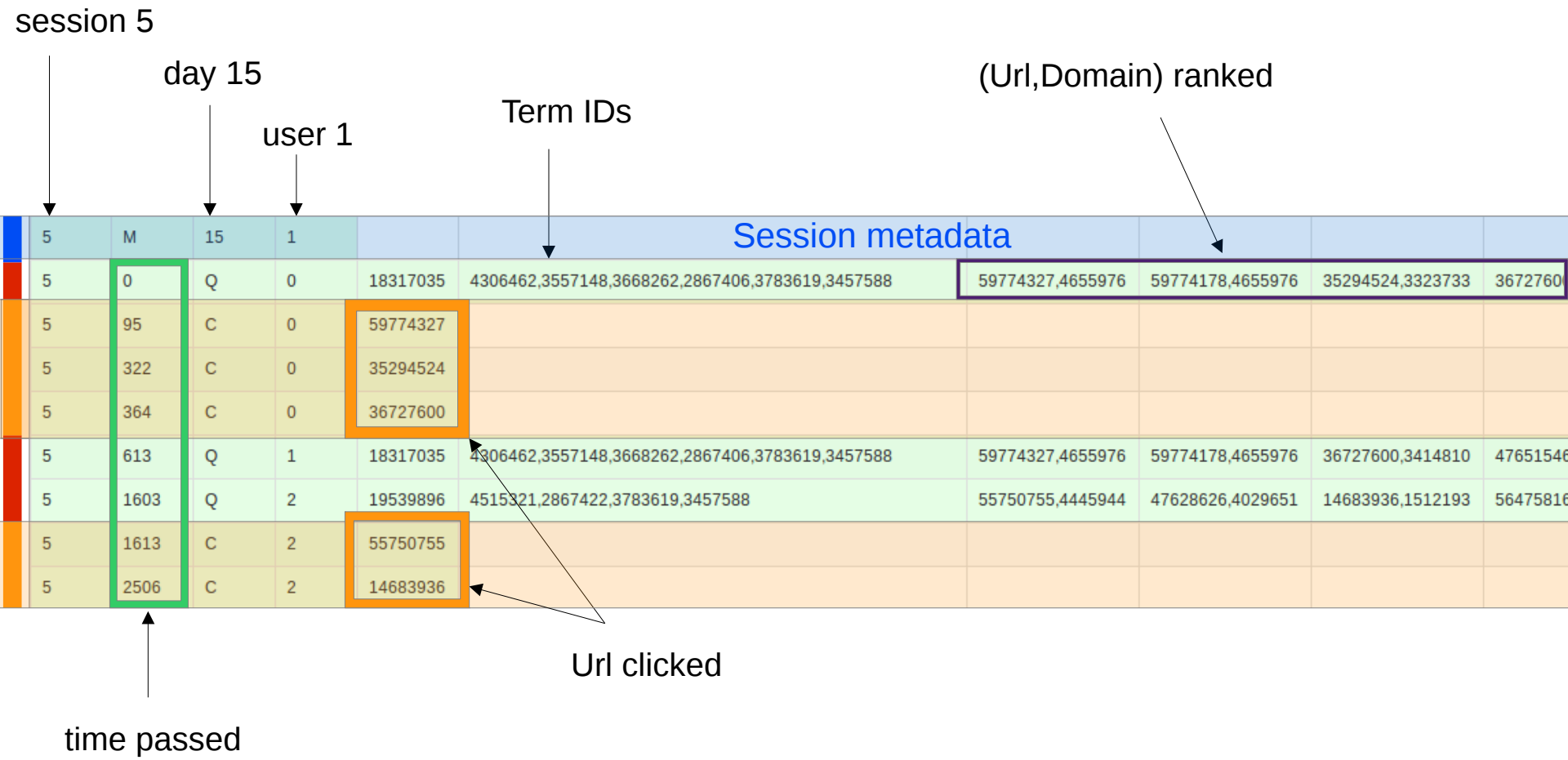
Session 1

Session 2

Session 3

Session 4

Sessions format.



5	M	15	1						
5	0	Q	0	18317035	4306462,3557148,3668262,2867406,3783619,3457588	59774327,4655976	59774178,4655976	35294524,3323733	3672760
5	95	C	0	59774327					
5	322	C	0	35294524					
5	364	C	0	36727600					
5	613	Q	1	18317035	4306462,3557148,3668262,2867406,3783619,3457588	59774327,4655976	59774178,4655976	36727600,3414810	4765154
5	1603	Q	2	19539896	4515321,2867422,3783619,3457588	55750755,4445944	47628626,4029651	14683936,1512193	5647581
5	1613	C	2	55750755					
5	2506	C	2	14683936					

Evaluation.

Dataiku

The URLs are labeled using 3 grades of relevance: {0, 1, 2}

The labeling is done automatically, based on dwell-time.

0 : irrelevant - no clicks and clicks with dwell time < 50 time units.

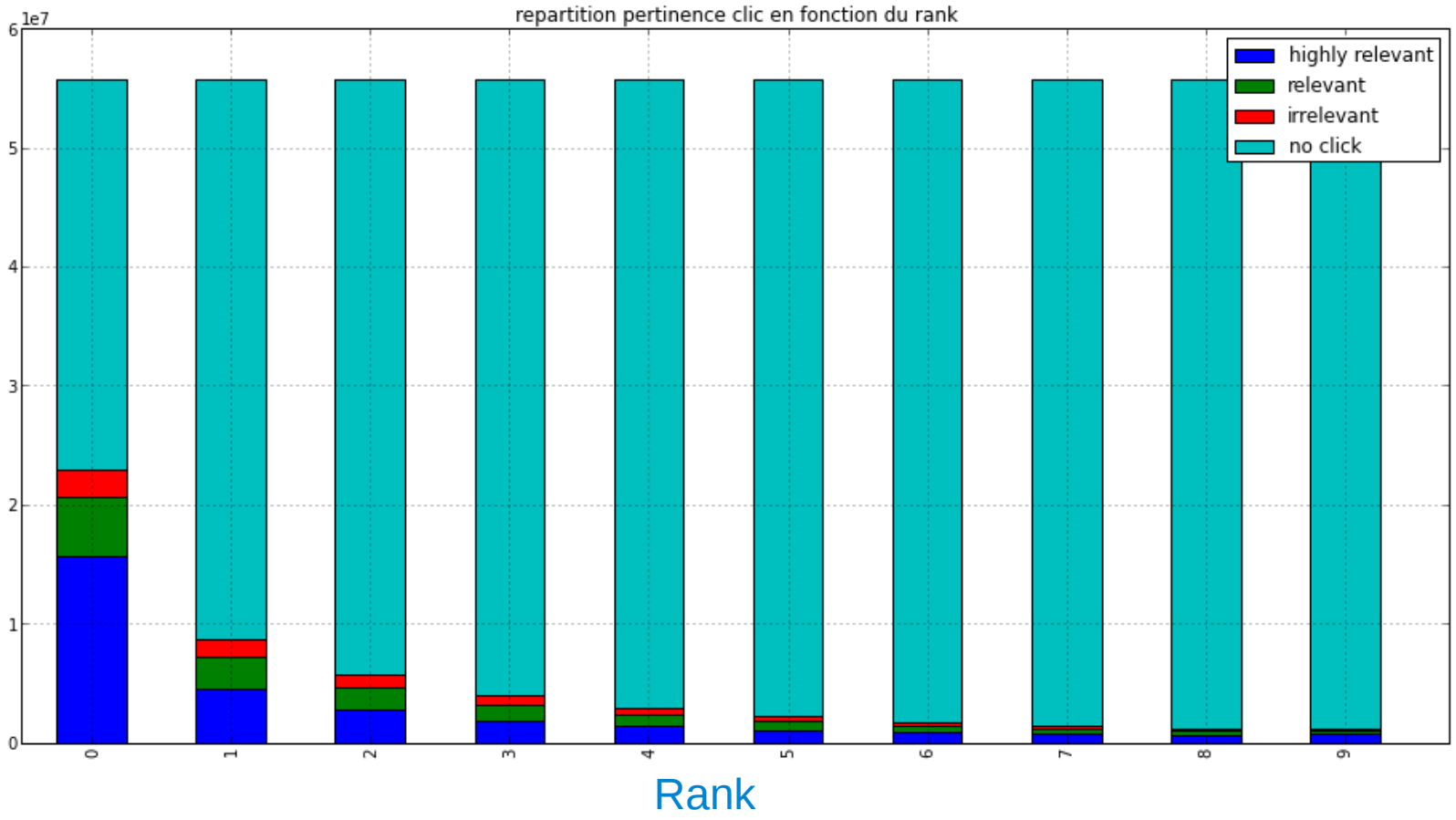
1 : relevant - clicks with dwell time > 50 and < 399 time units.

2 : highly relevant - clicks with dwell time > 400 time units.

How to beat Yandex?

Dataiku

Clicks count



So we have to sort better than that!

Step 1 : Reshape it!

Dataiku

user Number	day Number	url Number	domain Number	rank Number	pertinence Number
0	4	50504886	4217515	0	
0	4	9848058	1084315	1	
0	4	50534229	4217515	2	
0	4	50591618	4217515	3	
0	4	26242582	2597528	4	
0	4	34623075	3279130	5	
0	4	68893581	5149883	6	
0	4	50628761	4217517	7	2
0	4	32262001	3142702	8	
0	4	35443881	3339757	9	

For each user we would estimate his click probability on an url

Step 2 : Cross validate

Dataiku

Split your dataset like Yandex did:

- **On the last 3 days.**
- **Only one session by user.**

Goal: auto-evaluate our model.

Step 3 : Add new features

Dataiku

We add some informations on each user:

- Did he see this url in the past?
- Did he click on it?
- How many times?
- Did he skip it?
- Had he ever click on a rank 9 url in the past?

Step 3 : Add new features

Dataiku

The thing is, we don't want to re-rank all...

So we add click entropy:

For each query: $$H = -\sum p(x) \log p(x)$$

Where p(x) is the percentage of clicks on document x among all clicks.

Example:

Small click entropy query: yahoo, youtube.

Large click entropy query: photos, jobs.

Step 4 : the model.

Goal: Predict the probability of click of an user on an url.

Our training set:

session	url	features...					target

We use logistic regression and random forest.

The leaderboard.

Dataiku

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	-	pampampampam (ooc) *	0.80278	19	Sun, 08 Dec 2013 20:22:28 (-1.2h)
2	-	learner *	0.80158	20	Wed, 04 Dec 2013 16:18:50
3	-	E. G. Ortiz-García *	0.80044	33	Fri, 22 Nov 2013 19:38:12
4	-	AIMORE	0.79885	62	Mon, 02 Dec 2013 23:32:30 (-6d)
5	-	Martin Martin	0.79877	7	Tue, 03 Dec 2013 01:48:59 (-2.1d)
6	-	blue_maple	0.79778	51	Tue, 10 Dec 2013 08:18:19 (-7.8d)
7	↑4	bart	0.79743	34	Tue, 10 Dec 2013 05:37:00 (-4.8d)
8	↑1	**Dataiku Science Studio**	**0.79719**	**20**	**Thu, 05 Dec 2013 08:27:03 (-9.9h)**
9	↑63	camcamcam	0.79717	16	Tue, 10 Dec 2013 06:09:35
10	↓3	DuckTile	0.79683	7	Tue, 19 Nov 2013 11:44:47 (-4.3d)

	Default Ranking Baseline	0.79056
	Random Baseline	0.47972

Thanks !

Dataiku

If you want to enter with us in a future challenge:

contact@dataiku.com