

Message

**From:** Eric Lehman [Redacted@google.com]  
**Sent:** 6/6/2016 9:59:40 PM  
**To:** Maryam Tavafi [Redacted@google.com]; Mahsan Rofouei [Redacted@google.com]  
**CC:** Sundeep Tirumalareddy [Redacted@google.com]  
**Subject:** Re: Question about the open position in your team

# Redacted

/Eric

Redacted

REDACTED FOR PUBLIC FILING & ABRIDGED

Ex. No.  
UPX0192

1:20-cv-03010-APM

GOOG-DOJ-19327173

On Mon, Jun 6, 2016 at 1:47 PM, Maryam Tavafi <Redacted@[google.com](#)> wrote:

Redacted

Thanks,  
Maryam

On Mon, Jun 6, 2016 at 1:32 PM, Eric Lehman <Redacted@[google.com](#)> wrote:  
Hi Maryam,

Redacted

On Mon, Jun 6, 2016 at 1:09 PM, Maryam Tavafi <Redacted@[google.com](#)> wrote:  
Hi Eric,

Redacted

Thanks,  
Maryam



Unifying click prediction systems. Bubbled up.

Scope is large, so high-level.

Lots of details. Unresolved and hard.

Won't dweel, please bring up.

Insights. Tangential, so pointers.

## Clicks in Ranking

- Reliance on user feedback (“clicks”) in ranking has steadily increased over the past decade, despite anxieties around grandfathering, discoverability, etc.
- Showing results that users want to click is **NOT** the ultimate goal of web ranking. This would:
  - Promote low-quality, click-bait results.
  - Promote results with genuine appeal that are not relevant.
  - Be too forgiving of optionalization.
  - Demote official pages, promote porn, etc.

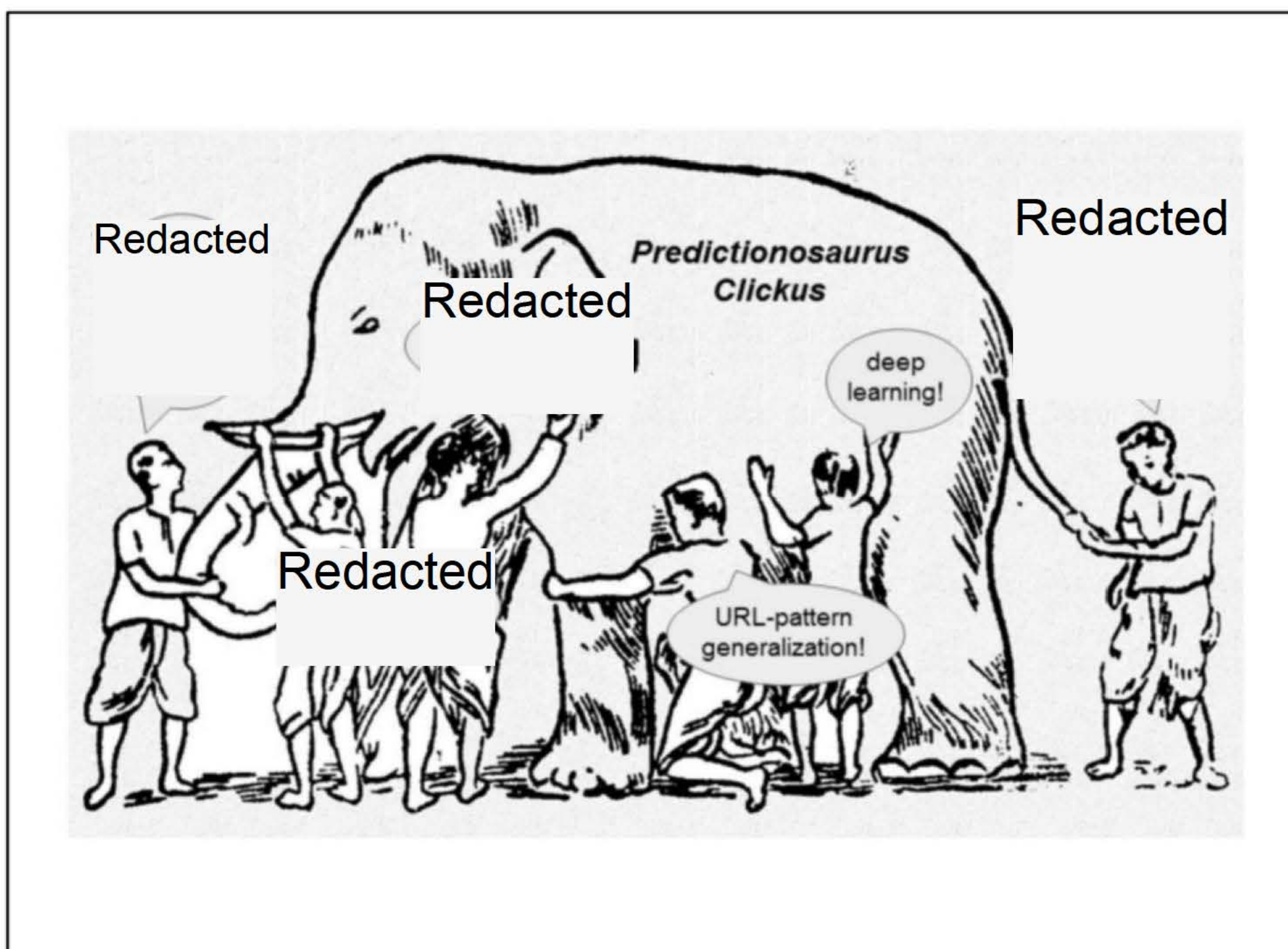
Step back, long view.

Like cheating: Memorizing instead of understanding the material.

## Clicks as a Proxy Objective

- But showing results that users want to click is **CLOSE** to our goal.
- And we can do this “almost right” thing extremely well by drawing upon trillions of examples of user behavior in search logs.
- This suggests a strategy for improving search quality:
  - Predict what results users will click.
  - Boost those results.
  - Patch up problems with page quality, relevance, optionalization, etc.
- Not a radical idea. We have done this for years.

Wrong thing to do, but really, really good at it.



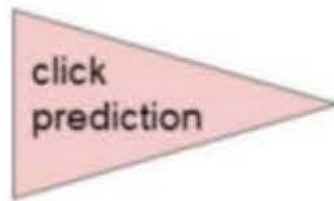
So I think we've maybe slipped into this situation.

There is this one unified, critical concept: figuring out what users would click, given a fair opportunity.

Many teams have come at that idea with radically different approaches.

Maybe we got caught up in the details of each approach and did not properly recognize the elephant in the room.

## Life Inside the Red Triangle



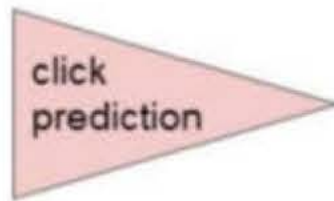
click  
prediction

- The “inner loop” for people working on click prediction becomes tuning on user feedback data. Human evaluation is used in system-level testing.
- We get about 1,000,000,000 new examples of user behavior every day, permitting high-precision evaluation, even in smaller locales. The test is:

***Were your click predictions better or worse than the baseline?***

- This is a fully-quantifiable objective, unlike the larger problem of optimizing search quality. The need to balance multiple metrics and intangibles is largely pushed downstream.

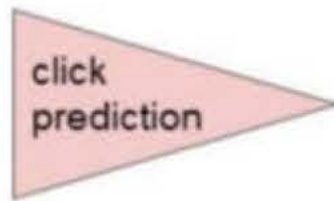
## Life Inside the Red Triangle



click  
prediction

- The evaluation methodology is “train on the past, predict the future”. This largely eliminates problems with over-fitting to training data.
- Continuous evaluation is on fresh queries and the live index. So the importance of freshness is built into the metric.
- The importance of localization and further personalization are also built into the metric, for better or worse.

## Life Inside the Red Triangle



click  
prediction

- This refactoring creates a monstrous and fascinating optimization problem: use hundreds of billions of examples of past user behavior (and other signals), to predict future behavior involving a huge range of topics.
- The problem seems too large for any existing machine learning system to swallow. We will likely need some combination of manual work, RankLab tuning, and large-scale machine learning to achieve peak performance.
- In effect, the metric quantifies our ability to emulate a human searcher. One can hardly avoid reflections on the Turing Test and Searle's Chinese Room.
- Moving from thousands of training examples to billions is game-changing...

### **~1,000,000 IS ratings**

are more than sufficient to superbly tune curves via RankLab and human judgment.

But this gives only a low-resolution picture of how people interact with search results.

Useful behavioral patterns may appear in few training instances and thus can not be learned from IS ratings.



**~100,000,000,000 clicks**

provide a vastly clearer picture of how people interact with search results.

A behavior pattern apparent in just a few IS ratings may be reflected in hundreds of thousands of clicks, allowing us to learn second and third order effects.



## Example

- Click data indicates that documents whose title contains **dvm** are currently under-ranked for queries that start with **[dr ...]**.
  - dvm = Doctor of Veterinary Medicine
  - There are a couple relevant examples in the 15K set.
  - There are about a million examples in click data.
- So the volume of click data for this special situation roughly equals the *total* volume of all human rating data.
- Learning this association is not only possible from the training data, but required to minimize the objective function.

Of Possible Interest...

# Redacted