

NDCG IS OVERRATED



Berlin Search Meetup
October, 2023

Rethinking offline search evaluation

Obligatory Bio Slide

👋 Hi I'm Doug
(@softwaredoug everywhere)

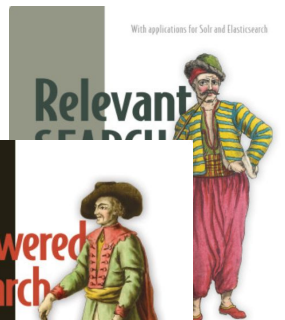
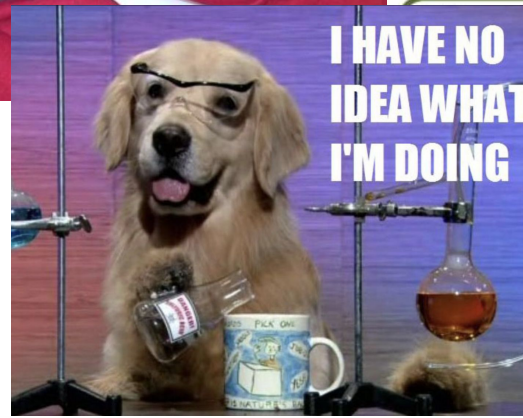
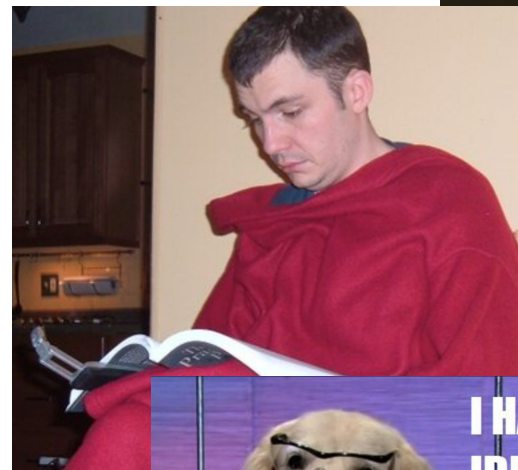
Long-time search enthusiast... Not
yet (never?) an expert

I wrote some search books, did some open
source

I work at Reddit

I worked at Shopify & OpenSource Connections
in search

I blog here: <http://softwaredoug.com>








Learning to Rank GO

Outline






- Judgments & NDCG
- Obvious problems with NDCG (and pals)
- Not so obvious problems w/ the judgment model
- What we're missing in offline evaluation
- A better way: Treatment fidelity in search relevance
- Conclusions... Science is Hard

JUDGMENT REGIME OF EVALUATION

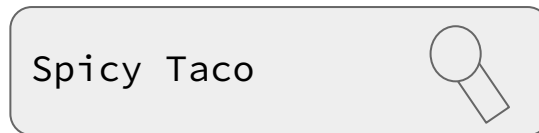
Judgments




Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2

(n)DCG what is it?






Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2

Our latest and greatest algorithm returns:








1. 
2. 
3. 

(n)DCG what is it?

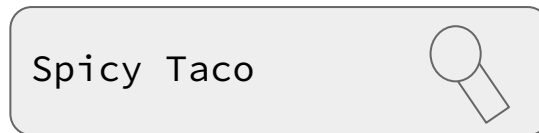
Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2






(n)DCG what is it?






Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2

Our latest and greatest algorithm returns:



1. 
2. 
3. 

Higher positions more important...

Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2

Spicy Taco



1.
2.
3.



Pos'n Discount
(1/posn)*






$$1 / 1 = 1$$

$$1 / 2 = 0.5$$

$$1 / 3 = 0.333$$

*Not actual discount used, just simpler math for illustration

Label each result with a grade...

Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2

Spicy Taco 






- 1.
- 2.
- 3.



Pos'n Discount (1/posn)*	Result Grade
$1 / 1 = 1$	0.4
$1 / 2 = 0.5$	0.9
$1 / 3 = 0.333$	0.2

*Not actual discount used, just simpler math for illustration

Multiply each row...

Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2

Spicy Taco








- 1.
- 2.
- 3.



Pos'n Discount (1/posn)	Result Grade	Posn Discounted Gr
$1 / 1 = 1$	0.4	$1 * 0.4 = 0.4$
$1 / 2 = 0.5$	0.9	$0.5 * 0.9 = 0.45$
$1 / 3 = 0.333$	0.2	$0.2 * 0.333 = 0.066$

Sum for DCG

Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2

Spicy Taco








1.
2.
3.



Pos'n Discount (1/posn)	Result Grade	Posn Discounted Gr
$1 / 1 = 1$	0.4	$1 * 0.4 = 0.4$
$1 / 2 = 0.5$	0.9	$0.5 * 0.9 = 0.45$
$1 / 3 = 0.333$	0.2	$0.2 * 0.333 = 0.066$

$$\begin{aligned} \text{DCG@3} &= 0.4 + 0.45 + 0.066 \\ &= 0.916 \end{aligned}$$

Compute ideal for this query

Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2

Spicy Taco



1.



2.



3.



Pos'n Discount (1/posn)	Result Grade	Posn Discounted Gr
$1 / 1 = 1$	0.9	$1 * 0.4 = 0.9$
$1 / 2 = 0.5$	0.4	$0.5 * 0.9 = 0.2$
$1 / 3 = 0.333$	0.2	$0.2 * 0.333 = 0.066$

$$\begin{aligned} \text{idCG@3} &= 0.9 + 0.45 + 0.066 \\ &= 1.416 \end{aligned}$$

$NDCG@3 = DCG@3 / IDC@3$

Spicy Taco



Query	Document	Grade (0-1)
Spicy Taco		0.9
Spicy Taco		0.4
Spicy Taco		0.2
Hamburger		0.9
Hamburger		0.2

1.



2.



3.






Pos'n Discount (1/posn)	Result Grade	Posn Discounted Gr
$1 / 1 = 1$	0.4	$1 * 0.4 = 0.4$
$1 / 2 = 0.5$	0.9	$0.5 * 0.9 = 0.45$
$1 / 3 = 0.333$	0.2	$0.2 * 0.333 = 0.066$

$$\begin{aligned} DCG@3 &= 0.4 + 0.45 + 0.066 \\ &= 0.916 \end{aligned}$$

$$\begin{aligned} NDCG@3 &= 0.916 / 1.416 \\ &= \mathbf{0.64} \end{aligned}$$

THE PROBLEMS


What if our ideal is terrible?

Query	Document	Grade (0-1)
Spicy Taco		0.4
Spicy Taco		0.4
Spicy Taco		0.2

- 1.
- 2.
- 3.






This ranking gives an **NDCG@3=1**

Spicy Taco 

Pos'n Discount (1/posn)	Result Grade	Posn Discounted Gr
$1 / 1 = 1$	0.4	$1 * 0.4 = 0.4$
$1 / 2 = 0.5$	0.4	$0.5 * 0.4 = 0.2$
$1 / 3 = 0.333$	0.2	$0.2 * 0.333 = 0.066$


$$\begin{aligned} \text{idCG@3} &= 0.4 + 0.2 + 0.066 \\ &= 0.666 \end{aligned}$$

Just stick to DCG?

Query	Document	Grade (0-1)
Spicy Taco		0.4
Spicy Taco		0.4
Spicy Taco		0.2

- 1.
- 2.
- 3.

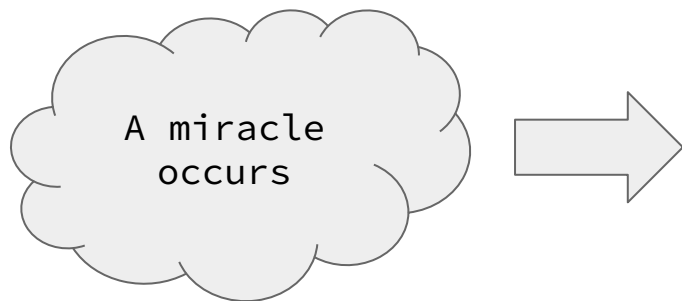





Spicy Taco 		
Pos'n Discount (1/posn)	Result Grade	Posn Discounted Gr
$1 / 1 = 1$	0.4	$1 * 0.4 = 0.4$
$1 / 2 = 0.5$	0.4	$0.5 * 0.4 = 0.2$
$1 / 3 = 0.333$	0.2	$0.2 * 0.333 = 0.066$

$$\begin{aligned} \text{DCG@3} &= 0.4 + 0.2 + 0.066 \\ &= \underline{\underline{0.666}} \end{aligned}$$

(Ok this seems 'lower' than other results)

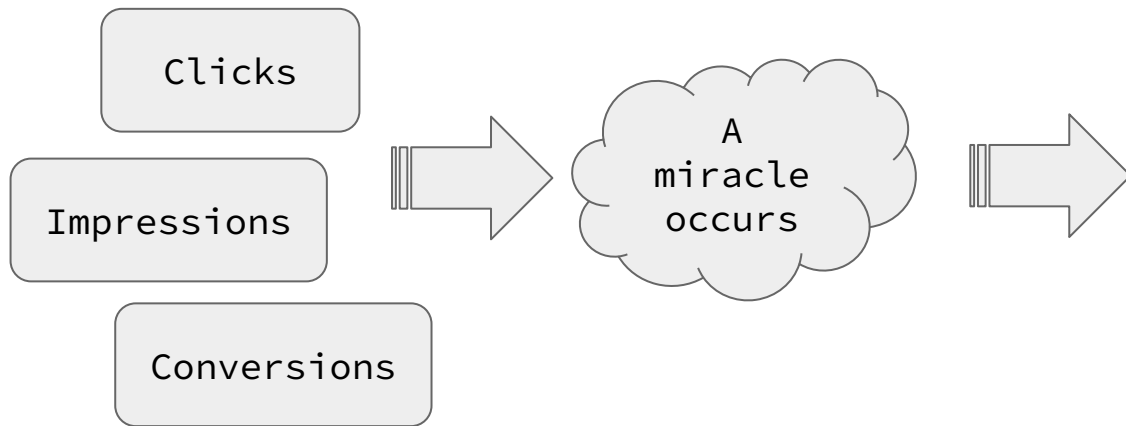
Grade quality






Query	Document	Grade (0-1)
Spicy Taco		0.4
Spicy Taco		0.4
Spicy Taco		0.2

Where do these
come from?

Focus on engagement based



Query	Document	Grade (0-1)
Spicy Taco		0.4
Spicy Taco		0.4
Spicy Taco		0.2

Biases galore...

Google

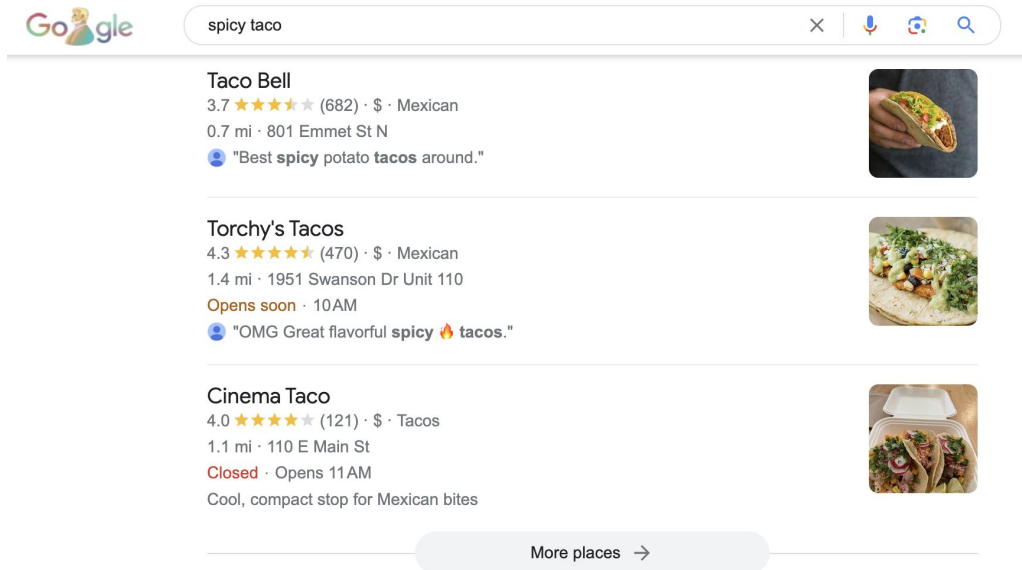
spicy taco

Taco Bell
3.7 ★★★★★ (682) · \$ · Mexican
0.7 mi · 801 Emmet St N
🗣️ "Best **spicy** potato **tacos** around."

Torchy's Tacos
4.3 ★★★★★ (470) · \$ · Mexican
1.4 mi · 1951 Swanson Dr Unit 110
Opens soon · 10AM
🗣️ "OMG Great flavorful **spicy** 🔥 **tacos**."

Cinema Taco
4.0 ★★★★★ (121) · \$ · Tacos
1.1 mi · 110 E Main St
Closed · Opens 11AM
Cool, compact stop for Mexican bites

More places →



Where to the clicks go?

Position Bias

Google search for "spicy taco".

- Taco Bell**
3.7 ★★★★★ (682) · \$ · Mexican
0.7 mi · 801 Emmet St N
"Best spicy potato **tacos** around."
- Torchy's Tacos**
4.3 ★★★★★ (470) · \$ · Mexican
1.4 mi · 1951 Swanson Dr Unit 110
Opens soon · 10AM
"OMG Great flavorful **spicy** 🌶️ **tacos**."
- Cinema Taco**
4.0 ★★★★★ (121) · \$ · Tacos
1.1 mi · 110 E Main St
Closed · Opens 11AM
Cool, compact stop for Mexican bites

More places →

- Chili Pepper Madness**
<https://www.chilipeppermadness.com> · Recipes ⋮
Bold and Spicy Taco Recipes from ...
Here you'll find my collection of homemade **taco** recipes that focus on big and bold flavors, many of them nice and **spicy**.

More eyes,
More clicks,
Yet... terrible tacos

Position Bias at weird places

The screenshot shows a Google search for "spicy taco". The results are as follows:

- Taco Bell**: 3.7 stars (682 reviews), Mexican, 0.7 mi away. Description: "Best spicy potato tacos around." Eye-tracking: 5 eyes.
- Torchy's Tacos**: 4.3 stars (470 reviews), Mexican, 1.4 mi away. Description: "OMG Great flavorful spicy tacos." Eye-tracking: 2 eyes.
- Cinema Taco**: 4.0 stars (121 reviews), Tacos, 1.1 mi away. Description: "Cool, compact stop for Mexican bites." Eye-tracking: 0 eyes.
- Chili Pepper Madness**: A recipe link with 4.0 stars. Description: "Here you'll find my collection of homemade taco recipes that focus on big and bold flavors, many of them nice and spicy." Eye-tracking: 4 eyes.

A "More places" button is visible between the restaurant and recipe results.

A seam in the UX, more eyes here?

Attractiveness Bias

Google

spicy taco

Taco Bell
3.7 ★★★★★ (682) · \$ · Mexican
0.7 mi · 801 Emmet St N
"Best spicy potato tacos around."

Torchy's Tacos
4.3 ★★★★★ (470) · \$ · Mexican
1.4 mi · 1951 Swanson Dr Unit 110
Opens soon · 10AM
"OMG Great flavorful spicy 🌶️ tacos."

Cinema Taco
4.0 ★★★★★ (121) · \$ · Tacos
1.1 mi · 110 E Main St
Closed · Opens 11AM
Cool, compact stop for Mexican bites

More places →

Chili Pepper Madness
<https://www.chilipeppermadness.com> Recipes

Bold and Spicy Taco Recipes from ...
Here you'll find my collection of homemade **taco** recipes that focus on big and bold flavors, many of them nice and **spicy**.



AN ACTUAL TACO!!
Nomnomnom
clickclickclick



WTF is this!?



Confidence bias



spicy taco

Taco Bell

3.7 ★★★★★ (682) · \$ · Mexican

0.7 mi · 801 Emmet St N

"Best spicy potato tacos around."



Torchy's Tacos

4.3 ★★★★★ (470) · \$ · Mexican

1.4 mi · 1951 Swanson Dr Unit 110

Opens soon · 10 AM

"OMG Great flavorful spicy tacos."



Cinema Taco

4.0 ★★★★★ (121) · \$ · Tacos

1.1 mi · 110 E Main St

Closed · Opens 11 AM

Cool, compact stop for Mexican bites



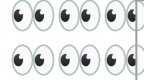
More places →

Chili Pepper Madness

<https://www.chilipeppermadness.com> · Recipes

Bold and Spicy Taco Recipes from ...

Here you'll find my collection of homemade taco recipes that focus on big and bold flavors, many of them nice and spicy.



Views	Clicks	Conversions
1000	135	32
10	2	1

Taco Bell – CTR 0.135

Torchy's – CTR 0.2

How much do we trust these stats given amount of data?

Presentation / Survivorship bias



spicy taco

Taco Bell

3.7 ★★★★★ (682) · \$ · Mexican

0.7 mi · 801 Emmet St N

"Best spicy potato tacos around."



Torchy's Tacos

4.3 ★★★★★ (470) · \$ · Mexican

1.4 mi · 1951 Swanson Dr Unit 110

Opens soon · 10 AM

"OMG Great flavorful spicy 🌶️ tacos."



Cinema Taco

4.0 ★★★★★ (121) · \$ · Tacos

1.1 mi · 110 E Main St

Closed · Opens 11 AM

Cool, compact stop for Mexican bites



Views	Clicks	Conversions
1000	135	32
10	2	1

Taco Bell – CTR 0.135

Torchy's – CTR 0.2

Brazos – 0 / 0 -> undefined

Meanwhile on page 5... **the good tacos!**:



Brazos Tacos

<https://store.brazostacos.com>

[Brazos Tacos - Charlottesville](#)

Select a location from **Brazos** and order onli direct!

Views	Clicks	Conversions
0	0	0

(n)DCG just shuffles the deckchairs



spicy taco

Taco Bell

3.7 ★★★★★ (682) · \$ · Mexican

0.7 mi · 801 Emmet St N

🗣️ "Best spicy potato tacos around."



Torchy's Tacos

4.3 ★★★★★ (470) · \$ · Mexican

1.4 mi · 1951 Swanson Dr Unit 110

Opens soon · 10AM

🗣️ "OMG Great flavorful spicy 🌶️ tacos."



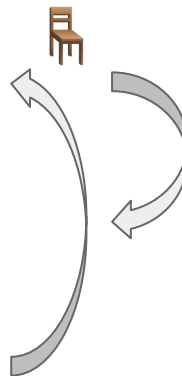
Cinema Taco

4.0 ★★★★★ (121) · \$ · Tacos

1.1 mi · 110 E Main St

Closed · Opens 11AM

Cool, compact stop for Mexican bites



We're restricted to answering questions within some top N labeled results

N is very small

NumDocs is very very large

(n)DCG struggles with recall



spicy taco



Taco Bell

3.7 ★★★★★ (682) · \$ · Mexican

0.7 mi · 801 Emmet St N

🗣️ "Best spicy potato tacos around."



Torchy's Tacos

4.3 ★★★★★ (470) · \$ · Mexican

1.4 mi · 1951 Swanson Dr Unit 110

Opens soon · 10 AM

🗣️ "OMG Great flavorful spicy 🌶️ tacos."



Cinema Taco

4.0 ★★★★★ (121) · \$ · Tacos

1.1 mi · 110 E Main St

Closed · Opens 11 AM

Cool, compact stop for Mexican bites



Brazos Tacos

<https://store.brazostacos.com>

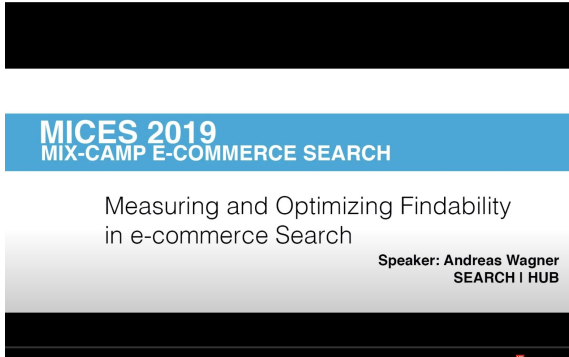
Brazos Tacos - Charlottesville

Select a location from **Brazos** and order onli direct!



A BIGGER PROBLEM

The judgment based model is broken?



MICES 2019
Andreas Wagner

<https://www.youtube.com/watch?v=xgHf9k272nc>



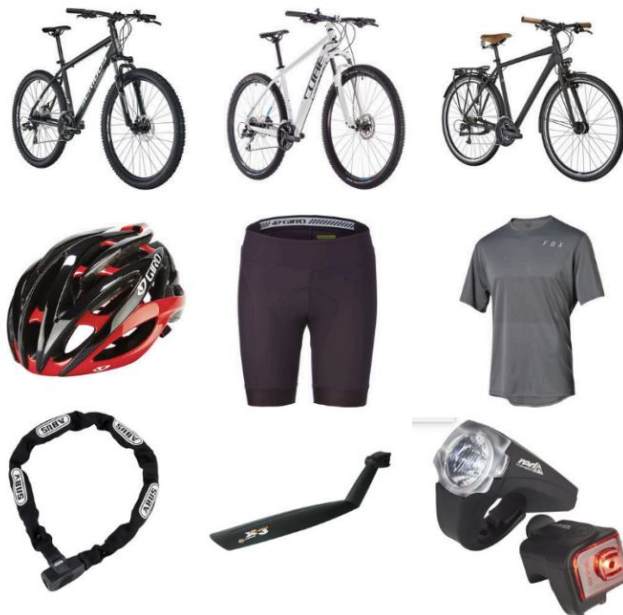
Haystack 2019
Tara Diedrichsen, Tito Sierra

<https://www.youtube.com/watch?v=7PjBSH6Wqhc>

Query = bicycle



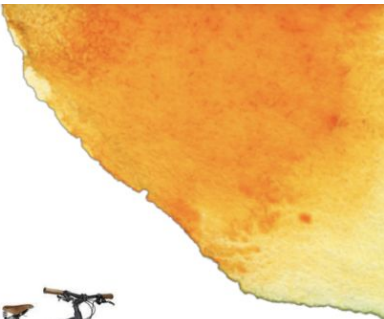
Expert Rating - 5



Expert Rating - 2

+21% Clicks

+17% GMV



Different Approaches to Search Results Evaluation

Results list: assess overall relevance of top ranking results list

Results for: tesla | Actions

News (10,000+)

Group/Duplicates: High Similarity | What's hot?

Sort by: Relevance

1. Buy Tesla, or sell Tesla? Barris analysis names it a Fresh Pick; calls negatively 'noisy'
News: Steps | Sep 11, 2010 | 1010 words | Steve Hardy
2. Sell Tesla and buy GM - seriously: Tesla often fails to deliver on its promises, and there's no momentum in the stock
MarketWatch (1/5) | Sep 15, 2010 | NEWS & COMMENTARY: Jeff Ravens' Strength in Numbers | 1340 words | Jeff Ravens, MarketWatch
3. Latest Tesla news: Researchers uncover security flaws in Tesla Model S keyless fobs
V3.co.uk | Sep 12, 2010 | 2140 words | Graeme Butler
4. The Tesla of China Might Have the Same Problems as Tesla
Barris's Drive | Sep 14, 2010 | ONLY | 380 words | By Graeme Auld

(View list of similar documents (1))

5. Tesla Owner is Driving to 40 States and all 100 Tesla Stores on Quest to Prove Teslas are the Best Road Trip Cars in the World
Automotive World/Hotline | Sep 10, 2010 | 875 words

Document level: assess relevance of individual top ranking documents

Results for: tesla | Actions

News (10,000+)

Group/Duplicates: High Similarity | What's hot?

Sort by: Relevance

1. Buy Tesla, or sell Tesla? Barris analysis names it a Fresh Pick; calls negatively 'noisy'
News: Steps | Sep 11, 2010 | 1010 words | Steve Hardy
2. Sell Tesla and buy GM - seriously: Tesla often fails to deliver on its promises, and there's no momentum in the stock
MarketWatch (1/5) | Sep 15, 2010 | NEWS & COMMENTARY: Jeff Ravens' Strength in Numbers | 1340 words | Jeff Ravens, MarketWatch
3. Latest Tesla news: Researchers uncover security flaws in Tesla Model S keyless fobs
V3.co.uk | Sep 12, 2010 | 2140 words | Graeme Butler
4. The Tesla of China Might Have the Same Problems as Tesla
Barris's Drive | Sep 14, 2010 | ONLY | 380 words | By Graeme Auld

(View list of similar documents (1))

5. Tesla Owner is Driving to 40 States and all 100 Tesla Stores on Quest to Prove Teslas are the Best Road Trip Cars in the World
Automotive World/Hotline | Sep 10, 2010 | 875 words

Why is (n)DCG overrated?

- NDCG itself is somewhat flawed, forcing us to 0-1 scale
- The underlying labels will always have biases, some very hard to overcome
- The judgment model itself assumes singular query-doc model matters most

Overrated != Useless... or even “bad”

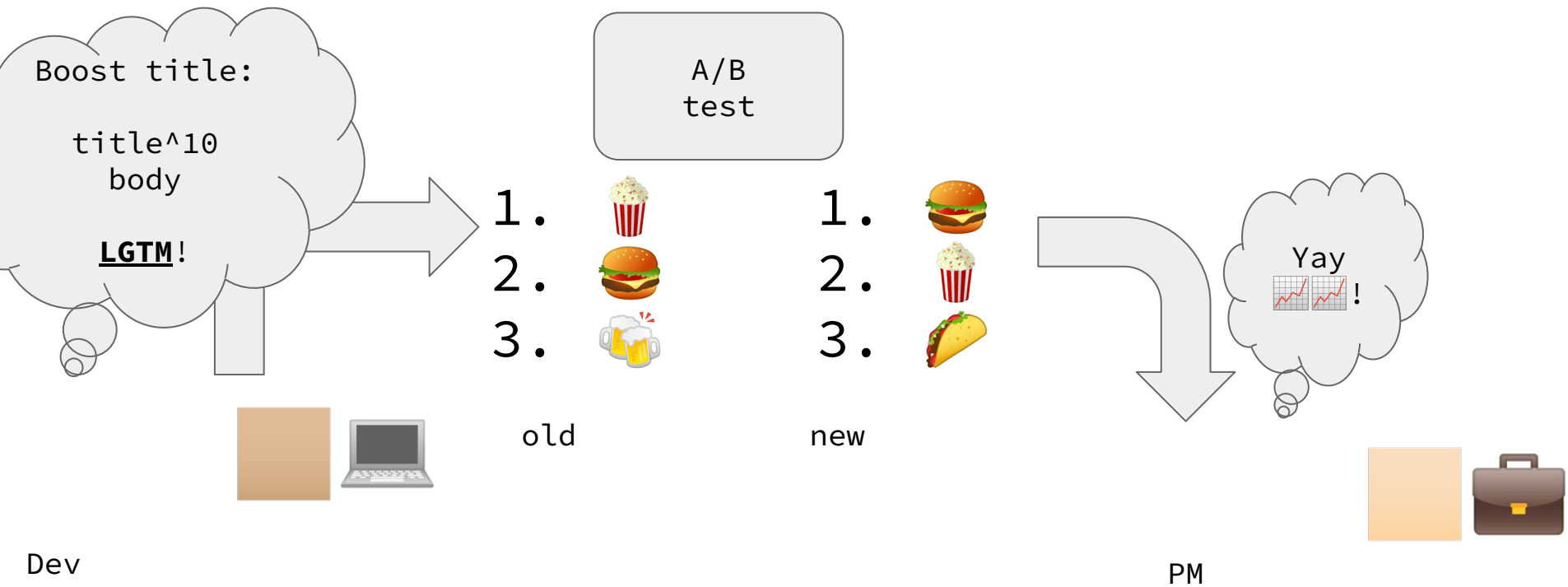
- NDCG is just one metric, amongst many to examine
- Learning to Rank depend on us having good NDCG

There is no “one true metric”

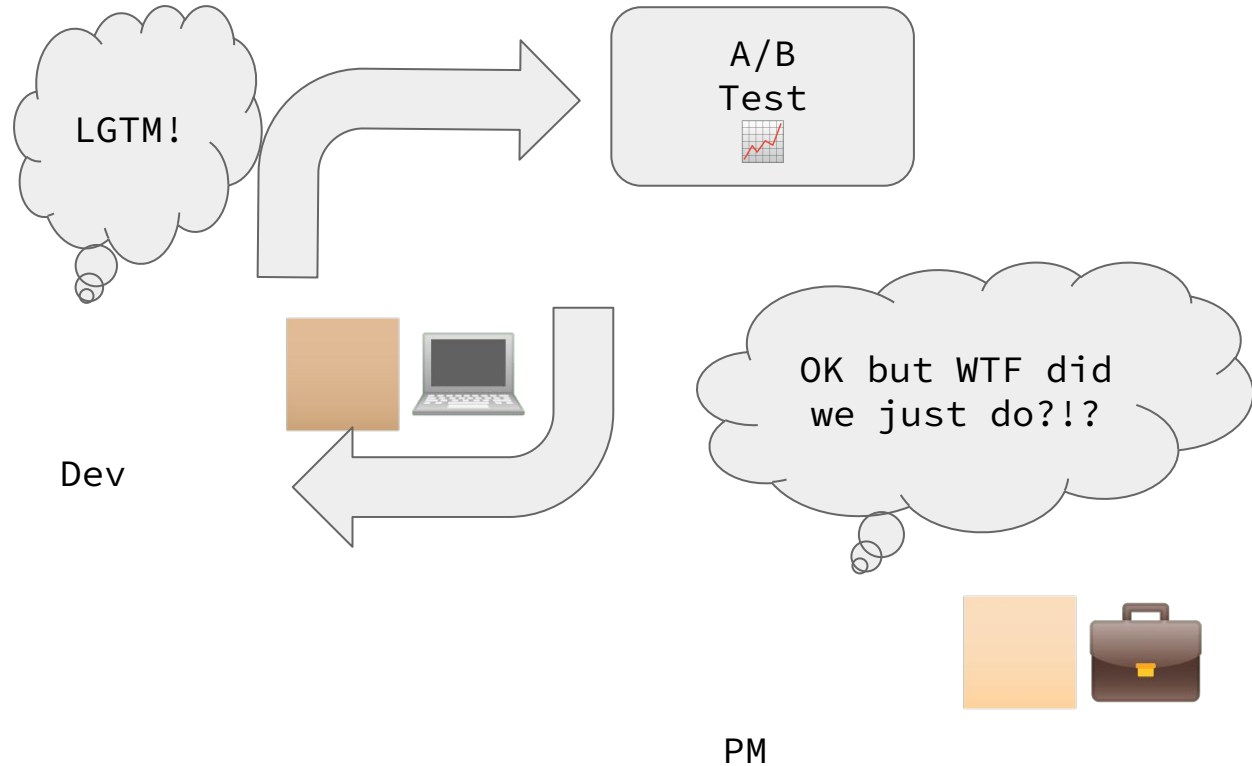
- Know what your metric actually measures
 - Improvement on human labels
 - Improvement on known CTR results, etc
- We can make decisions with many metrics

WHAT WE MISS IN
OFFLINE

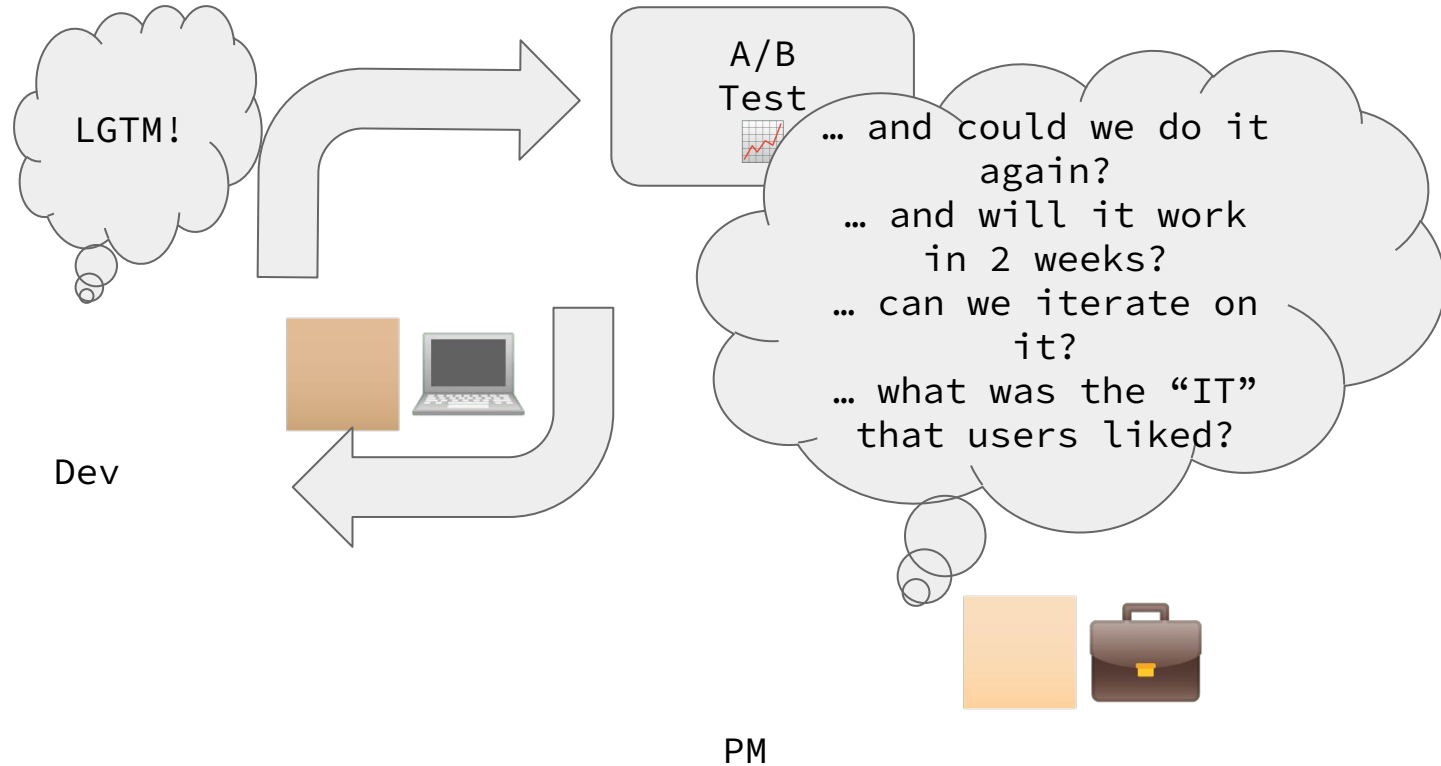
Why not YOLO ship to A/B?



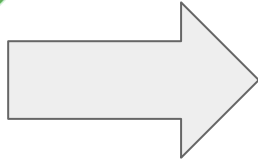
... because we don't gain (much) knowledge



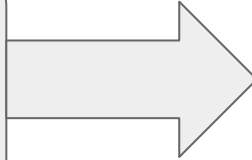
... because we don't gain (much) knowledge



Think about medical testing



Hypothesis:
Compound X
improves
outcomes



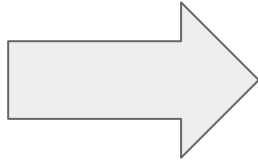
Am I actually
producing the right
compound?

- In search we often skip this

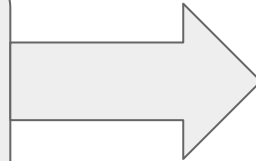
What is the impact of
the compound on
patients?

- And ship it straight to users

... not so good



Random mix
of chemicals
one guy did
in his
personal lab

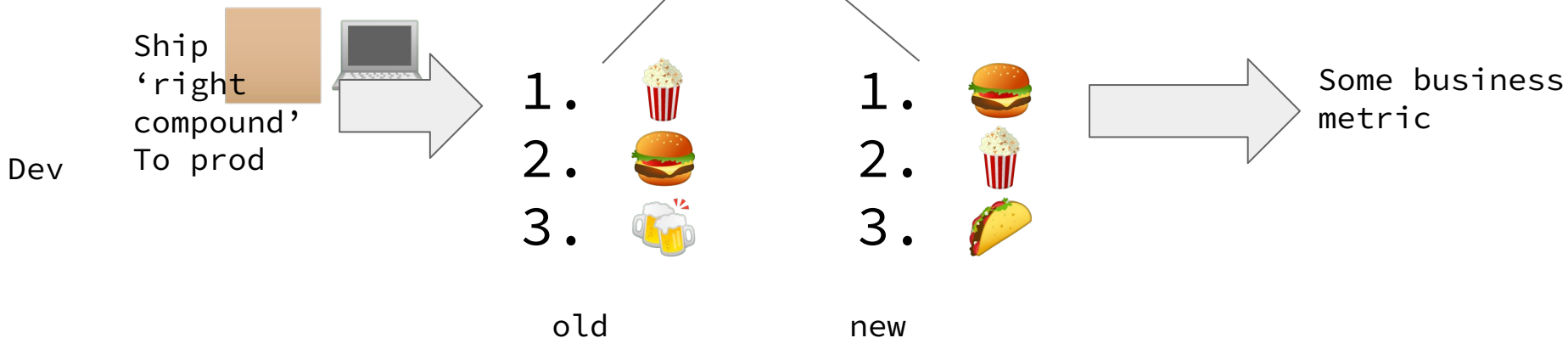


What is the impact of
the compound on
patients?

Our actual job: develop and test hypotheses







Assigning users to both ranking






Instead of is this a good change?


Dev

1.  
2. 
3. 

NDCG=0.7

1. 
2. 
3. 

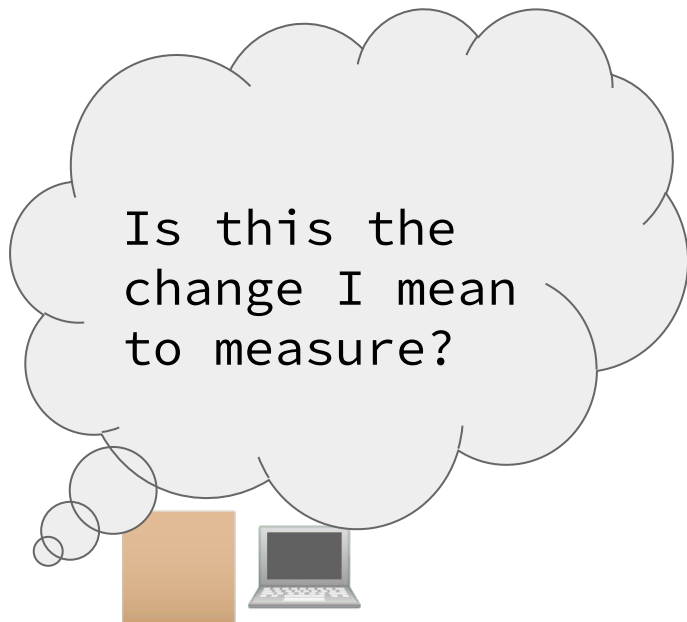
NDCG=0.75

 SHIPIT!!



A/B test

Is this the expected treatment?



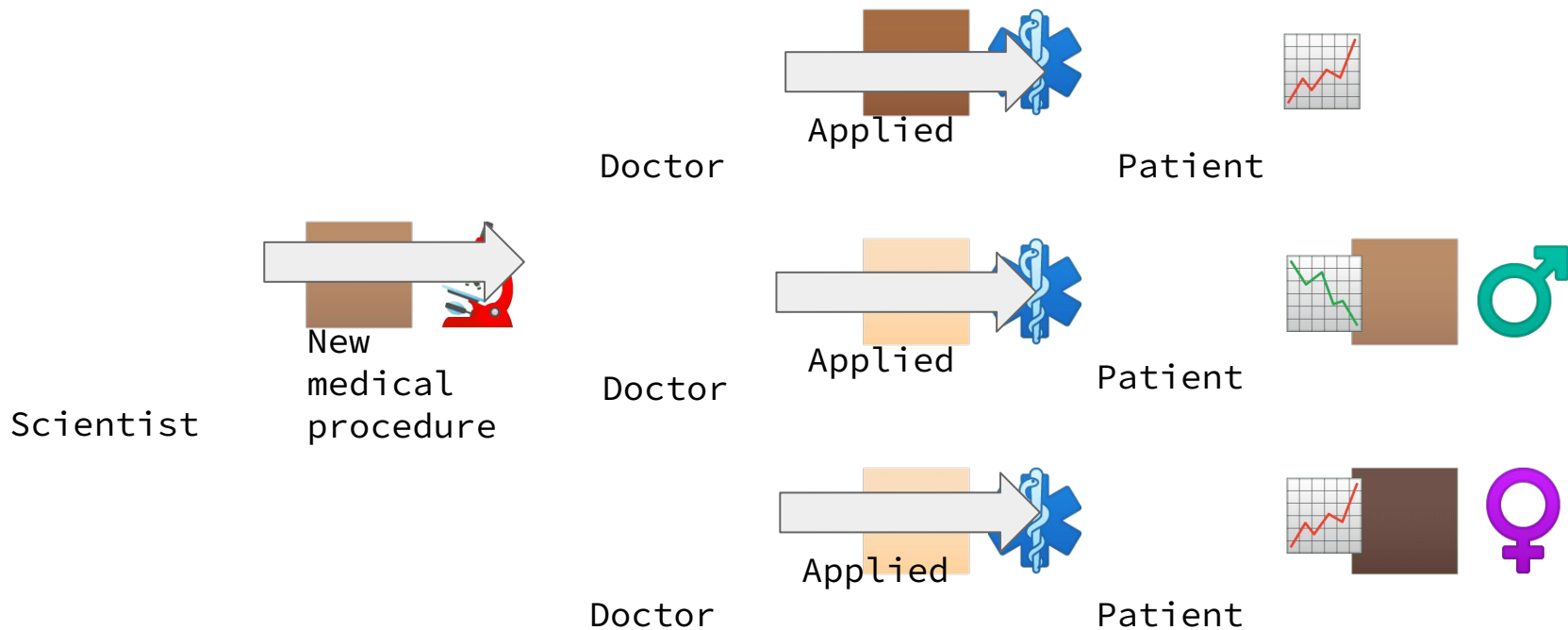
1. 🍿
2. 🍔
3. 🍺

1. 🍔
2. 🍿
3. 🌮

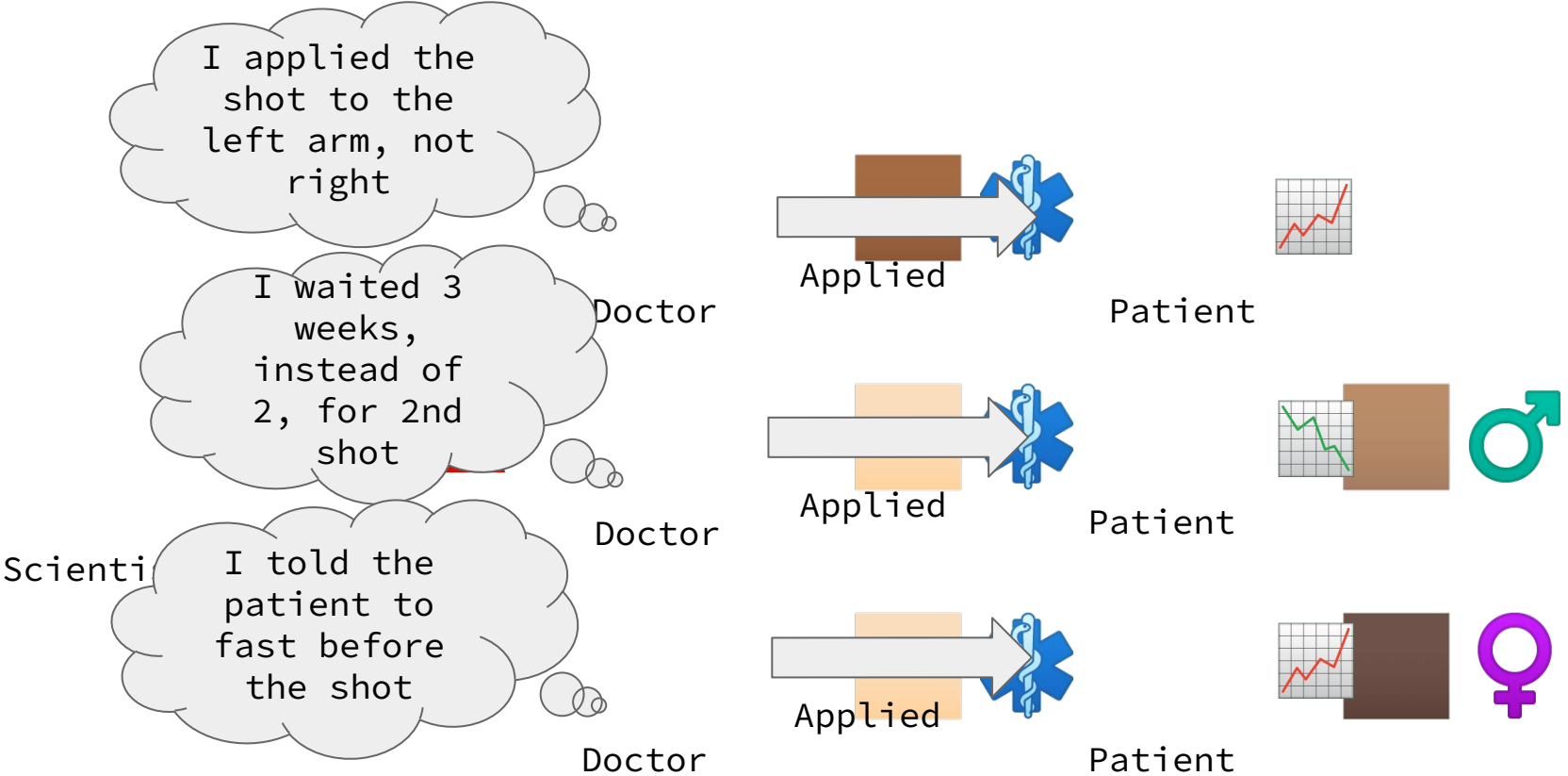
Hypothesis development:

- Do I change the expected queries?
- How much is that change?
- Functionally, is the change what I intended?

What we miss: treatment fidelity



Fidelity: Did we apply intervention as intended?

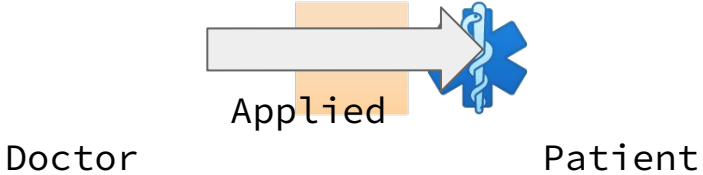
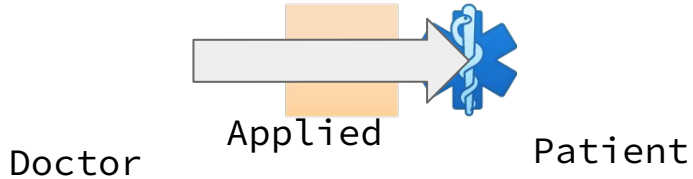
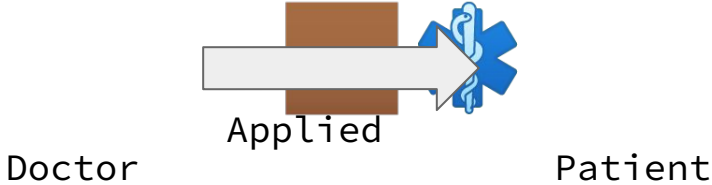


Fidelity: Did we apply intervention as intended?

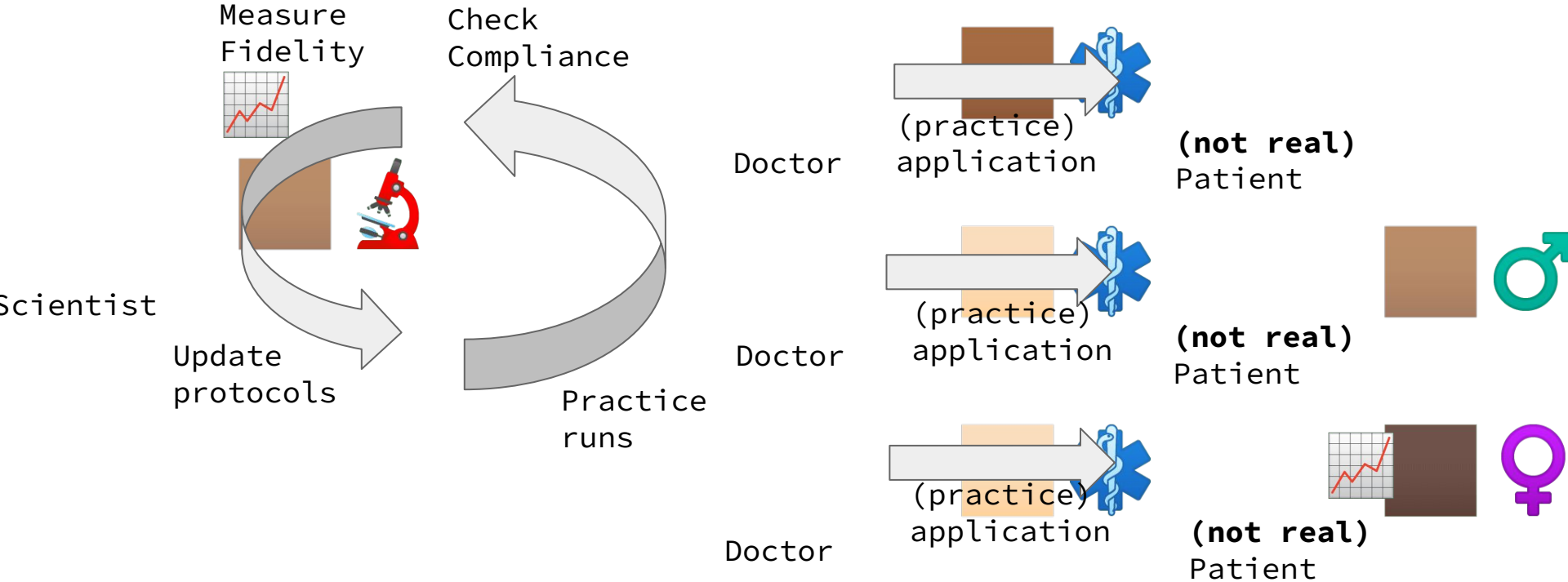
Can I trust the outcome of this trial?



Scientist



Solution: iterate on procedures -> repeatable



Similarly in search (and other ML systems)

Am I actually A/B testing what I think I'm testing?



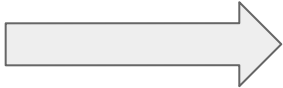
Dev

Model
Search Engine



Applied

User / Query



Applied

User / Query

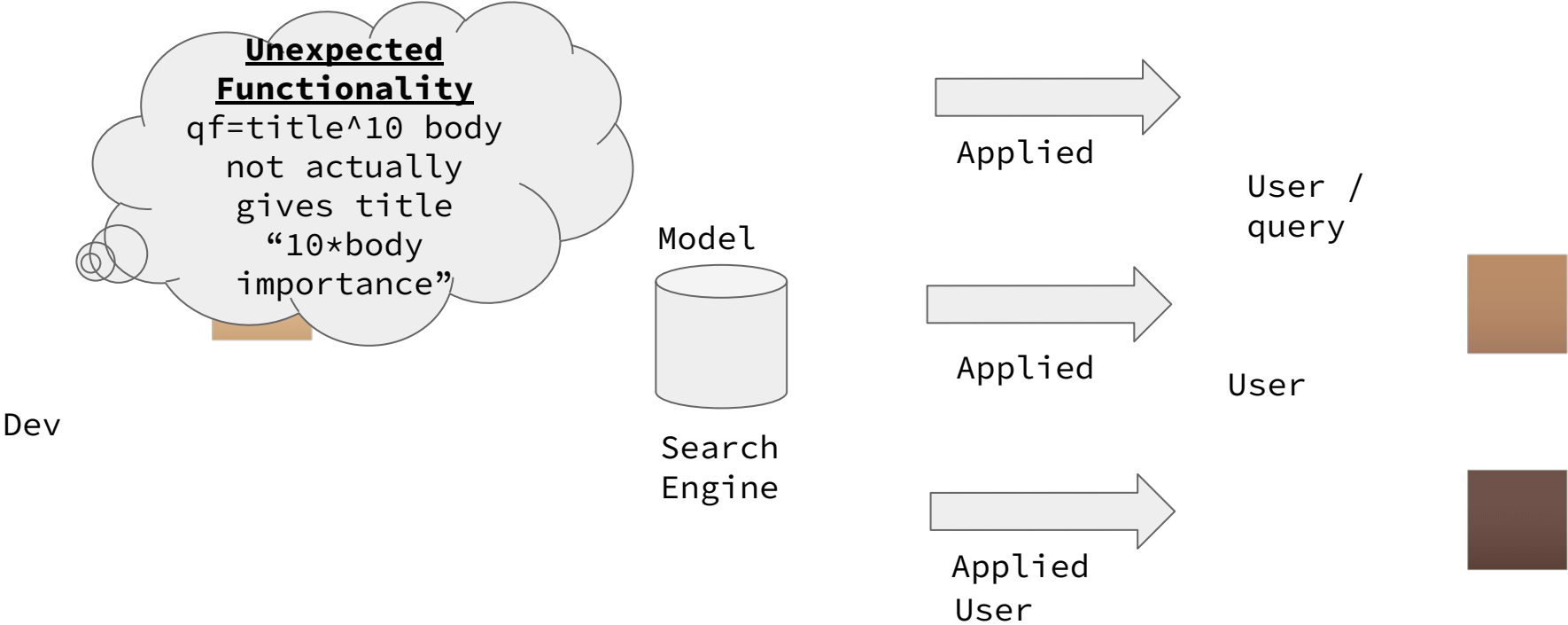


Applied User

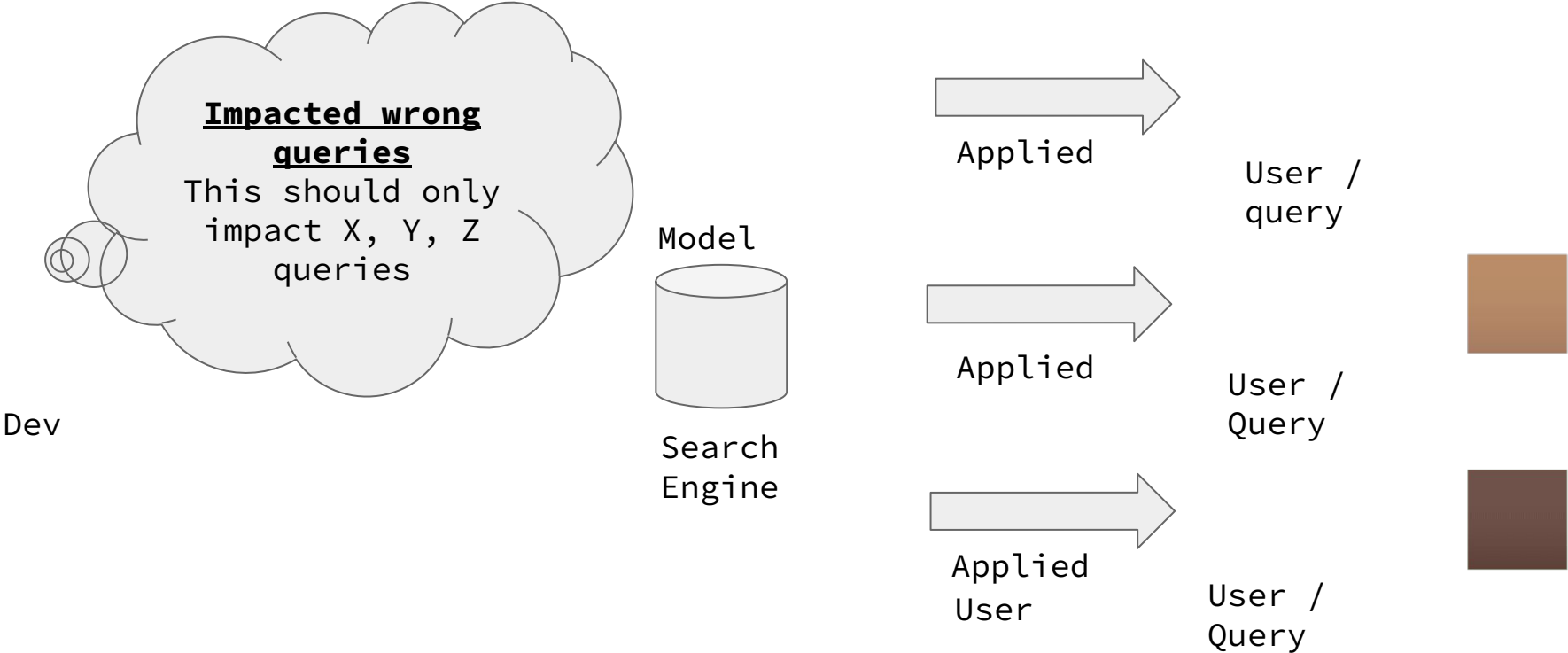
User / Query



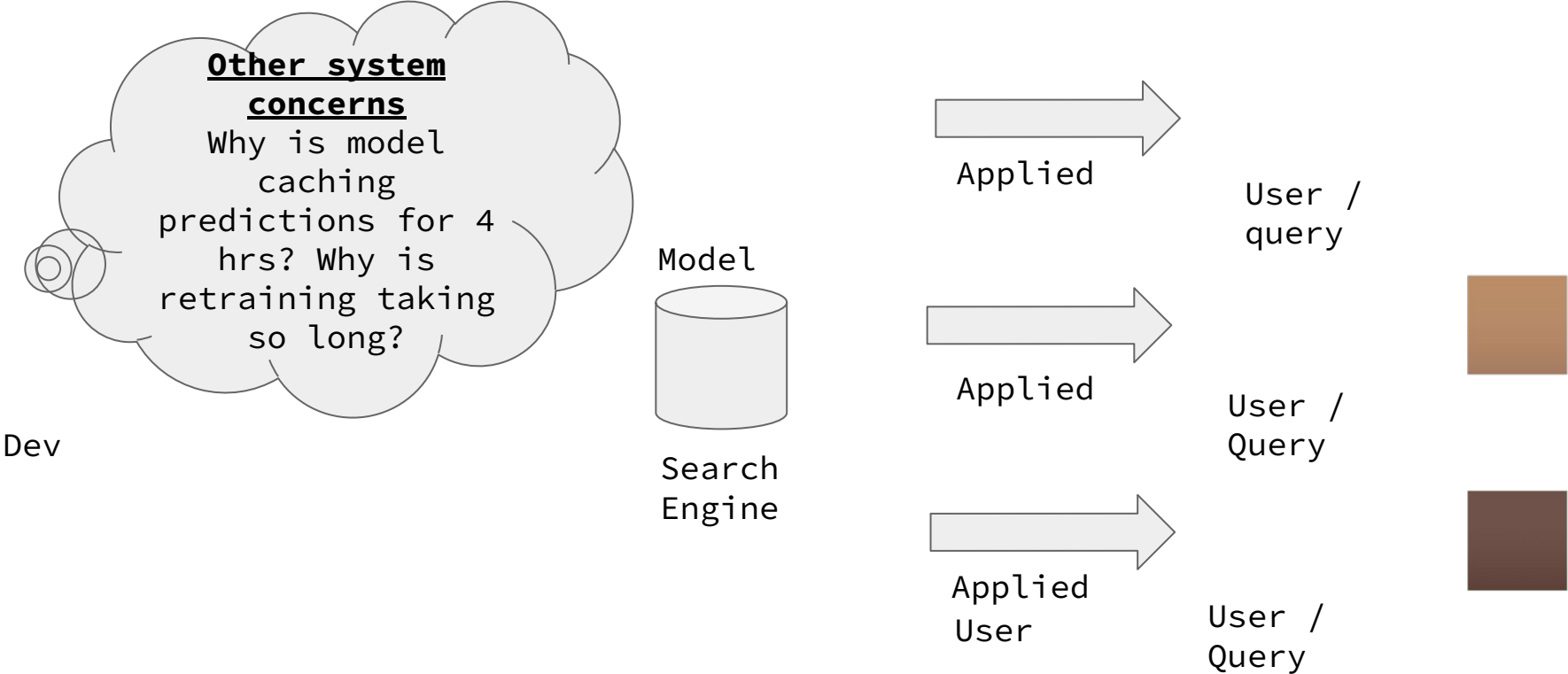
Similarly in search (and other ML systems)



Similarly in search (and other ML systems)



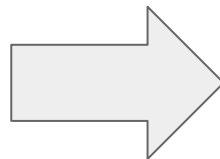
Similarly in search (and other ML systems)



TREATMENT FIDELITY
IN OFFLINE SEARCH
RELEVANCE

Q's for Treatment fidelity in search







- Did I change the expected queries?
- How much is that change?
- Functionally, is the change what I intended?



Hypothesis: this change will improve business outcome

Quantifying change

Jaccard

- | | | | |
|----|---|----|--|
| 1. |  | 1. |  |
| 2. |  | 2. |  |
| 3. |  | 3. |  |

$$\frac{A \cap B}{A \cup B} = \frac{2 \text{ shared results}}{4 \text{ total results}} = 0.5$$

Other Metrics, that account for ranking:







Rank-biased overlap:
<https://github.com/chaingyaochen/rbo>

Damage:
<https://github.com/o19s/search-metrics/blob/main/qual.py#L64>

Quantifying change

Just **change** we
don't know if
its good or bad

Jaccard

- | | | | |
|----|---|----|--|
| 1. |  | 1. |  |
| 2. |  | 2. |  |
| 3. |  | 3. |  |

$$\frac{A \cap B}{A \cup B} = \frac{2 \text{ shared results}}{4 \text{ total results}} = 0.5$$

Other Metrics, that
account for ranking:

Rank-biased overlap:
<https://github.com/chaingyaochen/rbo>

Damage:
<https://github.com/o19s/search-metrics/blob/main/qual.py#L64>

For which queries?

Did I target
'food'
queries?



<u>Query</u>	<u>Jaccard</u> (higher, less change)
Spicy taco	0.5
Beer	0.25

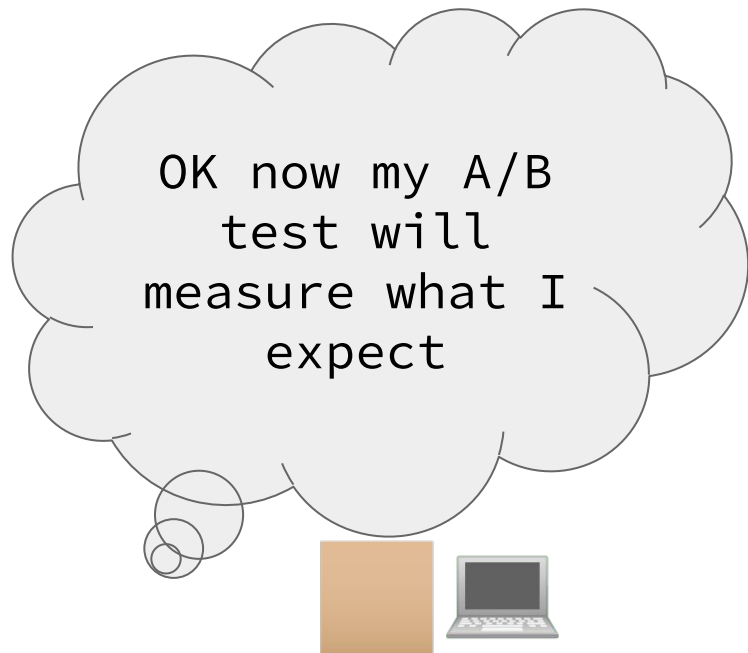
For which queries?

Uhoh, what's
happening...



<u>Query</u>	<u>Jaccard</u> (higher, less change)
Spicy taco	0.5
Beer	0.25
puppies	0.0
Tiger King	0.0
Tree pruning	1.0

Iterate to better target what we expect...



<u>Query</u>	<u>Jaccard</u> (higher, less change)
Spicy taco	0.5
Beer	0.25
puppies	0.8
Tiger King	0.9
Tree pruning	1.0

But is this the change we expect?

I expect to increase
'taxonomical
proximity'
between query
and doc







Assume we have a taxonomy:

q=spicy taco -> Food / in_bread / hinge

Food / in_bread / hinge -> taco

Food / in_bread / detached -> sandwich

1.  Food / in_bread / detached
2.  Food / ...
3.  Food / in_bread / detached
4.  Food / in_bread / detached

But is this the change we expect?

I expect to increase
'taxonomical
proximity'
between query
and doc







Invent a metric:

Assume we have a taxonomy:

q=spicy taco -> Food / in_bread / hinge

Food / in_bread / hinge -> taco

Food / in_bread / detached -> sandwich

1.  Food / in_bread / detached
2.  Food / ...
3.  Food / in_bread / detached
4.  Food / in_bread / detached

$$\text{TAX@4} = 1 + \sum 0.1 * (1 - \text{nodes_apart})$$

But is this the change we expect?

I expect to increase
'taxonomical
proximity'
between query
and doc







Invent a metric:

Assume we have a taxonomy:

q=spicy taco -> Food / in_bread / hinge

Food / in_bread / hinge -> taco

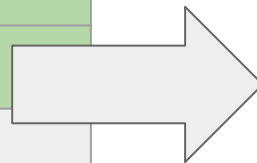
Food / in_bread / detached -> sandwich

1.  Food / in_bread / detached
2.  Food / ...
3.  Food / in_bread / detached
4.  Food / in_bread / detached

$$\begin{aligned} \text{TAX@4} &= 1 + (0.1 * -1) + (0.1 * -2) + (0.1 * 0) + (0.1 * 0) \\ &= \mathbf{0.7} \end{aligned}$$

Did we make expected change?

<u>Query</u>	<u>Jaccard</u> (higher, less change)	<u>Tax Sim Control</u>	<u>Tax Sim Test</u>
Spicy taco	0.5	0.2	0.7
Beer	0.25	0.1	0.5
puppies	0.8	0.0	0.0
Tiger King	0.9	0.0	0.0
Tree pruning	1.0	0.0	0.0

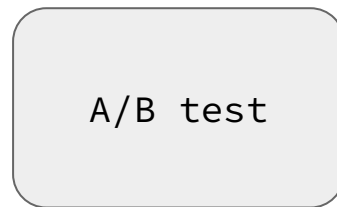
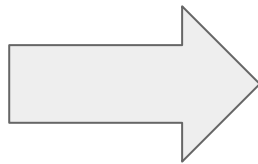


Did we increase this area of our ranking over control?

Hypothesis is valid, now test!

Offline Priorities:

- ✓ Only expected queries changed
- ✓ Expected change implemented
- ☹ Is good change



- ✓ Is good change?

Offline ensure a well-formed hypothesis

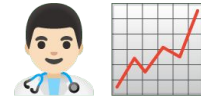
Hypothesis: If the system improves ‘taxonomic similarity’ between query and doc, on food queries

(As tested and clearly shown in this offline test)

...I will see an increase in business metrics

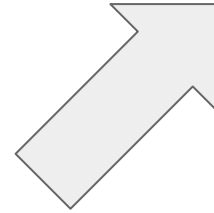
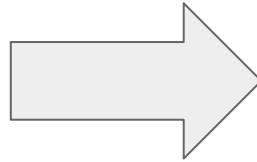
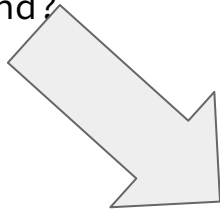
... SCIENCE IS HARD ...

Simulations still valuable



“Chemical Engineering”
Am I actually
producing the right
compound?

Clinical Trials on
humans



Computer
simulation

Animal
testing



Simulations let us...

Iterate on solutions for quality not just fidelity outside online testing

We still want to promote promising changes to A/B...

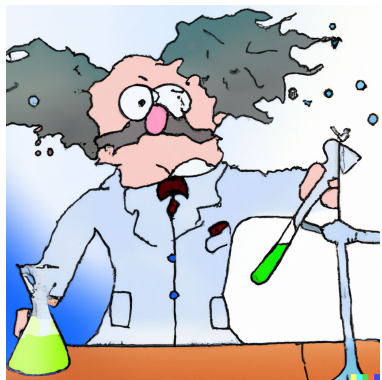
...but more importantly we want them to have the impact what where we expect

... But simulations will always be limited

Biases galore (presentation, attractiveness, the inherent flaws in the judgment-based models)

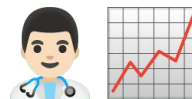
Accept their limitations, use simple mitigations for biases, but understand their flaws instead of fixing them.

... we gain no real knowledge



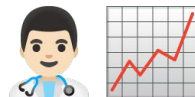
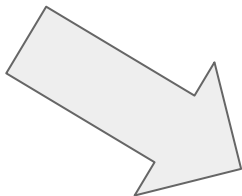
What did I even build!?

All that matters
is NDCG went up!

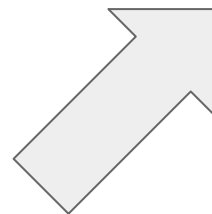


Successful A/B test

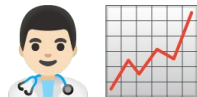
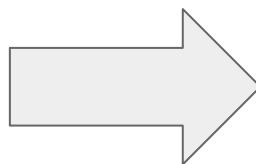
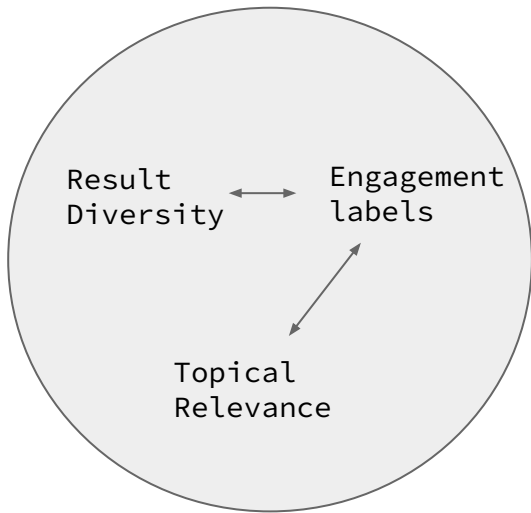
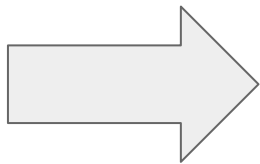
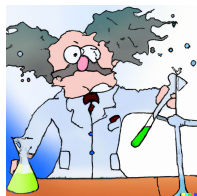
YOLO Again



NDCG went up!



Really these are all just models...

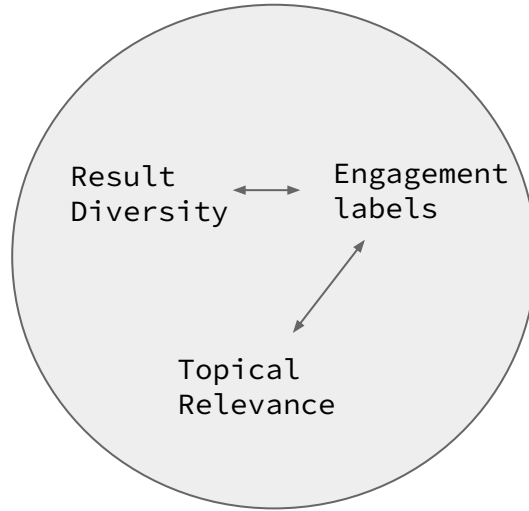
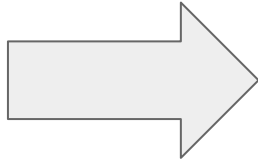
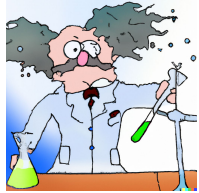


A/B test

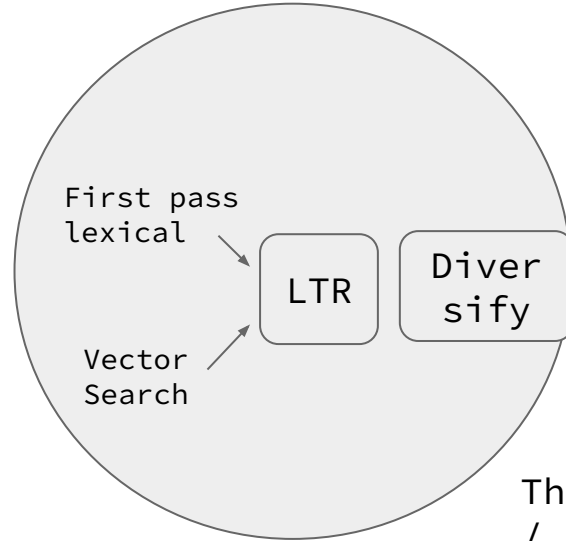


Some model “loss function” that attempts to explain offline

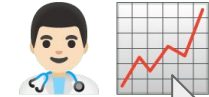
... we use to create systems that probe the world



Theoretical Framework evolves...



... system evolves according to success / failure of framework. Like a “probe”



A/B test

Theories validated / not