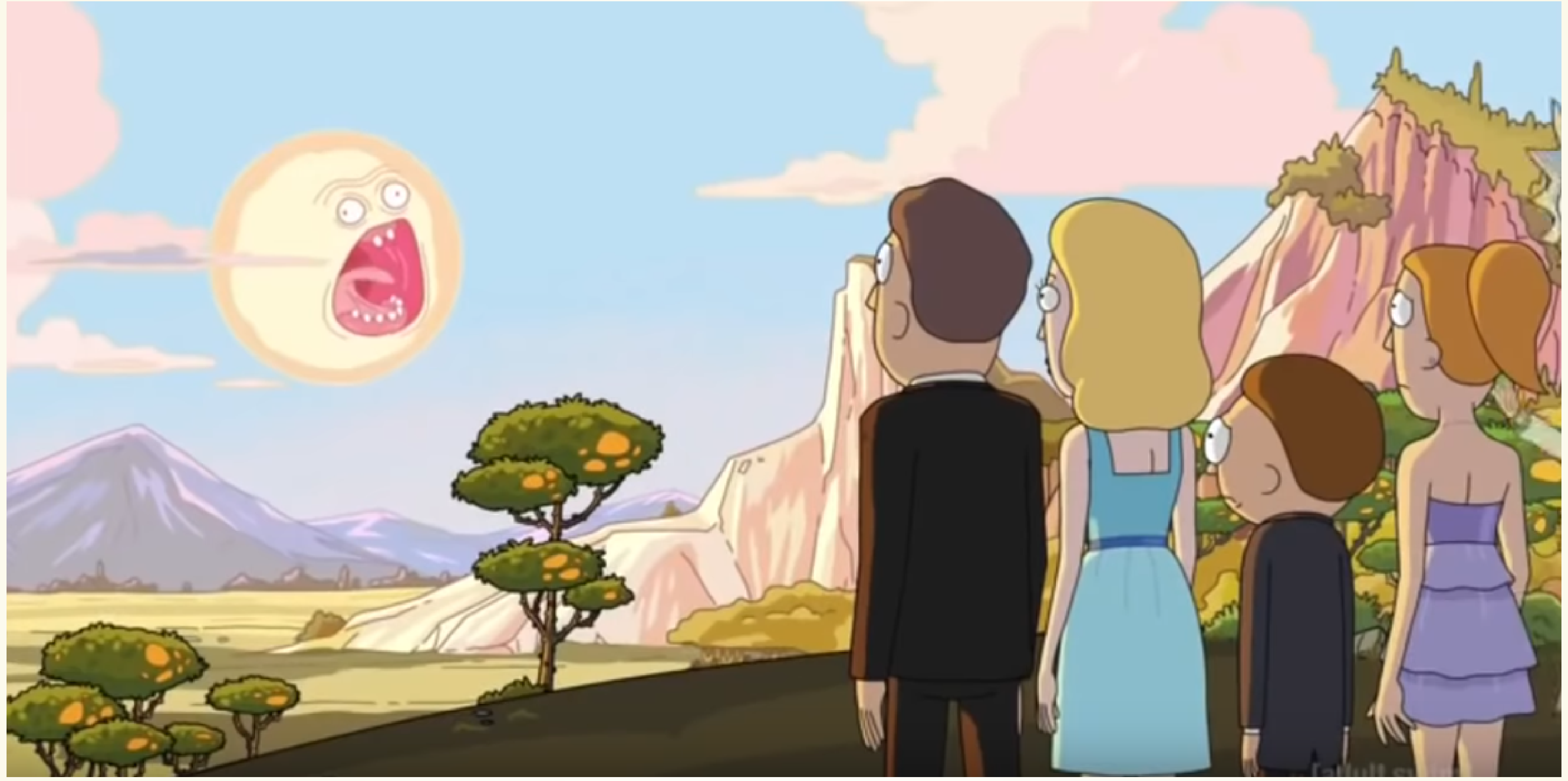
REINFORCEMENT LEARNING

IN SEARCH

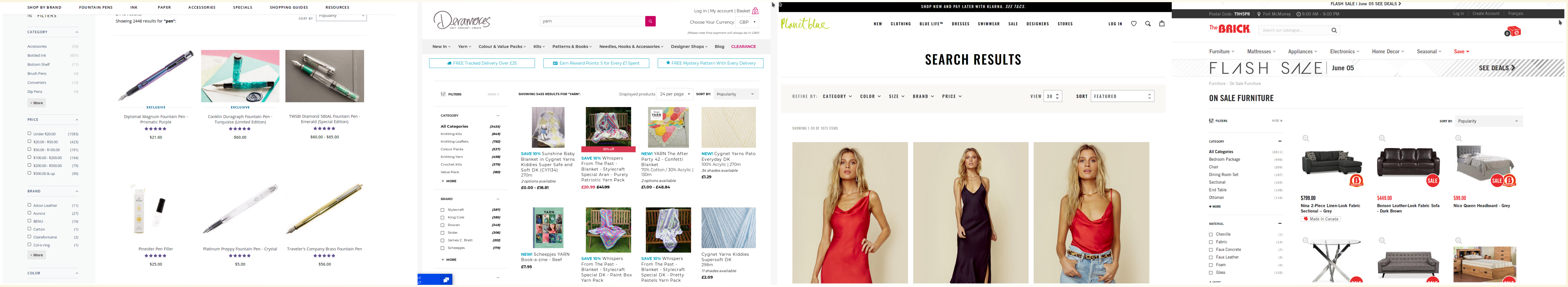
Roman Grebennikov | Findify AB | MICES 2021

ABOUT ME



- Doing ML for 12 years
- Quant trading → Credit risk → eCommerce
- Relevancy & ranking @ Findify

ABOUT FINDIFY



- white-label eCommerce SaaS search
- 20M searchable products
- 50M MAU

AGENDA

- Personalization
- Doing it wrong
- Doing it right *
- Methods and solutions

RANKING: CURRENT STATE

- Search: ElasticSearch/SOLR, TF/IDF, BM25
- Listings: sort by popularity
- Recommendations: CF/SVD/ASL

BM25: an intuitive view

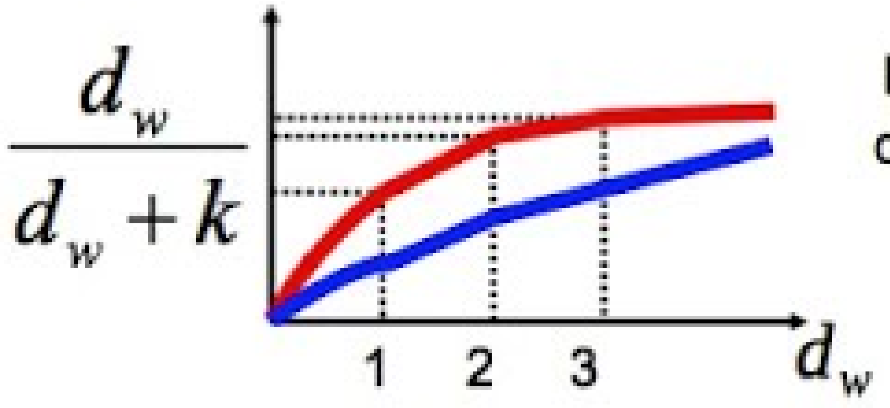
Common words less important

Repetitions of query words ➔ good

$$\log \frac{P(D \mid R=1)}{P(D \mid R=0)} \approx \sum_{w} \left(\underbrace{\frac{d_w(1+k)}{d_w + k\left((1-b) + \frac{b \cdot dl}{avg.dl}\right)}}_{\text{Repetitions less important than different query words}} \cdot \log \frac{N - N_w + \frac{1}{2}}{N_w + \frac{1}{2}} \right)$$

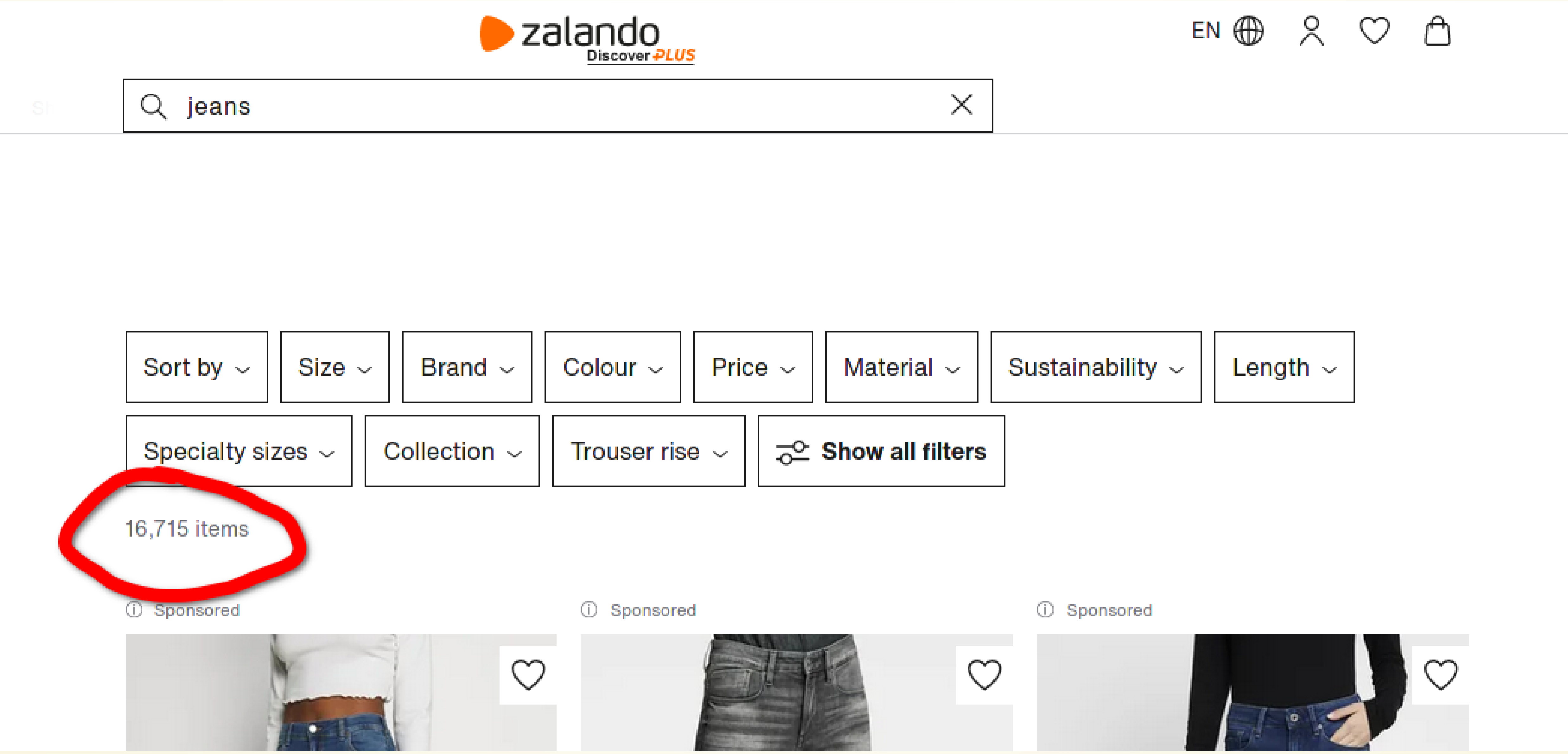
More words in common with the query ➔ good

But more important if document is relatively long (wrt. average)

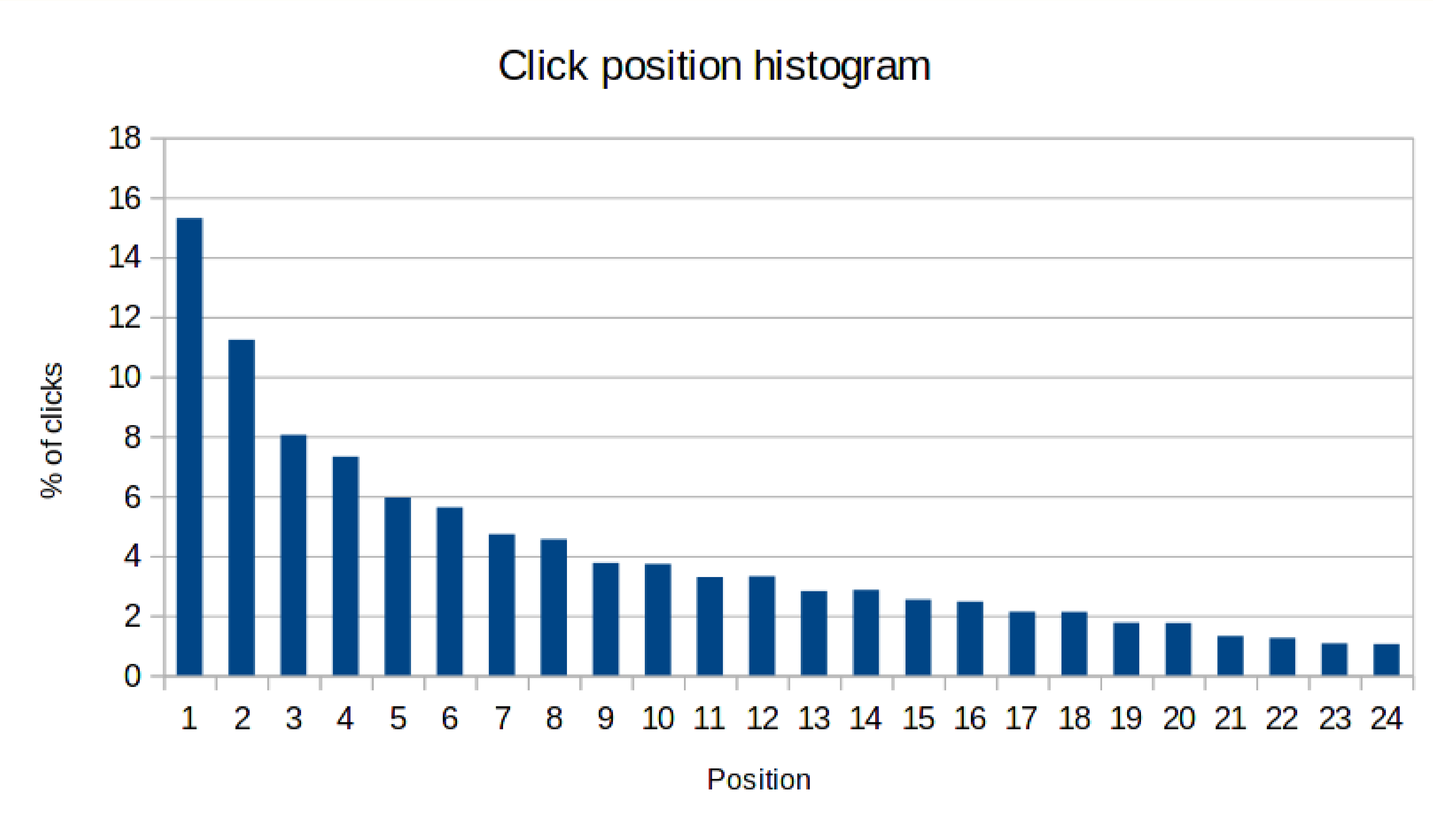


BM25

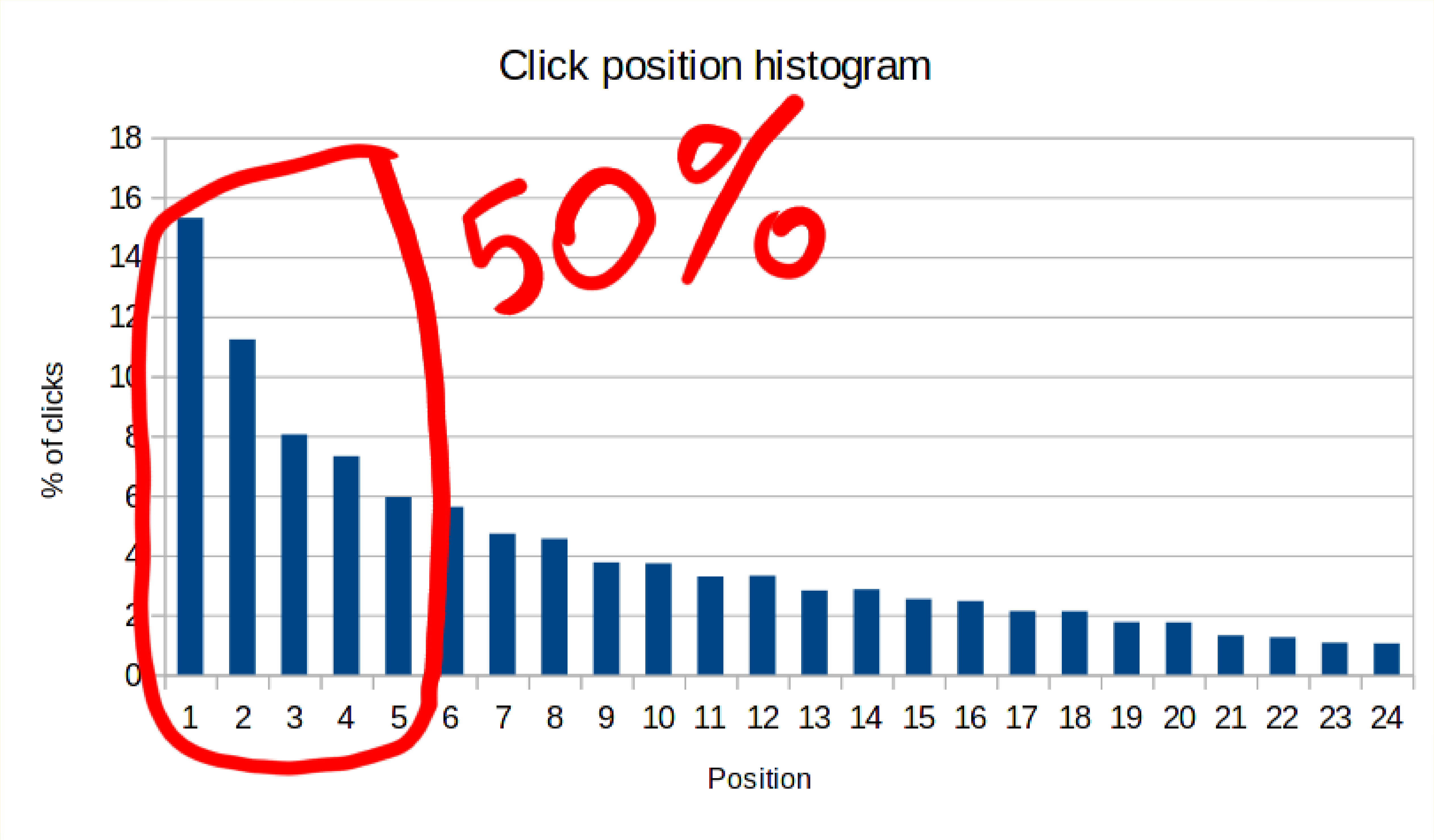
type "cat" - get cat



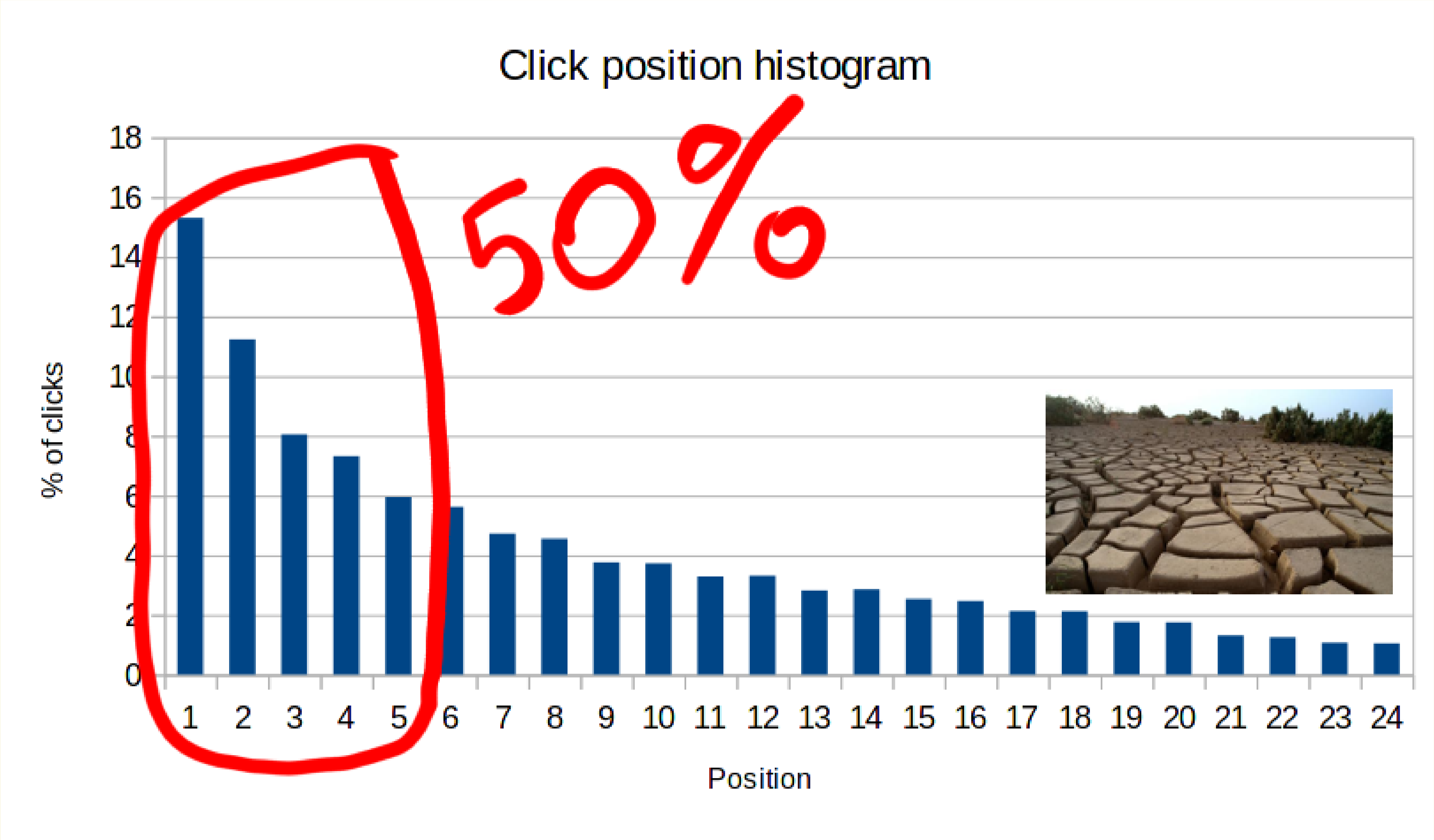
CUSTOMER FOCUS LOST



CUSTOMER FOCUS LOST



CUSTOMER FOCUS LOST



HOW TO LOOSE A CUSTOMER

- only first items are important
- irrelevant top-N = lost customer

BM25 WITH BENEFITS

Idea: rank by relevancy AND popularity

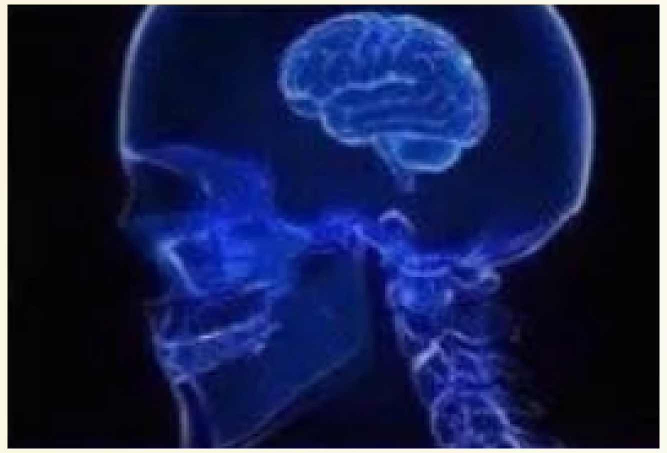
score = bm25 * log(1 + clicks)

Product	BM25	# of clicks	Score
p1	1.7	20	2.24
p2	1.6	350	4.07
p3	0.7	600	1.94

BM25 WITH BENEFITS

Product	BM25	# of clicks	Score
p2	1.6	350	4.07
p1	1.7	20	2.24
p3	0.7	600	1.94

BM25 WITH BENEFITS



score = bm25

score = bm25 * log(1 + clicks)

score = $c_1 + c_2$ * bm25 * log(c_3 + clicks)

A BETTER RANKING FORMULA?

- why only clicks?
  - let's add pageviews, purchases and carts!
- let's add windows for 1-7-14 days!
- more constants to tune!

A BETTER RANKING FORMULA?

$$
\begin{gathered}
\mathcal{L}_{SM} = -\tfrac{1}{2}\partial_\nu g^a_\mu \partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \tfrac{1}{4} g^2_s f^{abc} f^{ade} g^b_\mu g^c_\nu g^d_\mu g^e_\nu - \partial_\nu W^+_\mu \partial_\nu W^-_\mu - \\
M^2 W^+_\mu W^-_\mu - \tfrac{1}{2}\partial_\nu Z^0_\mu \partial_\nu Z^0_\mu - \tfrac{1}{2c^2_w} M^2 Z^0_\mu Z^0_\mu - \tfrac{1}{2}\partial_\mu A_\nu \partial_\mu A_\nu - igc_w(\partial_\nu Z^0_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - \\
Z^0_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + Z^0_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)) - igs_w(\partial_\nu A_\mu (W^+_\mu W^-_\nu - \\
W^+_\nu W^-_\mu) - A_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + A_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)) - \\
\tfrac{1}{2} g^2 W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \tfrac{1}{2} g^2 W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2 c^2_w (Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + \\
g^2 s^2_w (A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2 s_w c_w (A_\mu Z^0_\nu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - \\
2A_\mu Z^0_\mu W^+_\nu W^-_\nu) - \tfrac{1}{2}\partial_\mu H \partial_\mu H - 2M^2 \alpha_h H^2 - \partial_\mu \phi^+ \partial_\mu \phi^- - \tfrac{1}{2}\partial_\mu \phi^0 \partial_\mu \phi^0 - \\
\beta_h \left(\tfrac{2M^2}{g^2} + \tfrac{2M}{g} H + \tfrac{1}{2}(H^2 + \phi^0\phi^0 + 2\phi^+\phi^-)\right) + \tfrac{2M^4}{g^2}\alpha_h - g\alpha_h M \left(H^3 + H\phi^0\phi^0 + 2H\phi^+\phi^-\right) - \\
\tfrac{1}{8} g^2 \alpha_h \left(H^4 + (\phi^0)^4 + 4(\phi^+\phi^-)^2 + 4(\phi^0)^2\phi^+\phi^- + 4H^2\phi^+\phi^- + 2(\phi^0)^2 H^2\right) - gMW^+_\mu W^-_\mu H - \\
\tfrac{1}{2} g \tfrac{M}{c^2_w} Z^0_\mu Z^0_\mu H - \tfrac{1}{2} ig \left(W^+_\mu (\phi^0 \partial_\mu \phi^- - \phi^- \partial_\mu \phi^0) - W^-_\mu (\phi^0 \partial_\mu \phi^+ - \phi^+ \partial_\mu \phi^0)\right) + \\
\tfrac{1}{2} g \left(W^+_\mu (H \partial_\mu \phi^- - \phi^- \partial_\mu H) + W^-_\mu (H \partial_\mu \phi^+ - \phi^+ \partial_\mu H)\right) + \tfrac{1}{2} g \tfrac{1}{c_w} (Z^0_\mu (H \partial_\mu \phi^0 - \phi^0 \partial_\mu H) + \\
M \left(\tfrac{1}{c_w} Z^0_\mu \partial_\mu \phi^0 + W^+_\mu \partial_\mu \phi^- + W^-_\mu \partial_\mu \phi^+\right) - ig \tfrac{s^2_w}{c_w} M Z^0_\mu (W^+_\mu \phi^- - W^-_\mu \phi^+) + igs_w M A_\mu (W^+_\mu \phi^- - \\
W^-_\mu \phi^+) - ig \tfrac{1-2c^2_w}{2c_w} Z^0_\mu (\phi^+ \partial_\mu \phi^- - \phi^- \partial_\mu \phi^+) + igs_w A_\mu (\phi^+ \partial_\mu \phi^- - \phi^- \partial_\mu \phi^+) - \\
\tfrac{1}{4} g^2 W^+_\mu W^-_\mu \left(H^2 + (\phi^0)^2 + 2\phi^+\phi^-\right) - \tfrac{1}{8} g^2 \tfrac{1}{c^2_w} Z^0_\mu Z^0_\mu \left(H^2 + (\phi^0)^2 + 2(2s^2_w - 1)^2 \phi^+\phi^-\right) - \\
\tfrac{1}{2} g^2 \tfrac{s^2_w}{c_w} Z^0_\mu \phi^0 (W^+_\mu \phi^- + W^-_\mu \phi^+) - \tfrac{1}{2} ig^2 \tfrac{s^2_w}{c_w} Z^0_\mu H (W^+_\mu \phi^- - W^-_\mu \phi^+) + \tfrac{1}{2} g^2 s_w A_\mu \phi^0 (W^+_\mu \phi^- + \\
W^-_\mu \phi^+) + \tfrac{1}{2} ig^2 s_w A_\mu H (W^+_\mu \phi^- - W^-_\mu \phi^+) - g^2 \tfrac{s_w}{c_w} (2c^2_w - 1) Z^0_\mu A_\mu \phi^+\phi^- - g^2 s^2_w A_\mu A_\mu \phi^+\phi^- + \\
\tfrac{1}{2} ig_s \lambda^a_{ij} (\bar{q}^\sigma_i \gamma^\mu q^\sigma_j) g^a_\mu - \bar{e}^\lambda (\gamma\partial + m^\lambda_e) e^\lambda - \bar{\nu}^\lambda (\gamma\partial + m^\lambda_\nu) \nu^\lambda - \bar{u}^\lambda_j (\gamma\partial + m^\lambda_u) u^\lambda_j - \bar{d}^\lambda_j (\gamma\partial + m^\lambda_d) d^\lambda_j + \\
igs_w A_\mu \left(-(\bar{e}^\lambda \gamma^\mu e^\lambda) + \tfrac{2}{3}(\bar{u}^\lambda_j \gamma^\mu u^\lambda_j) - \tfrac{1}{3}(\bar{d}^\lambda_j \gamma^\mu d^\lambda_j)\right) + \tfrac{ig}{4c_w} Z^0_\mu \{(\bar{\nu}^\lambda \gamma^\mu (1+\gamma^5) \nu^\lambda) + (\bar{e}^\lambda \gamma^\mu (4s^2_w - \\
1 - \gamma^5) e^\lambda) + (\bar{d}^\lambda_j \gamma^\mu (\tfrac{4}{3} s^2_w - 1 - \gamma^5) d^\lambda_j) + (\bar{u}^\lambda_j \gamma^\mu (1 - \tfrac{8}{3} s^2_w + \gamma^5) u^\lambda_j)\} + \\
\tfrac{ig}{2\sqrt{2}} W^+_\mu \left((\bar{\nu}^\lambda \gamma^\mu (1+\gamma^5) U^{lep}{}_{\lambda\kappa} e^\kappa) + (\bar{u}^\lambda_j \gamma^\mu (1+\gamma^5) C_{\lambda\kappa} d^\kappa_j)\right) + \\
\tfrac{ig}{2\sqrt{2}} W^-_\mu \left((\bar{e}^\kappa U^{lep\dagger}{}_{\kappa\lambda} \gamma^\mu (1+\gamma^5) \nu^\lambda) + (\bar{d}^\kappa_j C^\dagger_{\kappa\lambda} \gamma^\mu (1+\gamma^5) u^\lambda_j)\right) + \\
\tfrac{ig}{2M\sqrt{2}} \phi^+ \left(-m^\kappa_e (\bar{\nu}^\lambda U^{lep}{}_{\lambda\kappa} (1-\gamma^5) e^\kappa) + m^\lambda_\nu (\bar{\nu}^\lambda U^{lep}{}_{\lambda\kappa} (1+\gamma^5) e^\kappa\right) + \\
\tfrac{ig}{2M\sqrt{2}} \phi^- \left(m^\lambda_e (\bar{e}^\lambda U^{lep\dagger}_{\lambda\kappa} (1+\gamma^5) \nu^\kappa) - m^\kappa_\nu (\bar{e}^\lambda U^{lep\dagger}_{\lambda\kappa} (1-\gamma^5) \nu^\kappa\right) - \tfrac{g}{2} \tfrac{m^\lambda_\nu}{M} H (\bar{\nu}^\lambda \nu^\lambda) - \\
\tfrac{g}{2} \tfrac{m^\lambda_e}{M} H (\bar{e}^\lambda e^\lambda) + \tfrac{ig}{2} \tfrac{m^\lambda_\nu}{M} \phi^0 (\bar{\nu}^\lambda \gamma^5 \nu^\lambda) - \tfrac{ig}{2} \tfrac{m^\lambda_e}{M} \phi^0 (\bar{e}^\lambda \gamma^5 e^\lambda) - \tfrac{1}{4} \bar{\nu}_\lambda M^R_{\lambda\kappa} (1-\gamma_5) \hat{\nu}_\kappa - \\
\tfrac{1}{4} \overline{\bar{\nu}_\lambda M^R_{\lambda\kappa} (1-\gamma_5) \hat{\nu}_\kappa} + \tfrac{ig}{2M\sqrt{2}} \phi^+ \left(-m^\kappa_d (\bar{u}^\lambda_j C_{\lambda\kappa} (1-\gamma^5) d^\kappa_j) + m^\lambda_u (\bar{u}^\lambda_j C_{\lambda\kappa} (1+\gamma^5) d^\kappa_j\right) + \\
\tfrac{ig}{2M\sqrt{2}} \phi^- \left(m^\lambda_d (\bar{d}^\lambda_j C^\dagger_{\lambda\kappa} (1+\gamma^5) u^\kappa_j) - m^\kappa_u (\bar{d}^\lambda_j C^\dagger_{\lambda\kappa} (1-\gamma^5) u^\kappa_j\right) - \tfrac{g}{2} \tfrac{m^\lambda_u}{M} H (\bar{u}^\lambda_j u^\lambda_j) - \tfrac{g}{2} \tfrac{m^\lambda_d}{M} H (\bar{d}^\lambda_j d^\lambda_j) + \\
\tfrac{ig}{2} \tfrac{m^\lambda_u}{M} \phi^0 (\bar{u}^\lambda_j \gamma^5 u^\lambda_j) - \tfrac{ig}{2} \tfrac{m^\lambda_d}{M} \phi^0 (\bar{d}^\lambda_j \gamma^5 d^\lambda_j)
\end{gathered}
$$

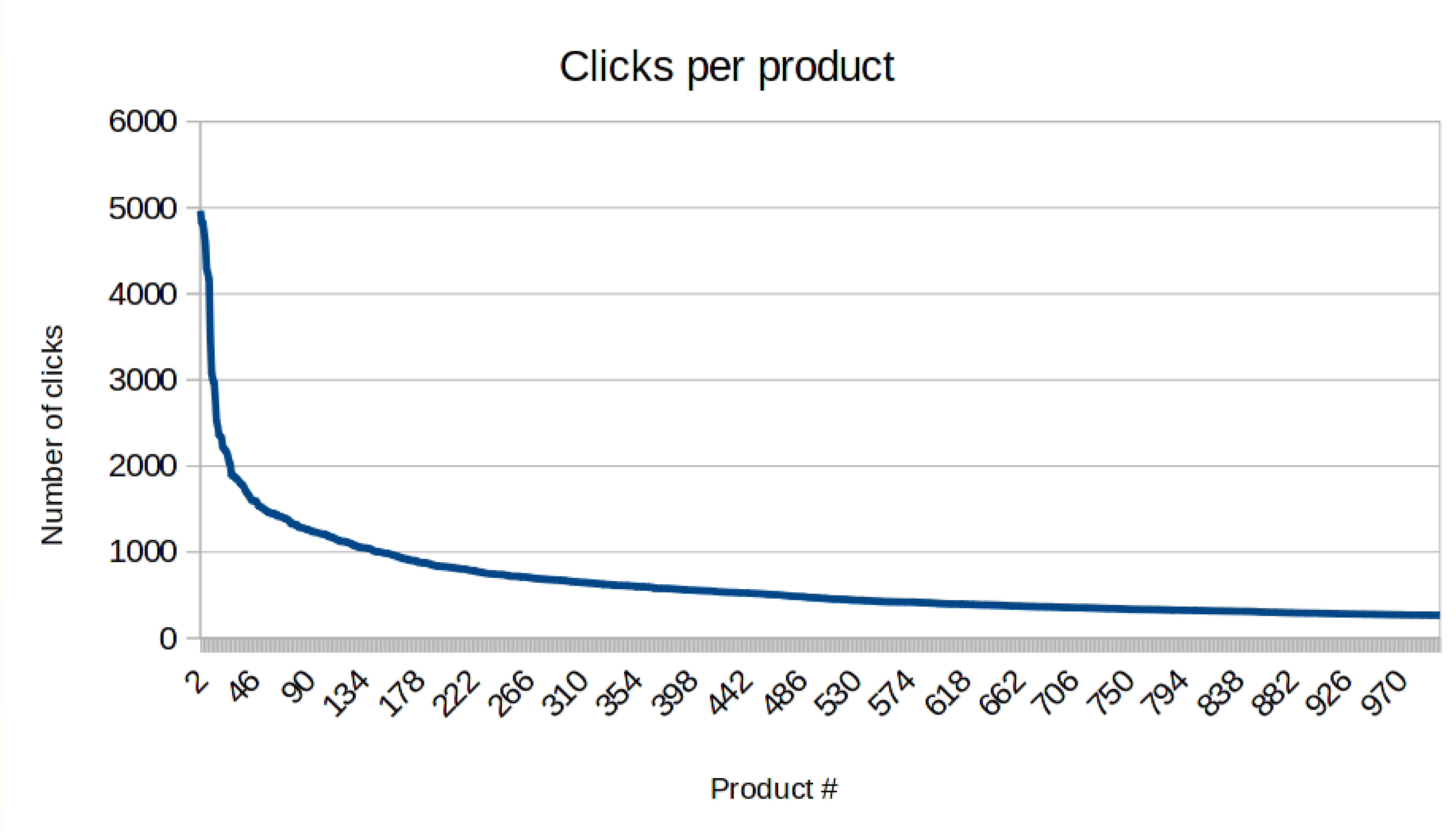
probably won't work out of the box -> needs training!


LEARN-TO-RANK
ME WITH
ELASTICSEARCH
A.WALLIN

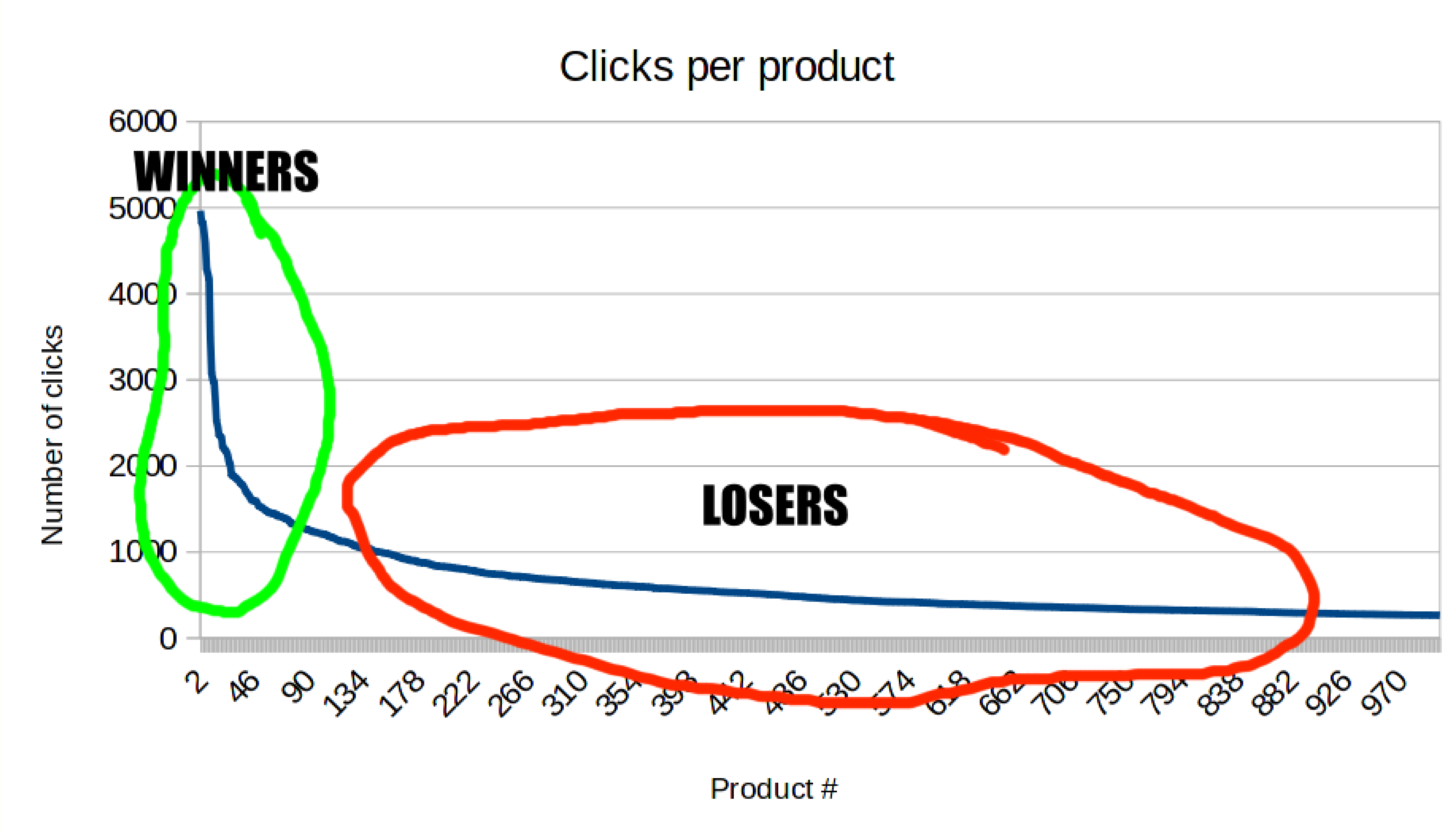
LTR: GOAL TO OPTIMIZE

Product	BM25	# of clicks	Interaction
p1	1.7	100	-
p2	0.9	20	click
p3	0.7	600	-

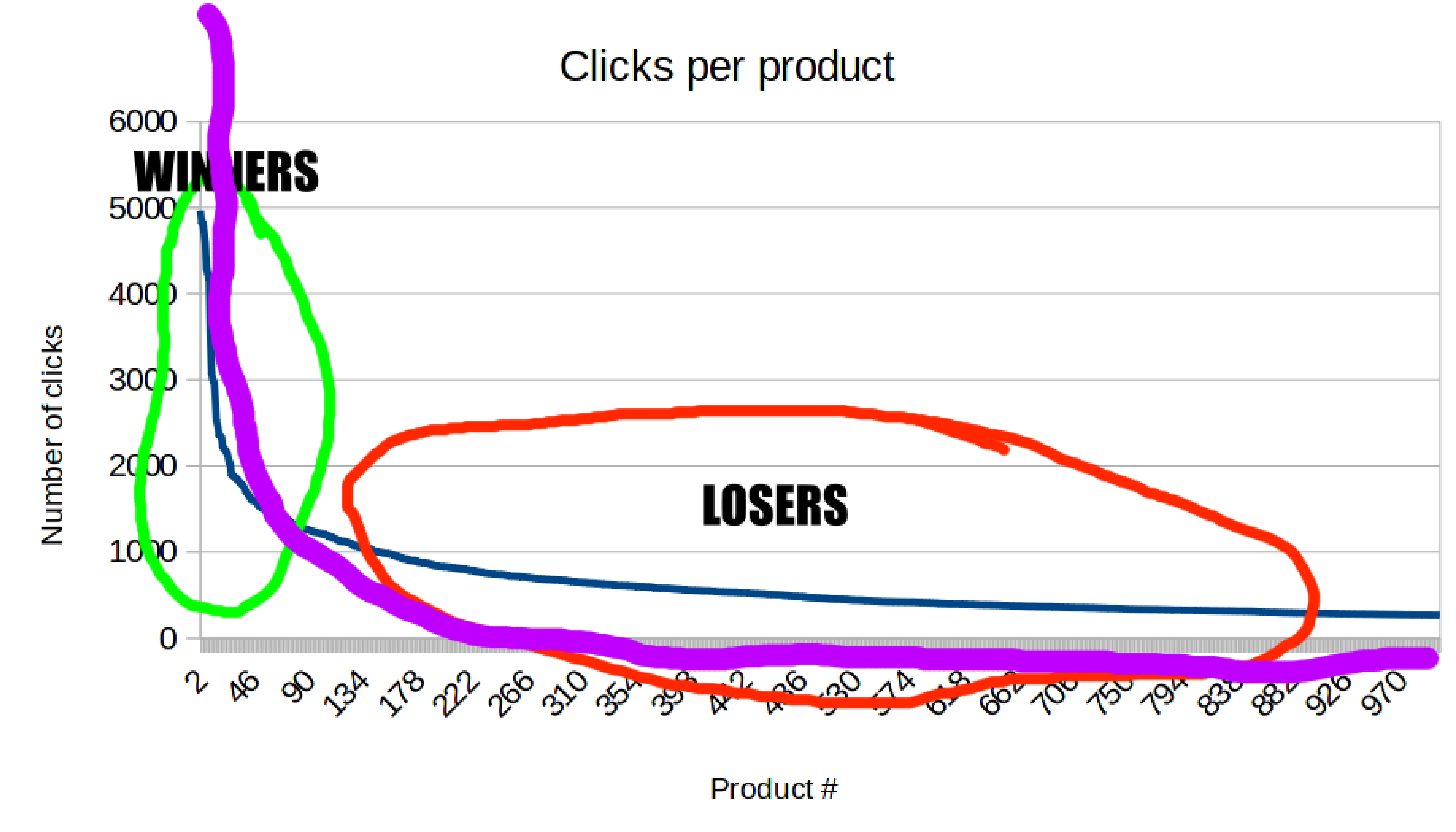
FEEDBACK LOOP



FEEDBACK LOOP OF BAD SEARCH

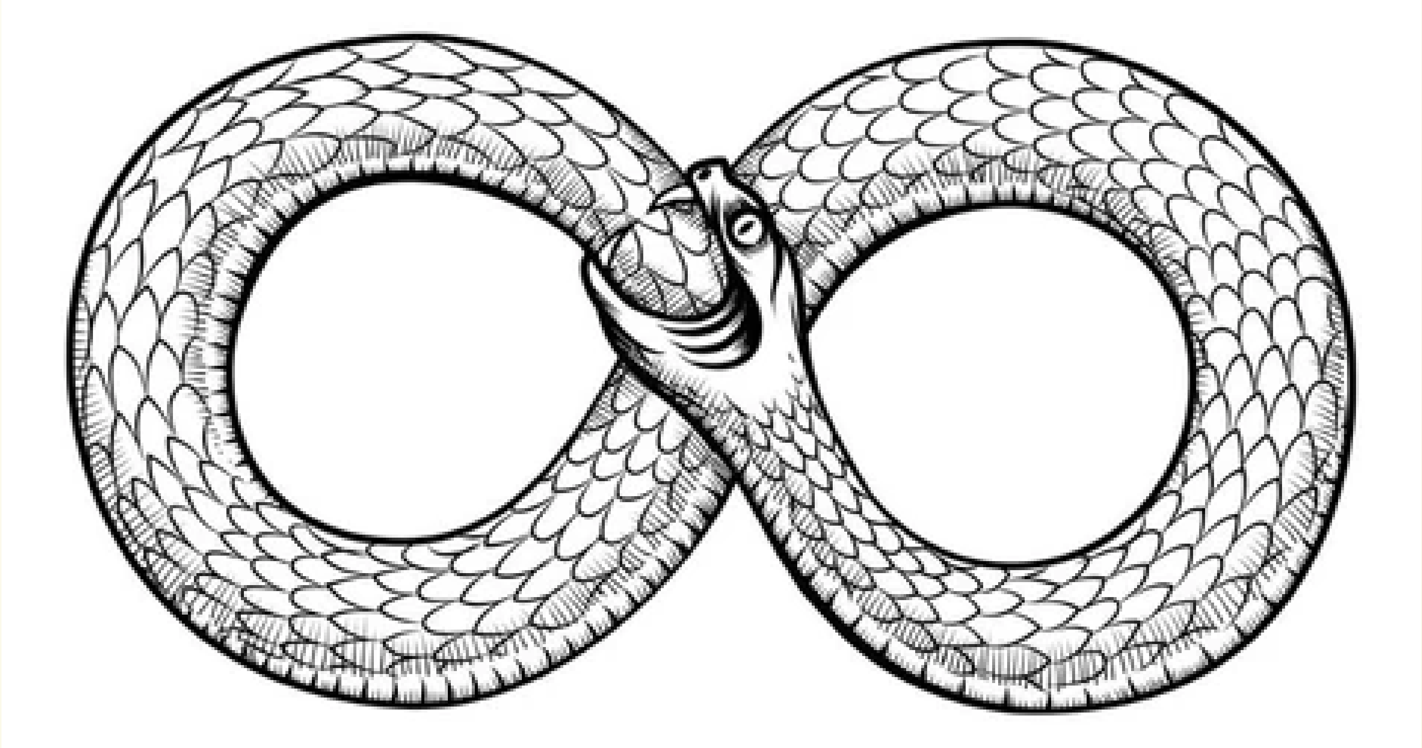


FEEDBACK LOOP OF BAD SEARCH

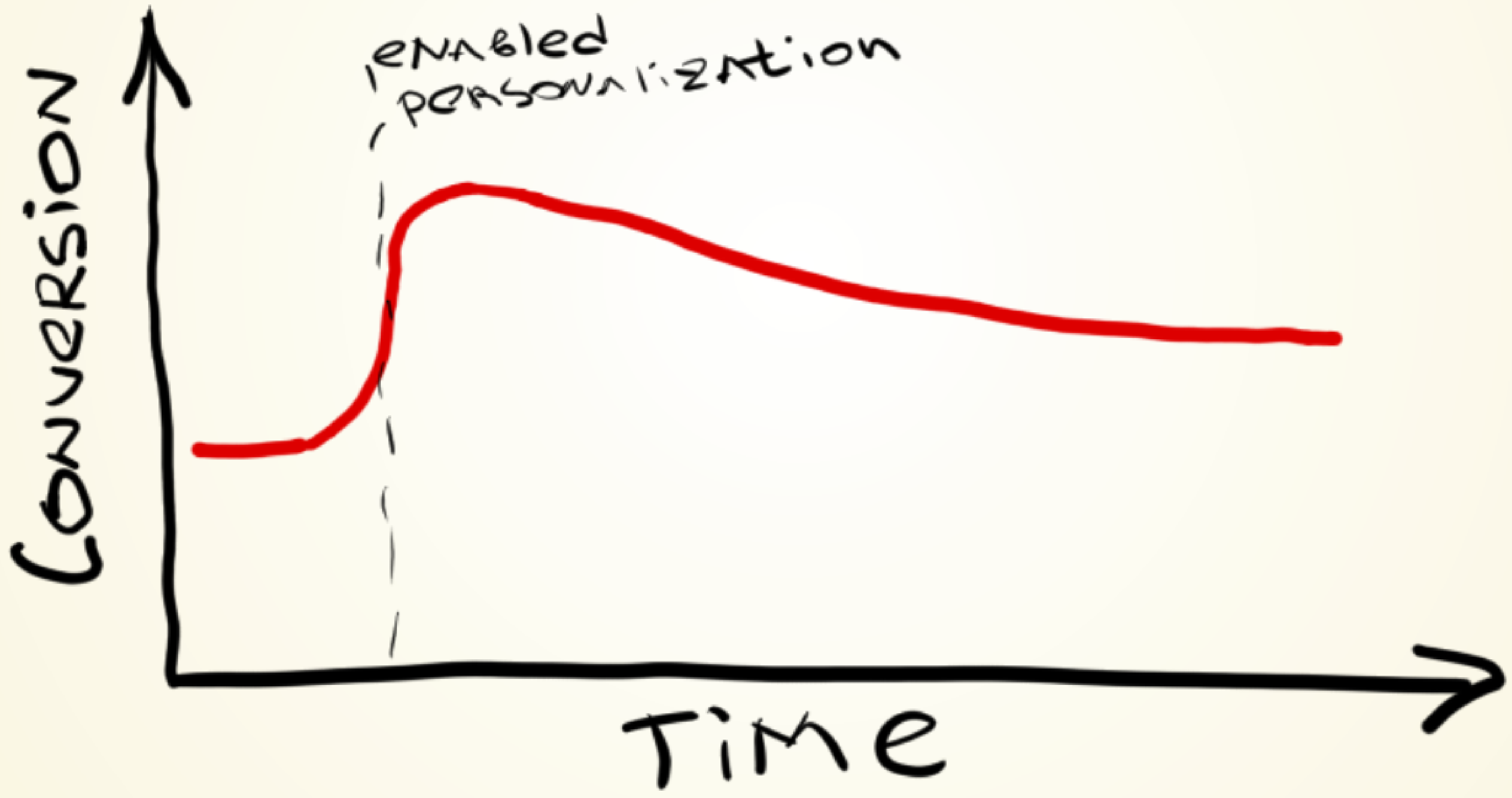


MODEL BIAS

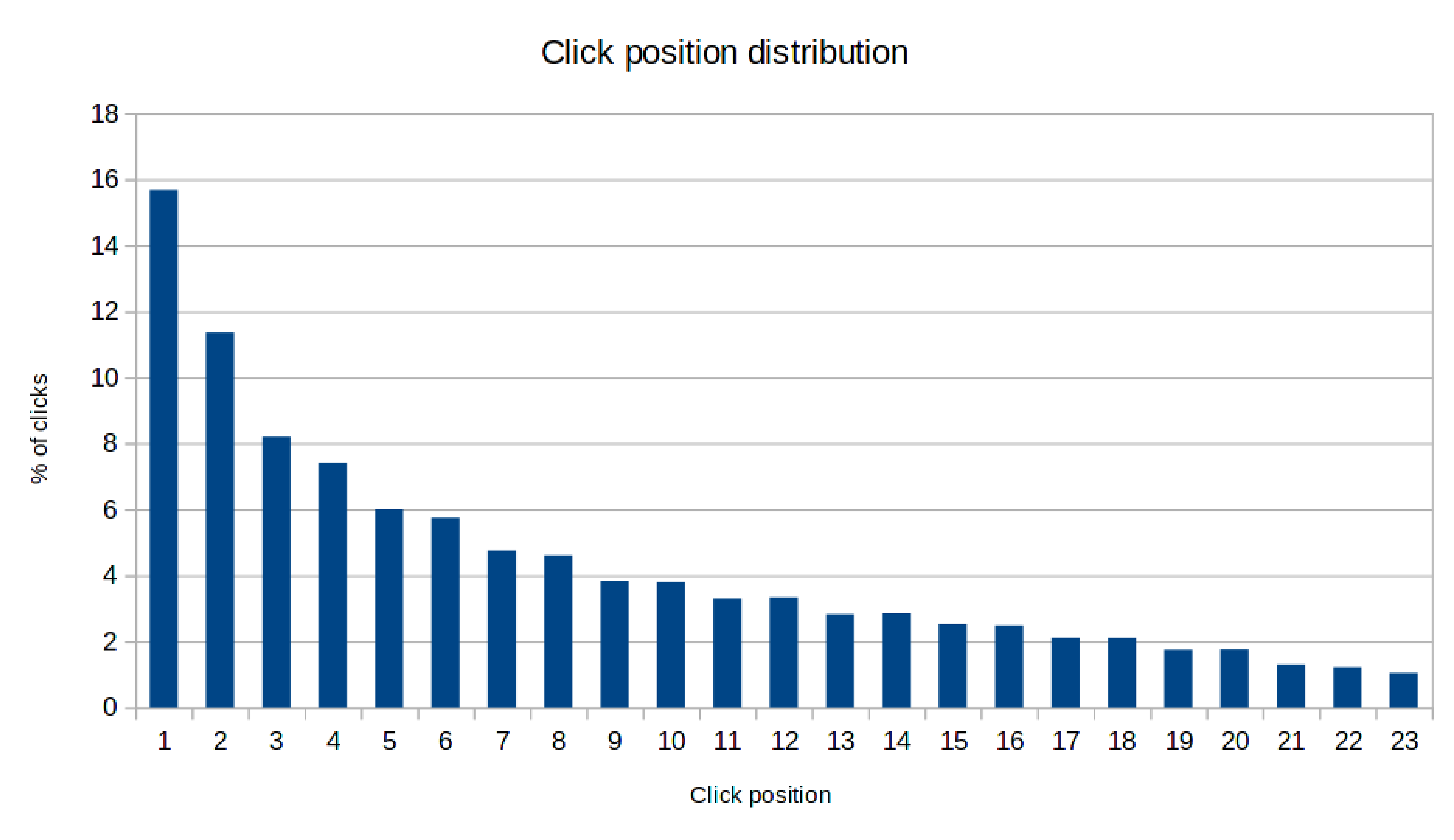
- Produce ranking based on past clicks
- Collect clicks influenced by previous ranking
- Retrain model on clicks
- Repeat



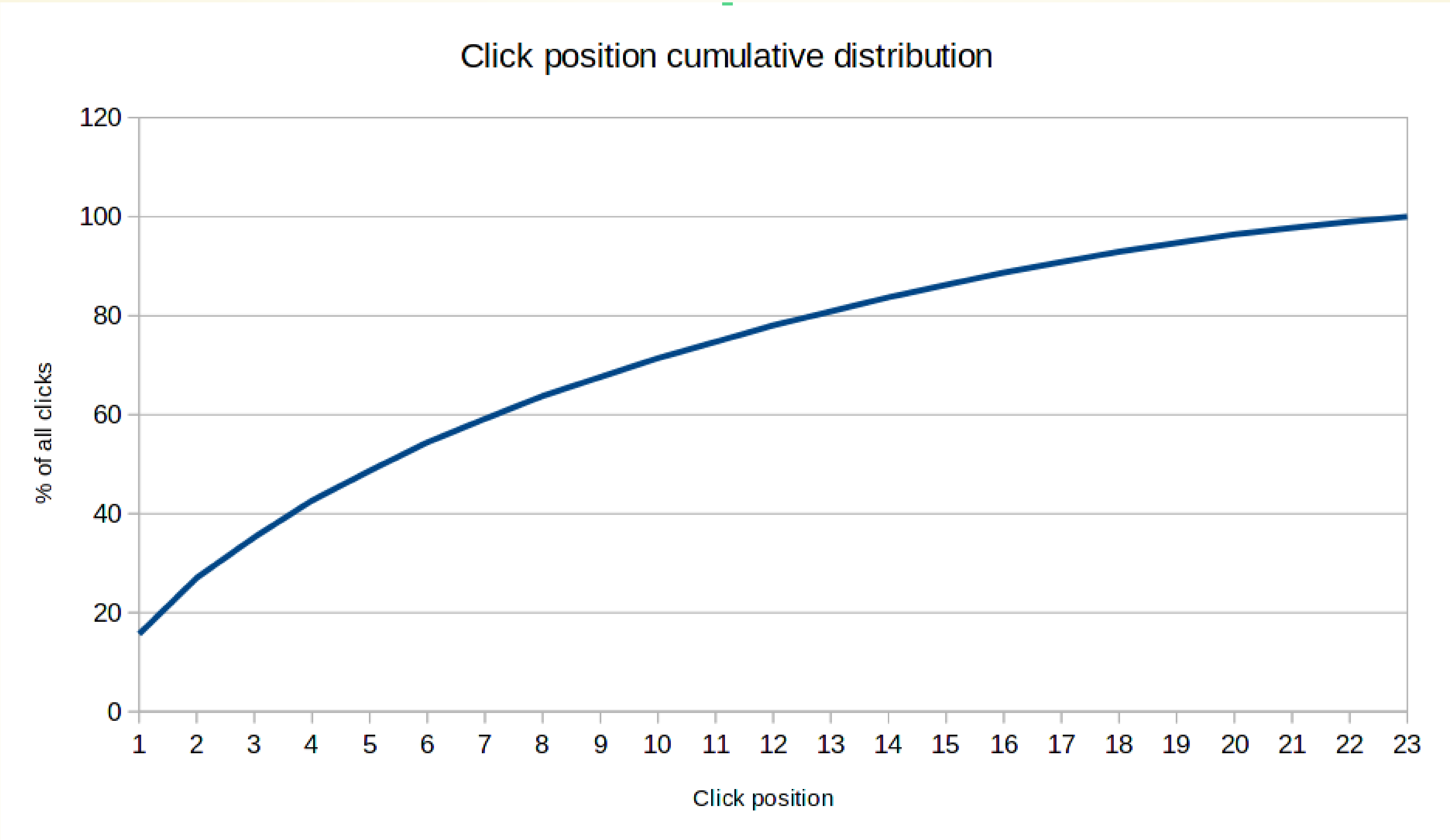
TRUE PRODUCTION STORY



POSITION BIAS

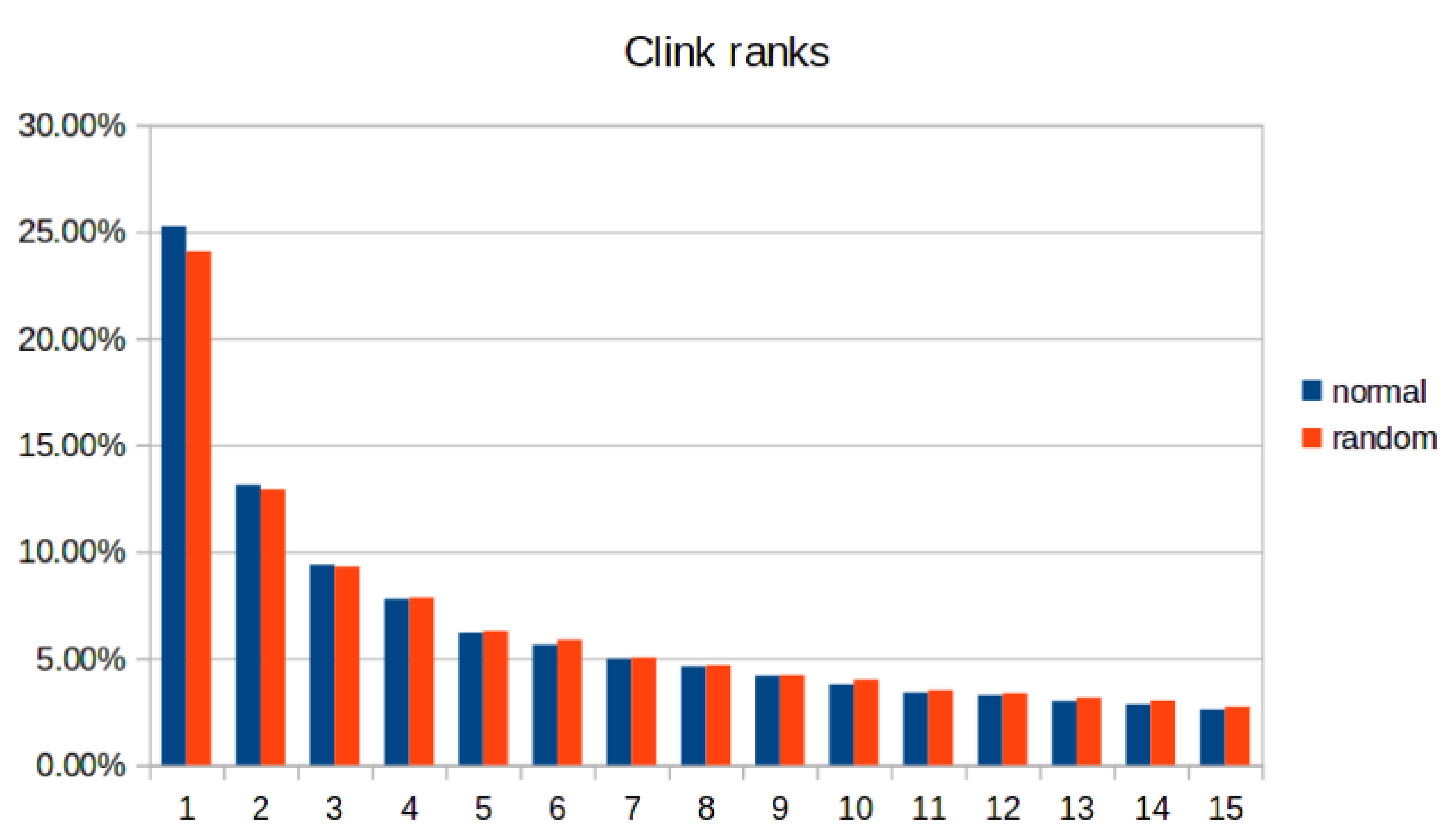


POSITION BIAS



POSITION BIAS: EXPERIMENT

default ordering vs random shuffling a/b test




MODEL BIAS
POSITION BIAS
Me with XGBoost

ESTIMATE THE BIAS?

observed relevancy = bias + true relevancy

1. Estimate the bias
2. Subtract the bias

   ...

3. ~~PROFIT~~ true relevancy

IPW: INVERSE PROPENSITY WEIGHTING

Wang X. - Position bias estimation for unbiased learning to rank in personal search

- Query-level Propensity:
  - downsample queries where avg(click_rank) is too high
- Document-level Propensity:
  - lower weight for top clicks

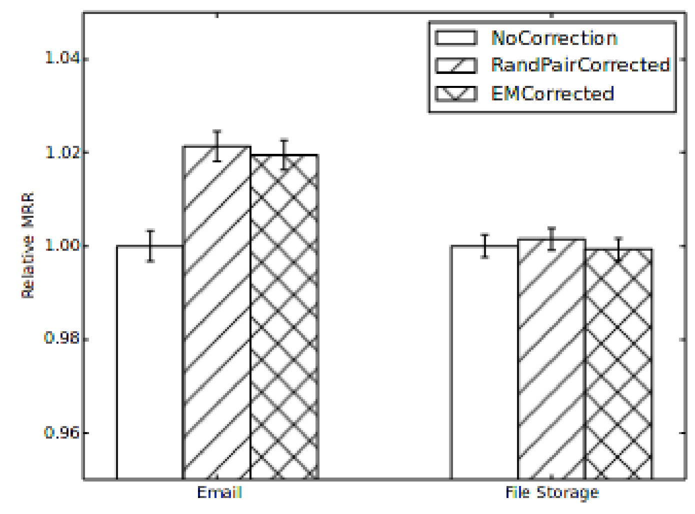


Figure 3: Ranking effectiveness comparison of pairwise approaches with different position bias correction methods.

DOCUMENT LEVEL PROPENSITY

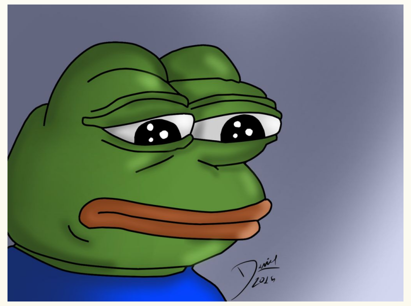
Query: socks

- "Star wars socks" on position 1: 50 clicks
- "Star wars socks" on position 5: 20 clicks
- "Star wars socks" on position 25: 3 clicks

IPW: PROS AND CONS

- Estimation depends on context
- More contexts - more data needed
- Improper estimation may produce another bias
- Need for shuffling

You're not google!



CURSE OF OFFLINE LEARNING

- production training data is biased
- de-biasing is hard

do we really need to stay offline?

REINFORCEMENT LEARNING

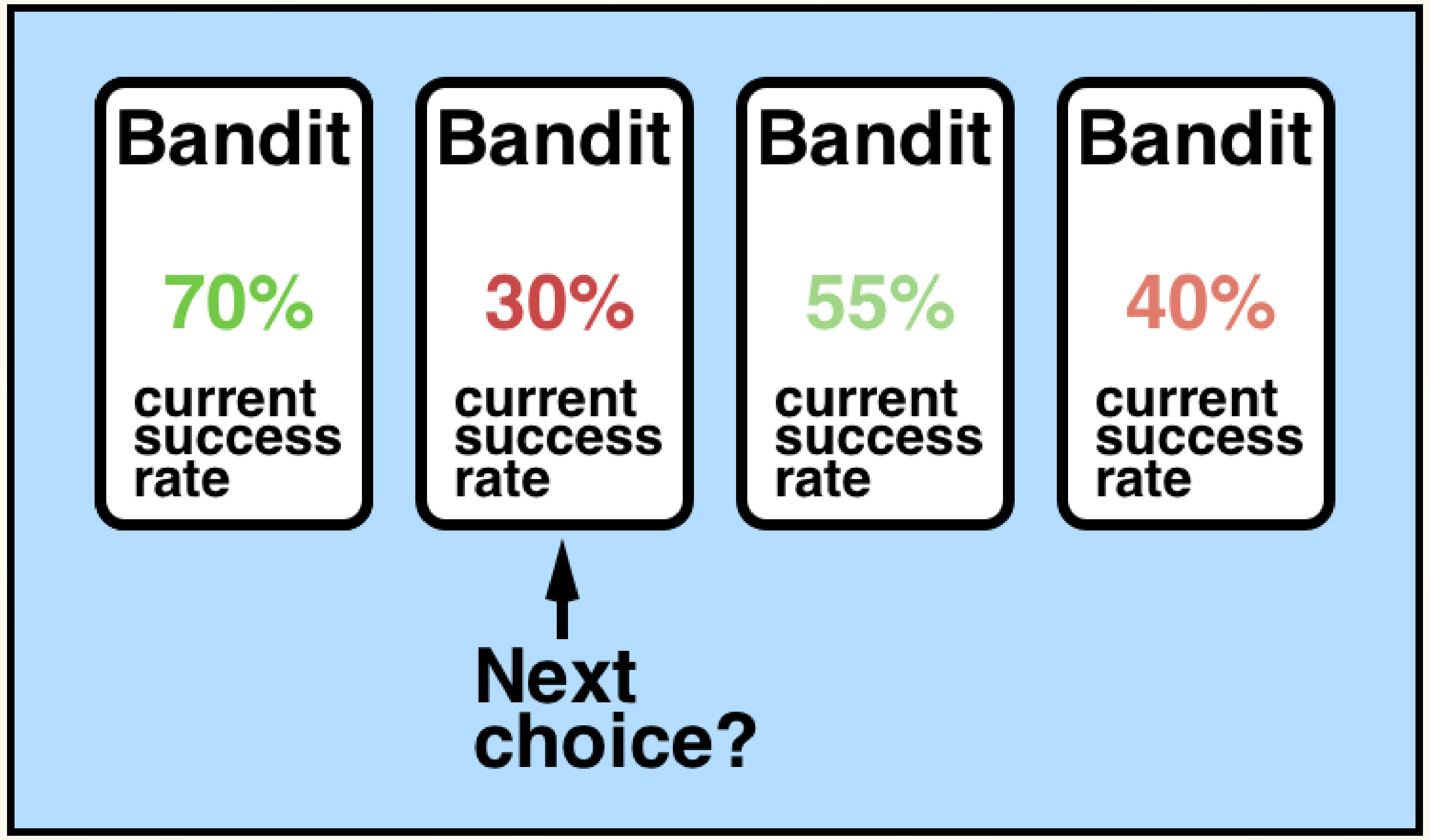
- Explore: poke real people, observe reaction
  - if you rank by popularity, noone knows what would happen if ranked by CTR
- Exploit: use it for better results
- Goal: minimize the **regret** / maximize the **reward**

MULTI-ARMED BANDITS

MULTI-ARMED BANDITS

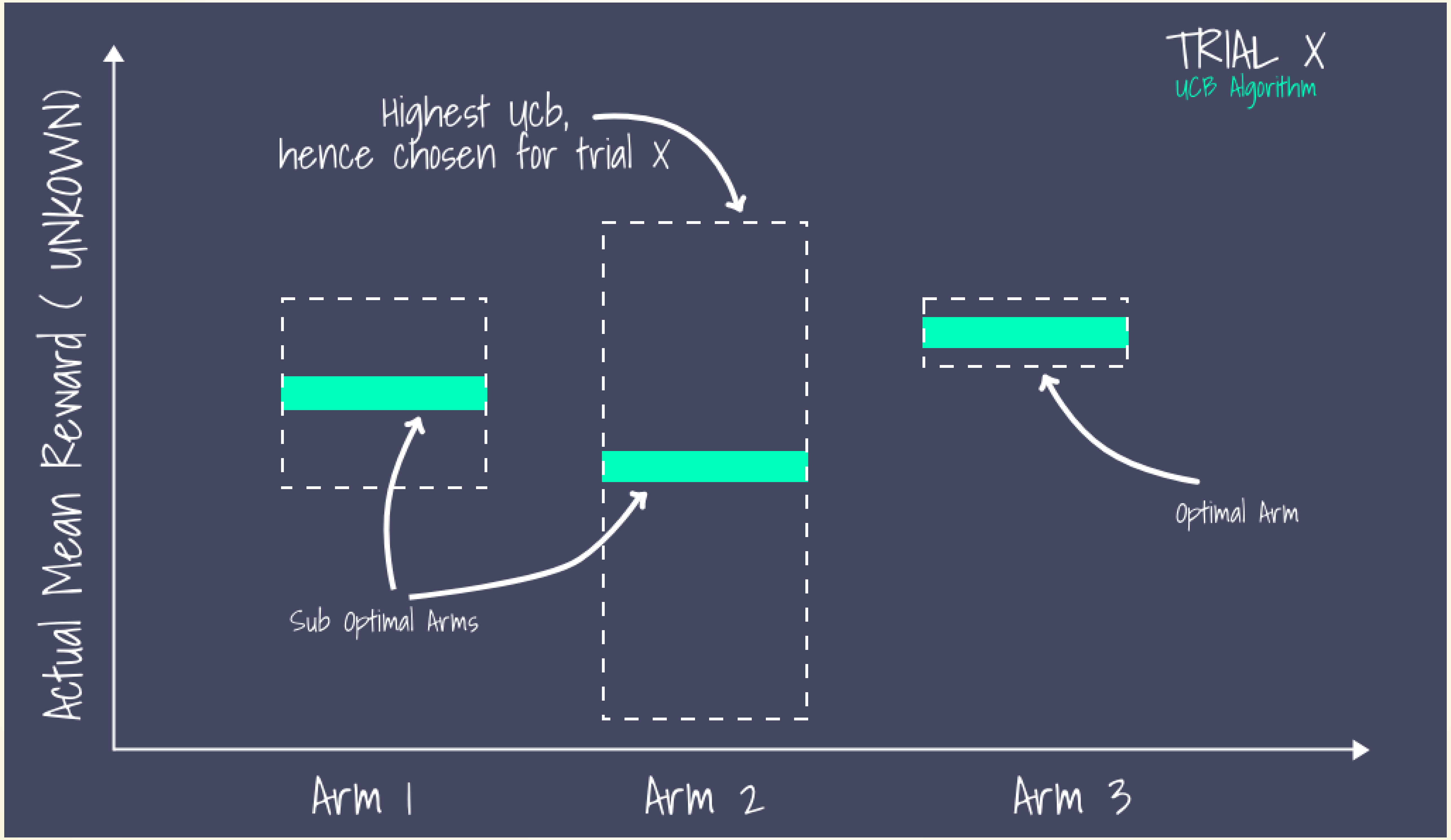
- Slot machines with unknown win probability
- Exploit: maximise the $$$ won
- Explore: which one has highest win chance?

MULTI-ARMED BANDITS



- Greedy: choose arm with max success rate
- ε-greedy: P(ε) = random, P(1-ε) = greedy
- LinUCB: choose arm with Upper Confidence Bound

LIN-UCB



TRIAL X
UCB Algorithm
Highest ucb,
hence chosen for trial X
Actual Mean Reward (UNKOWN)
Sub Optimal Arms
Optimal Arm
Arm 1
Arm 2
Arm 3

LIN-UCB

- Choose the arm with the highest optimistic reward
- Won? UCB goes up, eat cookie

- Lost? UCB goes down
  => Other arm on next iteration

LIN-UCB MEETS RANKING

What is an arm?

- Document?

  - Works great for small datasets like news articles

  - What if item set is large?

- Feature?

  - How rank influences reward?

CASCADE-LIN-UCB

S. Zhong: Cascading Bandits for Large-Scale Recommendation Problems

$$\mathbf{score} = c_1 * f_1 + \ldots + c_N * f_N$$

-

- Arm: feature value (# of clicks, purchases)
- Explore: increase-decrease feature weight

  - rank by score, see how it affects reward

- Reward: Increase $$$
- Choose next arm via LinUCB

RL: A/B TEST ON STEROIDS

- A/B test when segments are dynamic
- Compare previous iteration with candidate
- Reward goes up = success

CASCADE-LIN-UCB IN PRACTICE

- Reward doesn't depend on score
  - Combine multiple business objectives!
  - 50% CTR + 50% Conversion
  - Click probability?
- Linear model: easy to explain
- Requires a LOT of time to converge
- Linear model: tough with non-linearities

GOING NDCG

H. Oosterhuis: Differentiable Unbiased Online Learning to Rank

- Reward ~= NDCG
- NDCG is not smooth => use pairwise loss

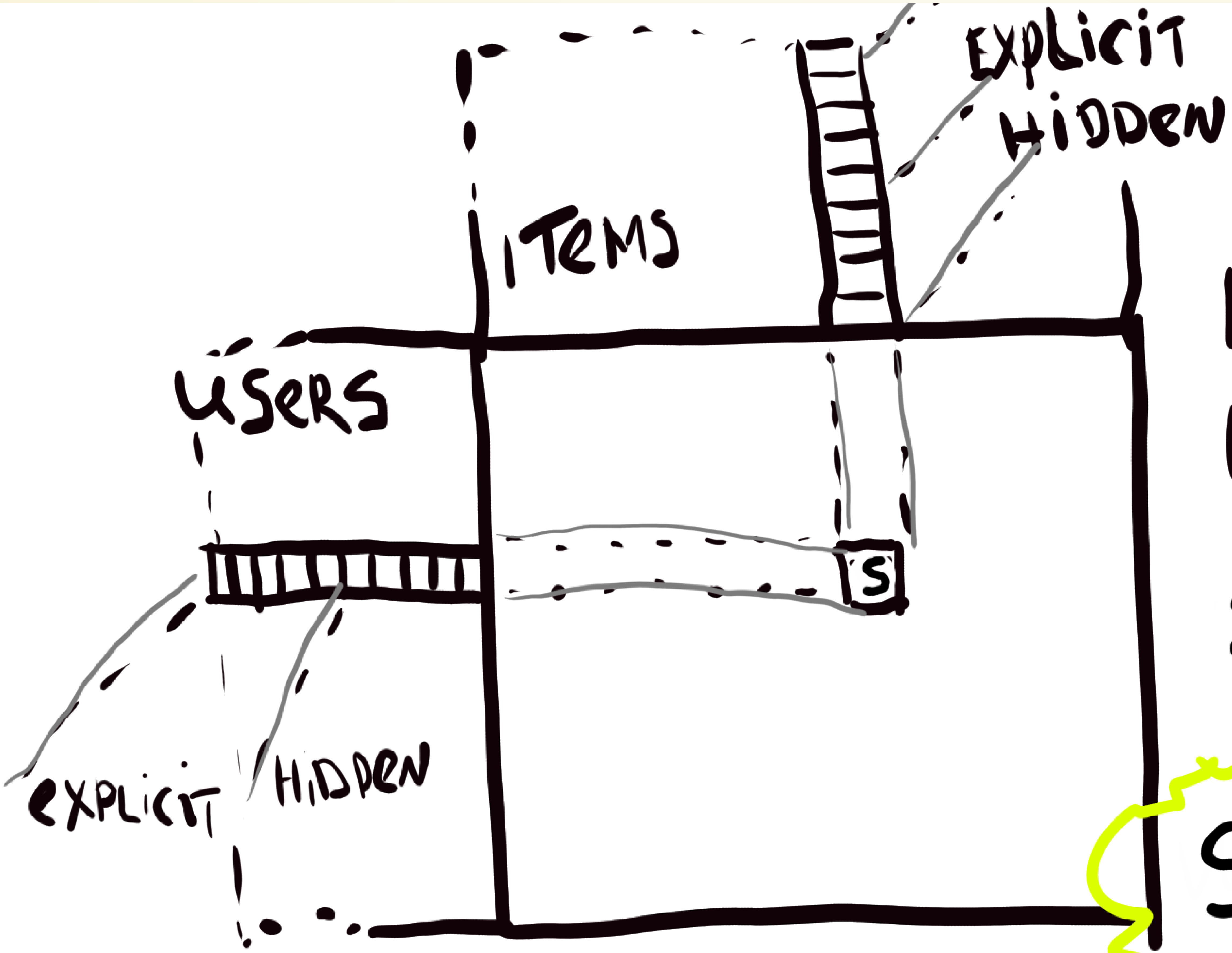
PDGD IN PRACTICE

- Much faster to converge
  - weights are updated on each interaction!
- No way to plug whatever reward :(
  - Focused on pairwise loss => better NDCG
- Linear model: easy to explain
- Linear model: tough with non-linearities

COLLABORATIVE RANKING

S. Gupta @ Amazon: CPR: Collaborative Pairwise Ranking for Online List Recommendations

- User/item explitict and latent features
- Weight ~ pairwise loss
- Like MF SVD/ALS for recommendations



ITEMS: $K \times N$

USERS: $K \times M$

SCORES: $N \times M$

$S = \text{item} \cdot \text{user}$

CPR IN PRACTICE

- Complex math: tricky to update in realtime
- Focus on pairwise loss
- Good luck with explainability


WELCOME TO THE LEARN-TO-RANK
PLEASE FOLLOW ME

CURRENT STATE OF AFFAIRS

- Heavy NDCG focus
- NDCG is not a business metric!

  - There were cases when NDCG goes up, but CTR & Conversion goes down

  - What if people click higher but less frequent?

- Focusing on AOV/Conv/CTR is hard :(

TOWARDS A SMARTER RANKING

- Which reward is important to you?

- Ensemble of models: combine strengths!

LINKS

- Slides: dfdx.me/slides/reinforcement-learning-in-search
- Me: linkedin.com/in/romangrebennikov/

QUESTIONS?

Derp Learning	100	200	300	400	500
Holywar	100	200	300	400	500
Code of Conduct	100	200	300	400	500
Cargo cult	100	200	300	400	500
TensorFlow	100	200	300	400	500
Haskell	100	200	300	400	500

a slide by @nikitonsky