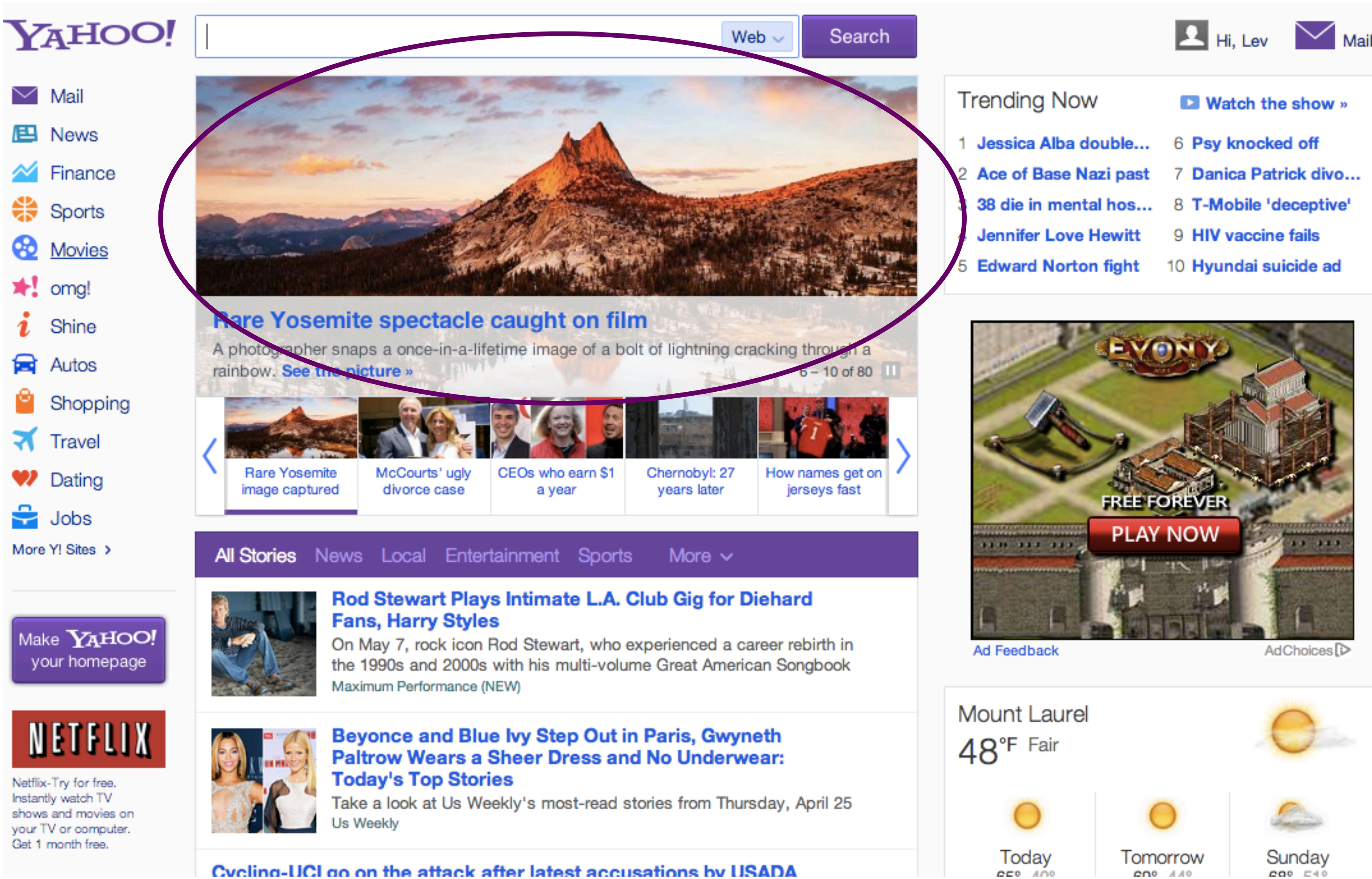
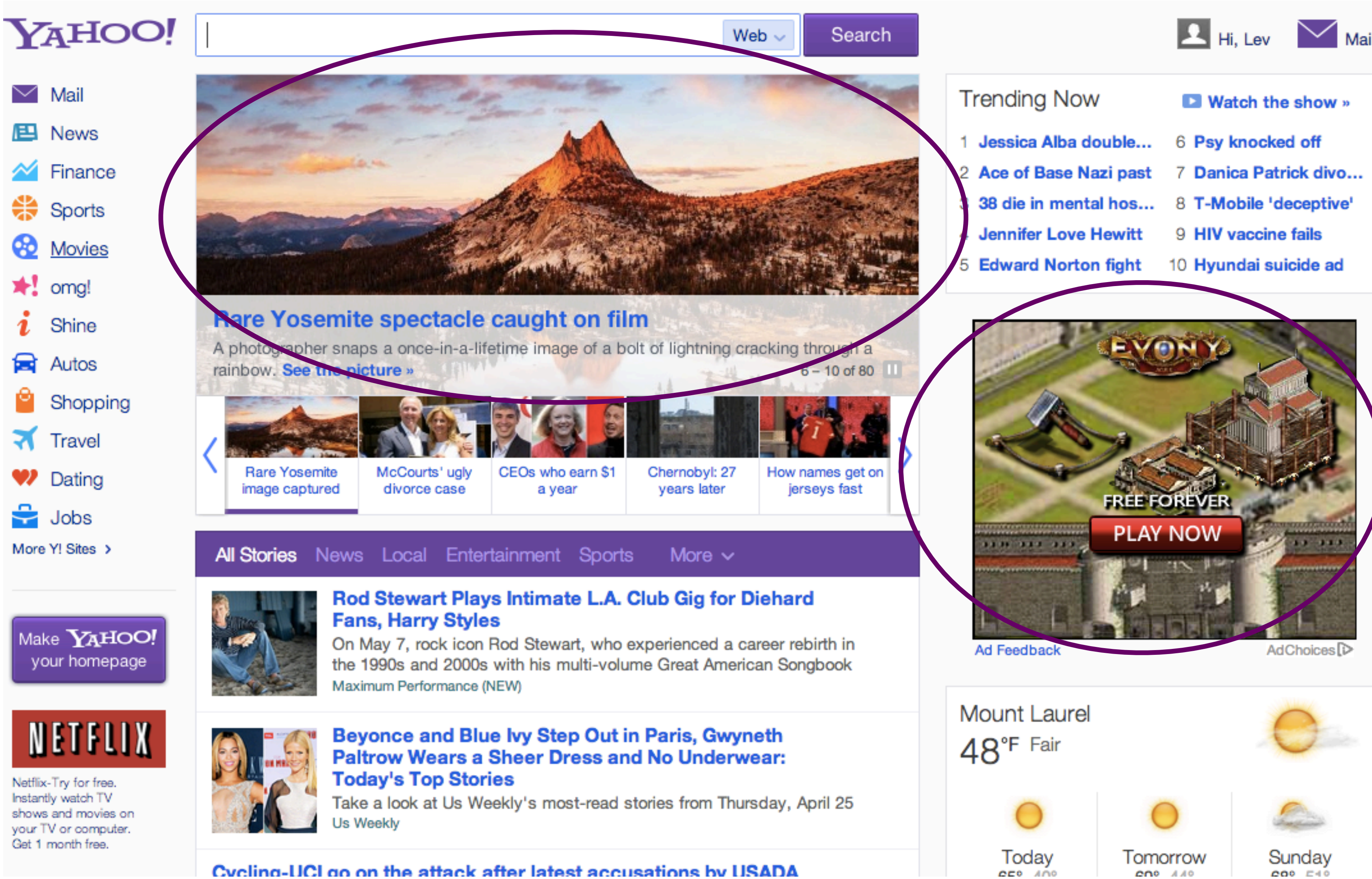
CONTEXTUAL BANDIT ALGORITHMS FOR INTERNET-SCALE APPLICATIONS

LEV REYZIN

UIC MATHEMATICS

ASA: RECENT ADVANCES IN MACHINE LEARNING


YAHOO!
Web
Search
Hi, Lev
Mail
Mail
News
Finance
Sports
Movies
omg!
Shine
Autos
Shopping
Travel
Dating
Jobs
More Y! Sites
Make YAHOO! your homepage
NETFLIX
Netflix-Try for free. Instantly watch TV shows and movies on your TV or computer. Get 1 month free.
Rare Yosemite spectacle caught on film
A photographer snaps a once-in-a-lifetime image of a bolt of lightning cracking through a rainbow. See the picture »
6 – 10 of 80
Rare Yosemite image captured
McCourts' ugly divorce case
CEOs who earn $1 a year
Chernobyl: 27 years later
How names get on jerseys fast
All Stories
News
Local
Entertainment
Sports
More
Rod Stewart Plays Intimate L.A. Club Gig for Diehard Fans, Harry Styles
On May 7, rock icon Rod Stewart, who experienced a career rebirth in the 1990s and 2000s with his multi-volume Great American Songbook
Maximum Performance (NEW)
Beyonce and Blue Ivy Step Out in Paris, Gwyneth Paltrow Wears a Sheer Dress and No Underwear: Today's Top Stories
Take a look at Us Weekly's most-read stories from Thursday, April 25
Us Weekly
Trending Now
Watch the show »
1 Jessica Alba double...
2 Ace of Base Nazi past
38 die in mental hos...
Jennifer Love Hewitt
5 Edward Norton fight
6 Psy knocked off
7 Danica Patrick divo...
8 T-Mobile 'deceptive'
9 HIV vaccine fails
10 Hyundai suicide ad
EVONY
FREE FOREVER
PLAY NOW
Ad Feedback
AdChoices
Mount Laurel
48°F Fair
Today
Tomorrow
Sunday


YAHOO!
Web
Search
Hi, Lev
Mail
Mail
News
Finance
Sports
Movies
omg!
Shine
Autos
Shopping
Travel
Dating
Jobs
More Y! Sites
Make YAHOO! your homepage
NETFLIX
Netflix-Try for free. Instantly watch TV shows and movies on your TV or computer. Get 1 month free.
Rare Yosemite spectacle caught on film
A photographer snaps a once-in-a-lifetime image of a bolt of lightning cracking through a rainbow. See the picture »
6 – 10 of 80
Rare Yosemite image captured
McCourts' ugly divorce case
CEOs who earn $1 a year
Chernobyl: 27 years later
How names get on jerseys fast
All Stories
News
Local
Entertainment
Sports
More
Rod Stewart Plays Intimate L.A. Club Gig for Diehard Fans, Harry Styles
On May 7, rock icon Rod Stewart, who experienced a career rebirth in the 1990s and 2000s with his multi-volume Great American Songbook
Maximum Performance (NEW)
Beyonce and Blue Ivy Step Out in Paris, Gwyneth Paltrow Wears a Sheer Dress and No Underwear: Today's Top Stories
Take a look at Us Weekly's most-read stories from Thursday, April 25
Us Weekly
Trending Now
Watch the show »
1 Jessica Alba double...
2 Ace of Base Nazi past
38 die in mental hos...
Jennifer Love Hewitt
5 Edward Norton fight
6 Psy knocked off
7 Danica Patrick divo...
8 T-Mobile 'deceptive'
9 HIV vaccine fails
10 Hyundai suicide ad
EVONY
FREE FOREVER
PLAY NOW
Ad Feedback
AdChoices
Mount Laurel
48°F Fair
Today
Tomorrow
Sunday

THE WALL STREET JOURNAL.
U.S. EDITION Friday, April 26, 2013 As of 8:52 AM EDT

LATEST HEADLINES — MORE HEADLINES »

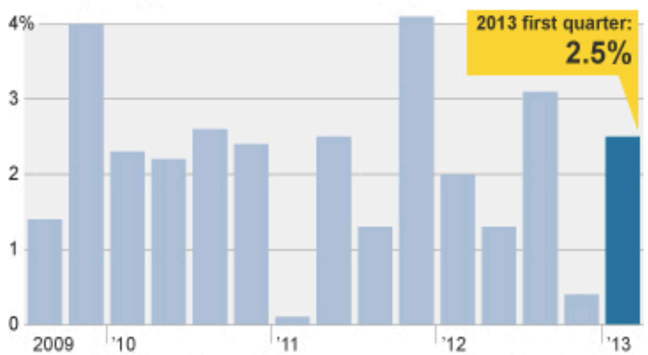
7:48 AM **Shiseido Sees Profit Next Year**

First-Quarter Growth, at 2.5%, Misses Expectations

The nation's gross domestic product, a measure of all goods and services produced in the economy, advanced at a 2.5% annual rate between January and March, the Commerce Department said. 17 min ago

U.S. GDP

Quarterly change at an annualized rate, adjusted for inflation



Source: Commerce Department — The Wall Street Journal

Disappointing GDP Data Weigh on Futures

Mansion: From Oil, a Land Rush | Einstein Proved Right--Again | NFL Draft

Markets ›

Overview | U.S. | Europe | Asia | FX | Rates | Futures

	LAST	CHG	%CHG
DJIA	14700.80	+24.50	0.17%
Nasdaq	3289.99	+20.33	0.62%
FTSE 100	6401.63	-40.96	0.64%
Nikkei 225	13884.13	-41.95	0.30%
Crude Oil	93.50	-0.14	0.15%
Gold	1470.50	+8.50	0.58%

RANGE: 1 DAY

4/26/13 8:44 AM EDT

Market Data | MoneyBeat | Watchlist | Portfolio | Customize

.wsj.com/home-page#

SERVING CONTENT TO USERS

IP address, browser properties, etc.

result (ie. ad, news story)

click or not

SERVING CONTENT TO USERS

IP address, browser properties, etc.

result (ie. ad, news story)

click or not

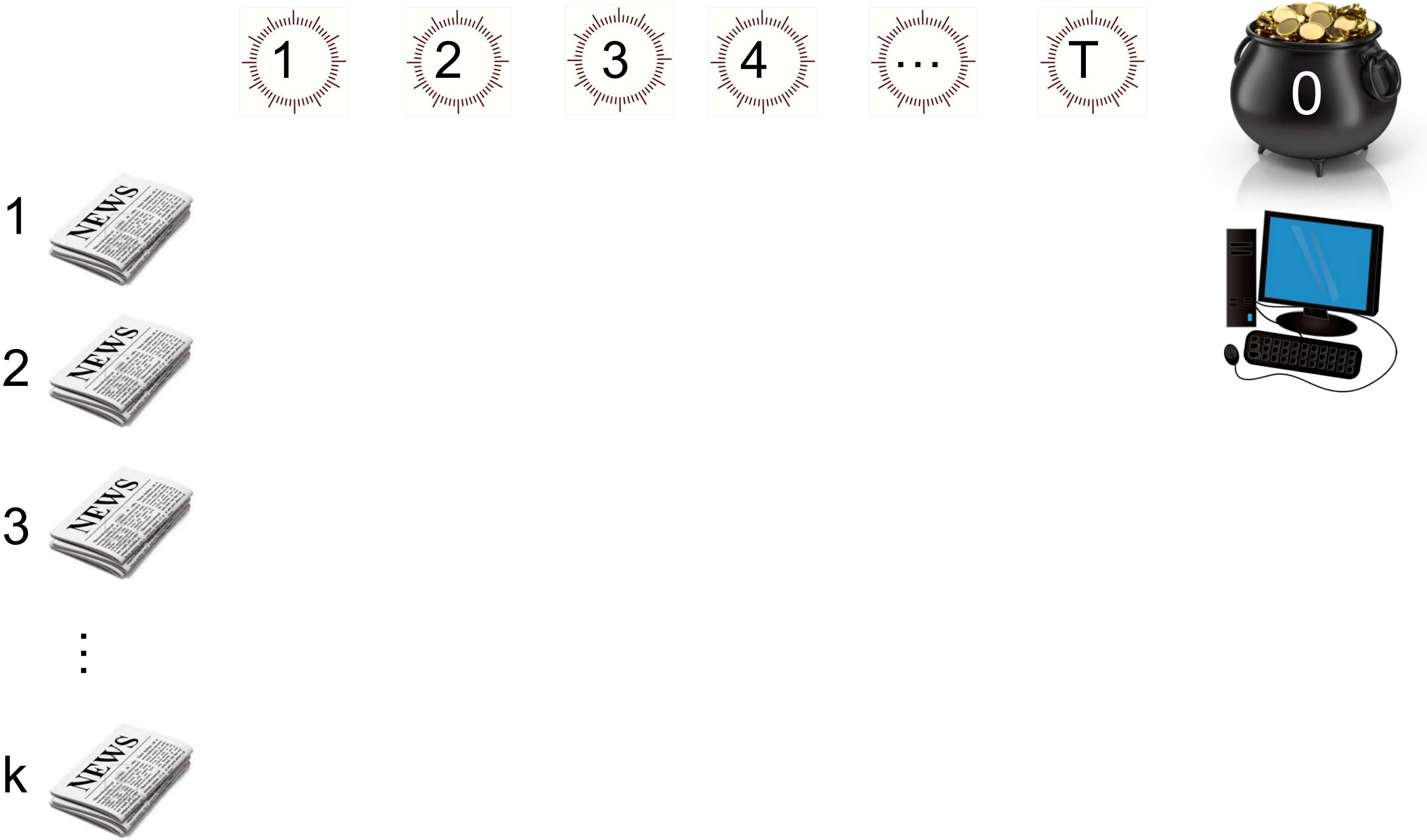
SERVING CONTENT TO USERS

context x_t

action j_t

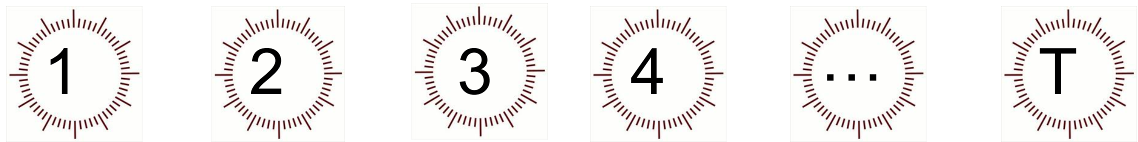
reward $r_{j_t}(t)$

MULTIARMED BANDITS
[ROBBINS '52]



MULTIARMED BANDITS

[ROBBINS ’52]



1 

2  click

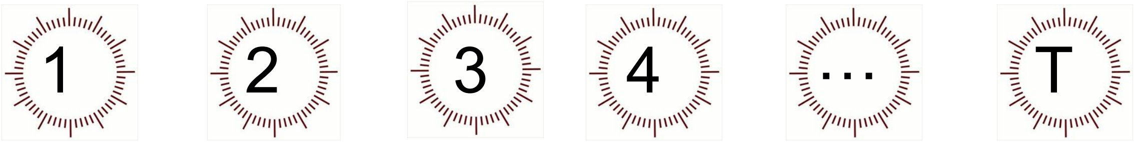
3 

⋮

k 

MULTIARMED BANDITS

[ROBBINS '52]



1 

2 

click

3 

click

⋮

k 

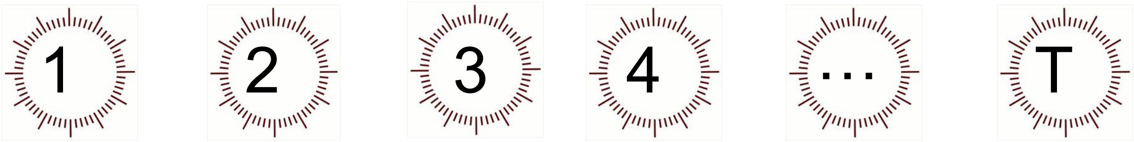
MULTIARMED BANDITS

[ROBBINS '52]

	1	2	3	4	…	T
1 						
2 	click		no			
3 		click				
⋮						
k 						

MULTIARMED BANDITS

[ROBBINS '52]



click

1 

2  click no

3  click

⋮

k  click

MULTIARMED BANDITS
[ROBBINS '52]

1  → 0.5T

2  → 0.2T

3  → 0.33T

⋮

k  → 0.1T

MULTIARMED BANDITS
[ROBBINS ’52]

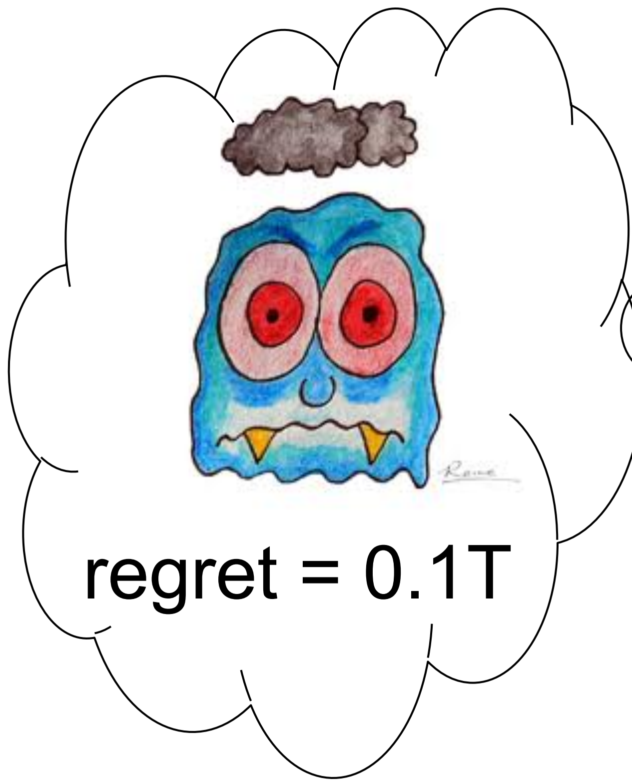
1  → **0.5T**

2  → 0.2T

3  → 0.33T

⋮

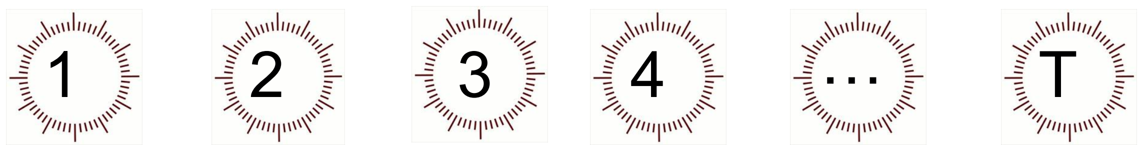
k  → 0.1T



CONTEXTUAL BANDITS

[AUER-CESABIANCHI-FREUND-SCHAPIRE ’02]

context:



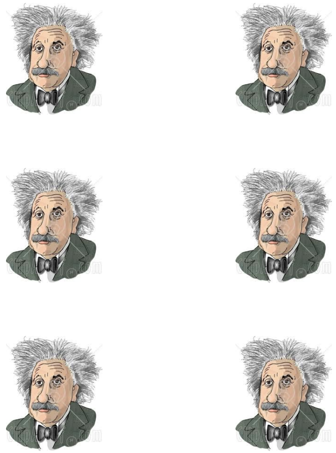
1 

2 

3 

⋮

k 



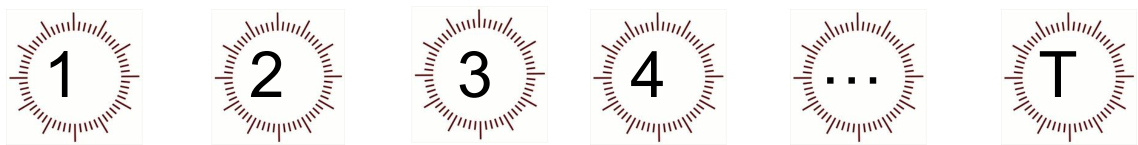
N experts/policies/functions

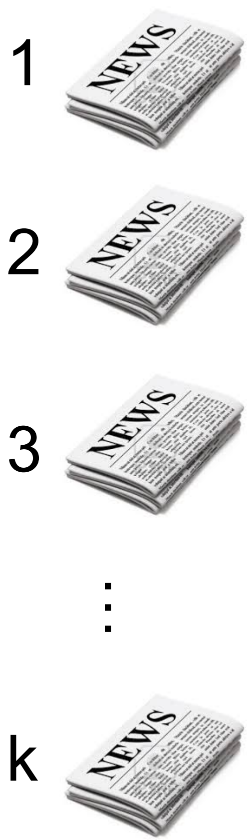
think of N >> K

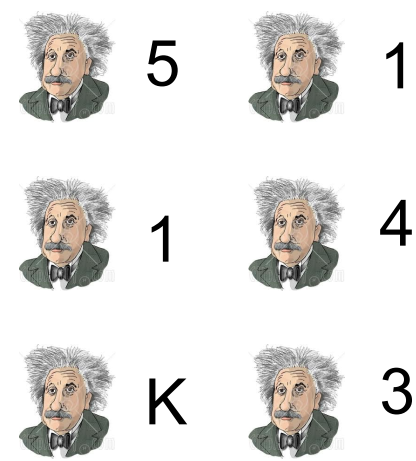
CONTEXTUAL BANDITS

[AUER-CESABIANCHI-FREUND-SCHAPIRE ’02]

context: x_1







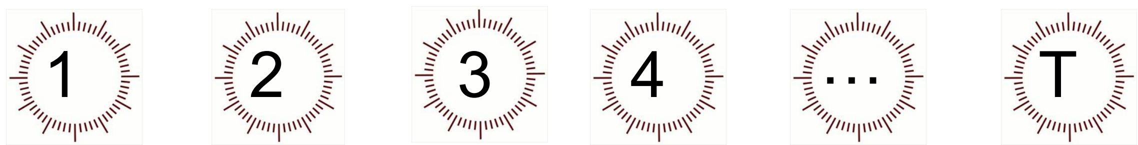
N experts/policies/functions

think of N >> K

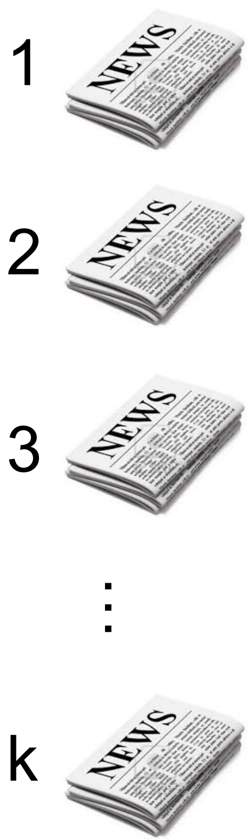
CONTEXTUAL BANDITS

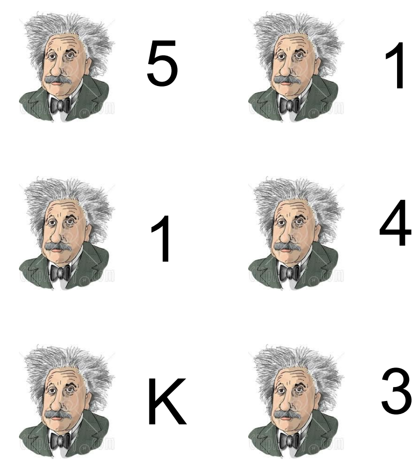
[AUER-CESABIANCHI-FREUND-SCHAPIRE ’02]

context: x_1



click

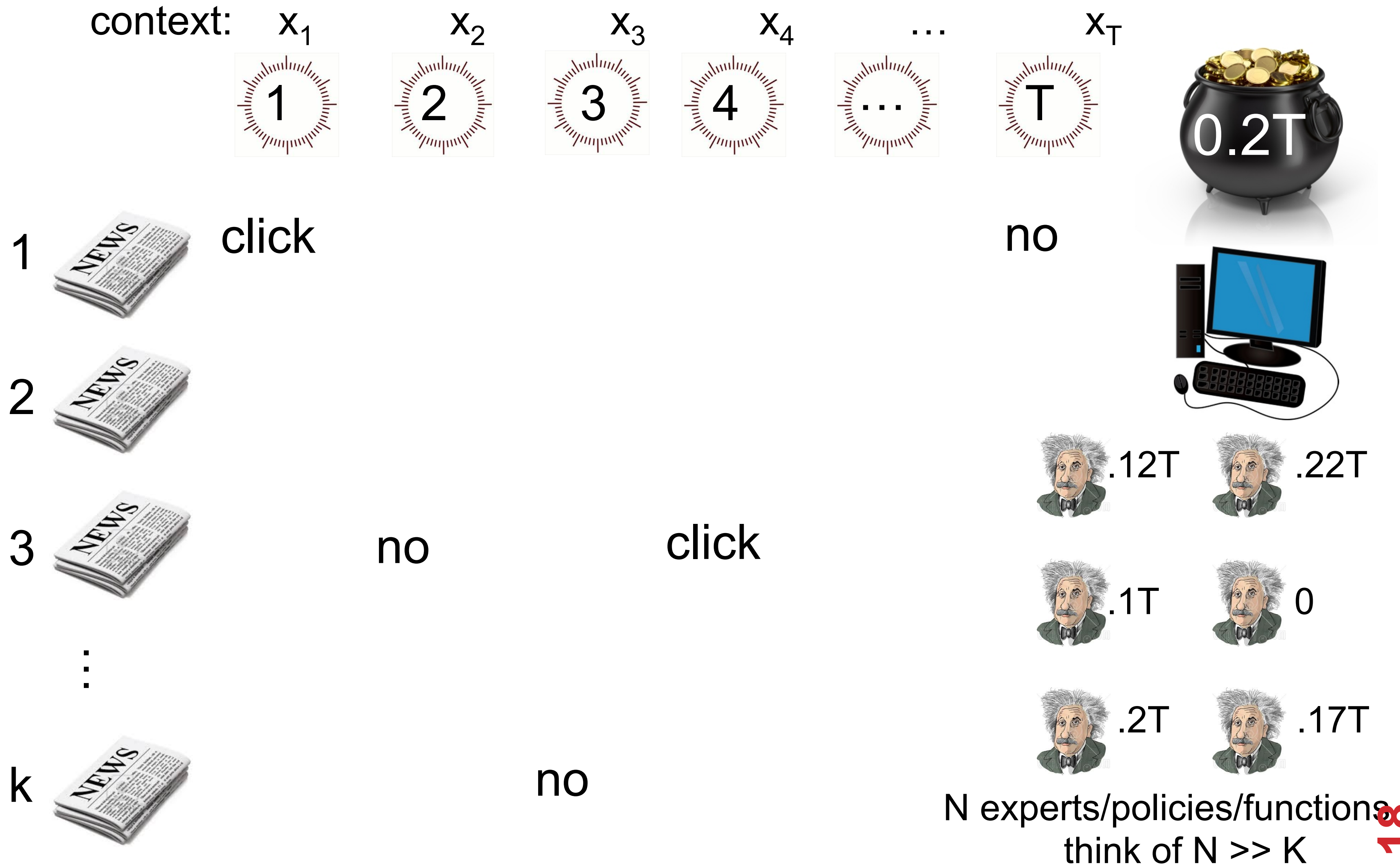




N experts/policies/functions
think of N >> K

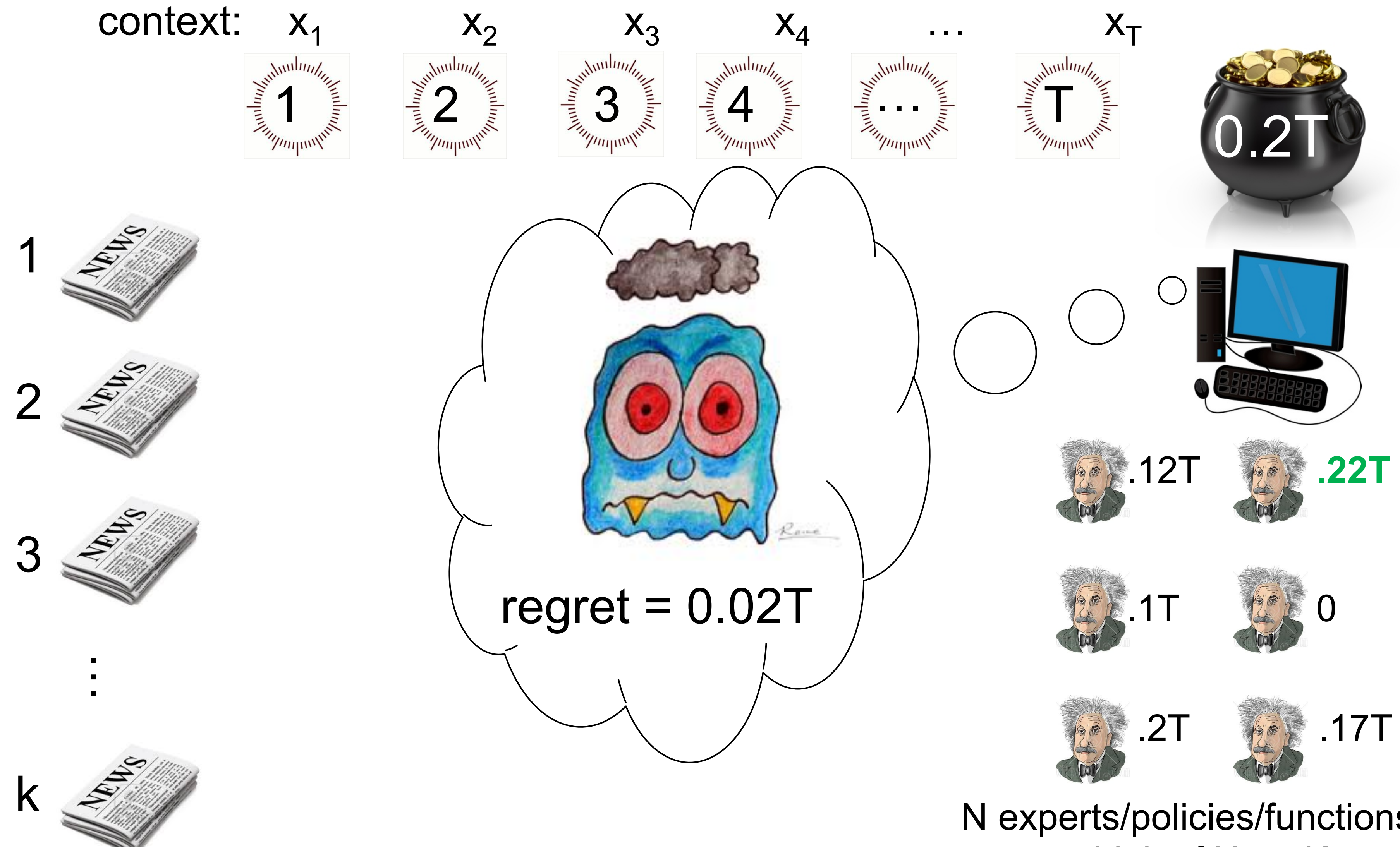
CONTEXTUAL BANDITS

[AUER-CESABIANCHI-FREUND-SCHAPIRE '02]



CONTEXTUAL BANDITS

[AUER-CESABIANCHI-FREUND-SCHAPIRE ’02]

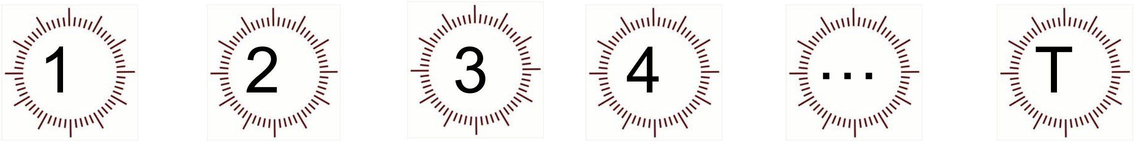


N experts/policies/functions
think of N >> K

CONTEXTUAL BANDITS

[AUER-CESABIANCHI-FREUND-SCHAPIRE ’02]

context: x_1 x_2 x_3 x_4 … x_T

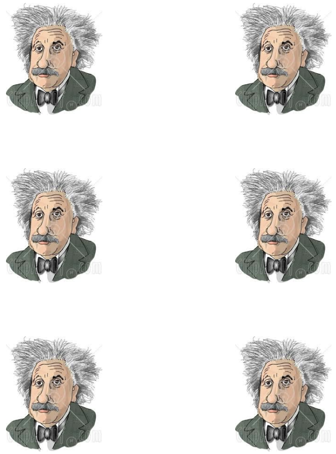


1
2
3
⋮
k

←

the clicks can come i.i.d. from a distribution or be arbitrary
stochastic / adversarial

The experts can be present or not. →
contextual / non-contextual



BANDITS

Harder than supervised learning:

In the bandit setting we do not know the rewards of actions not taken.

Many applications

Ad auctions, medicine, finance, …

Exploration/Exploitation

Can exploit expert/article you've learned to be good.

Can explore expert/article you're not sure about.

EPSILON-FIRST

Rough idea of ε-first (or ε-greedy): act randomly for ε rounds, then go with best (arm or expert).

Rough analysis: even for 2 arms, we suffer regret ε+ (T-ε)/(ε$^{1/2}$).

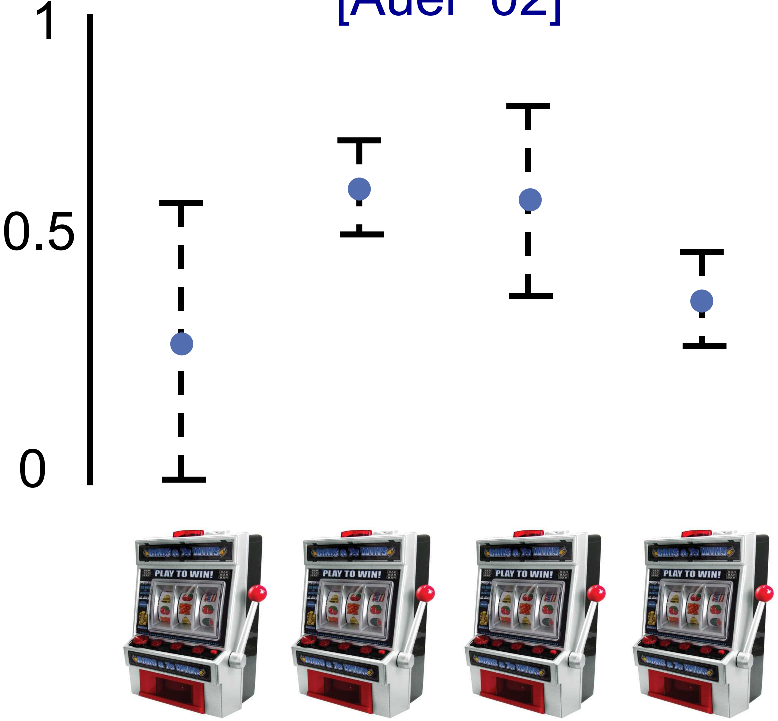
- ε≈ T$^{2/3}$ is optimal tradeoff, gives regret ≈ T$^{2/3}$

But actually O(T)$^{1/2}$ regret is possible!

TRADITIONAL BANDIT ALGORITHMS

UCB

[Auer '02]

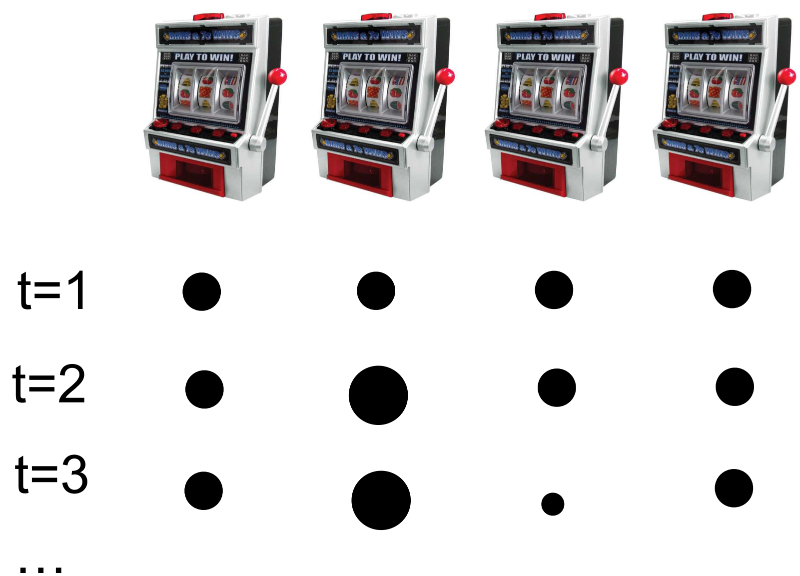


Algorithm: at every time step

1) pull arm with highest UCB
2) update confidence bound of the arm pulled.

EXP3 / EW

[Littlestone-Warmuth '94]
[Auer et al. '02]



Algorithm: at every time step

1) sample from distribution defined by weights (mixed w/ uniform)
2) update weights "exponentially"

UCB VS EXP3: A COMPARISON

UCB

[AUER '02]

- **Pros**
  - Optimal for the stochastic setting.
  - Succeeds with high probability.
- **Cons**
  - Does not work in the adversarial setting.
  - Is not optimal in the contextual setting.

EXP3 & FRIENDS

[ACFS '02]

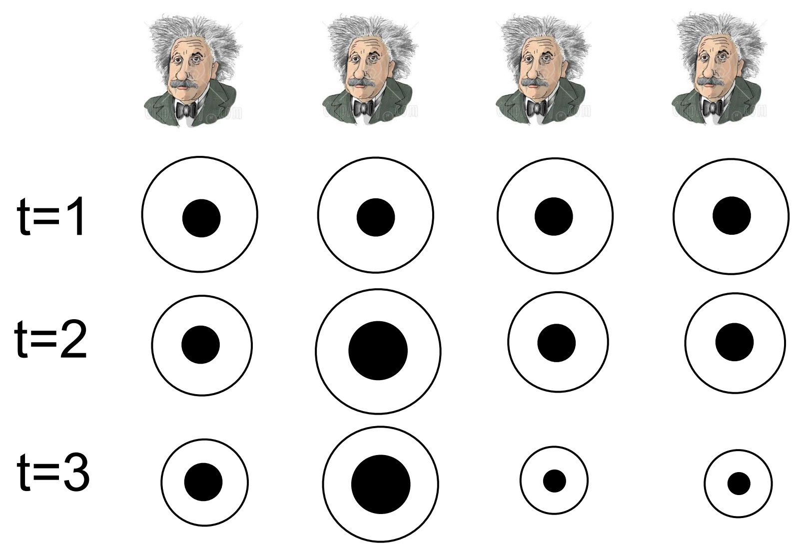
- **Pros**
  - Optimal for both the adversarial and stochastic settings.
  - Can be made to work in the contextual setting
- **Cons**
  - Does not succeed with high probability in the contextual setting (only in expectation).

EXP4.P

Main Theorem [Beygelzimer-Langford-Li-**R**-Schapire '11]:
For any δ>0, with probability at > 1-δ, EXP4P has "optimal" regret in the adversarial contextual bandit setting.

key insights
on top of UCB/ EXP

1) exponential weights and upper confidence bounds "stack"
2) generalized Bernstein's inequality for martingales



IDEAS BEHIND EXP4.P
(ALL APPEARED IN PREVIOUS ALGORITHMS)

exponential weights

- keep a weight on each expert that drops exponentially in the expert's (estimated) performance

upper confidence bounds

- use an upper confidence bound on each expert's estimated reward

ensuring exploration

- make sure each action is taken with some minimum probability

importance weighting

- give rare events more importance to keep estimates unbiased

Exponential Weight Algorithm for Exploration and Exploitation with Experts

Exp4.P [Beygelzimer, Langford, Li, R, Schapire ’10]

Initialization: $\forall \pi \in \Pi : w_t(\pi) = 1$

For each $t = 1, 2, \ldots$:

1. Observe x_t and let for $a = 1, \ldots, K$
$$p_t(a) = (1 - Kp_{\min}) \frac{\sum_\pi \mathbf{1}[\pi(x_t) = a]\, w_t(\pi)}{\sum_\pi w_t(\pi)} + p_{\min},$$
where $p_{\min} = \sqrt{\frac{\ln |\Pi|}{KT}}$.
2. Draw a_t from p_t, and observe reward $r_t(a_t)$.
3. Update for each $\pi \in \Pi$
$$w_{t+1}(\pi) = w_t(\pi) \exp\left(\frac{p_{\min}}{2}\left(\mathbf{1}[\pi(x_t) = a_t] \frac{r_t(a_t)}{p_t(a_t)} + \frac{1}{p_t(\pi(x_t))}\sqrt{\frac{\ln N/\delta}{KT}}\right)\right)$$

EXP4.P IN PRACTICE

- Application – Yahoo! front page
- We chose a special policy class for which we could efficiently keep track of the weights.
  - Created 5 clusters, with users (at each time step) getting features based on their distances to clusters.
  - Policies mapped clusters to article (action) choices.
  - Ran on personalized news article recommendations for Yahoo! front page.
- We used a learning bucket on which we ran the algorithms and a deployment bucket on which we ran the greedy (best) learned policy.

Reported estimated (normalized) click-through rates on front page news. Over 41M user visits. 253 total articles. 21 candidate articles per visit.

	EXP4P	EXP4	ε-greedy
Learning eCTR	1.0525	1.0988	1.3829
Deployment eCTR	**1.6512**	1.5309	1.4290

Reported estimated (normalized) click-through rates on front page news. Over 41M user visits. 253 total articles. 21 candidate articles per visit.

	EXP4P	EXP4	ε-greedy
Learning eCTR	1.0525	1.0988	1.3829
Deployment eCTR	**1.6512**	1.5309	1.4290

Why does this work in practice?

HOPE FOR AN EFFICIENT ALGORITHM?

[DUDIK-HSU-KALE-KARAMPATZIAKIS-LANGFORD-R-ZHANG '11]

For EXP4P, the dependence on N in the regret is logarithmic.

this suggests

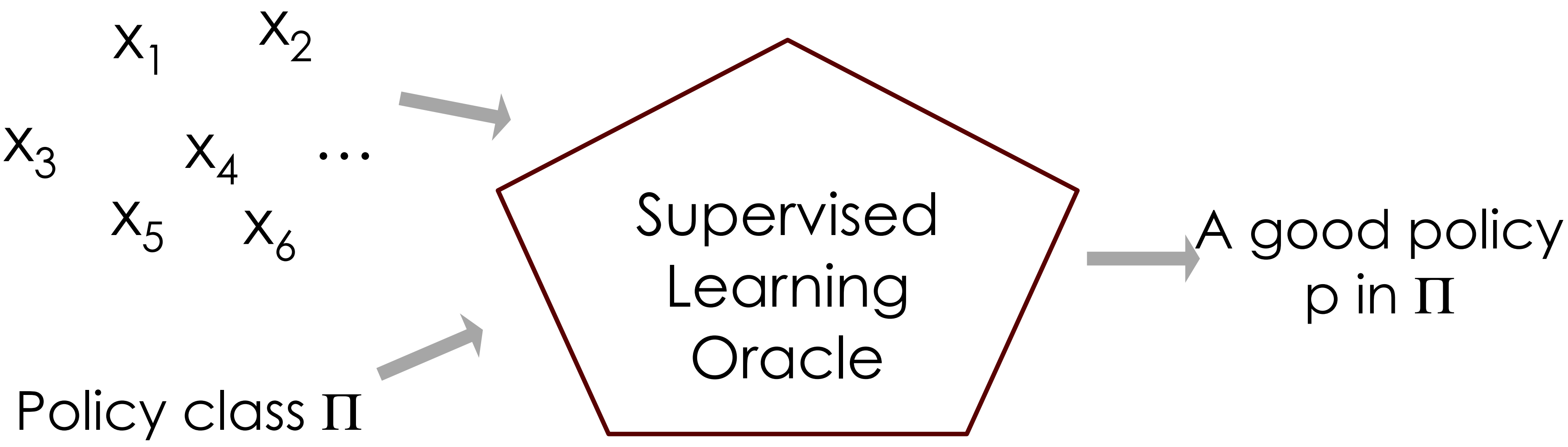
We could compete with a large, even super-polynomial number of policies! (e.g. $N=K^{100}$ becomes $10 \log^{1/2} K$ in the regret)

however

All known contextual bandit algorithms explicitly "keep track" of the N policies. Even worse, just reading in the N would take too long for large N.

Reduce to Supervised Learning!
(Idea from [Langford-Zhang '07])

- "Competing" with an exponentially large set of policies is commonplace in supervised learning.
- Recommendations of the policies/functions don't need to be explicitly read when the policy class has structure!

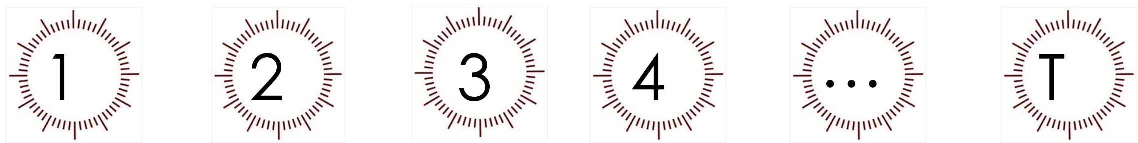


Reduce to Supervised Learning!

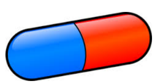
(Idea from [Langford-Zhang '07])

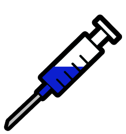
- "Competing" with an exponentially large set of policies is commonplace in supervised learning.
- Recommendations of the policies/functions don't need to be explicitly read when the policy class has structure!

context: x_1 x_2 x_3



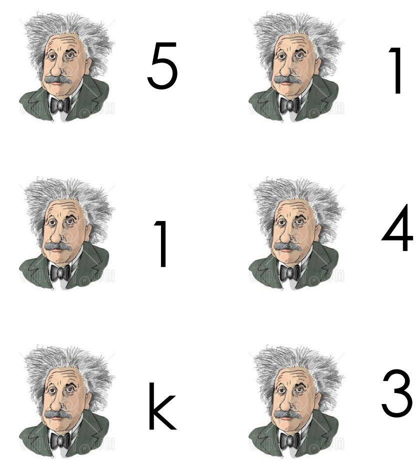
1  yes

2  no

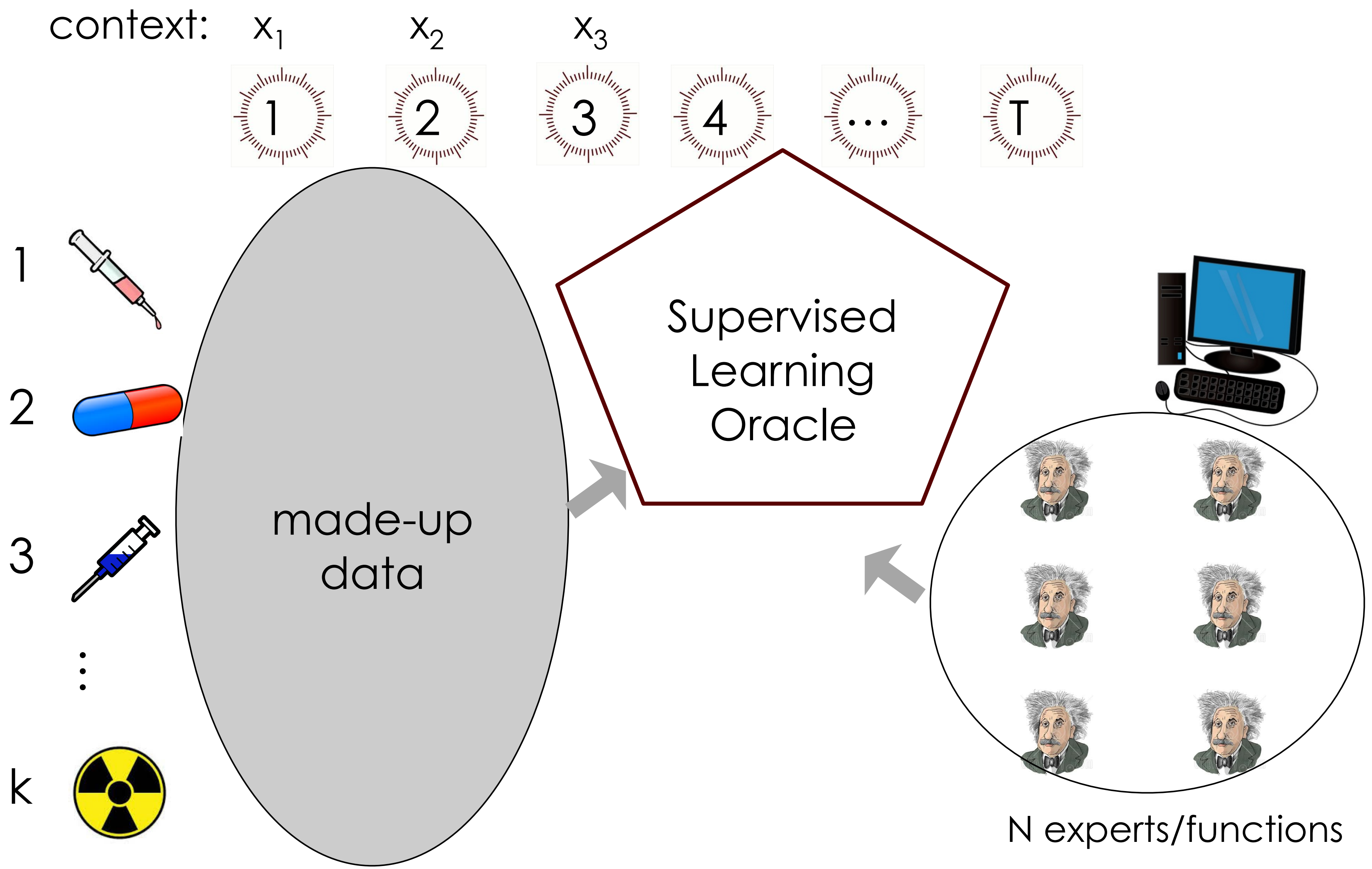
3 

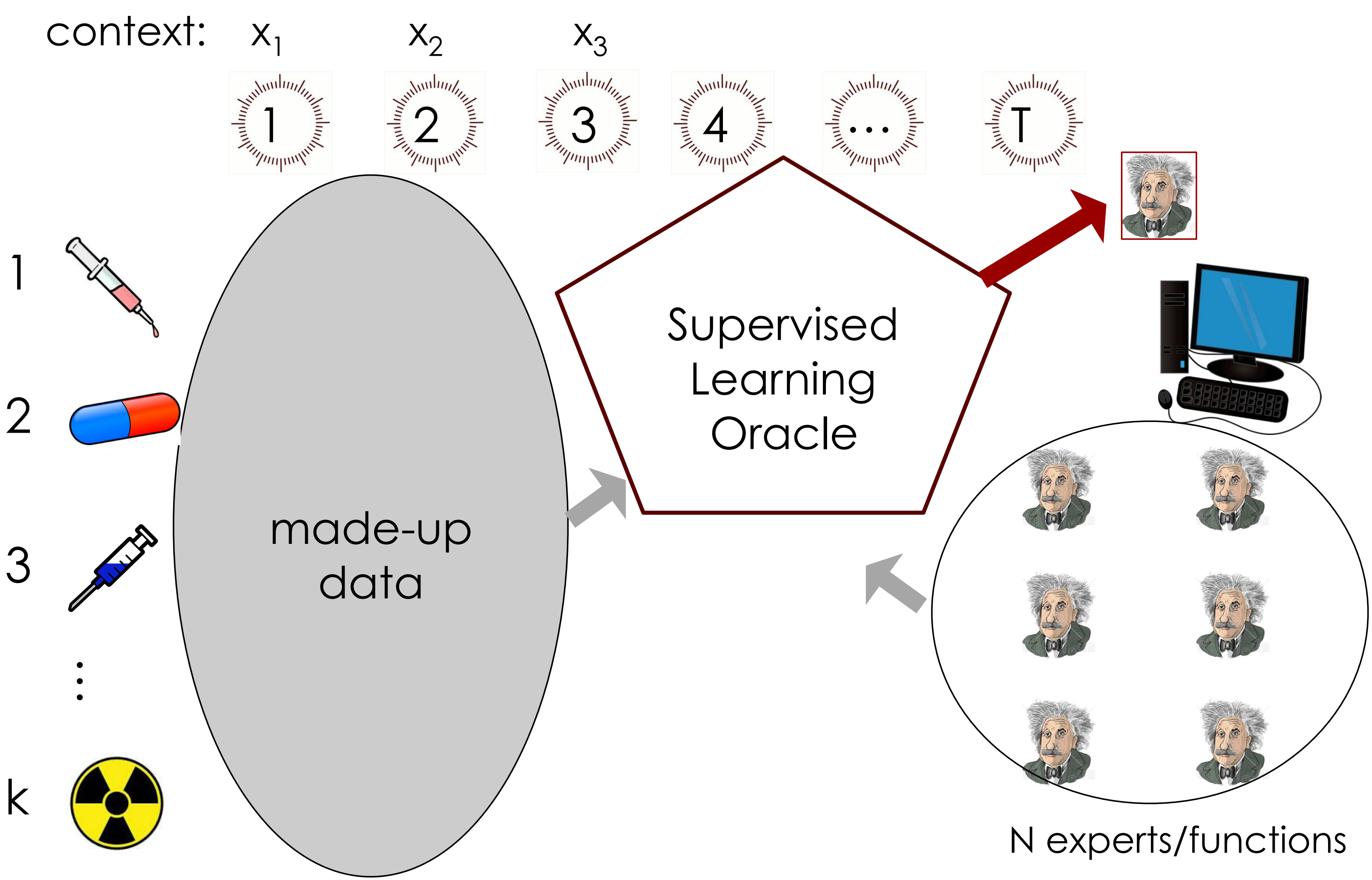
⋮

k 



N experts/functions


context:
x_1
x_2
x_3
1
2
3
4
...
T
1
2
3
⋮
k
made-up data
Supervised Learning Oracle
N experts/functions


context: x_1 x_2 x_3
1
2
3
4
…
T
1
2
3
⋮
k
made-up data
Supervised Learning Oracle
N experts/functions

Thm: [Dudik-Hsu-Kale-Karampatziakis-Langford-**R**-Zhang '11]:
For any $\delta>0$, w.p. at least $1-\delta$, given access to a learning oracle, R-UCB has regret $O((KT \ln (NT/\delta))^{1/2}$ in the stochastic contextual bandit setting and runs in time $\mathrm{poly}(K,T, \ln N)$.

Main idea:

make a convex program that optimally "solves" the bandit problem.

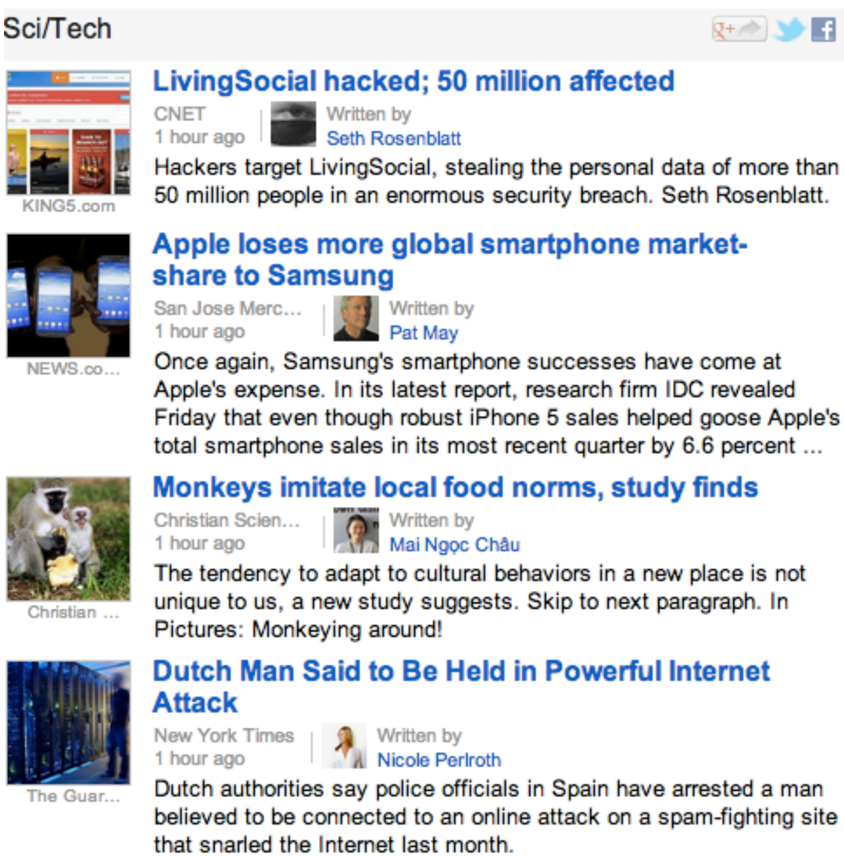
(Ab)use the supervised learning oracle to act as a separation oracle for this problem.

Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits
[Agarwal-et al. '14]

A research goal: make this work in adversarial model.

Bandit Slate Problems

[Kale-R-Schapire ’11]


Sci/Tech

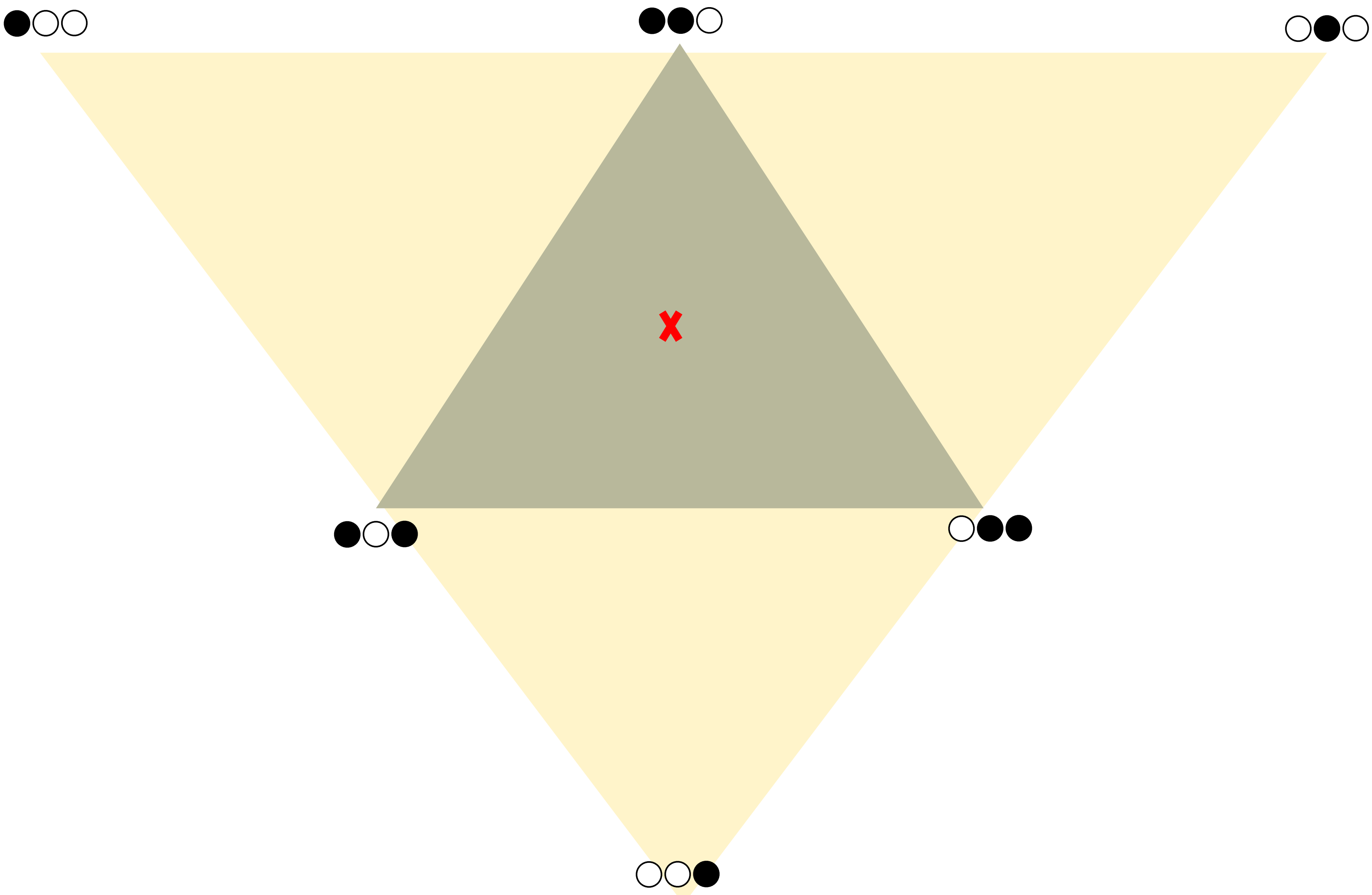
LivingSocial hacked; 50 million affected
CNET
1 hour ago
Written by
Seth Rosenblatt
Hackers target LivingSocial, stealing the personal data of more than 50 million people in an enormous security breach. Seth Rosenblatt.
KING5.com

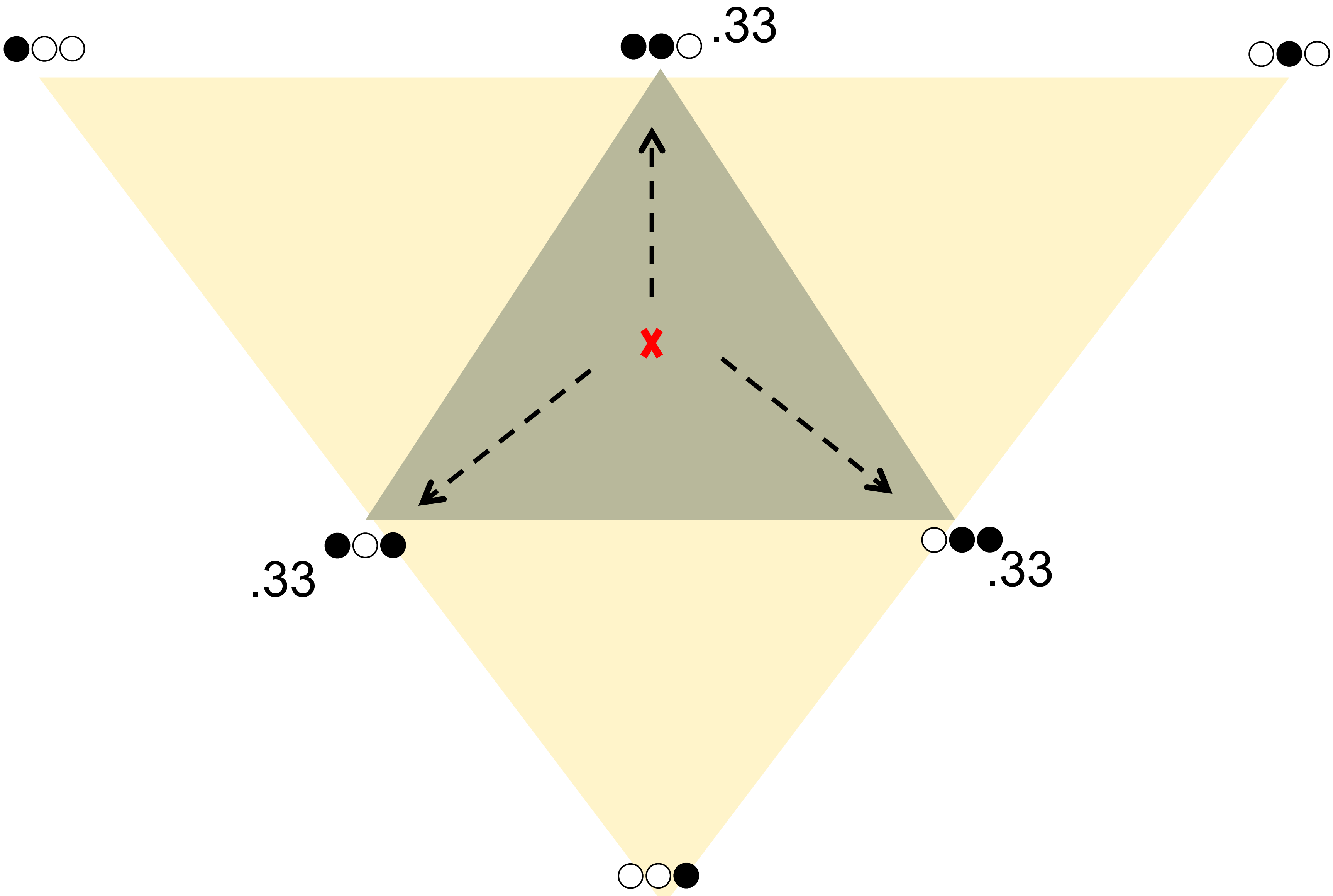
Apple loses more global smartphone market-share to Samsung
San Jose Merc…
1 hour ago
Written by
Pat May
Once again, Samsung's smartphone successes have come at Apple's expense. In its latest report, research firm IDC revealed Friday that even though robust iPhone 5 sales helped goose Apple's total smartphone sales in its most recent quarter by 6.6 percent …
NEWS.co…

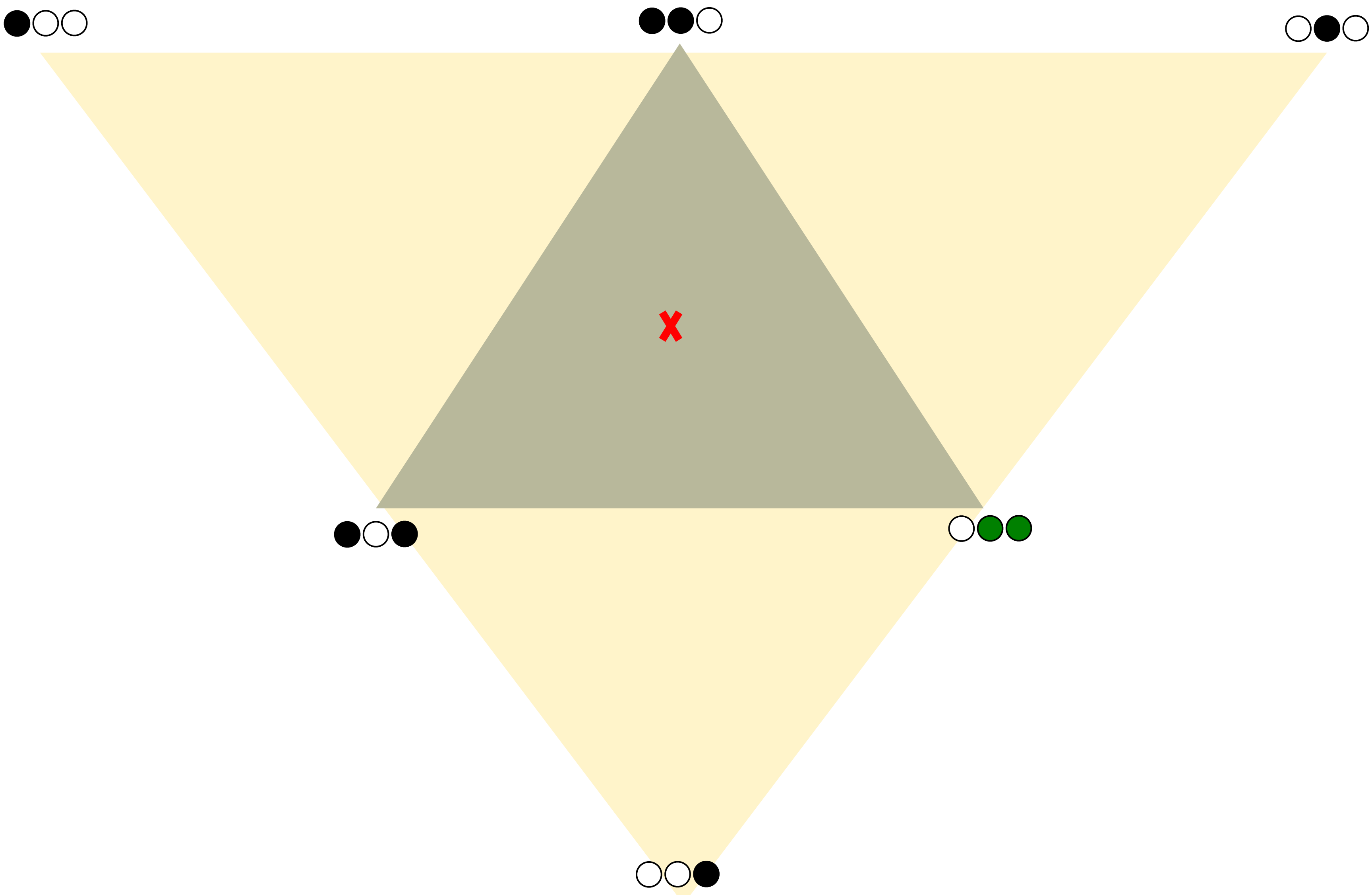
Monkeys imitate local food norms, study finds
Christian Scien…
1 hour ago
Written by
Mai Ngọc Châu
The tendency to adapt to cultural behaviors in a new place is not unique to us, a new study suggests. Skip to next paragraph. In Pictures: Monkeying around!
Christian …

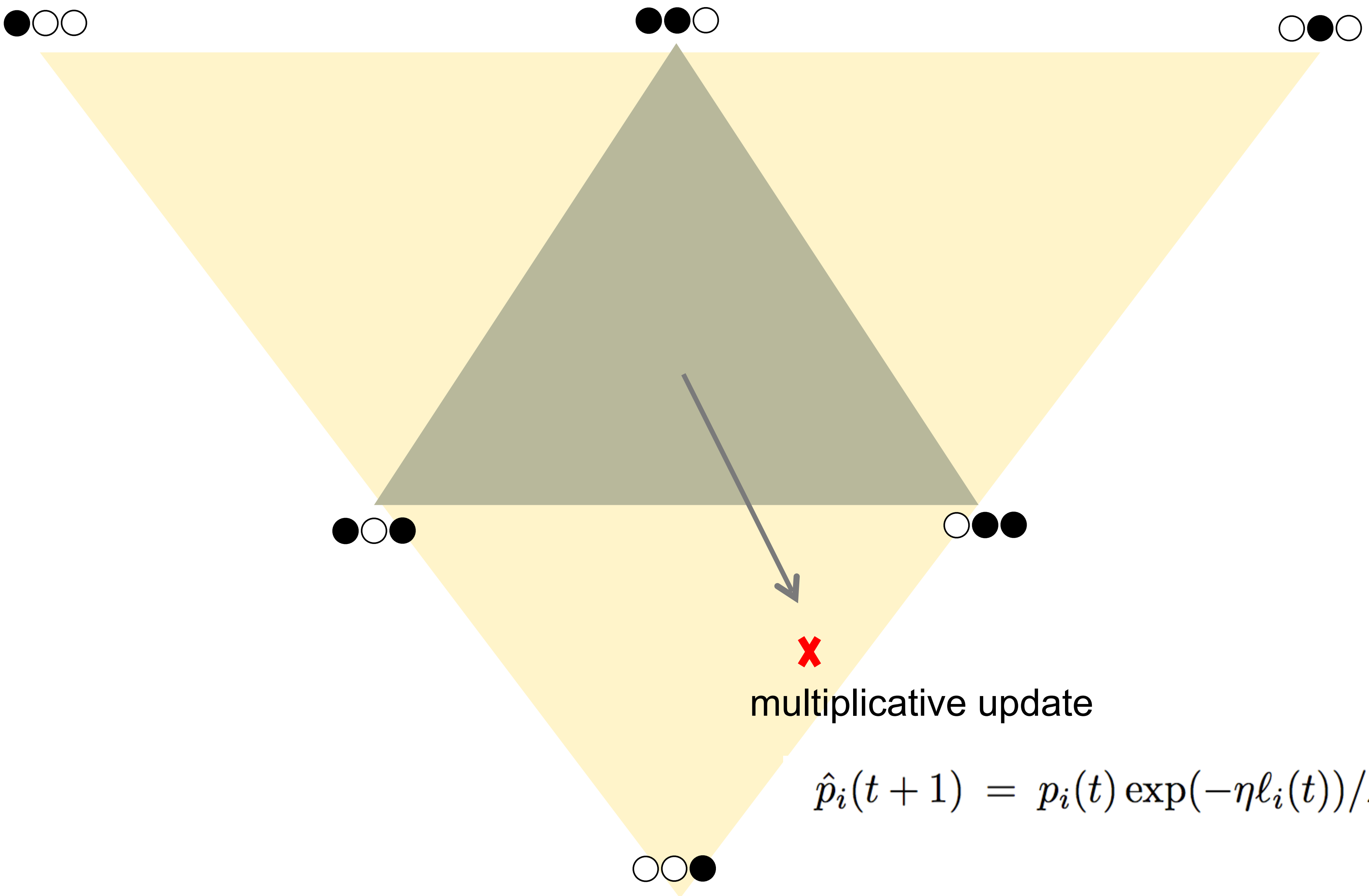
Dutch Man Said to Be Held in Powerful Internet Attack
New York Times
1 hour ago
Written by
Nicole Perlroth
Dutch authorities say police officials in Spain have arrested a man believed to be connected to an online attack on a spam-fighting site that snarled the Internet last month.
The Guar…

Instead of selecting one article, we need to select s ≥ 1, articles (possibly ranked). The motivation is web ads where a search engine shows multiple articles at once.



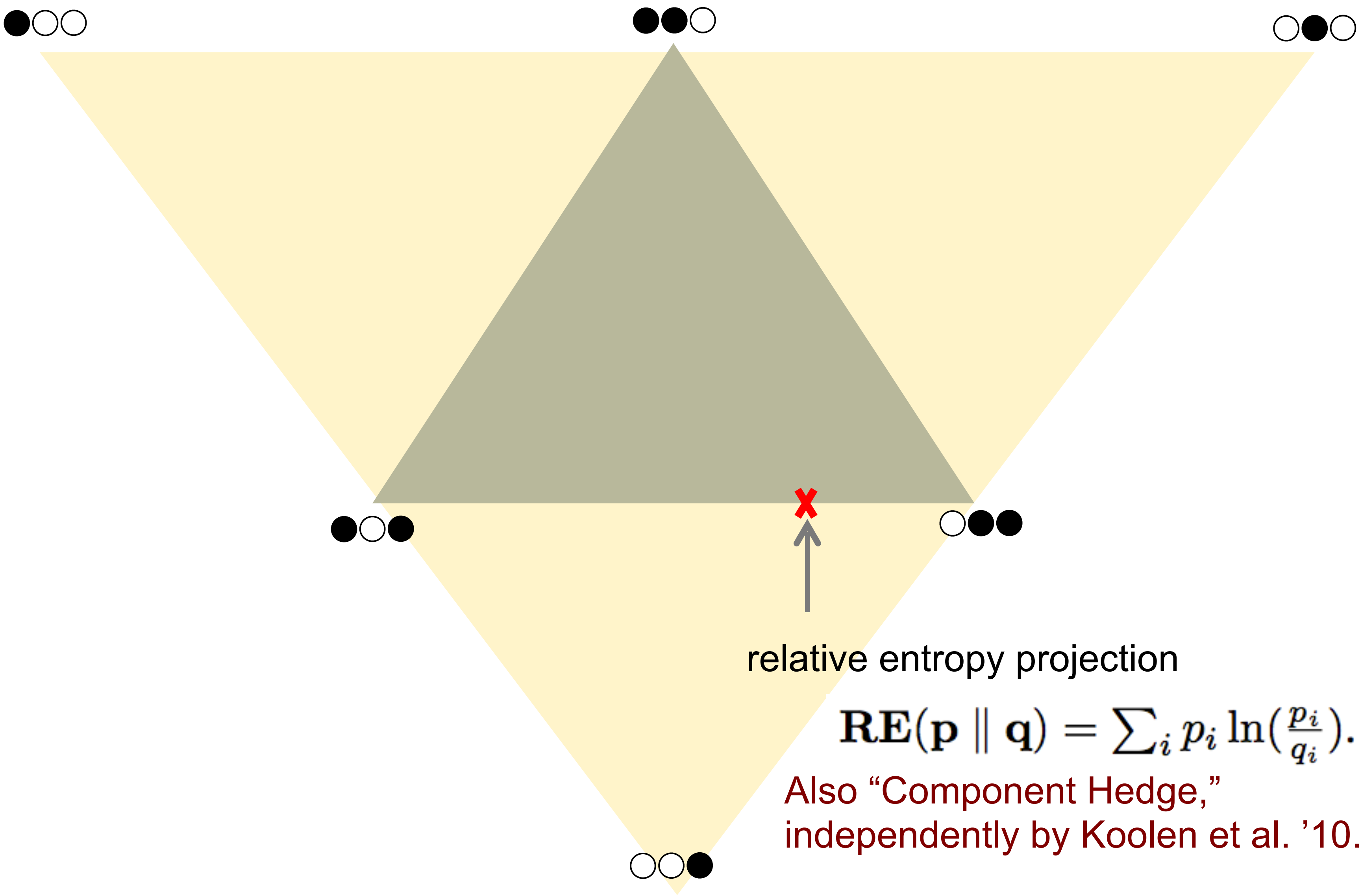

●○○
●●○ .33
○●○
.33 ●○●
○●● .33
○○●





multiplicative update

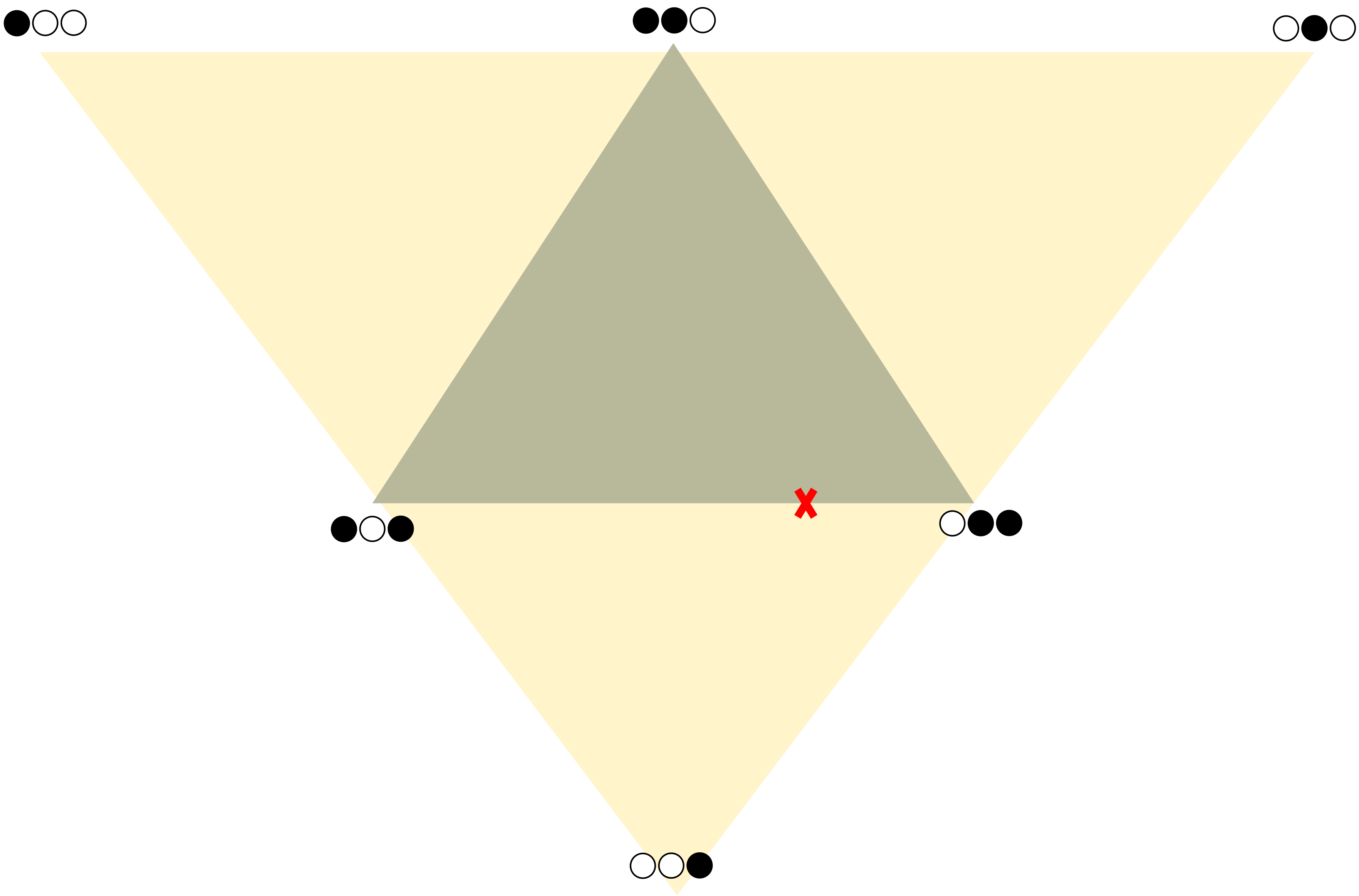
$$\hat{p}_i(t+1) \;=\; p_i(t)\exp(-\eta \ell_i(t))/Z(t)$$



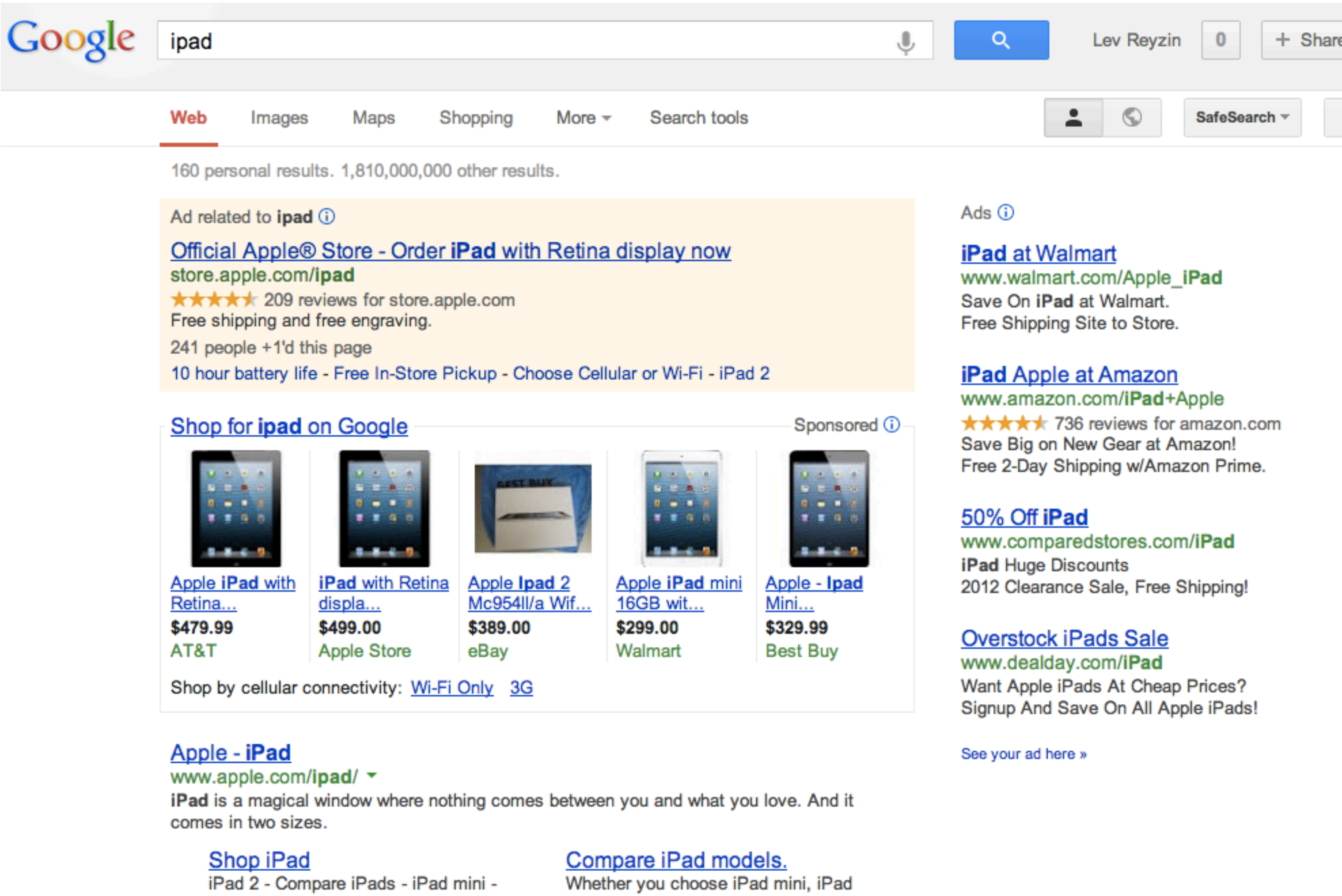
relative entropy projection

$$\mathbf{RE}(\mathbf{p} \parallel \mathbf{q}) = \sum_i p_i \ln(\frac{p_i}{q_i}).$$

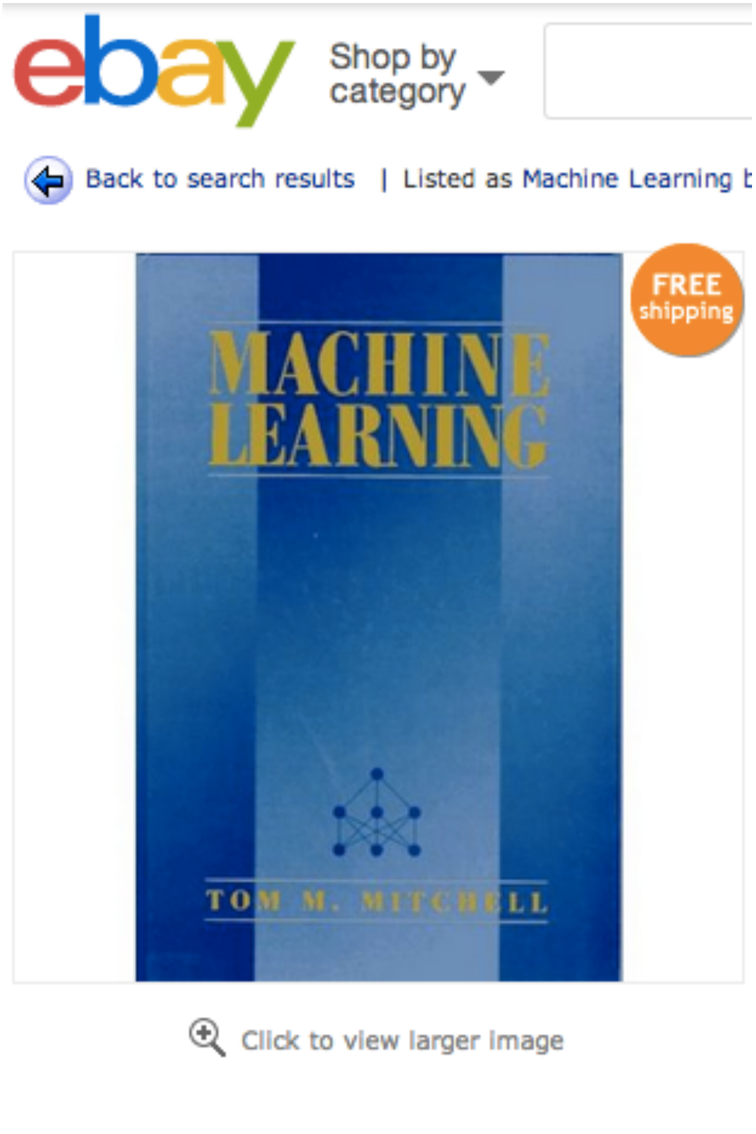
Also “Component Hedge,”
independently by Koolen et al. ’10.

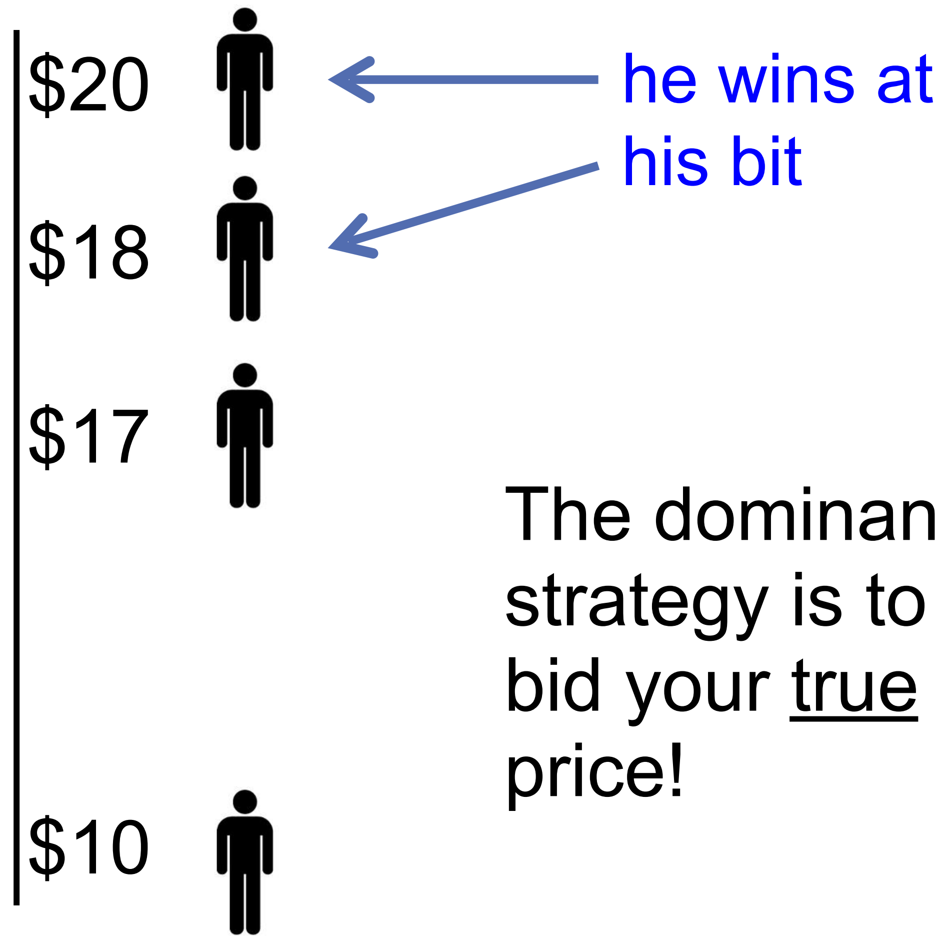


An Interesting Issue: Ads



SECOND PRICE, TRUTHFUL BIDDING





The dominant strategy is to bid your <u>true</u> price!

CONTEXTUAL ADVERTISING

1. **For an ad to be shown, it must have high expected earnings.**
   Earnings = clickthrough rate (CTR) x expected charged price

2. **CTR must be learned**
   a classic contextual bandits problem

3. **Charged prices are functions of the bids of advertisers.**
   e.g. Can't ever charge more than an advertiser's bid

4. **Ads must be shown so that CTR is learned quickly, but the auction should be truthful.**

SUMMARY

When dealing with many customers/ subscribers and many options, a smart automated strategy needs to be employed.

This is becoming true of nearly every company presenting content online.

Presents many important mathematical challenges, most of which are wide open.

THANK YOU!