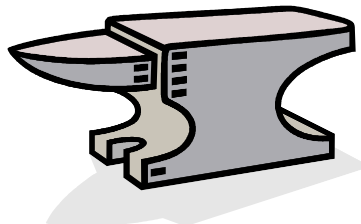


Throw(person, *x*) ?

Weight(x) < 50lbs ^
Max_dim(x) < 20ft ^ … ^
=>Throw(person, x)

Weight(baseball) = 5oz ^…. =>
Throw(person, baseball)

"throwable objects such as"

Web Images Maps Shopping Books More Search tools

About 5,050 results (0.19 seconds)

[Patent US5984812 - Grippable surface for throwable object - Google ...]

www.google.com/patents/US5984812

This invention relates to a grippable surface for **throwable objects such as** a football, baseball, etc. which enhances the ease with which the object may be **...**

[PDF] [Name Juggle.pdf - GOAL Consulting]

www.goalconsulting.org/page3/files/Name%20Juggle.pdf

Materials: Many soft **throwable objects such as** fleece balls, wadded up pieces of paper, Nerf™ balls. Level: Grades K and higher. Suggested Procedure. 1.

`Cities such as X`

`Y, mayor of X`

- The Web makes hard AI problems easier
- …but

Understanding Language is Hard

Michael Bloomberg, **mayor** of **New York City**, was born in …

Eric Johnson, CEO of **Texas Instruments**, **mayor** of **Dallas** from 1964-1971, and …

Redundancy enables Information Extraction

Michael Bloomberg, **mayor** of **New York City**, was born in …

Announced by **Michael Bloomberg**, the **mayor** of **New York City**…

Michael Bloomberg, **mayor** of **New York City**, **stated that**…

…+10,000s more

Eric Johnson, CEO of **Texas Instruments**, **mayor** of **Dallas** from 1964-1971, and …

(only once)

Mayor	City	*p*
Michael Bloomberg	New York City	0.99
Texas Instruments	Dallas	0.05
…	…	…

[Downey et al., AIJ 2010]

Redundancy: Single Pattern

Consider a single pattern suggesting C , e.g.,

`countries such as` x

If an extraction x appears k times in a set of n distinct occurrences of the pattern, what is the probability that $x \in C$?

Redundancy: Single Pattern

C = `Country`

n = 10 occurrences

"...countries such as Saudi Arabia..."
"...countries such as the United States..."
"...countries such as Saudi Arabia..."
"...countries such as Japan..."
"...countries such as Africa..."
"...countries such as Japan..."
"...countries such as the United Kingdom..."
"...countries such as Iraq..."
"...countries such as Afghanistan..."
"...countries such as Australia..."

Naïve Model: Noisy-Or

C = `Country`

n = 10

	k	$P_{\text{noisy-or}}$
Saudi Arabia	**2**	**0.99**
Japan	**2**	**0.99**
United States	**1**	**0.9**
Africa	**1**	**0.9**
United Kingdom	**1**	**0.9**
Iraq	**1**	**0.9**
Afghanistan	**1**	**0.9**
Australia	**1**	**0.9**

[Agichtein & Gravano, 2000; Lin et al. 2003]

$P_{\text{noisy-or}}(x \in C \mid x \text{ seen } k \text{ times})$

$$= 1 - (1 - p)^k$$

p = probability pattern yields a correct extraction, i.e.,

$$p = 0.9$$

Noisy-or ignores:

- Sample size (n)
- Distribution of C

Needed in Model: Sample Size

C = `Country`

$n = 10$

	k	$P_{\text{noisy-or}}$
Saudi Arabia	2	0.99
Japan	2	0.99
United States	1	0.9
Africa	1	0.9
United Kingdom	1	0.9
Iraq	1	0.9
Afghanistan	1	0.9
Australia	1	0.9

C = `Country`

n ~50,000

	k	$P_{\text{noisy-or}}$
United States	3899	0.9999…
China	1999	0.9999…
	…	
OilWatch Africa	1	0.9
Religion Paraguay	1	0.9
Chicken Mole	1	0.9
Republics of Kenya	1	0.9
Atlantic Ocean	1	0.9

As sample size increases, noisy-or becomes inaccurate.

Needed in Model: Distribution of C

C = **`Country`**

n ~50,000

	k	$P_{\text{noisy-or}}$
United States	**3899**	**0.9999…**
China	**1999**	**0.9999…**
…		
OilWatch Africa	**1**	**0.9**
Religion Paraguay	**1**	**0.9**
Chicken Mole	**1**	**0.9**
Republics of Kenya	**1**	**0.9**
Atlantic Ocean	**1**	**0.9**

$$P_{\text{freq}}(x \in C \mid x \text{ seen } k \text{ times})$$

$$= 1 - (1 - p)^{\alpha k/n}$$

Needed in Model: Distribution of C

C = **`Country`**

n ~50,000	k	P_{freq}
United States	3899	0.9999…
China	1999	0.9999…
…		
OilWatch Africa	1	0.05
Religion Paraguay	1	0.05
Chicken Mole	1	0.05
Republics of Kenya	1	0.05
Atlantic Ocean	1	0.05

$P_{\text{freq}}(x \in C \mid x \text{ seen } k \text{ times})$

$$= 1 - (1 - p)^{\alpha k/n}$$

Needed in Model: Distribution of *C*

C = `Country`

n ~50,000	k	P_{freq}
United States	**3899**	**0.9999…**
China	**1999**	**0.9999…**
…		
OilWatch Africa	**1**	**0.05**
Religion Paraguay	**1**	**0.05**
Chicken Mole	**1**	**0.05**
Republics of Kenya	**1**	**0.05**
Atlantic Ocean	**1**	**0.05**

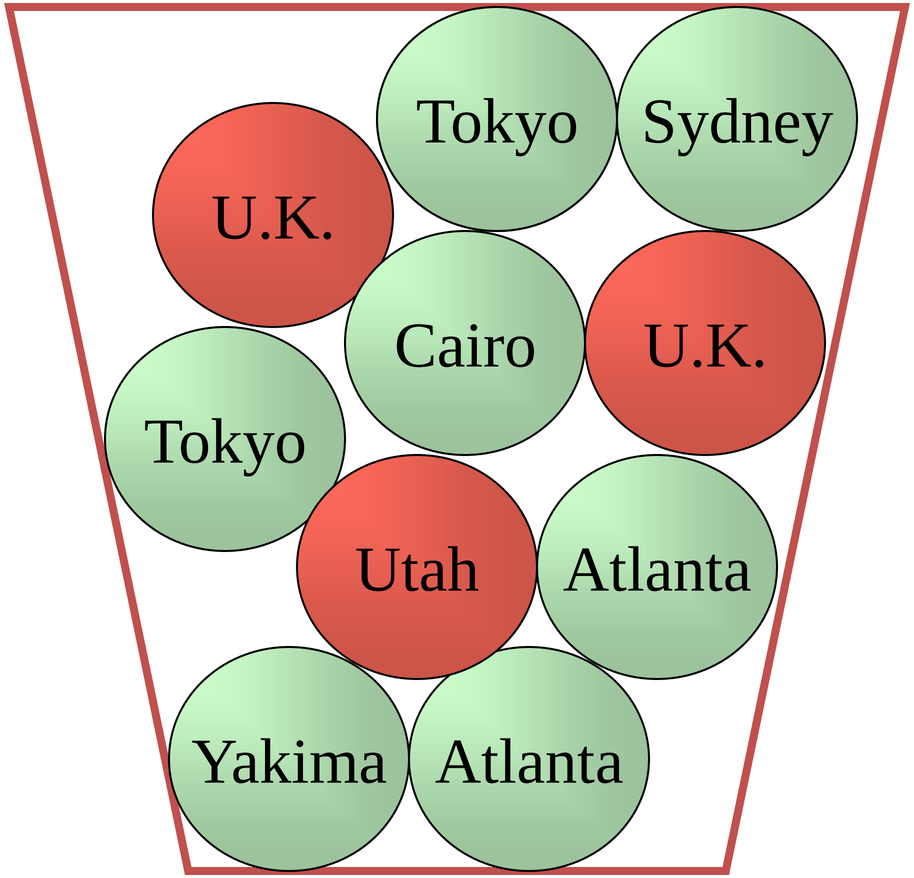
C = `City`

n ~50,000	k	P_{freq}
New York	**1488**	**0.9999…**
Chicago	**999**	**0.9999…**
…		
El Estor	**1**	**0.05**
Nikki	**1**	**0.05**
Ragaz	**1**	**0.05**
Villegas	**1**	**0.05**
Northeastwards	**1**	**0.05**

Probability $x \in C$ depends on the distribution of *C*.

Solution: URNS Model

Urn for C = **`City`**



…**`cities such as Tokyo`**…

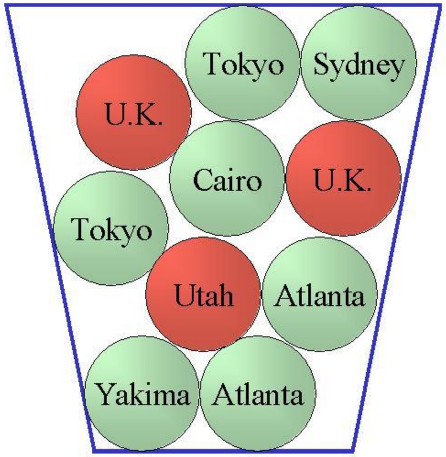
Urn – Formal Definition

C – set of unique target labels

E – set of unique error labels

num(C) – distribution of target labels

num(E) – distribution of error labels



Urn Example

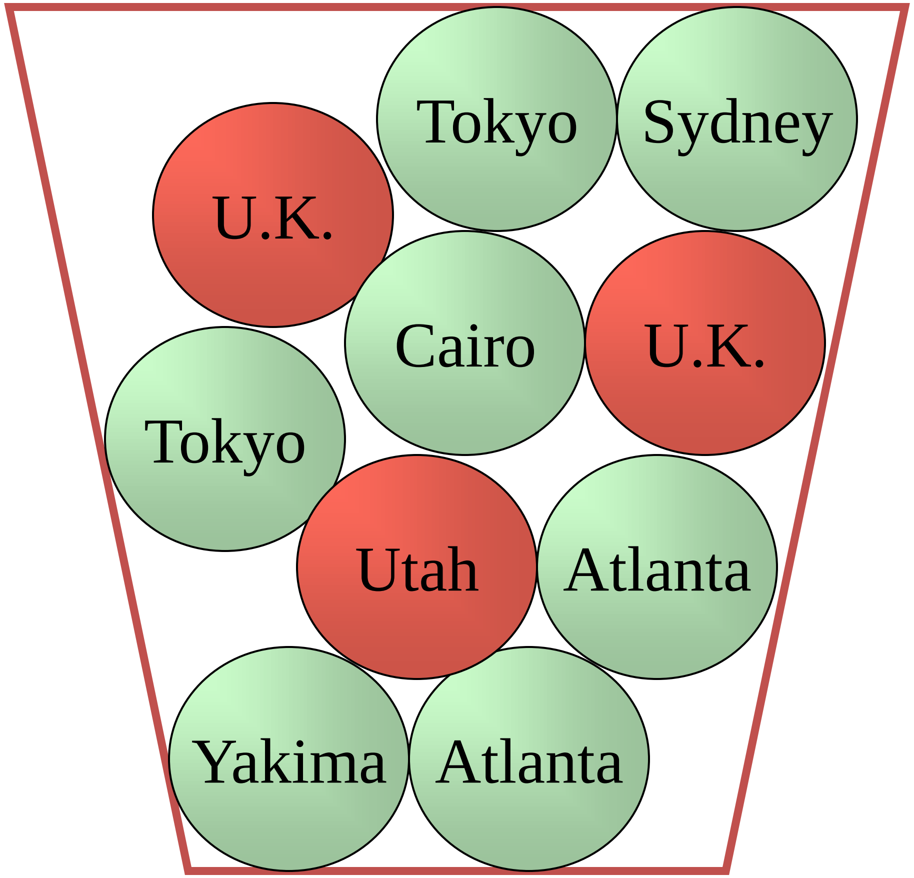
distribution of target labels:

num(*C*) = {2, 2, 1, 1, 1}

distribution of error labels:

num(*E*) = {2, 1}

Urn for *C* = `City`



Computing Probabilities

If an extraction x appears k times in a set of n distinct occurrences of the pattern, what is the probability that $x \in C$?

Computing Probabilities

Given that an extraction *x* appears *k* times in *n* draws from the urn (with replacement), what is the probability that *x* ∈ *C*?

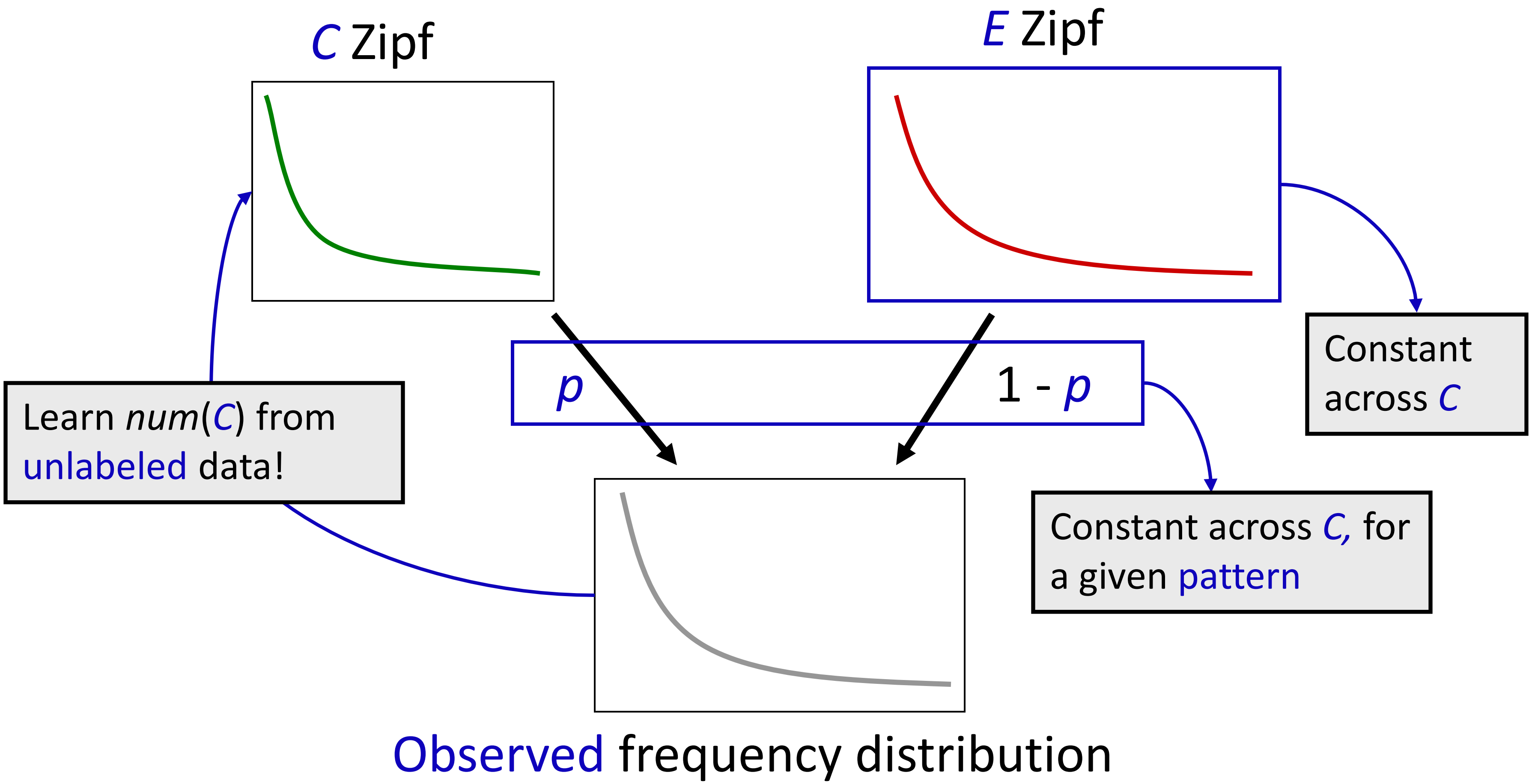
$$P(x \in C | x \,\mathrm{appears}\, k \,\mathrm{times\, in}\, n \,\mathrm{draws}) = \frac{\sum_{r \in num(C)} (\frac{r}{s})^k (1 - \frac{r}{s})^{n-k}}{\sum_{r' \in num(C \cup E)} (\frac{r'}{s})^k (1 - \frac{r'}{s})^{n-k}}$$

where *s* is the total number of balls in the urn

URNS without labeled data

- Needed: *num*(*C*), *num*(*E*)
- Assumed to be Zipf
  - Frequency of *i*th element $\propto i^{-z}$
- With assumptions, learn Zipfian parameters for any class *C* from unlabeled data alone

URNS without labeled data



Probabilities Assigned by URNS

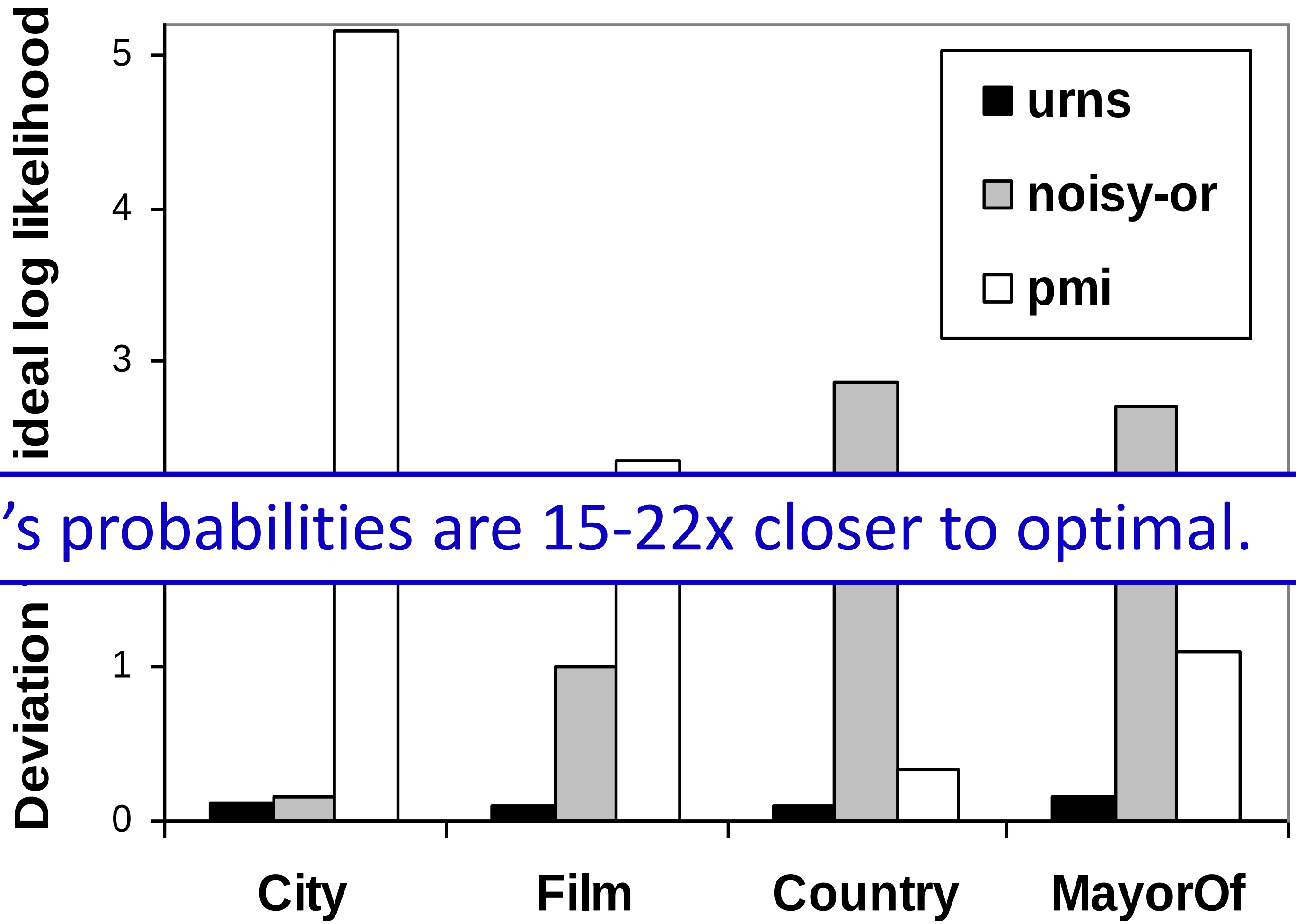
C = `Country`

n ~50,000	k	P_{URNS}
United States	3899	0.9999…
China	1999	0.9999…
…		
OilWatch Africa	1	0.03
Religion Paraguay	1	0.03
Chicken Mole	1	0.03
Republics of Kenya	1	0.03
Atlantic Ocean	1	0.03
New Zeland	1	0.03

C = `City`

n ~50,000	k	P_{URNS}
New York	1488	0.9999…
Chicago	999	0.9999…
…		
El Estor	1	0.63
Nikki	1	0.63
Ragaz	1	0.63
Villegas	1	0.63
Cres	1	0.63
Northeastwards	1	0.63

Probability Accuracy



URNS's probabilities are 15-22x closer to optimal.

Sensitivity Analysis

- URNS assumes *num*(*E*), *p* are constant
  - If we alter parameter choices substantially, URNS still outperforms noisy-or, PMI by at least 8x
- Most sensitive to *p*
  - *p* ~ 0.85 is relatively consistent across randomly selected classes from Wordnet
    (solvents, devices, thinkers, relaxants, mushrooms, mechanisms, resorts, flies, tones, machines, …)

Multiple Extraction Patterns

Phrase	Hits
"**Omaha** and other cities"	950
"**Illinois** and other cities"	24,400
"cities such as **Omaha**"	930
"cities such as **Illinois**"	6

- Multiple urns
  - Target label frequencies correlated across urns
  - Error label frequencies can be uncorrelated

Benefits from Multiple Urns

	Precision at *K*	
K	Single	Multiple
10	1.0	**1.0**
20	0.9875	**1.0**
50	0.925	**0.955**
100	0.8375	**0.845**
200	0.7075	**0.71**

Using multiple URNS reduces error by 29%.