Learning from Human-Generated Lists

Kwang-Sung Jun (deltakam@cs.wisc.edu)
Department of Computer Sciences, University of Wisconsin-Madison

Xiaojin (Jerry) Zhu (jerryzhu@cs.wisc.edu)
Department of Computer Sciences, University of Wisconsin-Madison

Burr Settles (burrsettles@gmail.com)
Duolingo

Timothy Rogers (ttrogers@wisc.edu)
Department of Psychology, University of Wisconsin-Madison

ICML'13

Example 1:

"List examples of animals without repetition for 60 seconds."

Order	Item
1	dog
2	cat
3	tiger
4	cow
…	…
7	lion
8	tiger
9	bear
…	…
11	armadillo

Example 1: Verbal Fluency

"List examples of animals without repetition for 60 seconds."

Order	Item
1	dog
2	cat
3	tiger
4	cow
…	…
7	lion
8	tiger
9	bear
…	…
11	armadillo

Example 2: Feature Volunteering

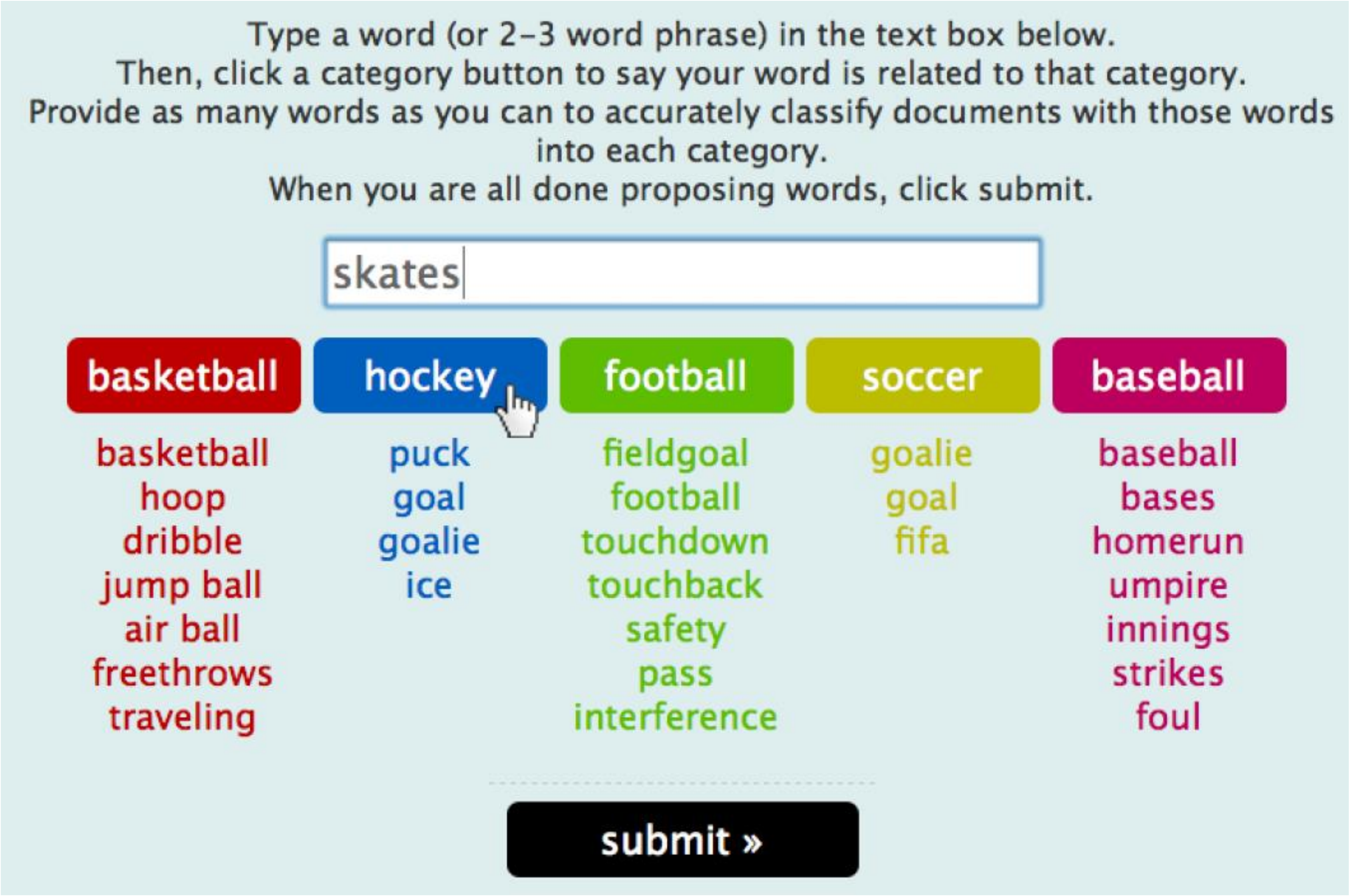
- Simple rules: e.g. skates ⇒ hockey
  - IF a document contains the word **skates**, THEN label the document as **hockey**.

Example 2: Feature Volunteering

Example 2: Feature Volunteering

Example 2: Feature Volunteering

Example 2: Feature Volunteering



Example 2: Feature Volunteering

Order	Item
1	baseball bat ⇒ Baseball
…	…
7	quarterback ⇒ Football
8	football field ⇒ Football
9	soccer ball ⇒ Soccer
…	…
23	basketball court ⇒ Basketball
24	football field ⇒ Football
25	soccer field ⇒ Soccer
…	…

Characteristics of Human-Generated Lists

- Order matters
- Repeats happen

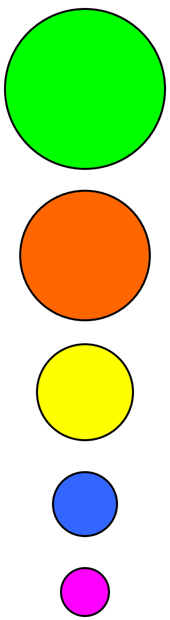
Order	Item
1	dog
2	cat
7	lion
8	tiger
9	bear
…	…
11	armadillo

Order	Item
1	baseball bat ⇒ Baseball
…	…
…	…
23	basketball court ⇒ Basketball
24	football field ⇒ Football
25	soccer field ⇒ Soccer
…	…

Sampling WIth Reduced repLacement (SWIRL)

SWIRL (Sampling WIth Reduced repLacement)

- s_i: size of the ball *i*



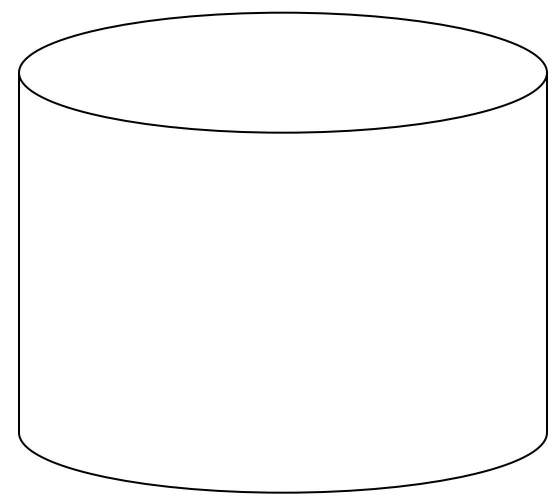
$s_{green} = 5$

$s_{orange} = 4$

$s_{yellow} = 2.5$

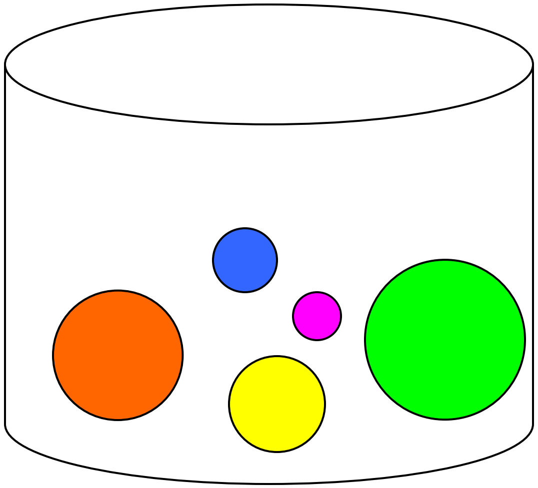
$s_{blue} = 1$

$s_{pink} = 0.5$



SWIRL (Sampling WIth Reduced repLacement)

- s_i: size of the ball i
- iteration 1:

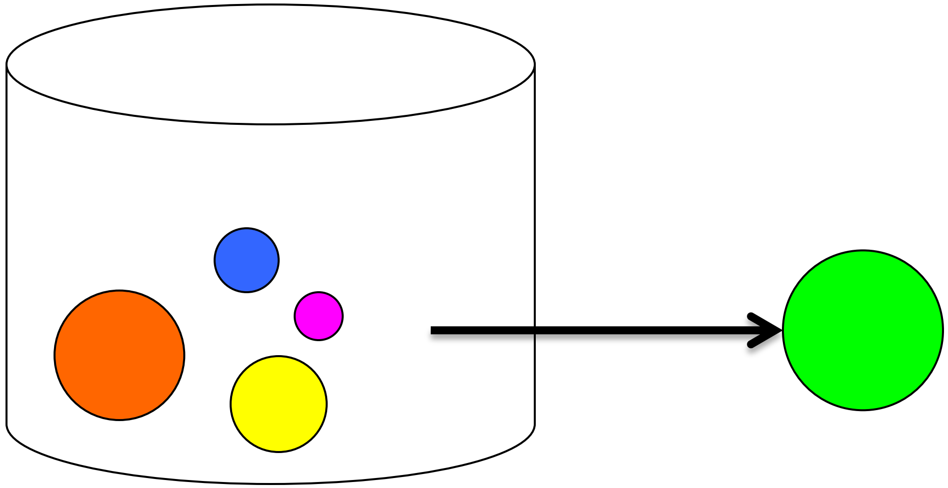


Probability ∝ Size

Order	Item

SWIRL (Sampling WIth Reduced repLacement)

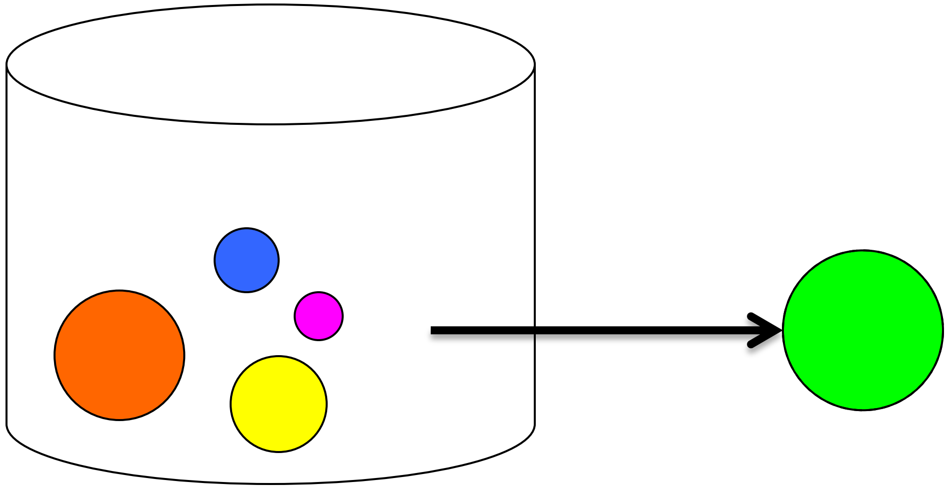
- s_i: size of the ball i
- iteration 1:



Order	Item

SWIRL (Sampling WIth Reduced repLacement)

- s_i: size of the ball i
- iteration 1:

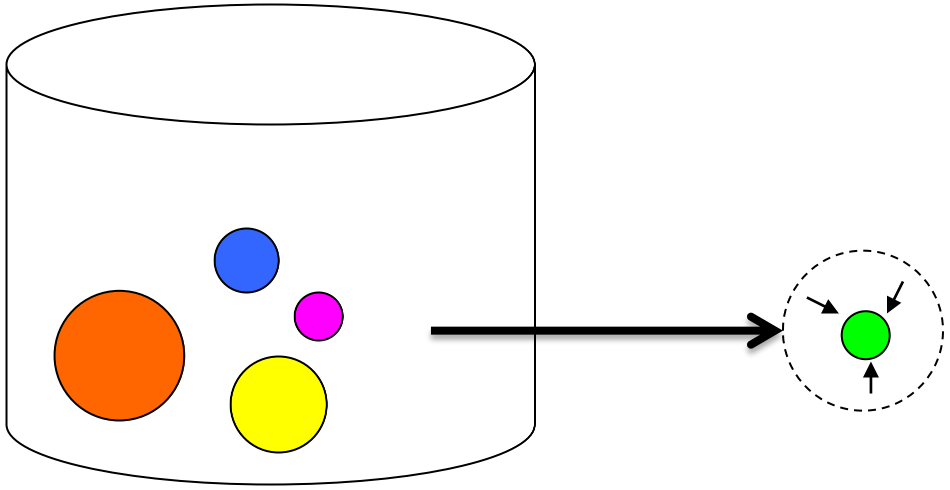


Order	Item
1	Green

SWIRL (Sampling WIth Reduced repLacement)

- s_i: size of the ball *i*
- iteration 1:

- α: discount factor

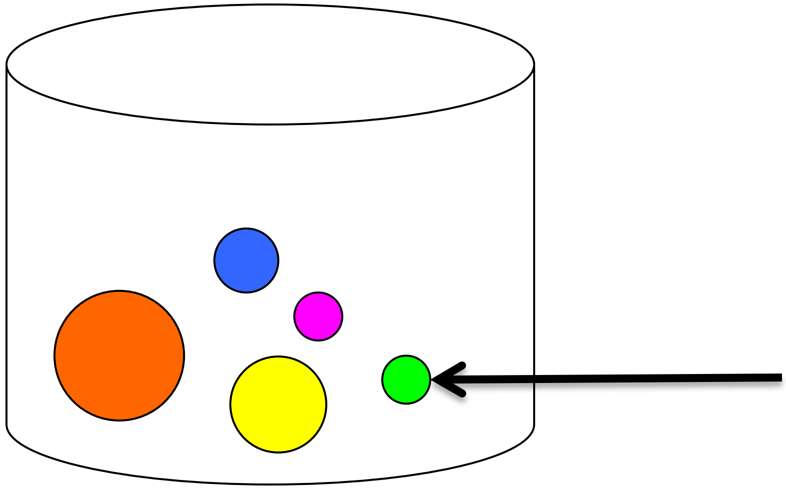


Order	Item
1	Green

$$s_{green} \leftarrow \alpha s_{green}$$

SWIRL (Sampling WIth Reduced repLacement)

- s_i: size of the ball i
- iteration 1:
- α: discount factor

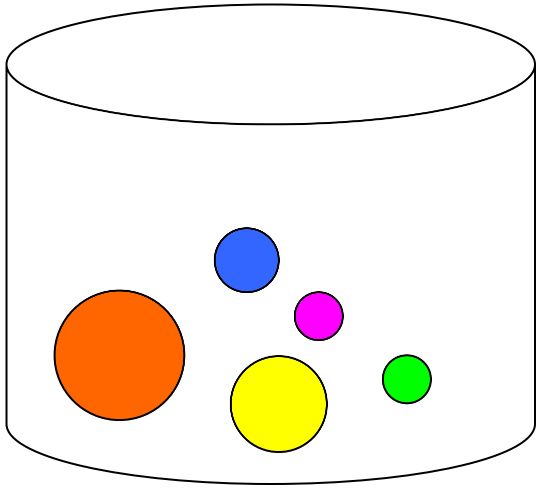


Order	Item
1	Green

SWIRL (Sampling WIth Reduced repLacement)

- s_i: size of the ball i
- iteration 1:

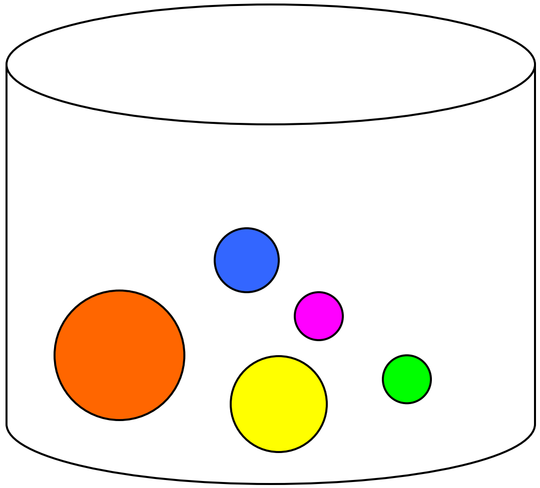
- α: discount factor



Order	Item
1	Green

SWIRL (Sampling WIth Reduced repLacement)

- s_i: size of the ball i
- iteration 2:
- α: discount factor

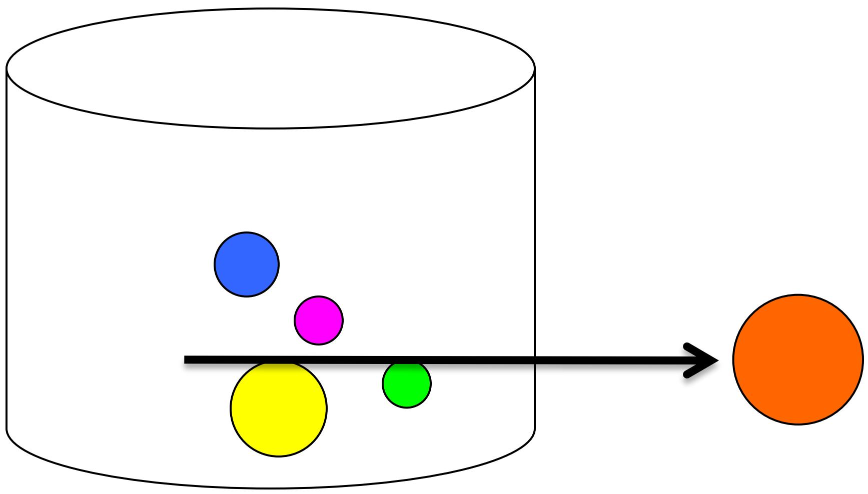


Probability ∝ Size

Order	Item
1	Green

SWIRL (Sampling WIth Reduced repLacement)

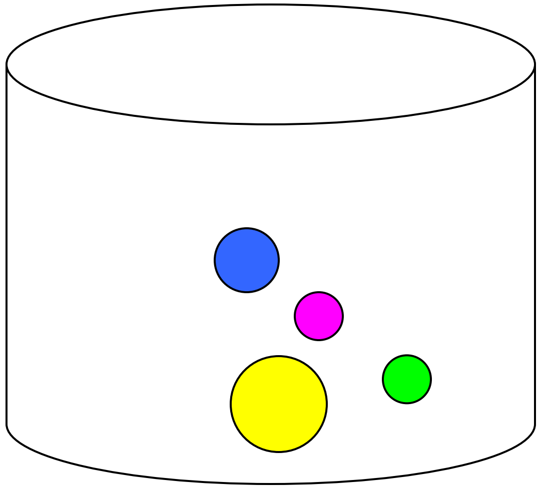
- s_i: size of the ball i
- iteration 2:
- α: discount factor

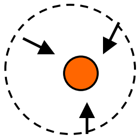


Order	Item
1	Green
2	Orange

SWIRL (Sampling WIth Reduced repLacement)

- s_i: size of the ball *i*
- iteration 2:
- α: discount factor



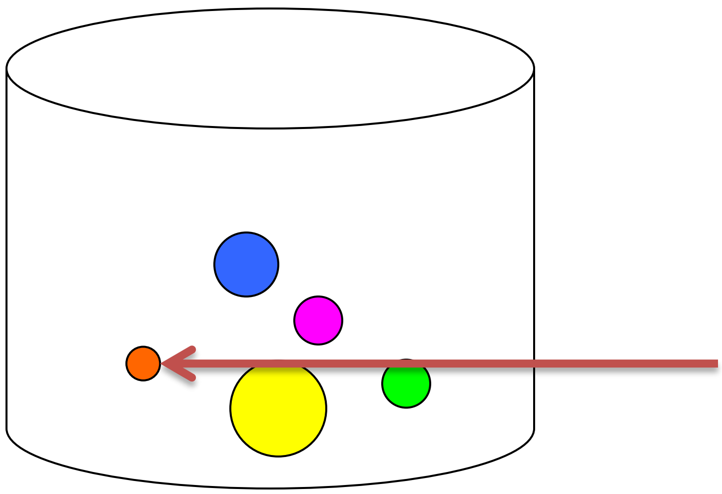


Order	Item
1	Green
2	Orange

$$s_{orange} \leftarrow \alpha s_{orange}$$

SWIRL (Sampling WIth Reduced repLacement)

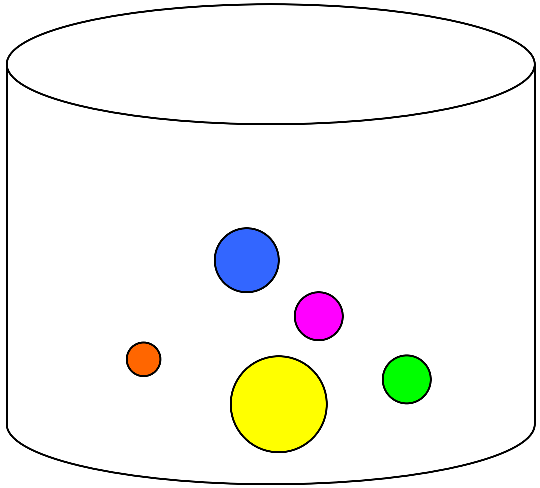
- s_i: size of the ball i
- iteration 2:
- α: discount factor



Order	Item
1	Green
2	Orange

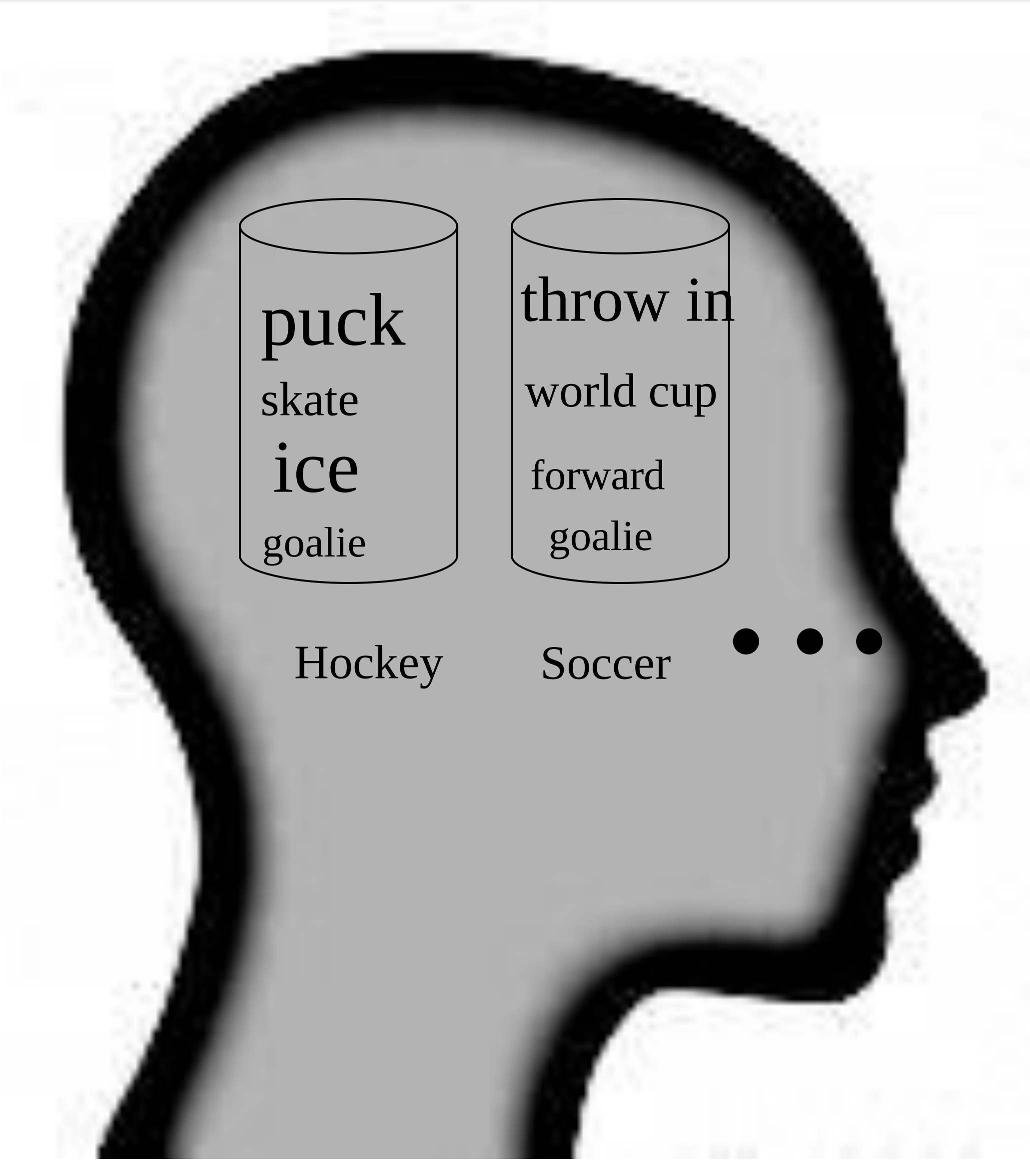
SWIRL (Sampling WIth Reduced repLacement)

- s_i: size of the ball i
- iteration 2:
- α: discount factor



Order	Item
1	Green
2	Orange
…	…

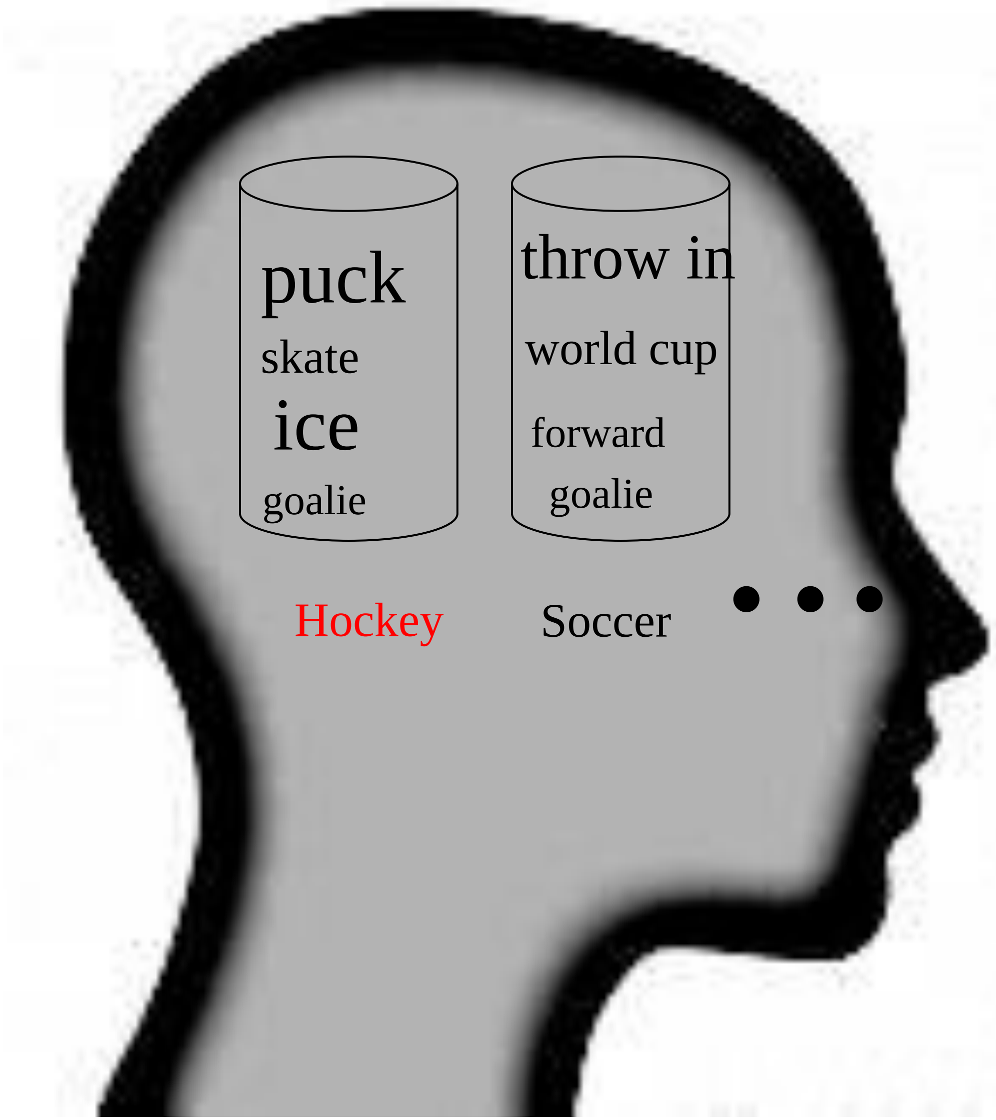
SWIRL for Feature Volunteering



Sports

Order	Item

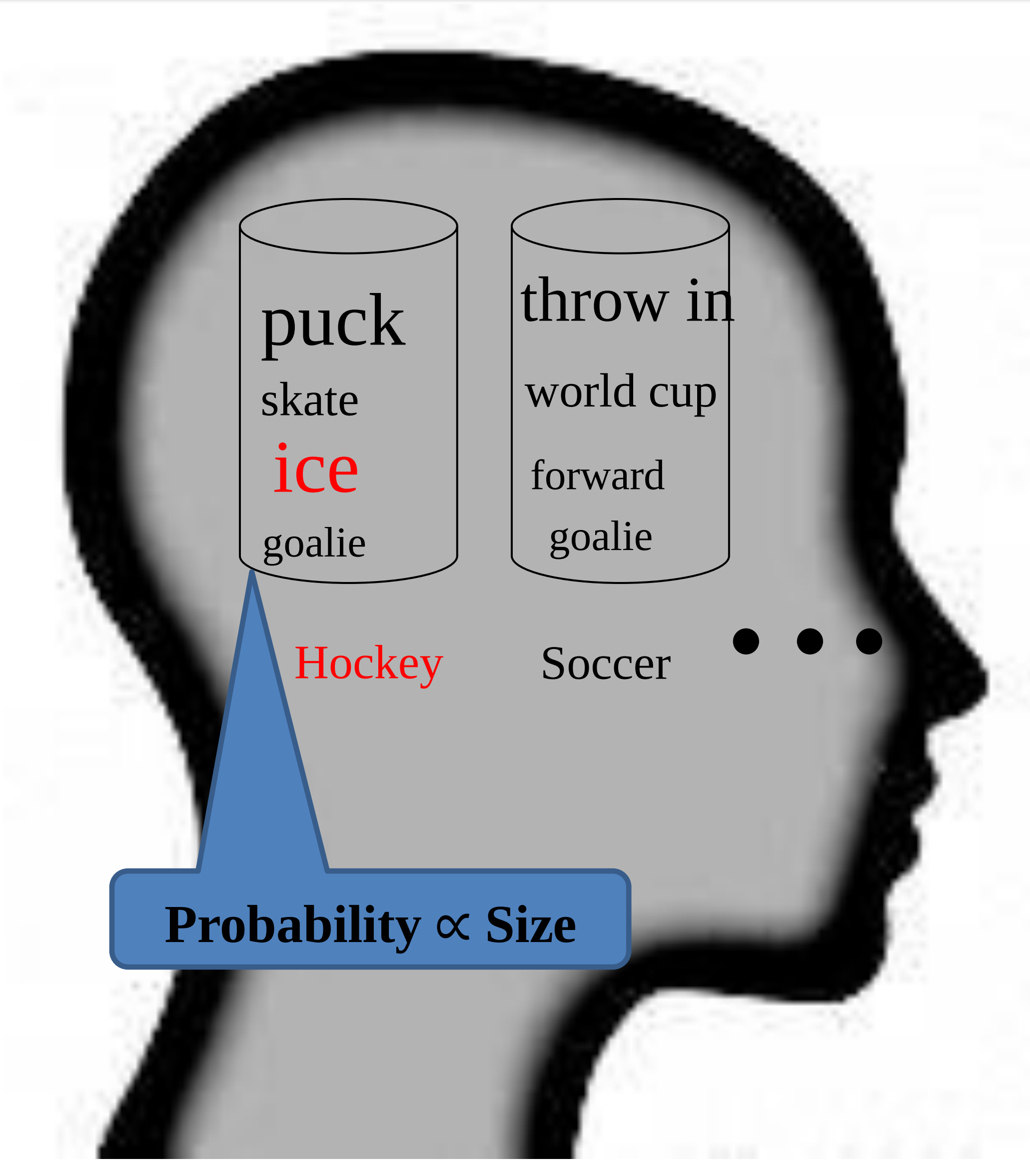
SWIRL for Feature Volunteering



Sports

Order	Item
1	

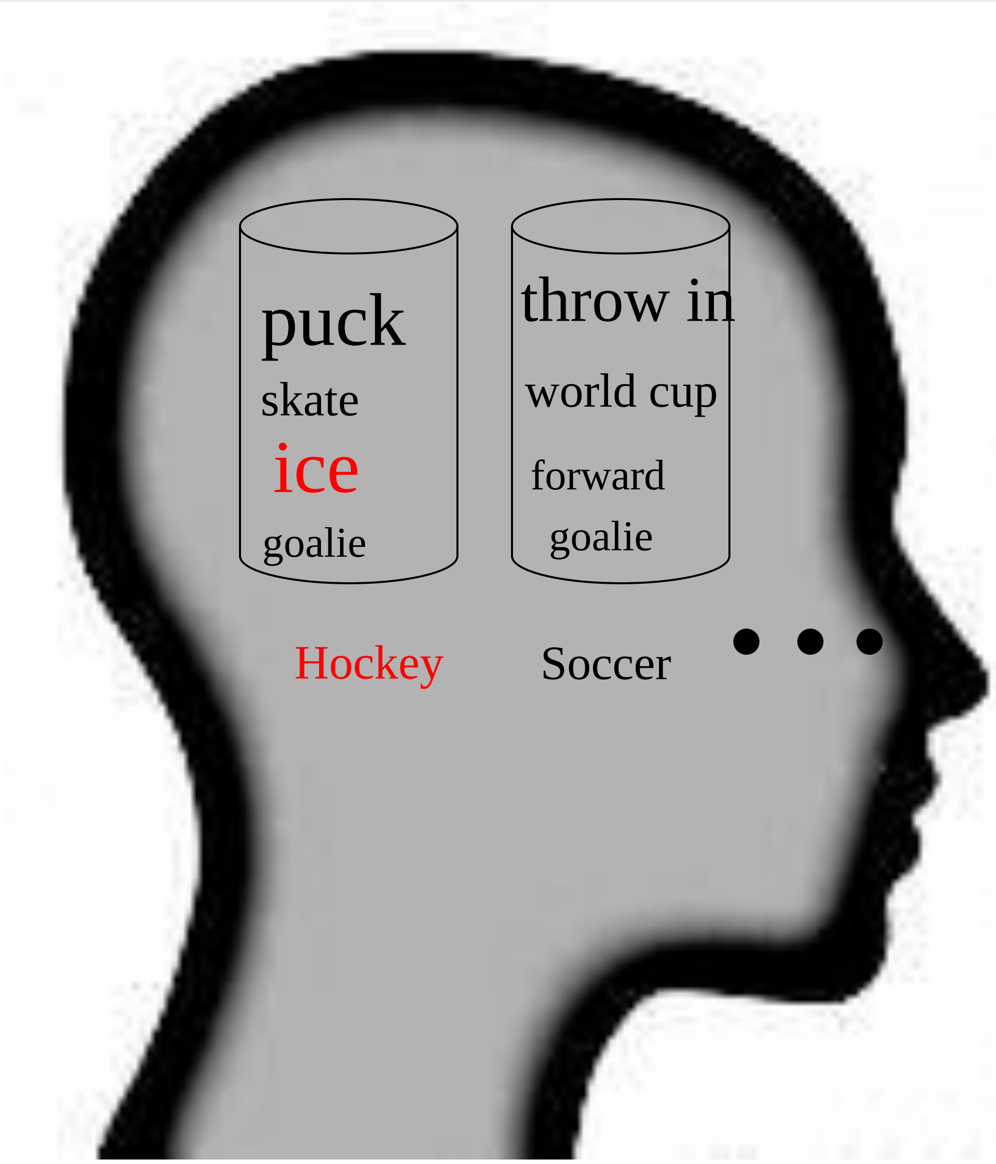
SWIRL for Feature Volunteering



Sports

Order	Item
1	

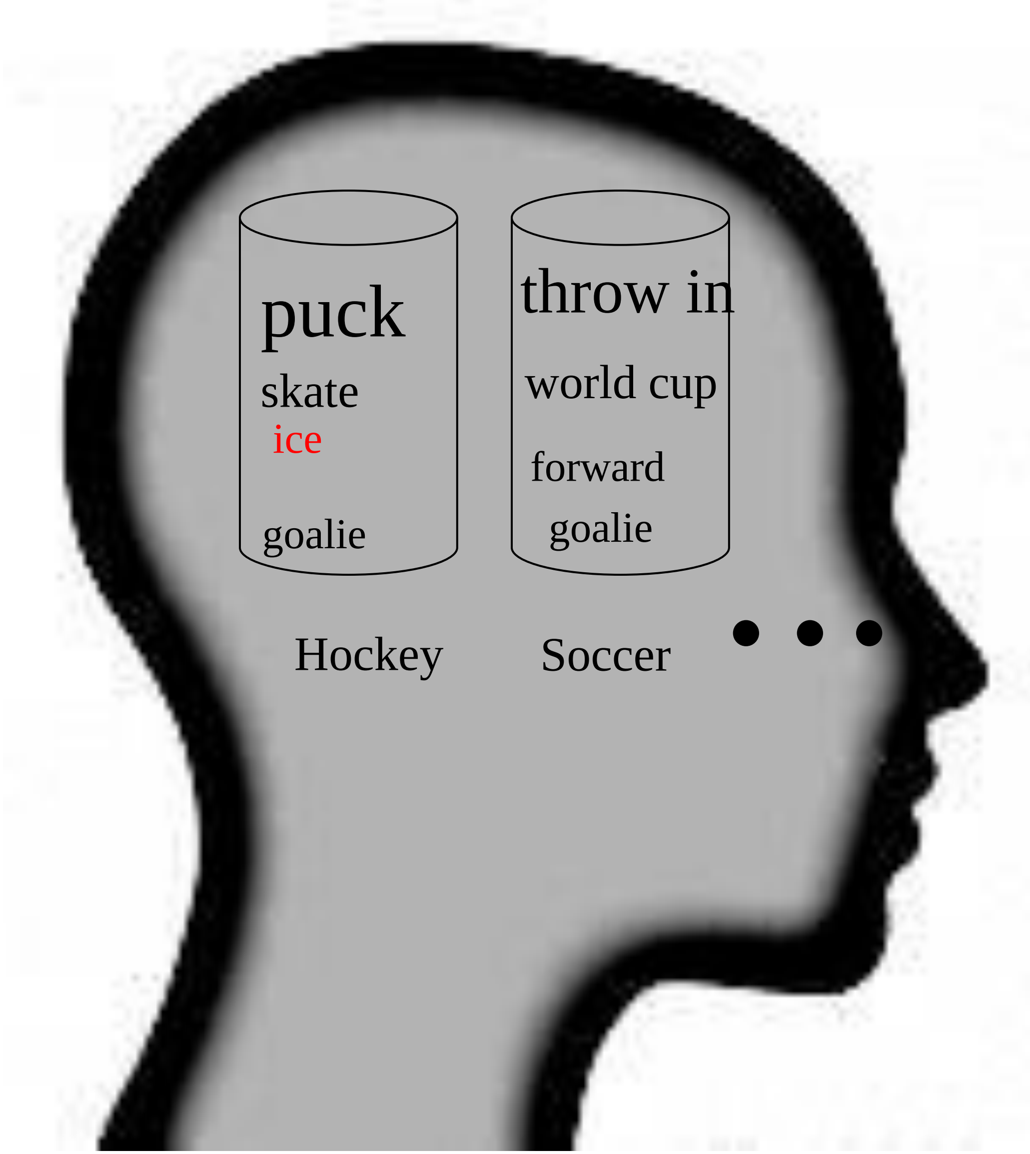
SWIRL for Feature Volunteering



Sports

Order	Item
1	ice →hockey

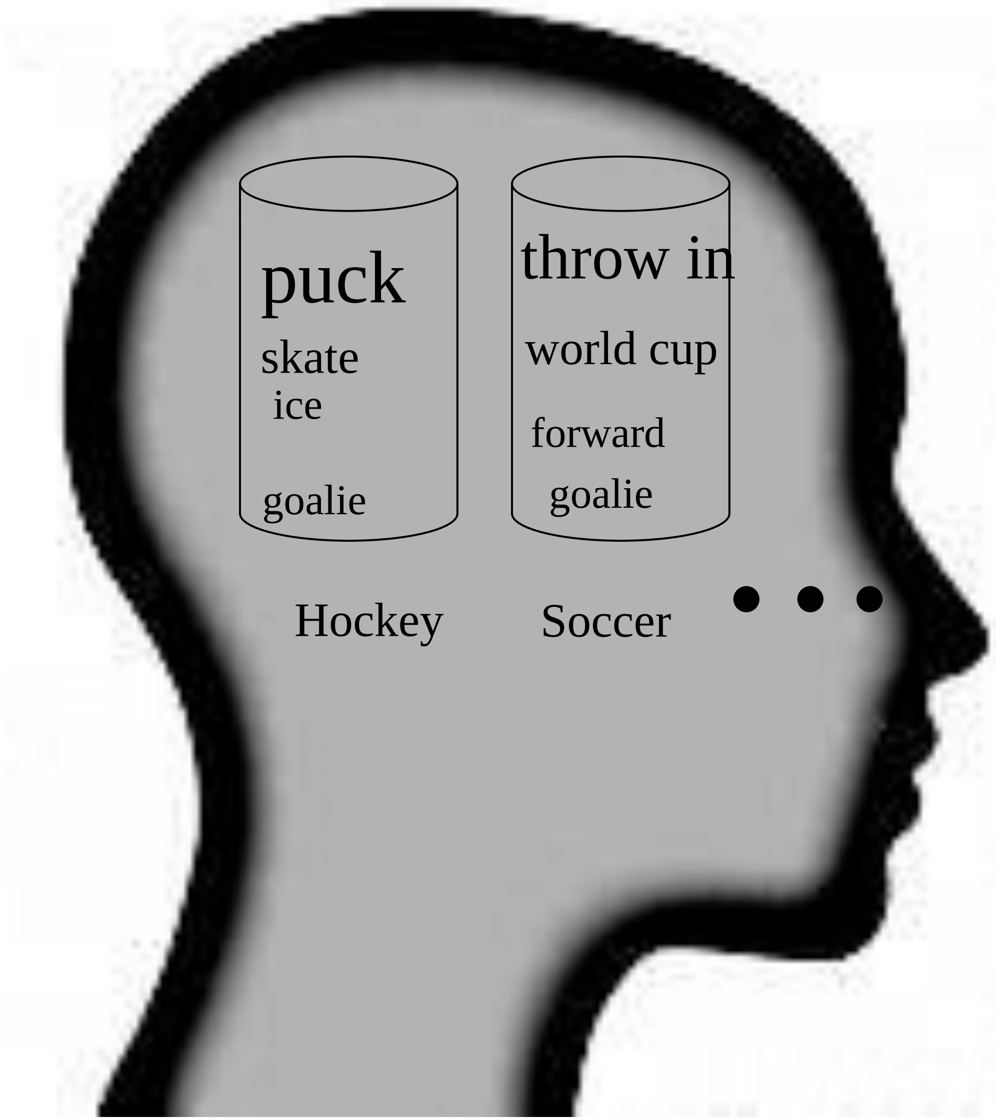
SWIRL for Feature Volunteering



Sports

Order	Item
1	ice →hockey

SWIRL for Feature Volunteering



Sports

Order	Item
1	ice →hockey
…	…

SWIRL Algorithm

- Input: $\boldsymbol{s} = \{s_i \mid i \in V\}, \lambda, \alpha$
- $n \sim \text{Poisson}(\lambda)$
- **for** $t = 1, \dots, n$ **do**
  - $z_t \sim \text{Multinomial}\left(\frac{s_i}{\sum_{j \in V} s_j} \mid i \in V\right)$
  - $s_{z_t} \leftarrow \alpha s_{z_t}$
- **end for**
- Output: $(z_1, \dots, z_n)$

Maximum Likelihood Estimate

- Observed Lists: $\mathbf{z}^{(1)} = \left(z_1^{(1)}, \ldots, z_{n^{(1)}}^{(1)}\right), \ldots, \mathbf{z}^{(N)} = \left(z_1^{(N)}, \ldots, z_{n^{(N)}}^{(N)}\right)$
- $n^{(j)}$: list length of $\boldsymbol{z}^{(j)}$

$$\ell = \sum_{j=1}^{N} n^{(j)} \log \lambda - \lambda + \sum_{t=1}^{n^{(j)}} \log P\left(z_t^{(j)} \mid z_{1:t-1}^{(j)}, \mathbf{s}, \alpha\right)$$

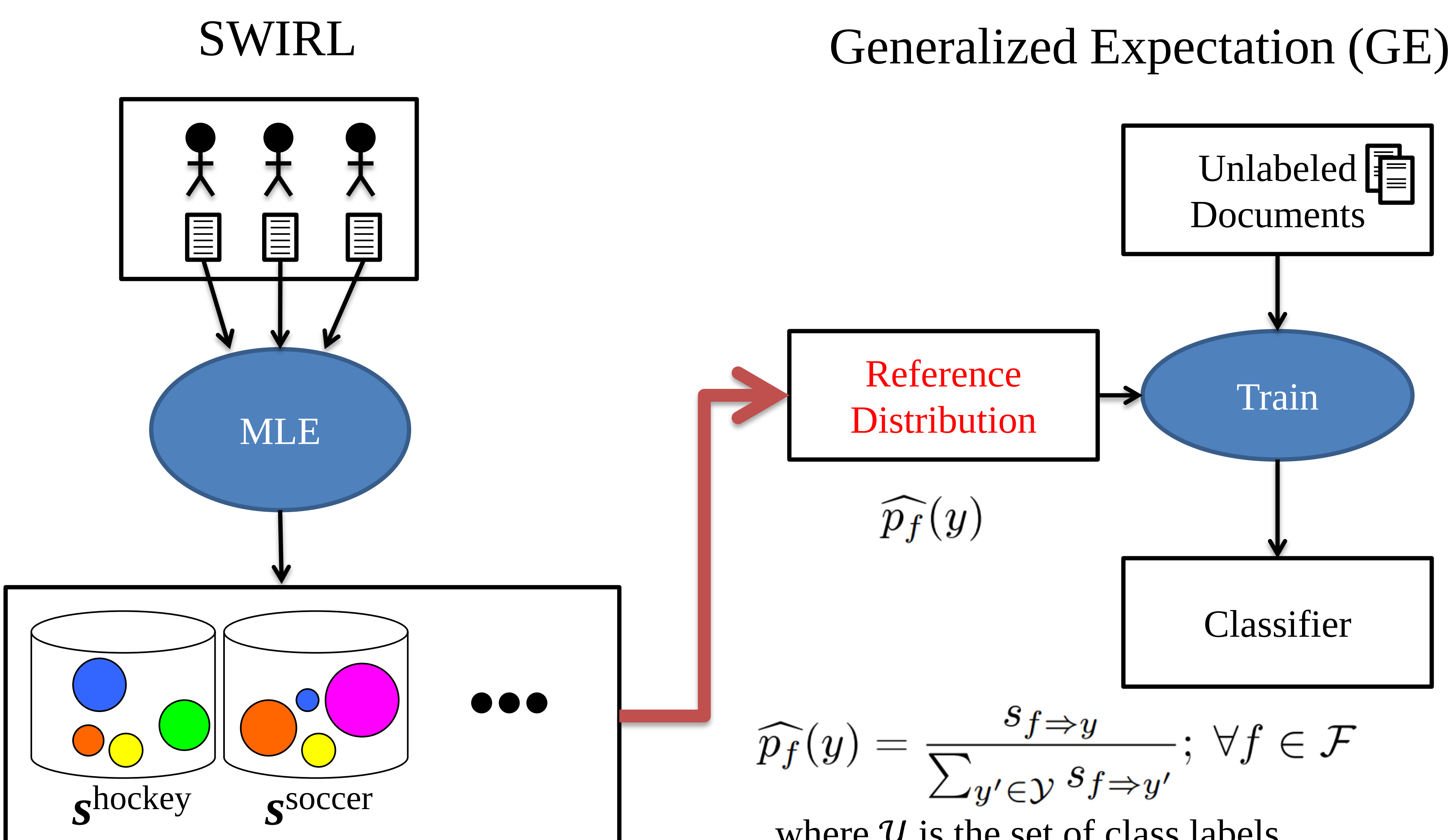
- MLE = $(\hat{\lambda}, \hat{\alpha}, \hat{\boldsymbol{s}})$
- Optimization
  - $\boldsymbol{s}$ is scale invariant: constrain most frequent item's size to 1
  - L-BFGS
  - Concave log likelihood

Application 1: Learning by Feature Volunteering

- Train a text classifier by volunteering feature-label pairs
  - Generalized Expectation (GE) [Druck08]
  - Informative Dirichlet Prior (IDP) [Settles11]

Order	Item
1	baseball bat ⇒ Baseball
…	…
7	quarterback ⇒ Football
8	football field ⇒ Football
9	soccer ball ⇒ Soccer
…	…
23	basketball court ⇒ Basketball
24	football field ⇒ Football
25	soccer field ⇒ Soccer
…	…

Application 1: Learning by Feature Volunteering



$$\widehat{p_f}(y) = \frac{s_{f \Rightarrow y}}{\sum_{y' \in \mathcal{Y}} s_{f \Rightarrow y'}}; \ \forall f \in \mathcal{F}$$

, where $\mathcal{Y}$ is the set of class labels.

Application 1: Learning by Feature Volunteering

Domain	Class Labels	Lists		Documents		Reference Distributions			FV
		N	$\lvert\mathcal{F}\rvert$	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{F}^+\rvert$	SWIRL	Equal	Schapire	
sports	baseball, basketball, football, hockey, soccer	52	594	1123	2948	**0.865**	0.847	0.795	**0.875**
movies	negative, positive	27	382	2000	2514	**0.733**	**0.733**	**0.725**	0.681
webkb	course, faculty, project, student	56	961	4199	2521	**0.463**	0.444	0.429	0.426

N: the # of subjects, $\mathcal{F}$: the set of features (phrases) volunteered in N lists, $\mathcal{U}$: the set of unlabeled documents, and $\mathcal{F}^+$: union of $\mathcal{F}$ and unigrams in $\mathcal{U}$.

SWIRL: $\widehat{p_f}(y) = \frac{s_{f\Rightarrow y}}{\sum_{y'\in\mathcal{Y}} s_{f\Rightarrow y'}};\ \forall f \in \mathcal{F}$, where $\mathcal{Y}$ is the set of class labels.

Equal: $\widehat{p_f}(y) = \frac{\mathbb{1}\{s_{f\Rightarrow y} > 0\}}{\sum_{y'\in\mathcal{Y}} \mathbb{1}\{s_{f\Rightarrow y'} > 0\}};\ \forall f \in \mathcal{F}$

Schapire: Smoothed **Equal** used in [Druck08]

FV: Feature Voting. Non-GE baseline.

Application 2: Verbal Fluency

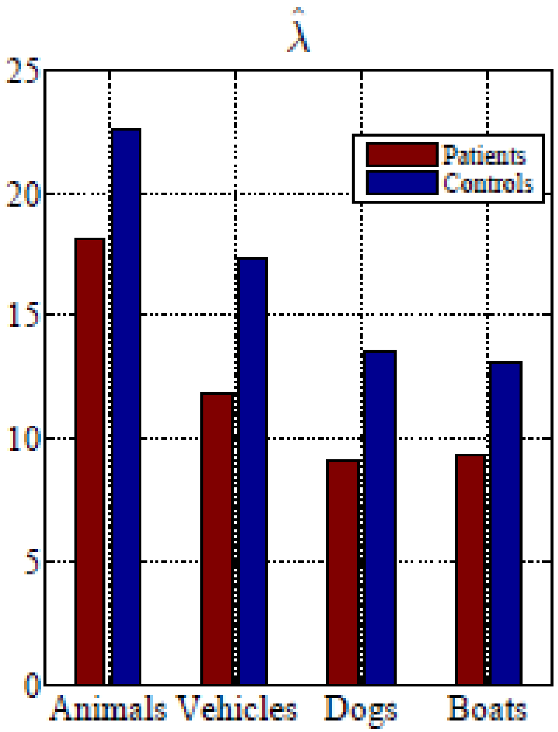
"List examples of animals without repetition for 60 seconds."

- 27 patients / 24 healthy people
- Categories:
  - animals
  - vehicles
  - dogs
  - boats

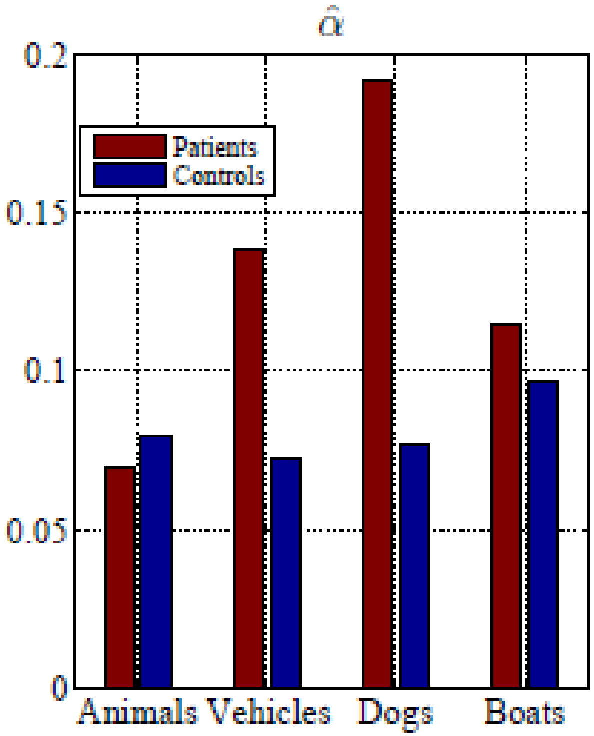
Order	Item
1	dog
2	cat
3	tiger
4	cow
…	…
7	lion
8	tiger
9	bear
…	…
11	armadillo

Application 2: Verbal Fluency

$$(\hat{\lambda}, \hat{\alpha}, \hat{\mathbf{s}})_{patients} \text{ vs. } (\hat{\lambda}, \hat{\alpha}, \hat{\mathbf{s}})_{healthy}$$



Item	$\mathbf{s}_P$	$\mathbf{s}_C$	$\mathbf{s}_W$
cat	.15	.13	.05
dog	.17	.11	.08
lion	.04	.04	.01
tiger	.04	.03	.01
bird	.04	.02	.03
elephant	.03	.03	.01
zebra	.01	.04	.00
bear	.03	.03	.03
snake	.02	.02	.01
horse	.02	.03	.06
⋮	⋮	⋮	⋮



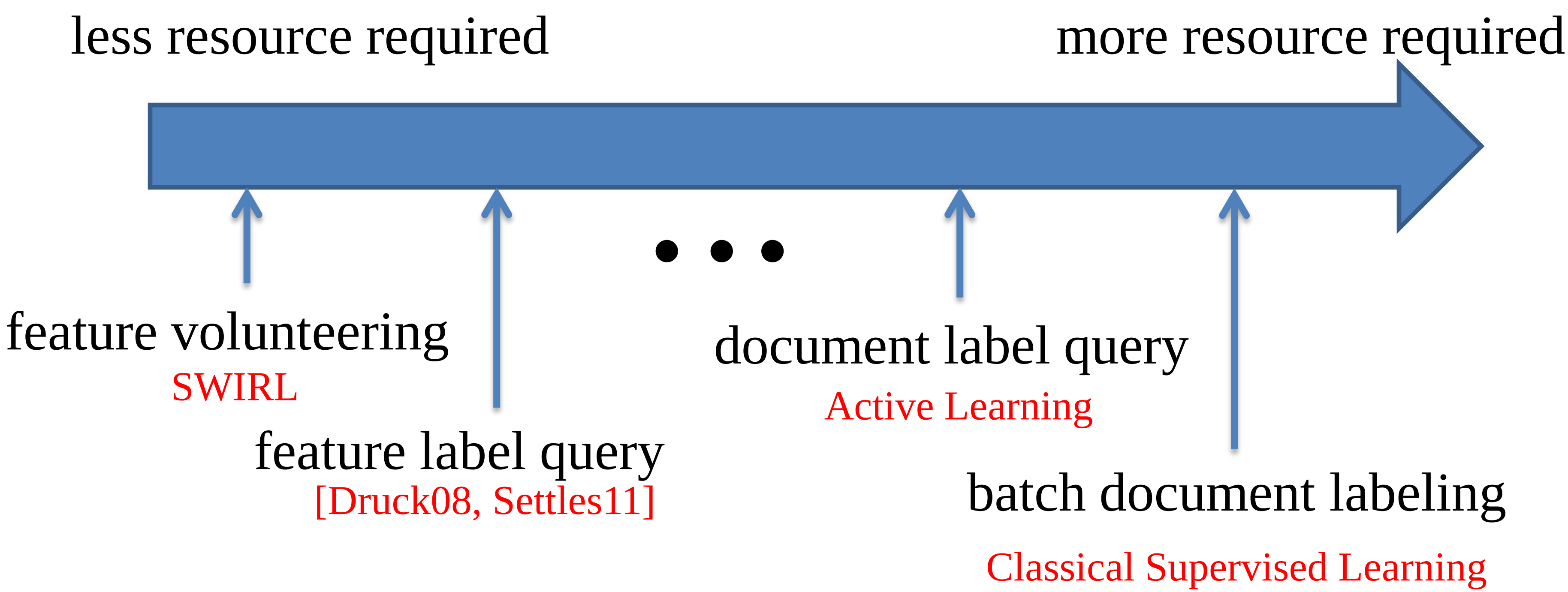
Application 2: Verbal Fluency

- Patient vs. healthy classification

Animals	Vehicles	Dogs	Boats	Majority Vote
0.647	0.706	**0.784**	0.627	0.529

Future Work

- Supervision pipeline



Future Work

- More applications
  - e.g. Hashtags

Tags

salisbury | collection | salisbury collection | slide | old | photo | 35mm | camera | found | found slide | found slides | 1959 | slides | vintage | car | cars | car ferry | ferry | tellsprung | mountain | mountains | june 1959 | june | europe 1959 | their car

Future Work

- Hierarchical SWIRL: individual level parameters as well as group level parameters.

- Structured SWIRL: "runs" of semantically-related items

Order	Item
1	dog
2	cat
3	tiger
4	cow
…	…
7	lion
8	tiger
9	bear
…	…
11	armadillo

Conclusion

Human-generated lists are interesting,
and SWIRL can make them useful!

Code & data are available at: http://pages.cs.wisc.edu/~deltakam (or google "Kwang-Sung Jun").

Acknowledgements

The authors were supported in part by National Science Foundation grants IIS-0953219, IIS-0916038, IIS-1216758, IIS-0968487, DARPA, and Google. The patient data was collected under National Institute of Health grant R03 NS061164 with TR as the PI. We thank Bryan Gibson for help collecting the feature volunteering data and students and faculties in UW-Madison for providing feedbacks.