Landmark Classification in Large-scale Image Collections

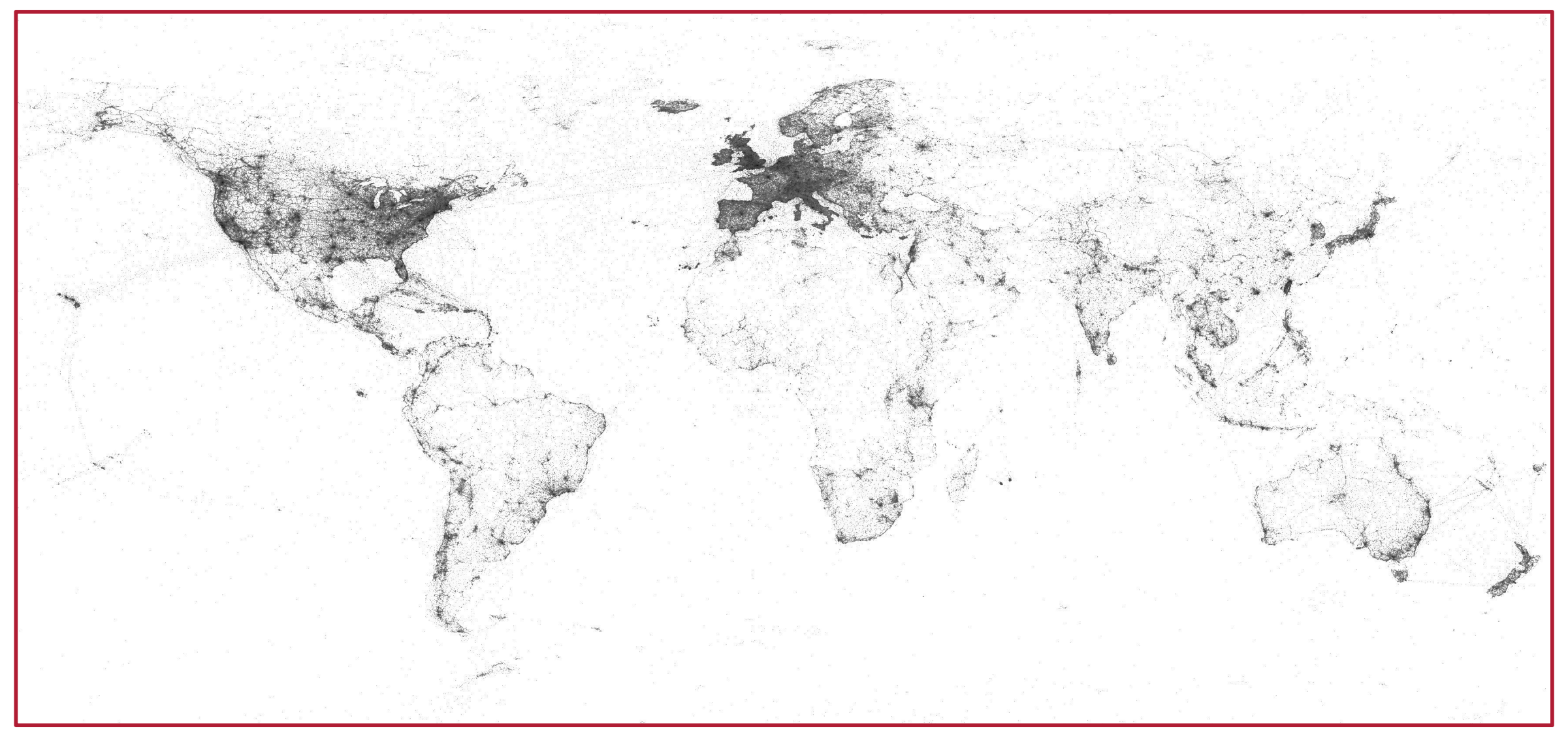
Yunpeng Li, David Crandall, Daniel Huttenlocher
Department of Computer Science, Cornell University, Ithaca, NY USA

Motivation

- Objective: Automatic image classification in web-scale photo collections
- Create and use **large labeled datasets** for performance evaluation
- Exploit **relational information** by jointly classifying a user's *photo stream*
- **Study scalability** of recognition techniques on labeled datasets with **hundreds of categories** and **millions of images**

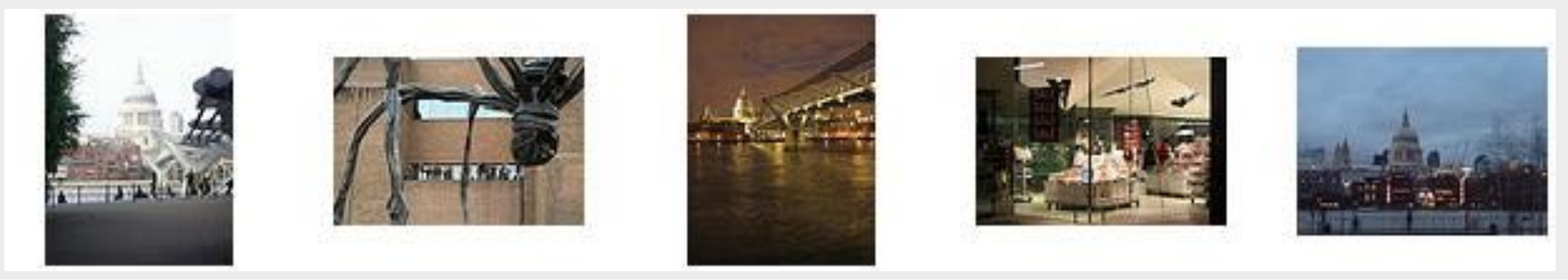
Automatically-generated labeled dataset

- Used **60 million geotagged photos** from Flickr.com, collected via public API
- Identified the 500 **most-photographed landmarks** by locating peaks in the geotag distribution using a **mean shift** procedure [1] with a kernel of radius ~100 meters



- This produced a set of **1.9 million images** each labeled into one of **500 categories**
- Also downloaded photos taken within the same photo stream as the above images (by the same user, within 48 hours), producing a total dataset of **6.5 million images**
- **Dataset generation was completely automatic**, avoiding bias that can be introduced by hand-selecting images, landmarks, or tags

Sample of our dataset: Top 10 (of 500) landmarks

Landmark[1]	Random tags	Random images
1. eiffel	eiffel, city, travel, night, street	
2. trafalgarsquare	london, summer, july, trafalgar, londra	
3. bigben	westminster, london, ben, night, unitedkingdom	
4. londoneye	stone, cross, london, day2, building	
5. notredame	2000, portrait, iglesia, france, notredamecathedral	
6. tatemodern	england, thames, greatbritian, streetart, vacation	
7. empirestate-building	manhattan, newyork, travel, scanned, evening	
8. venice[2]	tourists, slide, venecia, vacation, carnival	
9. colosseum	roma, england, stadium, building, italy	
10. louvre	places, muséedulouvre, eau, paris, canon	

[1] To describe each landmark we show the tag with the greatest ratio between the frequency within the landmark and the overall frequency worldwide.
[2] This landmark is Piazza San Marco in Venice.

Photo classification

- Most Internet photos include non-visual metadata, e.g. textual annotations (text tags), who took the photo (user id), when it was taken (EXIF timestamp), etc.
- We used both **visual features** and **text tags** for classification
- Additionally, we used **constraints on the the sequence of photos** taken by the same user at about the same time (a user's *photo stream*)
- Intuition: Some transitions are likely, while others are implausible, e.g.:

User X:

9:50am

9:55am (same day)

Labeling 'venice'-'venice' is likely

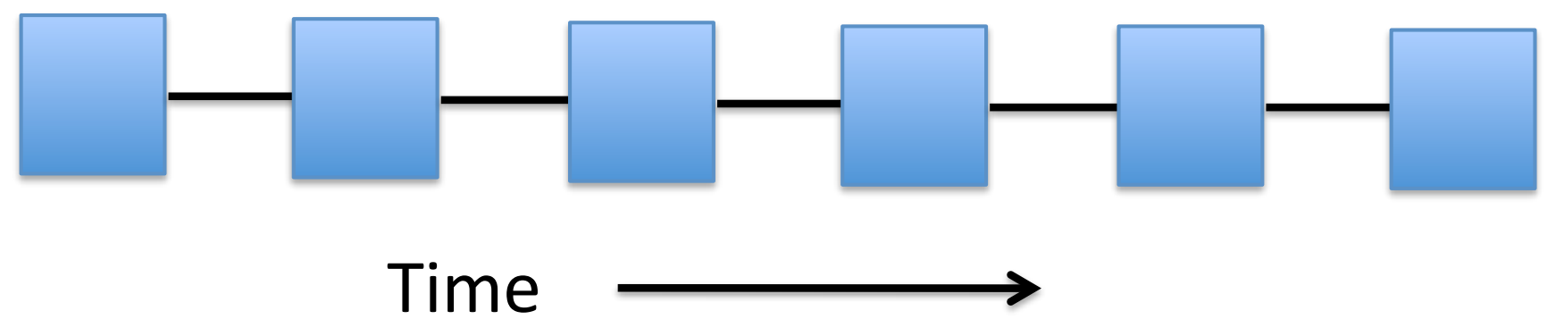
Labeling 'venice'-'bigben' is not (London to Venice in 5 minutes?!)

Classifying individual images

- **Bag-of-words model** [2], with vector-quantized SIFT descriptors as visual words
- Simple **vector space model** to represent text tags
- Linear classifier: **Multi-class SVM** (special-case of a structured SVM [3])
- Set of labels: Landmarks identified by above geo-clustering

Joint classification of a user's photo stream

- Modeled labeling of the photo steam as a **structured output**

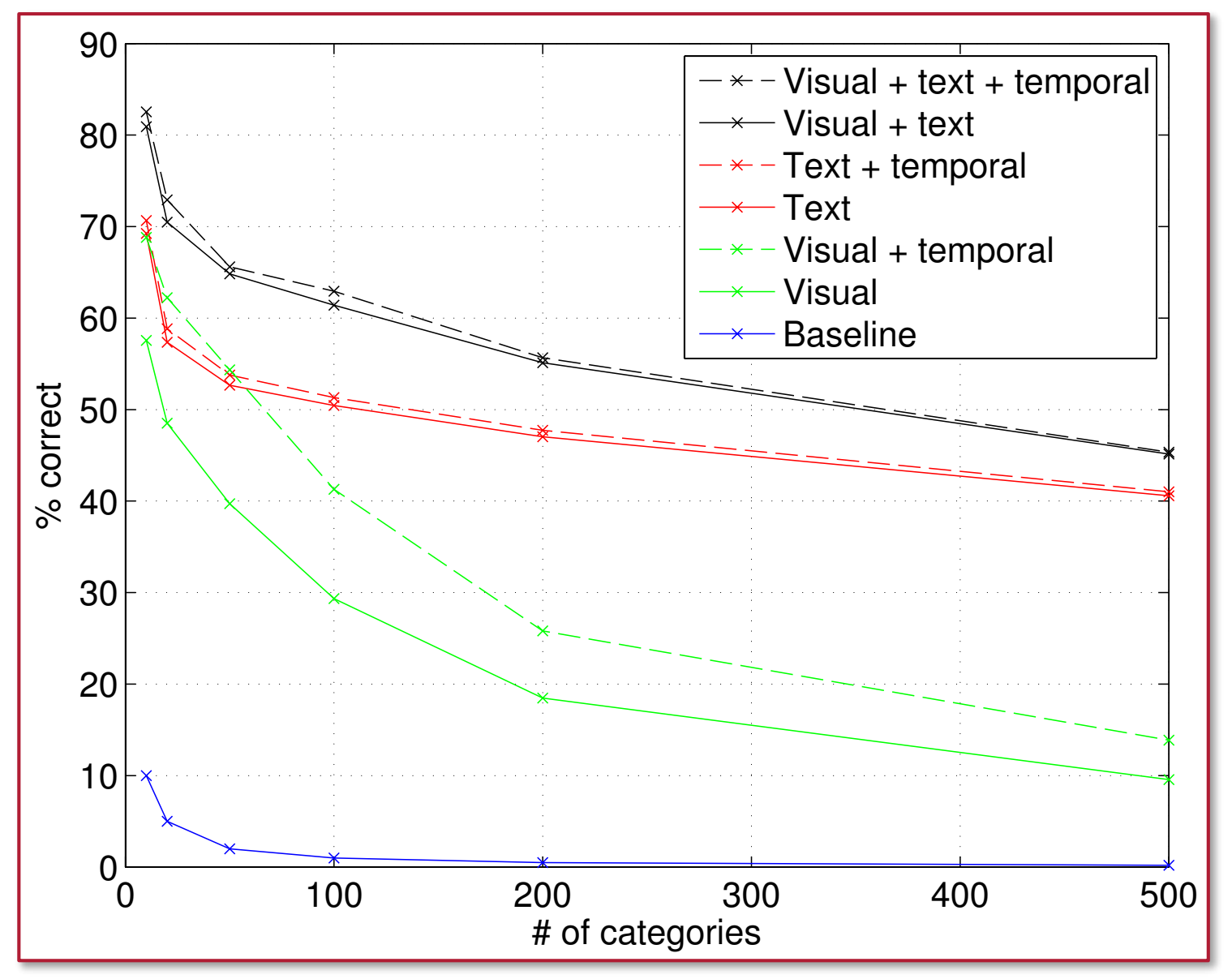


- Learned both transition and visual models together using **structured SVMs** [3]
- Used **the Viterbi algorithm** during learning (for finding most violated constraints) as well as for classification (for finding the best labeling)

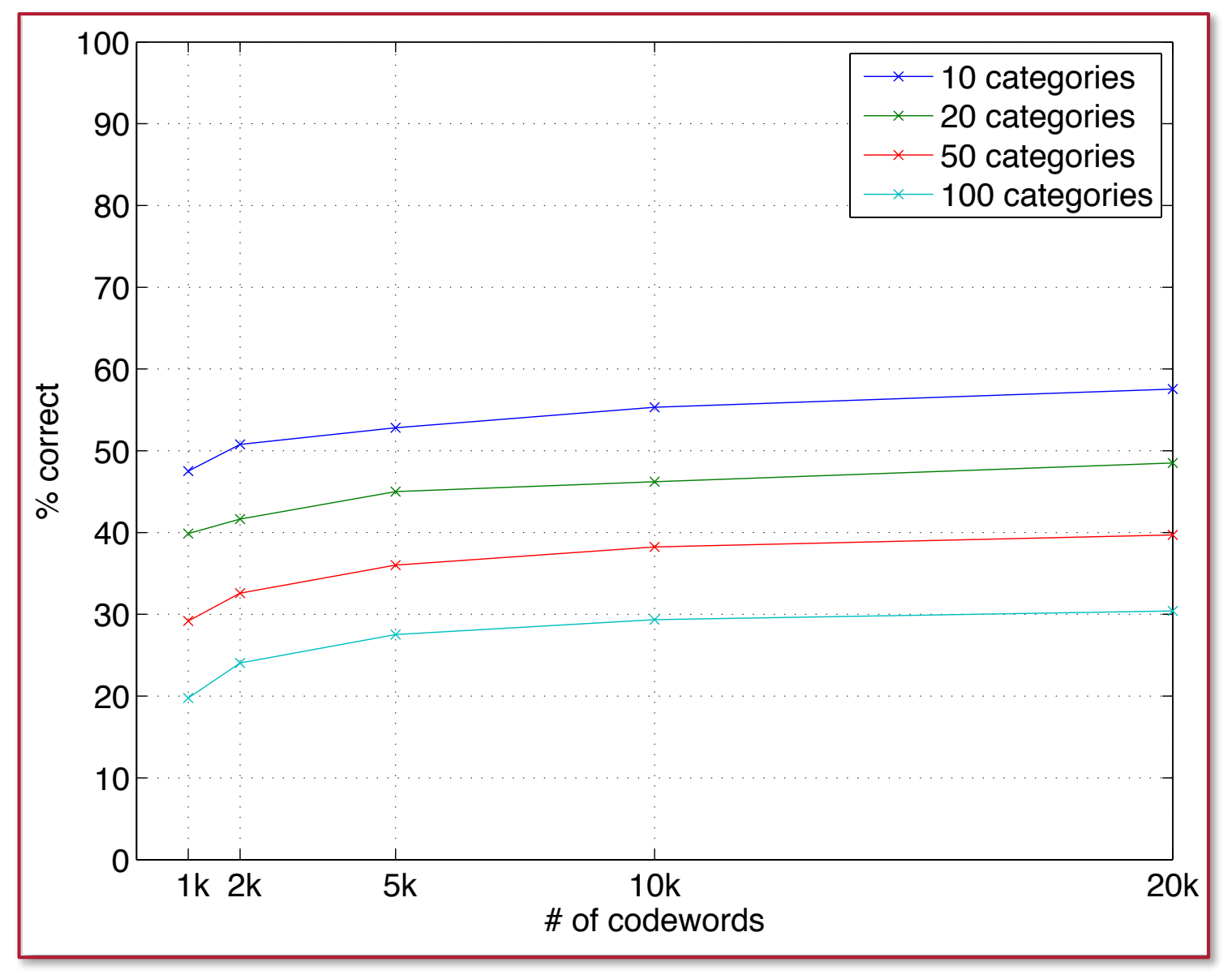
Experimental results

- **Measured landmark classification performance** with varying numbers of classes, and using combinations of **visual features, textual tags, and photo streams**
- We sampled from the dataset to produce **an equal number of images in each class**
- To prevent bias introduced by any single user, we partitioned test/training sets by photographer, and sampled a limited number of images from each photographer
- All experiments **involved tens or hundreds of thousands of images**

Classification rate vs. # of categories



Classification rate vs. vocabulary size



Percentage of images correctly classified

		Single images			Photo streams		
Categories	Baseline	Visual	Textual	Combined	Visual	Textual	Combined
Top 10 landmarks	10.00	57.55	69.25	80.91	68.82	70.67	82.54
Landmarks 200-209	10.00	51.39	79.47	86.53	60.83	79.49	87.60
Landmarks 400-409	10.00	41.97	78.37	82.78	50.28	78.68	82.83
Top 20 landmarks	5.00	48.51	57.36	70.47	62.22	58.84	72.91
Landmarks 200-219	5.00	40.48	71.13	78.34	52.59	72.10	79.59
Landmarks 400-419	5.00	29.43	71.56	75.71	38.73	72.70	75.87
Top 50 landmarks	2.00	39.71	52.65	64.82	54.34	53.77	65.60
Top 100 landmarks	1.00	29.35	50.44	61.41	41.28	51.32	62.93
Top 200 landmarks	0.50	18.48	47.02	55.12	25.81	47.73	55.67
Top 500 landmarks	0.20	9.55	40.58	45.13	13.87	41.02	45.34
Top 10 landmarks, human performance[3]	10.00	68.00	---	76.40	---	---	---
Top 10 landmarks, actual priors	14.86	53.58	---	79.40		---	---

[3] Mean performance on a small study of 20 people, with σ=11.61, 11.91.

Conclusions

- **Combination of vision and text tags does better** than either alone
- **Using photo streams improves visual classification significantly**, performing about the same as text tags. The improvement is minor when using only text tags
- Increasing size of visual vocabulary improves recognition (up to at least 80K words)
- **Classifier does about as well as humans** when using tags + visual features

[1] Crandall, Backstrom, Huttenlocher, Kleinberg. Mapping the World's Photos, WWW 2009.
[2] Csurka, Dance, Fan, Williamowski, Bray. Visual Categorization with Bags of Keypoints, ECCV 2004.
[3] Tsochantaridis, Hofmann, Joachims, Altun. Support Vector Machine Learning for Interdependent and Structured Output Spaces, ICML 2004.

Supported in part by NSF grants BCS-0537606, IIS-0705774, and IIS-0713185, and used the resources of the Cornell University Center for Advanced Computing, which receives funding from Cornell, New York State, NSF, and other agencies, foundations, and corporations.