# Do Objects in Real-World Images Have a Canonical Name?

Anonymous EMNLP-IJCNLP submission


## Abstract

While research in language & vision is well aware that there is variation in scene description, object naming has been addressed in a comparatively simplistic way: Typically, a single object label is provided for each object, and assumed to be its canonical category or name. Inspired by research in Cognitive Science, we study variation in object naming by eliciting dozens of object names for the same object. Unlike work in that field, we do so for natural images (25K objects in the VisualGenome dataset). The resulting dataset, ManyNames, reveals a substantial level of agreement in how humans call individual instances, and lower agreement across instances. Analysis shows that most of the name variants do not correspond to taxonomic relations (as encoded in WordNet), which have been the focus of Psycholinguistic studies, and that referential uncertainty is a major source of variation. We investigate whether a state-of-the-art model of object labeling implicitly encodes similar variation in object naming and discuss implications for research in language & vision.


## 1 Introduction

Expressions describing or referring to objects in visual scenes typically include object names: e.g., *cheesecake* or *dessert* in Figure **??**. Determining these object names is a core aspect of virtually every language & vision task, ranging from e.g. referring expression generation to visual dialogue (**?**). We investigate the extent to which there is variation in the names chosen by different people for the same object, and its implications for research in language & vision.

Our paper puts together two strands of research that have mostly been pursued independently to date. On the one hand, state-of-the-art computer vision systems are able to accurately classify images into thousands of different categories (e.g.

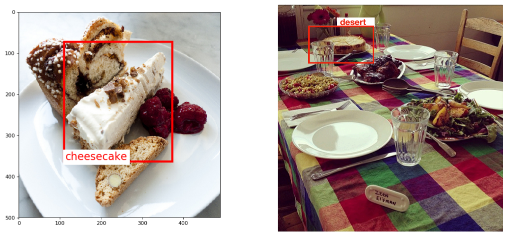


Figure 1: Two objects of the same type of cake, with different names in VisualGenome

Szegedy et al. (2015)), where the task is often to predict the name for a given object. *//g: Is this true? Imagenet task asks for synsets, which can be taken to be categories. . . To refine//* However, they mostly adopt very simple assumptions with respect to the underlying lexicon, which is implemented as a simple, flat labeling scheme: A standard object recognition system would be trained to classify the left object in Figure 1 as *cheesecake*, the right one as *dessert*, and using *dessert* for the left picture would be considered incorrect. On the other hand, research on object naming in Cognitive Science has shown that people choose different names depending on the circumstances, with factors such as context or the prototypicality of the object with respect to the category playing a role (**?**). *//g: This research also argues that there is high agreement in how people name objects; to do: make coherent.//* However, this research typically uses stylized drawings are used, and is focused on taxonomic relations (*sparrow-bird*). *//sz: It is thus unclear how findings from these stylized settings generalize to tasks in language & vision like referring expression generation, where naming is a core aspect. Therefore, in contrast to traditional naming norm studies in Cognitive Science we study object naming in realistic scenes where objects are situated in a natural context! (This comes with additional challenges, like potential* *object occlusion, background/foreground confusion etc.)//*

In our study, we collect large-scale object naming data via crowdsourcing. Like object naming studies in Cognitive Science, we collect multiple names per object (concretely, 36); like most work on language & vision, we use natural images *//sz: (showing objects in complex visual contexts, surrounded by other objects, not ImageNet-like images)//* on a large scale, annotating objects in 25K images from the Visual Genome dataset. We analyze the agreement in object naming across subjects, and the sources of variation. We find that: *//g: To be put in paragraph form//*

- there is quite a high level of agreement in the task, with the relative frequency of the most common name being 70% on average. This is in accordance with previous results in Cognitive Science (**?**);
- the level of agreement in object naming is much higher in certain domains than in others; as it happens, the domains that have been traditionally used in object naming research (e.g. animals) seem to display the highest amount of agreement in our data set;
- most of the variation in our dataset comes from alternative names that do not stand in a taxonomic relation, suggesting that the previous work in Cognitive Science is missing much of the empirical ground.

  our datasets contains a lot of variability for names coming from different parts of the taxonomy (*dessert* vs. *cake*, *bottle* vs. *wine*)

Moreover, we analyze whether current models implicitly encode the variation in naming, by doing XXX. We find YYY.

## 2 Data collection

We take data from VisualGenome (**?**), which contains a dense region-based labeling of $108k$ images with object descriptions, attributes, and relationships, as well as question-answer pairs, all linked to WordNet synsets (**?**). VisualGenome is suitable for our purpose of collecting names for a relatively large amount of instances of common objects in naturalistic images, as it has images of varying complexity, with close-ups as well as complex images with many objects. As common in Computer Vision, objects are identified via bounding boxes (see red boxes in Figure 1).[1]

### 2.1 Sampling of Instances

We selected seven domains: six from McRae et al.'s feature norms (REF), a dataset widely used in Psycholinguistics that consists of common objects of different categories (e.g., ANIMALS, FURNITURE), and PERSON, because it is very frequent category in VisualGenome.

Within each domain, we aimed at collecting instances at different levels in a taxonomy to cover a wide range of phenomena, but it is not straightforward to do so because ontological taxonomies do not align well with the lexicon (for instance, *dog* and *cow* are both mammals, but *dog* has many more common subcategories), and most domains are not organized in a clear taxonomy in the first place (e.g. HOME). Instead, we defined a set of synsets that we would use to collect our object instances from VisualGenome, balancing variability. From the set of synsets that match or subsume the concepts in the McRae norms, we kept those that had a high number of VisualGenome object instances of different names. For example, VisualGenome instances subsumed by McRae's *dog* were named *beagle*, *greyhound*, *puppy*, *bulldog*, etc., while McRae's *duck*, *goose*, or *gull* did not have name variants in VisualGenome, so we kept *dog* and *bird* (which subsumes *duck*, *goose*, or *gull*) as collection synsets.

We then retrieved all VG images depicting an object whose name matches or is subsumed by words in one of these synsets; we refer to these words as *seeds*, and we had XXX*//g: Carina?//* of them. We did not consider objects with names in plural form, with parts-of-speech other than nouns[2], or that were multi-word expressions (e.g., *pink bird*). We further only considered objects whose bounding box covered an area of $20-90\%$ of the image.

Because of the Zipfian distribution of names, and to balance the collection, we sampled instances depending on the size of the seeds: up to 500 instances for seeds with up to 800 objects, and up to 1000 instances for larger seeds. **double-check** This yielded a dataset with $31,093$ in-

[1] We use image and object interchangeably in the following, since we only selected one target object per image (i.e., each object and image in VG is chosen at most once).

[2] (REF to tagger)

stances, which was further pruned during annotation (see next subsection).

*//g: Checked up to here//*.

Table ?? gives an overview of the collection synsets, XXX, XXX, grouped into 7 domains. (**Report only final dataset, with a note in caption/footnote referring to the checkpoint pruning.**)

Number of images/objects: 25,596
Number of object names: 450
Number of collection nodes (synsets): 52

## 2.2 Procedure

describe the crowdsourcing set-up and the task TODO: Footnote: we ran pilot experiments to design our experiment and instructions.

### Collection Method

- instructions;
  Appendix ?? gives the instructions as they were presented to the annotators.
- each round: HIT of 10 instances, collect 9 annotations for each HIT
- round 0 (with opt-outs) → pruning → rounds 1-3 (no opt-outs)
  pruning: Based on given opt-outs: keep images with no OCCLUSION, at most BBOX is ambiguous twice, at most 17% of names in plural form, most frequent names is of same domain as VG name (gives $25,596$, i.e., discard $5,497$ instances)
- workers could only participate in one round, such as to avoid workers annotating an instance more than once.

Overall $XX$ participants, each annotated between $XX$ and $XX$ instances. *//g: Put mean or median instead of min-max//*

## 2.3 Data

give an overview of the collected data

# 3 Analysis: Agreement

In this section, we investigate to what extent names annotated in VisualGenome and elicited in ManyNames can be considered canonical, i.e. to what extent speakers agree in their naming choices. Whereas traditional picture naming studies typically use a prototypical image per category and, hence, are mostly interested in the agreement on concept or category-level, we carry out an anaylsis on two different levels: First, we will look at instances and see to what extent names overlap for the same object. Second, we will uses the existing annotation of names in VisualGenome to analyze agreement on the level of categories.

## 3.1 Measures

We compute the following agreement measures:

- **% top**: the average relative frequency of the most frequent response (shown in percent)
- $H$: the $H$ agreement measure used previously in the psycholinguistic literature
- **N**: the average number of types in the response set of ManyNames
- $\mathbf{N}_{>1}$: the average number of types, excluding types that have been annotated only once
- **top=VG**: the proportion of items where the top response in ManyNames corresponds to the VisualGenome name
- **% VG**: the average relative frequency of the VisualGenome name in the response set

For measuring **instance-level agreement**, we consider all names annotated for an object as a response set and then average over these response sets. Furthermore, we compute **category-level agreement** by merging the response sets for all objects that have the same VisualGenome name and compute the measures over these aggregated response sets. *//g: I'd call it "name-level agreement" – it's not really category, is it?//*

## 3.2 Results

Table 3 shows the analysis of the instance-level and category-level agreement. On the instance-level, our annotators achieve a fair amount of overlap in their object naming choices. Thus, for roughly 70% of our objects, the most frequent response in ManyNames corresponds to the original VisualGenome name and, similarly, the average frequency of the top response is also 70%. Generally, this seems to suggest that indeed many objects in our data set have a canonical name. At the same time, the average number of name types per

| vehicles | food | animals_plants | home | buildings | people | clothing |
|---|---|---|---|---|---|---|
| train (954) | pizza (518) | giraffe (915) | bed (888) | house (340) | boy (853) | shirt (904) |
| car (642) | cake (261) | horse (822) | bench (714) | bridge (274) | man (806) | jacket (451) |
| plane (485) | bread (186) | cat (754) | table (687) | dugout (91) | woman (766) | coat (267) |
| airplane (479) | sandwich (153) | dog (654) | desk (672) | tent (53) | girl (650) | dress (190) |
| motorcycle (466) | bun (143) | zebra (461) | counter (516) | restaurant (33) | lady (342) | hat (77) |
| truck (465) | cheese (110) | cow (324) | couch (366) | overpass (23) | guy (330) | t-shirt (62) |
| boat (450) | donut (78) | bird (295) | chair (365) | grill (22) | child (230) | tie (51) |
| jet (106) | salad (70) | sheep (216) | carpet (307) | garage (18) | batter (110) | blazer (43) |
| aircraft (85) | sauce (68) | bull (48) | bowl (219) | hotel (16) | kid (85) | hood (26) |
| van (76) | apple (33) | flower (40) | curtain (182) | castle (14) | skateboarder (80) | cap (20) |

Table 1: Overview of our dataset: Top-10 VG names for each domain (number of instances in parentheses). **double-check**

| animals_plants | vehicles | home | clothing | buildings | food | people |
|---|---|---|---|---|---|---|
| ungulate$_1$ (2037) | aircraft$_1$ (1208) | furnishing$_2$ (5355) | shirt$_1$ (968) | house$_1$ (364) | dish$_2$ (812) | woman$_1$ (1768) |
| horse$_1$ (833) | train$_1$ (957) | vessel$_3$ (525) | overgarment$_1$ (786) | bridge$_1$ (297) | baked_goods$_1$ (770) | man$_1$ (1167) |
| feline$_1$ (763) | car$_1$ (727) | kitchen_utensil$_1$ (132) | dress$_1$ (199) | shelter$_1$ (169) | foodstuff$_2$ (280) | male_child$_1$ (853) |
| dog$_1$ (688) | motorcycle$_1$ (564) | crockery$_1$ (92) | headdress$_1$ (135) | restaurant$_1$ (58) | vegetable$_1$ (48) | athlete$_1$ (396) |
| bird$_1$ (389) | truck$_1$ (559) | cutlery$_2$ (82) | neckwear$_1$ (65) | outbuilding$_1$ (31) | edible_fruit$_1$ (42) | child$_1$ (333) |
| flower$_1$ (44) | boat$_1$ (499) | tool$_1$ (72) | robe$_1$ (27) | hotel$_1$ (19) | beverage$_1$ (23) | creator$_2$ (11) |
| rodent$_1$ (27) | ship$_1$ (38) | lamp$_1$ (34) | glove$_2$ (7) | housing$_1$ (17) | | professional$_1$ (5) |
| insect$_1$ (12) | | | footwear$_1$ (5) | place_of_worship$_1$ (12) | | |
| fish$_1$ (11) | | | | | | |

Table 2: Overview of our dataset: Synset nodes for each domain (subscript indicates synset number; number of instances in parentheses). **double-check**

object (5.7, or 2.9 when excluding low-frequency types in each response set) suggests that there is a stable amount of naming variants that is elicited for instances. Furthermore, the agreement varies quite considerably among domains: in the animal domain, which is often discussed in the object naming literature, annotators achieve a very stable and robust agreement over 90% and an $H$ agreement which comes close to 0 (where 0 is perfect agreement). The people domain, on the other hand, is subject to much more variation and agreement is dramatically lower here, and comes close to 50% for % top.

*//g: Super-interesting results.//* Finally, the category-level agreement figures tell yet another story: when aggregating the responses for all objects with the same VisualGenome name, we obtain on average 28 types (with $n > 1$), i.e. 27 variants of the original VG name. Surprisingly, here, only 29.4% of the aggregated response sets still have the VG name as the most frequent response, which means that for 70% of the VG names, annotators in ManyNames, on average, prefer a different name. Likewise, the relative frequency of the top response drops considerably and $H$ increases from 1.3 for instance-level agreement to 2.4 on object-level agreement. What does this discrepancy between the instance-level and category-level agreement in VisualGenome and ManyNames naming choices mean? First of all, it suggests that the same original VisualGenome name can trigger very different variants depending on the visual instance, leading to a drastic increase of variants elicited for categories as compared to instances. Second, this clearly shows that annotators in VG do not generally annotate the most canonical name and that many names annotated for objects in VG do not correspond to the overall most preferred variant. *//sz: think more ...// //g: I don't think we can conclude this second part – we do have the 70% top=VG figure that says that VG annotators annotate the most canonical name. What this suggests to me is that instance-level properties are more important than category-level properties, somehow. That is, there are systematic properties of instances that make them have a single most salient name. However, I expect that this result will be very influenced by referential uncer-*

*tainty (in single images, it will mostly be clear that it's a man, but in some it may be unclear →high instance agreement, low category agrement.//*

### 3.3 Qualitative Analysis

*//sz: put qualititative discussion here//* Table 4 shows examples for canonical and non-canonical VG names in our data set, where canonical means that the name was the top response in the aggregated response set in ManyNames.

*//g: The non-canon. VG names suggest that people prefer more general names ("car > sedan", "horse > pony", "tie > necktie"). Could be due to lexical availability (more general →more frequent →more available). This could be verified (using frequency). Hypothesis: In cases where top name != VG, the VG name is less general. Could be also a more general hypothesis: see if people prefer more frequent names in general.//*

## 4 Analysis: Taxonomy

Previous work on large-scale collections of labels or names of objects has (explicitly or implicitly) assumed that once naming data is canonical, linguistic alternatives of the canonical name can simply be retrieved from existing taxonomies like e.g. WordNet. If this was indeed the case, it would be feasible (and probably even desirable) to canonicalize object names during dataset collection, without loosing too much information about linguistic variations in natural object naming scenarios (like e.g. referring expression generation). Hence, in this Section, we investigate to what extent the variation in object naming that we find in our MN data set (see previous Section) is covered by WordNet.

### 4.1 Lexical relations

In this section, we take a closer look at the lexical variation we observe in our data set. We analyze the data points where participants attributed different names to the same object and extract a set of pairwise **naming variants**. These naming variants correspond to pairs of words that can be used interchangeably to name certain objects. For each object, we extract the set of naming variants $s = \{(w_{top}, w_2), (w_{top}, w_3), (w_{top}, w_4), ...\}$ where $w_{top}$ is the most frequent name annotated for the object and $w_2 ... w_n$ constitute the less frequent alternatives of $w_{top}$. The **type frequency** of a naming variant $(w_{top}, w_x)$ corresponds to the number of objects where this variant occurs. The **token frequency** of $(w_{top}, w_x)$ corresponds the count of all annotations where $w_x$ has been used instead of $w_{top}$. In Table **??**, we show the the naming variants with the highest raw token frequency for each domain.

The naming variants can be grouped according to their lexical relation, as follows:

- **synonymy**: e.g. aircraft vs. airplane
- **hyponymy**: e.g. man vs. person
- **co-hyponymy**: e.g. swan vs. goose
- **no relation**: e.g. desk vs. apple

Research on object naming following the idea of entry-level categories has, essentially, exclusively looked at names that stand in a hierarchical relation (i.e. hyponymy/hypernymy).

We use WordNet to extract lexical relations between the naming variants in our data set. Unfortunately, this means that we have to exclude a certain portion of the data as either (i) one of the name is not covered in WordNet, (ii) we cannot find a lexical relation between the two names (see below). Also, we had to be relatively permissive with respect to the definition of hyponymy/co-hyponymy. For instance, to analyze *giraffe* as a hyponym of *animal* we have to look at the closure of the hyponyms of *animal* with a depth of 8 (in WordNet). *//sz: should we call this co-hyponymy or co-hierarchical relation?//*

*//sz: include Table that reports counts of the naming variants, coverage in WordNet etc.// //g: I think it'd be best to put the out-of-wordnet info in the Lexical relations table – this way we have everything in one place.//*

Table 5 shows the distribution of lexical relations for those naming variants that we were able to analyze with WordNet. Both in terms of their types and token frequency, the naming variants that instantiate a (loose) co-hyponymy relation are by far the most frequent. *//sz: discuss in more detail, discuss: to what extent is this an artefact of WordNet?//* This is really interesting: most research on object naming, to date, has focussed on hyponymy/hypernymy, i.e. variation that relates to hierarchical relations between object names. Our data suggests that co-hierarchical variation is really important too.

| | Instance-level agreement | | | | | | | Category-level agreement | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| domain | % top | $H$ | N | $N_{>1}$ | top=VG | % VG | # Obj | % top | $H$ | N | $N_{>1}$ | top=VG | % VG | # Cat |
| people | 51.9 | 2.1 | 8.6 | 4.3 | 49.8 | 32.3 | 4533 | 43.8 | 2.9 | 88.5 | 45.1 | 20.0 | 10.9 | 55 |
| clothing | 63.9 | 1.6 | 6.4 | 3.2 | 70.2 | 52.6 | 2192 | 50.6 | 2.5 | 68.3 | 32.5 | 38.5 | 24.6 | 39 |
| home | 66.4 | 1.5 | 6.3 | 3.1 | 78.5 | 58.8 | 6292 | 50.7 | 2.7 | 90.6 | 42.6 | 39.3 | 24.9 | 89 |
| buildings | 66.9 | 1.5 | 6.9 | 3.0 | 72.6 | 55.5 | 967 | 47.8 | 2.9 | 59.9 | 27.2 | 27.8 | 19.2 | 36 |
| food | 71.3 | 1.3 | 5.5 | 2.9 | 62.9 | 52.1 | 1975 | 47.0 | 2.5 | 31.5 | 15.0 | 29.3 | 19.3 | 92 |
| vehicles | 72.0 | 1.1 | 4.7 | 2.4 | 71.1 | 60.2 | 4552 | 56.5 | 2.0 | 63.3 | 30.0 | 18.4 | 17.9 | 49 |
| animals,plants | 91.3 | 0.4 | 2.7 | 1.5 | 93.8 | 88.0 | 4804 | 67.6 | 1.5 | 26.5 | 12.3 | 28.1 | 25.7 | 89 |
| all | 69.7 | 1.3 | 5.7 | 2.9 | 72.8 | 58.7 | 25315 | 52.8 | 2.4 | 58.2 | 27.8 | 29.4 | 20.9 | 449 |

Table 3: Agreement in naming measured on the level of instances and on the level of VG categories (i.e. after grouping objects by their VG name)

| VG name | top5 MN names | $n_{obj}$ |
|---|---|---|
| | *Canonical VG names with max agreement in MN* | |
| giraffe | giraffe (96.8), animal (1.2), zebra (0.4), camel (0.3), pole (0.1) | 915 |
| zebra | zebra (96.3), animal (1.0), giraffe (0.9), horse (0.2), microwave (0.2) | 461 |
| cat | cat (94.8), animal (0.9), kitten (0.8), dog (0.4), laptop (0.2) | 754 |
| | *Canonical VG names with min agreement in MN* | |
| booth | booth (19.3), table (12.3), phone booth (9.8), bench (6.7), building (4.4) | 11 |
| cabbage | cabbage (21.4), lettuce (17.0), hotdog (11.9), food (10.7), salad (10.4) | 9 |
| robe | robe (22.1), shirt (16.8), jacket (13.3), dress (5.7), clothing (3.2) | 19 |
| | *Non-canon. VG names with max agreement in MN* | |
| sedan | car (88.4), wheel (3.1), vehicle (2.3), automobile (1.3), dog (0.8) | 11 |
| pony | horse (83.9), pony (9.1), animal (2.9), donkey (1.1), cow (1.1) | 8 |
| necktie | tie (81.4), necktie (10.2), shirt (4.6), ties (1.5), jacket (0.5) | 11 |
| | *Non-canon. VG names with min agreement in MN* | |
| shelter | umbrella (9.7), shelter (8.8), roof (8.0), tent (7.1), building (6.8) | 10 |
| bath | shower (13.3), elephant (9.9), bird-bath (8.1), water (7.2), trough (7.2) | 10 |
| vegetable | food (15.7), broccoli (13.1), sandwich (10.6), salad (9.3), pizza (7.8) | 25 |

Table 4: Examples for VisualGenome (VG) names and their most frequent corresponding responses in the ManyNames data set (MN; percentages shown in brackets). "Canonical" means that the VisualGenome name is the top name in ManyNames, and non-canonical vice versa.

| | all MN variants | | MN $n > 10$ | |
|---|---|---|---|---|
| relation | %token | %types | %token | %types |
| *easy to recover* | | | | |
| meronymy | 0.3 | 0.9 | 0.8 | 1.0 |
| synonymy | 1.8 | 6.4 | 2.5 | 7.2 |
| hypernymy | 8.8 | 28.2 | 11.0 | 31.3 |
| *difficult to recover* | | | | |
| holonymy.1 | 0.2 | 0.8 | 0.3 | 0.9 |
| co-hyponymy | 4.8 | 6.2 | 6.0 | 6.4 |
| hyponymy | 4.9 | 6.6 | 5.5 | 6.9 |
| *not recoverable* | | | | |
| name not covered | 7.8 | 2.8 | 5.8 | 2.1 |
| rel not covered | 71.3 | 48.1 | 68.0 | 44.2 |

Table 5: Lexical relations between naming variants in WN and the VG name according to WordNet

## 4.2 The "no relation" case

We manually annotated the 100 most frequent name pairs in the "no relation" case. Table **??** shows that, in this category, one third of the pairs do refer to the same object, but the relationship is not captured in WordNet. Most of these cases are arguably coverage issues of WordNet, which doesn't capture the co-hyponymy of *horse-donkey* or the fact that *vehicle* is hypernym of *train*.*//g: I find this really weird... also some other cases I annotated. It sounds like I should have listened more carefully to Carina when she suggested going down and up in the wordnet hierarchy (cf. the example of food-fruit). :/ Maybe we'd capture quite a bit of them if we did a more sophisticated querying of WordNet. To discuss.//* However, a substantial group is constituted by names whose denotations overlap even if they don't belong to the same category. These are typically alternative conceptualizations of objects: as a cat or a toy, as a kind of building or its function (*building-home*), or as a portion or a kind of food (*pizza-slice*).

Still, 69% of the annotated pairs arguably do not denote the same object. Here we find problems HUMANS MAKE SAME "ERRORS" AS MACHINES – REFERENTIAL UNCERTAINTY IN THE ABSENCE OF CONTEXT (discuss as planned with Carina).

Interesting name pairs:

- storefront - store: strictly speaking it's part-whole, but how can one distinguish between the two?
- field - grass: same (reverse); how to distinguish?
- dog - pet (different conceptualizations; classified as "hypernym.2")
- airplane - flight, plane - flight (classified as "other").

Most of the cases are co-hyponyms with categories that are easily confused, such as *horse-donkey*, *truck-jeep*. In some cases, the visual cues are not enough to distinguish between the categories, but the frequency of this phenomenon suggests that co-hyponyms can be used interchangeably.

## 5 Conclusions

## References

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR 2015*, Boston, MA, USA.