



# A Large-Scale Image Dataset Collected via Google Image Search

Yu Su & Frédéric Jurie

GREYC, University of Caen, France

# Outline

- Background
- Collection
- Statistics
- Annotation

# Background



- The Quaero Still Images Dataset

- Corpus project, WP 8.1

- Objectives

- Provide training and evaluation data for CTC.WP8, e.g. scene annotation and object recognition.
    - Should be useful in the application where image content has to be searched.
    - Allow to evaluate automatic annotations tools.

# Background



- Some existing image datasets

- CalTech-256

- 256 object categories containing 30,607 images.

- PASCAL VOC

- 20 object categories containing 14,743 images.

- ImageNet

- 14,841 concepts organized by WordNet hierarchy.
    - About 10 millions images.

- ...

# Background



- Differences with previous datasets
  - Addressed task: improve the ranking of existing text based image search engines
    - In large scale problem using text to retrieve an initial set of images is reasonable
    - Images can be re-ranked according to their visual content
  - Previous dataset does not include any textual description (contextual information)

# Background



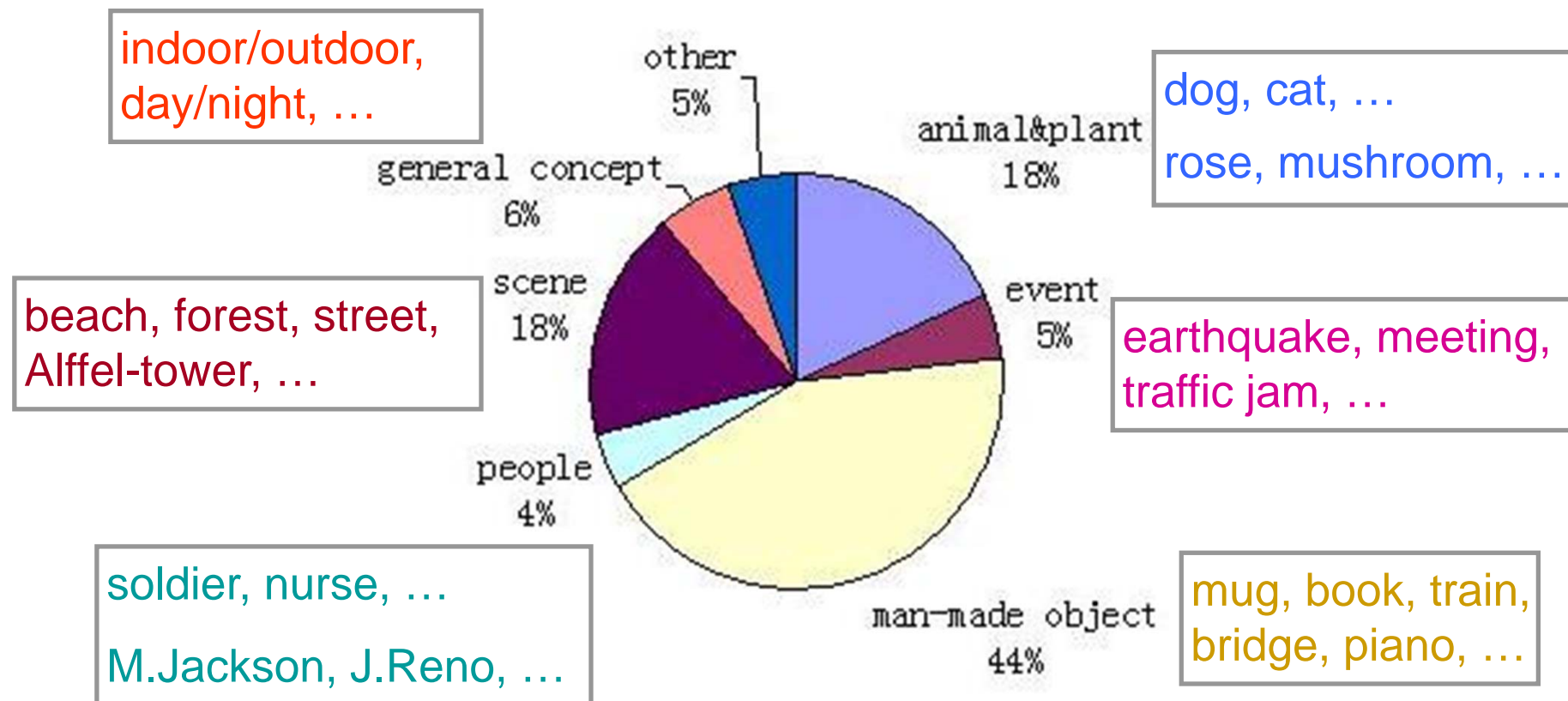
- Features of Quaero Still Images Dataset
  - Large scale
    - 518 concepts, ~1,000 images per concepts
  - Diverse concepts
    - Objects, people, scenes, events, ...
  - Collection source
    - Google Image Search
  - Contextual text and images
    - URLs, HTML tags and surrounding words
    - Other images on web pages

# Outline

- Background
- **Collection**
- Statistics
- Annotation

# Dataset Collection

- 518 pre-defined concepts\*



\*cooperating with R. Landais and G. Quénot, some concepts are the same as those in TRECVID



# Dataset Collection

- Exclude some “bad” concepts

**rice**



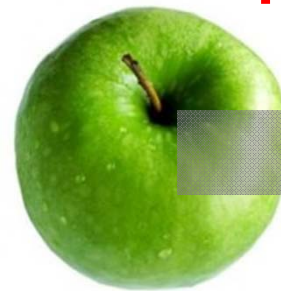
**cooked rice**



**rice plant**



**apple**



**fruit**

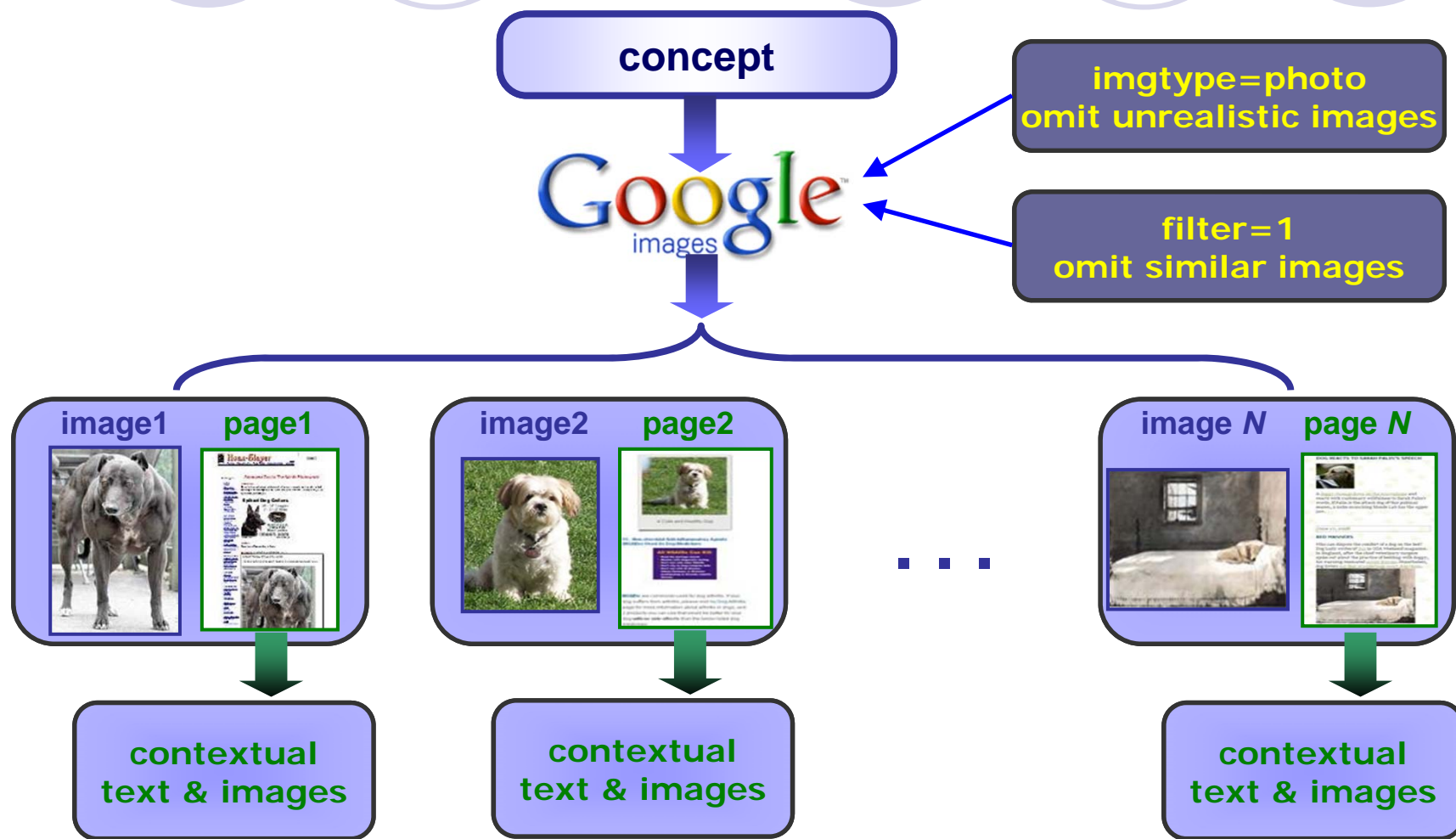


**trademark**



**non-consistent visual contents**

# Collection Flowchart



# Extract Contextual Info: An example


image *i*




concept: dog

page *i*

doors and never leave our houses again. Life is risky. Terribly risky.

 Ads by Google

*Be gone cat!*

 dog\_olive1

*Be gone dog!*

Counselling in France  
English speaking Counsellor Glenys Forrester  
(M.Sc. Psych)  
www.mgfcounselling.com

from → just for fun

ENDORSEMENT OF  
EVERYTHING ON THIS  
SITE

I often publish authors  
whose articles I  
appreciate but I may  
not always agree with  
all that is said. My  
pieces, also, cannot  
stand alone. Pieces of  
my work contain ideas  
that may sometimes  
seem to contradict  
other pieces because  
all cannot be said in a  
short essay on any  
given subject. It's wise  
to not make  
assumptions based on  
a few pieces of writing.  
The same holds true  
for me blogroll. I by no  
means agree with

<http://bipolarblast.wordpress.com/2009/03/27/why-is-the-fact-that-we-all-trip-over-our-dogs-and-cats-such-big-news/>

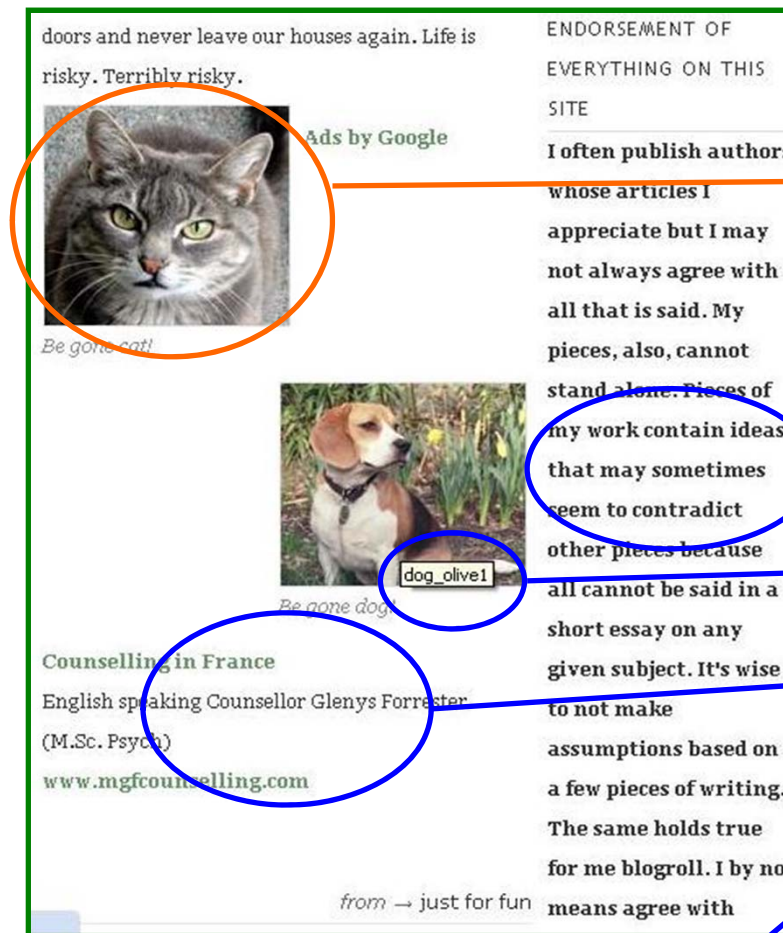
# Extract Contextual Info: An example

image  $i$



concept: dog

page  $i$



contextual  
image

surrounding  
words

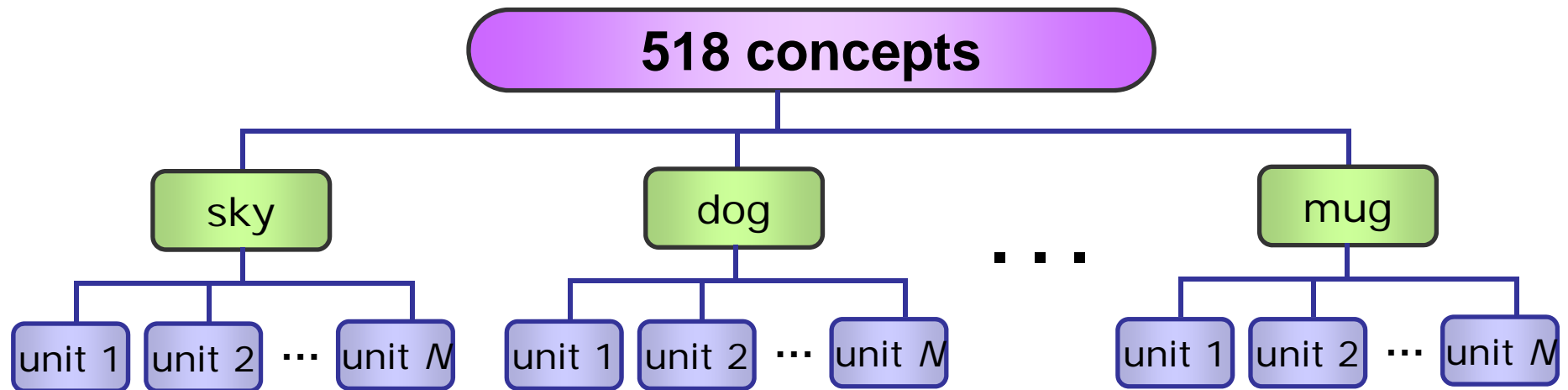
HTML tag

surrounding  
words

URL

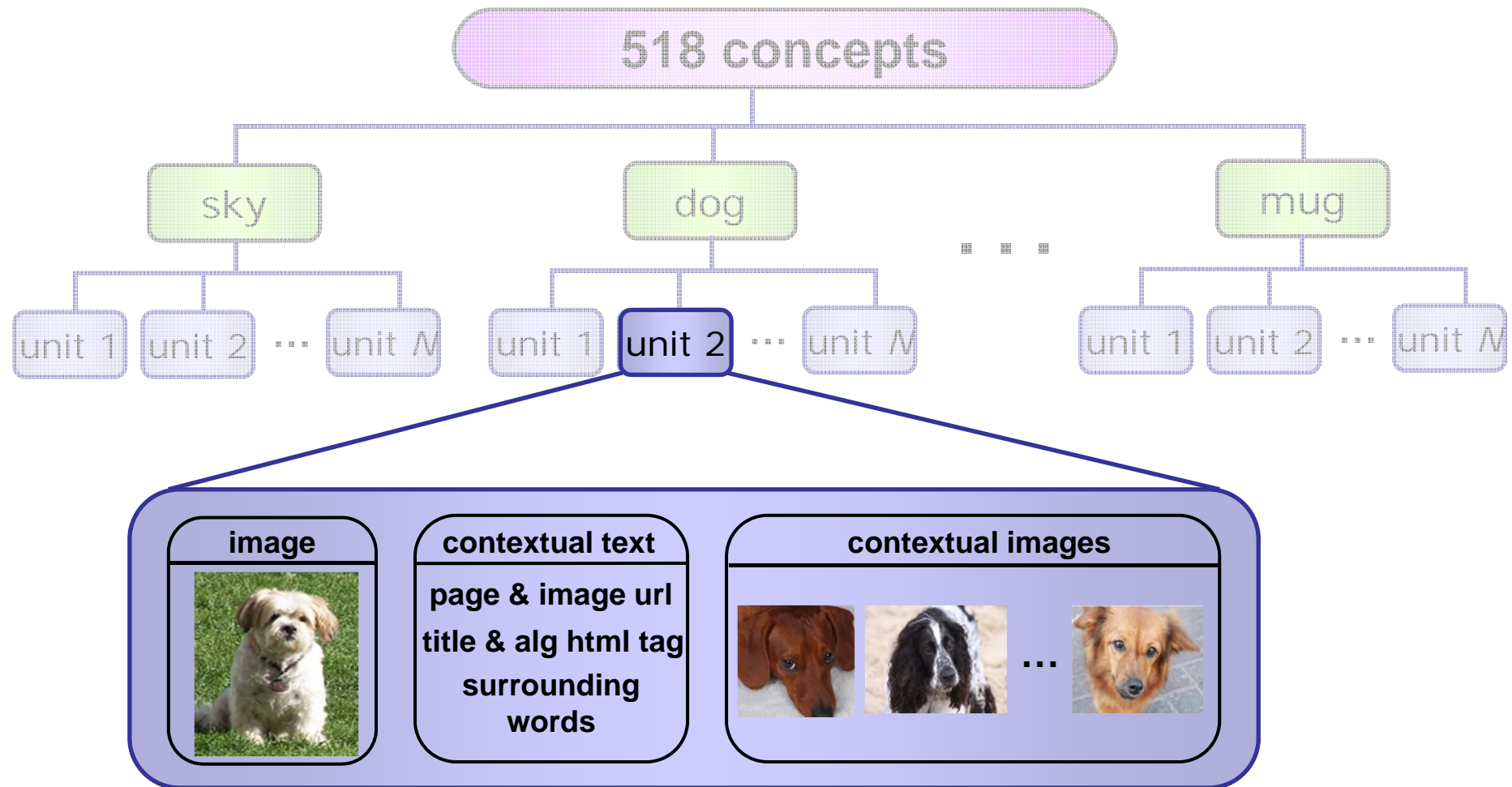
<http://bipolarblast.wordpress.com/2009/03/27/why-is-the-fact-that-we-all-trip-over-our-dogs-and-cats-such-big-news/>

# Dataset Structure





# Dataset Structure



# Outline

- Background
- Collection
- **Statistics**
- Annotation

# Dataset Statistics



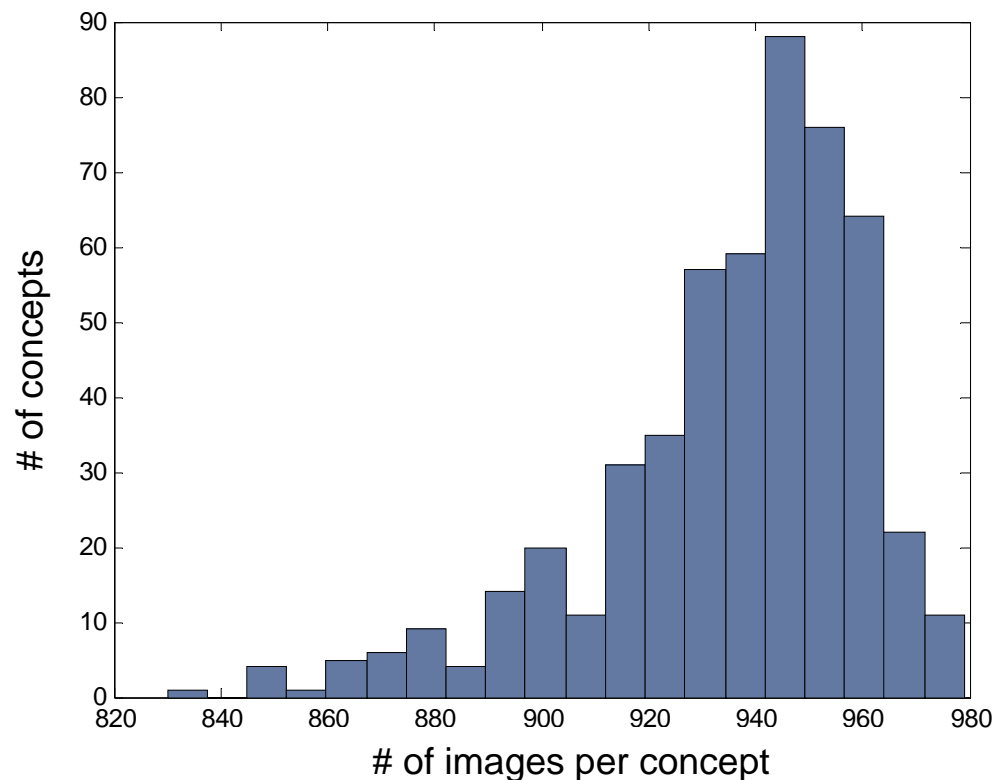
- Overview

- 518 concepts containing 484,747 images
- 482,007 pages and 1,543,766 contextual images
- Data volume: ~170G bytes
- Time cost: ~1500 hours



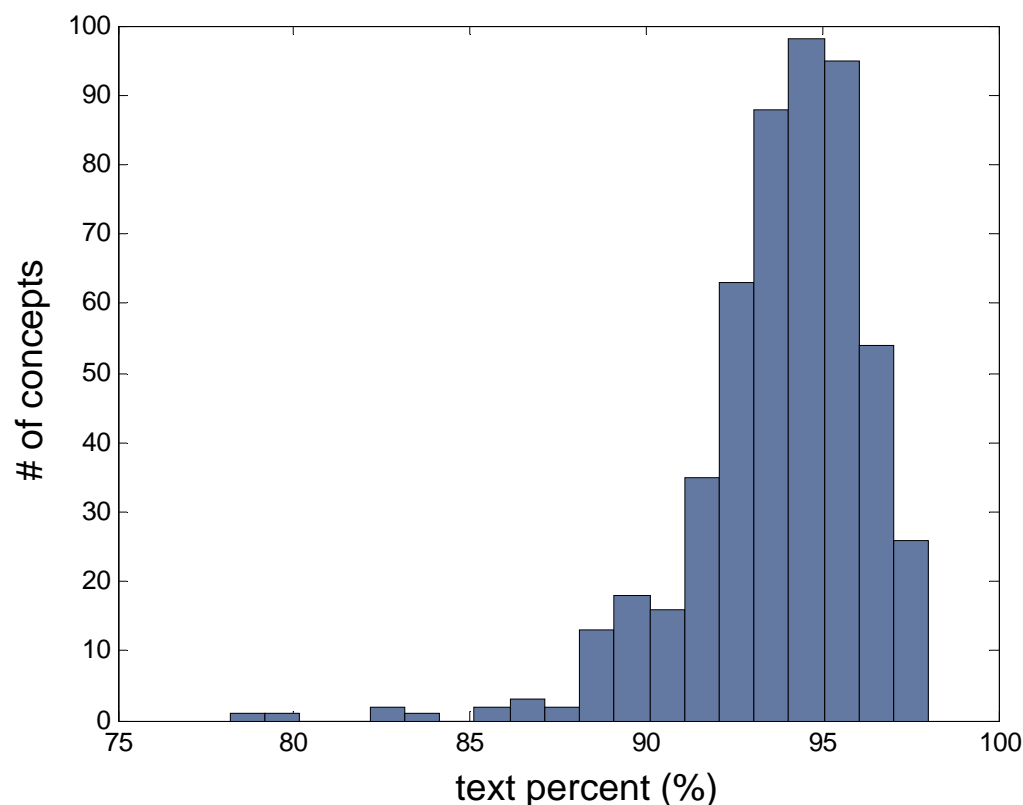
# Dataset Statistics

- 935 returned images per concept
  - For each concept, returned images are less than 1,000 due to bad URLs, unrealistic images and duplications.



# Dataset Statistics

- 94% returned images have contextual text
  - Not all the returned images have contextual text due to the bad URLs of some web pages



$$\text{text percent} = \frac{\# \text{ of images with text}}{\# \text{ of images}}$$

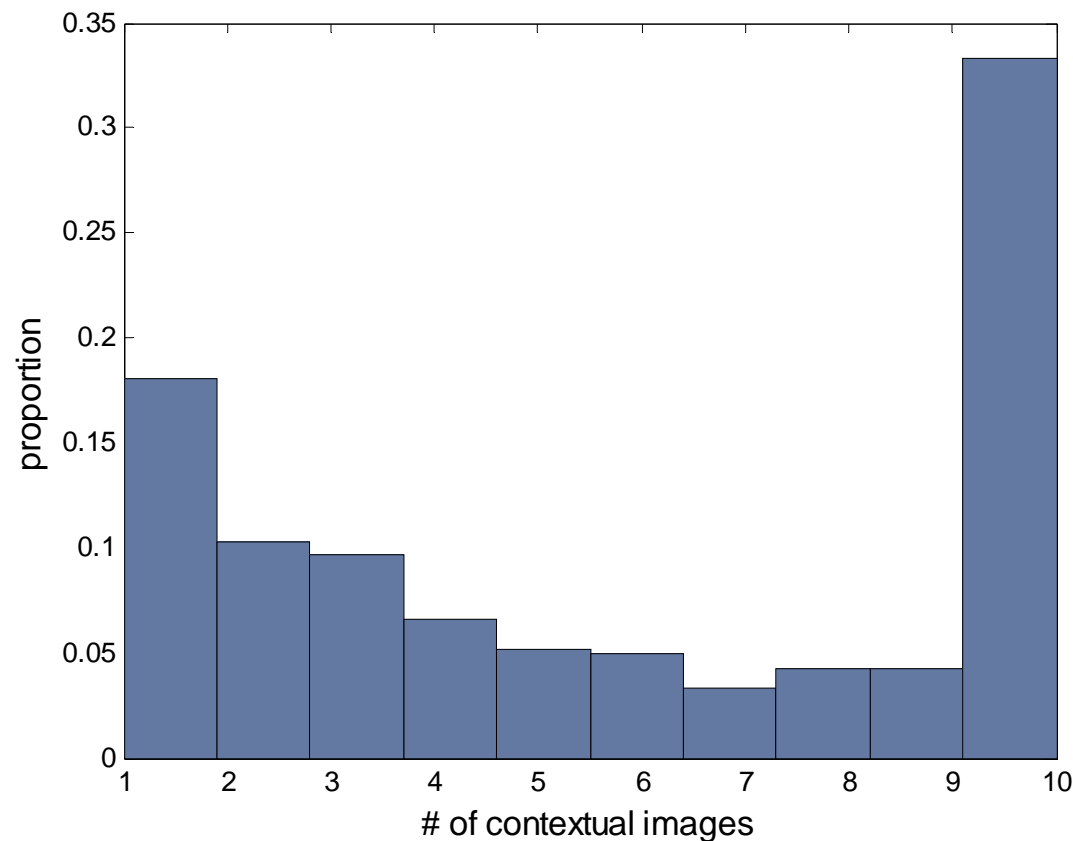
Example:

dog: 928 images, 892 with text

text percent is  $892/928=96\%$

# Dataset Statistics

- 59% returned images have contextual images
- Averagely, each of them have 5 contextual images.



# Outline

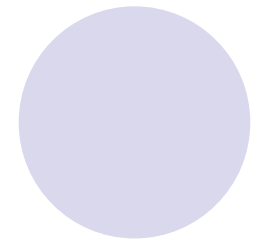
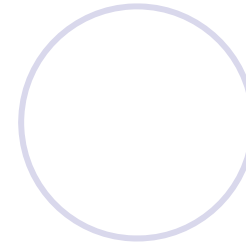
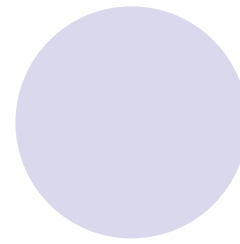
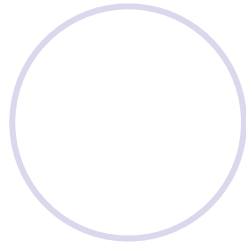
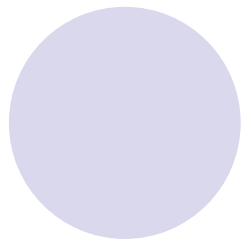
- Background
- Collection
- Statistics
- Annotation

# Annotation (not finished yet)

query : dog



~2 hours to annotate each concept



Thanks for your attention!