# Big datasets - promise or Big data, shmig data

D.A. Forsyth, UIUC

# Conclusion

- Q: What do big datasets tell us?
- A: Not much, if the emphasis is on size

- Collecting datasets is highly creative
  - rather than a nuisance activity
  - tools are getting better by the day

- Bias, weird frequencies are a major issue
  - There are no best practices for avoiding problems
  - May shape our representations

- Recognition problems are hard to frame
  - excess certainty may be dangerous

# Bias

Should not be perjorative

- Frequencies in the data may misrepresent the application
  - Because the labels are often wrong

    Label error
  - Because of what gets labelled

    Label bias
    - P(labelled|X) is not uniform
    - eg obscure but important objects in complex clutter
    - eg pedestrians in crowds
  - Because of what gets collected

    Curation bias
    - eg. pictures from the web are selected - not like a camera on head
    - eg. “Profession” labelling for faces in news pictures

X=data

# Induction is why bias matters

- Fundamental principle of machine learning
  - if the world is like the dataset, then future performance will be like training
    - Chernoff bounds, VC dimension, etc., etc.
- But what if the world can't be like the dataset?

# Pedestrian Detection

- Pedestrian detection:
  - We may not run down people who behave strangely
    - want “will fail to detect with frequency ...”
    - can do “...” IF test set is like training set
  - There is a large weight of easy cases which may conceal hard cases
- Resolution (frankly implausible)
  - ensure that training set is like test set
- Resolution (perhaps)
  - try only to learn things that are “fairly represented” in datasets
  - i.e. build models

Google

lion Search SafeSearch off ▼

About 23,100,000 results (0.05 seconds) Advanced search

Everything
Images
Videos
More

**Any size**
Medium
Large
Icon
Larger than...
Exactly...

**Any type**
Face
Photo
Clip art
Line drawing

**Any color**
Full color
Black and white

Related searches: lion roaring lioness lion drawing lion tattoo


**Lions Kill Giraffe**
479 × 450 - 48k - jpg
abolitionist.com
Find similar images


**Lion on Horseback**
468 × 393 - 39k - jpg
raincoaster.com
Find similar images


**3, Lion**
434 × 341 - 41k - jpg
bluepyramid.org
Find similar images


**Interestingly, the**
470 × 324 - 30k - jpg
bostonherald.com
Find similar images


**Description : Aslan**
792 × 768 - 99k - jpg
photocase.org
Find similar images


**I was doing research on**
400 × 300 - 27k - jpg
lowkayhwa.com
Find similar images


**Lion Tiger Size**
500 × 553 - 65k - jpg
indrajit.wordpress.com
Find similar images


**Lion Park, South**
450 × 300 - 30k - jpg
africa-nature-photog...
Find similar images


**Lion Limited**
500 × 500 - 76k - jpg
onlineartdemos.co.uk
Find similar images


**Lion**
395 × 480 - 47k - jpg
ibexinc.wordpress.com
Find similar images


**lions**
1200 × 800 - 243k - jpg
lifeasastudentnurse...
Find similar images


**African Lion**
500 × 333 - 57k - jpg
itsnature.org
Find similar images


**LIONS:**
604 × 800 - 225k - jpg
edge.org
Find similar images


**Lion. Panthera leo**
459 × 480 - 35k - jpg
shoarns.com
Find similar images


**lions, cuddle**
620 × 400 - 70k - jpg
telegraph.co.uk
Find similar images


**lion**
350 × 504 - 28k - jpg
sodahead.com
Find similar images


**LION!**
500 × 385 - 74k - jpg
firemice.wordpress.com
Find similar images


**Starring horse-riding**
800 × 626 - 53k - jpg
dailymail.co.uk
Find similar images


**Picture: 17 stone**
468 × 602 - 93k - jpg
dailymail.co.uk
Find similar images


**human-lion**
470 × 324 - 31k - jpg
seesdifferent...
Find similar images


**Lion at Sunset**
400 × 318 - 25k - jpg
art.com
Find similar images

# Representation is a response to bias

- Attributes
- Semantic parts
- Tying
- Example
  - Ramanan's activity example
    - where you are often reveals what you are doing
    - but how do we encode where you are
      - x-y coords?
      - near the stove?






# Conclusion

- Q: What do big datasets tell us?
- A: Not much, if the emphasis is on size

- Collecting datasets is highly creative
  - rather than a nuisance activity
  - tools are getting better by the day

- Bias, weird frequencies are a major issue
  - There are no best practices for avoiding problems
  - May shape our representations

- Recognition problems are hard to frame
  - excess certainty may be dangerous

Platonism?

# One belief space about recognition

- Categories are fixed and known
  - Each instance belongs to one category of k

Obvious nonsense
Obvious nonsense

- Object recognition=k-way classification

- current data sets ok in principle
  - improve coverage
  - collect unbiased datasets with fair coverage

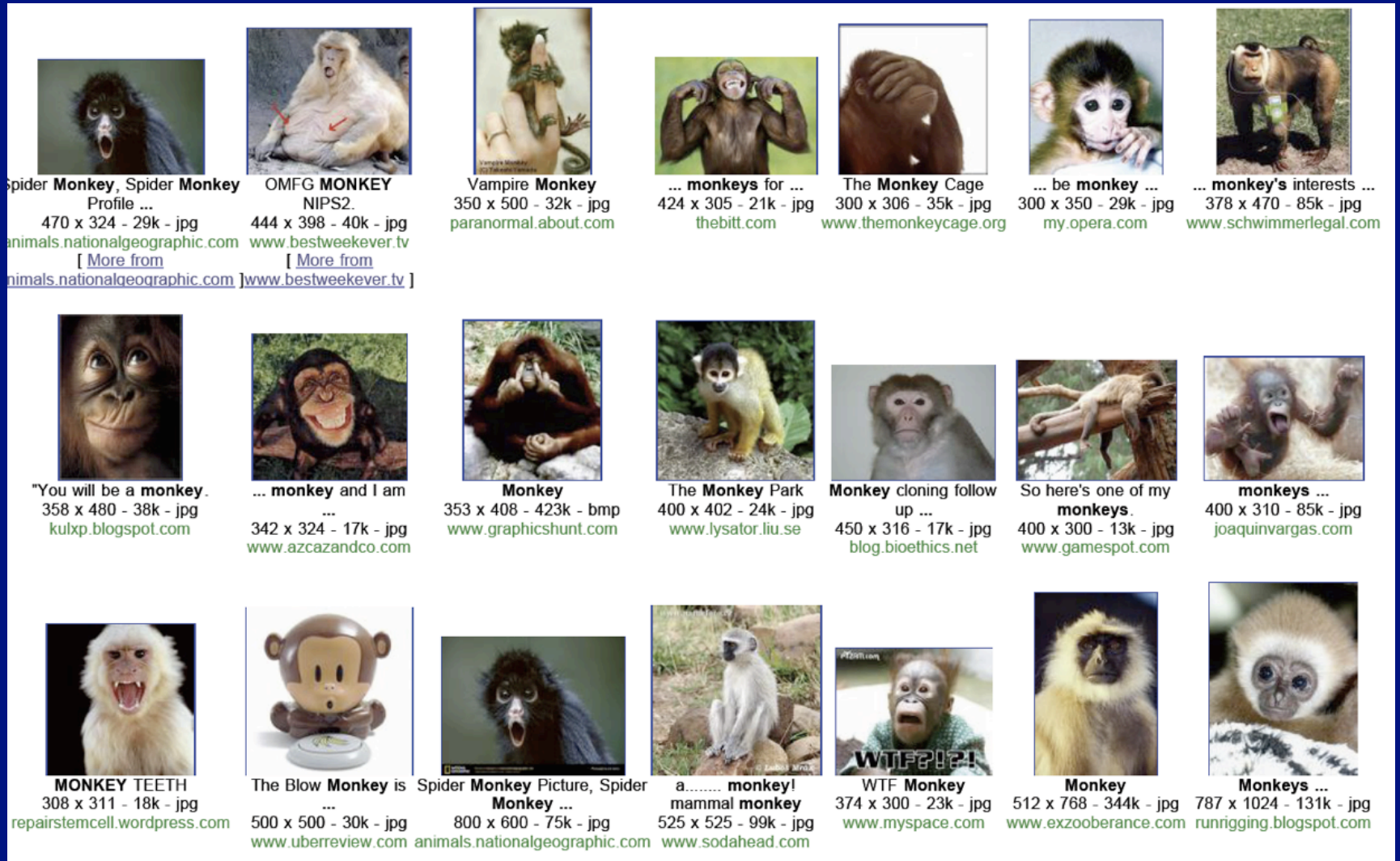
I doubt this is possible
I doubt this is possible

- research agenda:
  - more features, better classifiers:
  - perhaps category hierarchies for statistical leverage (tying)

# What have we inherited from this view?

- Deep pool of information about feature constructions
- Tremendous skill and experience in building classifiers
- Much practice at empiricism
  - which is valuable, and hard to do right

# Are these monkeys?

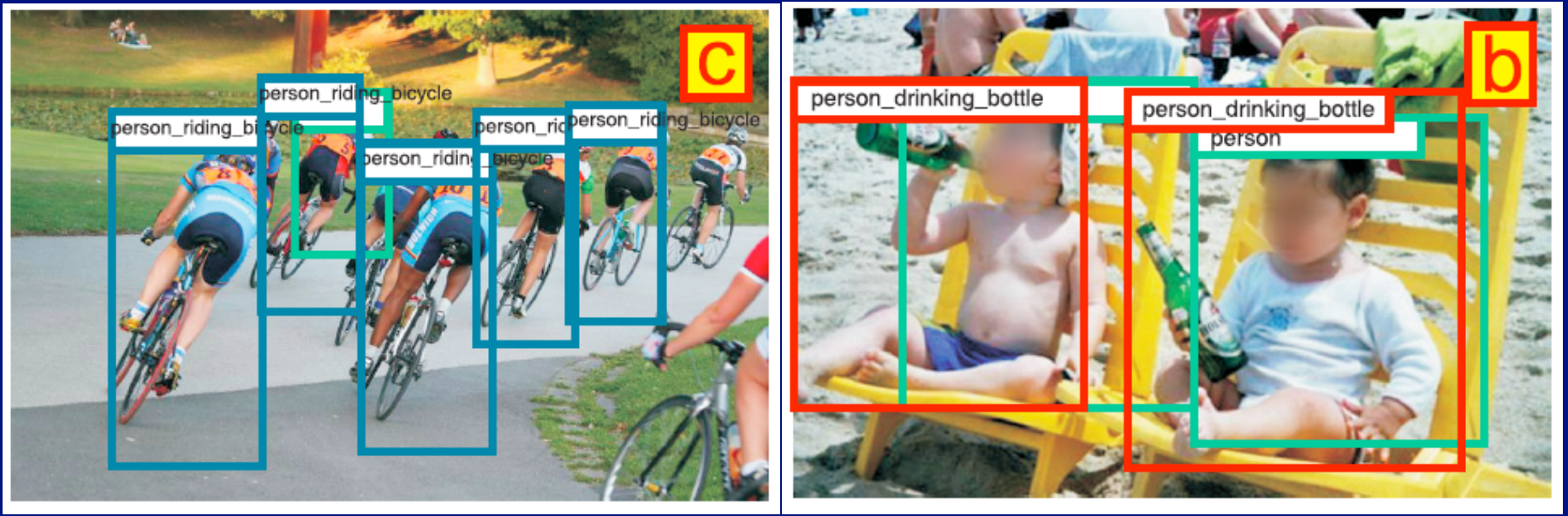


# Another belief space about recognition

- Categories are highly fluid
  - opportunistic devices to aid generalization
    - affected by current problem
  - instances can belong to many categories
    - simultaneously
  - at different times, the same instance may belong to different categories
  - categories are shaded
    - much “within class variation” is principled
  - Most categories are rare
  - Many might be personal, many are negotiated
- Understanding (recognition)
  - constant coping with the (somewhat) unfamiliar
  - bias is pervasive, affects representation

# Visual complexity

- Some “categories” hard to detect,  others easy?



# Co-existing category systems

Monkey or Plastic toy or  both or irrelevant

Some of this depends on what you're trying to do, in ways we don't understand

Person or child or beer drinker or beer-drinking child or tourist or holidaymaker or obstacle or potential arrest or irrelevant or...

# Research agenda

- How do we build bias-robust representations?
- What should we mean by “category”?
  - how are categories created?
  - how can multiple category systems co-exist?
  - how can we sew together categorization and utility?
- What should we report about pictures?
  - What kind of clumps of meaning should we detect?
  - What should we say about things?