Size isn't important
or
What do big visual datasets tell us ?

D.A. Forsyth, UIUC
(and I'll omit the guilty)

Conclusion

- 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

What could big datasets tell us?
(by virtue of being big)

- Good magnitude estimates of small effects

So what

- A more accurate estimate of what the world is like
  - frequencies, etc

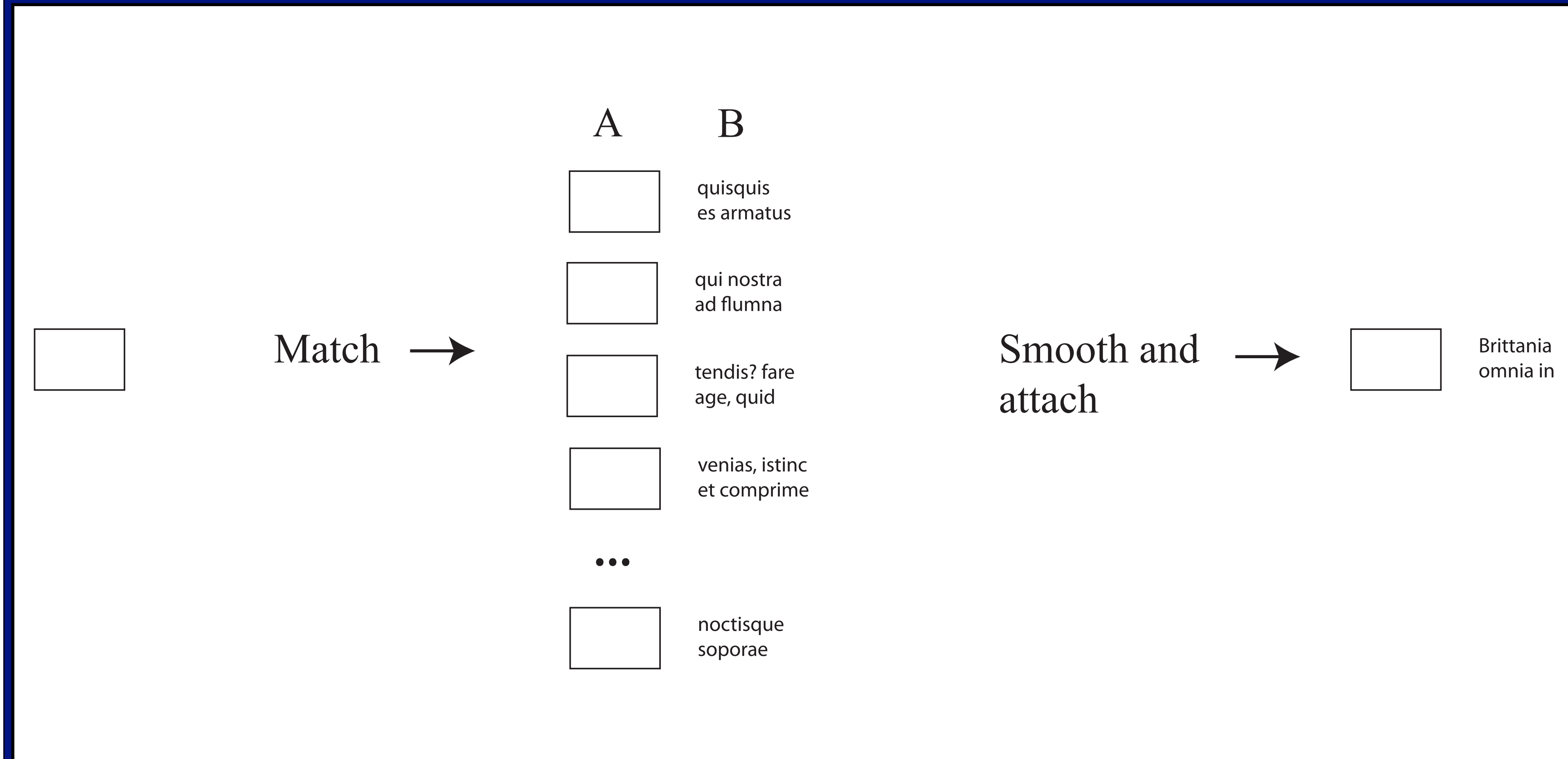
Seems unlikely, might go the other way

- Collective search is more significant than it gets credit for
  - Problem:
    - publish a dataset
    - people try methods, keep ones that do well
    - hence, results suffer from intense selection bias
  - Bigger datasets -> weaker recognition statistics
    - Because the categories are genuinely harder?
    - Because collective search is much harder?

Conclusion

- 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

Non-parametric regression

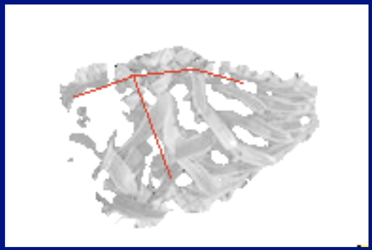


With a broad view of “match”, “smooth”, all classifiers fit into this story

A=picture, B=category

- Far too many to select one!
- Fergus et al 05; Fergus et al 04;Fei-Fei 06; Berg 05; Everingham et al Pascal Challenge reports 06, 07, 08;
  - etc etc etc etc etc

Table 1. Overall classification performance of the system, in various configurations, to 4289 control images and 565 test images. Configuration F is the primary configuration of the grouper, fixed before the experiment was run, which reports a nude present if either a girdle, a limb-segment girdle or a spine group is present, but not if a limb group is present. Other configurations represent various permutations of these reporting conditions; for example, configuration A reports a person present only if girdles are present. There are fewer than 15 cases, because some cases give exactly the same response.



Forsyth etal 96, 01

Conclusion

- 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

Bias isn’t always bad

- If all the faces on the web are politicians
  - one needs only to be good at politicians to be good at the web
- If people really only want to search videos for “kissing”
  - then you don’t need a general activity recognition strategy

Bias is pervasive

Torralba+Efros 11

Size doesn't make bias go away

- And could make it worse...
  - eg your dataset collector really likes red cars
- cf next slide

Web **Images** Videos Maps News Shopping Gmail more ▼ Search settings | Sign in

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

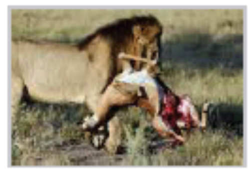

I was doing research on
400 × 300 - 27k - jpg
lowkayhwa.com
Find similar images


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

Label error

- Fact of life
  - people label things wrong
- Can fix when there are many instances
  - consistency (Zhao et al 08)
  - smoothing (Berg, 06; Li, 06; Wang 08; Collins 08)
- Might be able to fix with hierarchy+generalization
  - we should never mix up “cat”’s and “truck”’s

Label bias: the choice of what is labelled

- P(labelled|X) is not uniform
    - or P(X|labelled) is not the same as P(X|not labelled)
- There are models
    - problem sometimes called dataset shift, see (Quinonero-Candela 09)
    - can be addressed with, say, large unlabelled datasets
        - build smoothed estimate of p(labelled|X), reweight
- Important effect
    - can make high capacity classifiers generalize better than low capacity
    - (maybe) be very cautious about linear SVM's

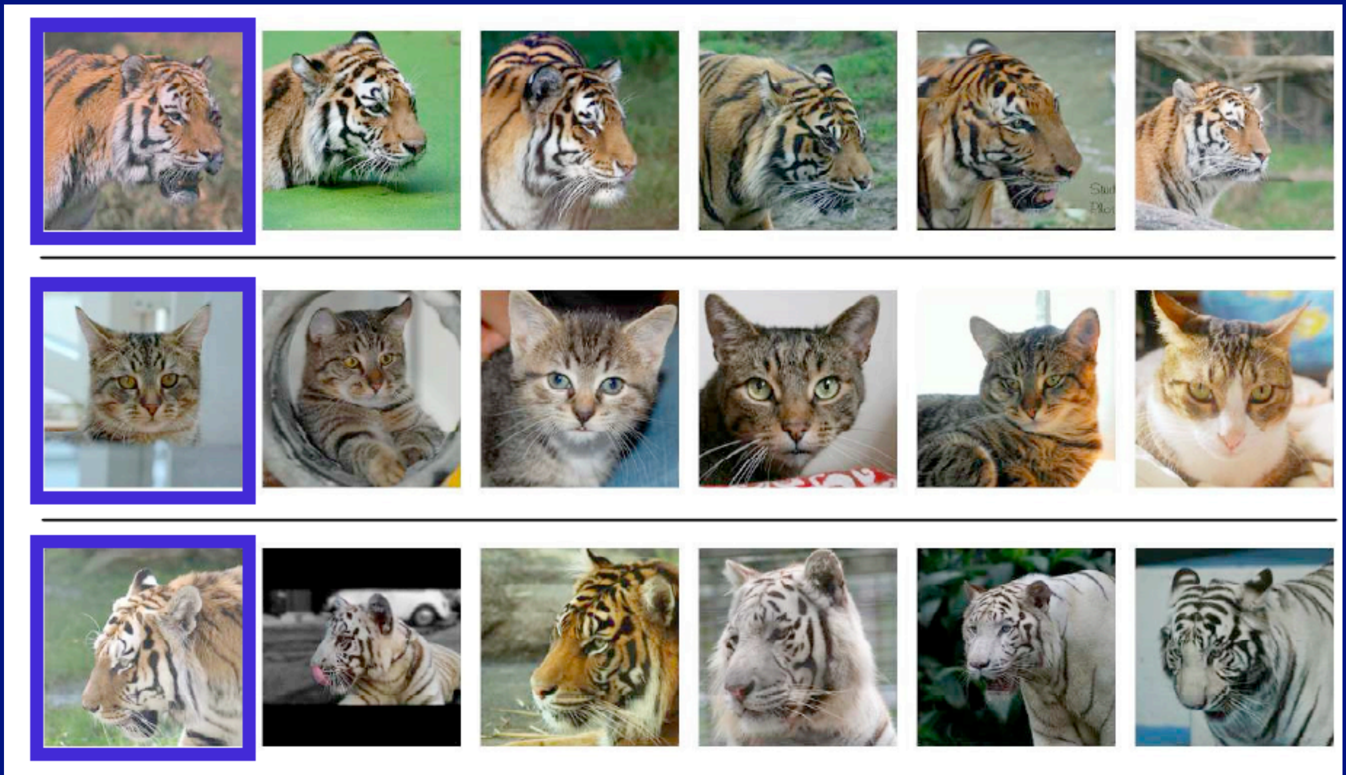
Curation bias

Loeff et al, 06

- Collected data is not a fair sample of X
  - labelled AND unlabelled data
- Images on the web are “curated”
- Iconography seems to be a big effect here
  - visual “modes” of representation
    - see Berg+Berg 09
  - we might not see them all
  - cf Google image search with Flickr

X=data
Y=labels
X_i = unlabelled examples
(X_j, Y_j)=labelled examples

Iconographic phenomena



Berg+Berg 09; see Jing+Baluja 08

Web Images Videos Maps News Shopping Gmail more ▼ Search settings | Sign in

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

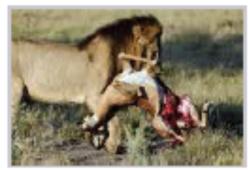

I was doing research on
400 × 300 - 27k - jpg
lowkayhwa.com
Find similar images


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

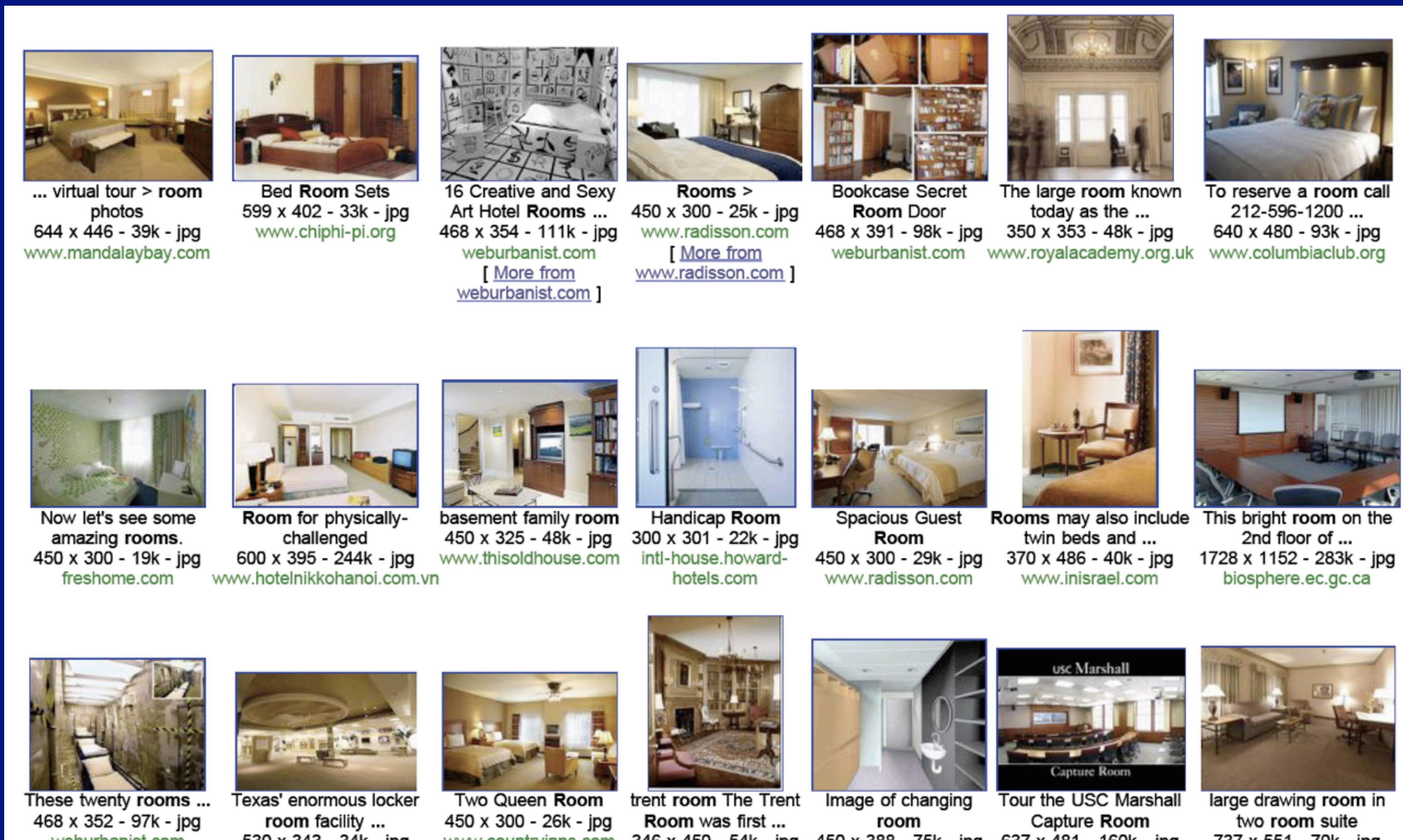

Starring horse-riding
800 × 626 - 53k - jpg
dailymail.co.uk
Find similar images


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

Google “rooms”



Flickr “rooms”

Conclusion

- 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

Induction

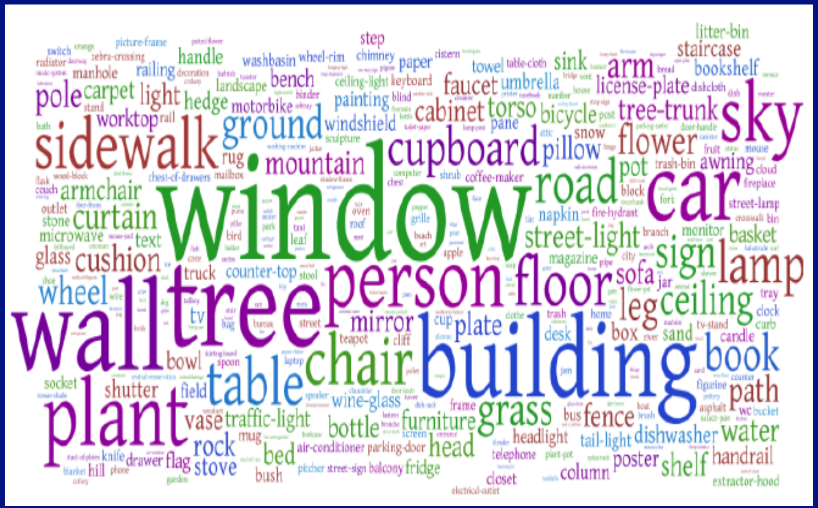
- Fundamental principle of machine learning
  - if the world is like the dataset, then future performance will be like training
    - Chernoff bounds, VC dimension, etc., etc.
- But what if the world can’t be like the dataset?

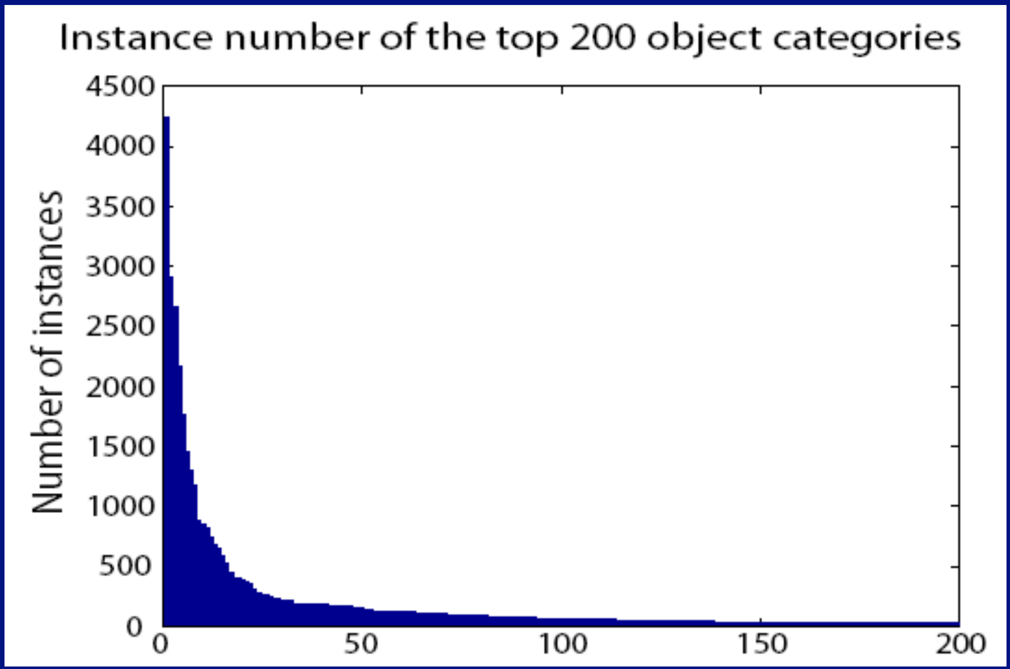
Pedestrian Detection

- 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

Object recognition

- The world can't be like the dataset because
  - many things are rare
  - this exaggerates bias





Wang et al, 10

Distributional semantics

- Most words are unusual
- Don’t know a word?
  - nearby words can tell you what it means
  - or how similar it is to a word you do know

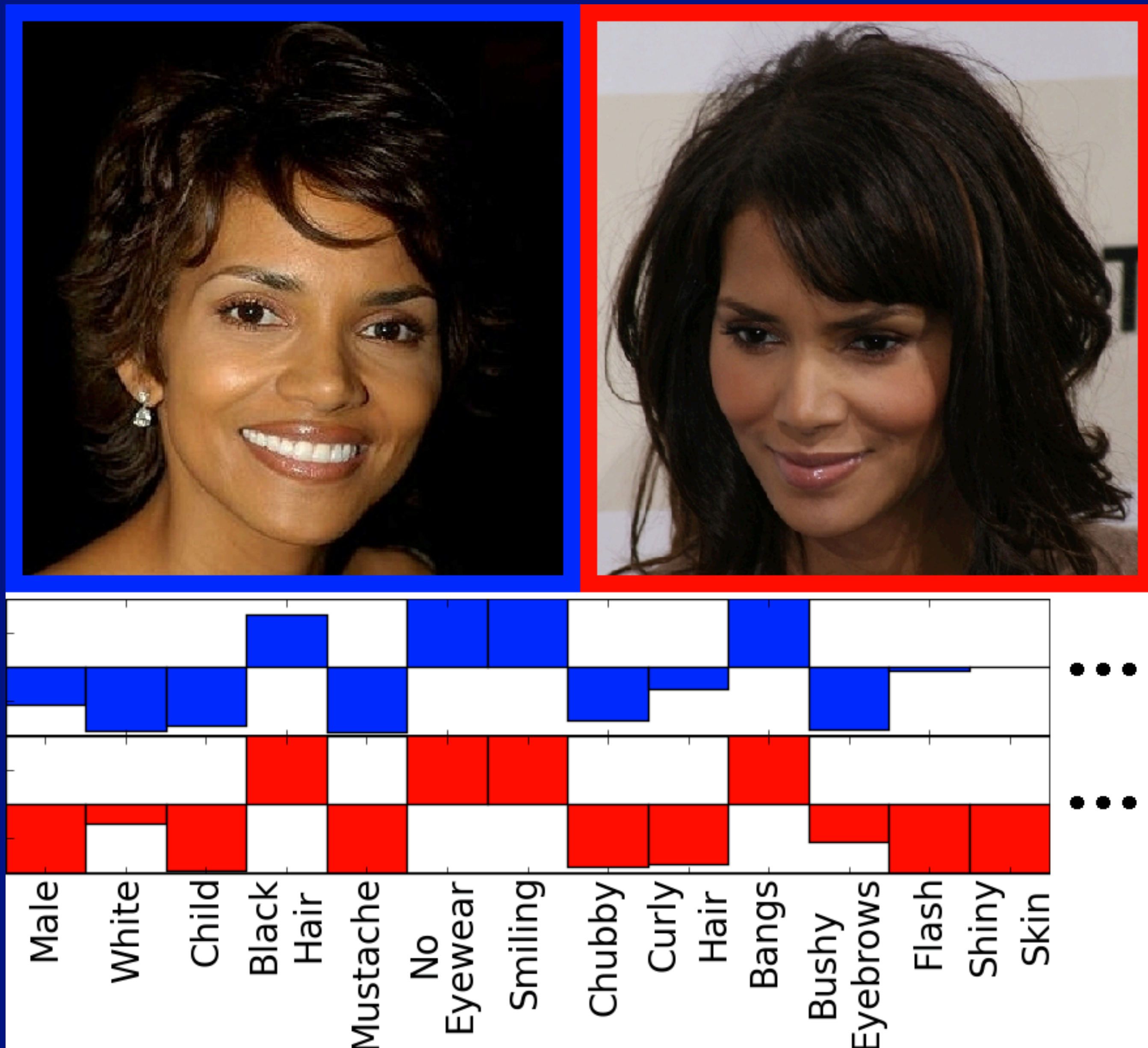
“No; this my hand will rather the multitudinous seas incarnadine, making the green one red.”

“In one routine, describing his “ludicrously alpha” surfing instructor for the Forgetting Sarah Marshall shoot, he exclaims, “The sea were incarnadine wiv his testosterone!””

Bias affects representation

- Attribute style representations
  - because each attribute may have large unbiased training set
  - even when each category does not

Farhadi et al 09; Lampert 09



“Attribute and Simile Classifiers for Face Verification,” ICCV 2009. (N. Kumar, A. Berg, P. Belhumeur, S. K. Nayar)

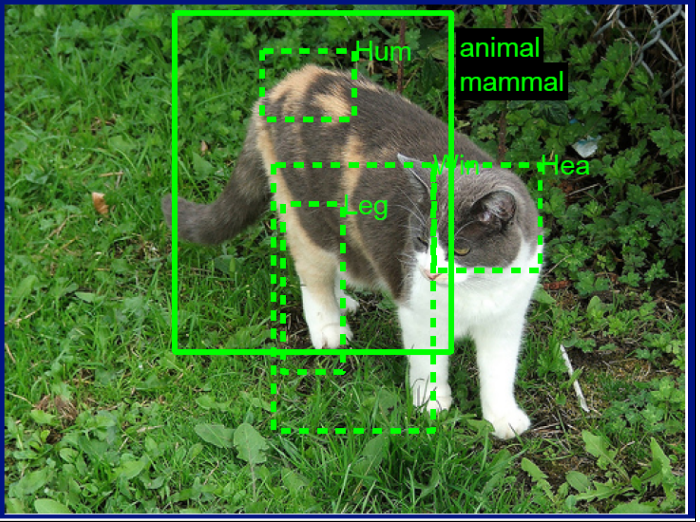
Bias affects representation

- Semantic parts
  - as opposed to variance suppressing
  - because many animals have legs, vehicles have wheels, etc.
    - again, may have large unbiased training set

Green box **Animal**

Red Box **Vehicle**

Farhadi et al 10
Endres et al 10



Bias affects representation

- Other kinds of semantics
  - 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

- 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

Are these monkeys?

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

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

Research agenda

- 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?
- What information is important?
  - Texture, yes; but: support? shape? geometry? context?
  - Goals and intentions?

Co-existing category systems

Monkey or Plastic toy or both or irrelevant

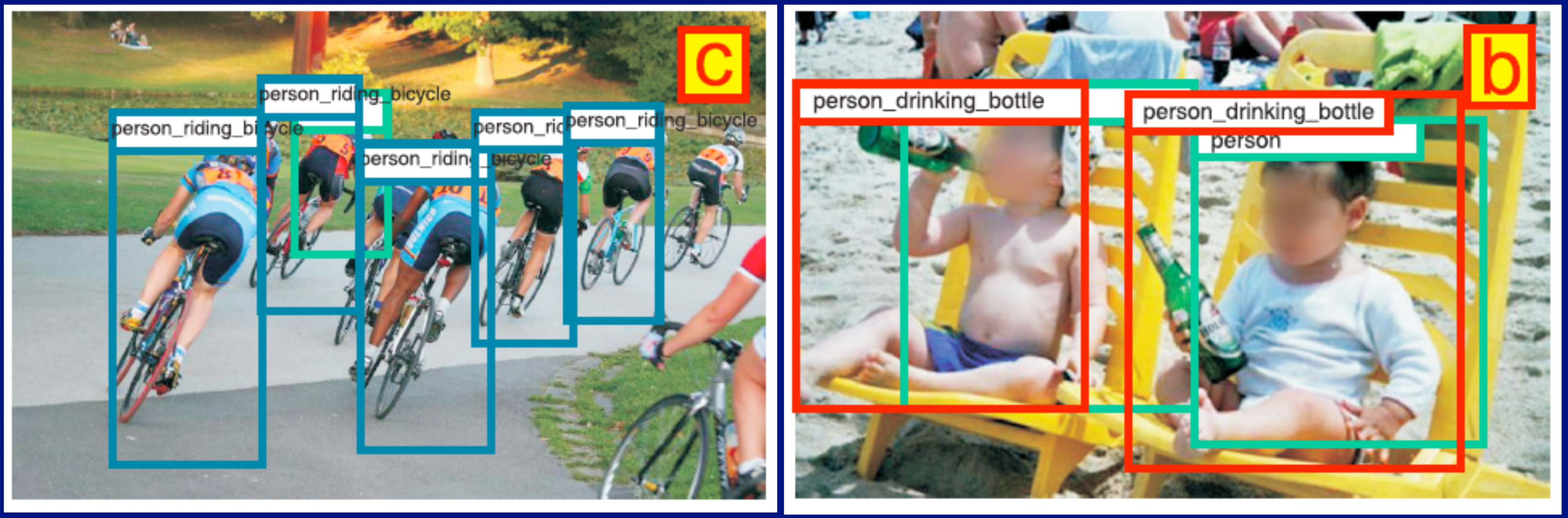
Some of this depends on what you're trying to do, in ways we don't understand

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

Clumps of meaning

"Sledder"
Is this one thing?
Should we cut her off her sled?

Clumps of meaning



Farhadi + Sadeghi 11

What should we report?

Two girls take a break to sit and talk .

Two women are sitting , and one of them is holding something .

Two women chatting while sitting outside

Two women sitting on a bench talking .

Two women wearing jeans , one with a blue scarf around her head , sit and talk .

Sentences from Julia Hockenmaier's work

Rashtchian ea 10

Reporting Sentences

Farhadi ea 10

A man stands next to a train on a cloudy day
A backpacker stands beside a green train
This is a picture of a man standing next to a green train
There are two men standing on a rocky beach, smiling at the camera.
This is a person laying down in the grass next to their bike in front of a strange white building.

Selection

- (No-one was hurt; I checked)


10/2009
PM

How many adults were on the platform and what were they doing?

What’s going to happen to the baby?

What outcome do we expect?

How are other people feeling?

What will they do?


/10/2009
PM

What should we do about datasets?

- Recognize and beware of fallacies
  - Good datasets are big implies big datasets are good
  - If you know your problem well, you can collect an unbiased dataset
- Always train on dataset A and test on B
  - this isn't the same as a train/test split of A
- Throw away more data than we're doing
  - it tends to go off, and when it has gone off, it's poisonous
- Come up with new methods to identify and manage bias
  - How?
- Come up with richer notions of categorical annotation

Conclusion

- Not much, if the emphasis is on size
  - strong classification methodologies are no substitute for thought
- 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

Obtain dataset

Build features

Mess around with classifiers, probability, etc

Produce representation

Computer vision

Obtain dataset

Build features

Light entertainment
(the way we do it)

Mess around with classifiers, probability, etc

Computer vision

Produce representation

Big questions

Computer vision

- What signal representation should we use ?

PLUMBING

MODELS

Computer vision

- What should we say about visual data?

Taxonomy

The Unfamiliar

- What do you say about it?
  - Attributes?
- Are many categories rare?
  -

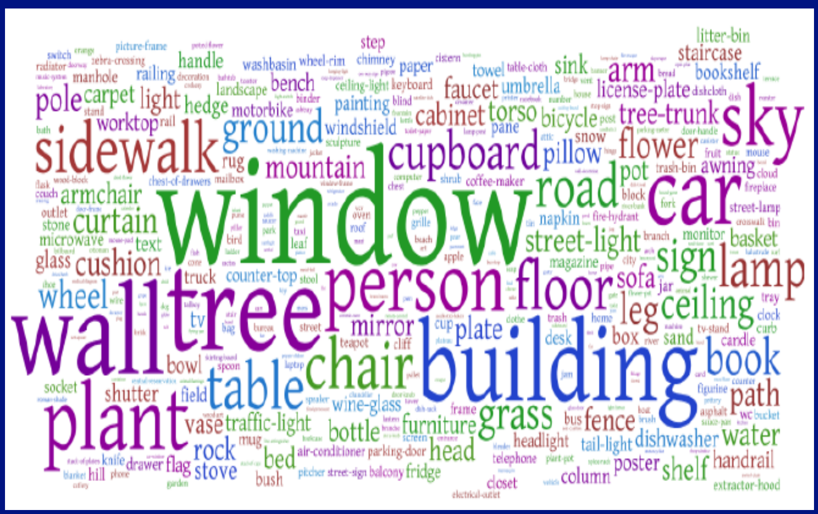
Distributional semantics

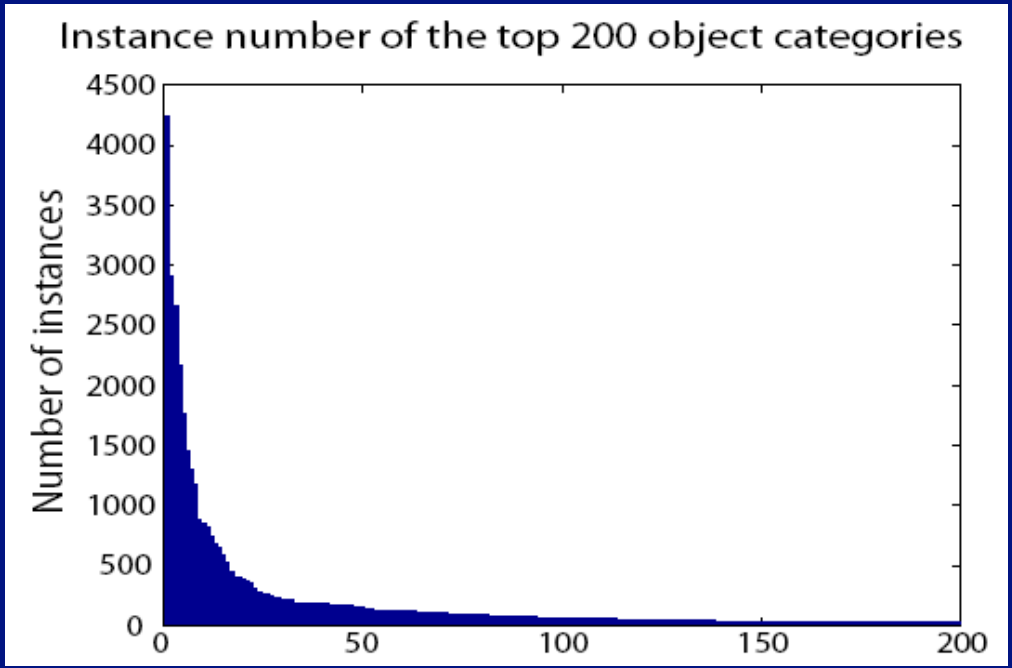
- Most words are unusual
- Don’t know a word?
  - nearby words can tell you what it means
  - or how similar it is to a word you do know

“No; this my hand will rather the multitudinous seas incarnadine, making the green one red.”

“In one routine, describing his “ludicrously alpha” surfing instructor for the Forgetting Sarah Marshall shoot, he exclaims, “The sea were incarnadine wiv his testosterone!””

Are most things unfamiliar?





Wang ea 10; labelme data

Collective search

- Problem:
  - publish a dataset
  - people try methods, keep ones that do well
  - hence, results suffer from intense selection bias
- Bigger datasets -> weaker recognition statistics
  - Because the categories are genuinely harder?
  - Because collective search is much harder?

Fallacy

Good datasets are big

implies

Big datasets are good

Fallacy

If you know your problem well

you can collect an unbiased dataset

Conclusion

- 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

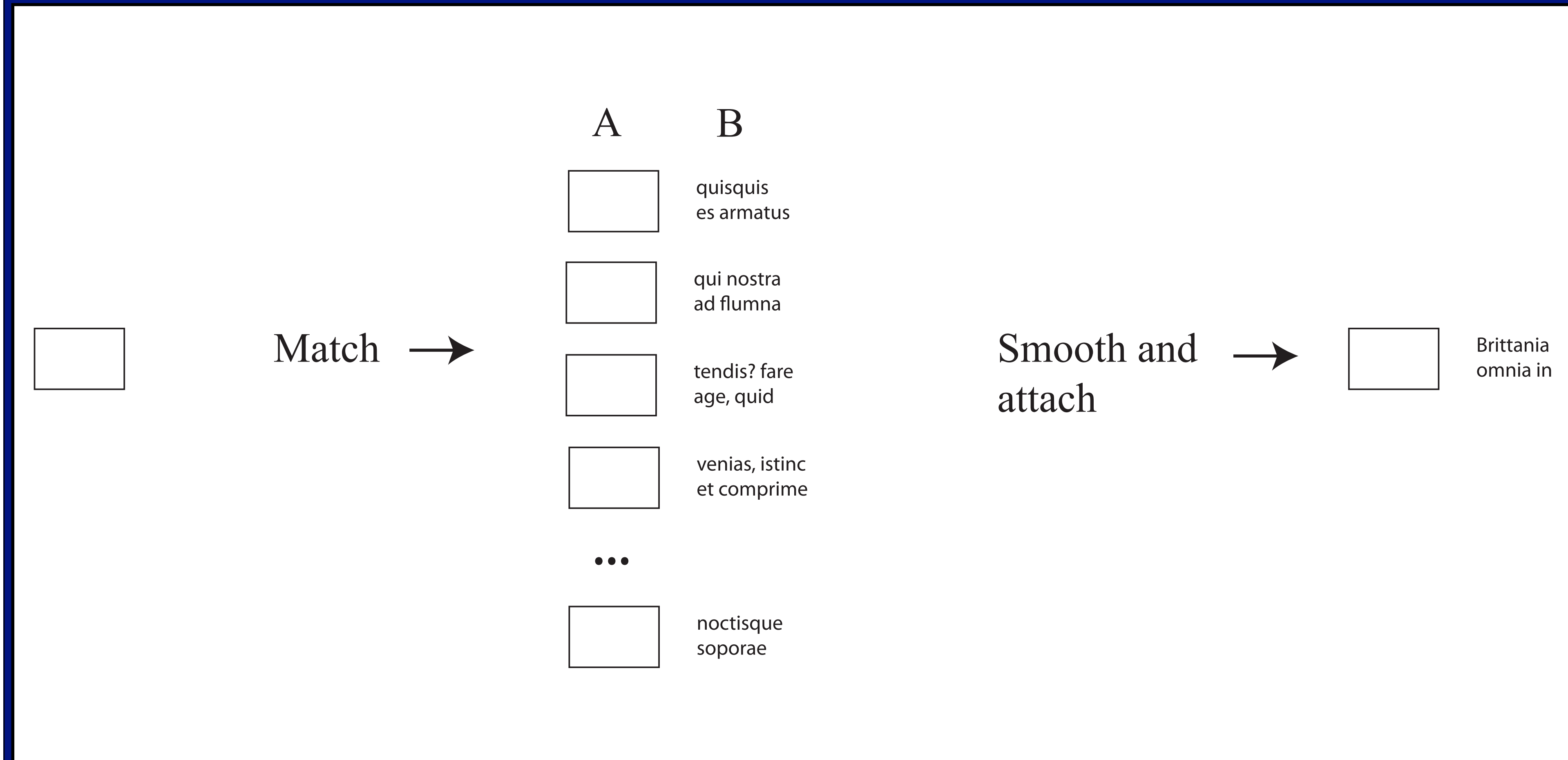
How do we assess different datasets?

- By what they are for
  - activity vs category
- By what they cover
  - many cases vs few
- By how well they represent the problem
  - in some special cases, it is easy to tell
  - what is the problem?
- By how big they are
  - easy!

Conclusion

- 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

Non-parametric regression



With a broad view of “match”, “smooth”, all classifiers fit into this story

A=Image, B=Body pose

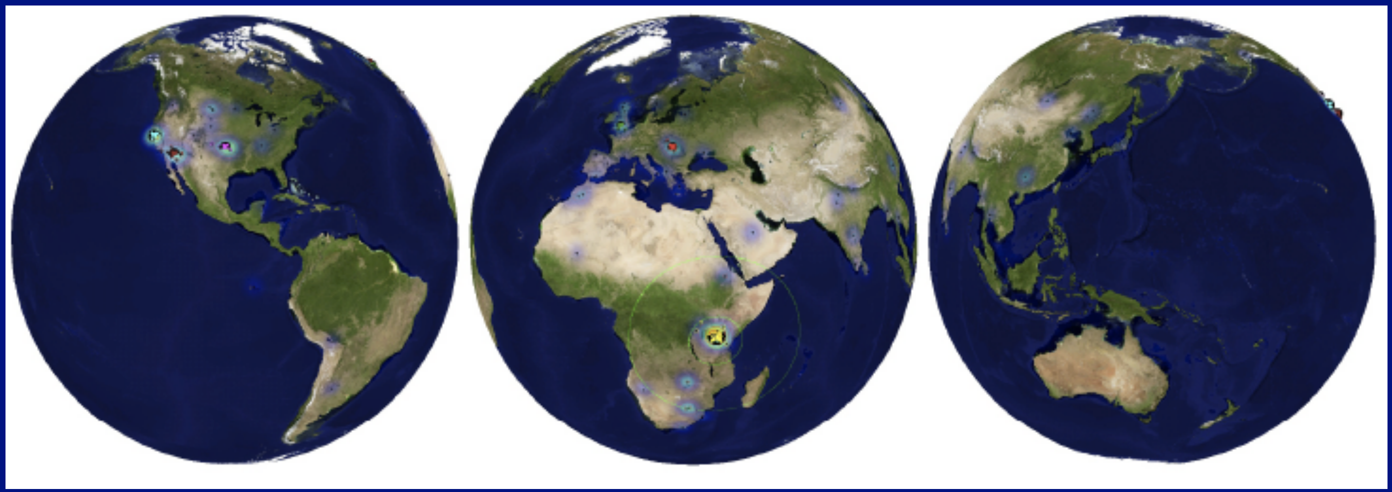
- Rosales+Sclaroff, 00; Shakhnarovich+Darrell, 03

A=Image with hole, B=fill-in

Efros+Leung, 99; Hays+Efros 07

Input Scene Matches Output

A=picture, B=location



Hays+Efros, 08

A=motion window, B=words

Laptev Perez 2007; see also Laptev et al 08

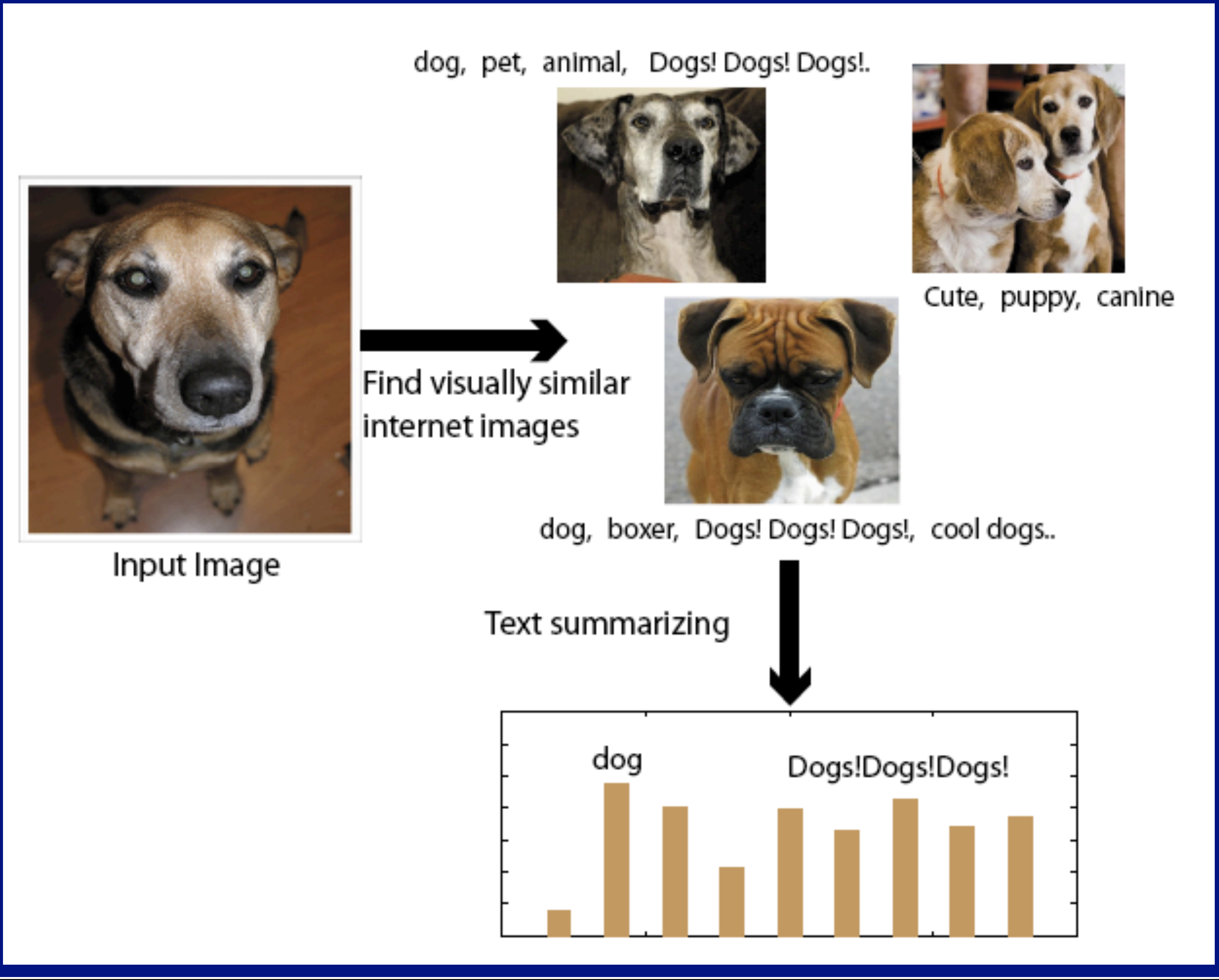
A=face image, B=name

President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters

Berg et al 04, 05; Guillaumin et al 08; Everingham et al 06; Ozkan et al 06; Zhao et al 08; Yagnik et al 07; lots of others

A=picture, B=words

Wang et al 09



A=picture, B=Sentence

A man stands next to a train on a cloudy day
A backpacker stands beside a green train
This is a picture of a man standing next to a green train
There are two men standing on a rocky beach, smiling at the camera.
This is a person laying down in the grass next to their bike in front of a strange white building.

Farhadi et al 10

Recognition datasets

- Collection strategies
  - Web pix + fix
    - Flickr
    - Google image search
    - Microsoft image search
  - Existing collections
    - Corel
  - Photograph yourself
  - Photograph isolated, then enrich

Gotchas!

- Web pix+fix
  - Bias (more later!)
  - Might be few of the right kind (Sapp et al 08)

This difficulty probably exaggerated

A great **hammer** to **hammer ...**
400 x 378 - 67k - jpg
www.drukhier.nl

... hammer in **...**
386 x 385 - 7k - jpg
rubayeet.wordpress.com

The **Hammer** is the most basic of all **...**
386 x 385 - 9k - jpg
homerepair.about.com

hammer
300 x 400 - 15k - jpg
bombmatt.wordpress.com

If I had a **hammer**
490 x 433 - 7k - gif
www.edspresso.com

... Hammer ...
600 x 710 - 172k - jpg
uzar.wordpress.com

Hammer
400 x 340 - 20k - jpg
www.bbc.co.uk

Geological **hammer ...**
307 x 307 - 15k - png
commons.wikimedia.org
[More from upload.wikimedia.org]

RIP **HAMMER** HAM1
350 x 350 - 13k - jpg
www.ancinterproducts.com

Hammer OS Certified Systems Engineer
300 x 450 - 23k - jpg
hammeros.wordpress.com

Ultimate Geeks Multi Tool **Hammer**
382 x 351 - 39k - jpg
nexus404.com

Stone **Hammer ...**
360 x 360 - 9k
www.germes-online.com

6) **Hammer** :
1600 x 1200 - 66k - jpg
library.thinkquest.org

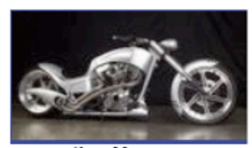
... hammer beer bottle opener.
450 x 381 - 9k - jpg
www.geekologie.com

shingler's **hammers**
300 x 300 - 9k - jpg
www.daviddarling.info

Claw **Hammer**
360 x 360 - 7k - jpg
www.lakewoodconferences.com
[More from www.lakewoodconferences.com]

Machinists **Hammer** With Fibre Glass **...**
360 x 360 - 7k
zhukeqiang.en.alibaba.com

We didn't really use **hammers** much **...**
500 x 362 - 29k - jpg
ocw.mit.edu



... the **Hammer**.
620 x 344 - 47k - jpg
www.didntyouhear.com

Hammer toe
400 x 320 - 11k - jpg
www.mdconsult.com

Large Chocolate **Hammer**:
504 x 262 - 20k - jpg
www.creativechocolatesofvt.com

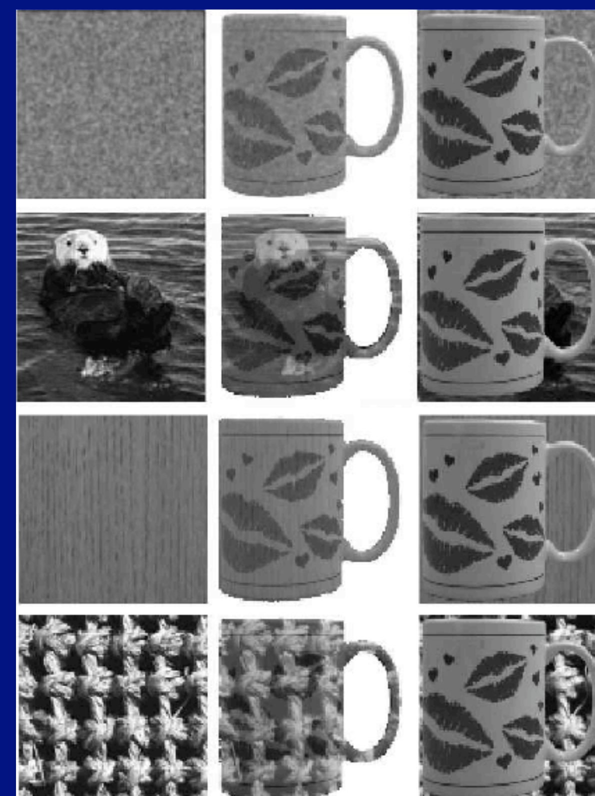
Gotchas!

- Existing collections
  - mainly stock photo's like Corel
  - Massive bias issues with corel
    - one can predict CD number from color histogram rather accurately (Chappelle et al, 99)

- Photograph yourself
  - hard work

Gotchas!

- Enriching
  - Use a probabilistic "model" to
    - enrich background
    - vary foreground
  - DANGER
    - strong unnatural high frequencies at blend
    - unnatural illumination relations
    - no surface texture distortion



- Random
  - Example: aspect and symmetry

Sapp Saxena Ng, 08 AAAI

Recognition datasets

- Taxonomy strategies
  - Choose some categories (Fei-Fei 04; Griffin 07; Everingham 06)
  - Wordnet (Deng 09)
  - Other?
- Labelling strategies
  - query image search, check responses (Fei-Fei 04; Griffin 07; Everingham 06)
  - tagging by volunteers
    - benevolent people (Antonio's mom) (Russell 08)
    - game players (von Ahn 04)
  - tagging by paid annotators (Yao 07; Sorokin 08)
    - Go to Alex and Fei-Fei's tutorial on Friday
  - active learning (Berg, 06; Li, 06; Wang 08; Collins 08)

Turk experience outside vision

- HLT-NAACL workshop 2010
  - proceedings out two weeks ago
  - competition: make a nice NLP dataset for less than $100
- http://behind-the-enemy-lines. blogspot.com/2010/03/ new-demographics-of-mechanical-turk. html

Why do you complete tasks in MTurk?	US	India
To spend free time fruitfully and get cash (e.g., instead of watching TV)	70%	60%
For "primary" income purposes (e.g., gas, bills, groceries, credit cards)	15%	27%
For "secondary" income purposes, pocket change (for hobbies, gadgets)	60%	37%
To kill time	33%	5%
The tasks are fun	40%	20%
Currently unemployed or part time work	30%	27%

Turk experience outside vision

- http://behind-the-enemy-lines. blogspot.com/2010/03/ new-demographics-of-mechanical-turk. html

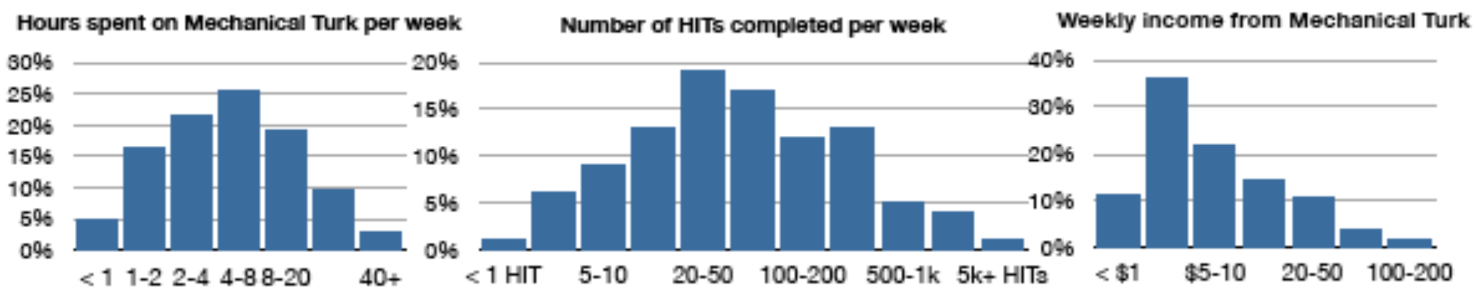


Figure 1: Time spent, HITs completed, and amount earned from a survey of 1,000 Turkers by Ipeirotis (2010).

Design remains hard

- When we get poor results, is it because
  - the interface is poor (e.g. confusing buttons)
  - the task is hard (e.g. mark all pixels such that ...)
  - the task is unnatural (e.g. are red cats heavier than blue dogs)

Conclusion

- 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

Conclusion

- 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

You can't get away from bias by saying you must know your problem well before you collect

Big questions

Computer vision

- What signal representation should we use ?

PLUMBING

Computer vision

- What should we say about visual data?

Taxonomy/Category problems

- A choice of taxonomy is a profound commitment
  - which may enhance/distort future research
- Examples:
  - Recognition

Object recognition = k class classification

- current data sets ok,
  - 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)

Are these monkeys?

Object recognition = describing what objects are like

- most current datasets
  - are largely of the wrong form
    - and no declarative data about objects
  - bias is intrinsic
    - and intertwined with representation agendas
- research agenda
  - learning by reading
  - similarity
  - coping with induction issues
  - sensible responses to objects of unknown category
  - within class variance has semantics
  - architectures, representations, semantics

Representational agenda
may be driven by bias
in datasets

Conclusion

- 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

You can't get away from bias by saying you must know your problem well before you collect