Which way forward? AI + vision

Larry Zitnick

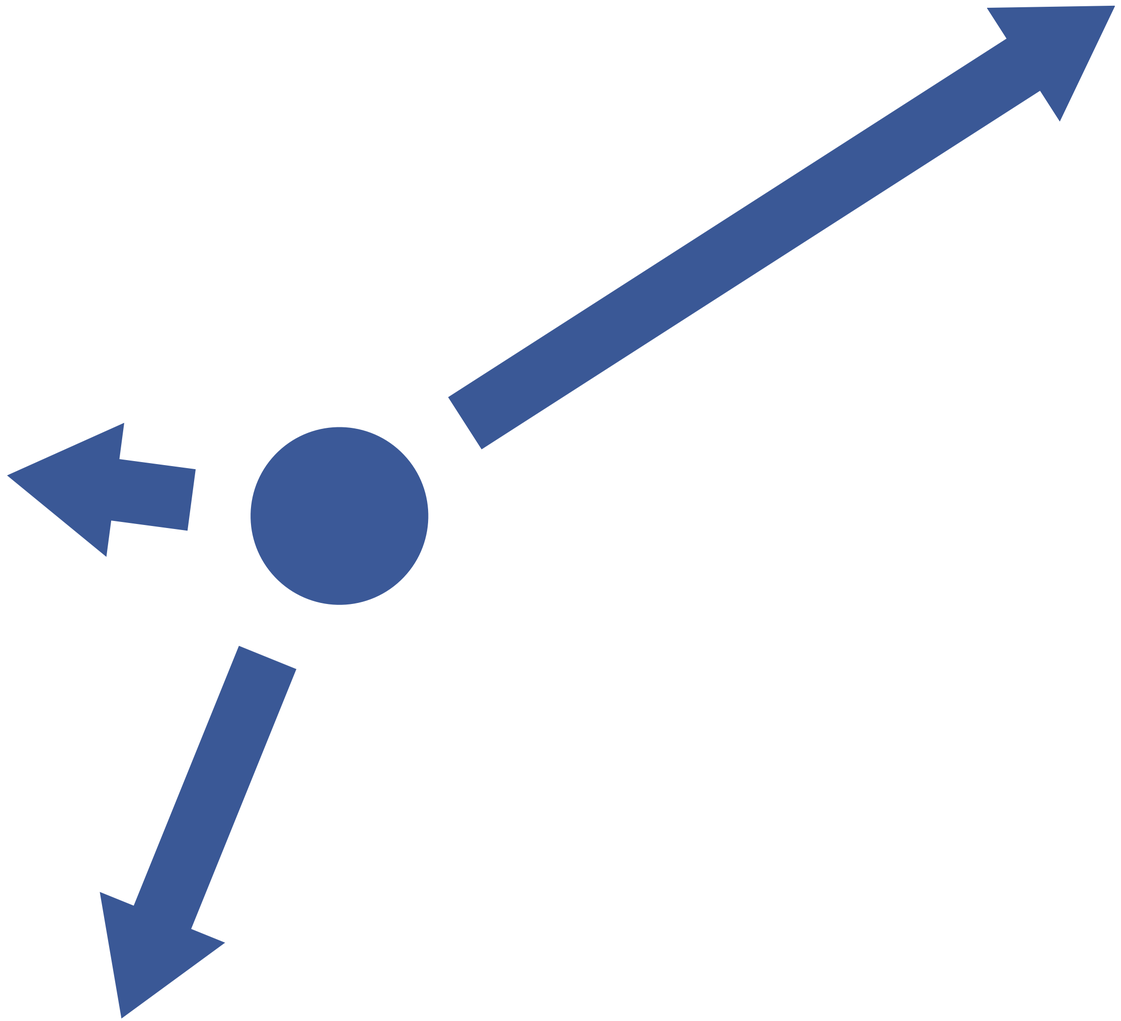
Lead, Facebook AI Research

95% of research isfailure

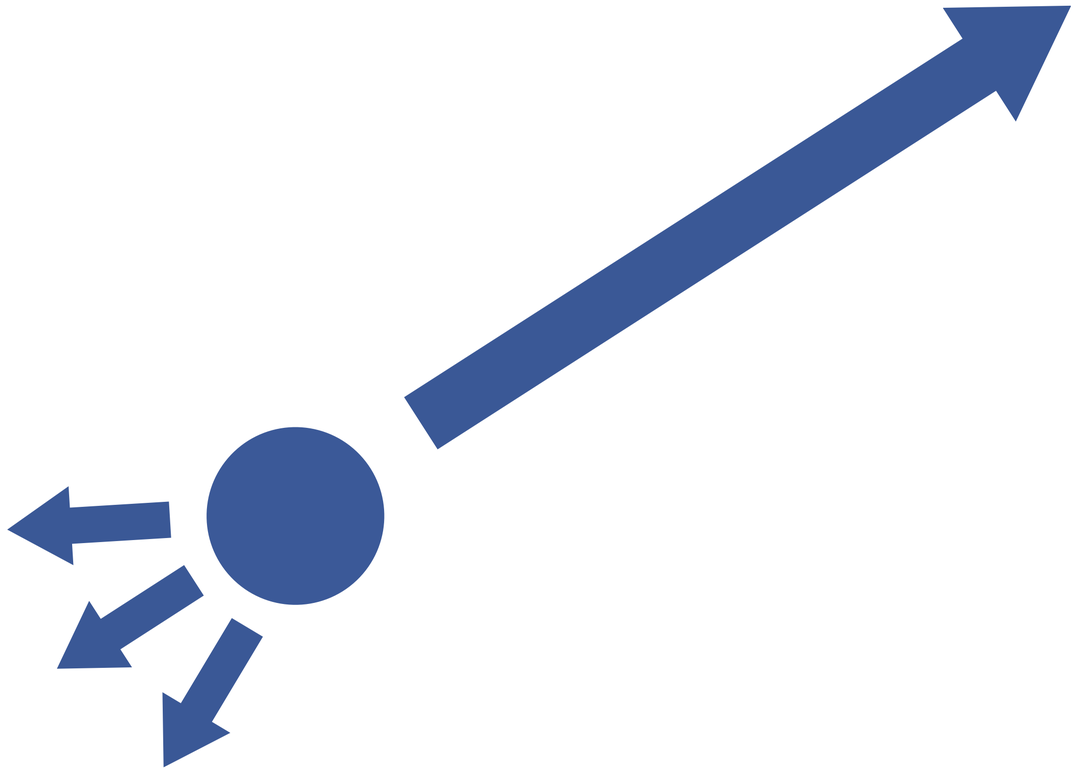
50% of internships fail

The point of research is not to publish…

… it’s to have impact.



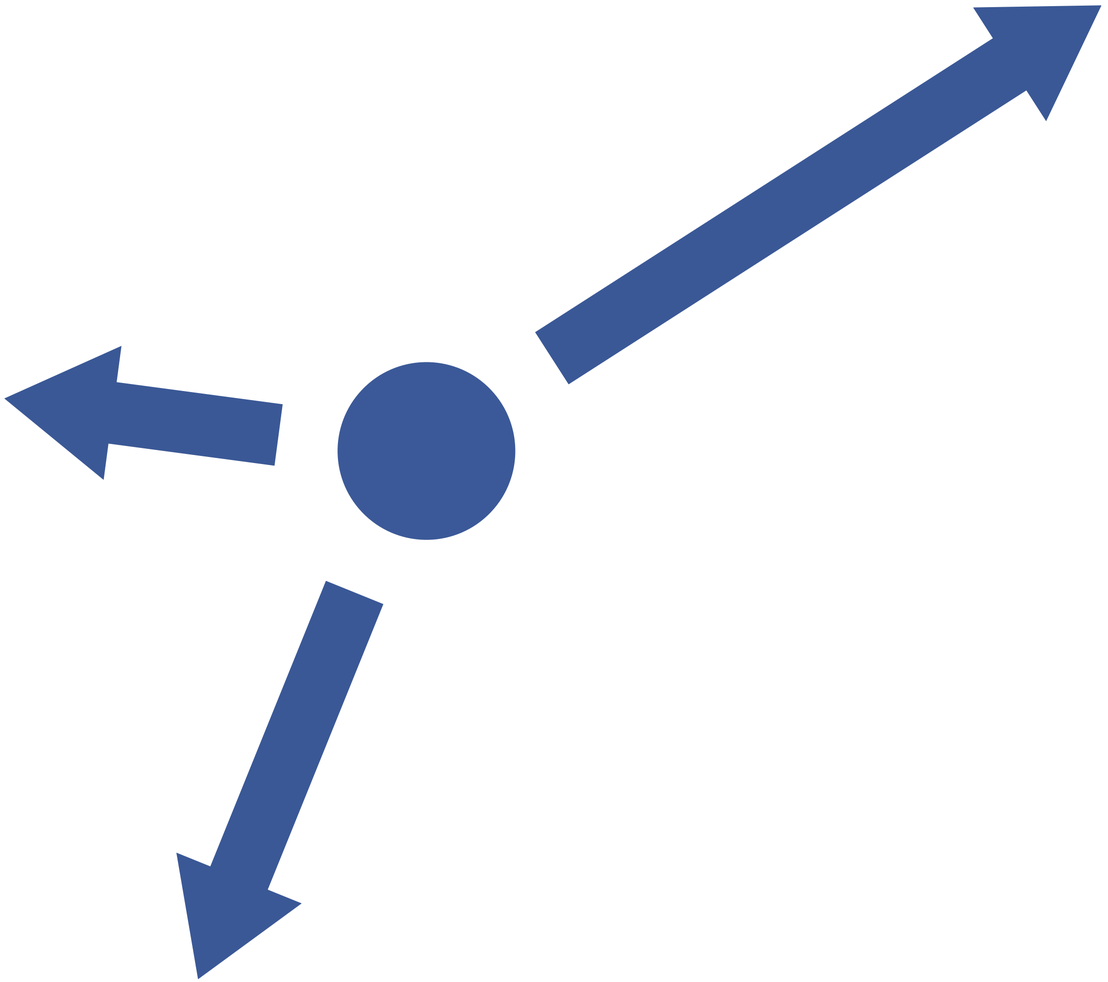
Negative ideas

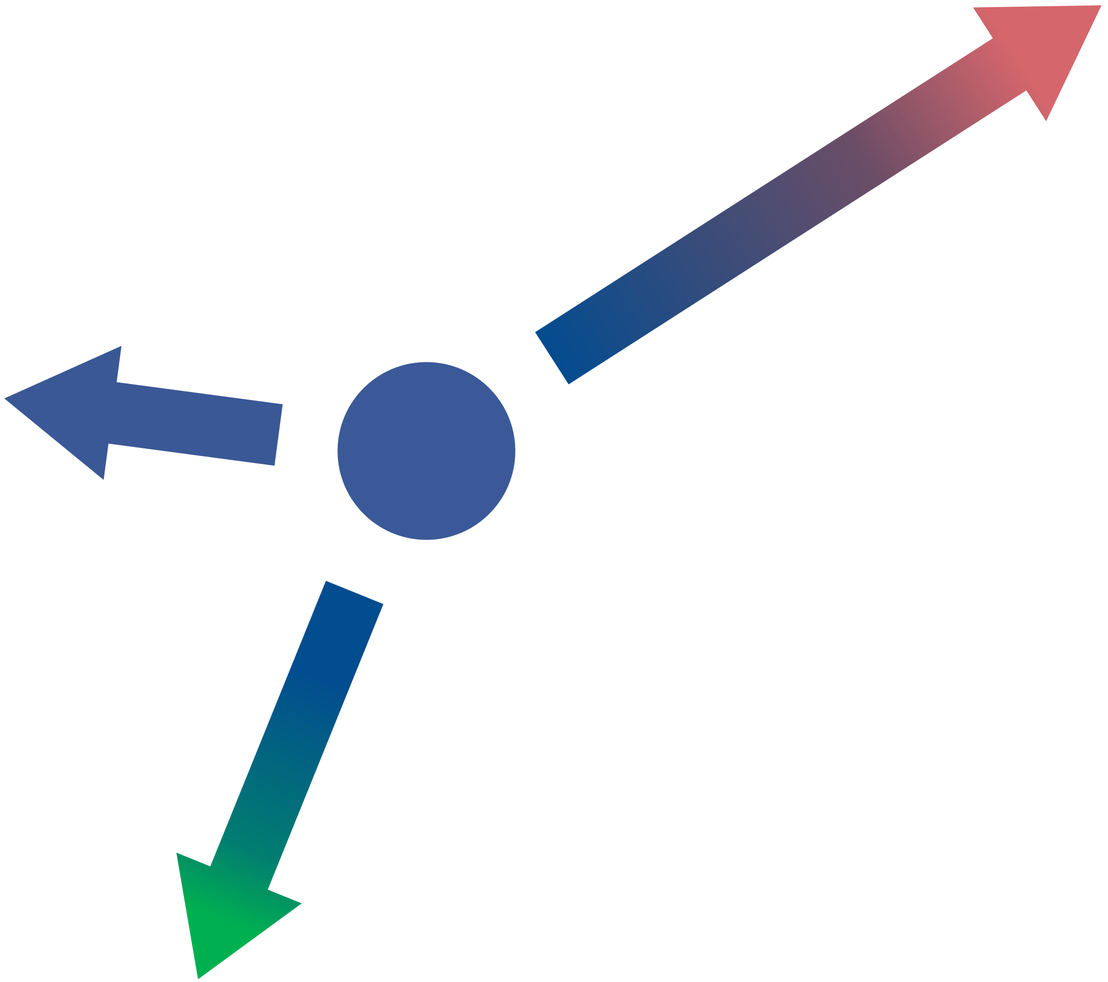


Impact shapes the research field.

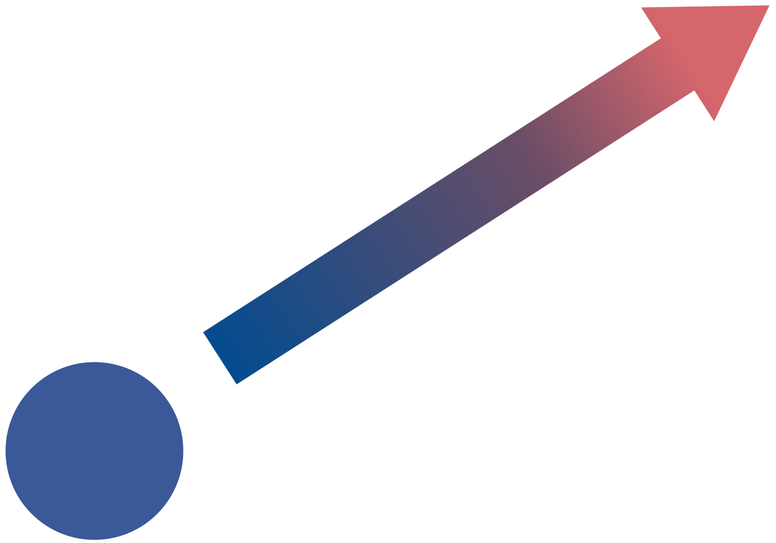
Impact may be positive …

… and it may be negative.



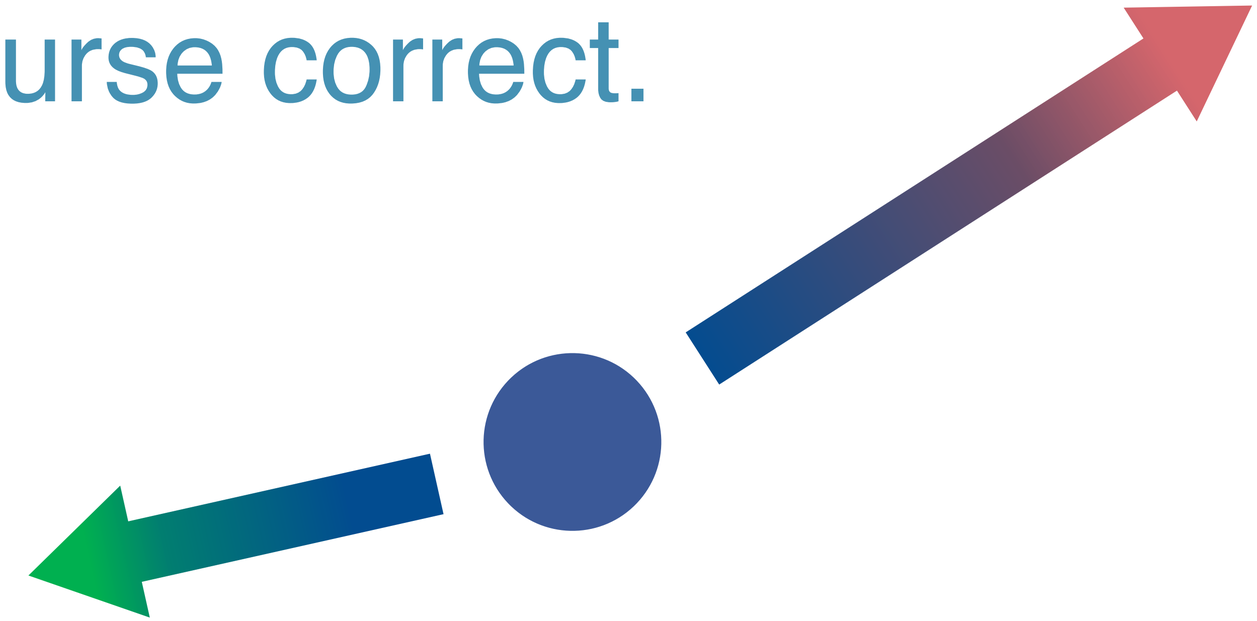


What does it mean to have negative impact?



No external impact.
Uncertain impact.
Misinterpreted impact.

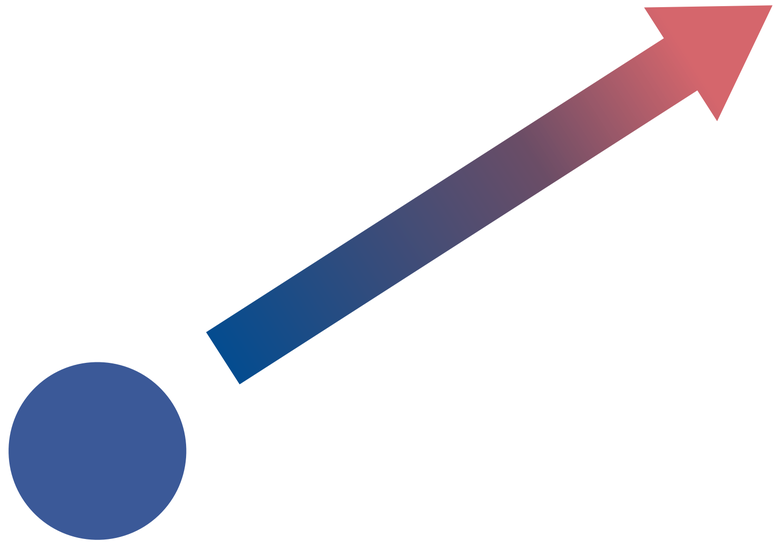
Need to course correct.



Good negative ideas go counter to the prevailing wisdom.

Negative results are commonly not impactful.

How to avoid this?

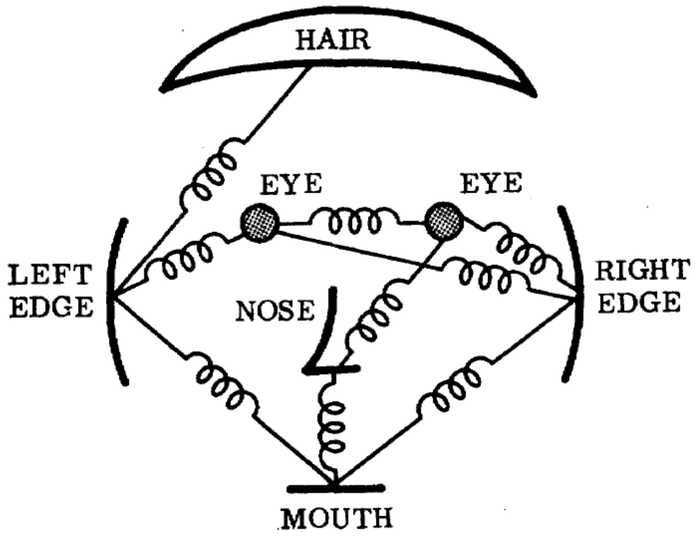


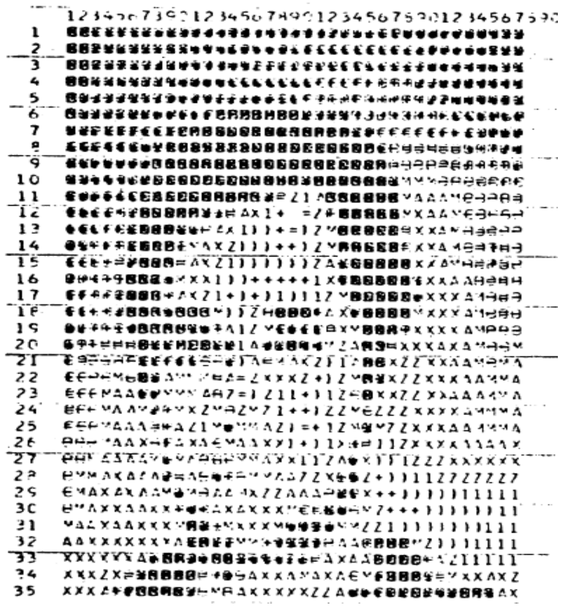
1. A case study in language + vision tasks

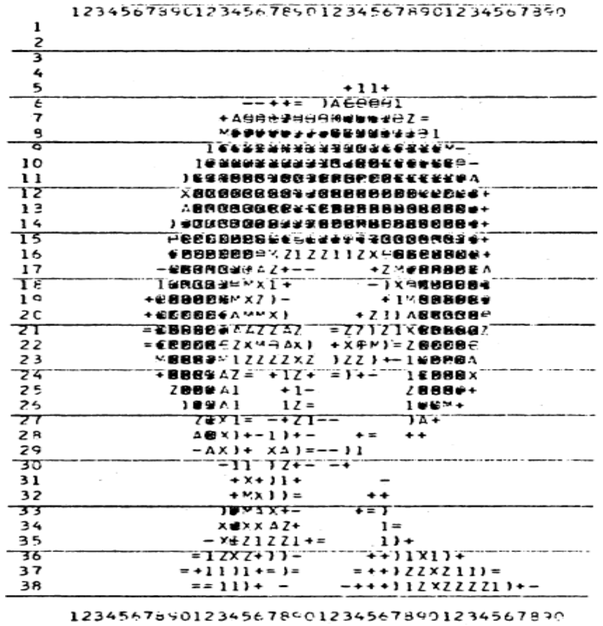
2. The approaching challenges with AI + vision

Let’s look back…

1973

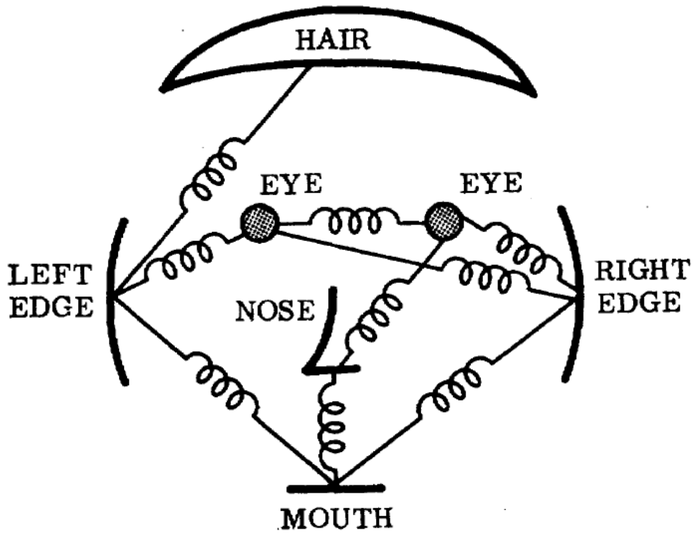


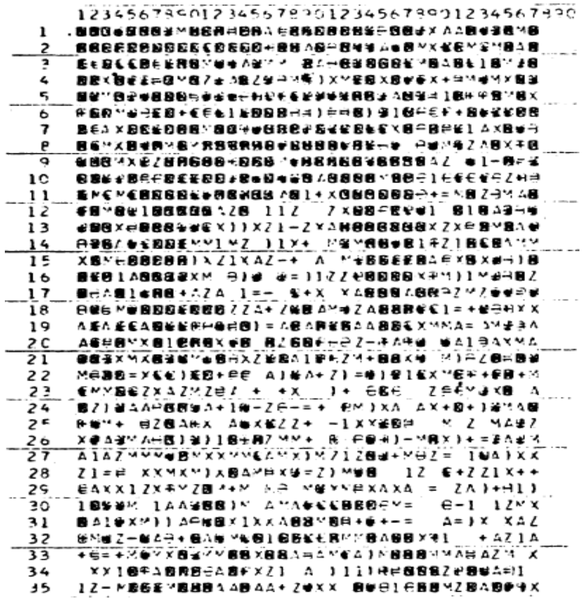


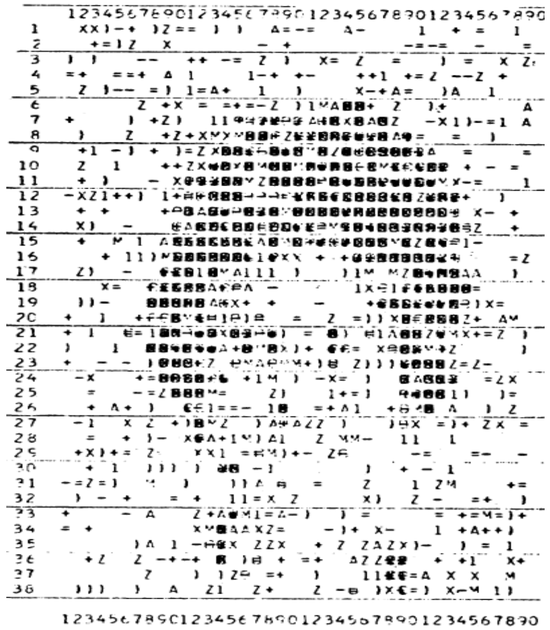


The representation and matching of pictorial structures, Fischler and Elschlager, 1973

1973

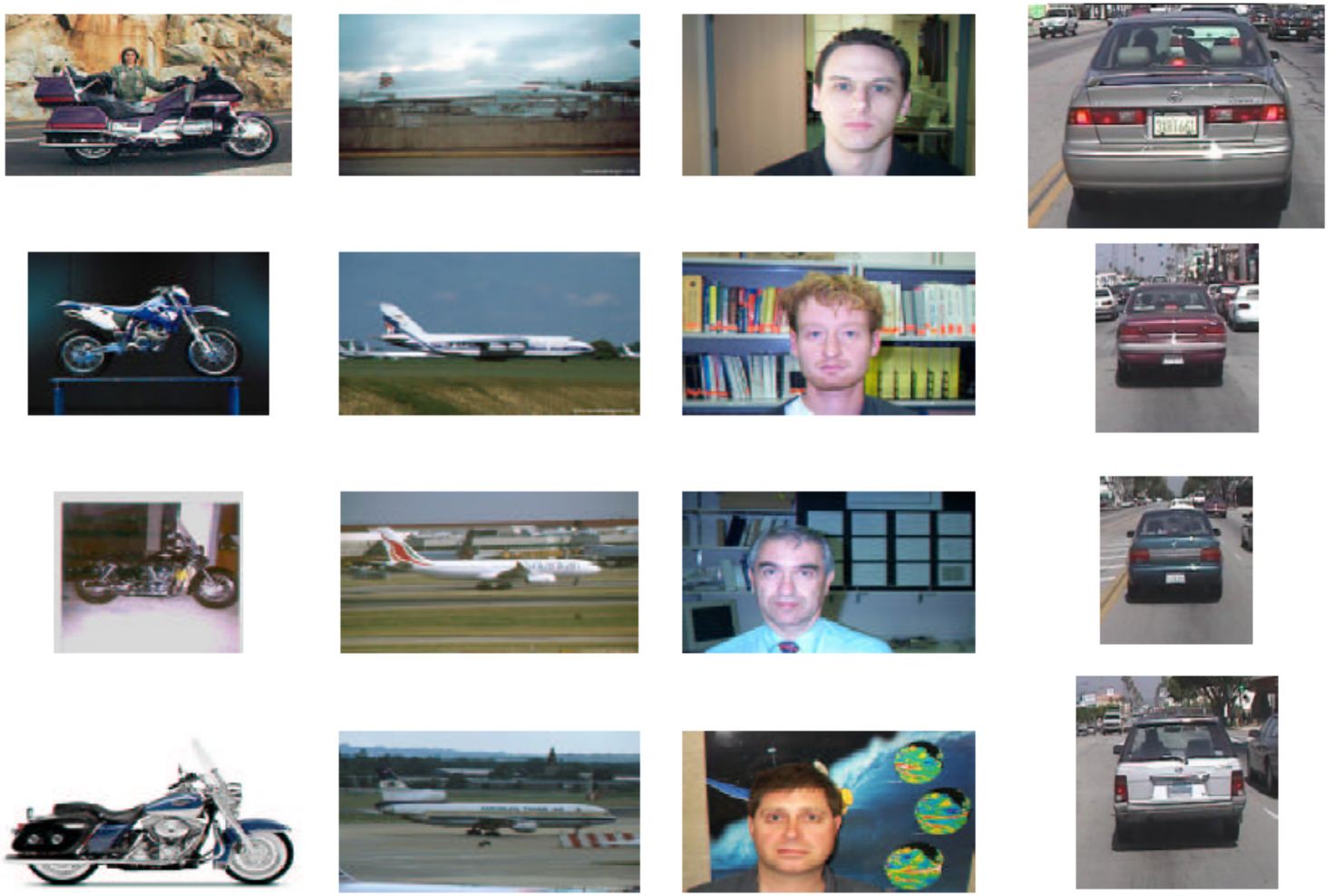






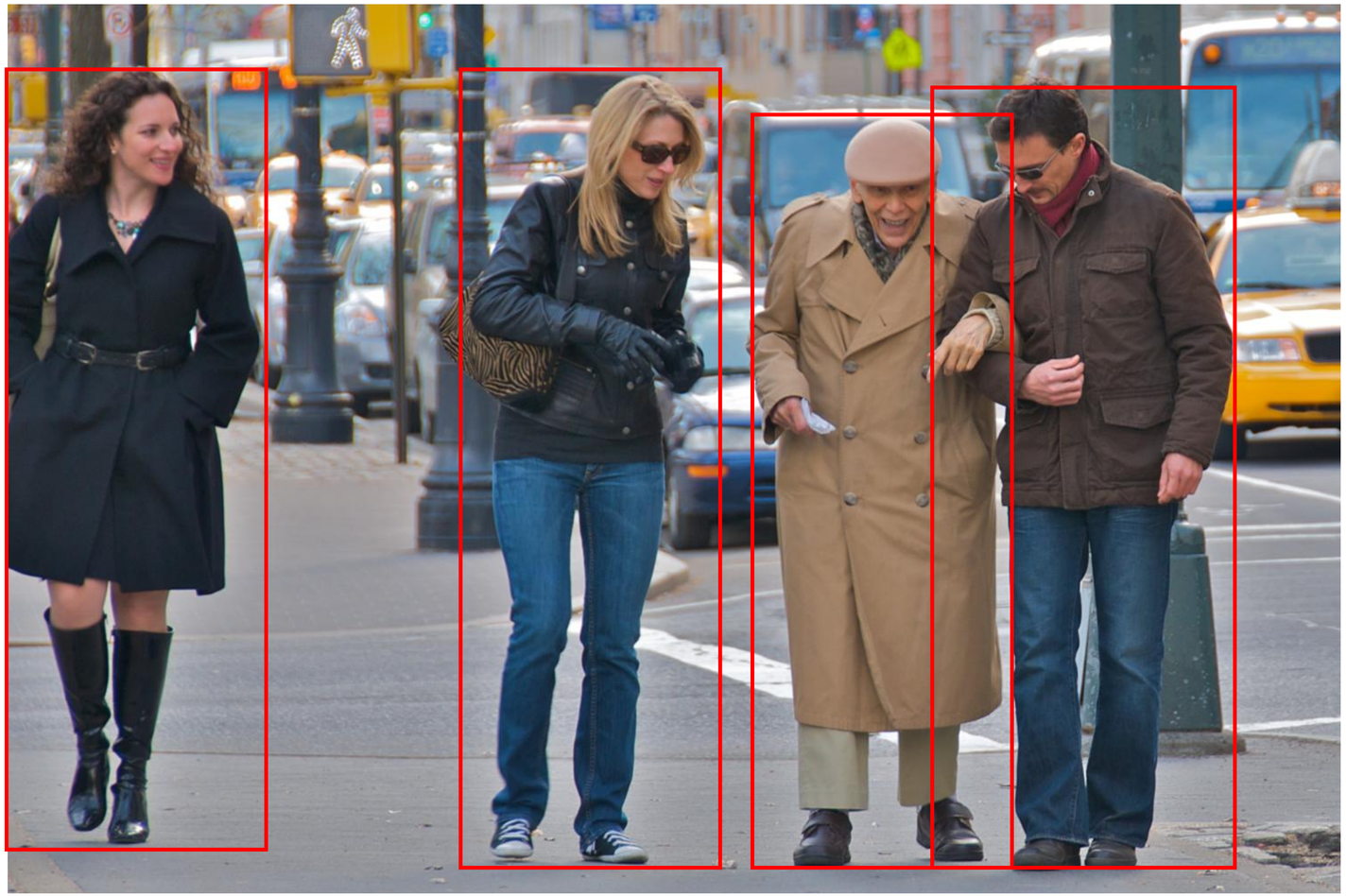
The representation and matching of pictorial structures, Fischler and Elschlager, 1973

2003



Object Class Recognition by Unsupervised Scale-Invariant Learning, Fergus et al., *CVPR* 2003.

INRIA Person Dataset



Histograms of oriented gradients for human detection, Dalal and Triggs, CVPR 2005.

How to write a paper:

1. Come up with algorithm.

2. Find/create dataset that works.

The beginning of a new era…


IMAGENET


PASCAL2
Pattern Analysis, Statistical Modelling and
Computational Learning


VQA
Visual Question Answering


COCO
Common Objects in Context

How to write a paper:

1. Pick a dataset.

2. Find an algorithm that works.

How to create a dataset:

1. Pick a problem.

2. Create a challenging LARGE dataset.

Image Captions

COCO
Common Objects in Context

160,000 images
5 captions per image

A man checking out a parked black scooter.
A person standing near a small motorcycle on a city street.
A man in a white shirt is looking at a three wheeled motorcycle.
A man looks down at two low riding motor bikes.
A guy staring at a weird looking bike.

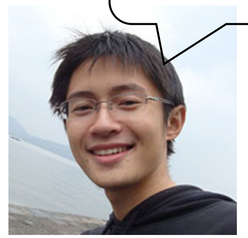
Timeline

August, 2014

120,000 images x 5 captions per image =
600,000 captions

The Great Freak Out

August, 2014 October



COCO
Common Objects in Context

Hao Fang
UW

Tsung-Yi Lin
Cornell Tech

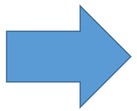
Xinlei Chen
CMU

Rama Vedantam
VT

Evaluation server = Hidden GT test data

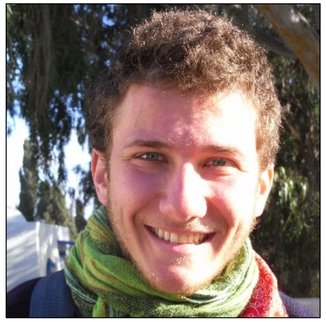
The Reckoning

August, 2014 October — April, 2015

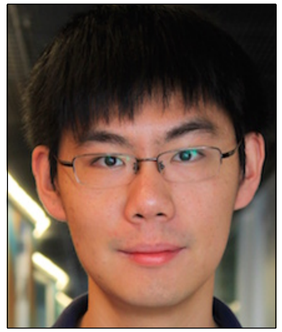


	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MSR[5]	0.912	0.247	0.519	0.695	0.526	0.391	0.291
Berkeley LRCN[1]	0.869	0.242	0.517	0.702	0.528	0.384	0.277
Human[3]	0.854	0.252	0.484	0.663	0.469	0.321	0.217
Google[2]	0.834	0.236	0.498	0.673	0.493	0.362	0.272
m-RNN (Baidu/ UCLA)[8]	0.819	0.229	0.504	0.685	0.512	0.376	0.279
MLBL[4]	0.74	0.219	0.499	0.666	0.498	0.362	0.26
NeuralTalk[6]	0.674	0.21	0.475	0.65	0.464	0.321	0.224
Tsinghua Bigeye[7]	0.673	0.207	0.49	0.671	0.494	0.35	0.241

COCO
Common Objects in Context



Matteo Ronchi
Caltech



Yin Cui
Cornell

Tsung-Yi Lin
Cornell Tech

Advisors:

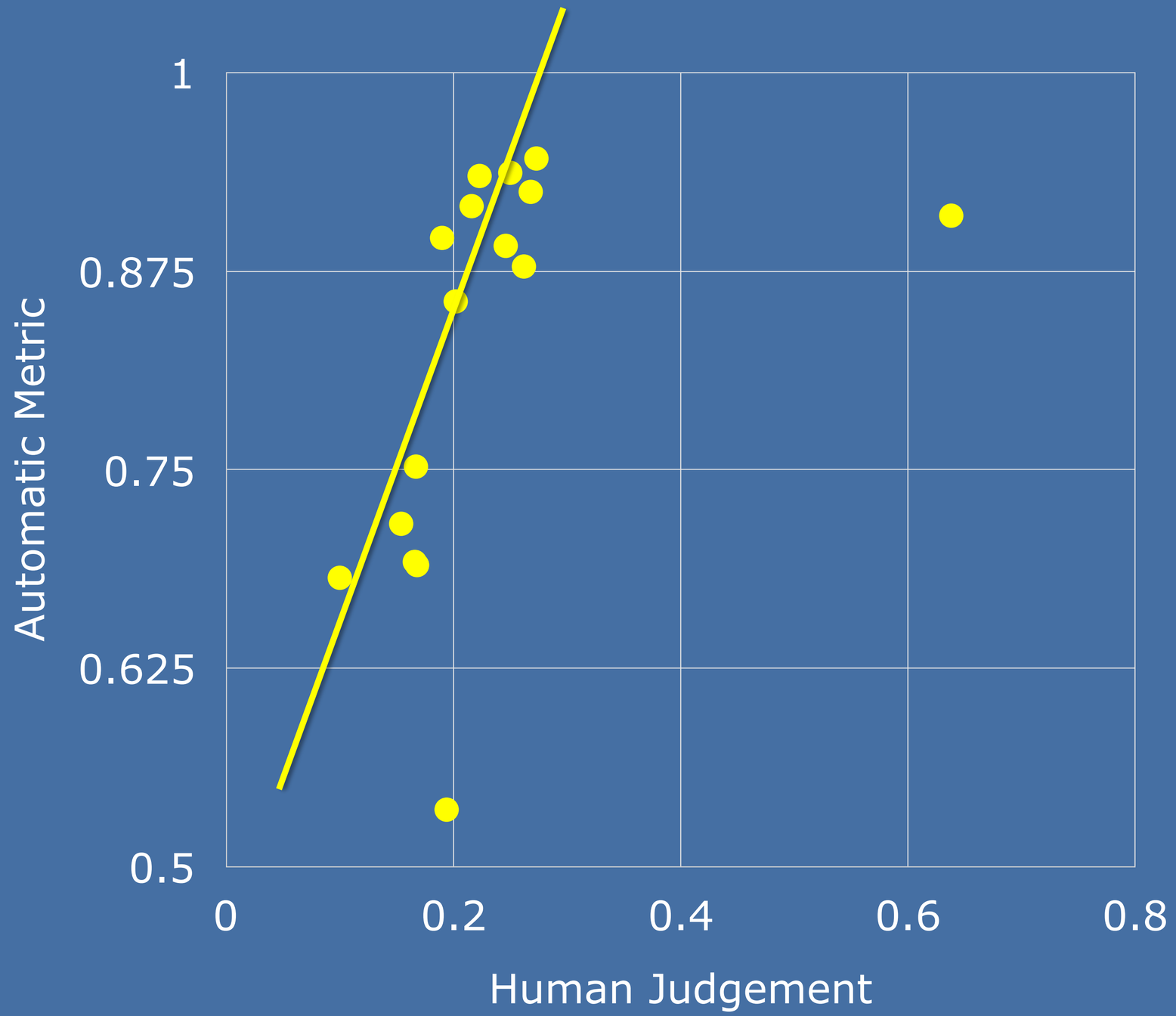
Tamara Berg
Piotr Dollar
Desmond Elliott
Julia Hockenmaier
Meg Mitchell
Devi Parikh
Larry Zitnick

The Enlightening

How do humans rate the captions?

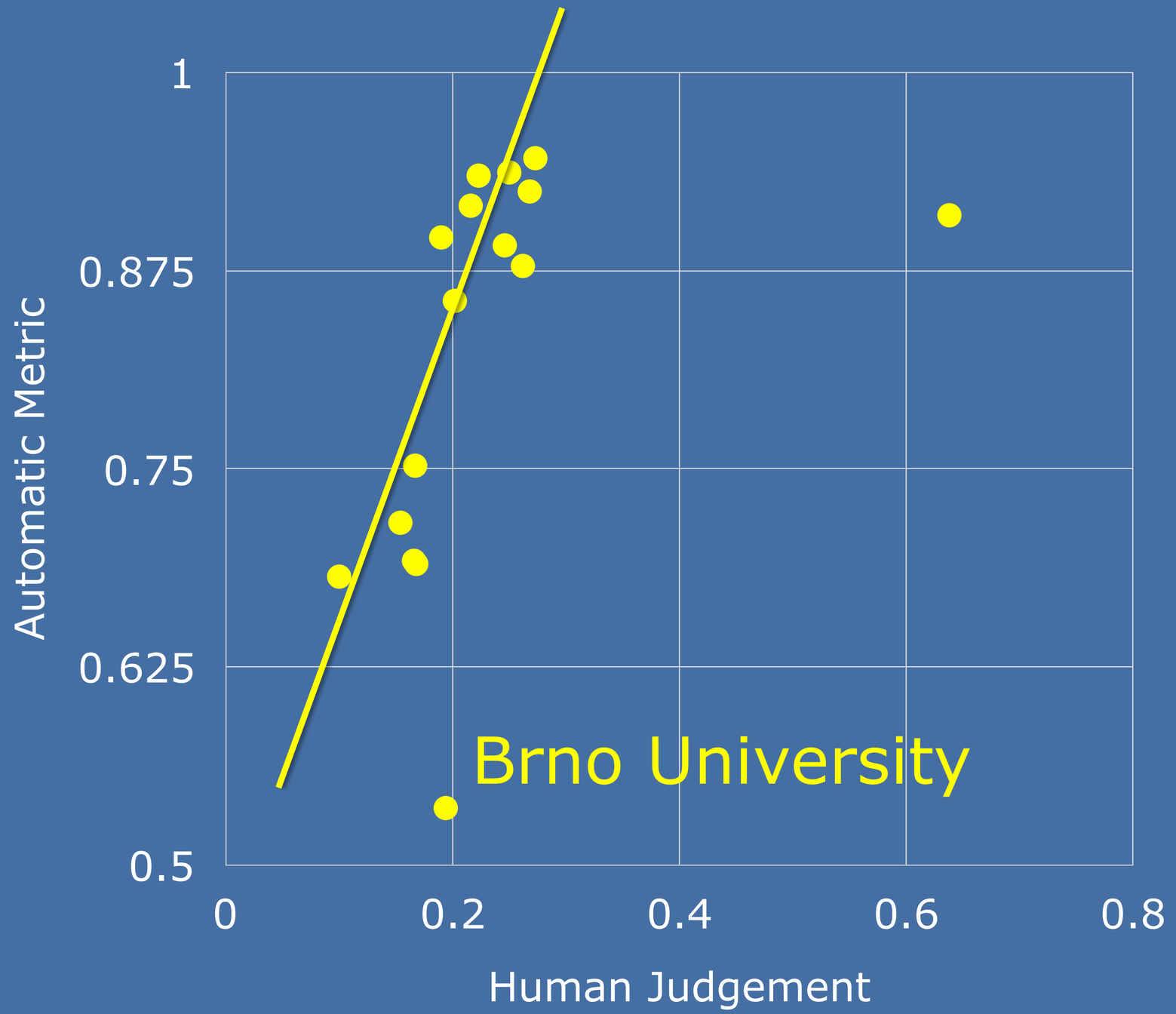
Evaluation

COCO Caption Challenge



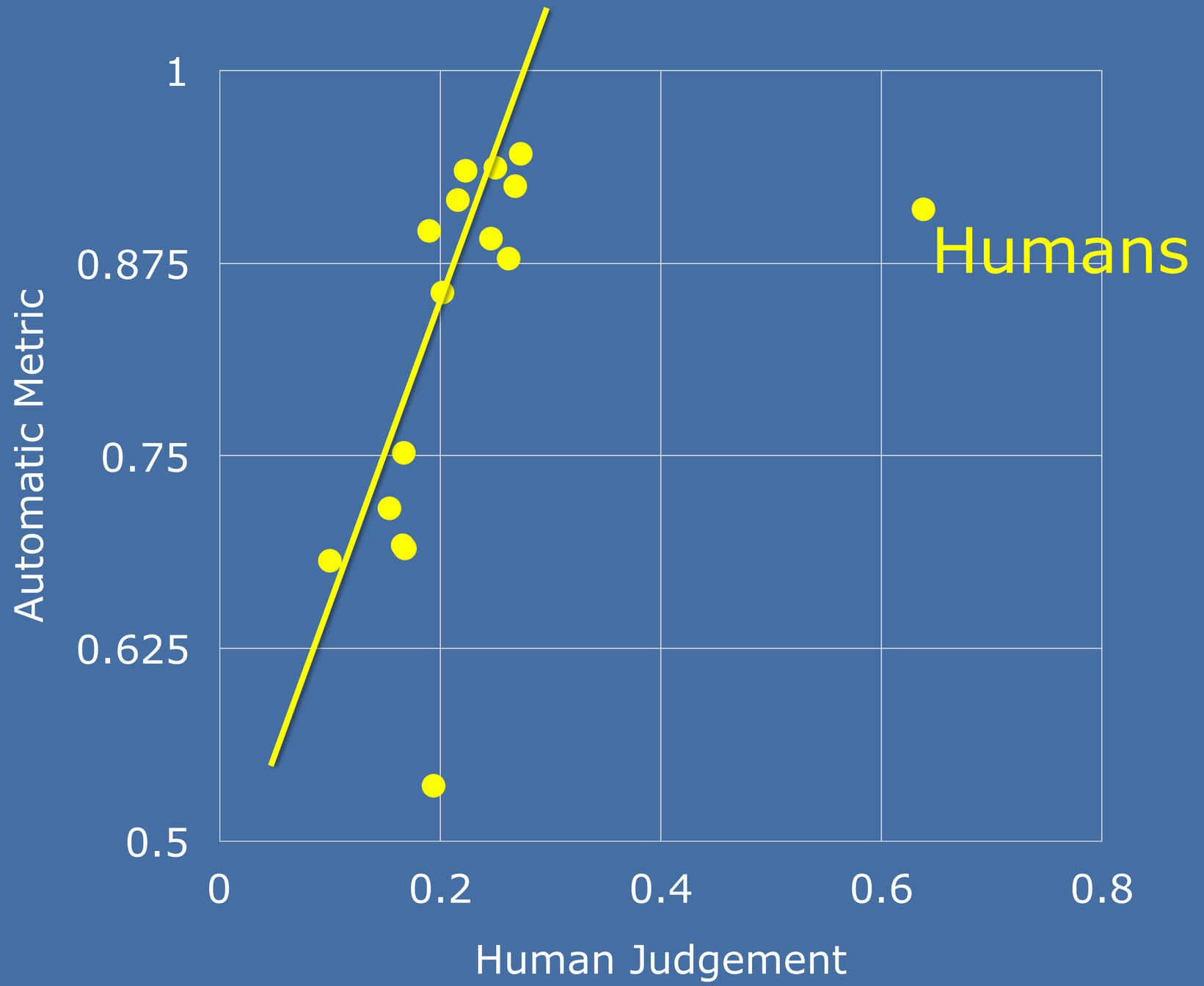
Evaluation

COCO Caption Challenge



Evaluation

COCO Caption Challenge

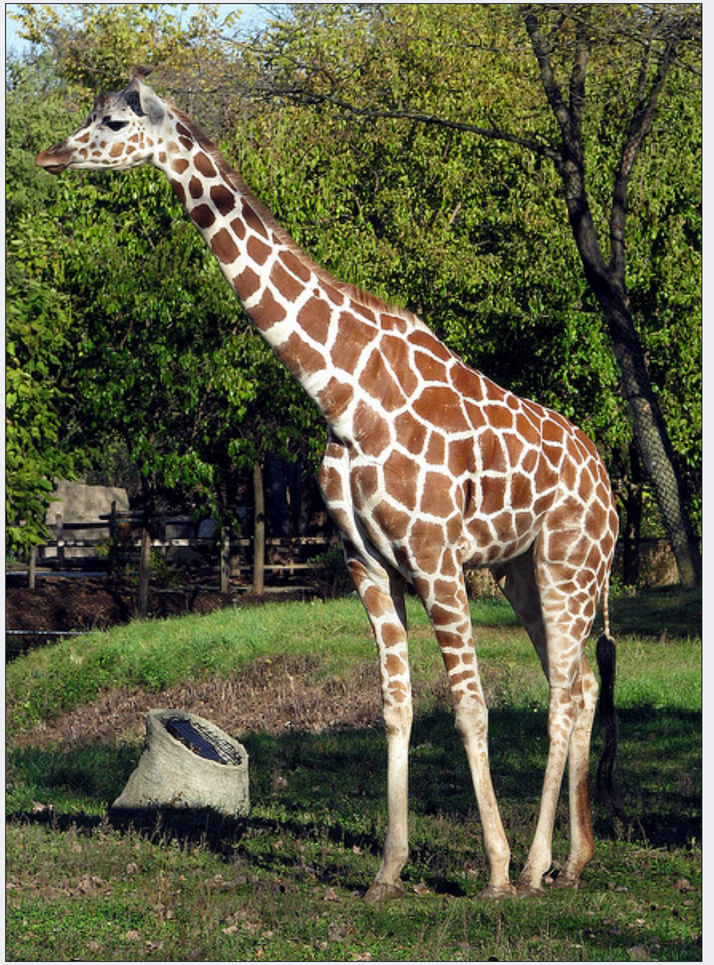


The Enlightening (part 2)

Baselines?

A man riding a wave on a surfboard in the water.

A giraffe standing in the grass next to a tree.



Mind’s Eye: A Recurrent Visual Representation for Image Capt
Generation, Chen and Zitnick, CVPR 2015.

https://www.youtube.com/watch?v=ZUIEOUoCLBo

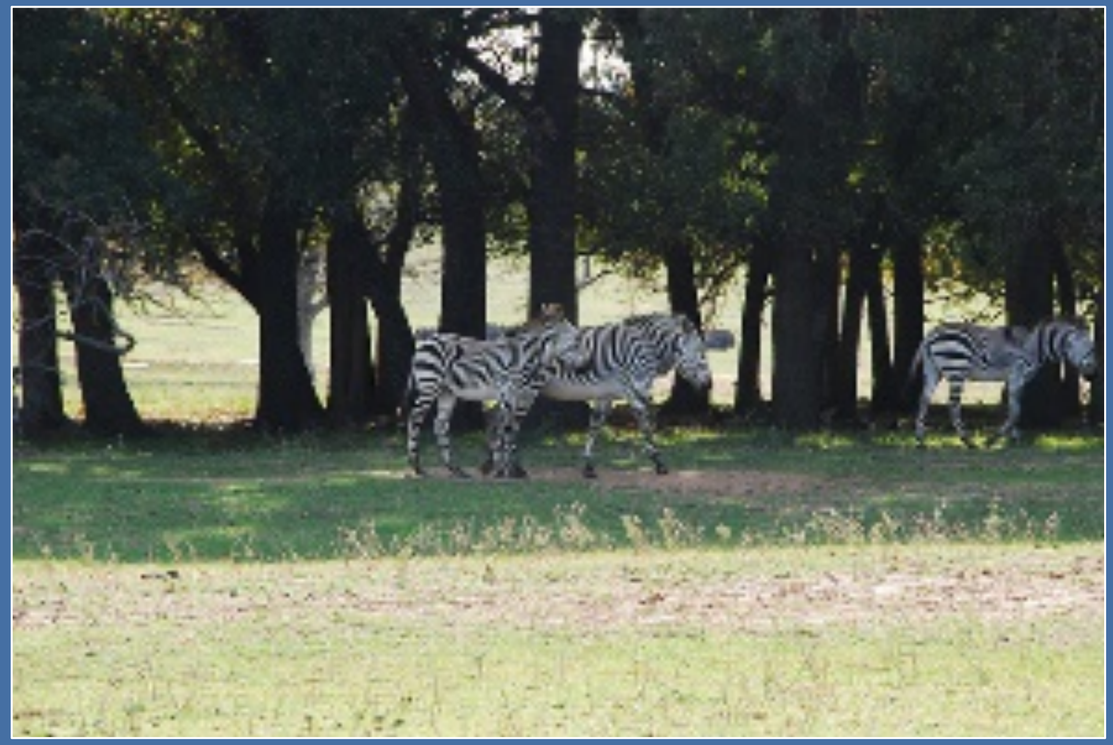
Nearest Neighbor

Test

Train

Nearest Neighbor

A black and white cat
sitting in a bathroom sink.



Two zebras and a giraffe in a field.

See mscoco.org for image information

Results

COCO Caption Challenge

	CIDEr-D	Meteor	ROUGE-L	BLEU-4
Google[4]	0.943	0.254	0.53	0.309
MSR Captivator[9]	0.931	0.248	0.526	0.308
m-RNN[15]	0.917	0.242	0.521	0.299
MSR[8]	0.912	0.247	0.519	0.291
Nearest Neighbor[11]	0.886	0.237	0.507	0.280
m-RNN (Baidu/ UCLA)[16]	0.886	0.238	0.524	0.302
Berkeley LRCN[2]	0.869	0.242	0.517	0.277
Human[5]	0.854	0.252	0.484	0.217
Montreal/Toronto[10]	0.85	0.243	0.513	0.268
PicSOM[13]	0.833	0.231	0.505	0.281
MLBL[7]	0.74	0.219	0.499	0.26
ACVT[1]	0.709	0.213	0.483	0.246
NeuralTalk[12]	0.674	0.21	0.475	0.224
Tsinghua Bigeye[14]	0.673	0.207	0.49	0.241
MIL[6]	0.666	0.214	0.468	0.216
Brno University[3]	0.517	0.195	0.403	0.134

A summary of what we messed up

No evaluation metric

Flawed evaluation metrics

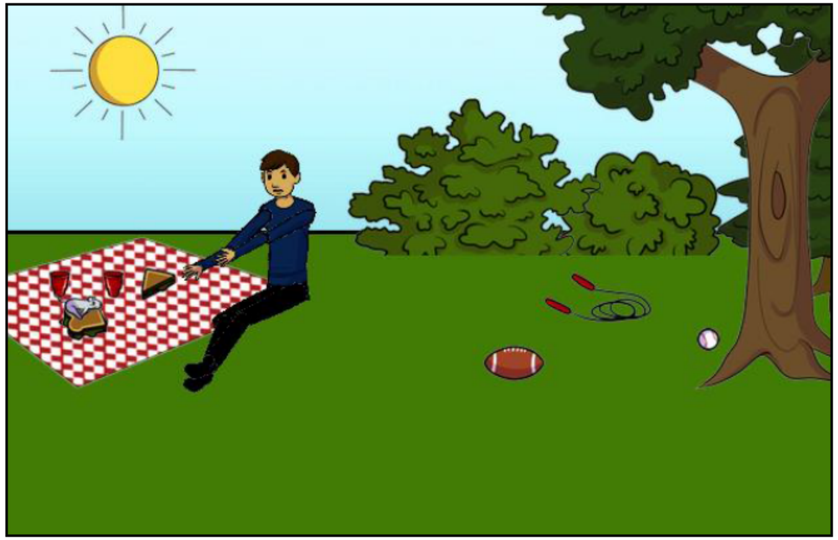
No baselines

Vision + Language (part 2)

VQA: Visual Question Answering

How many slices of pizza are there?
Is this a vegetarian pizza?

Does it appear to be rainy?
Does this person have 20/20 vision?



Is this person expecting company?
What is just under the tree?

VQA: Visual Question Answering
Antol et al., ICCV, 2015.

Bias

VQA Leaderboard

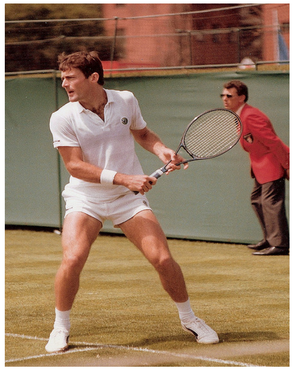
	By Answer Type			Overall
	Yes/No	Number	Other	
UC Berkeley & Sony[14]	83.79	38.9	58.64	66.9
Naver Labs[10]	83.78	37.67	54.74	64.89
DLAIT[5]	83.65	39.18	52.62	63.97
snubi-naverlabs[25]	83.64	38.43	51.61	63.4
POSTECH[11]	81.85	38.02	53.12	63.35
Brandeis[3]	82.53	36.54	51.71	62.8
VTComputerVison[19]	80.31	37.87	52.16	62.23
MIL-UT[7]	82.39	36.7	49.76	61.82
klab[23]	81.9	38.85	49.23	61.59
SHB_1026[13]	82.51	36.77	48.06	61.05
MMCX[8]	81.22	35.9	48.74	60.76
VT_CV_Jiasen[20]	80.86	37.85	47.99	60.46
LV-NUS[6]	81.75	35.14	45.96	59.56

Aishwarya Agrawal
Virginia Tech

What sport is … ?

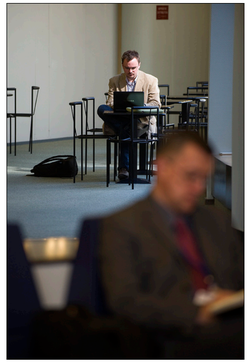
‘tennis’ 41%

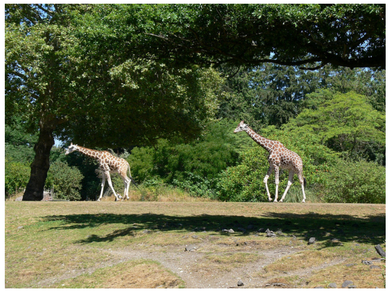


……


How many … ?

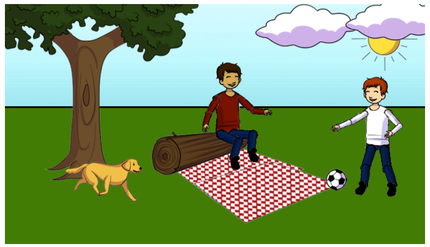
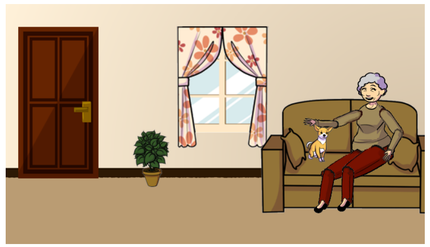
‘2’ 39%



……


What animal is … ?

‘dog’ 35%



……


Slide credit: Yash Goyal and Peng Zhang

Is there a clock … ?

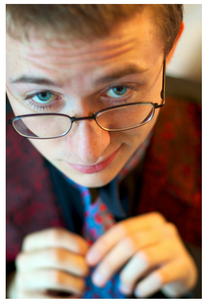
‘yes’ 98%



……


Is the man wearing glasses … ?

‘yes’ 94%



……


Are the lights on … ?

‘yes’ 85%



……


Do you see a … ?

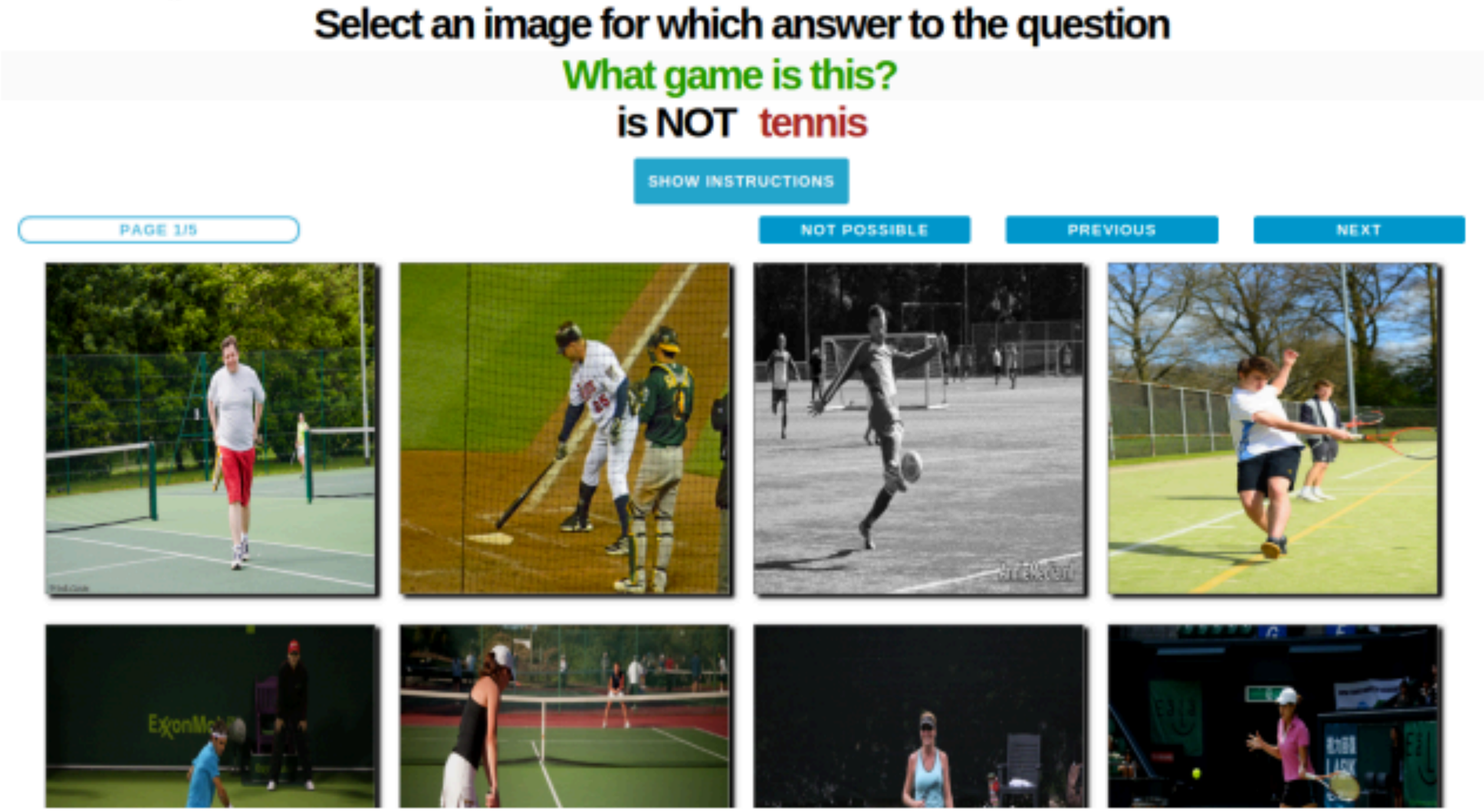
‘yes’ 87%



……

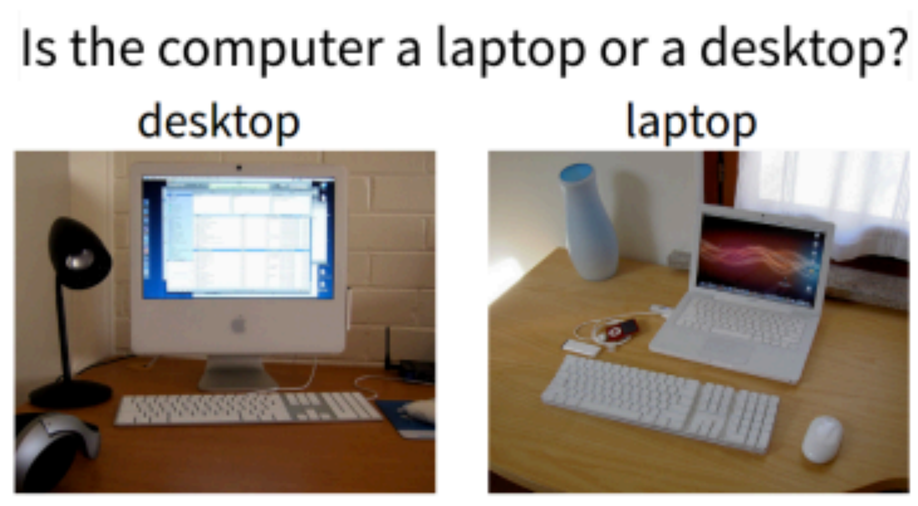

Slide credit: Yash Goyal and Peng Zhang

Balancing



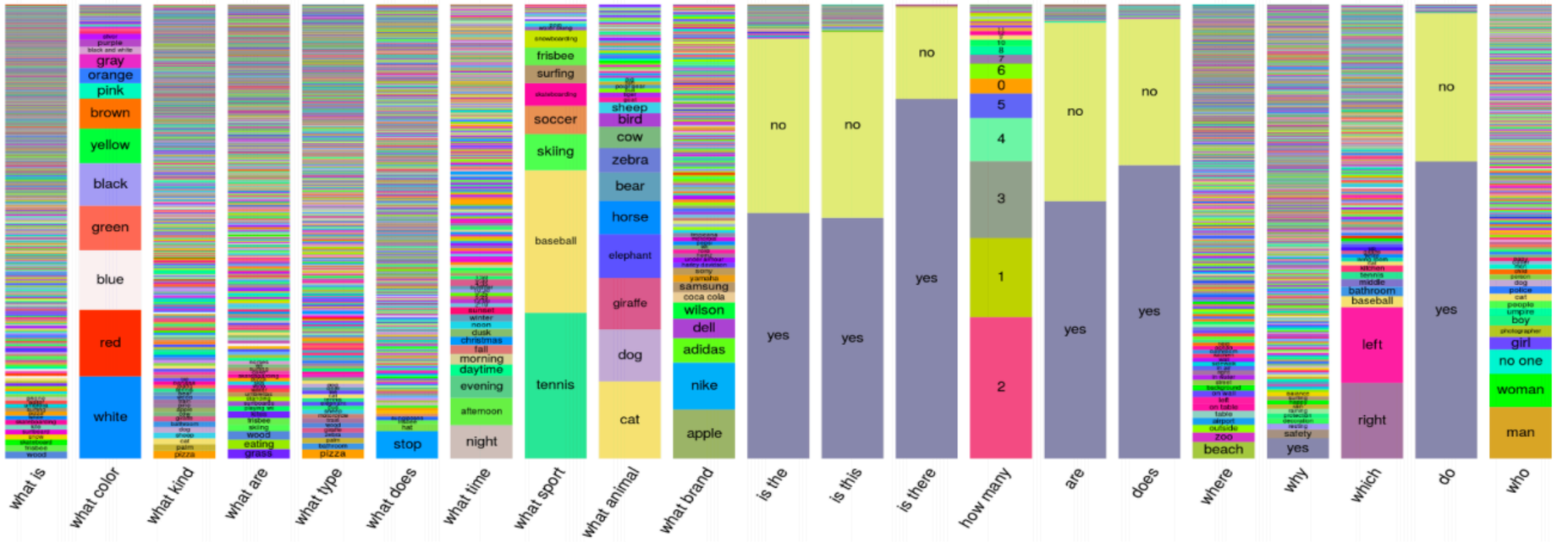
Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, Goyal, Khot, Summers-Stay, Batra, Parikh, 2016.

Balancing

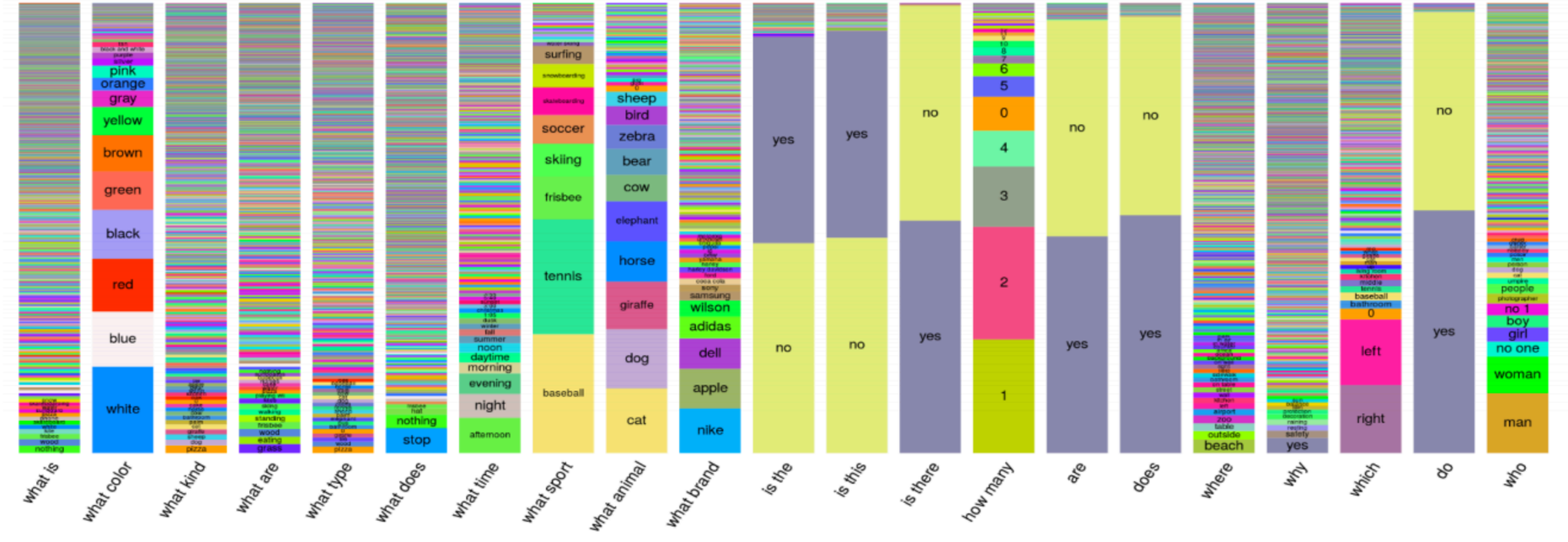


Slide credit: Devi Parikh

Answers from unbalanced dataset



Answers from balanced dataset

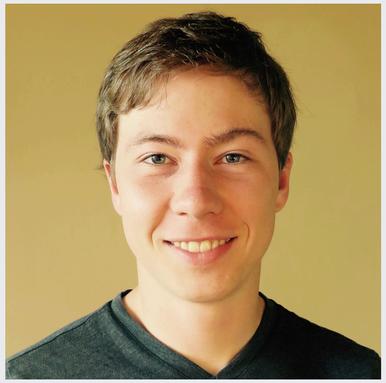


Slide credit: Devi Parikh

What fundamental problems was the dataset actually studying?

Clever Hans, 1907

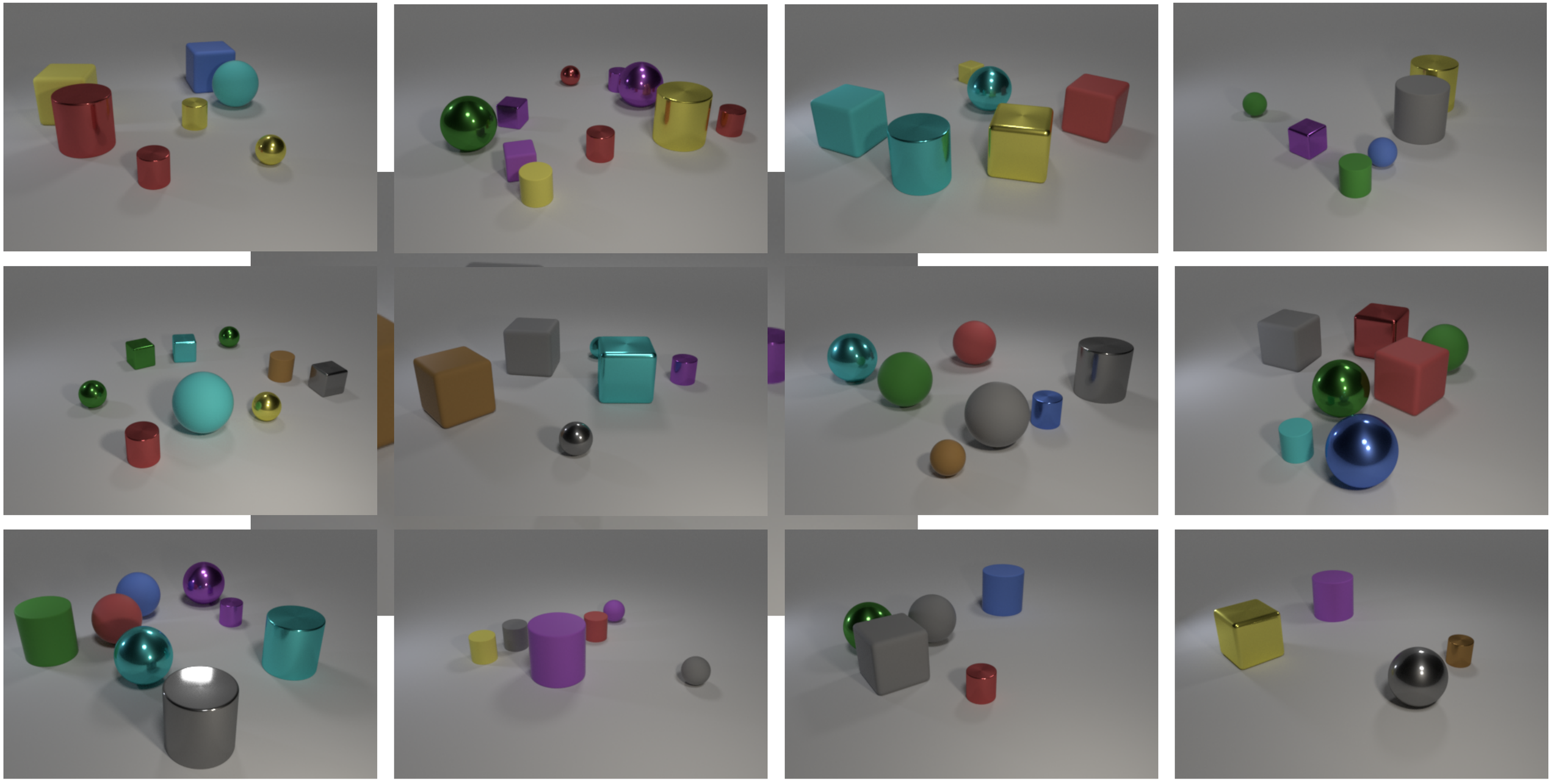
CLEVR: Compositional Language and Elementary Visual Reasoning

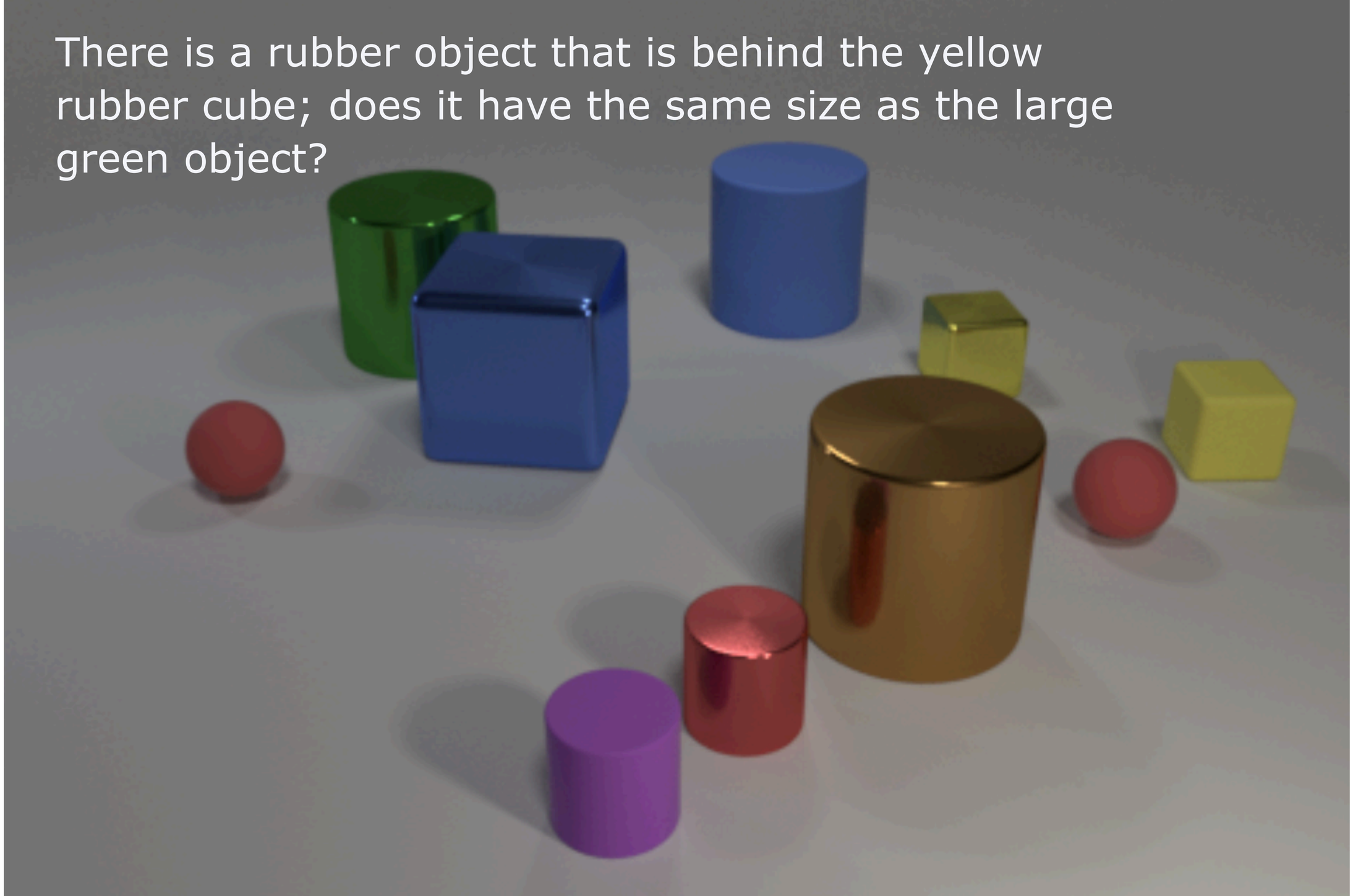


Justin Johnson,
Stanford

What is the man wearing on his face?

Is the plate white?




There is a rubber object that is behind the yellow rubber cube; does it have the same size as the large green object?

Q: Are there an equal number of large things and metal spheres?

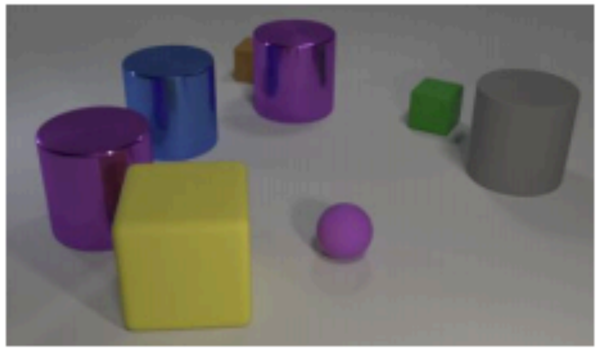
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere?

Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?

Q: How many objects are either small cylinders or metal things?

attribute identification, counting, comparison,
multiple attention, logical operations

Visual Reasoning

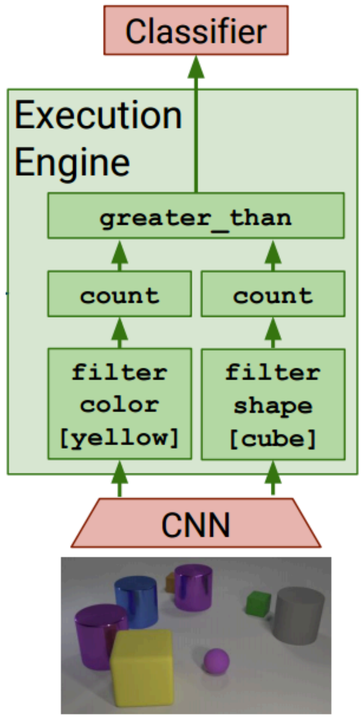


Are there more cubes than yellow things?

1. Predict program

```
greater than

count

filter color [yellow]

<SCENE>

count

filter shape [cube]

<SCENE>
```

2. Execute



Inferring and Executing Programs for Visual Reasoning, Johnson et al., ICCV 2017

Concurrent papers

A simple neural network module for relational reasoning, Santoro et al., arXiv 2017.

Learning Visual Reasoning Without Strong Priors, Perez et al., arXiv 2017.

Learning to Reason: End-to-End Module Networks for Visual Question Answering, Hu et al., arXiv 2017

The approaching challenges:

AI + vision

How do we evaluate AI?

Many “AI” tasks are hard to evaluate

Storytelling

GANs

Image captioning

Problem blindness

Why do we want to recognize chairs?

Intelligent agents must interact with the world.

Plan and reason

Visual Dialog,

Das at al., CVPR 2017

Visual Dialog

A man and a woman are holding umbrellas

What color is his umbrella?

His umbrella is black

What about hers?

Hers is multi-colored

How many other people are in the image?

I think 3. They are occluded

How many are men?

ParlAI

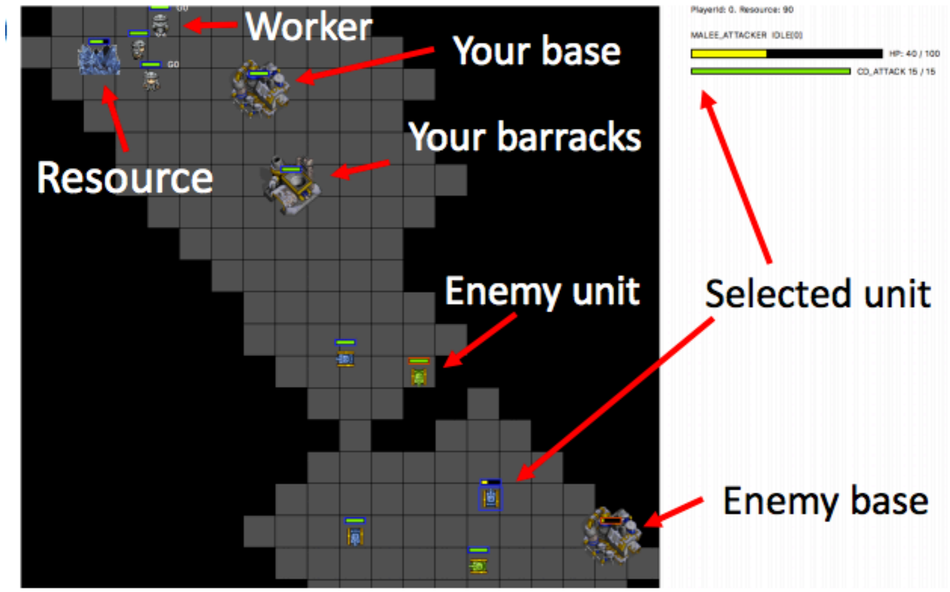
One-stop shop for dialog research

Integration with Mechanical Turk

- data collection
- training
- evaluation

ParlAI: A Dialog Research Software Platform
A. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, J. Weston

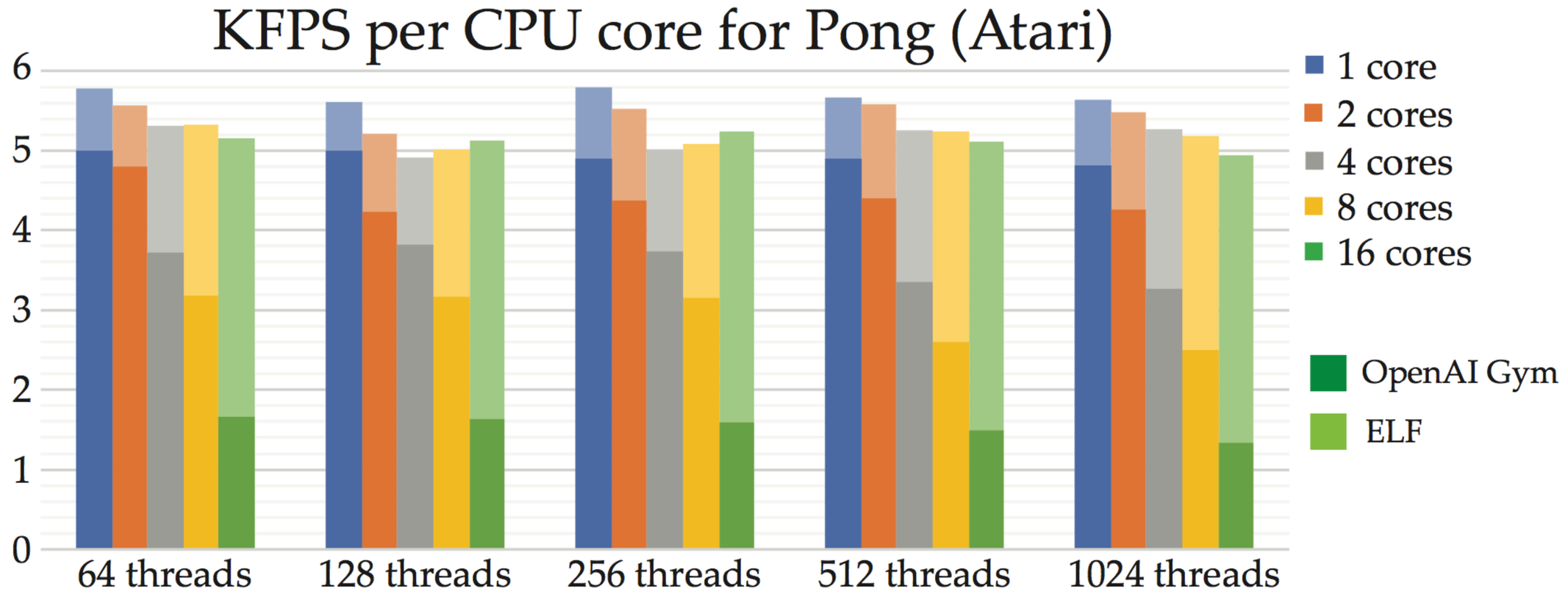
ELF: An Extensive, Lightweight and Flexible Research Platform



Game Name	Descriptions	Avg Game Length
Mini-RTS	Gather resource and build troops to destroy opponent's base.	1000-6000 ticks
Capture the Flag	Capture the flag and bring it to your own base	1000-4000 ticks
Tower Defense	Builds defensive towers to block enemy invasion.	1000-2000 ticks

ELF: An Extensive, Lightweight and Flexible Research Platform for Real-time Strategy Games

Y. Tian, Q. Gong, W. Shang, Y. Wu, L. Zitnick


KFPS per CPU core for Pong (Atari)
6
5
4
3
2
1
0
64 threads
128 threads
256 threads
512 threads
1024 threads
1 core
2 cores
4 cores
8 cores
16 cores
OpenAI Gym
ELF

Question: What color is the car?

Type answer here

SUBMIT

Yi Wu

Yuxin Wu

Georgia Gkioxari

Yuandong Tian

Summary

Baselines are critical

Study problems that can be evaluated

Always ask “why”

Understand what isn’t being studied

Facebook AI Research


Kaiming He


Piotr Dollar


Rob Fergus


Ross Girshick


Bharath Hariharan


Iasonas Kokkinos


Dhruv Batra


Camille Couprie


Yaniv Taigman


Manohar Paluri


Devi Parikh


Natalia Neverova


Marcus Rohrbach


Laurens van der Maaten


Herve Jegou


Yuandong Tian


Lior Wolf


Larry Zitnick

Facebook AI Research

facebook