

Magic of Data: Memory, Data++, Datasets, Bias



CS280: Computer Vision
A. Efros, UC Berkeley, Spring 2024

Humans and Data...

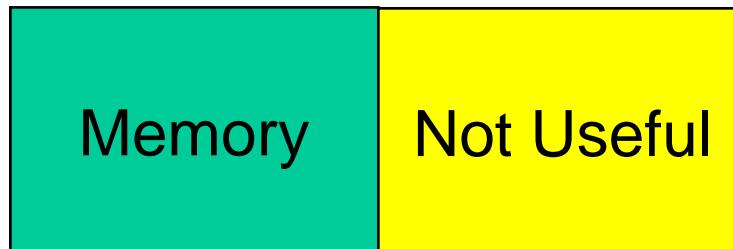
Definition of memory

“what is retained when the stimulus
is no longer present” (Palmer)

Why have memories?

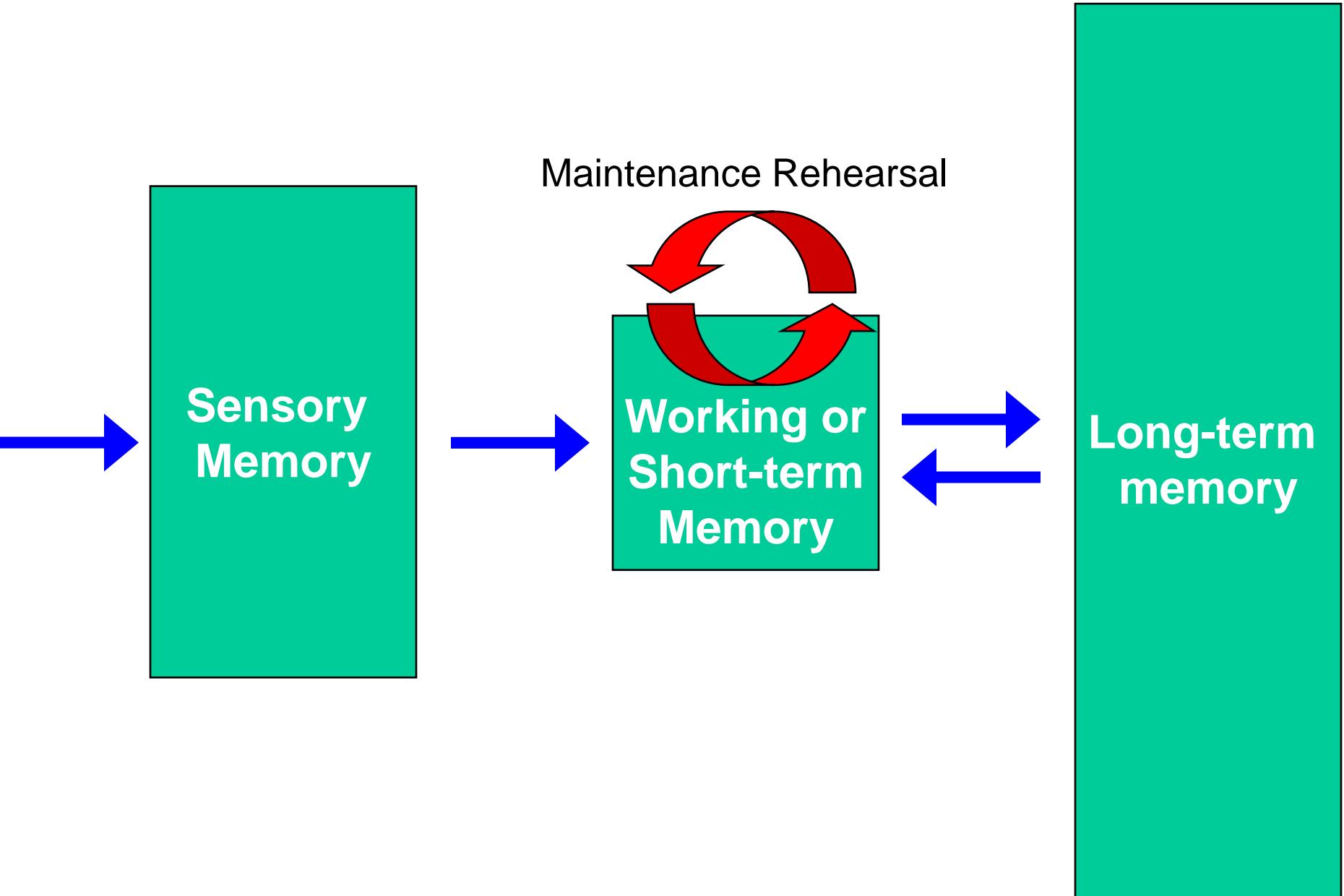
“The past is only relevant to the extent that it helps foresee, that is simulate, various futures.
That is why your “memory” is not a depository.
(...) all that memory is good for—biologically speaking—is to render your future behavior even more efficacious than your past behavior.”

-- Jan Koenderink

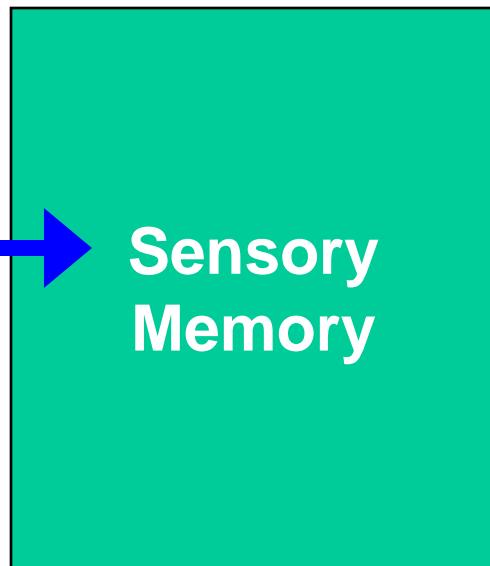


Sensation disentanglement

Three Stages of Memory



Sensory Memory



Function —process for basic physical characteristics

Capacity—large

- can hold many items at once

Duration—very brief retention of images

- .3 sec for visual info
- 2 sec for auditory info

Divided into two types:

- iconic memory—visual information
- echoic memory— auditory information

Attention is needed to transfer information to working memory

Iconic memory capacity

G R C P N S

I E M K W X

Y Z L J B D

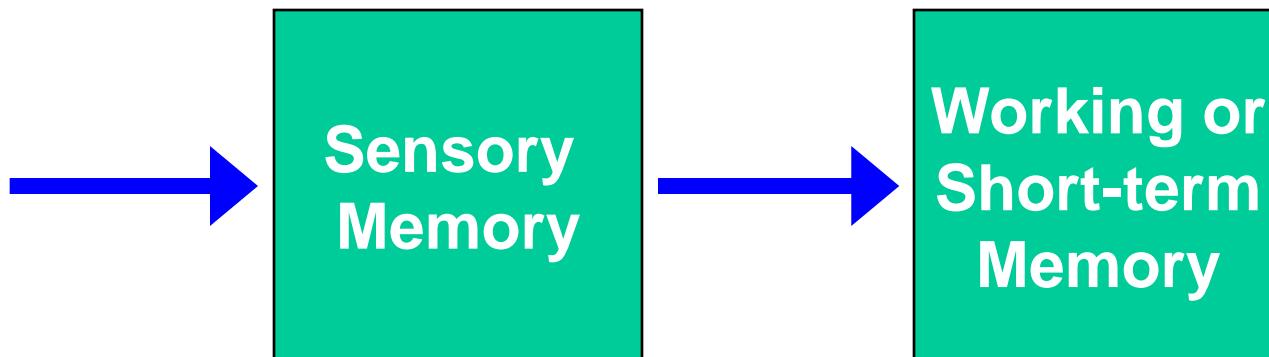
Short-Term Memory

Function—conscious processing of information

- where information is actively worked on

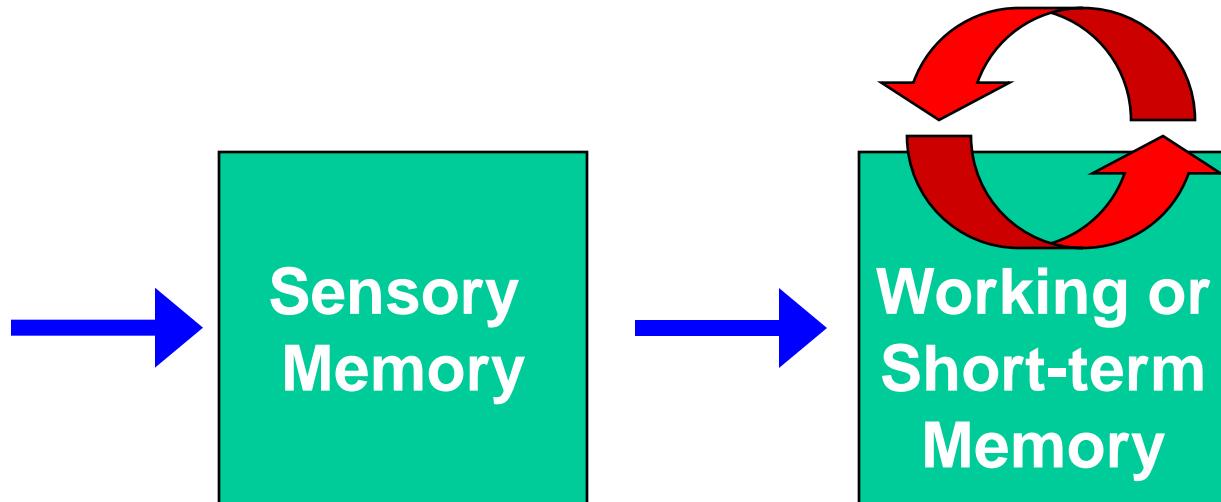
Capacity—limited (holds 7+/-2 items)

Duration—brief storage (about 30 seconds)



Maintenance Rehearsal

Mental or verbal repetition of information allows information to remain in working memory longer than the usual 30 seconds



Chunking

Grouping small bits of information
into larger units of information

- expands working memory load

Which is easier to remember?

- 4 8 3 7 9 2 5 1 6
- 483 792 516

G R C P N S
I E M K W X
Y Z L J B D

C H A R T S
W O N D E R
F L I M S Y

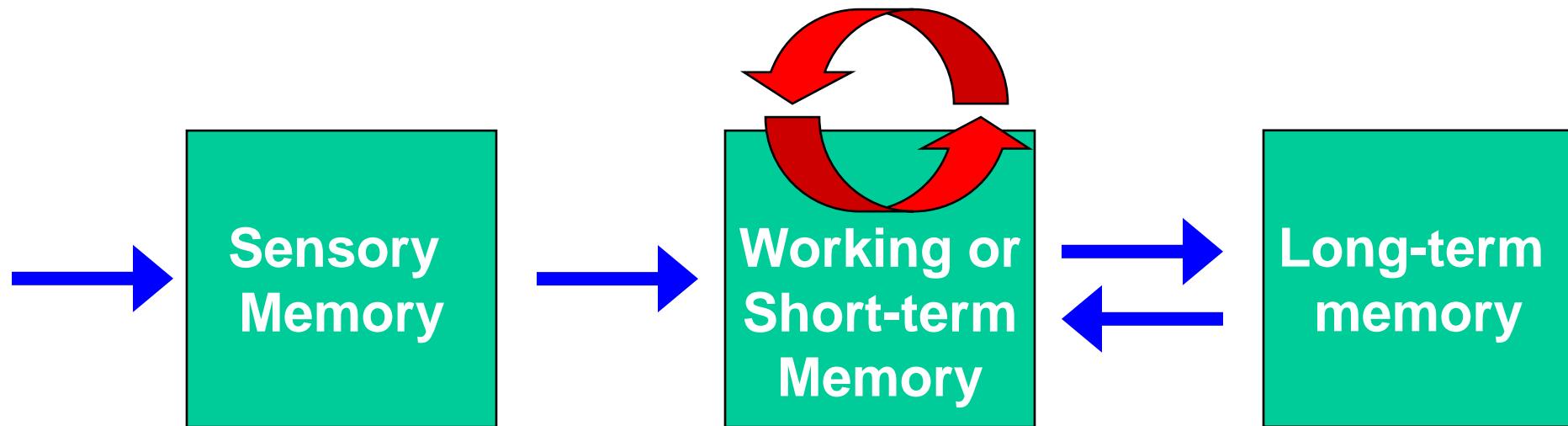
Long-Term Memory

Function—organizes and stores information

- more passive form of storage than working memory

Unlimited capacity

Duration—thought by some to be permanent



Nature of human long term visual memory

What we know in 2008

Standing (1973)

10,000 images

83% Recognition

... people can remember thousands of images

What we don't know in 2008...

... what people are remembering for each item?



According to Standing

"Basically, my recollection is that we just separated the pictures into **distinct thematic categories**: e.g. cars, animals, single-person, 2-people, plants, etc.) Only a few slides were selected which fell into each category, and they were visually distinct."



"Abstract Only"



Sparse Details



Highly Detailed

Massive Memory Experiment I

A stream of objects will be presented on the screen for
~ 3 second each.

Your primary task:

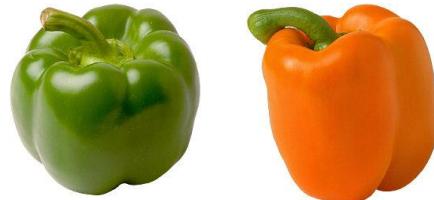
Remember them ALL!

afterwards you will be tested with...

*Completely
different objects...*



*Different exemplars
of the same kind of object...*



*Different states of
the same object...*



Massive Memory Experiment I

Your other task:

Detect exact repeats
anywhere in the stream



Ready?

(Seriously, get ready to clap. The images go by fast...)





<clap!>













<clap!>







10 Minutes Later...









<clap!>





<clap!>





30 Minutes Later...





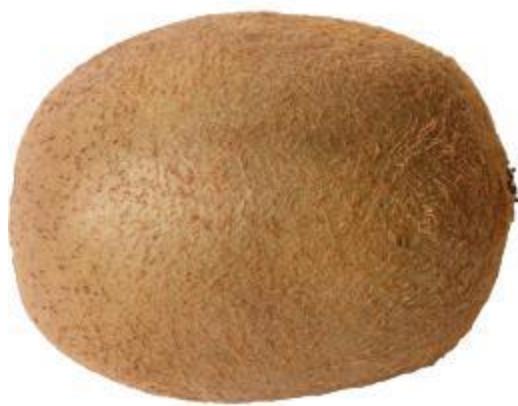








1 Hour Later...











<clap!>





2 Hours Later...





<clap!>











4 Hours Later...











<clap!>





5:30 Hours Later...



Massive Memory I: Methods



Showed 14 observers 2500 categorically unique objects

1 at a time, 3 seconds each

800 ms blank between items

Study session lasted about 5.5 hours

Repeat Detection task to maintain focus

Followed by 300 2-alternative forced choice tests

Which one did you see?

(go ahead and shout out your answer)



-A-



-B-



-A-



-B-

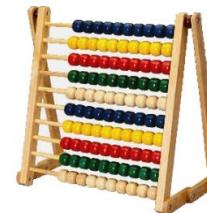


-A-

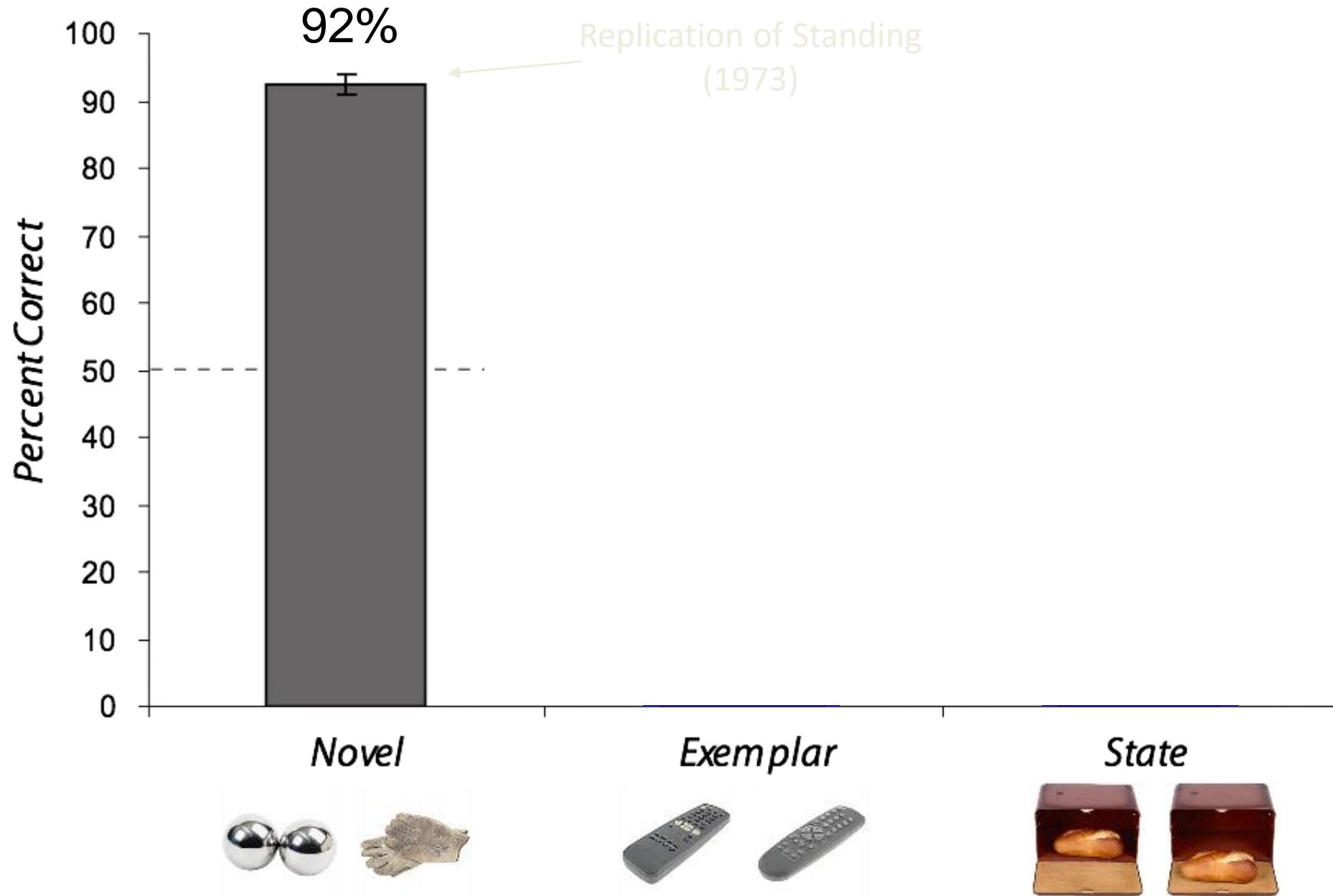


-B-

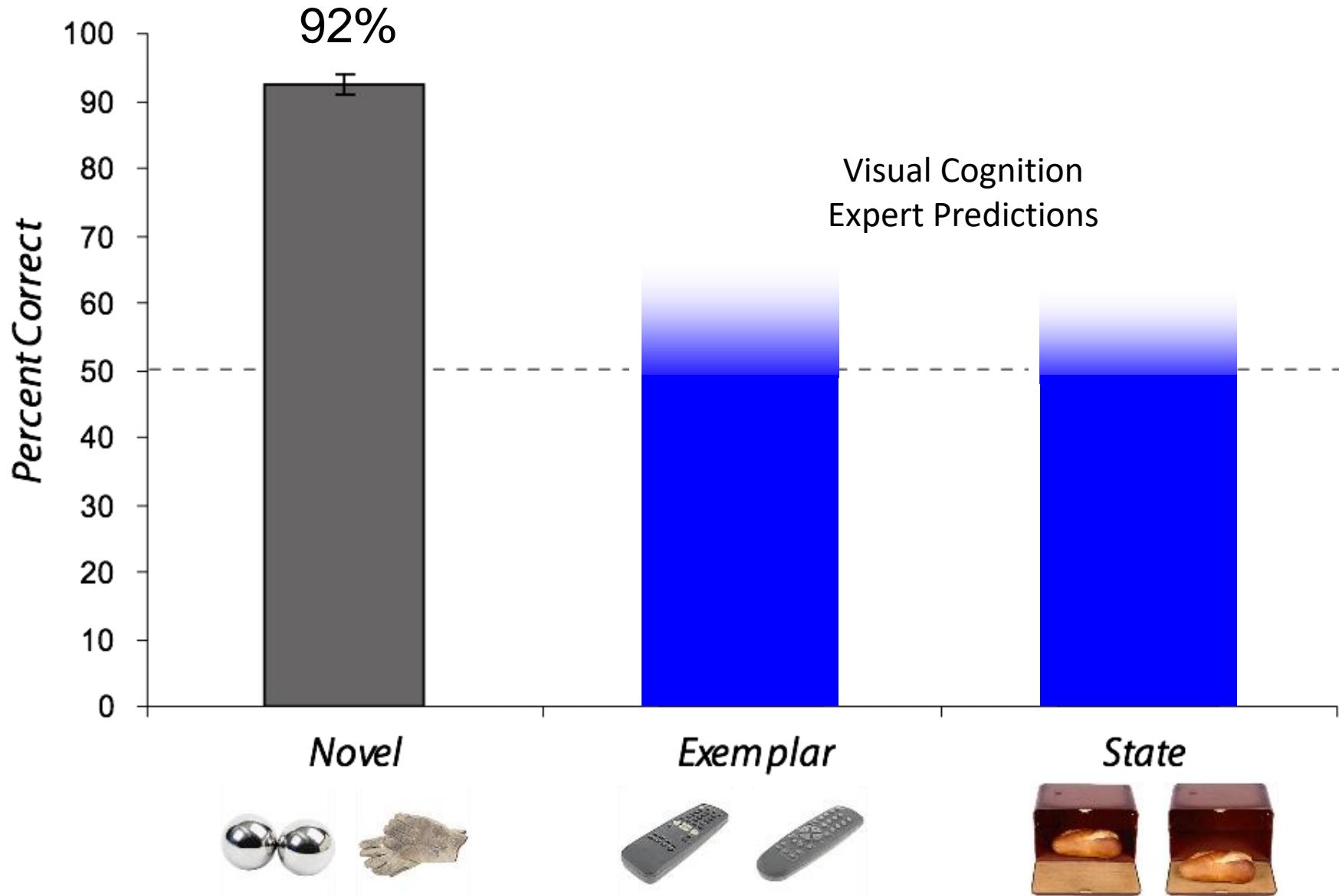
Examples of State memory test



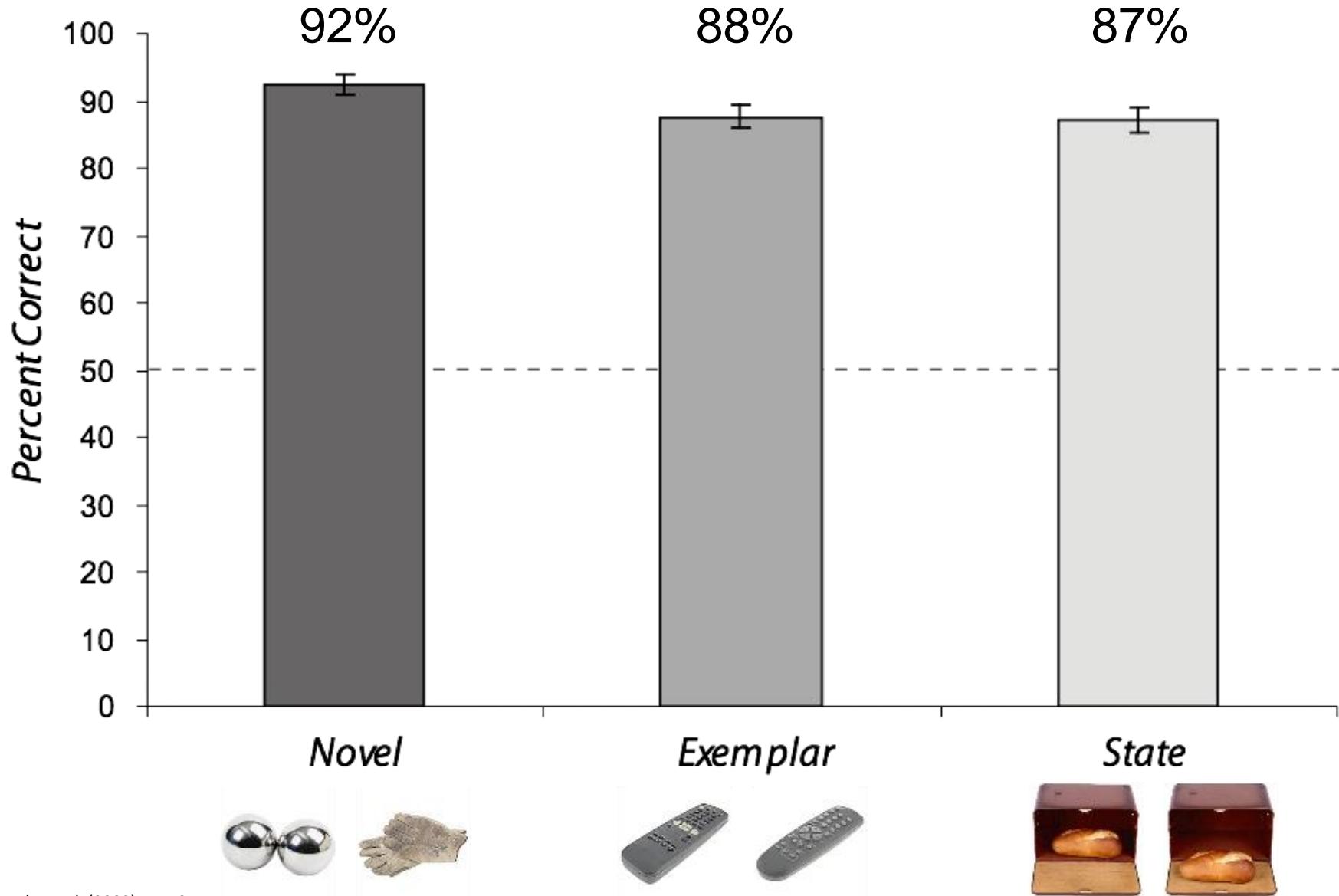
Recognition Memory Results



Recognition Memory Results

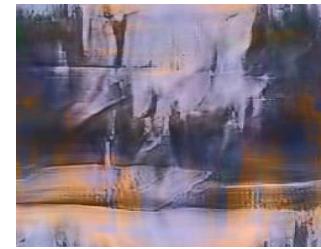
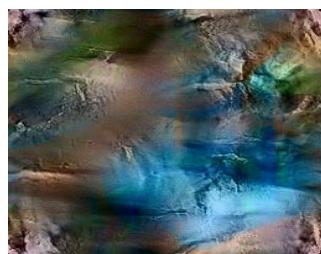


Recognition Memory Results



So, do humans have
“photographic memory”?

“No meaning” textures matching 1st & 2nd order statistics(by Portila & Simmonceli)



Clap your hands when
you see an image repeat

Ready?







<clap!>













<clap!>







<clap!>

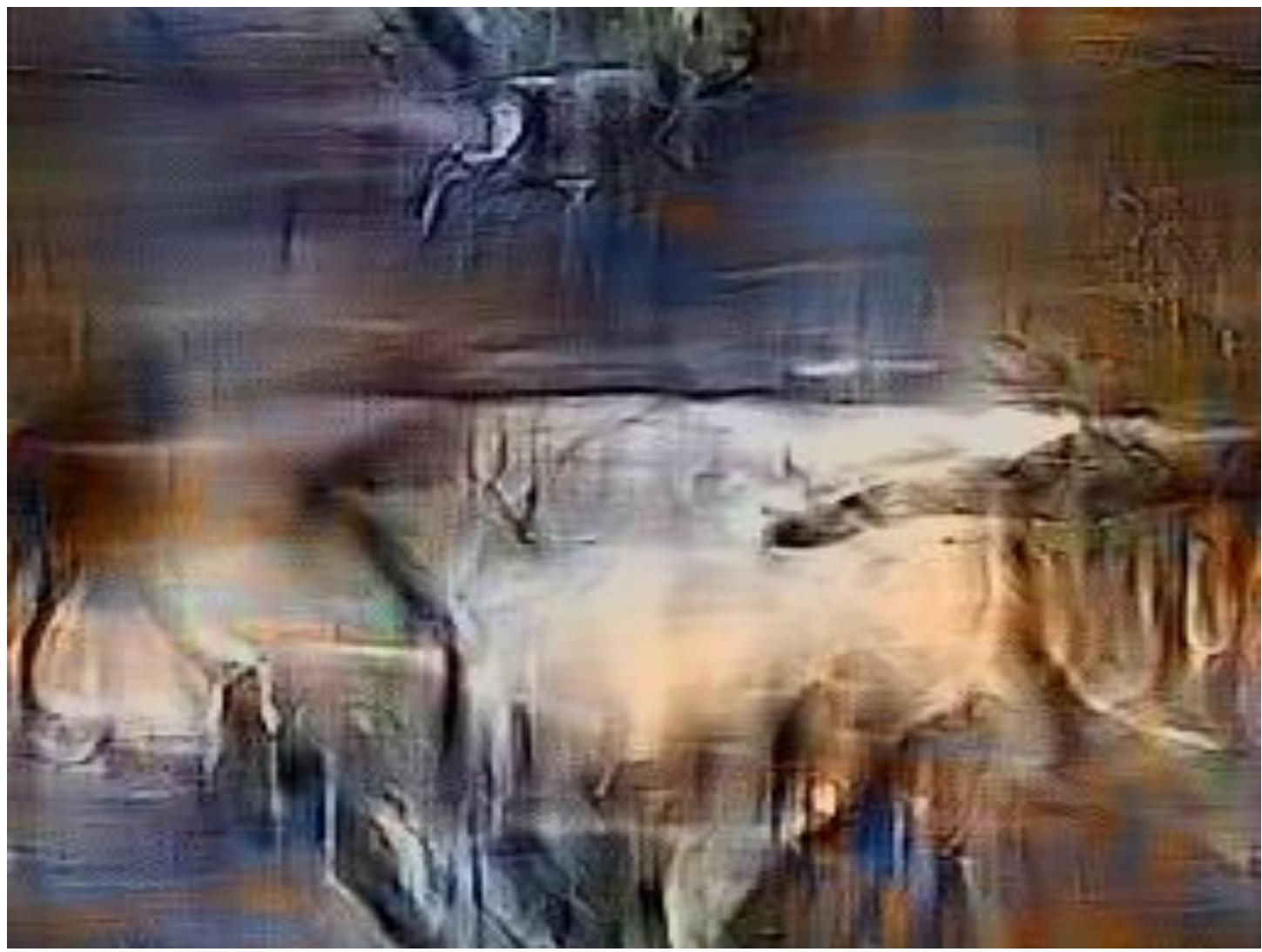




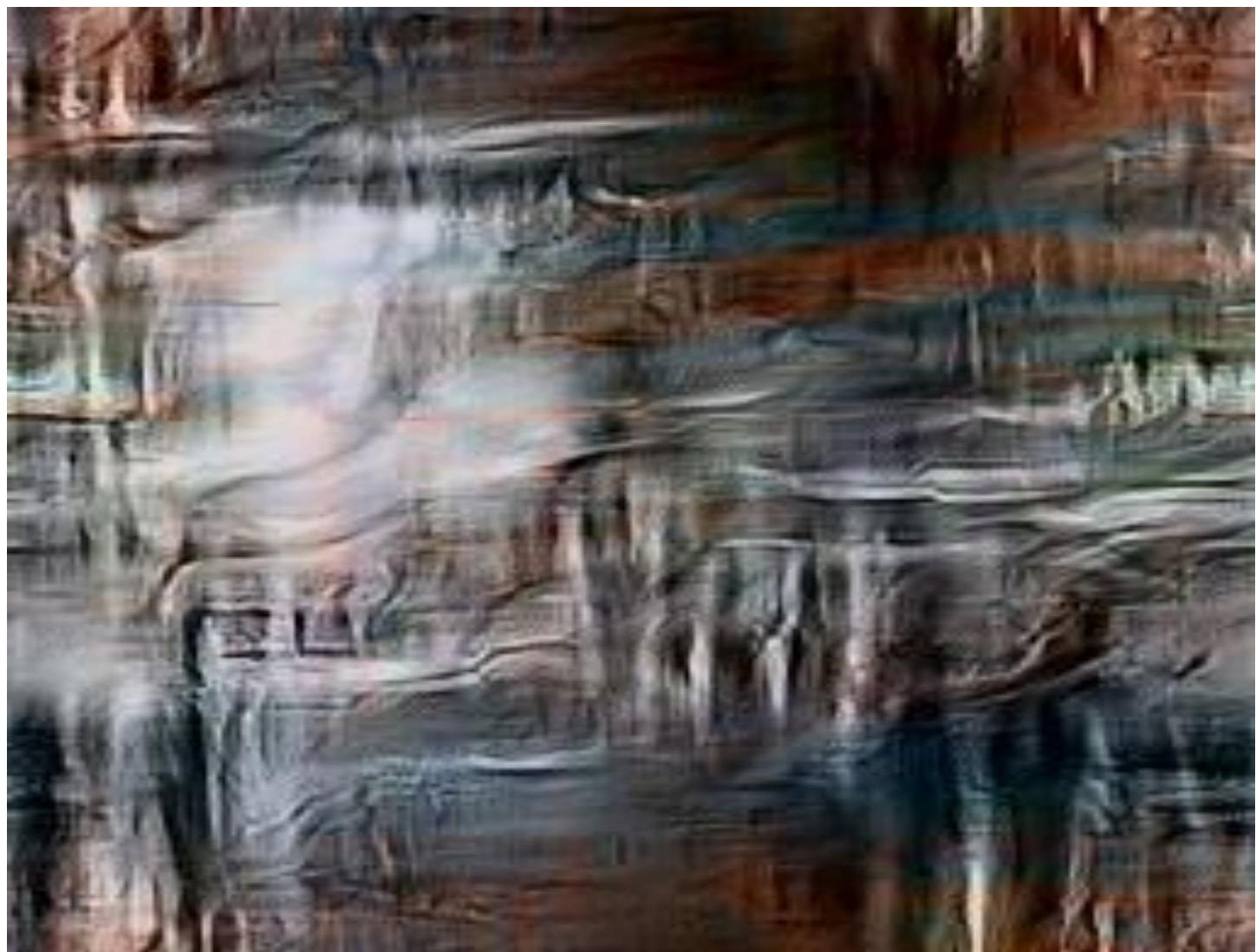




<clap!>



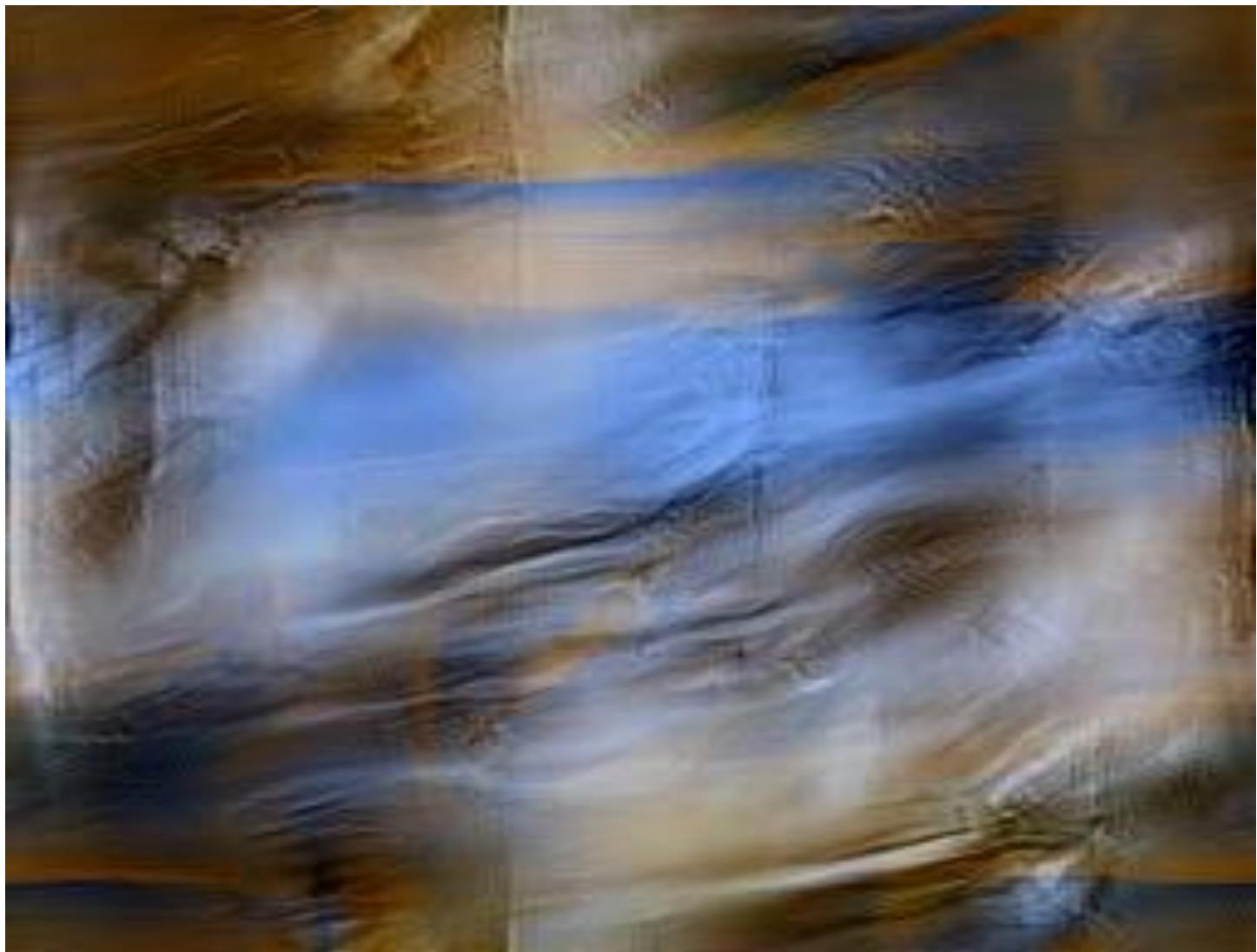


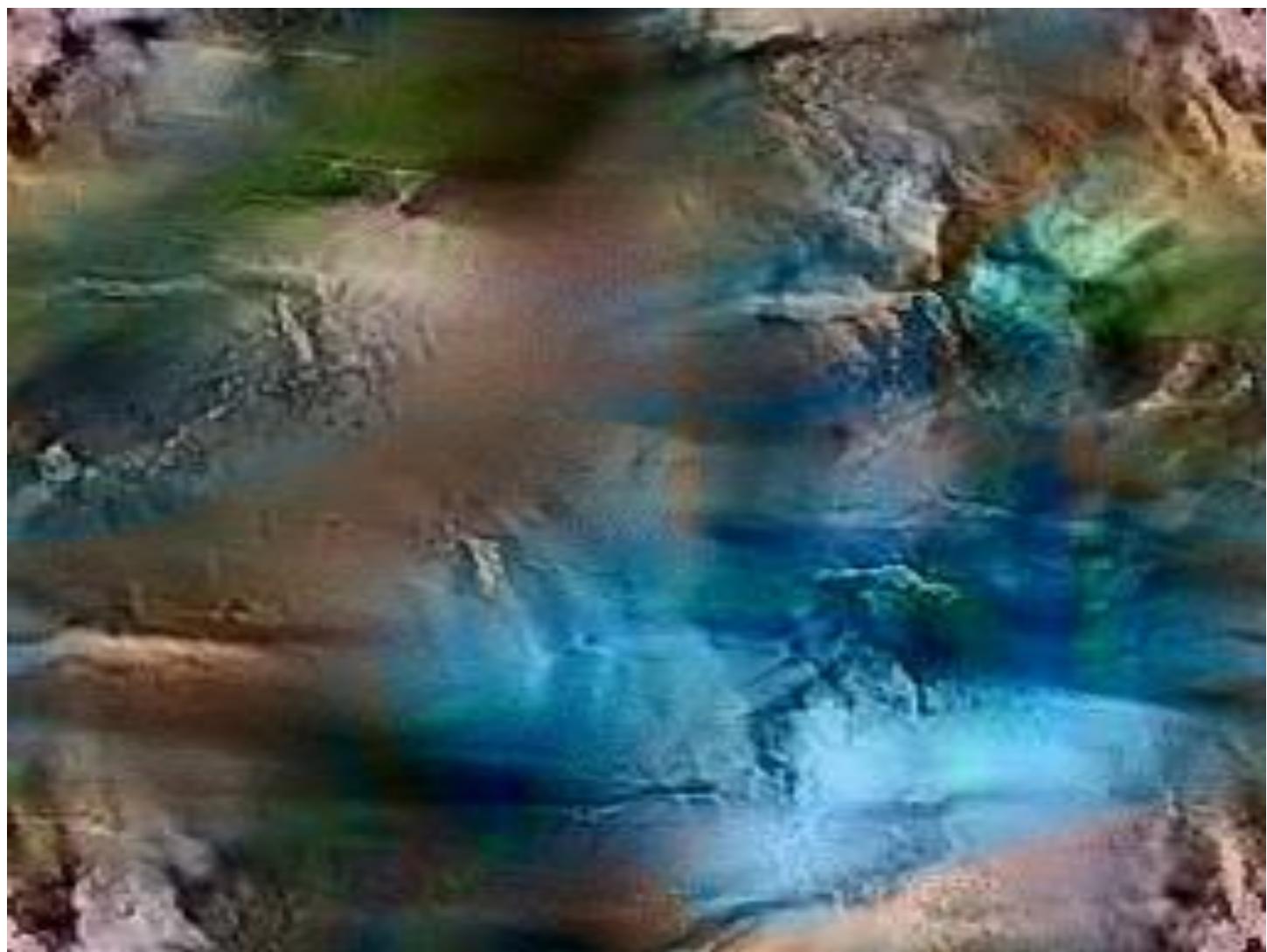






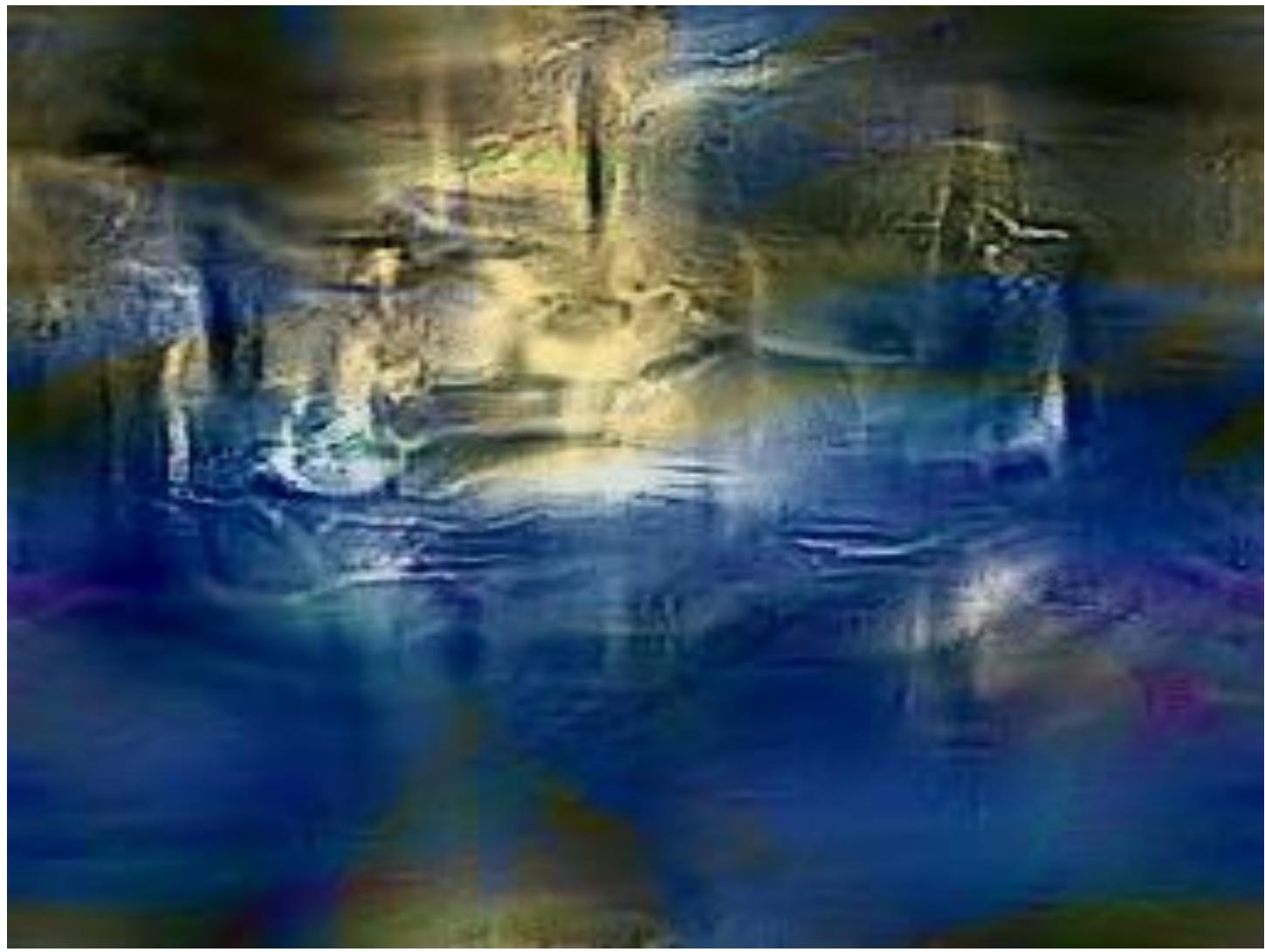
<clap!>







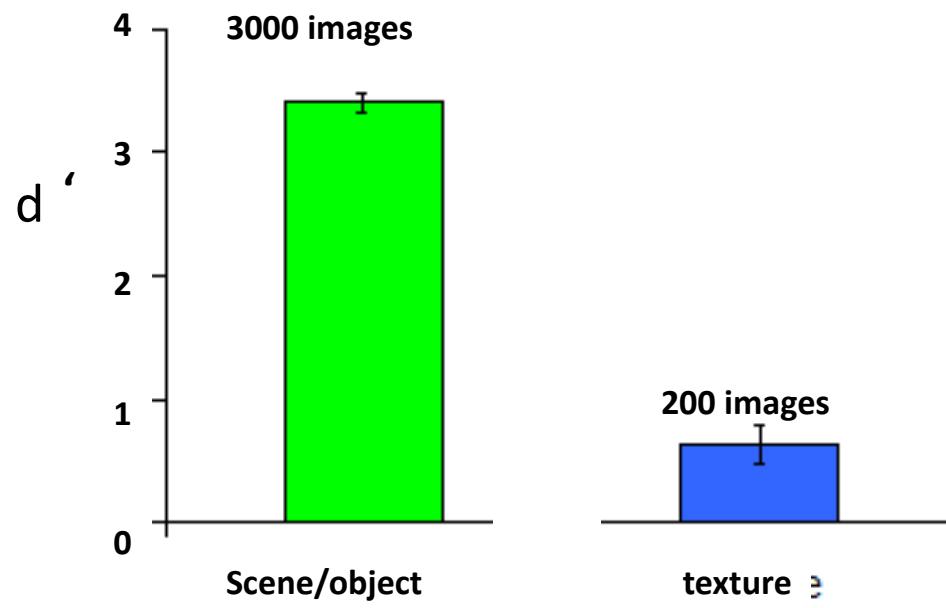
<clap!>





<clap!>





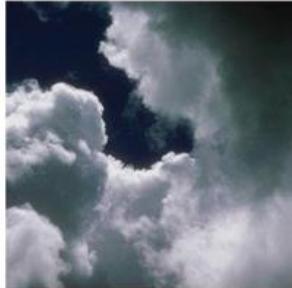
Isola & Oliva,
unpublished



Memorable

Hit rate: 67/70

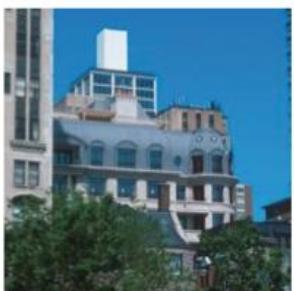
False alarm rate: 4/80



Average

Hit rate: 59/81

False alarm rate: 7/92



Forgettable

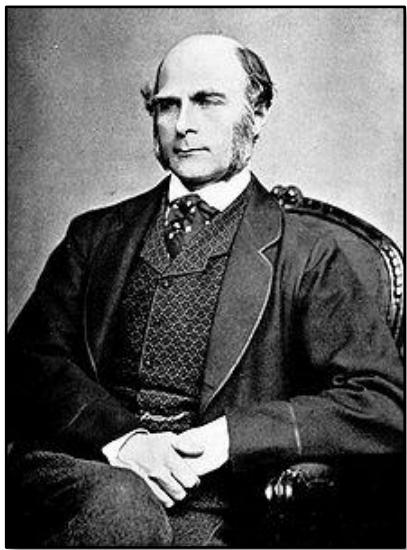
Hit rate: 21/68

False alarm rate: 3/82

Memorability (Isola et al)

Magical Properties of Big Data

Image Averaging



Multiple Individuals



Composite



Sir Francis Galton

1822-1911

[Galton, "Composite Portraits", Nature, 1878]

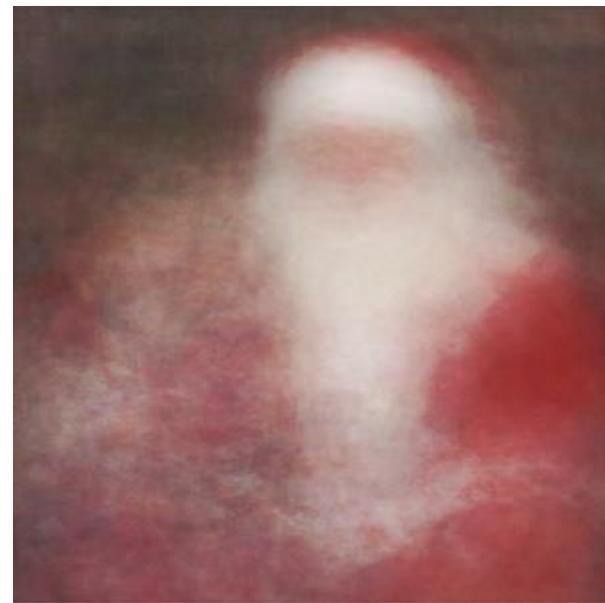
“100 Special Moments” (2004) by Jason Salavon



Newlyweds



Little Leaguer



Kids with Santa

“Object-Centric Averages” (2001) by Antonio Torralba



...

Manual Annotation and Alignment

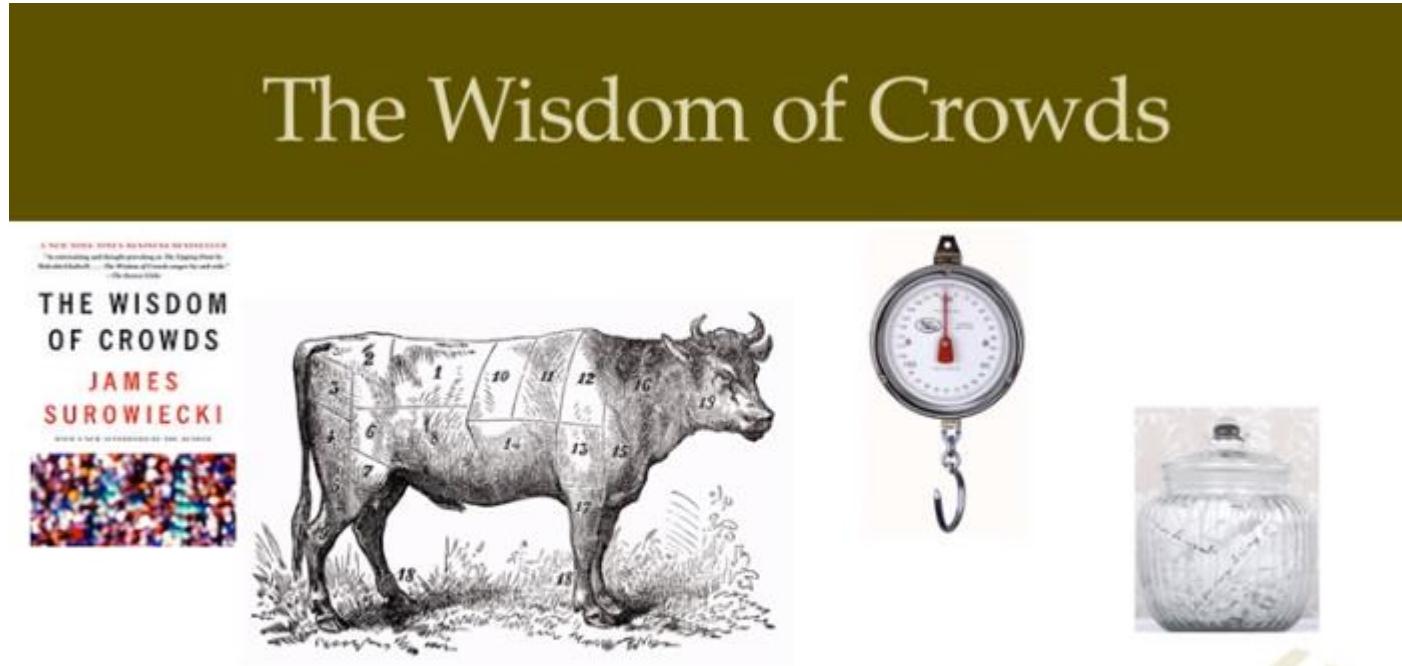


Average Image

“Wisdom of Crowds” (Francis Galton)

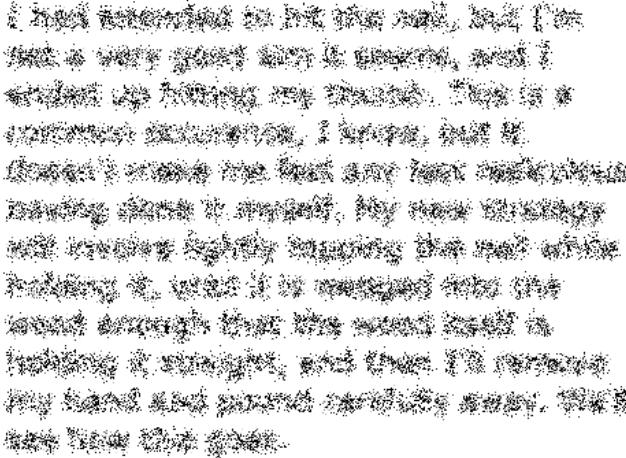
http://en.wikipedia.org/wiki/Wisdom_of_the_crowd

- Many idiots (“weak learners”) are often better than one expert



<http://www.npr.org/2015/08/20/432978431/wighty-issue-cow-guessing-game-helps-to-explain-the-stock-market>

Sequential Wisdom of Crowds



Experiment by Greg Little

Sequential Wisdom of Crowds

- Please transcribe this text.
 - Put words you are unsure about in **(parentheses)**.
 - Use (?) for words you don't know.

non-iterative

- ???
??
????. ~1287 seconds
 - - unreadable - ~60 seconds
 - (unreadable) ~59 seconds
 - ? ~74 seconds
 - words are blurred ~17 seconds
 - ***Experiment stopped here, still waiting on 3 responses.***

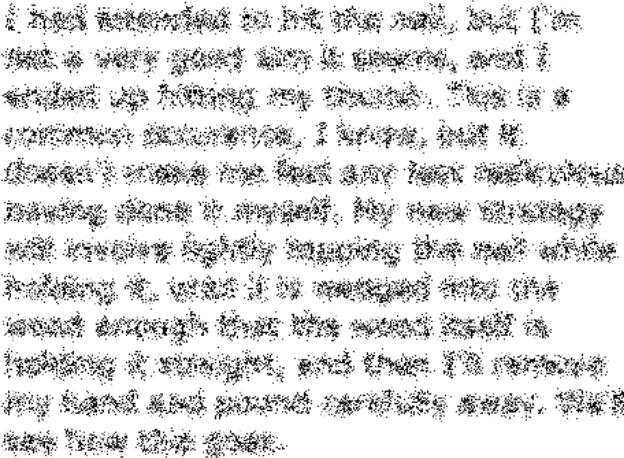
As a result, the number of people who have been infected with the virus has increased rapidly, leading to a significant increase in the number of deaths. The World Health Organization (WHO) has declared the situation a global emergency, and governments around the world are taking steps to contain the spread of the virus.

character count: 0/500 please type something above

Submit

Experiment by Greg Little

Sequential Wisdom of Crowds



- Please transcribe this text.
- Put words you are unsure about in **(parentheses)**.
- Use (?) for words you don't know.

character count: 0/500 please type something above

iterative

- I had attended to the nail, but I'm not a very good ~ 431 seconds
 - I had (resorted) to (hit) (the) (wall), (but) I'm (not) (a) (very) (good) (?) (is) (?), (?) (I) (?) (up) (?) (my) (?). This is a (?) (?), I (know), but (it) (doesn't) (mean) (the) (?) (sure) (?) (?) (?) (it) (?). My new (strategy) (will) (involve) (lightly) (?) then (?) (while) (?) (it) (is) (?) (?) (this) (wasn't) (enough) (?) (?) (?) (is) (?) (?) (straight), and then I'll (?) my hand and (?) (?) (?) (away). We'll (?) (?) (?) (?). ~ 1349 seconds
 - I (had) (resorted) to (hit) (the) (wall), (but) I'm (not) (a) (very) (good) (quitter). (I) (gave) (up) (on) (my) (ability). This is a (shame), I (know), but (it) (doesn't) (mean) (the) (end) (surely) (will) (come) (anytime) (soon). My new (strategy) (will) (involve) (lightly) (thinking) then (doing) (while) (trying) (to) (make) (it) (happen). (This) (wasn't) (supposed) (to) (happen) (this) (way). ~ 353 seconds
 - I had intended to hit the nail, but I'm not a very good (aim) it seems, and I ended up hitting my thumb. This is a common occurrence, I know, but it doesn't (make) me feel any (less) (?) having done it myself. My new strategy will involve lightly tapping the nail while holding it, until it is (?) into the (?) (enough) that the (?) itself is holding it straight, and then I'll remove my hand and pound (carefully) away. We'll see how (this) goes. ~ 645 seconds
 - I had intended to hit the nail, but I'm not (a) very good (aim) in it (seems), and I ended up hitting my thumb. This is a common occurrence, I know, but it doesn't make me feel (any) (less) (?) having done it myself. My new strategy will involve lightly tapping the nail (while holding it), until it is (?) into the (?) (enough) that the (?) itself is holding it straight, and then I'll remove my hand and pound (carefully) away. We'll see how (this) goes. ~ 318 seconds
 - I had intended to hit the nail, but I'm not a very good (aim) it seems, and I ended up hitting my thumb. This is a common occurrence, I know, but it doesn't (make) me feel any (less) (?) having done it myself. My new strategy will involve lightly tapping the nail while holding it, until it is (?) into the (?) (enough) that the (?) itself is holding it straight, and then I'll remove my hand and pound (carefully) away. We'll see how (this) goes. ~ 103 seconds
 - I had intended to hit the nail, but I'm not a very good aim it seems, and I ended up hitting my thumb. This is a common occurrence, I know, but it doesn't make me feel any less ridiculous having done it myself. My new strategy will involve lightly tapping the nail while holding it, until it is embedded into the wood enough that the wood itself is holding it straight, and then I'll remove my hand and pound carefully away. We'll see how this goes. ~ 159 seconds

Experiment by Greg Little

Sequential Wisdom of Crowds

Impossible HITS and iterations

I had intended to hit the nail, but I'm not a very good aim it seems, and I ended up hitting my thumb. This is a common occurance, I know, but it doesn't make me feel any less rediculous having done it myself. My new strategy will involve lightly tapping the nail while holding it, until it is **wedged** into the wood enough that the wood itself is holding it straight, and then I'll remove my hand and pound carefully away. We'll see how this goes.

original:

I had intended to hit the nail, but I'm not a very good aim it seems, and I ended up hitting my thumb. This is a common occurance, I know, but it doesn't make me feel any less rediculous having done it myself. My new strategy will involve lightly tapping the nail while holding it, until it is **wedged** into the wood enough that the wood itself is holding it straight, and then I'll remove my hand and pound carefully away. We'll see how this goes.

Iterative:

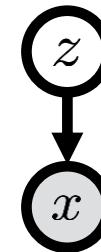
I had intended to hit the nail, but I'm not a very good aim it seems, and I ended up hitting my thumb. This is a common occurance, I know, but it doesn't make me feel any less rediculous having done it myself. My new strategy will involve lightly tapping the nail while holding it, until it is **embedded** into the wood enough that the wood itself is holding it straight, and then I'll remove my hand and pound carefully away. We'll see how this goes.

Experiment by Greg Little

Data++ (Isola et al)



Data



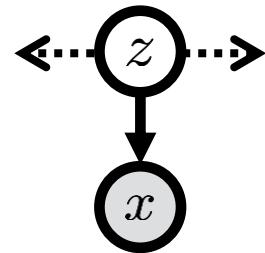
Data++



Data++ (Isola et al)



Data



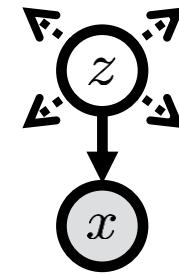
Data++



Data++ (Isola et al)

x

Data



Data++



But you still need the actual data!

Correspondence

<https://doi.org/10.1038/s41587-023-02103-0>

The perpetual motion machine of AI-generated data and the distraction of ChatGPT as a ‘scientist’

 Check for updates

 is a longtime researcher at the

https://www.nature.com/articles/s41587-023-02103-0.epdf?sharing_token=WEJJTtsNXob_tkzkzhzIdRgN0jAjWeI9jnR3ZoTv0MJp-NuyZtJcJihsrQSHmrTt-9BTduYs8MW0Riih8dxaKY55tWAeaCsG8IkB78_o8x1m64LcS5vuaTMmTSFlwUS6SOwKgXultunlgVHG9AedfmEWJzZnw3LXeysNG_WCC8%3D

Datasets and their discontents

Back in the *Pre-Dataset Era*

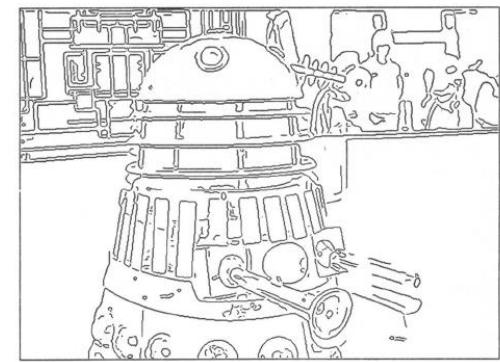
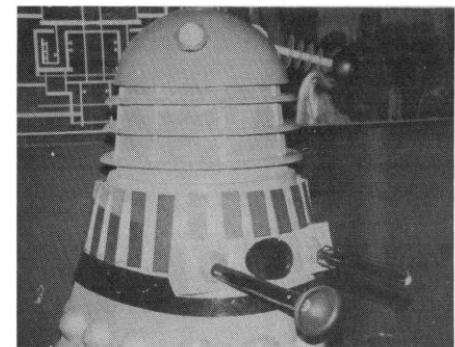
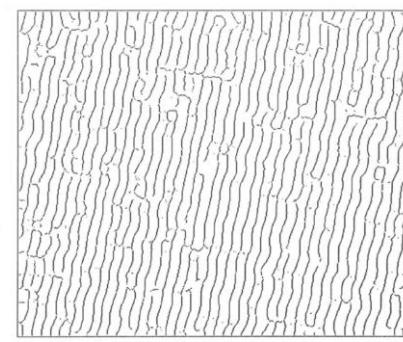
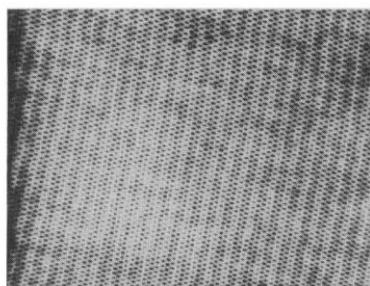
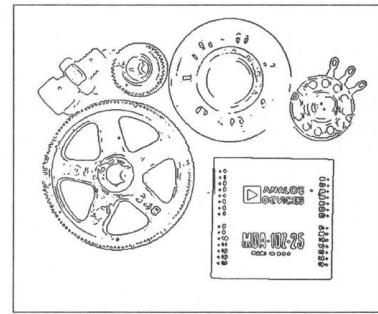
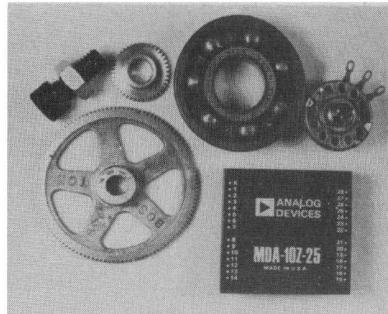


- I started my PhD in Computer Vision in 1998
- It was a completely different world
 - No numbers, just images
- e.g. CVPR joke from the 1990s:
 - *If paper has just one result image, it's a **hack***
 - *If it has 2 images, it's an **algorithm***
 - *If it has 3 images, it's a **framework!***

A Computational Approach to Edge Detection

JOHN CANNY, MEMBER, IEEE

35K citations



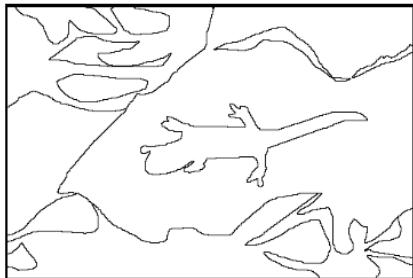
A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics

ICCV 2001

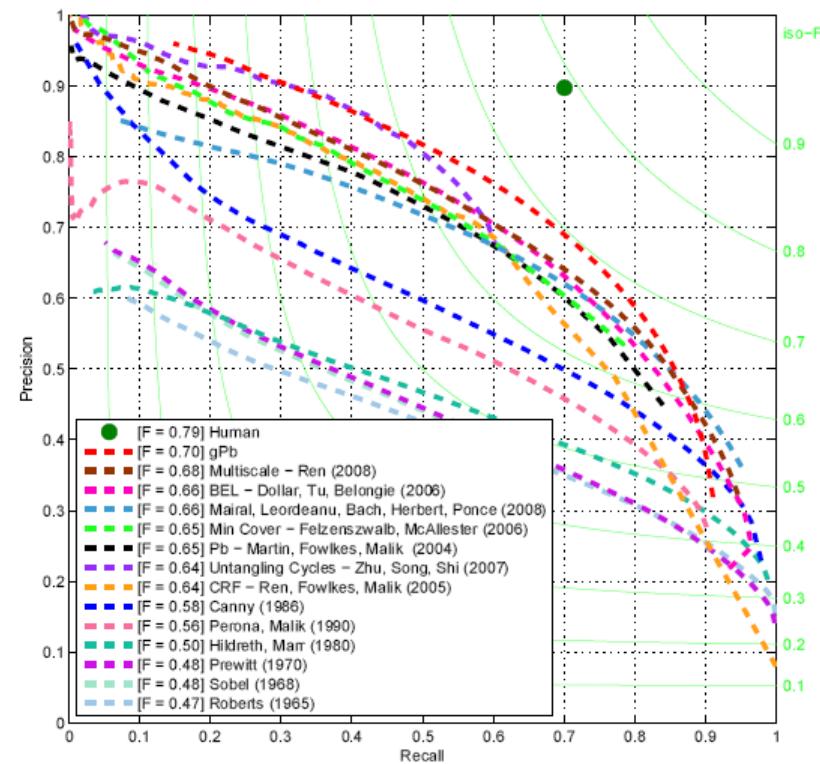
David Martin Charless Fowlkes
Department of Electrical Engineering
University of California
Berkeley, CA
{dmartin,fowlkes,doron,malil}



Original Image



Subject 1



The Dataset Era (2000-present)

"When you can measure what you are speaking about and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of the meager and unsatisfactory kind."

Lord Kelvin

- Dataset-driven Breakthroughs:
 - SIFT (thanks to Mikolajczyk & Schmid, CVPR'03)
 - DPMs (thanks to PASCAL VOC)
 - ConvNets (thanks to ImageNet and PASCAL VOC)
- At first, datasets were mainly for evaluation

Excesses of the “Data Revolution”

Goodhart's law:

"When a measure becomes a target, it ceases to be a good measure"

- Are we getting too obsessed with evaluation?
 - The dictatorship of the PR curve over the pixels...
 - Hard to jump out of algorithmic local minima
 - Too much value for “winning” a challenge
 - Easy to overfit over time
- These are all behavioral problems
 - Can be fixed with proper “Best Practices”

Are datasets measuring the right thing?

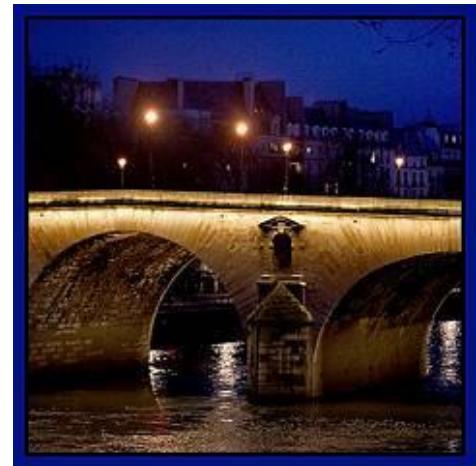
- In Machine Learning:
Dataset is The World
- In Recognition
Dataset is a *representation* of The World
- Do datasets provide a good representation?

Visual Data is Inherently Biased

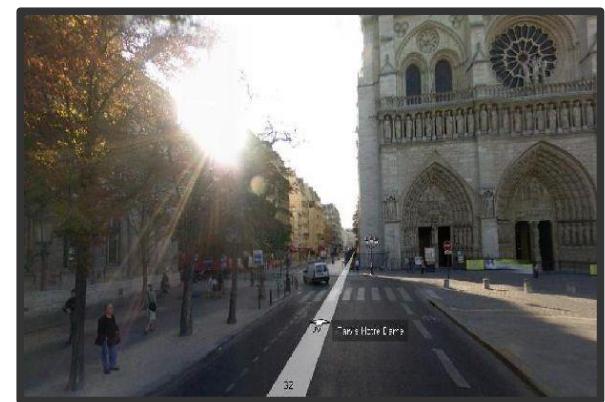
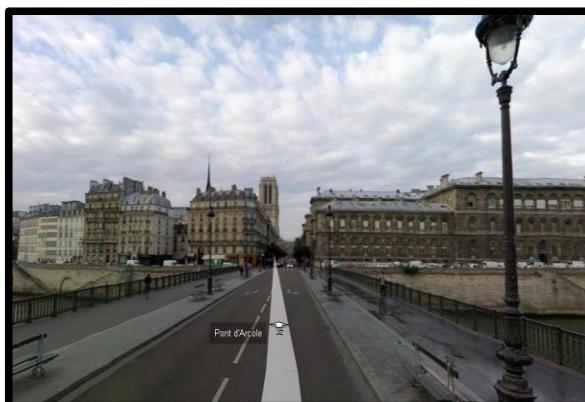
- Internet is a tremendous repository of visual data (Flickr, YouTube, Picassa, etc)
- But it's not random samples of visual world
 - And is “random” the right thing anyway?



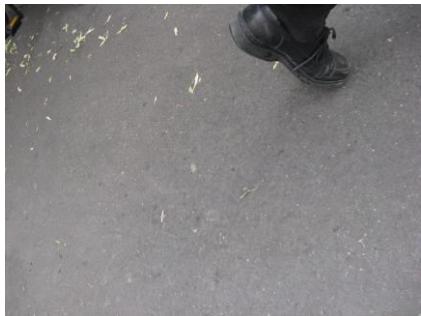
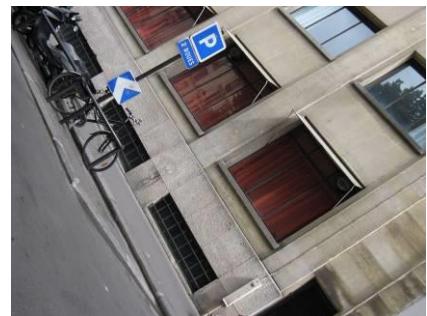
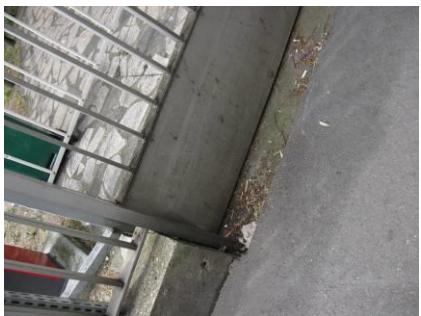
Flickr Paris



Google StreetView Paris



“Uniformly Sampled” Paris



Most “collection bias” is subconscious

mug

About 10,100,000 results (0.09 seconds)

Search SafeSearch moderate ▾ Advanced search

59¢ Logo Coffee Mugs
[www.DiscountMugs.com](#) Lead Free & Dishwasher Safe. Save 40-50%. No Catch. Factory Direct!

Custom Mugs On Sale
[www.Vistaprint.com](#) Order Now & Save 50% On Custom Mugs No Minimums. Upload Photos & Logos.

Promotional Mugs from 69¢
[www.4imprint.com/Mugs](#) Huge Selection of Styles Colors- Buy 72 Mugs @ \$1.35 ea-24hr Service Sponsored

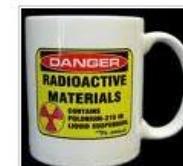
Related searches: [white mug](#) [coffee mug](#) [mug root beer](#) [mug shot](#)


Representational
500 × 429 - 91k - jpg
[eagereyes.org](#)
Find similar images


Ceramic Happy Face
300 × 300 - 77k - jpg
[larose.com](#)
Find similar images


Here I go then, trying
600 × 600 - 35k - jpg
[beeper.wordpress.com](#)
Find similar images


The Chalk Mug »
304 × 314 - 17k - jpg
[coolest-gadgets.com](#)
Find similar images


mug
300 × 279 - 54k - jpg
[reynosawatch.org](#)
Find similar images


Bring your own
500 × 451 - 15k - jpg
[cookstownunited.ca](#)
Find similar images


ceramic mug
980 × 1024 - 30k - jpg
[diytrade.com](#)


Dual Purpose Drinking
490 × 428 - 16k - jpg
[freshome.com](#)
Find similar images


This coffee mug,
300 × 300 - 22k - jpg
[gizmodo.com](#)
Find similar images


Back to Ceramic
400 × 400 - 11k - jpg
[freshpromotions.com.au](#)
Find similar images


Back to Ceramic
400 × 400 - 8k - jpg
[freshpromotions.com.au](#)
Find similar images


Coffee Mug as a
303 × 301 - 10k - jpg
[dustbowl.wordpress.com](#)
Find similar images


SASS Life Member
300 × 302 - 6k - jpg
[sassnet.com](#)


personalized coffee
400 × 343 - 15k - jpg
[wallyou.com](#)
Find similar images


We like our mugs
290 × 290 - 6k - jpg
[kitchencontraptions.com](#)
Find similar images

Sampling Bias

- People like to take pictures on vacation



Photographer Bias

- People want their pictures to be recognizable and/or interesting



vs.



Social Bias



Little Leaguer



Kids with Santa



The Graduate



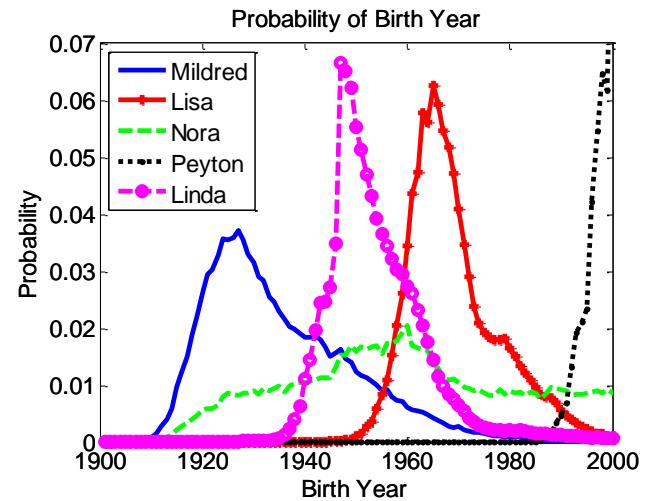
Newlyweds

“100 Special Moments” by Jason Salavon

Social Bias



Mildred and Lisa



Source: U.S. Social Security Administration

How much does this bias affect standard recognition datasets?

Unbiased Look at Dataset Bias

Alyosha Efros (UC Berkeley)

Antonio Torralba (MIT)



Disclaimer: no graduate students have been harmed in the production of this work

“Name That Dataset!” game



Caltech 101

Caltech 256

MSRC

UIUC cars

Tiny Images

Corel

PASCAL 2007

LabelMe

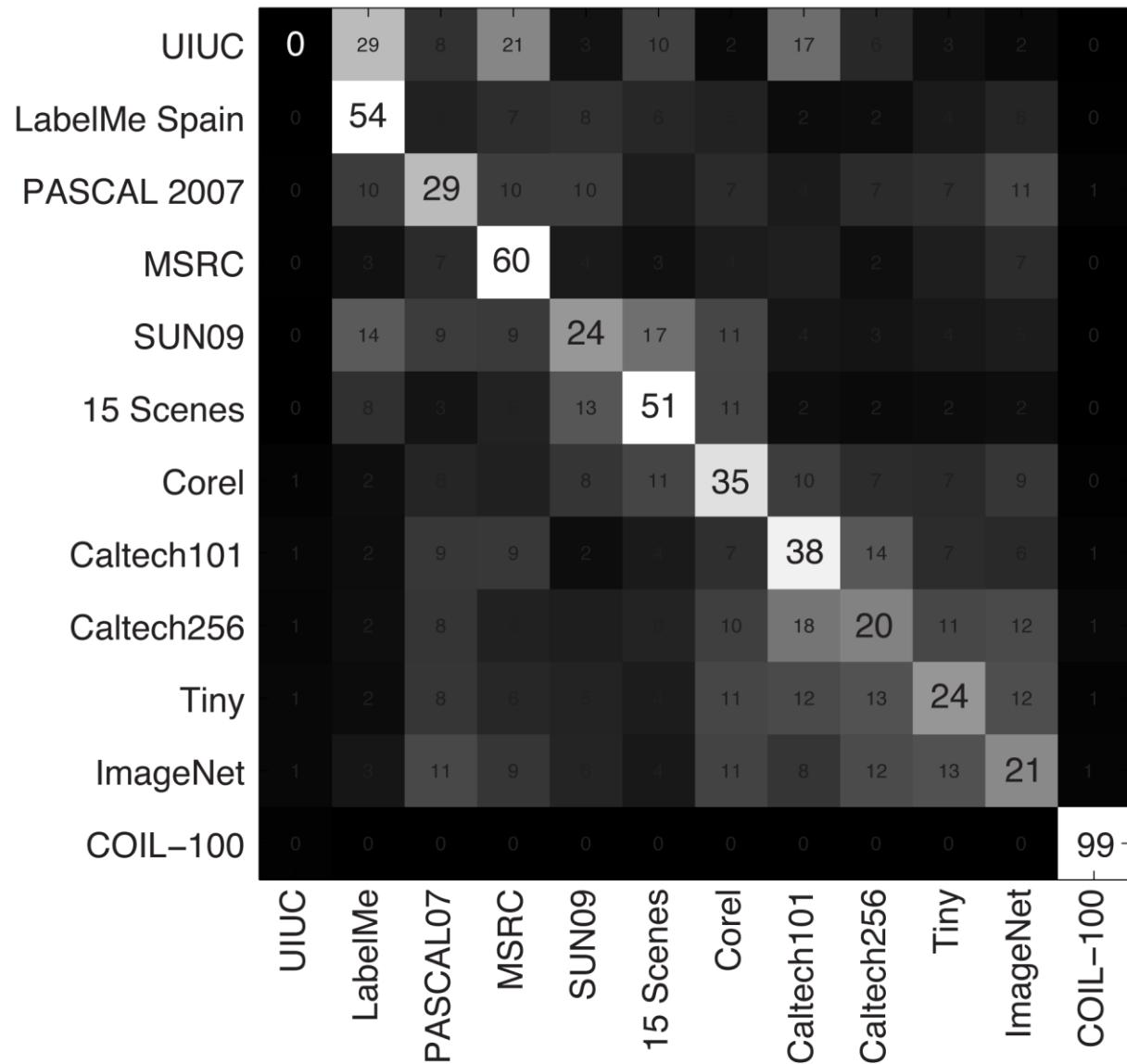
COIL-100

ImageNet

15 Scenes

SUN'09

SVM plays “*Name that dataset!*”



- 12 1-vs-all classifiers
- Standard full-image features
- 39% performance (chance is 8%)

Poor Cross-Dataset Generalization

MSRC



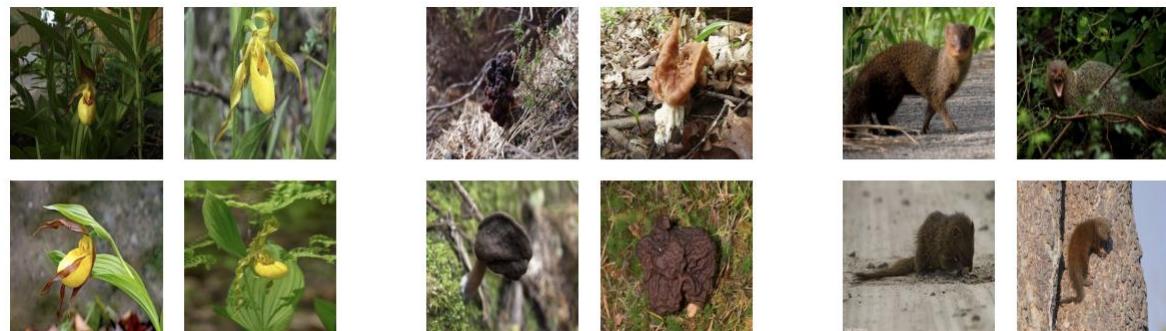
Classifier trained on MSRC cars

ImageNet vs. ImageNetV2

ImageNet



ImageNetV2



Comparing Datasets

In 2019, a decade after ImageNet was collected, Recht et al. introduced ImageNetV2 [1], which attempted to mirror the original ImageNet collection process, including restricting data to images uploaded in a similar timeframe.

However, models trained on ImageNet showed a consistent 11-14% accuracy drop on ImageNetV2, and the reasons for this have remained unclear.

While some studies have employed statistical tools to reveal a distributional difference between ImageNet and ImageNetV2 [10], we aim to discover more interpretable differences between these two datasets.

**Dataset Bias is here to stay.
How do we live with it?**

Describing Differences in Image Sets with Natural Language

“cool, even though it has language” - Alexei Efros

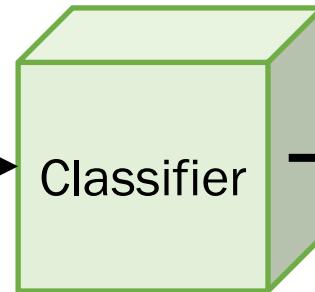
Lisa Dunlap*, Yuhui Zhang*, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E. Gonzalez, Serena Yeung

ML Pipeline

Training Data



Train



Deploy



Model failure



Model success

Analyze



"The model fails when the plane is on a runway"

What is in the failures that aren't in the successes?



+

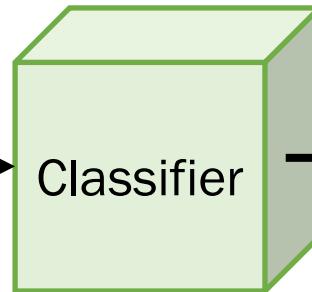
Collect Data

ML Pipeline

Training Data



Re-
Train



Re-
Deploy



Model
~~failure~~
success

Model
success

AGI



Bankrupt

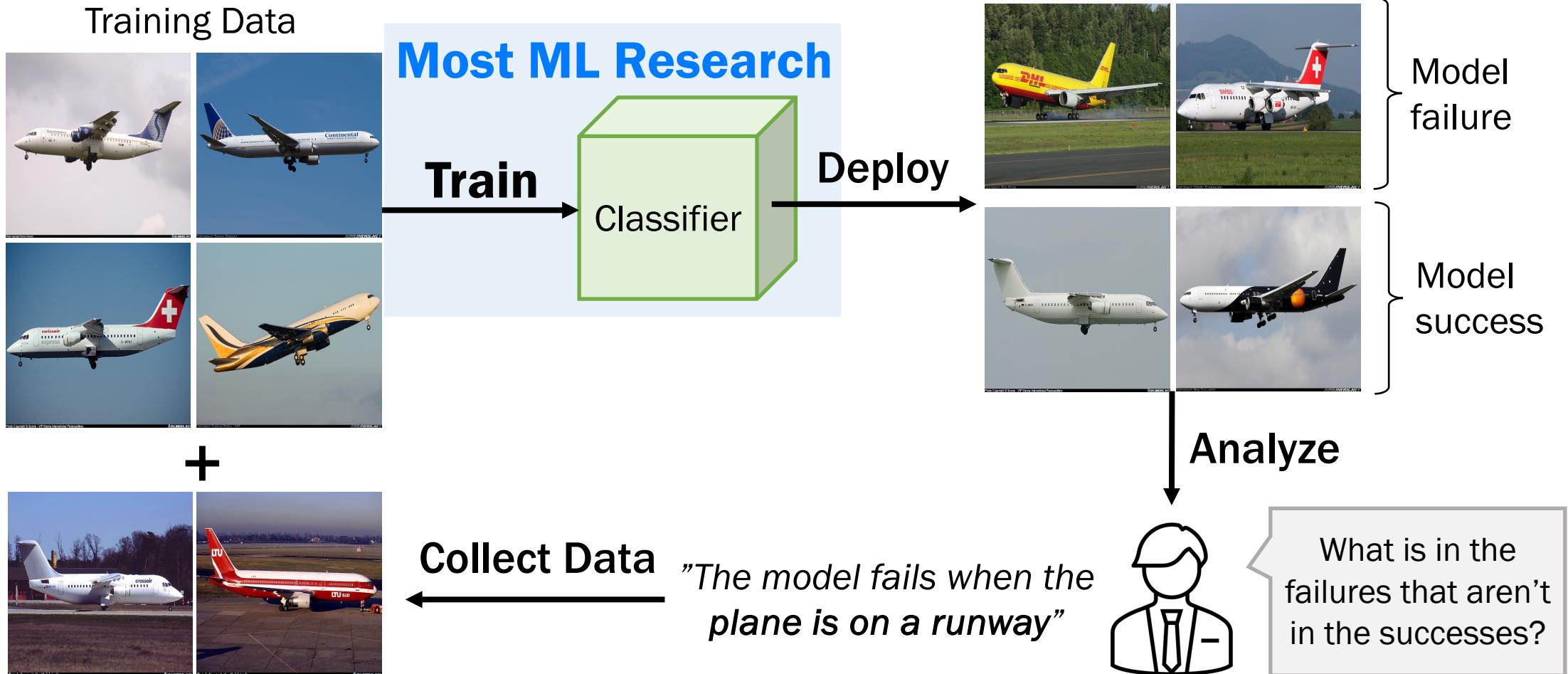
Collect Data

"The model fails when the plane is on a runway"



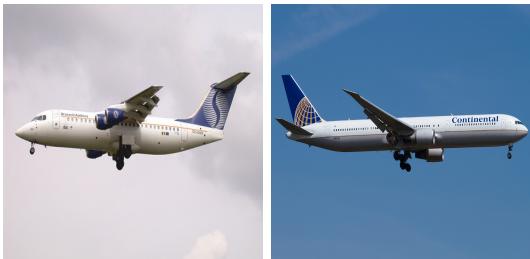
What is in the failures that aren't in the successes?

ML Pipeline



ML Pipeline

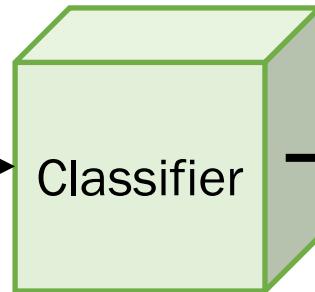
Training Data



+



Train



Deploy



What is the difference
between these sets?

Model
failure

Model
success

Biggest Bottleneck

Analyze

"The model fails when the
plane is on a runway"



What is in the
failures that aren't
in the successes?

Introducing Set Difference Captioning (SDC)

Given two sets of images D_A and D_B , output natural language descriptions which are more often true for D_A than D_B



Output

D_A contains more... "airplanes on runways"

SO MANY Tasks can be framed as SDC

- **How do datasets differ?**

D_A = dataset 1 D_B = dataset 2

- **How do two models compare to one another?**

D_A = model A outputs D_B = model B outputs

- **How have concepts changed in the past decade?**

D_A = images from 2023 D_B = images from 2013

- **What types of Ads get more engagement?**

D_A = Ads w/ high engagement D_B = Ads w/ meh engagement

How do we find differences?

+ Annoy Alyosha in the process

Understand Big Data with Big Models



Gemini



Please describe this image in detail.



The image features a cargo airplane, which appears to be a variant of the Boeing 737, known for its distinct hump-shaped forward fuselage which houses the cockpit above the main cargo area. This design is characteristic of a Boeing 737 Converted Freighter (BCF). The aircraft has a white fuselage with the words "AIR FREIGHT" prominently displayed in blue on the side, indicating its use for cargo transport. There is also a blue stripe that runs along the windows line, and a red stripe beneath it.

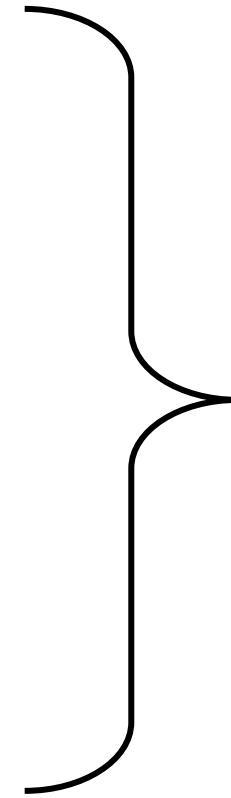
The plane is equipped with two jet engines, one under each wing. The wings are low-mounted with noticeable flaps and ailerons, typical of a commercial jet. The tail is standard with a vertical stabilizer and horizontal stabilizers at the rear end of the fuselage.

The airplane is on the ground, parked on a tarmac, and there is a red building with a grey roof visible in the background, which could be a cargo terminal or hangar. There are no people or other vehicles visible in the picture, which focuses solely on the aircraft. The sky is clear blue, suggesting good weather conditions, and it appears to be daytime.

Understand Big Data with Big Models



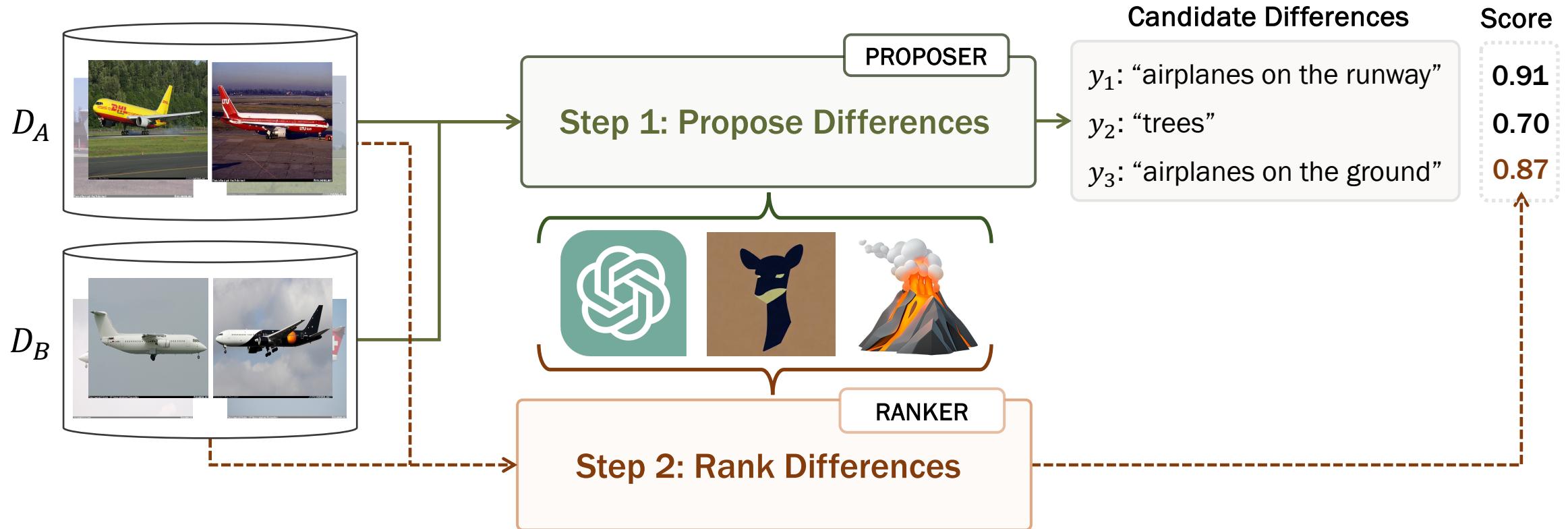
Gemini



Utilize all these
models to reason over
large sets of images

General SDC Framework

Two stage framework consisting of a **Proposer** which generates candidate differences and a **Ranker** which scores each difference

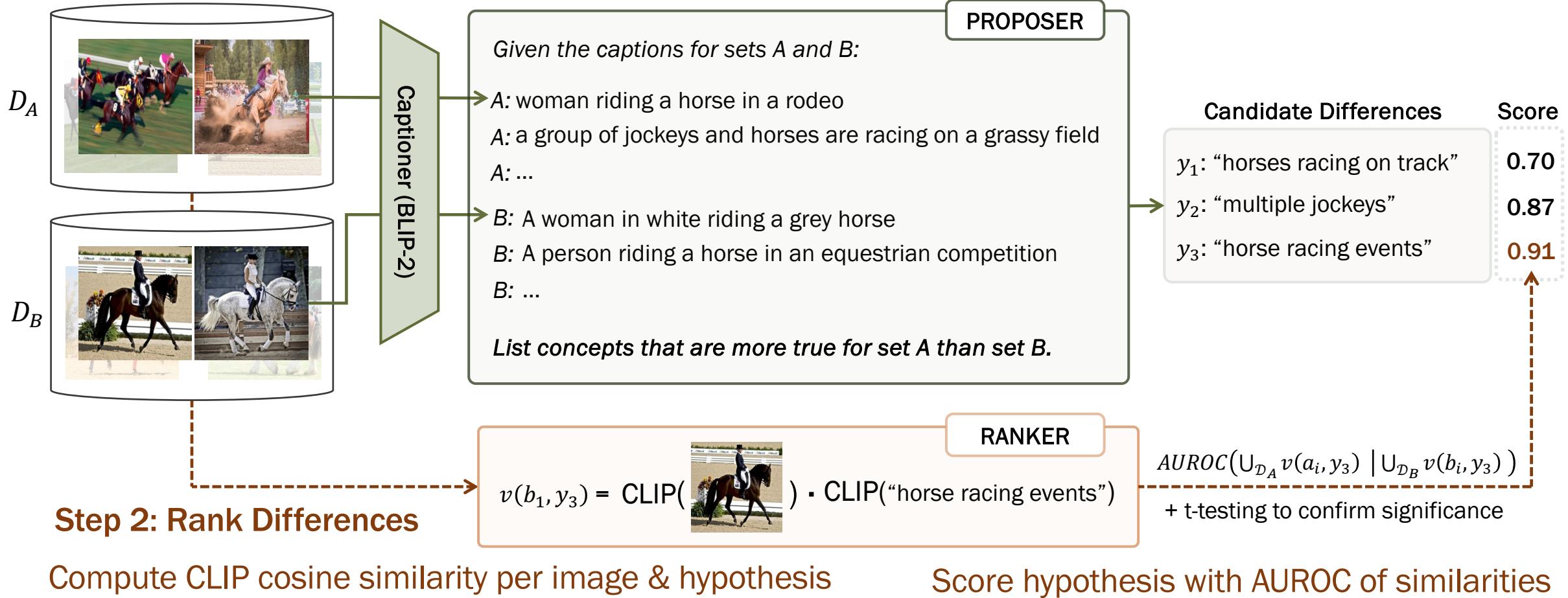


VisDiff : Describing Differences with Multi-Modal Models

Step 1: Propose Differences

Caption images in each set

Feed captions in LLM to generate differences



Applications

We apply VisDiff to applications where:

1. You have two large sets of images you want to compare
2. You have many sets of images you want to compare

Goal: mass-produce interpretable insights about your data, which a human can then confirm through existing work or further validation experiments

Applications

Comparing Datasets

Summarizing Model Failure

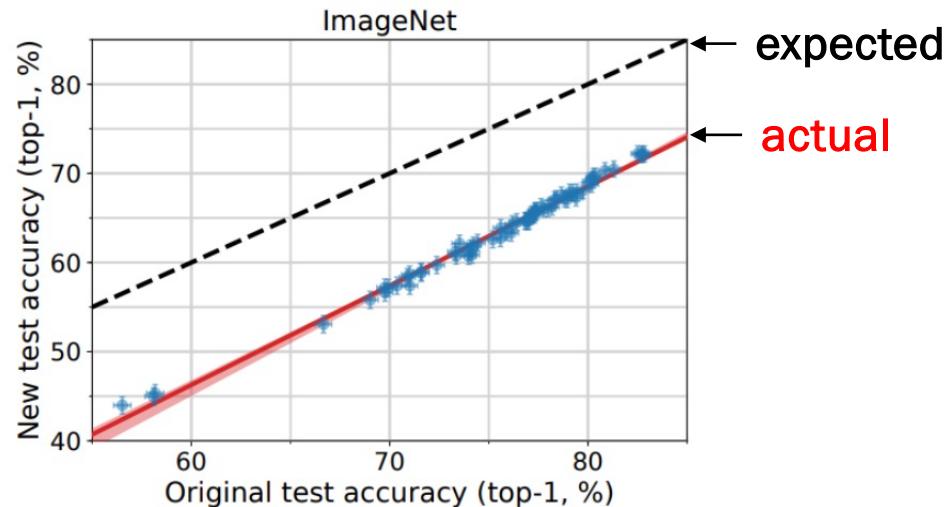
Comparing Model Behavior (classification)

Comparing Model Behavior (generation)

**What exactly IS the
difference between
ImageNet and ImageNetV2?**

ImageNetV2

- New test set for ImageNet which follows the original ImageNet dataset collection process as closely as possible*
- TL;DR accuracy on ImageNetV2 is much lower than ImageNet



Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht*
UC Berkeley Rebecca Roelofs
UC Berkeley Ludwig Schmidt
UC Berkeley Vaishaal Shankar
UC Berkeley

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models’ inability to generalize to slightly “harder” images than those found in the original test sets.

1 Introduction

The overarching goal of machine learning is to produce models that *generalize*. We usually quantify generalization by measuring the performance of a model on a held-out test set. What does good performance on the test set then imply? At the very least, one would hope that the model also performs well on a new test set assembled from the same data source by following the same data cleaning protocol.

In this paper, we realize this thought experiment by replicating the dataset creation process for two prominent benchmarks, CIFAR-10 and ImageNet [10, 35]. In contrast to the ideal outcome, we find that a wide range of classification models fail to reach their original accuracy scores. The accuracy drops range from 3% to 15% on CIFAR-10 and 11% to 14% on ImageNet. On ImageNet, the accuracy loss amounts to approximately five years of progress in a highly active period of machine learning research.

ImageNet VS ImageNetV2

ImageNet



ImageNetV2



Comparing Datasets

Investigating per-class differences

ImgNtV2



Class: Palace

VisDiff: *ImageNetV2 contains more..*

“East Asian architecture”,
“The Forbidden City in Beijing”*

ImgNt



ImgNtV2



Class: High-bar, gymnastics

VisDiff: *ImageNetV2 contains more..*

“Men’s gymnastics events”,
“London 2012 Olympics”

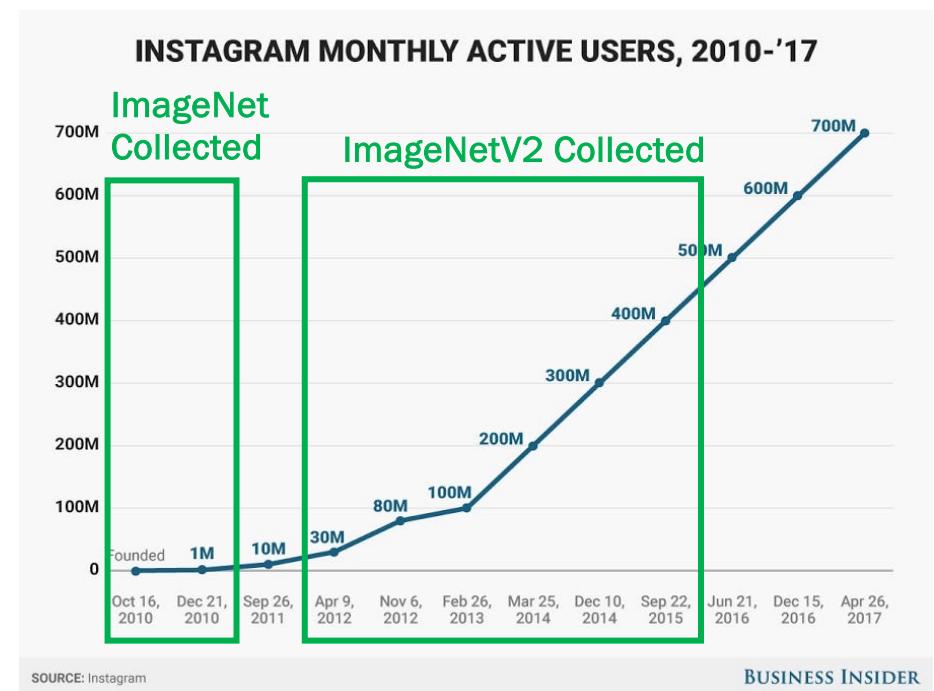
ImgNt

Comparing Datasets

Investigating global dataset differences

Setup: $D_A = \text{ImageNetV2}$ $D_B = \text{ImageNet}$

Top Difference: “photos taken from Instagram”



Applications

Summarizing Model Failure

Comparing Model Behavior (generation)

Computational CogSci

On what types of ImageNet
images is ResNet likely to fail
on?

Summarizing ResNet Model Failure

Setup: Run pretrained ResNet50 & ResNet101 on ImageNet,

D_A = misclassifications D_B = correct classifications

Incorrectly
Classified



Correctly
Classified

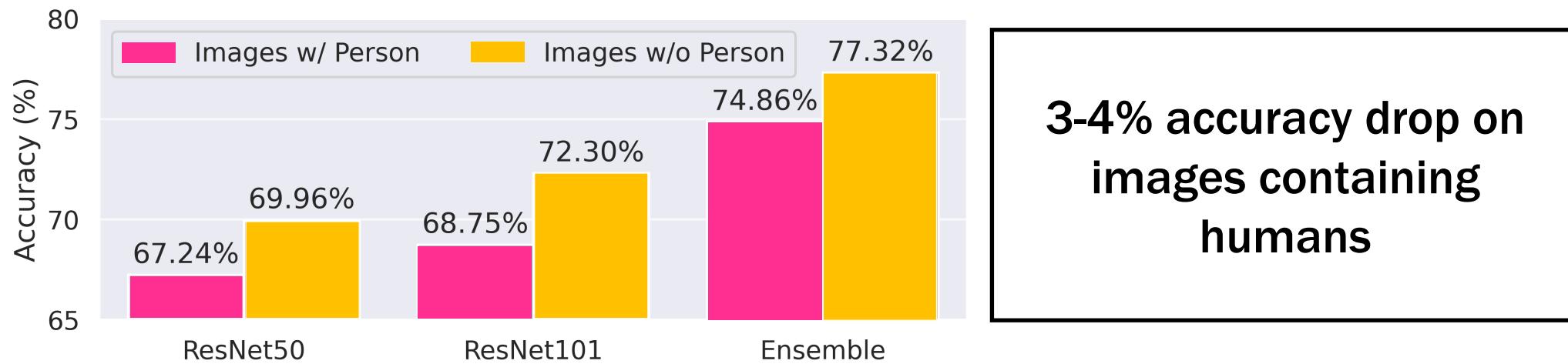


VisDiff : *misclassified images have more “humans interacting with objects”*

Summarizing ResNet Model Failure

Setup: Run pretrained ResNet50 & ResNet101 on ImageNet,

D_A = misclassifications D_B = correct classifications



VisDiff : misclassified images have more “humans interacting with objects”

Applications

Summarizing Model Failure

Comparing Model Behavior (generation)

Computational CogSci

What are the differences
between images generated
by StableDiffusionV1 and V2?

Comparing Different Text-2-Image Models

D_A

Stable DiffusionV2



“Infinity Pool”



“black holes in the center of a galaxy”



“yin-yang”

D_B

Stable DiffusionV1

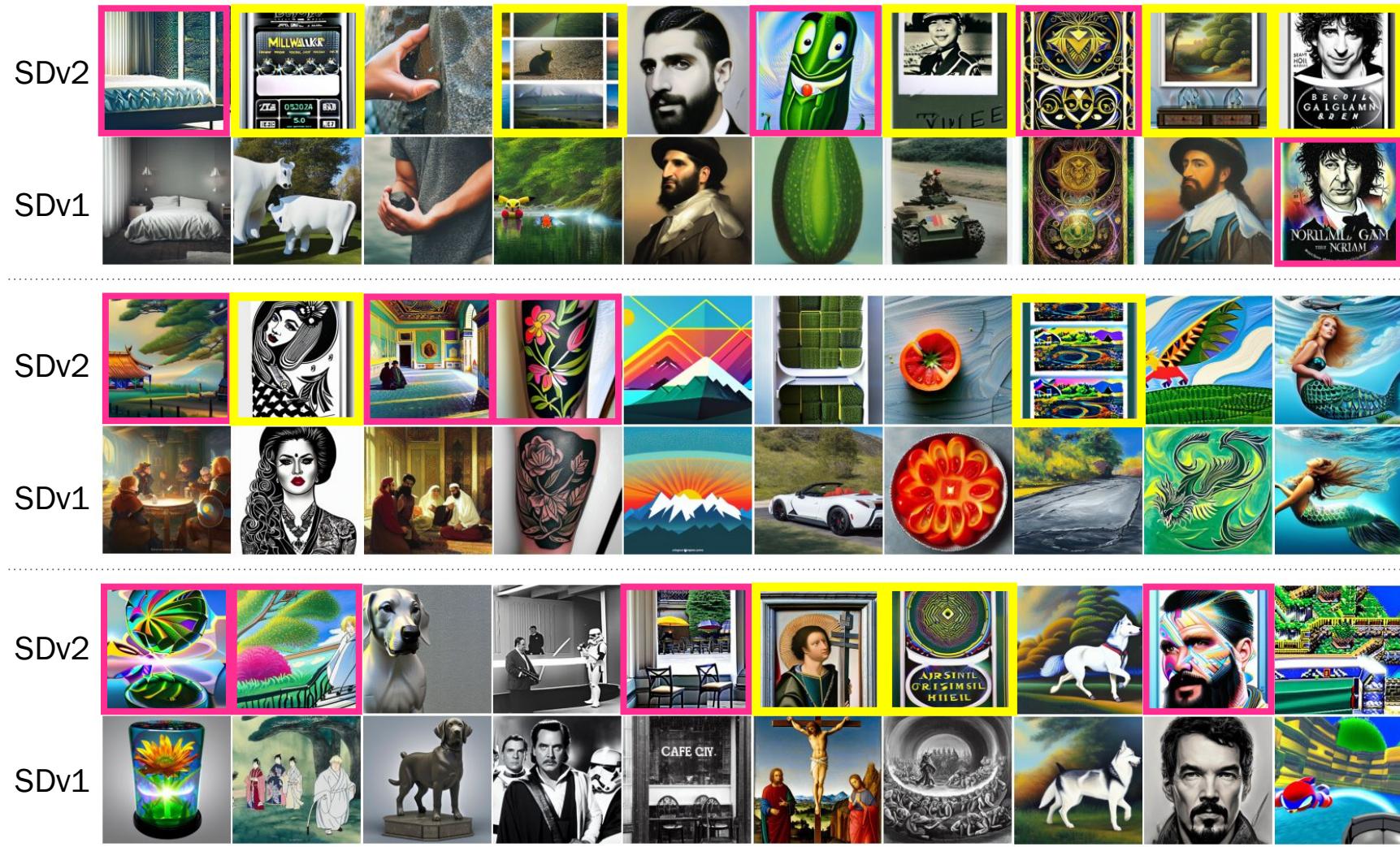


Comparing Different Text-2-Image Models

Setup: $D_A = \text{SDv2}$ images $D_B = \text{SDv1}$ images

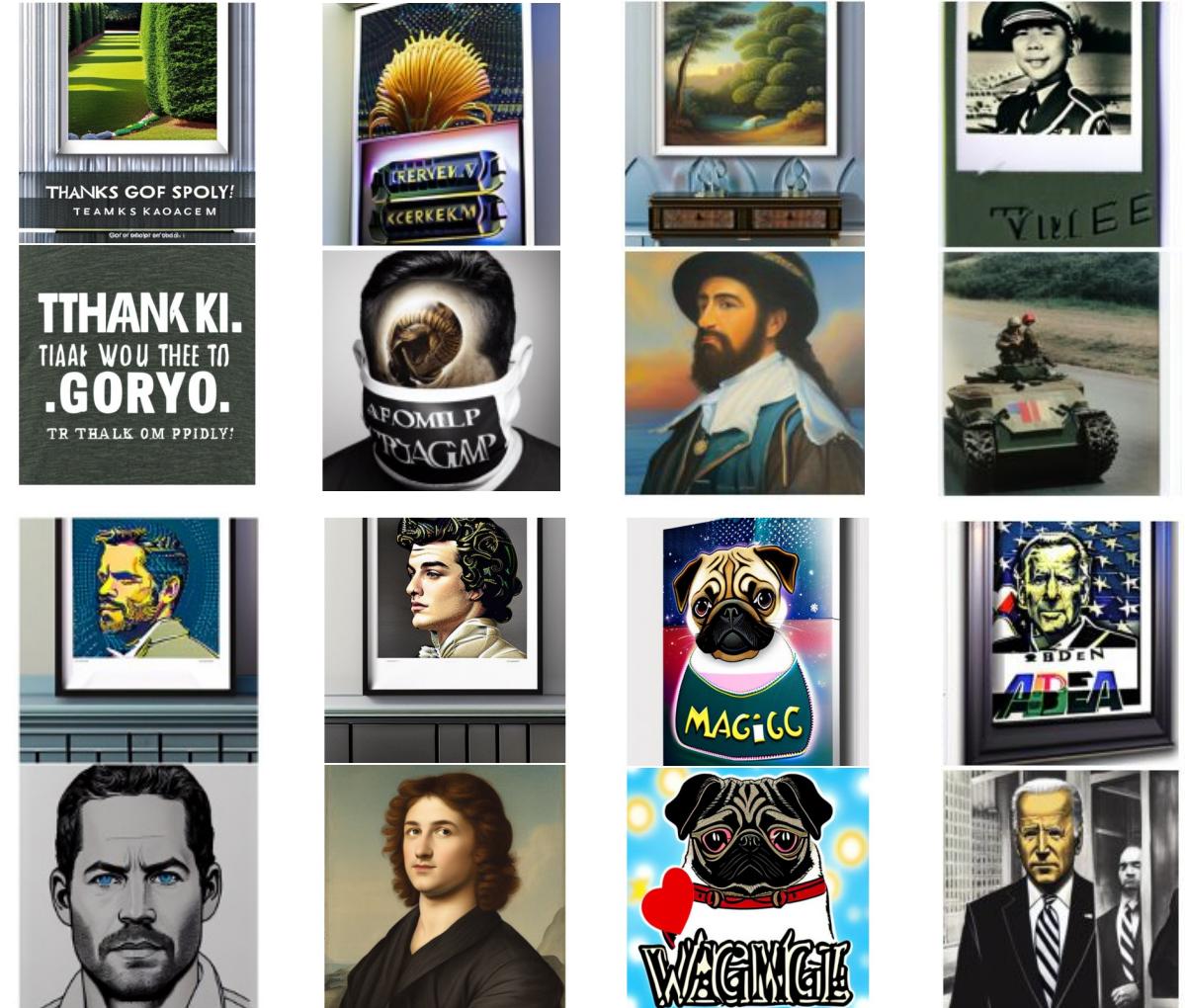
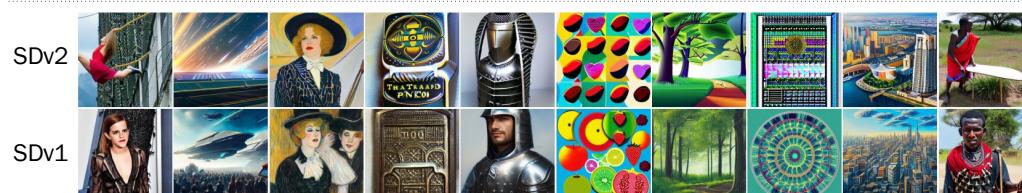
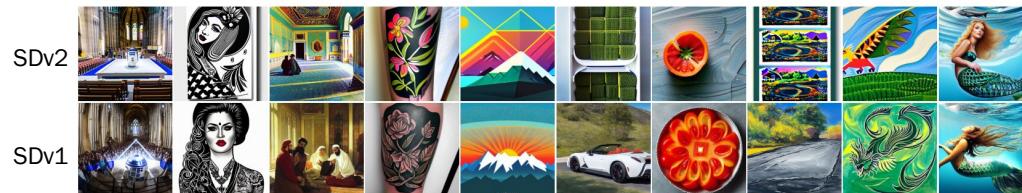


VisDiff : SDv2 images contain more “vibrant colors”
and “images in stands or picture frames”



VisDiff : SDv2 images contain more "vibrant colors"
and "images in stands or picture frames"

Comparing Different Text-2-Image Models



Applications

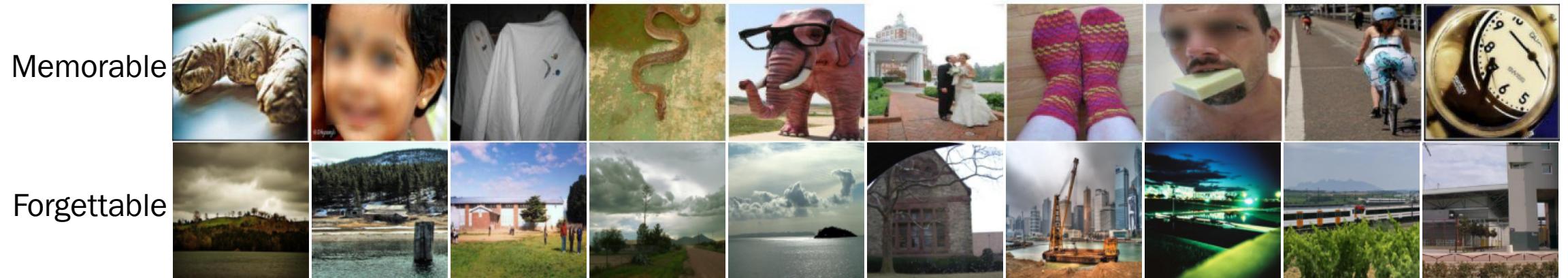
Summarizing Model Failure

Comparing Model Behavior (generation)

Computational CogSci

**What makes an image
memorable or forgettable?**

What makes an image memorable?



VisDiff Output (D_A = memorable images D_B = forgettable images):
"close-up of individual people", "use of accessories or personal items", "humorous elements"

VisDiff Output (D_A = forgettable images D_B = memorable images):
"landscape or beach settings", "abandoned or isolated structures, "sunset or nighttime scenes"

Classify memorability with CLIP using VisDiff outputs results in 90% accuracy

Chatbot Arena – Play around with VLMs!

<https://chat.lmsys.org/>

Visualizations!

Good visualizations are key to understanding your data

- Use [Weights and Biases](#) (not sponsored)
- VQA models are pretty good so **utilize them**
 - Use GPT to label images with attributes you are interested in
- For big datasets, group similar images together and summarize each group
 - Clustering on image embeddings and summarizing the cluster with GPT
- Compare groups/datasets with one another to find out distribution shift (sponsored)

What about describing differences in text?

Describing Differences between Text Distributions with Natural Language

Ruiqi Zhong¹ Charlie Snell¹ Dan Klein¹ Jacob Steinhardt¹

Abstract

How do two *distributions* of text differ? Humans are slow at answering this, since discovering patterns might require tediously reading through hundreds of samples. We propose to automatically describe the differences by “learning a natural language hypothesis”: given two distributions D_0 and D_1 , we search for a description that is more often true for D_1 , e.g., “*is military-related*.” To tackle this problem, we fine-tune GPT-3 to propose descriptions with the prompt: “[samples of D_0] + [samples of D_1] + *the difference between them is* __.” We then re-rank the descriptions by checking how often they hold on a larger set of samples with a learned verifier. On a benchmark of 54 real-world binary classification tasks, while GPT-3 Curie (13B) only generates a description similar to human annotation 7% of the time, the performance reaches 61% with fine-tuning and re-ranking, and our best system using GPT-3 Davinci (175B) reaches 76%. We apply our system to de-



Figure 1. Given two distributions (top), our system automatically discovers their differences and describes them with natural language (bottom). Grey/white background represents D_0/D_1 and red/blue represents whether a sample matches the description s .

Goal Driven Discovery of Distributional Differences via Language Descriptions

Ruiqi Zhong*, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, Jacob Steinhardt

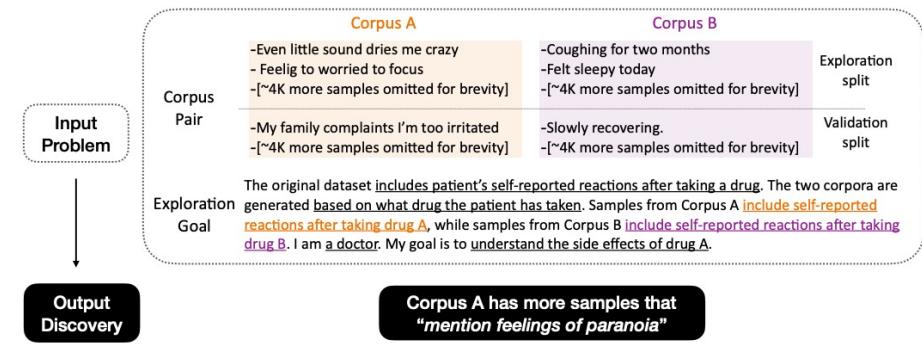
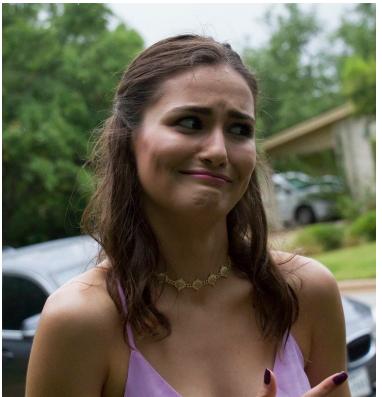


Figure 1: Each problem in OPEND5 contains 1) a corpus pair, which has ~17K samples on average and is partitioned into two halves called “exploration split” and “validation split”, and 2) a natural language description of the exploration goal, which also contains information about how the corpus pair was collected. A D5 system takes the goal and the exploration split as inputs and generates valid and relevant discoveries in natural language as outputs. The underlined texts in the exploration goal vary across problems, while the rest are templates.

It takes a village...



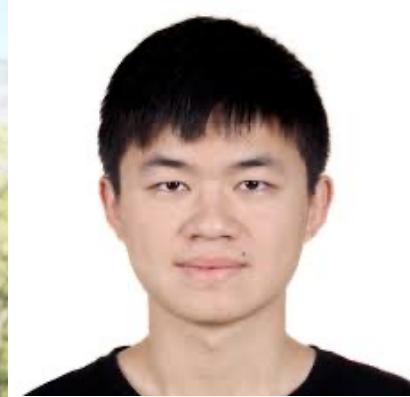
Lisa Dunlap



Yuhui Zhang



Xiaohan Wang



Ruiqi Zhong



Salem



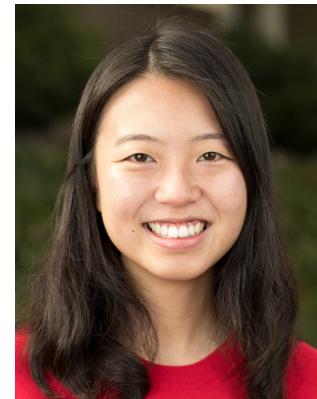
Trevor Darrell



Jacob Steinhardt



Joseph Gonzalez



Serena Yeung-Levy

+ many more!