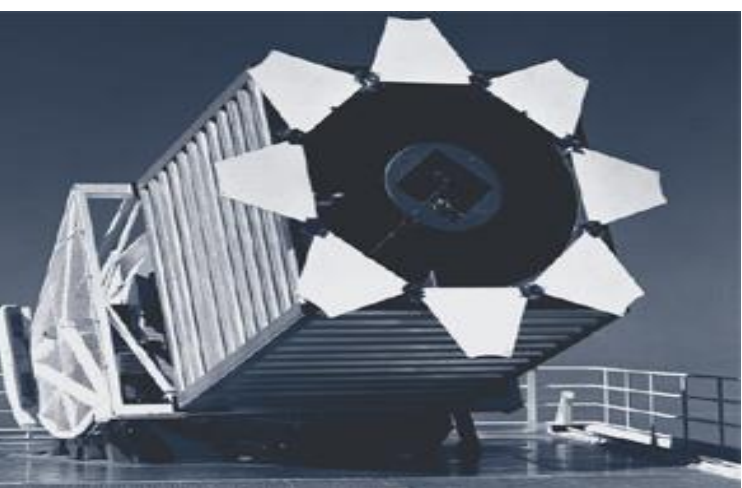




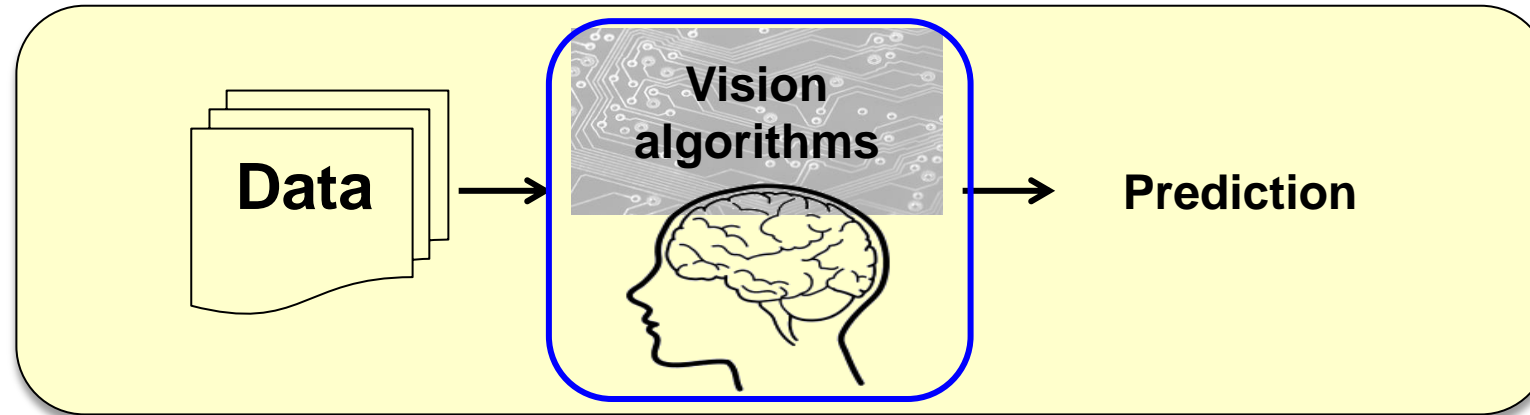
Which images need human attention?

Kristen Grauman
Department of Computer Science
University of Texas at Austin

Work with Yong Jae Lee, Sudheendra
Vijayanarasimhan, Prateek Jain, and Lu Zheng



Interactive visual analysis



Key question:

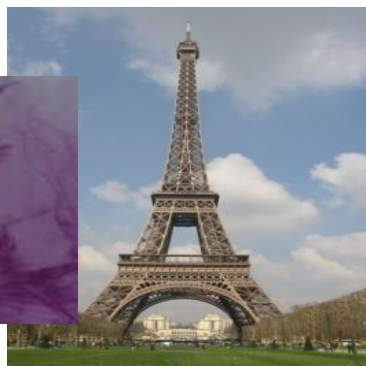
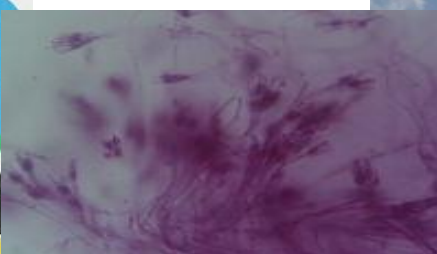
- Which visual data deserves human attention?

Two examples:

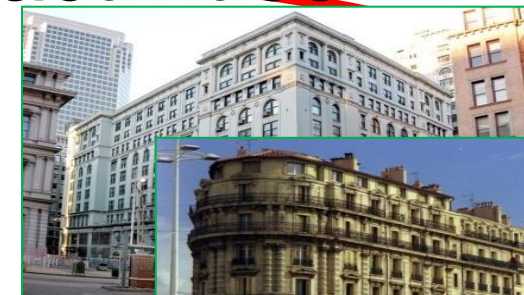
1. Supervised learning of object categories
2. Unsupervised video summarization

Visual recognition

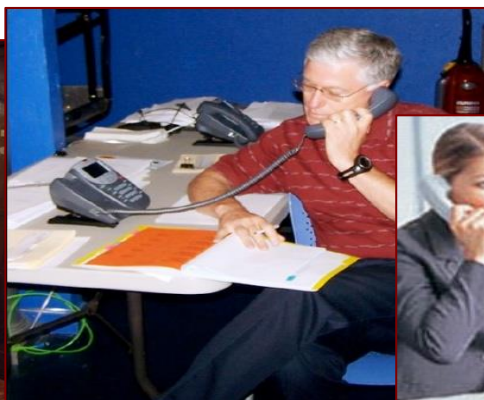
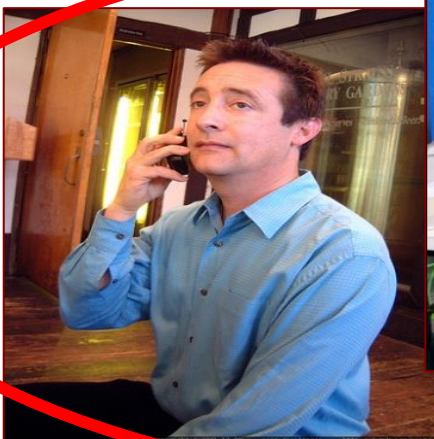
Recognition of objects, categories, scenes, activities



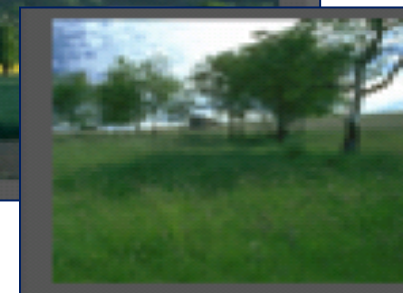
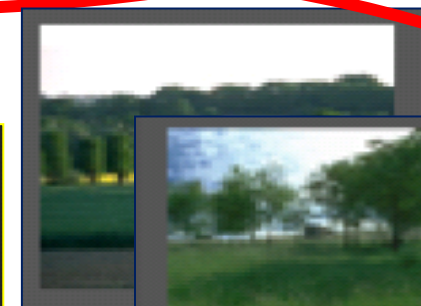
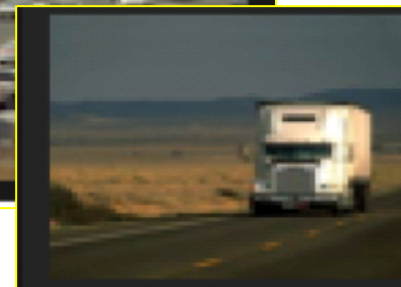
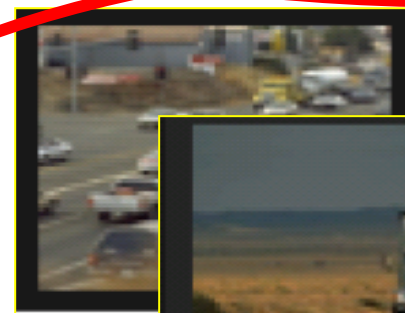
Specific objects



Object categories



Activities



Scenes

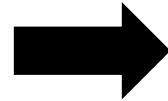
The importance of data in recognition

Best approaches today rely on discriminative learning

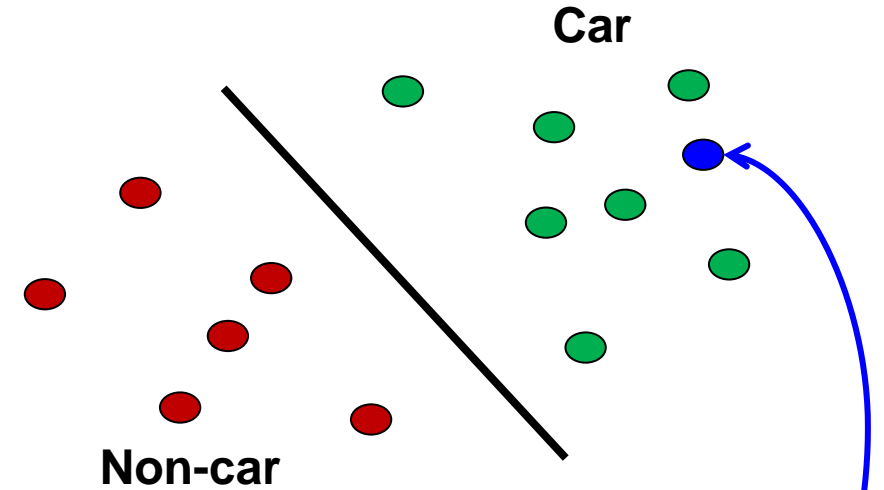


■
■
■

Training images



Annotator



Novel test image

The importance of data in recognition

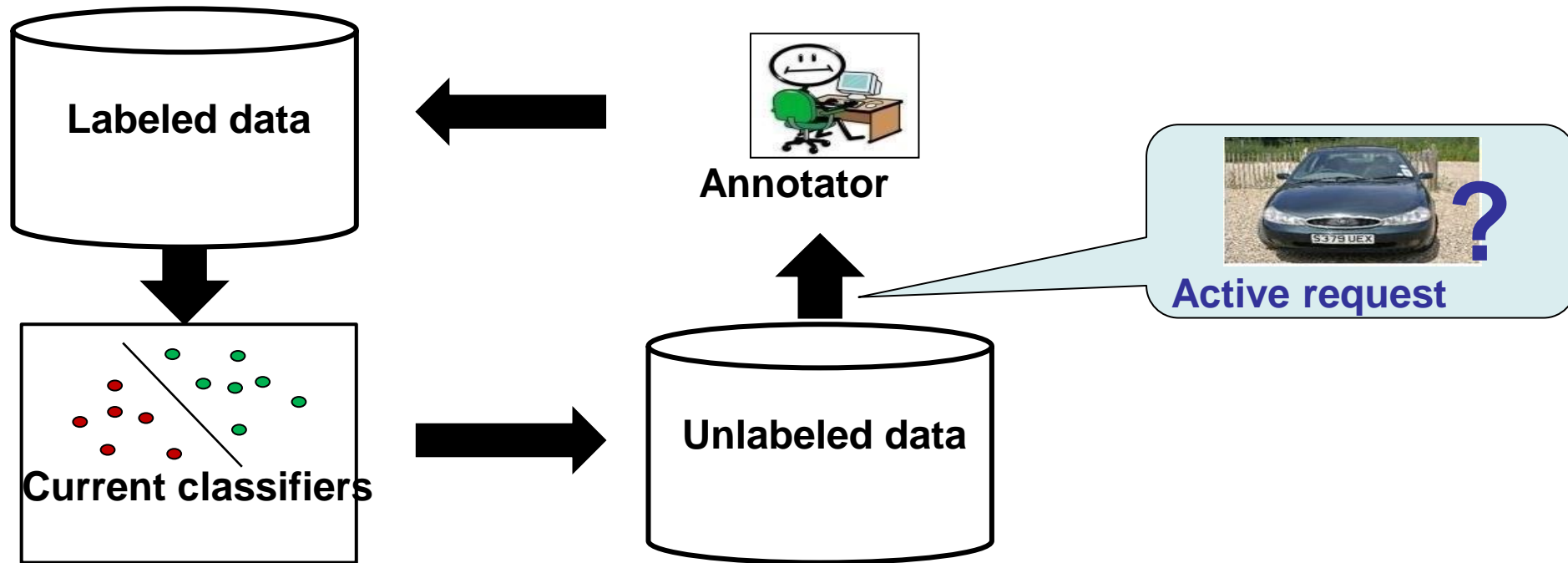
- Dataset creation

[LabelMe - Russell et al. 2005, Caltech - Griffin et al. 2007, Image-Net – Deng et al. 2010, PASCAL VOC – Everingham et al.,...]

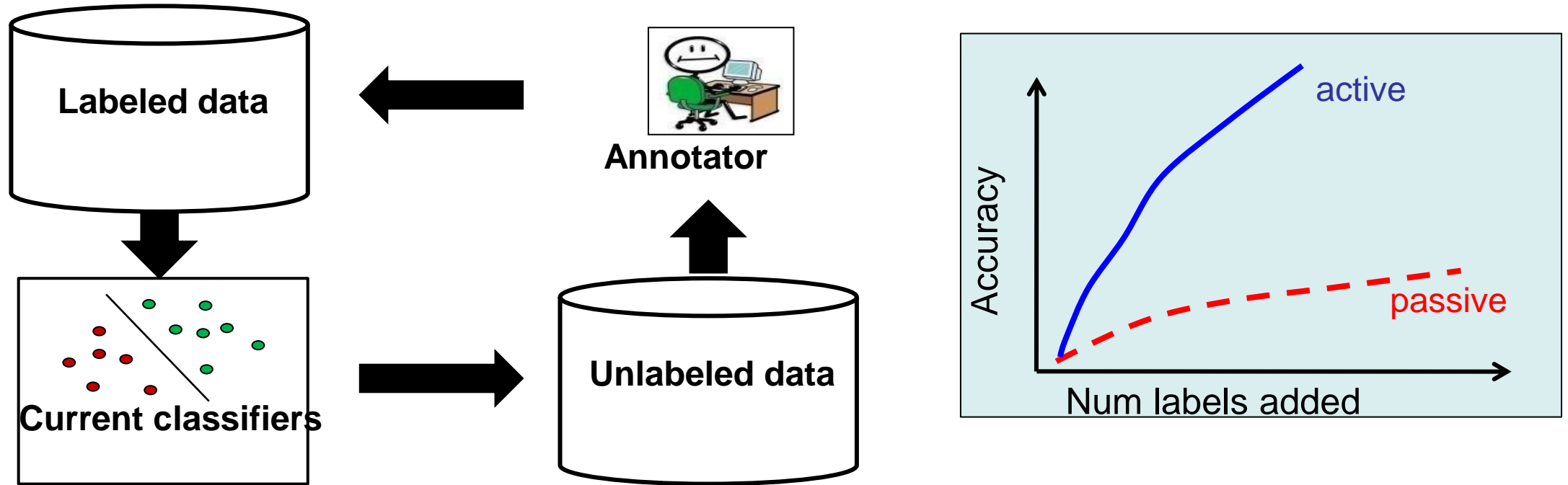
- Gathering annotations from “crowds”

[Sorokin et al. 2009, Vijayanarasimhan et al. 2009, Deng et al 2009, Endres et al. 2010, Branson et al. 2010, Welinder et al. 2010, ...]

Active learning for image annotation



Active learning for image annotation

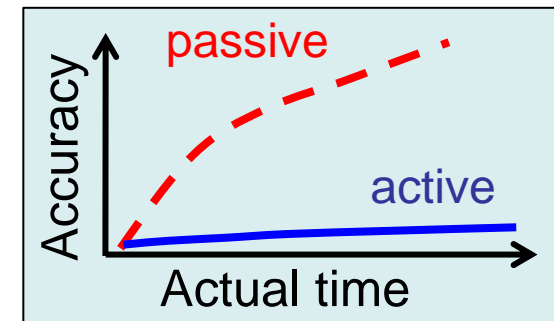


Intent: better models, faster/cheaper

Problem: “Sandbox” learning

Thus far, tested only in artificial settings:

- Unlabeled data already fixed, small scale, biased
- Computational cost ignored



Our idea: **Live** active learning

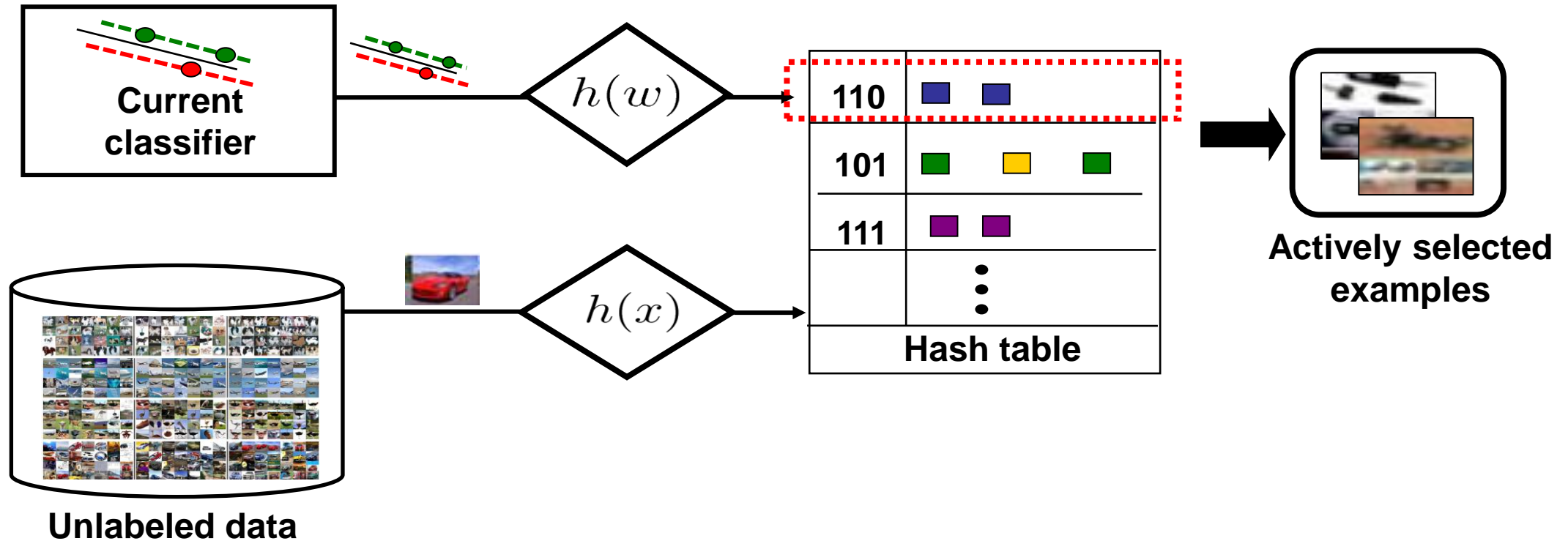
Large-scale active learning of object detectors
with **crawled data** and **crowdsourced labels**.

Key technical challenge:

How to scale active learning to massive unlabeled data?

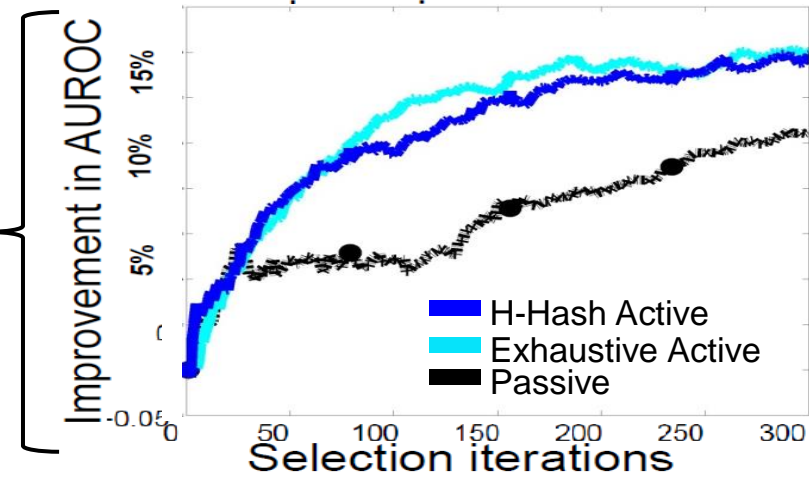
Sub-linear time active selection

We propose a novel hashing approach to identify the most uncertain examples in sub-linear time.

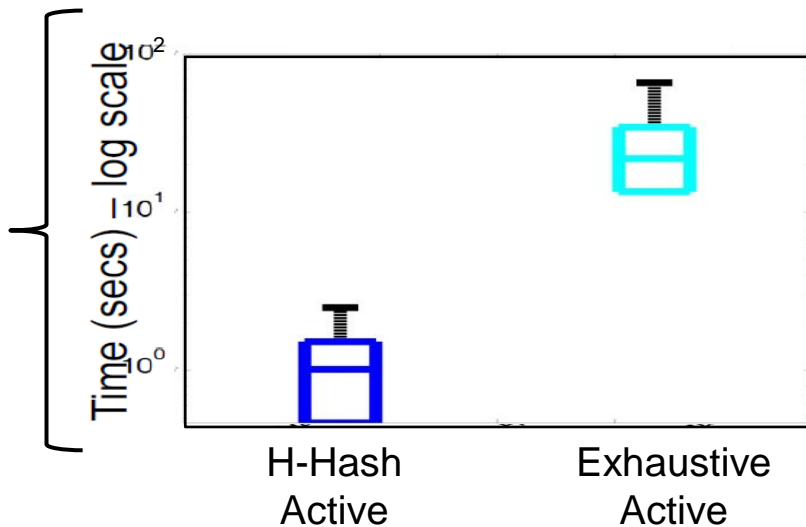


Sub-linear time active selection

Accuracy
improvements as
more data labeled



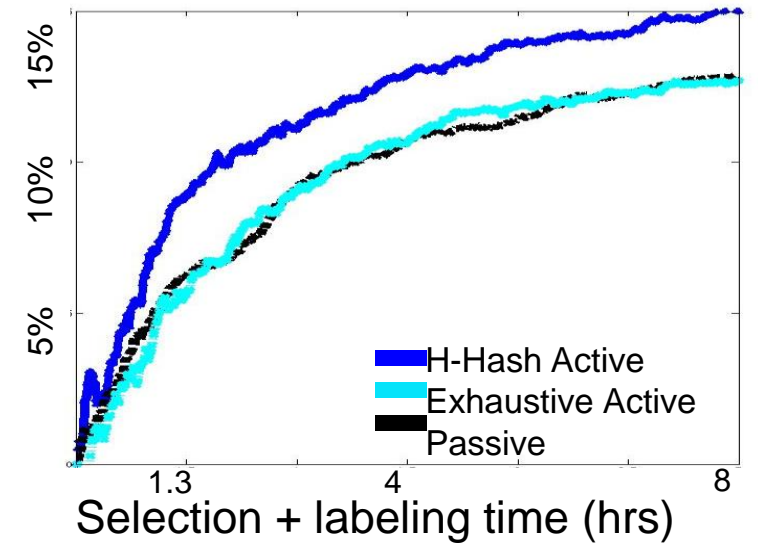
Time spent
searching for
selection



H-Hash result on 1M Tiny Images

Accounting for all costs

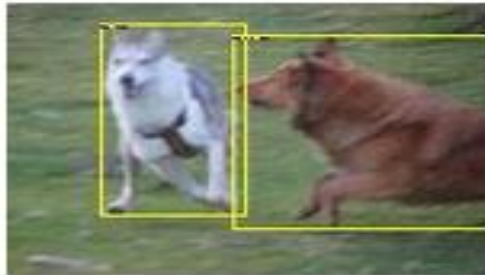
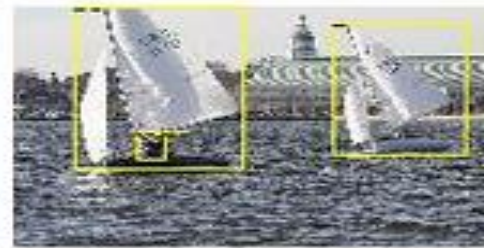
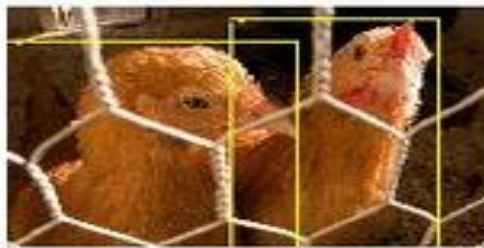
Improvement in AUROC



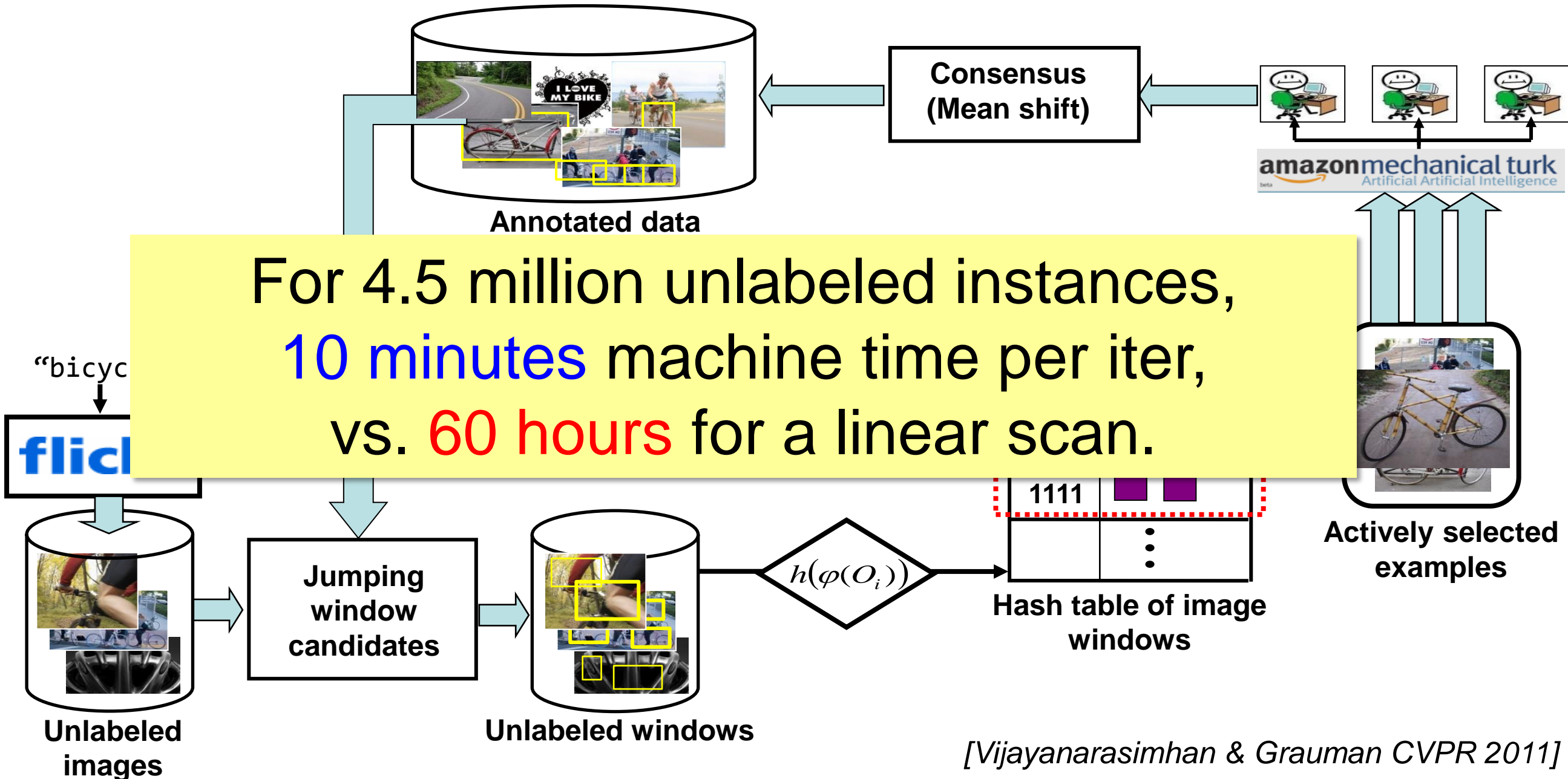
By minimizing **both**
selection and labeling time,
obtain the best accuracy
per unit time.

PASCAL Visual Object Categorization

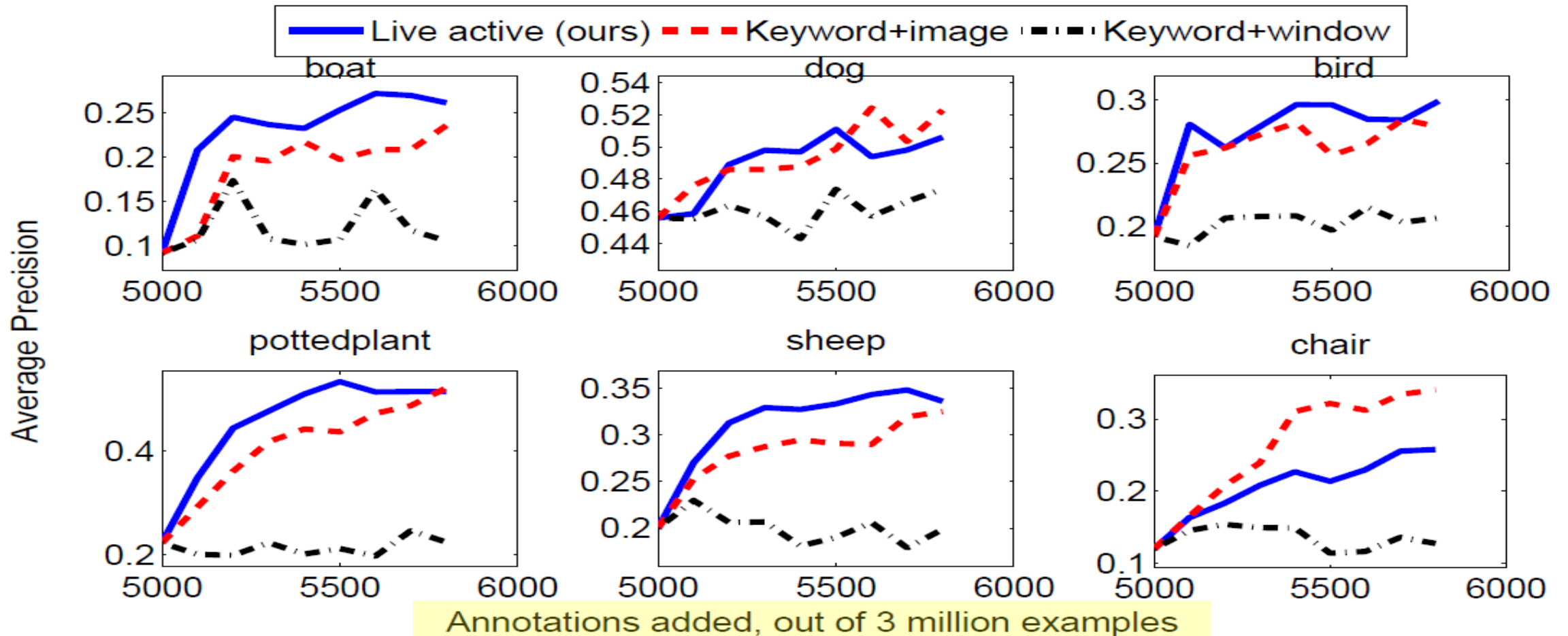
- “The” object detection benchmark
- Original image data from Flickr



Live active learning



Live active learning results



PASCAL VOC objects - Flickr test set

Outperforms status quo data collection approach

Live active learning results

First selections made when learning “boat”:

Ours: live active learning



Keyword+image baseline



Interactive learning for visual recognition



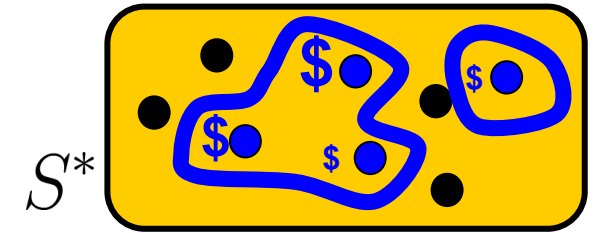
Label propagation in video

[Vijayanarasimhan & Grauman, ECCV 2012]



Joint learning w/attributes

[Kovashka et al. ICCV 2011]



Budgeted batch

[Vijayanarasimhan et al., CVPR 2010]



Active attribute discovery

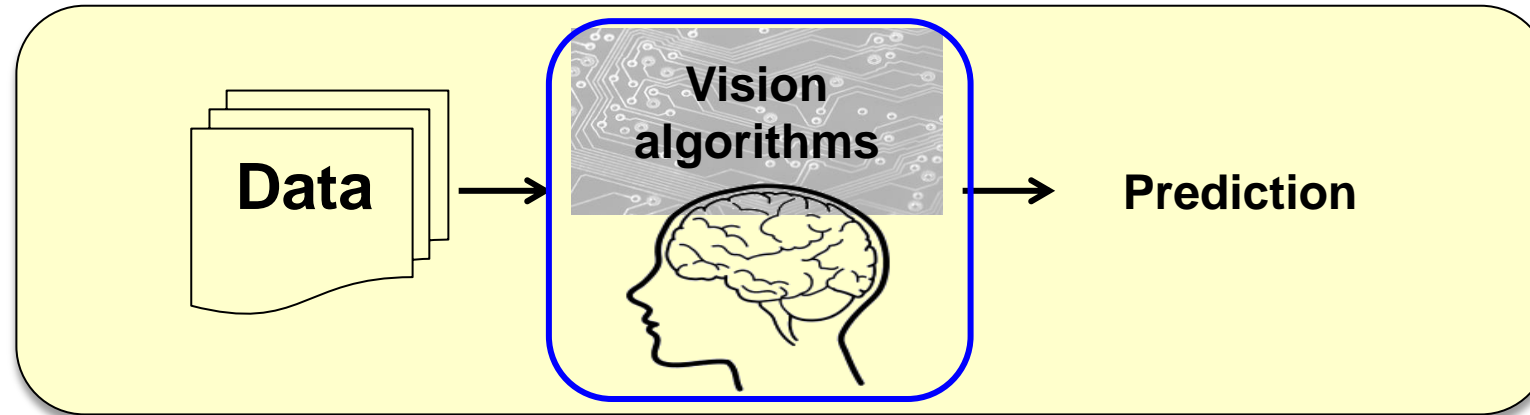
[Parikh & Grauman, CVPR 2011]



Choosing among annotation types

[Vijayanarasimhan & Grauman, NIPS 2008]

Interactive visual analysis



Key question:

- Which visual data deserves human attention?

Two examples:

1. Supervised learning of object categories
2. Unsupervised video summarization

Goal: Generate a visual summary



Wearable camera



Input: Egocentric video of the camera wearer's day



9:00 am

10:00 am

11:00 am

12:00 pm

1:00 pm

2:00 pm

Output: Storyboard (or video skim) summary

~1990

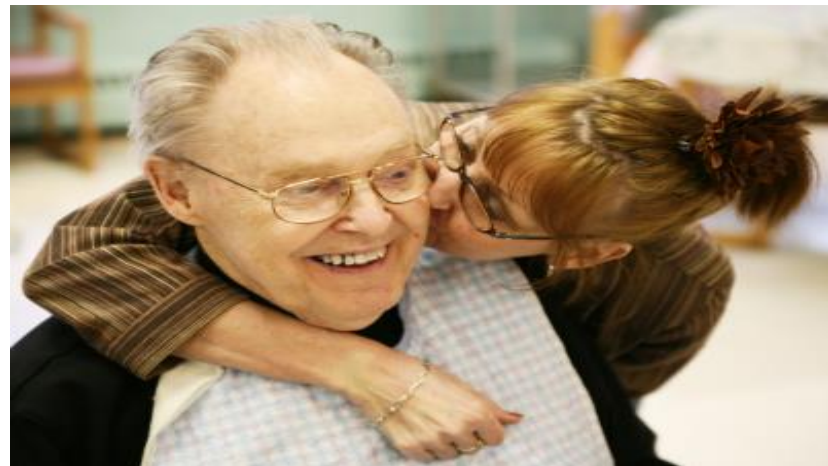


Steve Mann

2013



Potential applications of egocentric video summarization



Memory aid



Law enforcement



Mobile robot discovery

Prior work

- **Egocentric recognition**

[Starner et al. 1998, Doherty et al. 2008, Spriggs et al. 2009, Jojic et al. 2010, Ren & Gu 2010, Fathi et al. 2011, Aghazadeh et al. 2011, Kitani et al. 2011, Pirsiavash & Ramanan 2012, Fathi et al. 2012]

- **Video summarization**

[Wolf 1996, Zhang et al. 1997, Ngo et al. 2003, Goldman et al. 2006, Caspi et al. 2006, Pritch et al. 2007, Laganiere et al. 2008, Liu et al. 2010, Nam & Tewfik 2002, Ellouze et al. 2010]

→ **Low-level cues, stationary cameras**

→ **Consider summarization as a *sampling* problem**

Our idea: Story-driven summarization

Go

1.

2.



akest

Egocentric subshot detection

Define 3 generic ego-activities:



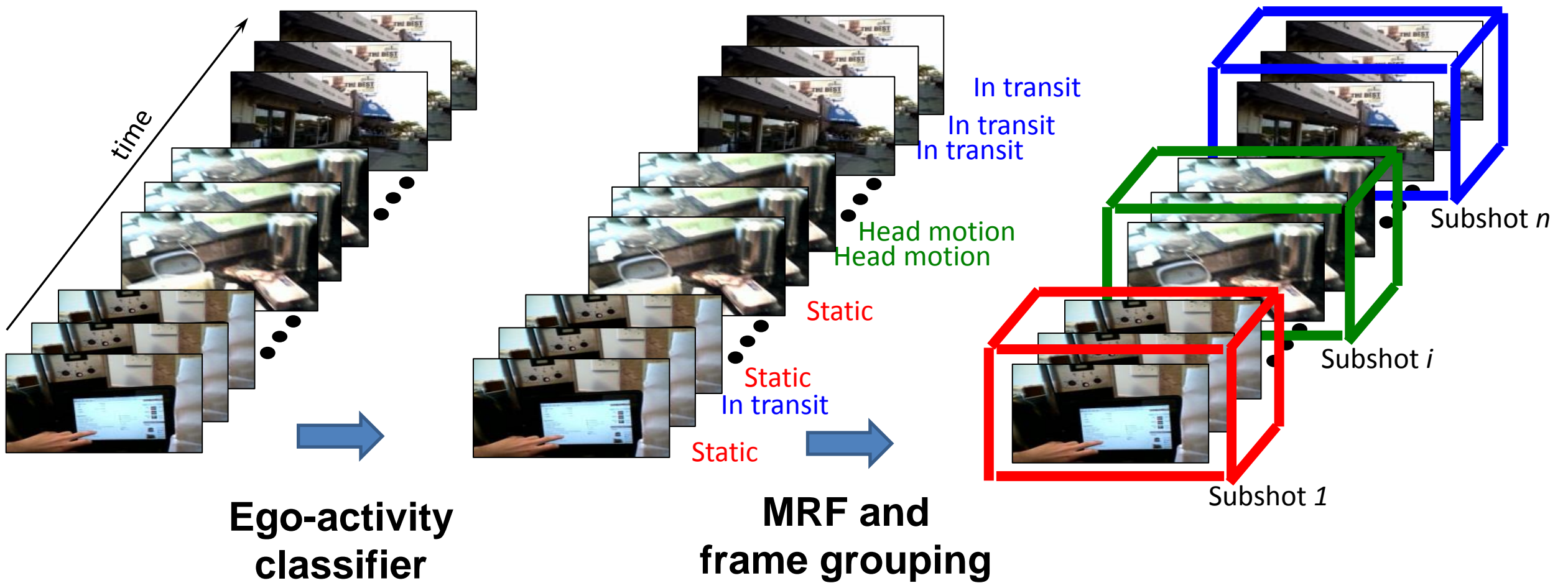
~Static

In transit

Head moving

- Train classifiers to predict these activity types
- Features based on flow and motion blur

Egocentric subshot detection

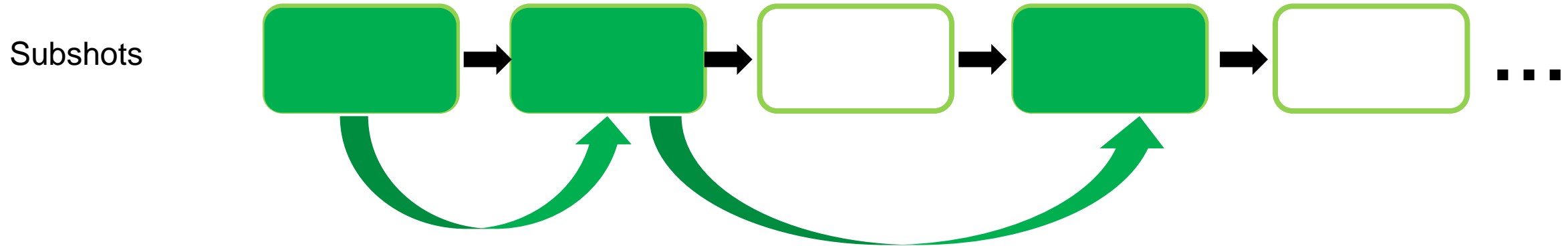


Subshot selection objective

Good summary = chain of k selected subshots in which each influences the next via some subset of key objects

$$S^* = \arg \max_{S \subset \mathcal{V}} \lambda_s \mathcal{S}(S) + \lambda_i \mathcal{I}(S) + \lambda_d \mathcal{D}(S)$$

influence **importance** **diversity**



Document-document influence

[Shahaf & Guestrin, KDD 2010]



CNN Money
A Service of CNN, Fortune & Money

FORTUNE Money

Home Video Business News Markets Term Sheet Economy Tech Personal Finance

REAL ESTATE
Mortgage Meltdown [Archive](#)

Home prices post record decline

S&P/Case-Shiller index of 10 major cities fell 6.7% in October. Housing markets remain 'grim.'

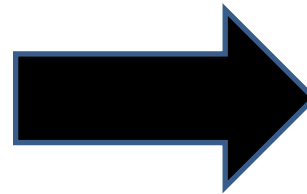
By Les Christie, CNNMoney.com staff writer
December 26 2007: 3:03 PM EST

EMAIL | PRINT | DIGG | RSS

NEW YORK (CNNMoney.com) -- Home prices fell 6.7 percent in October, compared with a year ago, according to the S&P/Case-Shiller 10-city home-price index. It was the largest drop recorded since the index began in 1987.

It marked the 10th consecutive month of price depreciation and 23 months of decelerating returns.

Special Report
FORECLOSURE MORTGAGE MELTDOWN
Seniors face grim choices amid



CNN Politics

Justice Entertainment Tech Health Living Travel Opinion iReport Money

HEALTH CARE

Health-care debate heats up as Senate, House grapple with plans

June 08, 2009

Share Twitter Email

Recommend 44 people recommend this what your friends recomm

As the debate on health-care reform heats up on Capitol Hill, it's clear lawmakers don't see eye-to-eye on the issue -- with each other or President Obama.

Obama told Congress this past weekend that it's time to deliver on health-care reform, and he wants a bill on his desk by October at the latest. But this week already is demonstrating just how difficult and complex coming up with a nuts-and-bolts bill is.

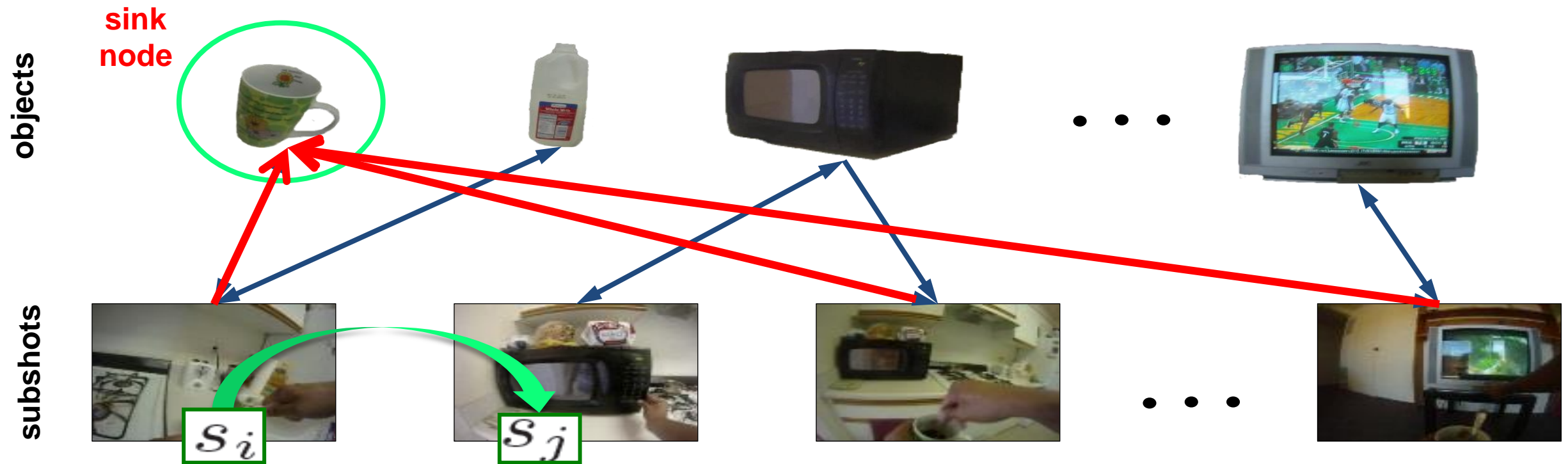
In the Senate, key negotiators broke up a session Monday still stuck on whether to create a government-run health-insurance plan to compete with private insurers -- something Obama and most Democrats want, and most Republicans oppose.

AMBULANCE

President Obama says a public health plan consumers and keep costs down.

Connecting the dots between news articles. D. Shahaf and C. Guestrin. In KDD, 2010.

Estimating visual influence



$$\text{INFLUENCE}(s_i, s_j | o) = \prod_i(s_j) - \prod_i^o(s_j)$$

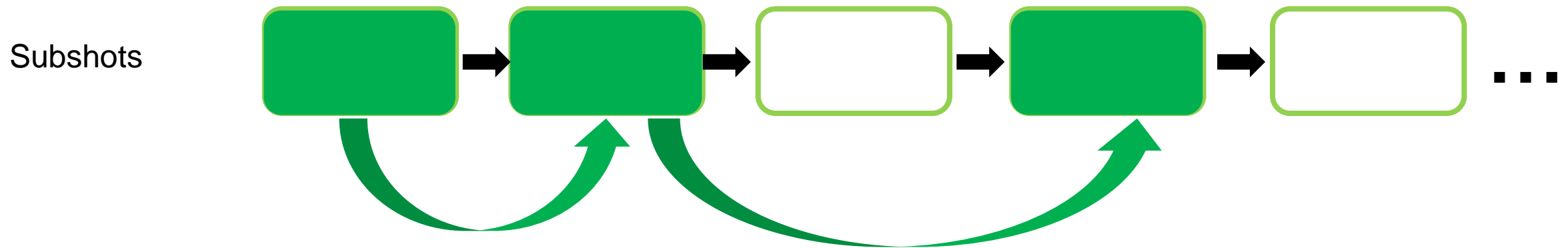
Captures how reachable subshot j is from subshot i , via any object o

Subshot selection objective

Good summary = chain of k selected subshots in which each influences the next via some subset of **key objects**

$$S^* = \arg \max_{S \subset \mathcal{V}} \lambda_s \mathcal{S}(S) + \lambda_i \mathcal{I}(S) + \lambda_d \mathcal{D}(S)$$

influence importance diversity



Learning object region importance

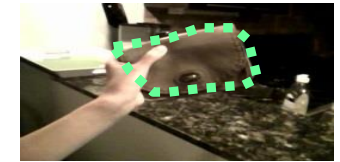
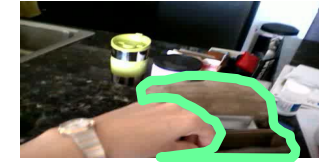
Egocentric features:



distance to hand



distance to frame center



frequency

Learning object region importance

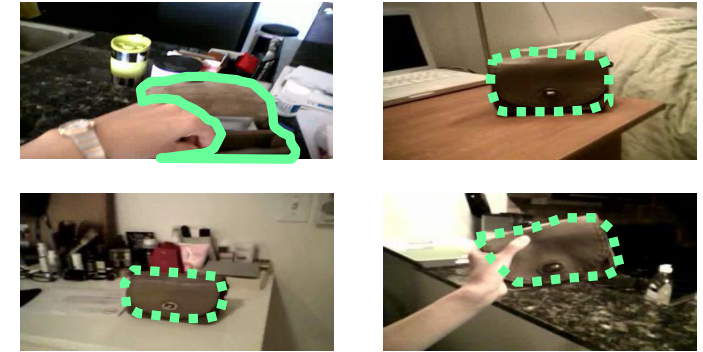
Egocentric features:



distance to hand

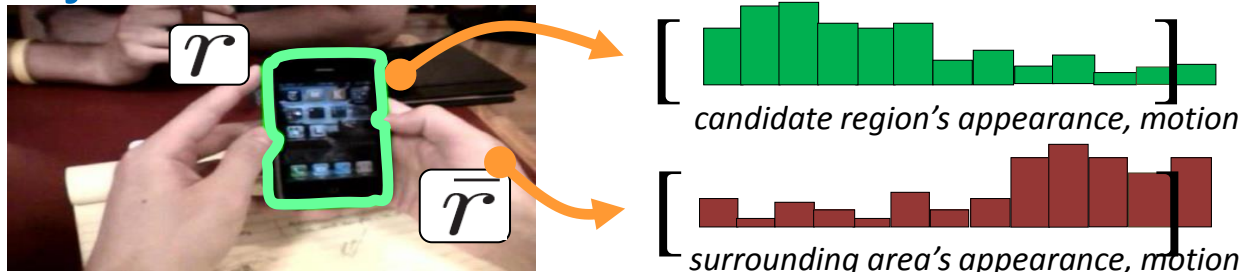


distance to frame center



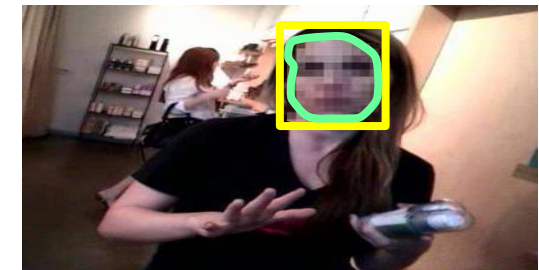
frequency

Object features:



"Object-like" appearance, motion

[Endres et al. ECCV 2010, Lee et al. ICCV 2011]



overlap w/ face detection

Region features: size, width, height, centroid

[Lee et al. CVPR 2012]

Egocentric video datasets

UT Egocentric (UTE)

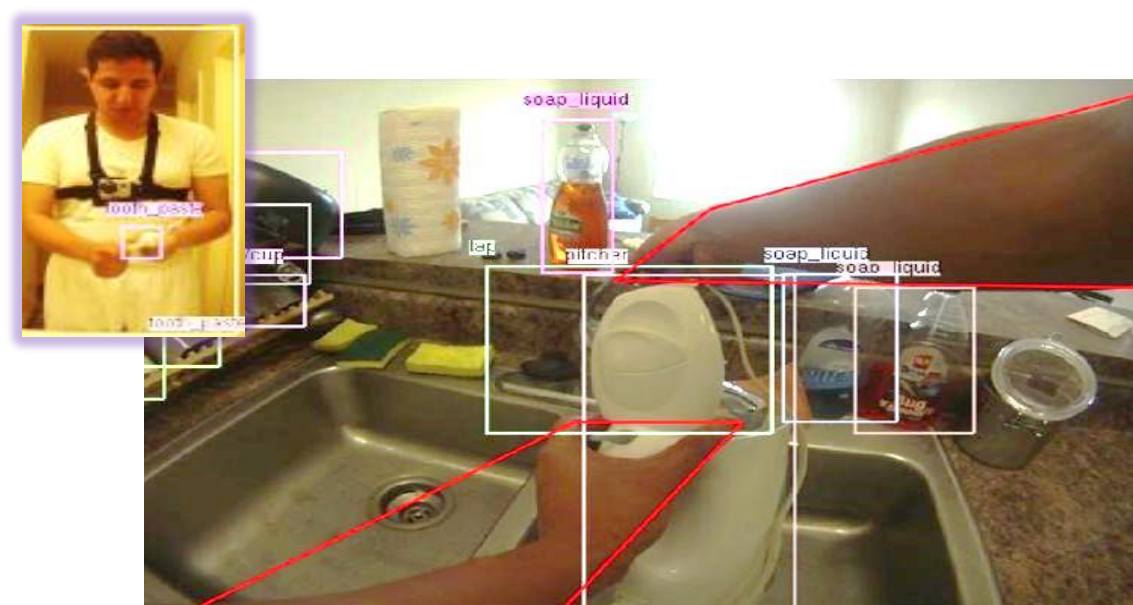
[Lee et al. 2012]



4 videos, each 3-5 hours long, uncontrolled setting.

Activities of Daily Living (ADL)

[Pirsiavash & Ramanan 2012]



20 videos, each 20-60 minutes, daily activities in house.

Human subject results: Blind taste test

How often do subjects prefer our summary?

Data	Uniform sampling	Shortest-path	Object-driven
UTE	90.0%	90.9%	81.8%
ADL	75.7%	94.6%	N/A

34 human subjects, ages 18-60

12 hours of original video

Each comparison done by 5 subjects

Total 535 tasks, 45 hours of subject time

Example keyframe summary

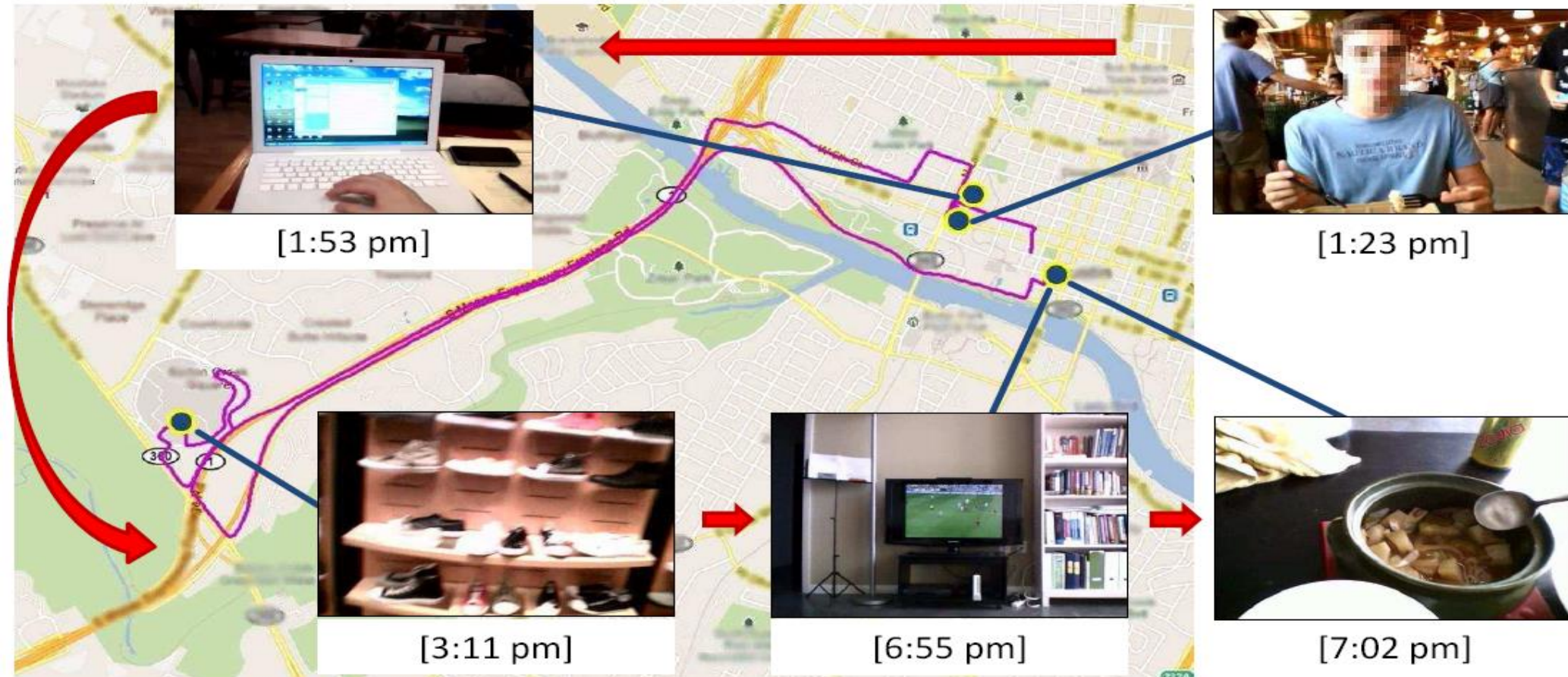


Original video (3 hours)



Our summary (12 frames)

Automatic storyboard maps



Augment keyframe summary with geolocations

Summary

- **Learn to focus human attention on the right data**
 - Actively train object detector with human in the loop
 - Summarize videos for fast human consumption
- **Key challenges**
 - Predicting what is important
 - Scaling to large-scale data collections
- **Semi-automating computer vision → new applications in large-scale visual analysis**