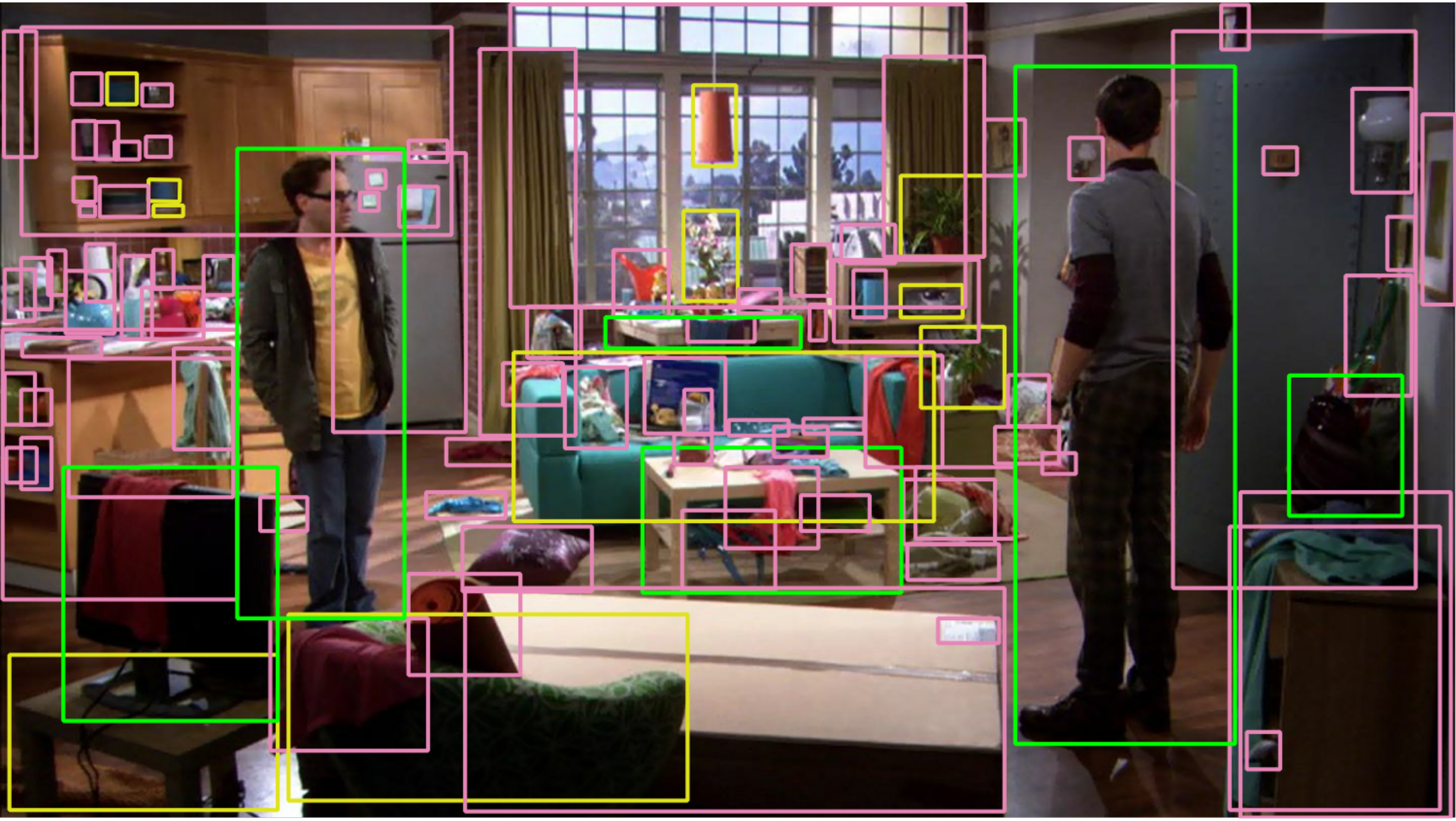
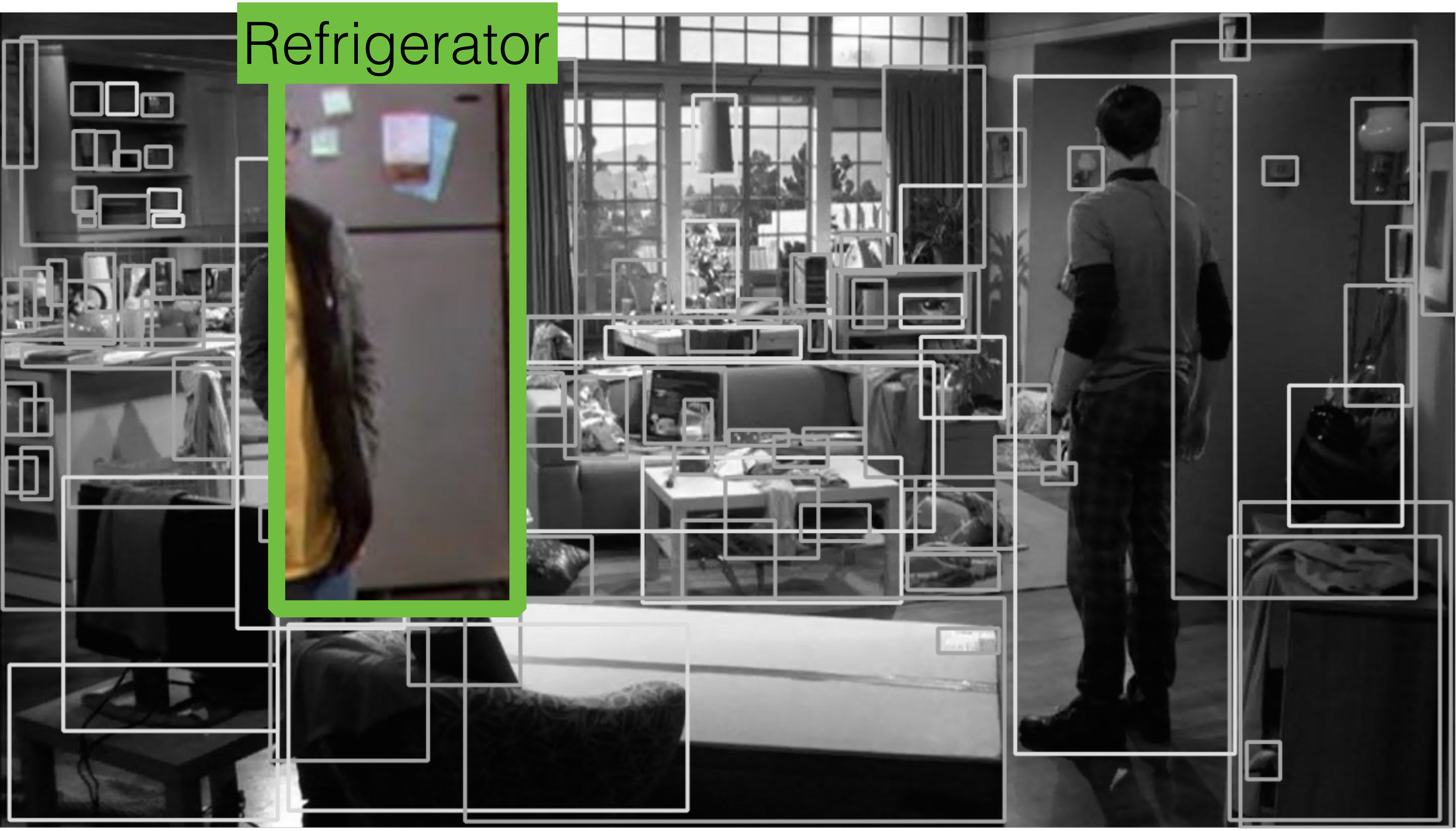
Scaling up object detection

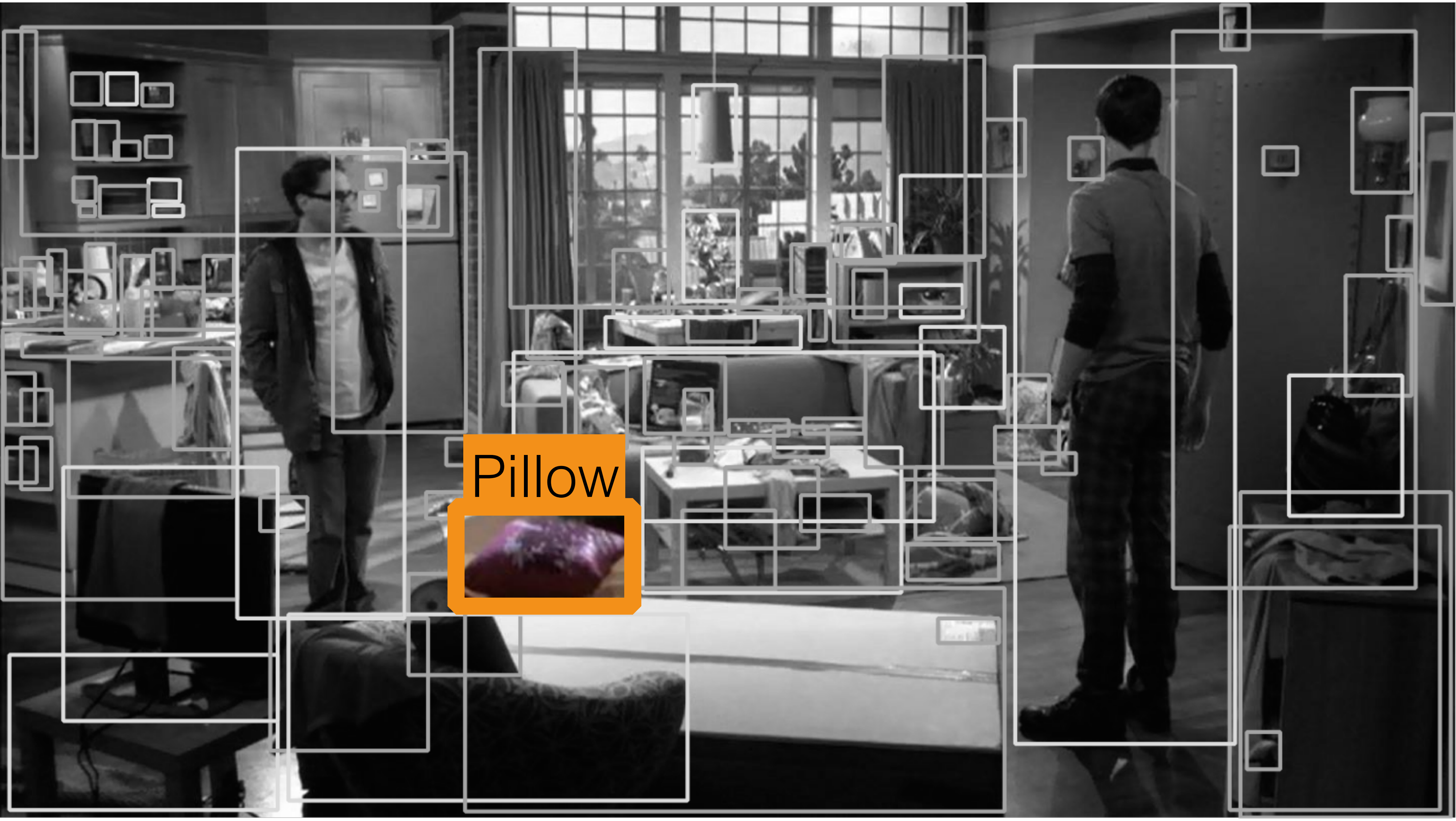
Olga Russakovsky
CMU

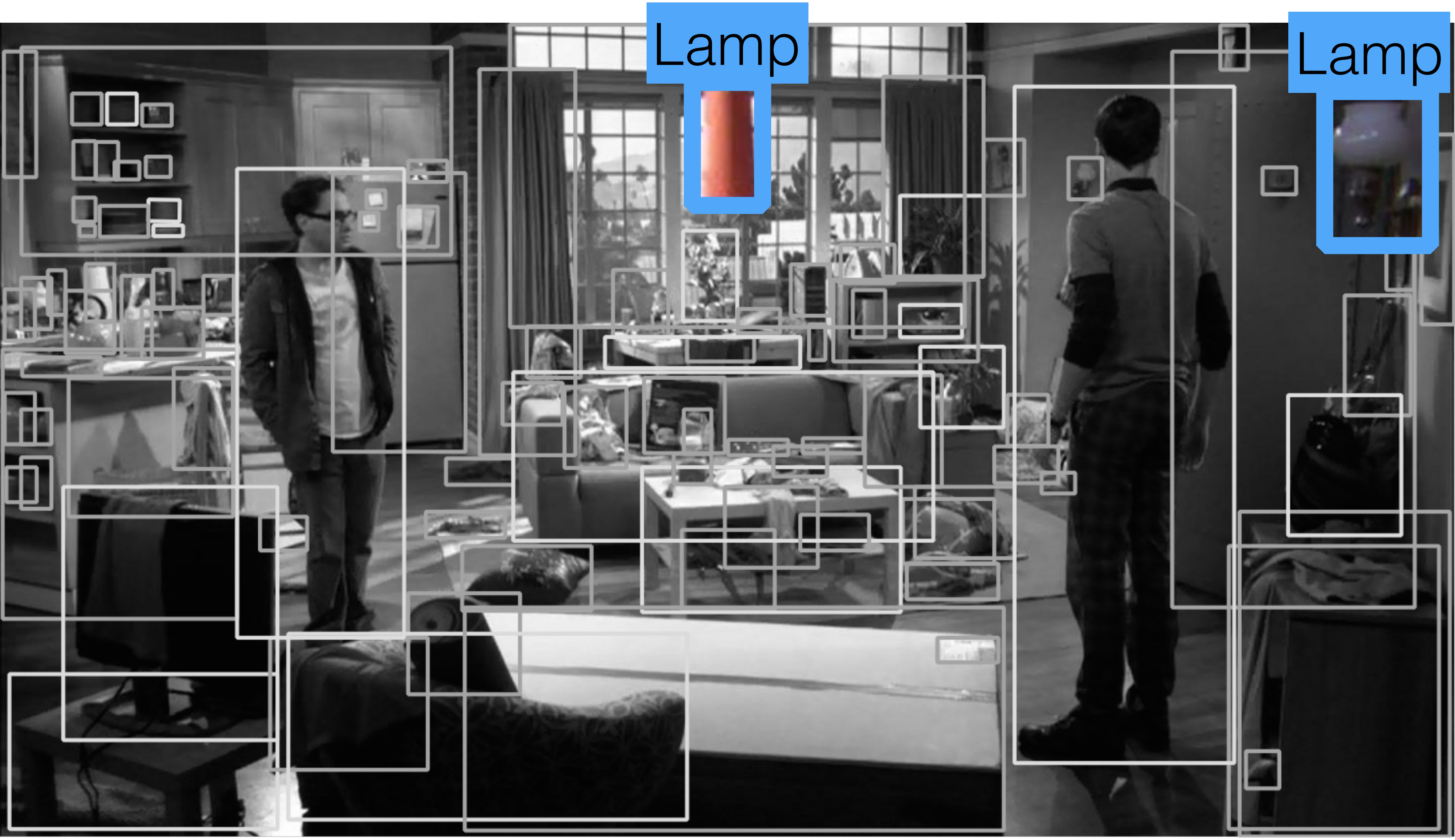
Stanford University

Carnegie Mellon University



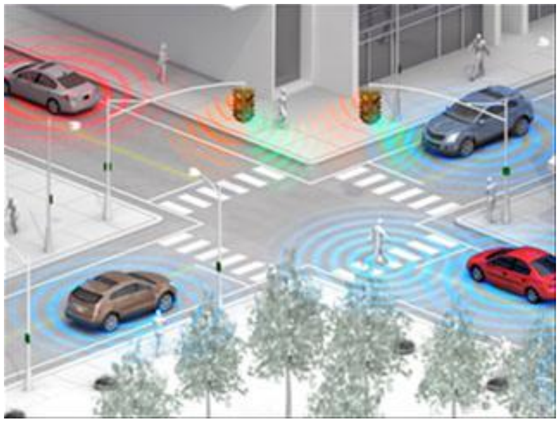

Refrigerator


Pillow

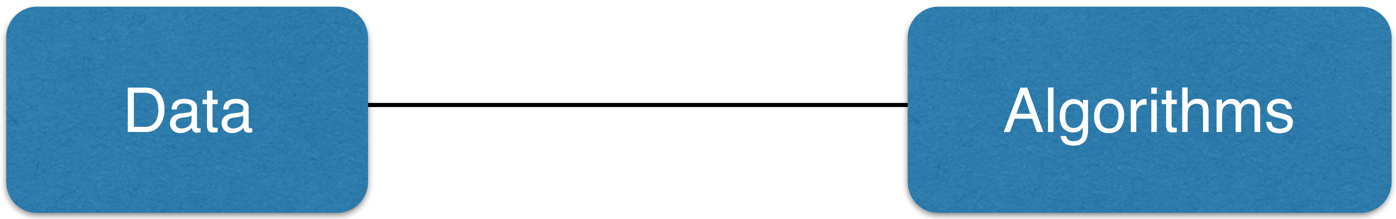

Lamp
Lamp

Scaling up object detection

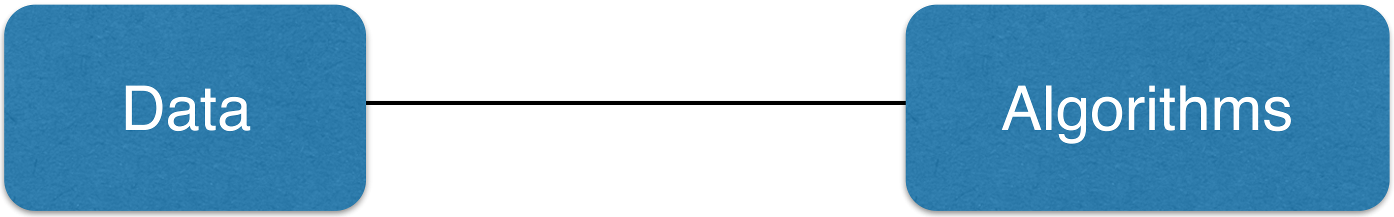
Why scale up object detection?



How to scale up object detection?

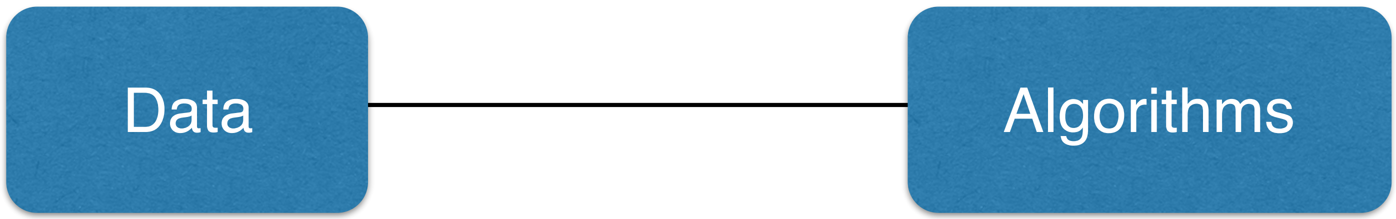


How to scale up object detection?



Traditionally, computer vision mostly focused on algorithms

How to scale up object detection?



Traditionally, computer vision mostly focused on algorithms

I claim data is **at least** as important

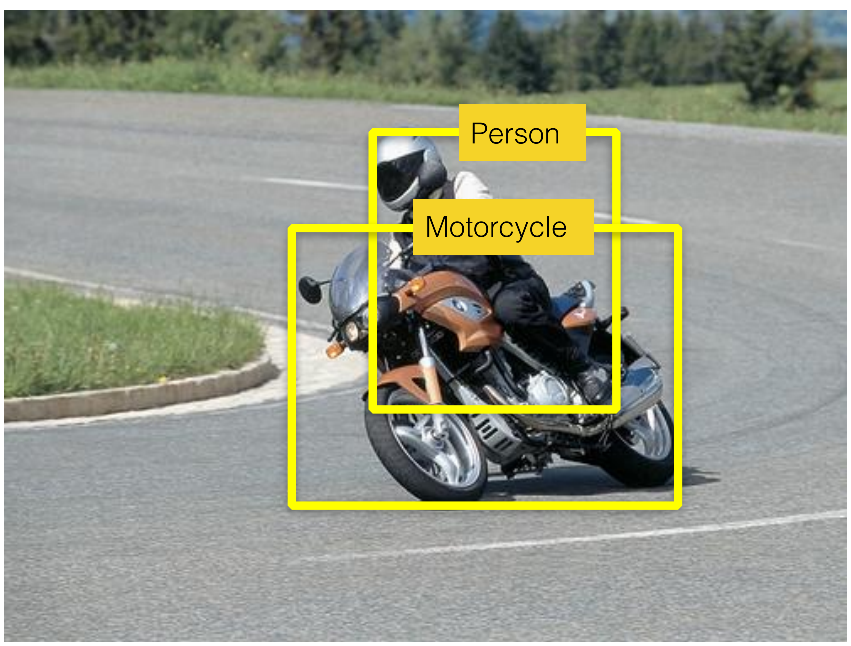
Year 2012

State of object detection

Data — Algorithms

PASCAL VOC

20 object classes 22,591 images



[Everingham et al. IJCV 2010]

Viola-Jones 01,
Fergus 03
Torralba 04,
Dalal-Triggs 05,
Chum 07,
Lampert 08,
Gall 09,
Maji 09,
Harzallah 09,
Felzenszwalb 10,
vanDeSande 11,
Song 11,
Malisiewicz 11,
…

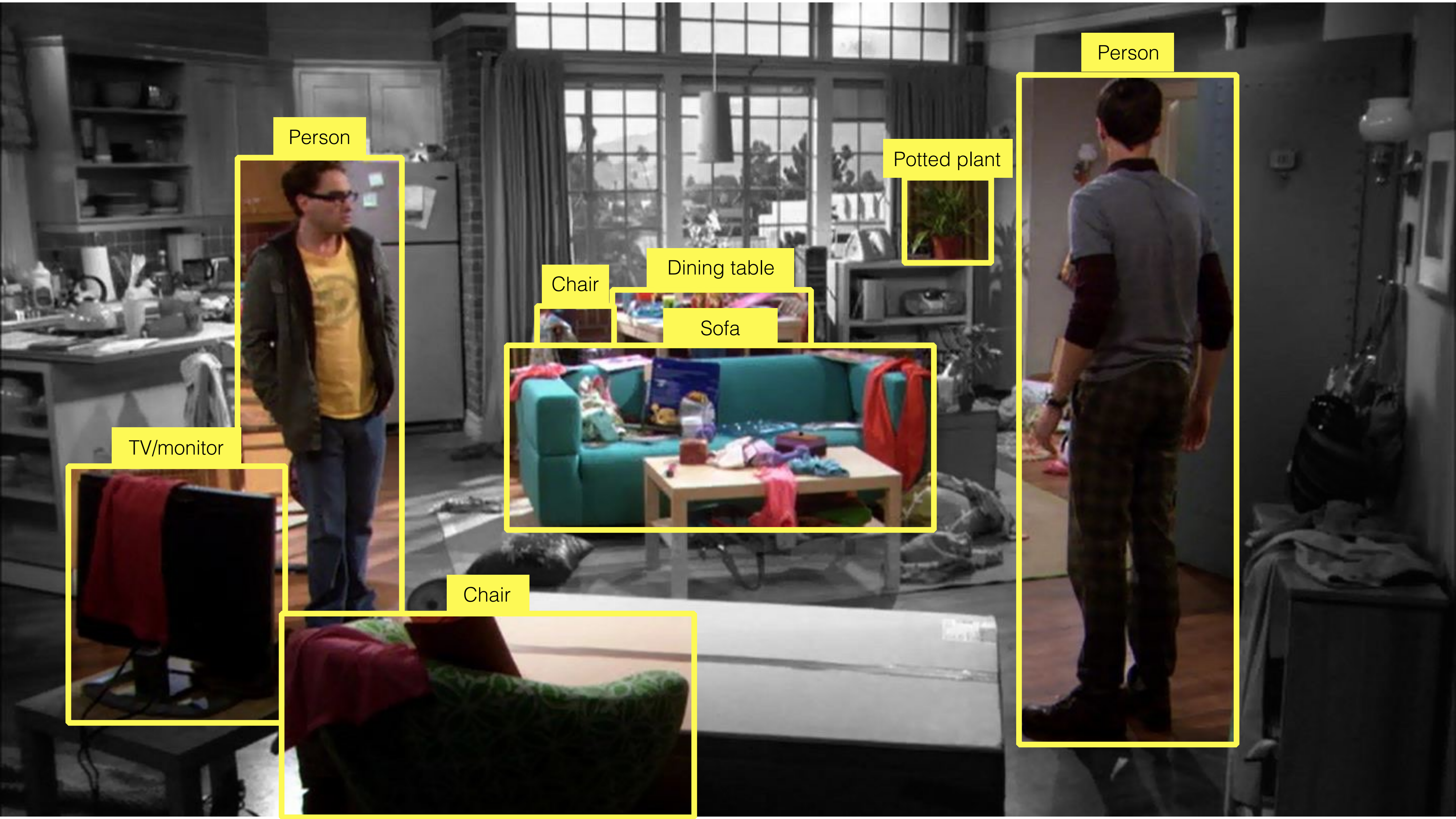
Year 2012

“State-of-the-art” results

[DPM, Felzenszwalb 2010]

Year 2012

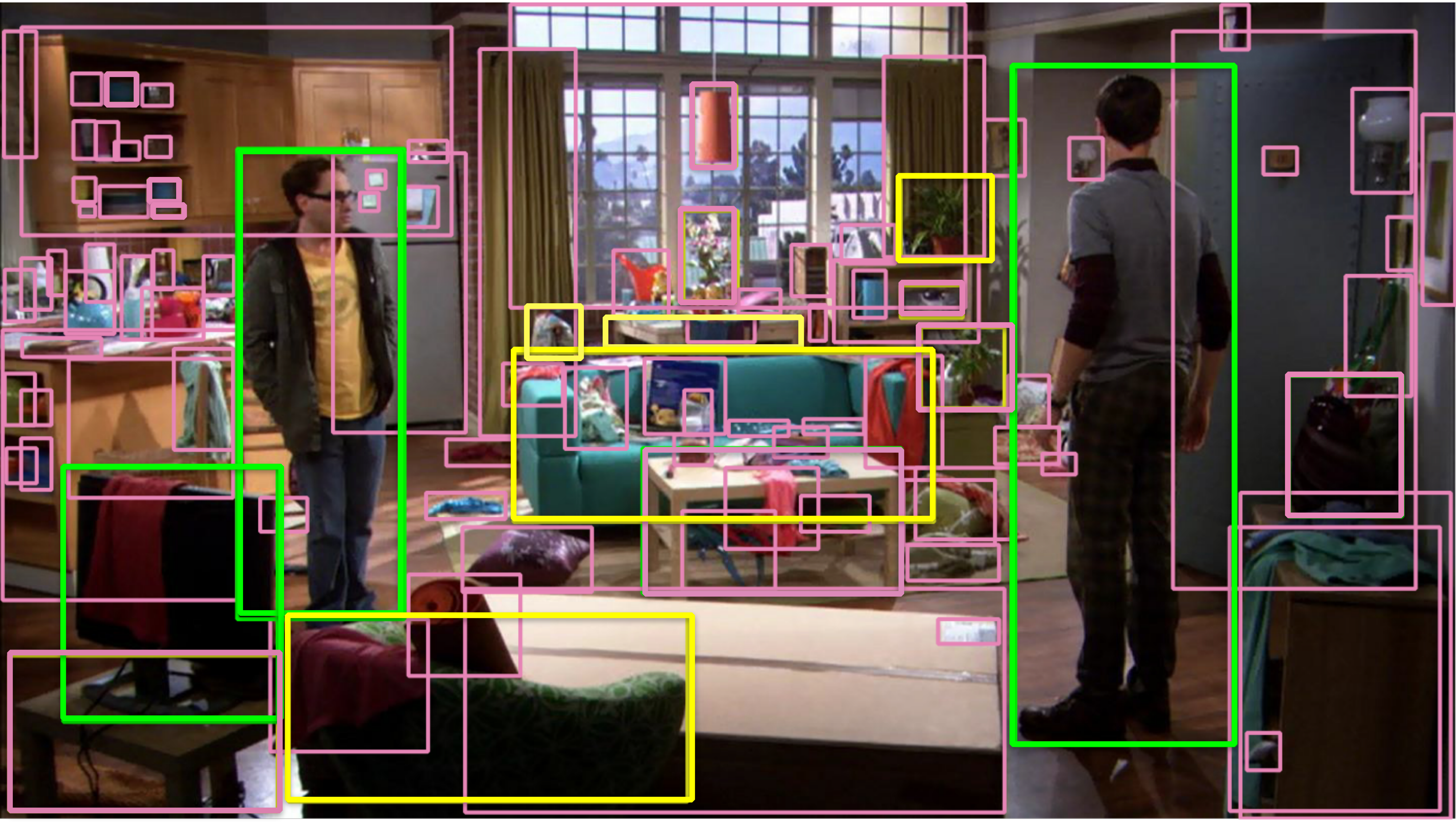
Upper bound given available data



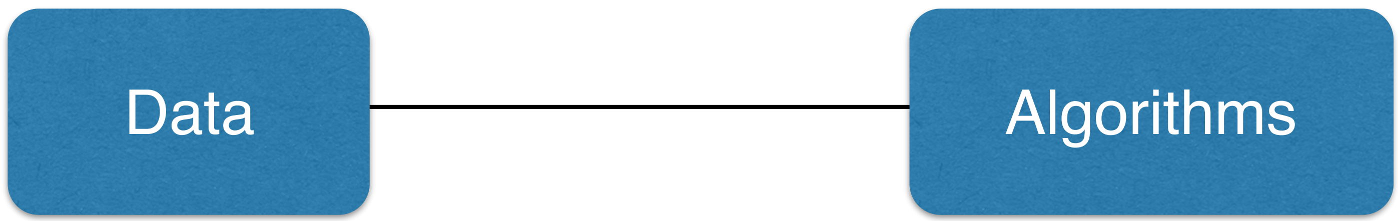
[Objects from PASCAL VOC, Everingham 10

Year 2012

Nowhere near this…



Scaling up object detection

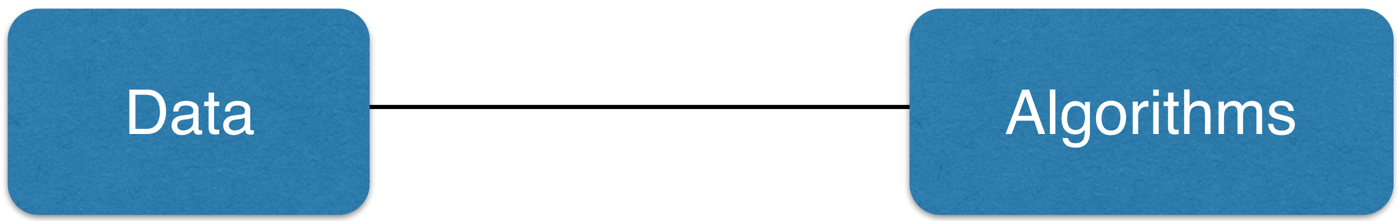


1) Scale up the data

2) Develop and analyze
the algorithms

3) Combine insights from both

Scaling up object detection



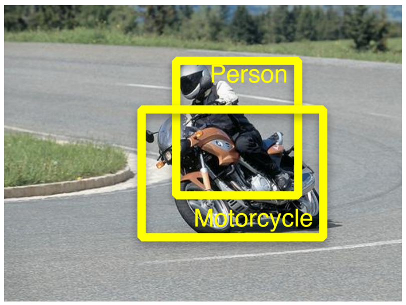
1) Scale up the data

2) Develop and analyze the algorithms

3) Combine insights from both

Scaling up the data

PASCAL VOC
20 classes
22,591 images



[Everingham10]

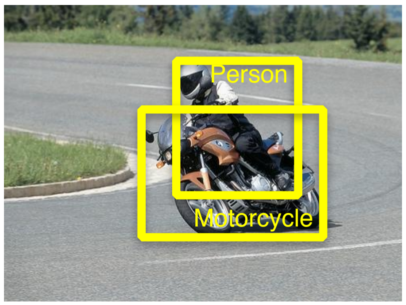
Detail of annotation

Scale of dataset

Scaling up the data

Detail of annotation

PASCAL VOC
20 classes
22,591 images



[Everingham10]

ImageNet
21,841 classes
14,197,122 images

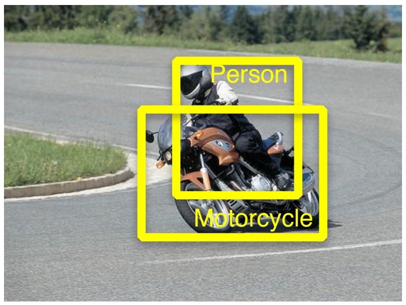
Dalmatian

[Deng09]

Scale of dataset

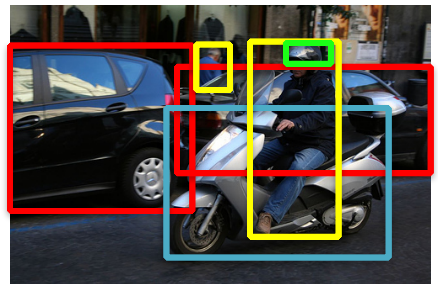
Scaling up the data

PASCAL VOC
20 classes
22,591 images



[Everingham10]

ILSVRC object detection
200 object classes
120,931 images



[CHI14, IJCV15]

ImageNet
21,841 classes
14,197,122 images

Dalmatian

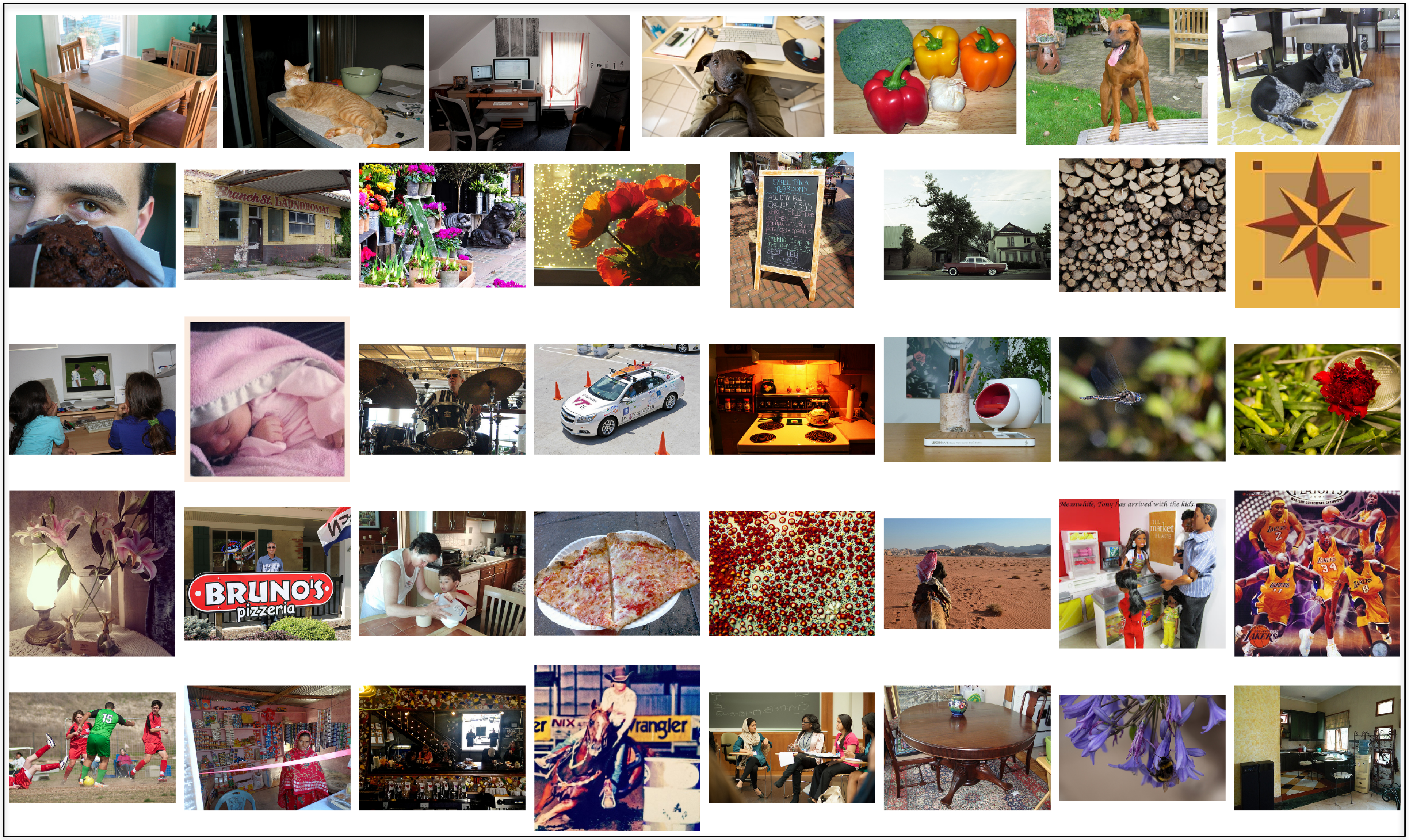
[Deng09]

Detail of annotation

Scale of dataset

ILSVRC object detection annotation

Step 1: Image collection



O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

ILSVRC object detection annotation

Step 1: Image collection

O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

ILSVRC object detection annotation

Step 1: Image collection

Step 2: Annotation

?

powered by

amazon mechanical turk™
Artificial Artificial Intelligence

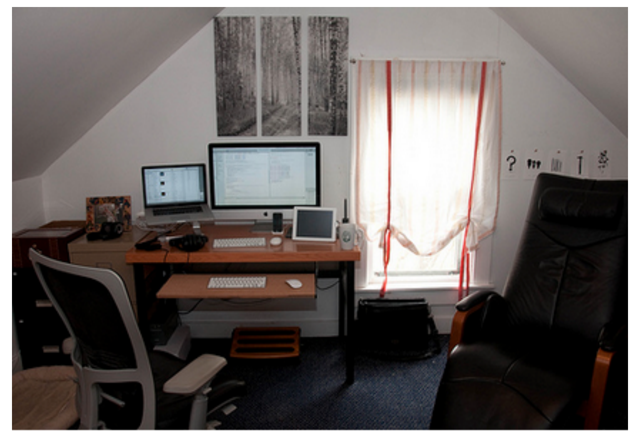
O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

ILSVRC object detection annotation

Step 1: Image collection

Attempt a)

Draw bounding boxes around all objects and name them



Step 2: Annotation

?

powered by

amazon mechanicalturk™
Artificial Artificial Intelligence

O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

ILSVRC object detection annotation

Step 1: Image collection

Attempt a)

Draw bounding boxes around all objects and name them

difficult to use the data

Step 2: Annotation

?

powered by

amazon mechanical turk™
Artificial Artificial Intelligence

O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

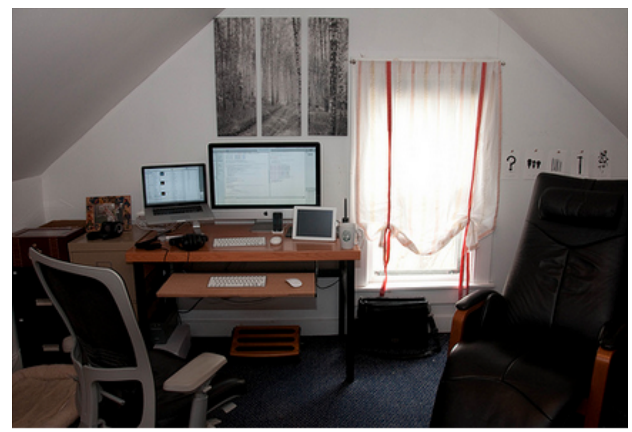
ILSVRC object detection annotation

Step 1: Image collection

Attempt b)

Draw bounding boxes around all instances of:

accordion, airplane, ant, antelope, apple, armadillo, artichoke, axe, baby bed, … zebra



Step 2: Annotation

?

powered by

amazon mechanical turk™
Artificial Artificial Intelligence

O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

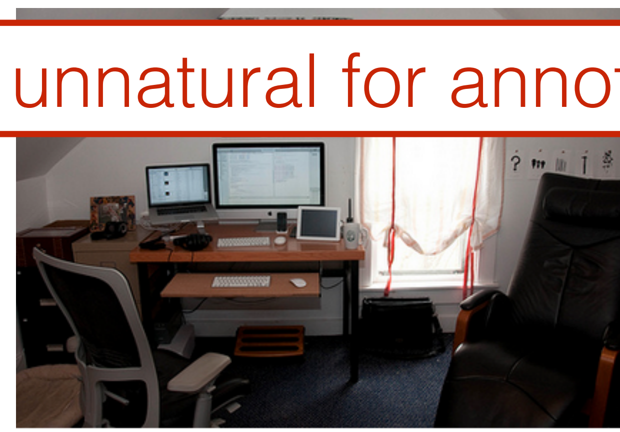
ILSVRC object detection annotation

Step 1: Image collection

Attempt b)

Draw bounding boxes around all instances of:

accordion, airplane, ant, antelope, apple, armadillo, artichoke, axe, baby bed, … zebra



very unnatural for annotators

Step 2: Annotation

?

powered by

amazon mechanical turk™

Artificial Artificial Intelligence

O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

ILSVRC object detection annotation

Step 1: Image collection

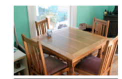
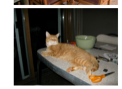
Step 2: Annotation

Decompose into
short, focused tasks

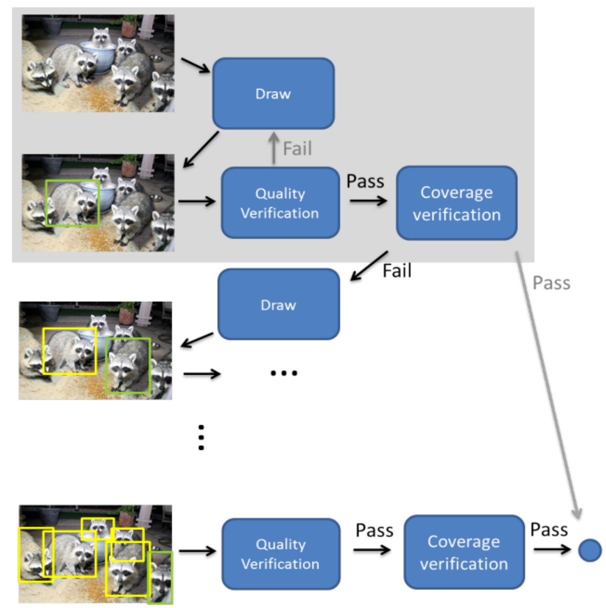
powered by

amazon mechanical turk™
Artificial Artificial Intelligence

Step 2a: Binary annotation

Input \ Labels	Table	Chair	Bowl	Dog	Cat	...
	+	+	-	-	-	-
	+	-	+	-	+	-

Step 2b: Location annotation



O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

ILSVRC object detection annotation

Step 1: Image collection

Step 2: Annotation

Decompose into
short, focused tasks

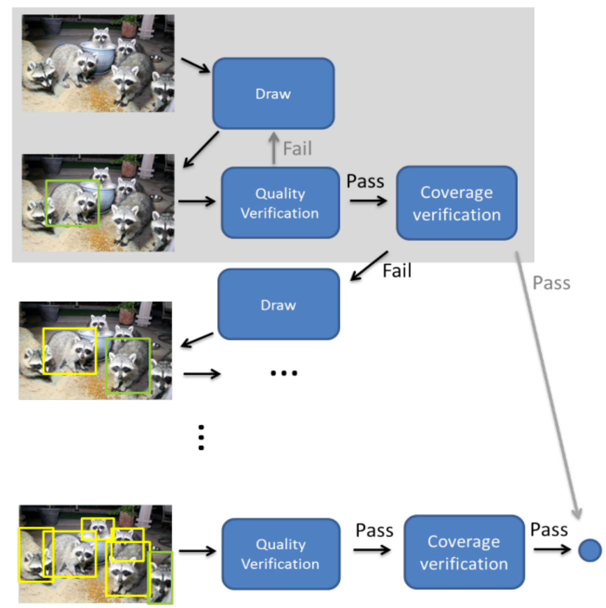
powered by

amazon mechanical turk™
Artificial Artificial Intelligence

Step 2a: Binary annotation

Input \ Labels	Table	Chair	Bowl	Dog	Cat	...
	+	+	-	-	-	-
	+	-	+	-	+	-

Step 2b: Location annotation



O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

Scale of ILSVRC detection annotation

≈

Scale of IMAGENET annotation

Scale of ILSVRC detection annotation
≈
Scale of IMAGENET annotation

ImageNet:
14M images x 1 class/image = **14M** binary questions

Scale of ILSVRC detection annotation
≈
Scale of IMAGENET annotation

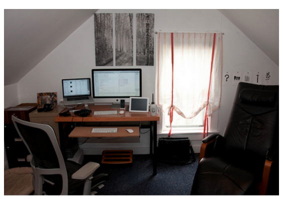
ImageNet:
14M images x 1 class/image = **14M** binary questions

ILSVRC detection:
120K images x 200 classes/image = **24M** binary questions

Multi-label annotation

Labels

Input

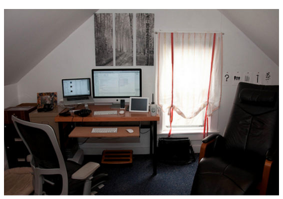
	Table	Chair	Bowl	Dog	Cat	...
						
						
						
						

(200 objects)

(120K images)

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Multi-label annotation

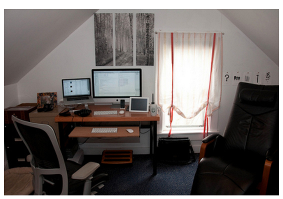
Labels / Input	Table	Chair	Bowl	Dog	Cat	...
	+	+	-	-	-	-
	+	-	+	-	+	-
	+	+	-	-	-	-
	-	-	-	+	-	-

(200 objects)

(120K images)

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Multi-label annotation

Input \ Labels	Table	Chair	Bowl	Dog	Cat	...
	+	+	-	-	-	-
	+	-	+	-	+	-
	+	+	-	-	-	-
	-	-	-	+	-	-

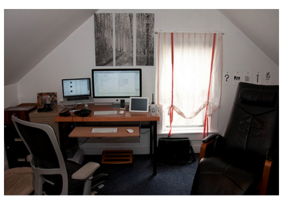
(120K images)

(200 objects)

Label sparsity

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Multi-label annotation

Labels / Input	Table	Chair	Bowl	Dog	Cat	...
	+	+	-	-	-	-
	+	-	+	-	+	-
	+	+	-	-	-	-
	-	-	-	+	-	-

(200 objects)

(120K images)

Label correlation

Label sparsity

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Man-made objects

Furniture

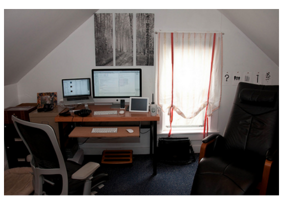
Animals

Label hierarchy

Labels

Input

(200 objects)

	Table	Chair	Bowl	Dog	Cat	...
	+	+	-	-	-	-
	+	-	+	-	+	-
	+	+	-	-	-	-
	-	-	-	+	-	-

Label sparsity

(120K images)

Label correlation

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Selecting the Right Question

Goal:

Get as much utility (new labels) as possible,
for as little cost (worker time) as possible,
given a desired level of accuracy

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Selecting the Right Question

Goal: $\boxed{U(Q) = \mathbf{E}\|y\|_1}$

Get as much **utility** (new labels) as possible,

for as little cost (worker time) as possible,

given a desired level of accuracy

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Selecting the Right Question

Goal: $$\boxed{U(Q) = \mathbf{E}\|y\|_1}$$

Get as much utility (new labels) as possible,
for as little **cost** (worker time) as possible,
given a desired level of accuracy

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Selecting the Right Question

Goal: $\boxed{U(Q) = \mathbf{E}\|y\|_1}$

Get as much utility (new labels) as possible,
for as little cost (worker time) as possible,
given a desired level of **accuracy**

Number of workers:

$$\min\{n : \sum_{i=n+1}^{2n+1} \binom{2n+1}{i} p^i (1-p)^{2n+1-i} > 1-\epsilon\}$$

$$1-\epsilon = \text{acceptable accuracy}$$

$$p = \text{worker accuracy}$$

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Multi-label annotation can be efficient

- Dataset:
  - 20K images from ILSVRC2013, split evenly into train/test
  - 200 classes (dog, table, …)
  - 64 internal nodes in hierarchy
- Baseline: Naïve approach

J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

Multi-label annotation can be efficient

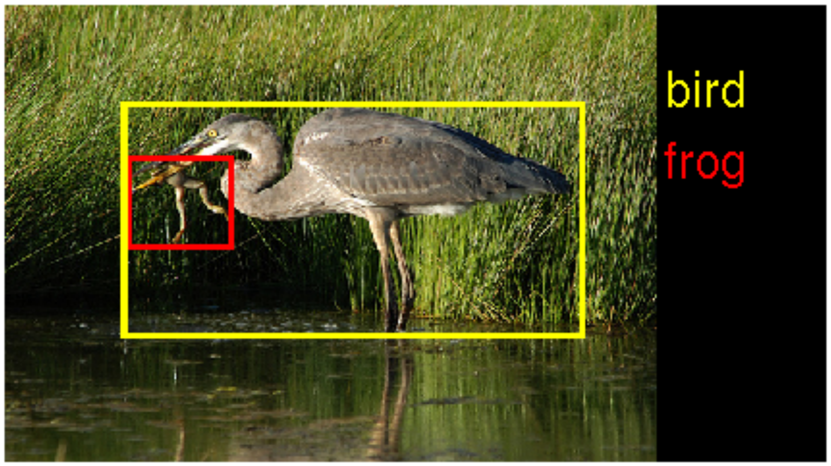
- Dataset:
  - 20K images from ILSVRC2013, split evenly into train/test
  - 200 classes (dog, table, …)
  - 64 internal nodes in hierarchy
- Baseline: Naïve approach

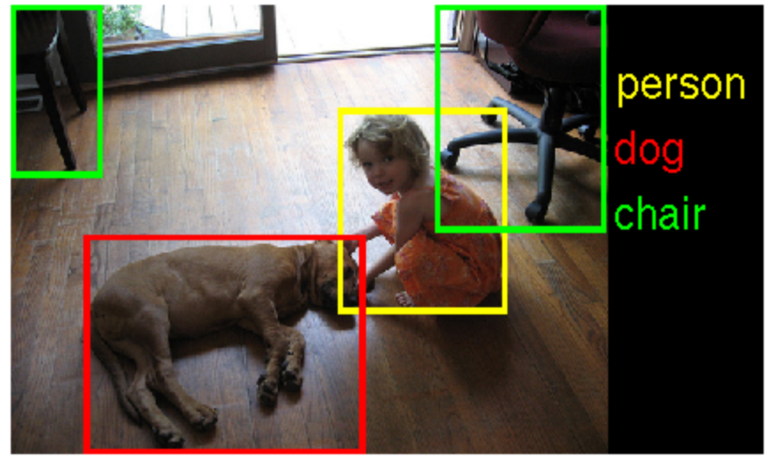
- **Result: 6.2x savings in human annotation time**

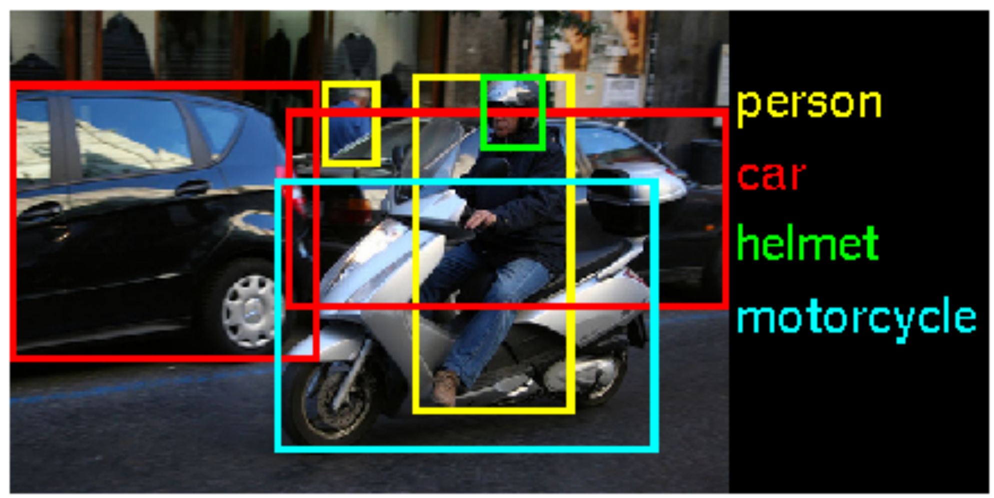
J Deng, O. Russakovsky et al. Scalable multi-label annotation. CHI, 2014

ILSVRC object detection data

200 object classes, 120,931 images







O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

Impact of ILSVRC

The New York Times

NUS National University of Singapore

Sun Yat-Sen University

The University of California UCLA

NYU

facebook

CloudCV: Object Detection

Tsing Hua University

University of Amsterdam

EUVISION technologies

MIT Technology Review

Toyota Technological Institute Chicago

Centrale Paris

Clarifai

Google

Microsoft Research

Andrew Howard Consulting

IBM Research

orange™

University of Oxford

CRIPAC Center for Research on Intelligent Perception and Computing

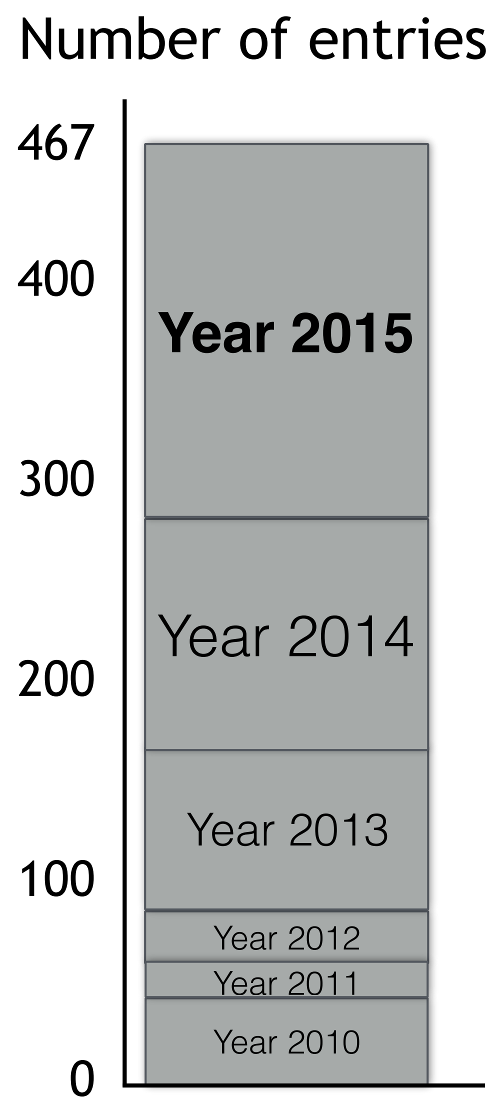
NEC Laboratories America

Relentless passion for innovation

TRIMPS

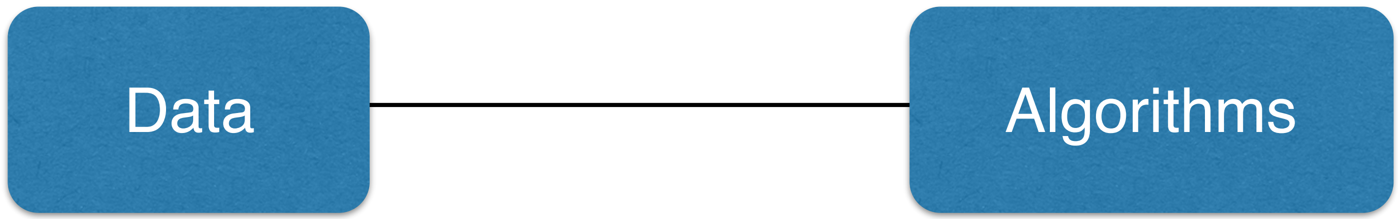
Decaf / Caffe

a Berkeley Vision Project



O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

Scaling up object detection

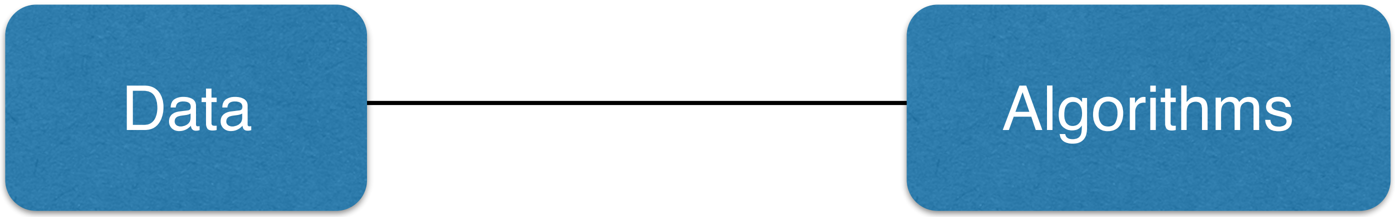


1) Scale up the data

2) Develop and analyze
the algorithms

3) Combine insights from both

Scaling up object detection

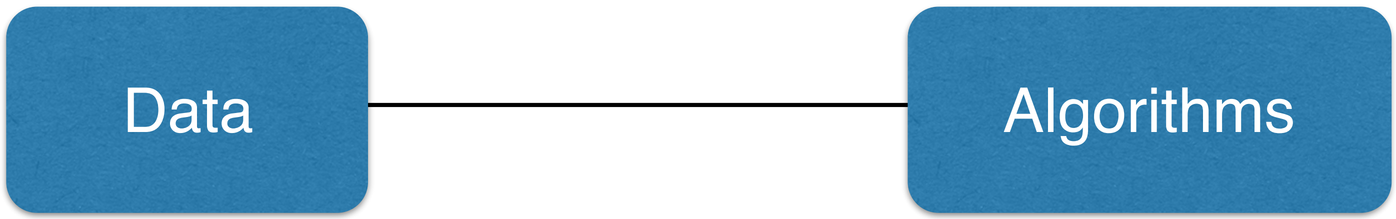


1) Scaled up the data by formulating data annotation as an optimization [CHI14, IJCV15]

2) Develop and analyze the algorithms

3) Combine insights from both

Scaling up object detection



1) **Scaled up the data** by formulating data annotation as an optimization [CHI14, IJCV15]

2) Develop and analyze the algorithms

3) Combine insights from both

Sidebar

Some object detection algorithmic work

Improving efficiency

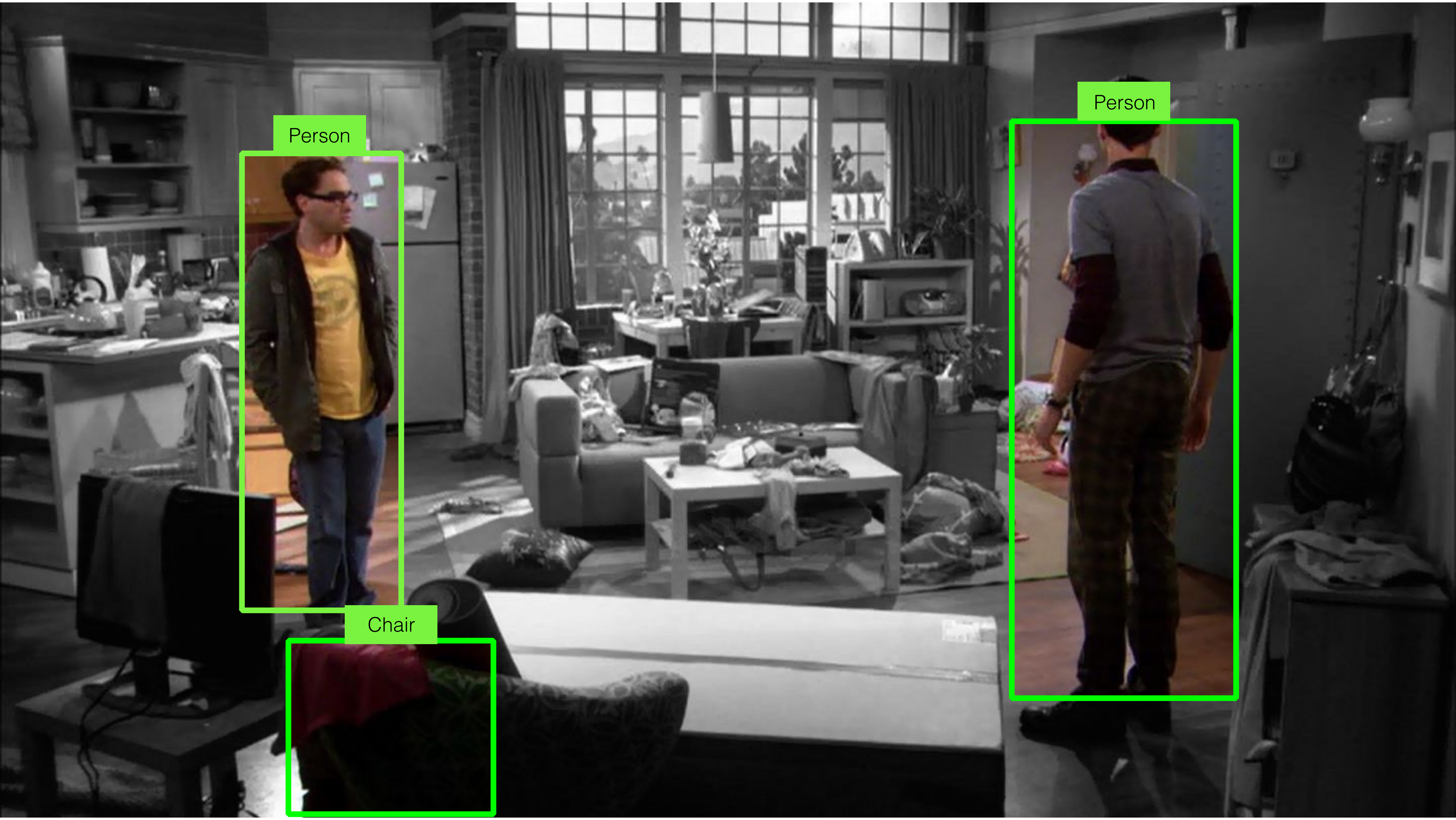
- Russakovsky and Ng. CVPR10

Improving accuracy

- Klingbeil, Carpenter, Russakovsky, Ng. ICRA10
- Russakovsky, Lin, Yu, Fei-Fei. ECCV12
- Modolo, Vezhnevets, Russakovsky, Ferrari. CVPR15

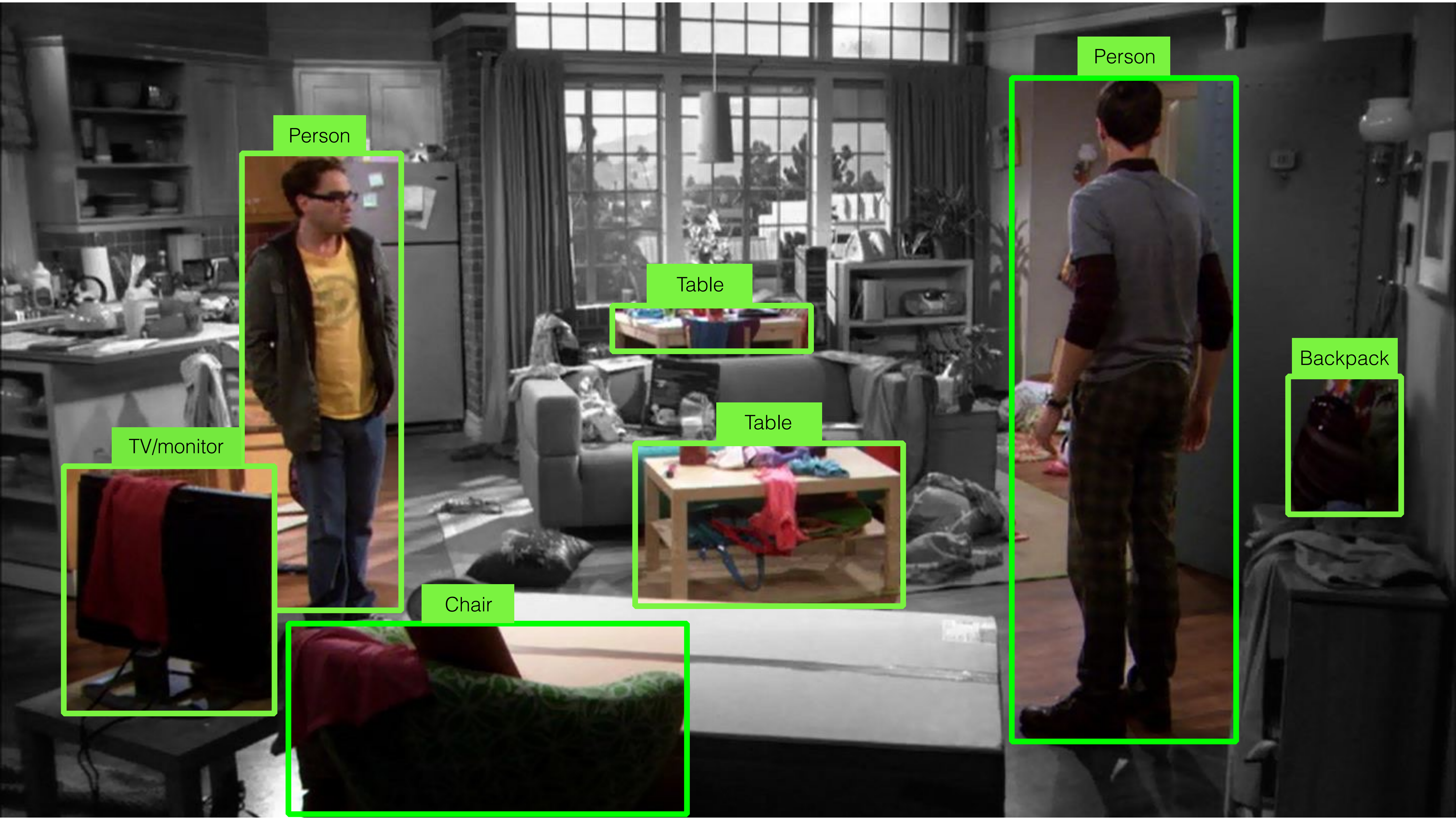
Let’s come back to this image:

“State-of-the-art” results in 2012



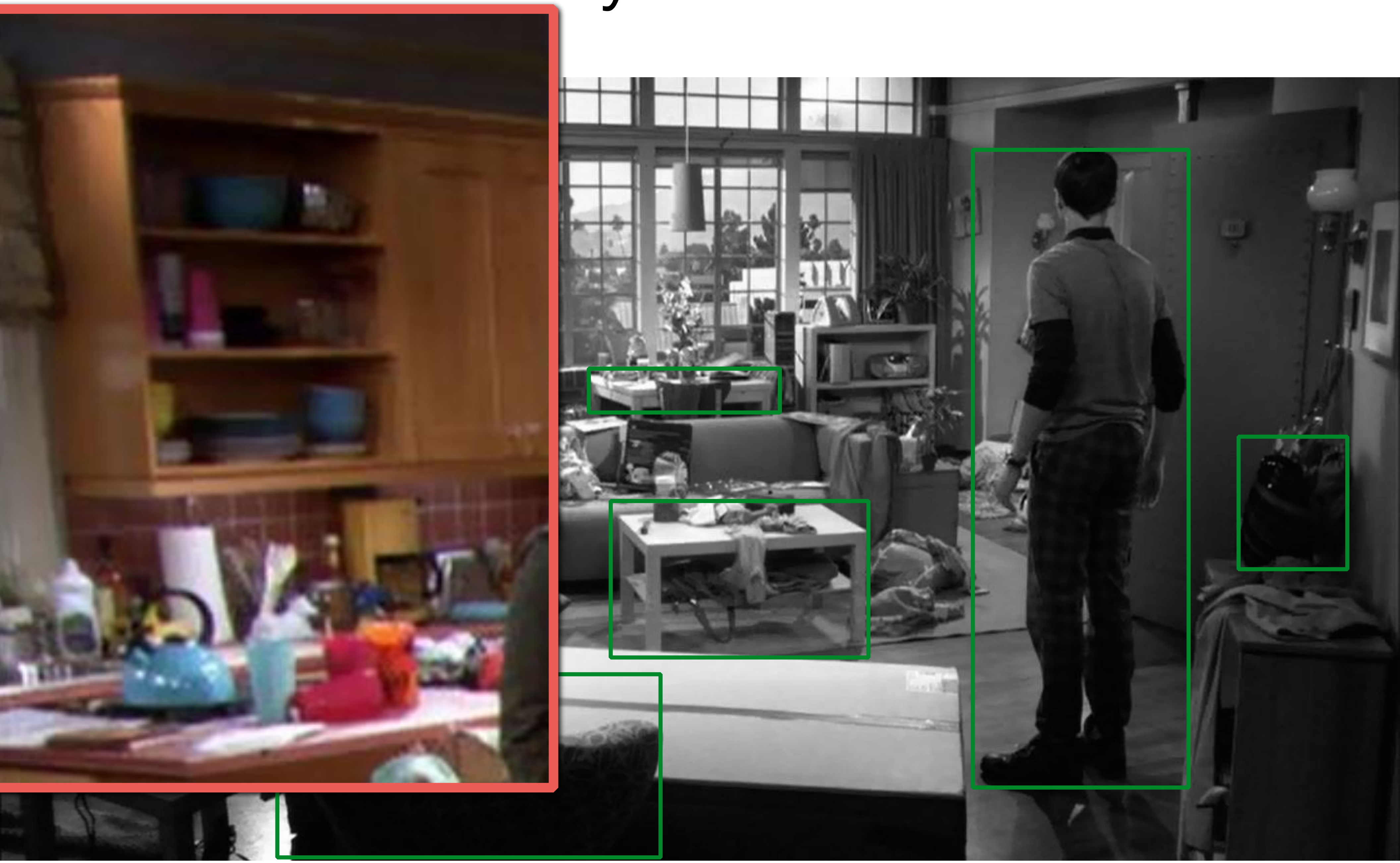
[DPM, Felzenszwalb 2010]

“State-of-the-art” results in 2014



[RCNN, Girshick 2014]

But why not better?



Easiest and hardest classes

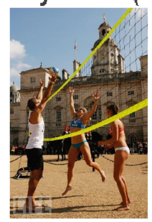
(Highest average precision in percent of any method in ILSVRC 2013-2014)

Easiest

butterfly (93)

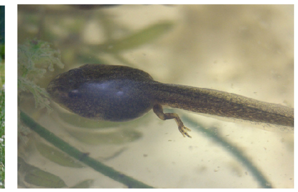
dog (84)

volleyball (83)



rabbit (83)

frog (82)



basketball (80)

snowplow (80)

bird (78)

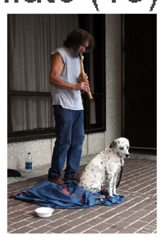
tiger (77)

zebra (77)

Hardest

lamp (15)

flute (15)



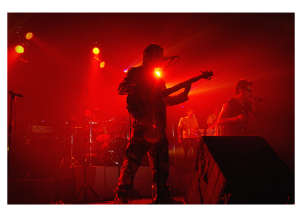
horizontal bar (14)

spatula (13)

nail (13)

ski (12)

microphone (11)



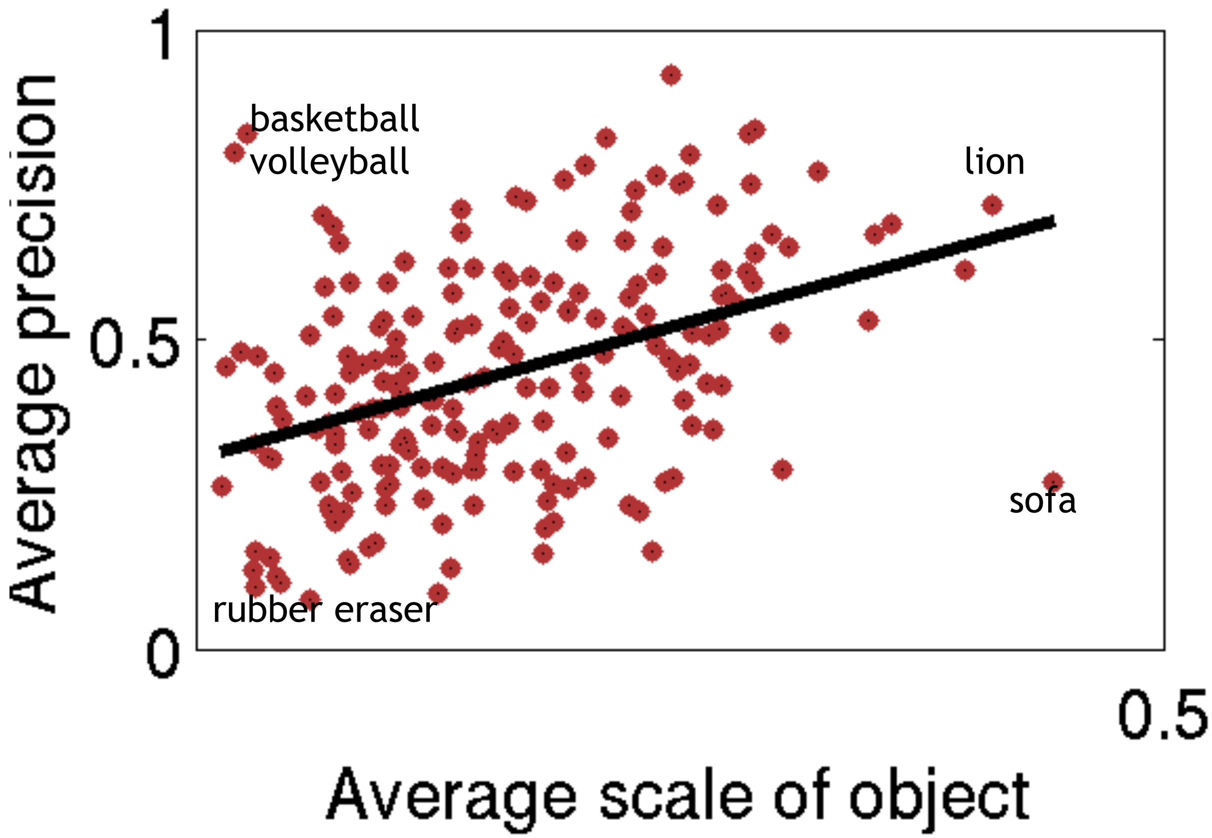
rubber eraser (10)

ladle (9)

backpack (8)

O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

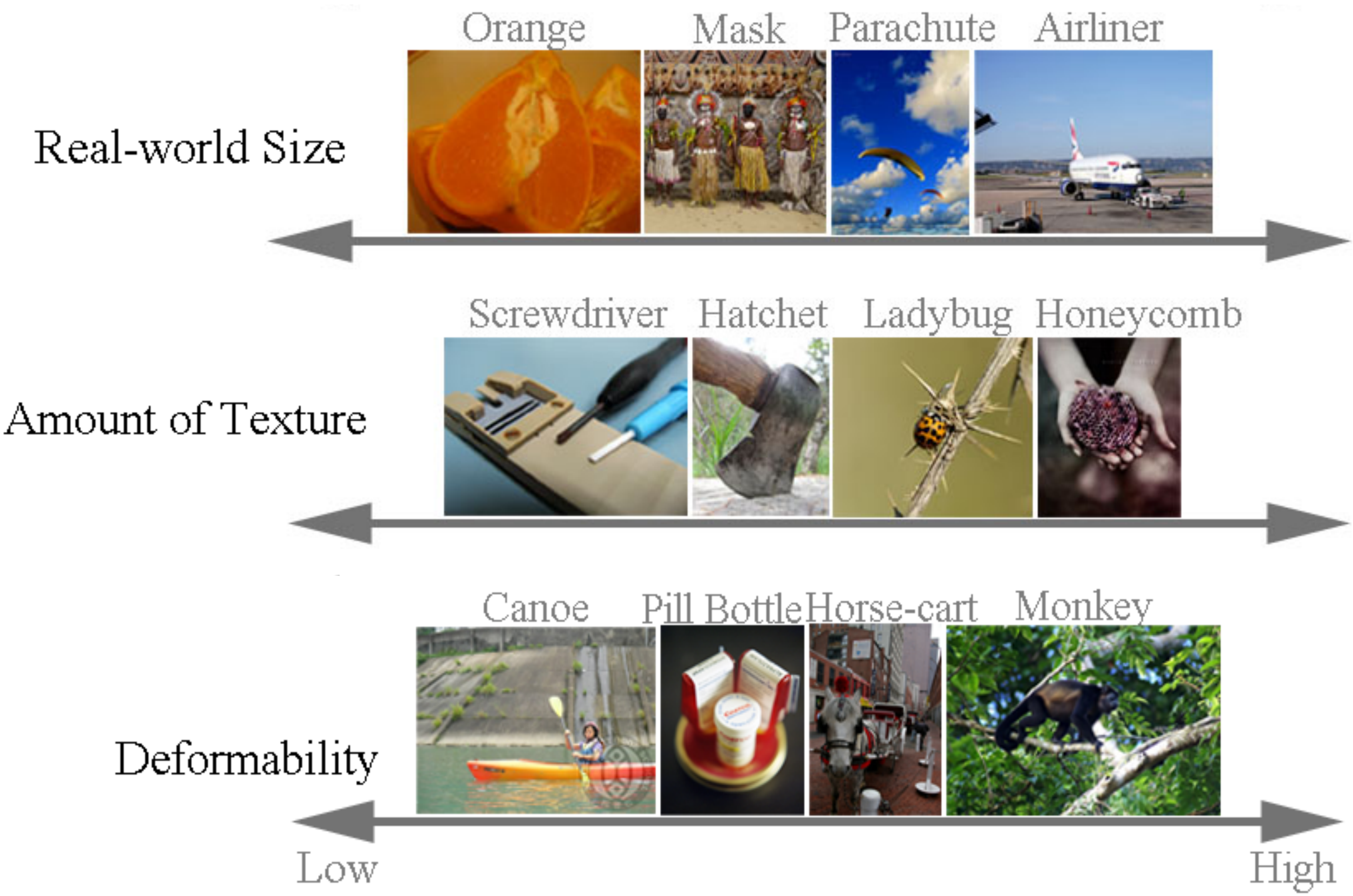
Object detection results per-class



- Each dot is an object class
- X-axis: average fraction of image area occupied by an instance of that class on the validation set
- Y-axis: highest average precision achieved by any method in ILSVRC2013 and ILSVRC2014

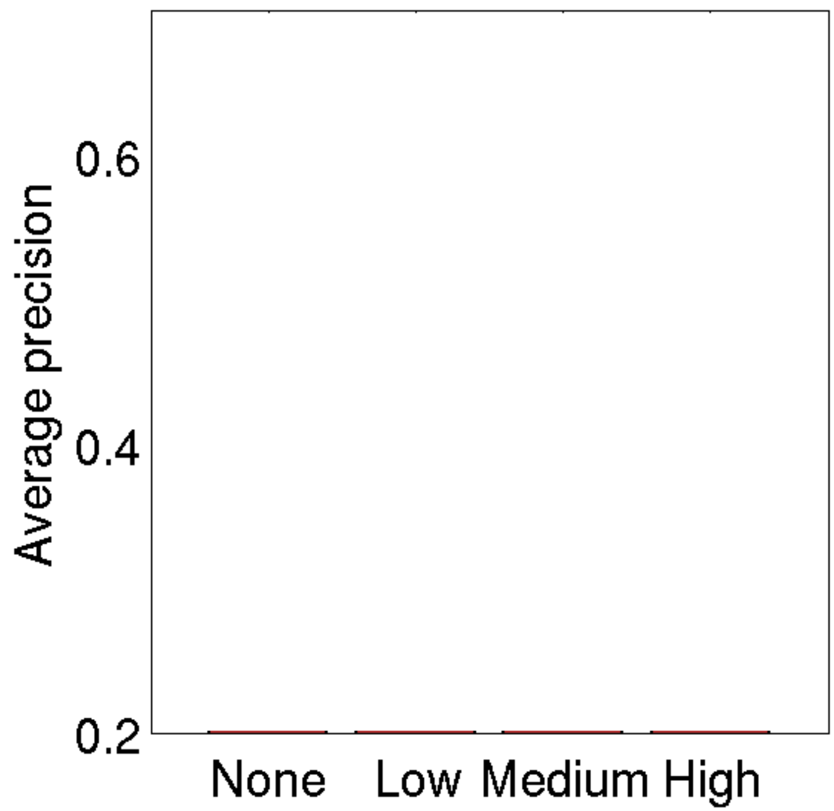
O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

Variety of object classes in ILSVRC



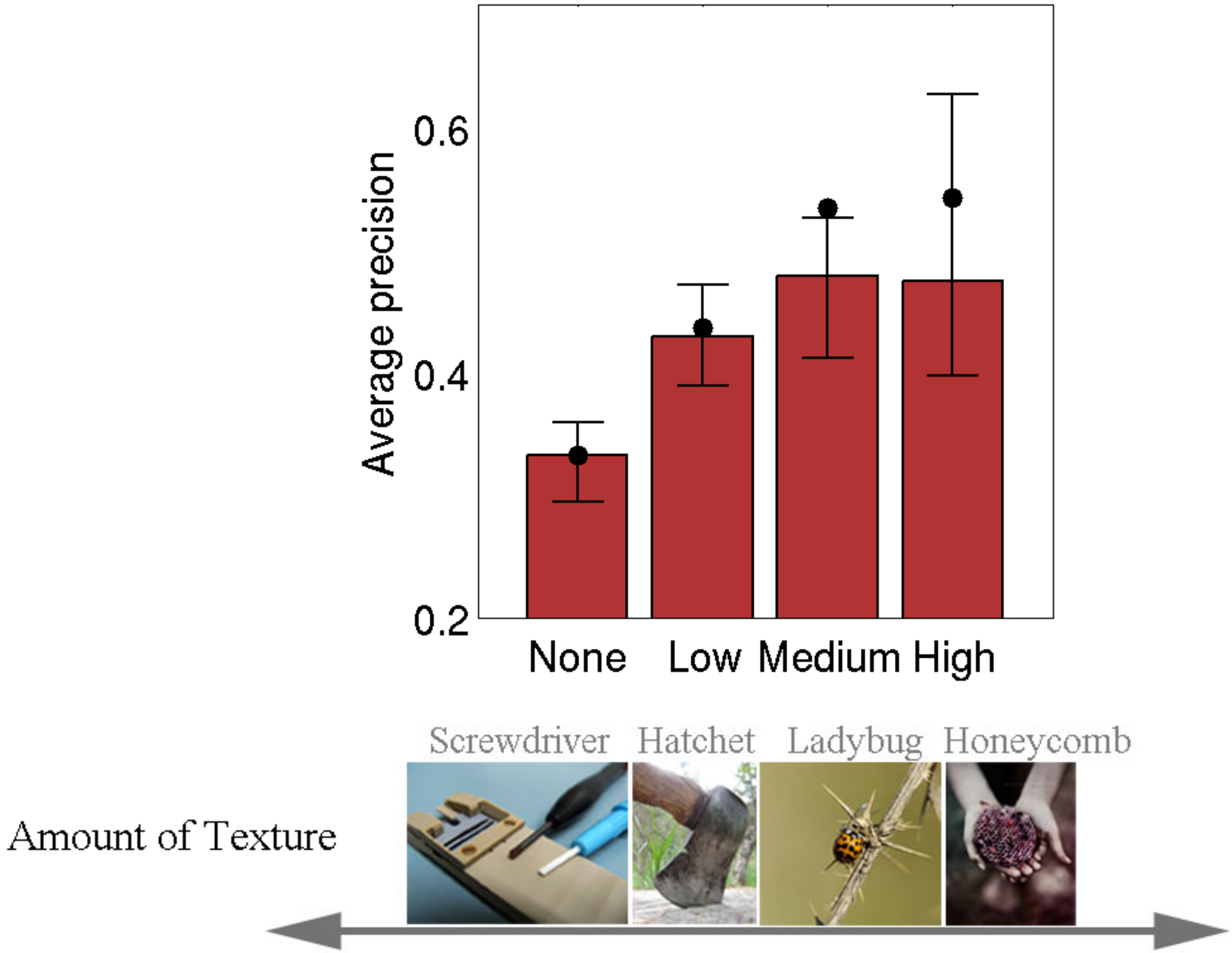
O Russakovsky et al. Detecting avocados to zucchinis: what have we done, and where are we going? ICCV 2013.

Impact of object texture



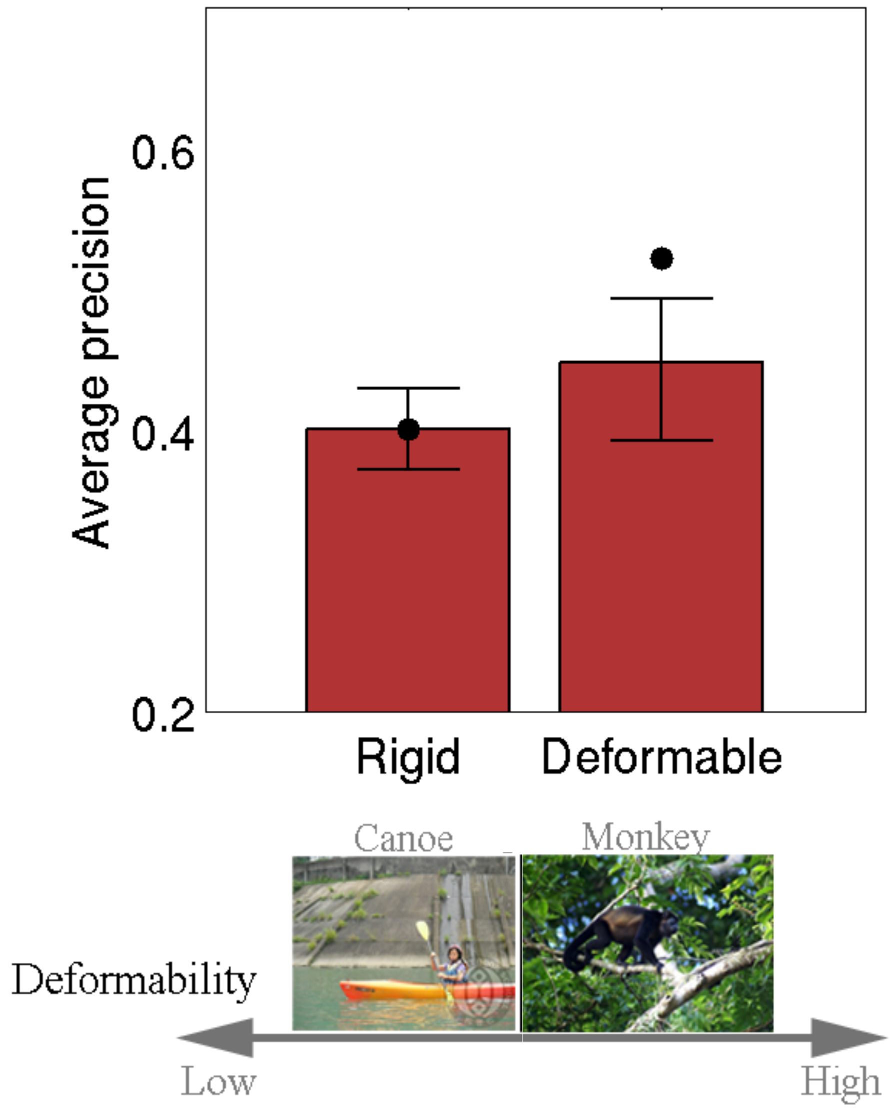
O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

Textured objects are easier



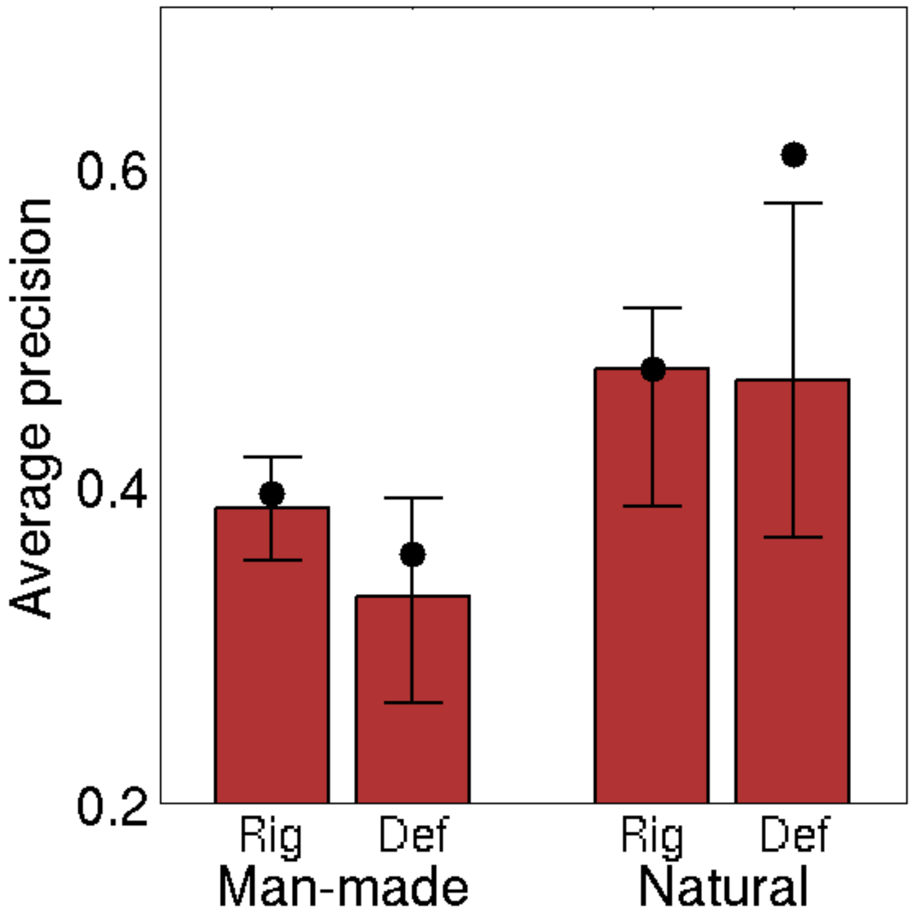
O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

Deformable objects are easier (?!)



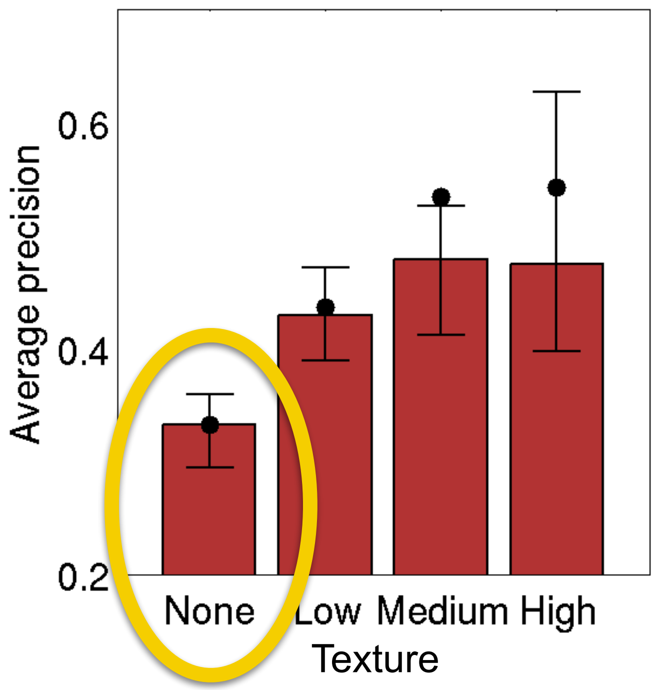
O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

Actually, **natural** objects are easier



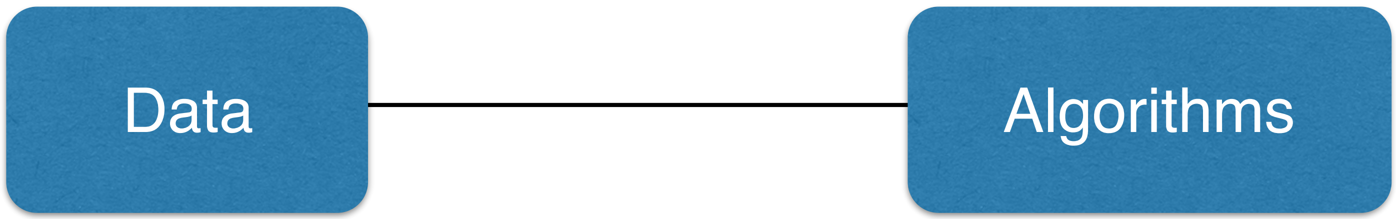
O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

Next frontier: untextured, man-made objects?



O Russakovsky* and J Deng* et al. ImageNet Large Scale Visual Recognition Challenge. IJCV 2015.

Scaling up object detection

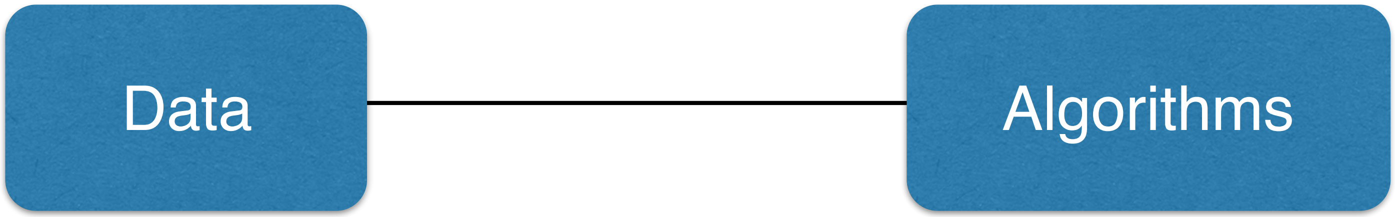


1) **Scaled up the data** by formulating data annotation as an optimization [CHI14, IJCV15]

2) Develop and analyze the algorithms

3) Combine insights from both

Scaling up object detection

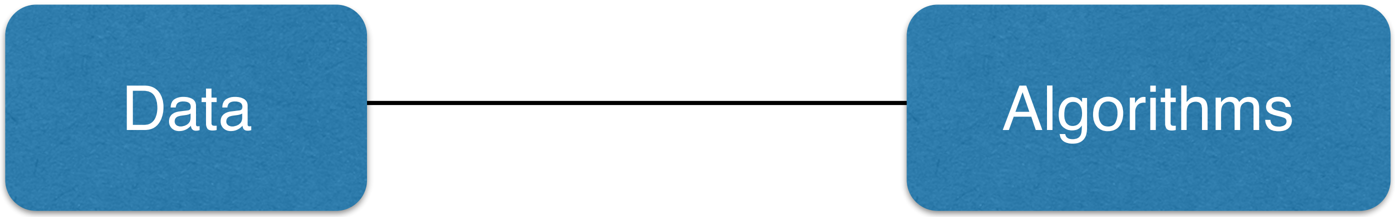


1) **Scaled up the data** by formulating data annotation as an optimization [CHI14, IJCV15]

2) **Developed algorithms** [CVPR10, ECCV12,CVPR15b] **and performed large-scale analysis to gain insight into the state of the field** [ICCV13, IJCV15]

3) Combine insights from both

Scaling up object detection

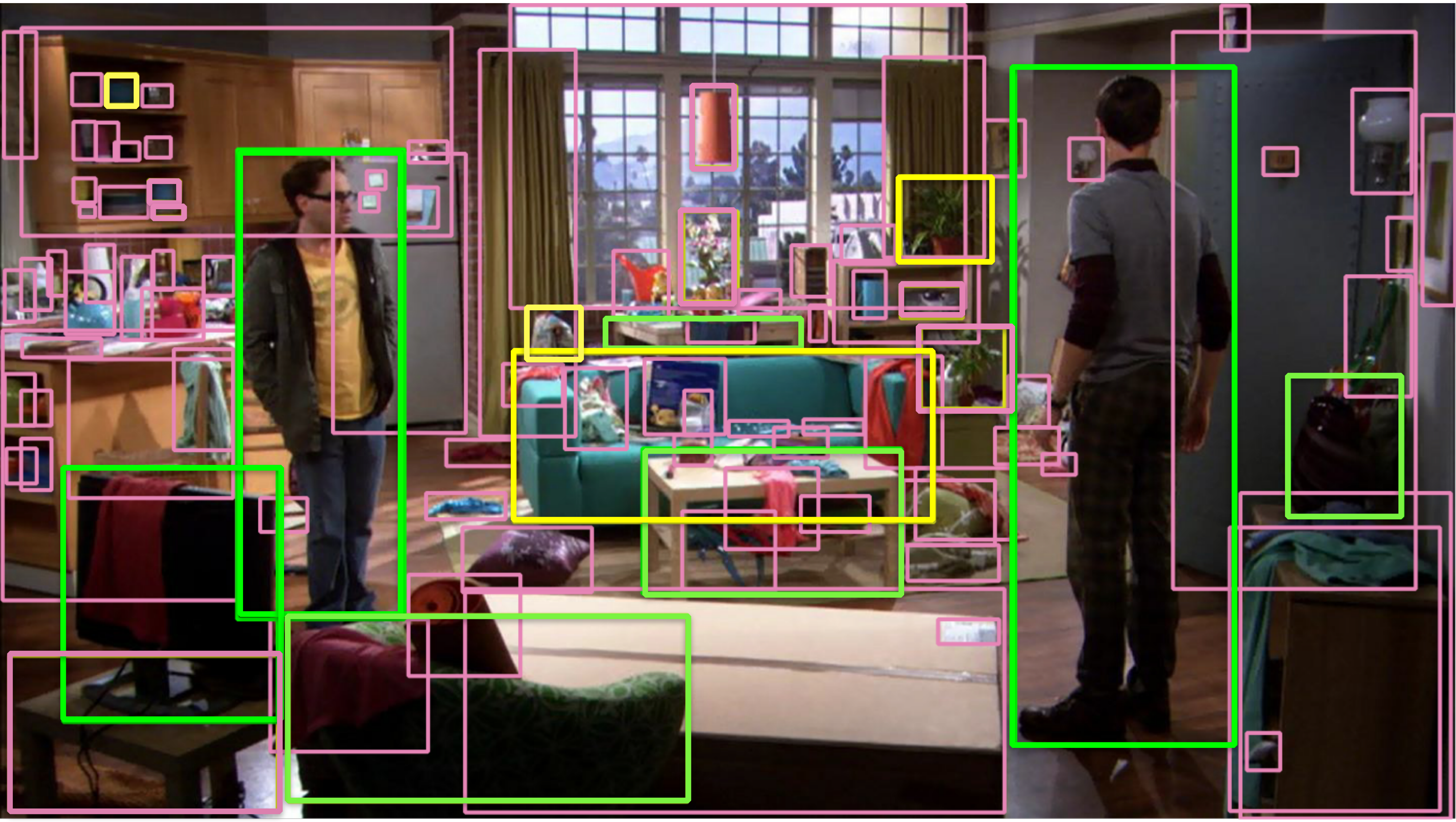


1) **Scaled up the data** by formulating data annotation as an optimization [CHI14, IJCV15]

2) **Developed algorithms** [CVPR10, ECCV12, CVPR15b] and **performed large-scale analysis** to gain insight into the state of the field [ICCV13, IJCV15]

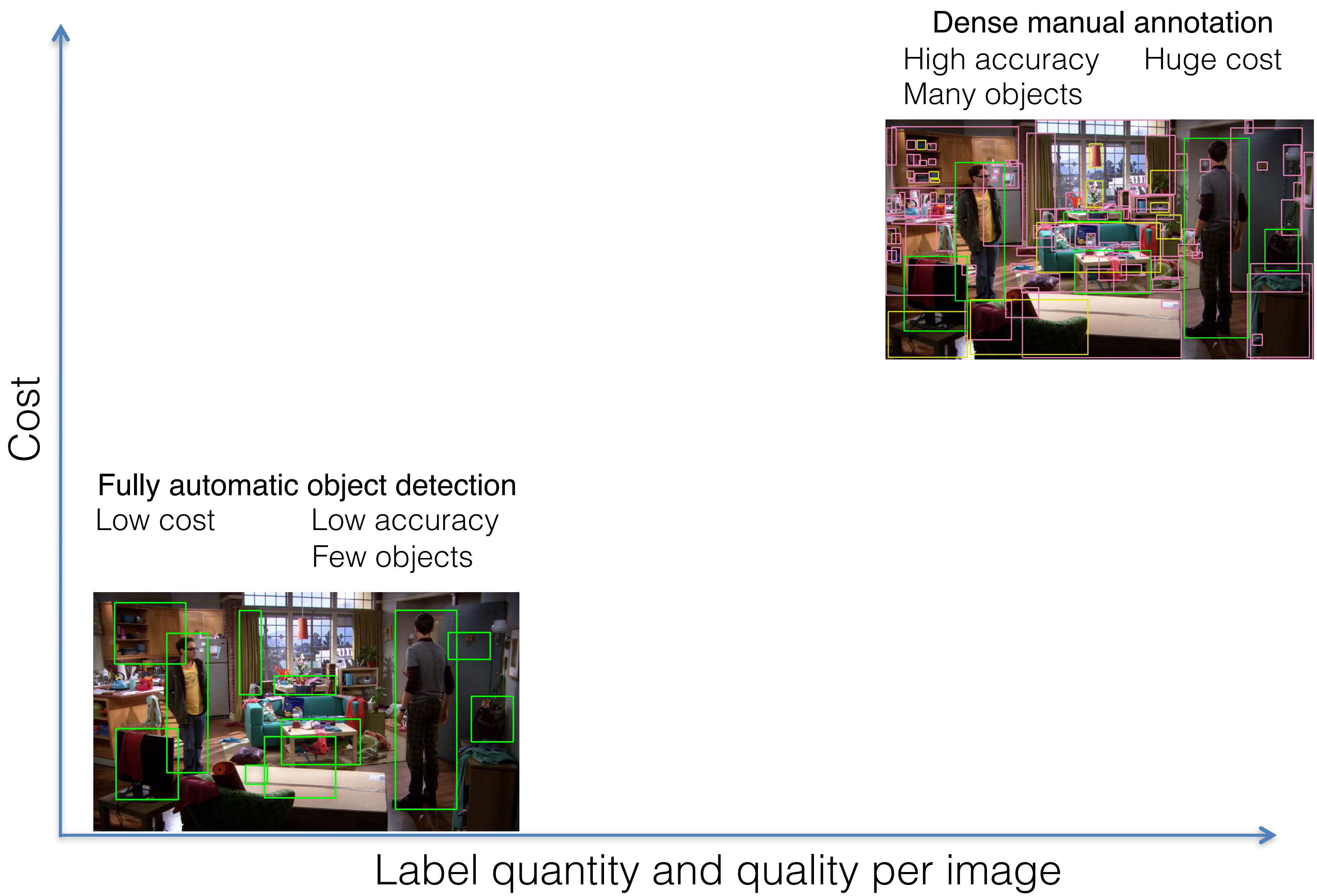
3) Combine insights from both

What would it take to detect all objects here?

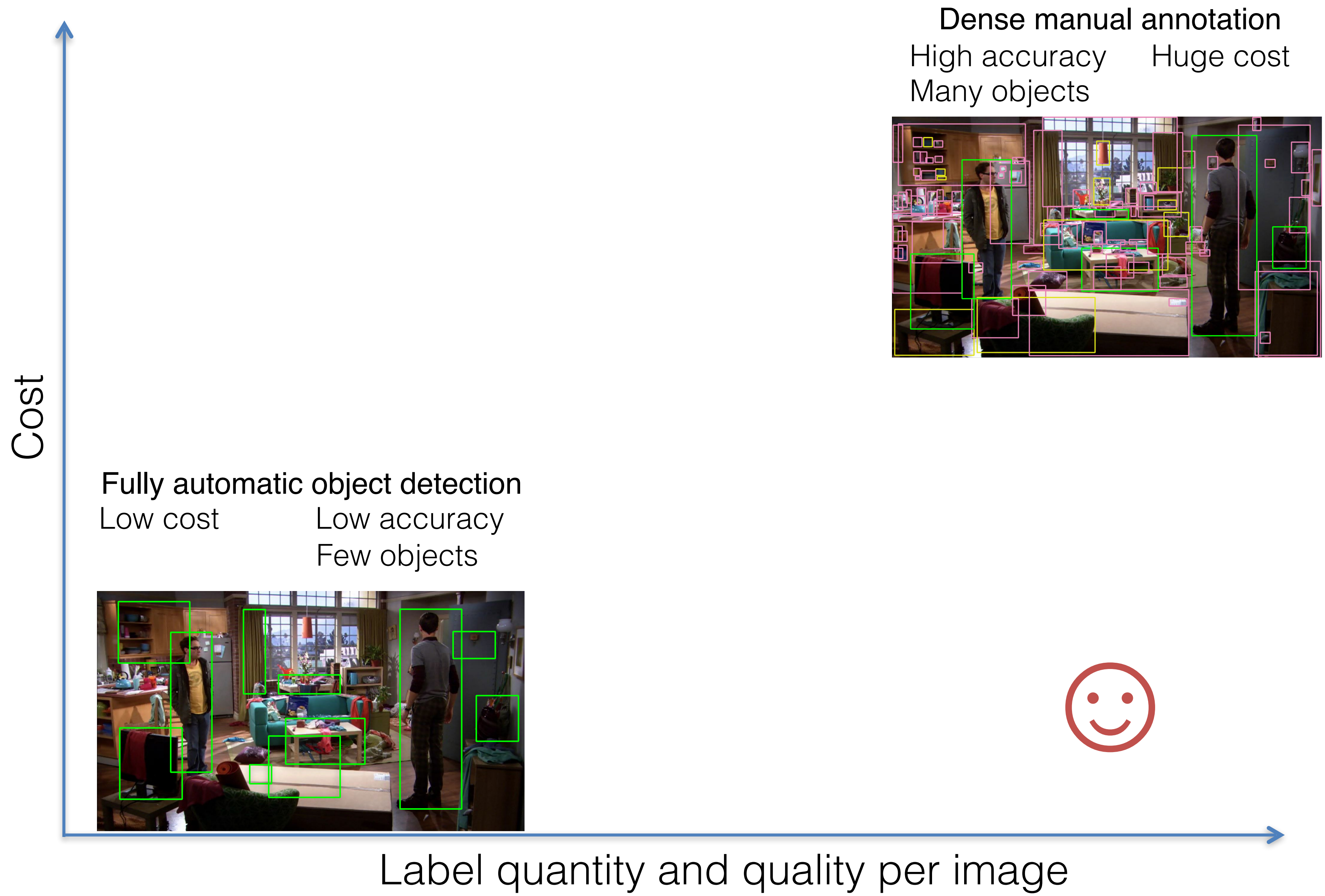


The accuracy/cost tradeoff

The accuracy/cost tradeoff



The accuracy/cost tradeoff

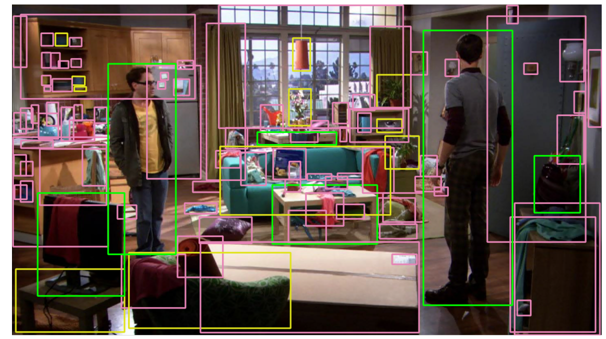


The accuracy/cost tradeoff

Crowd engineering is improving

Dense manual annotation

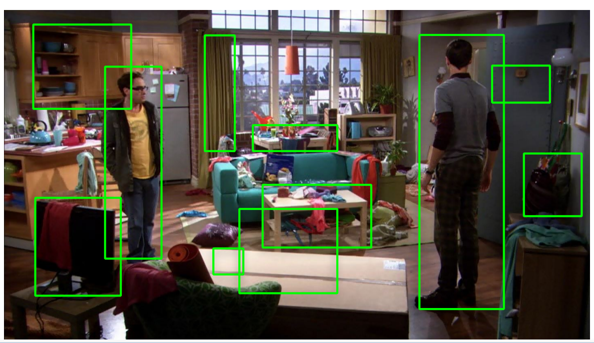
High accuracy   Huge cost
Many objects



Cost

Fully automatic object detection

Low cost   Low accuracy
Few objects



Label quantity and quality per image

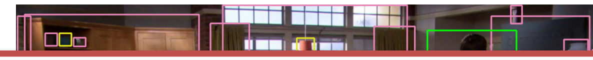
The accuracy/cost tradeoff

Crowd engineering is improving

Dense manual annotation

High accuracy Huge cost

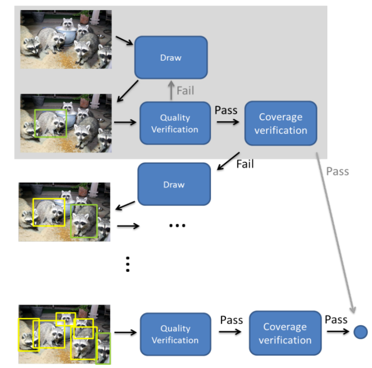
Many objects



Data

Humans need short, focused annotation tasks

Input \ Labels	Table	Chair	Bowl	Dog	Cat	...
	+	+	-	-	-	-
	+	-	+	-	+	-

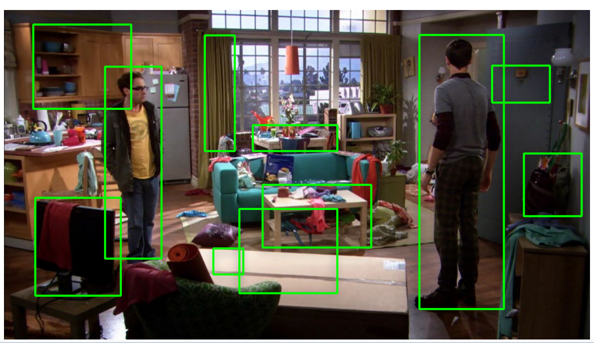


Cost

Fully automatic object detection

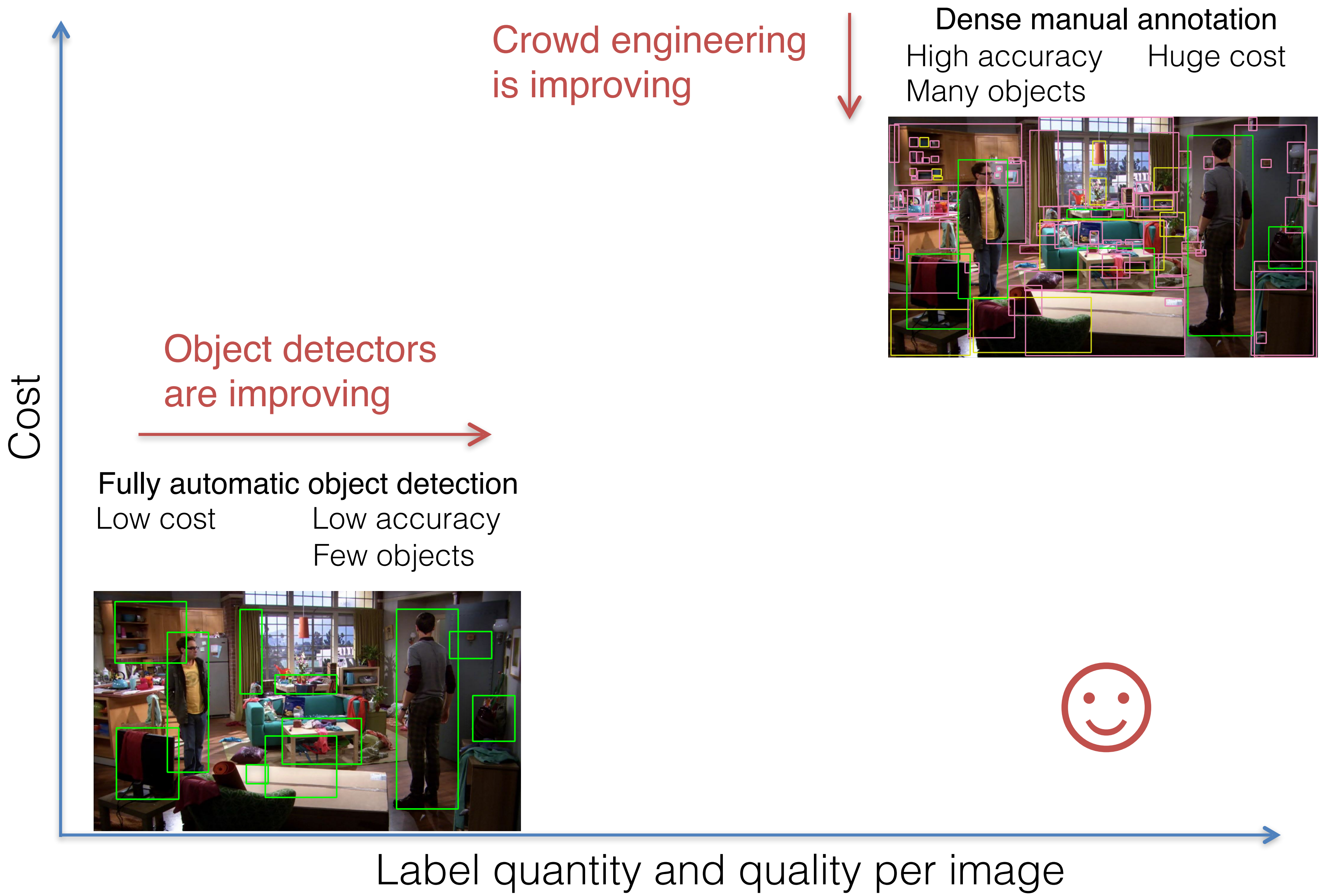
Low cost Low accuracy

Few objects

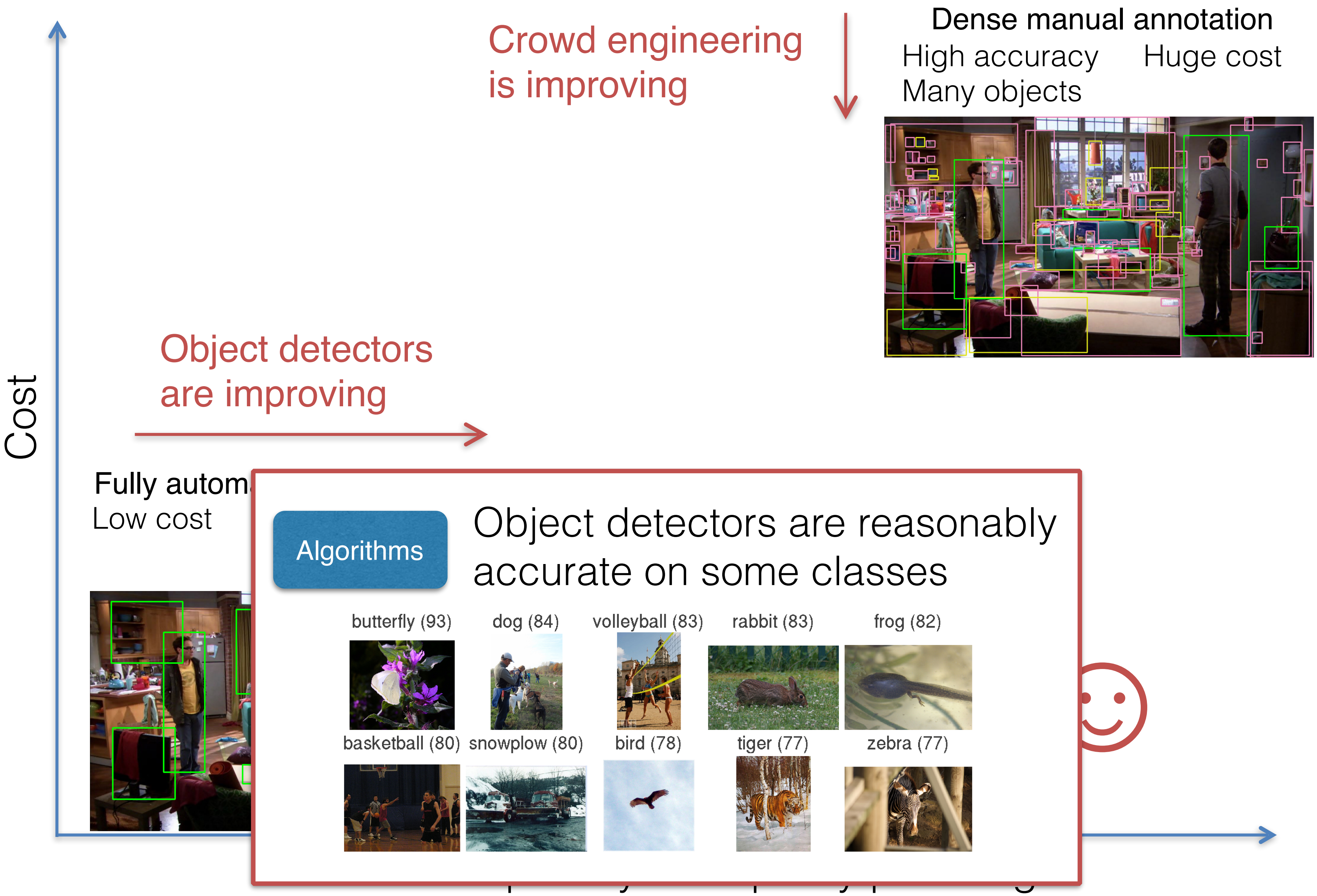


Label quantity and quality per image

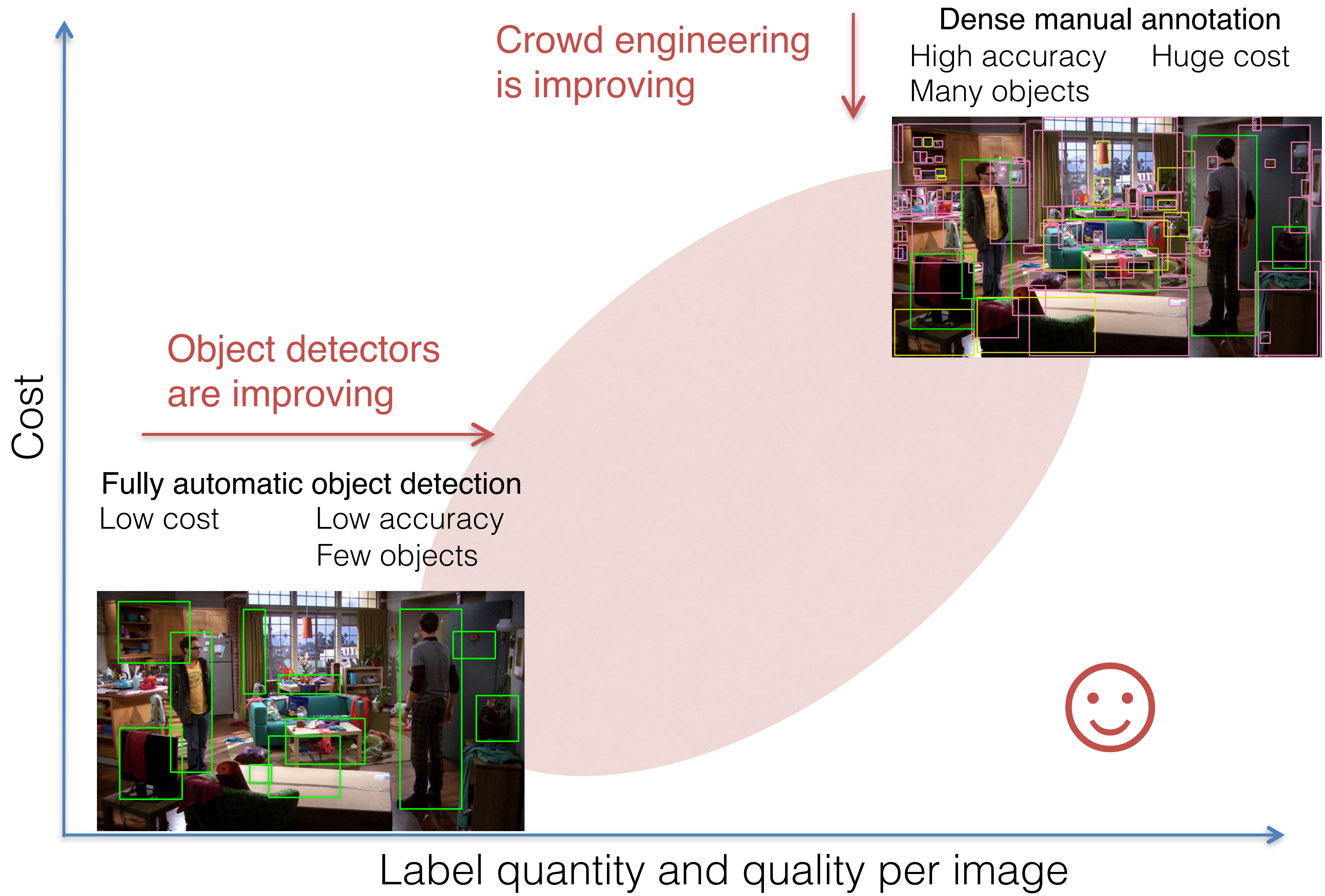
The accuracy/cost tradeoff



The accuracy/cost tradeoff



The accuracy/cost tradeoff



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Human-machine collaboration for object annotation

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Human-machine collaboration for object annotation

Input image
and constraints

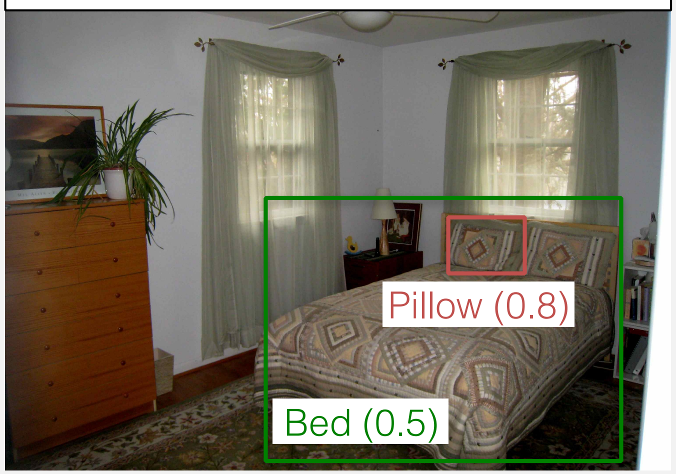
O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Human-machine collaboration for object annotation

Input image
and constraints

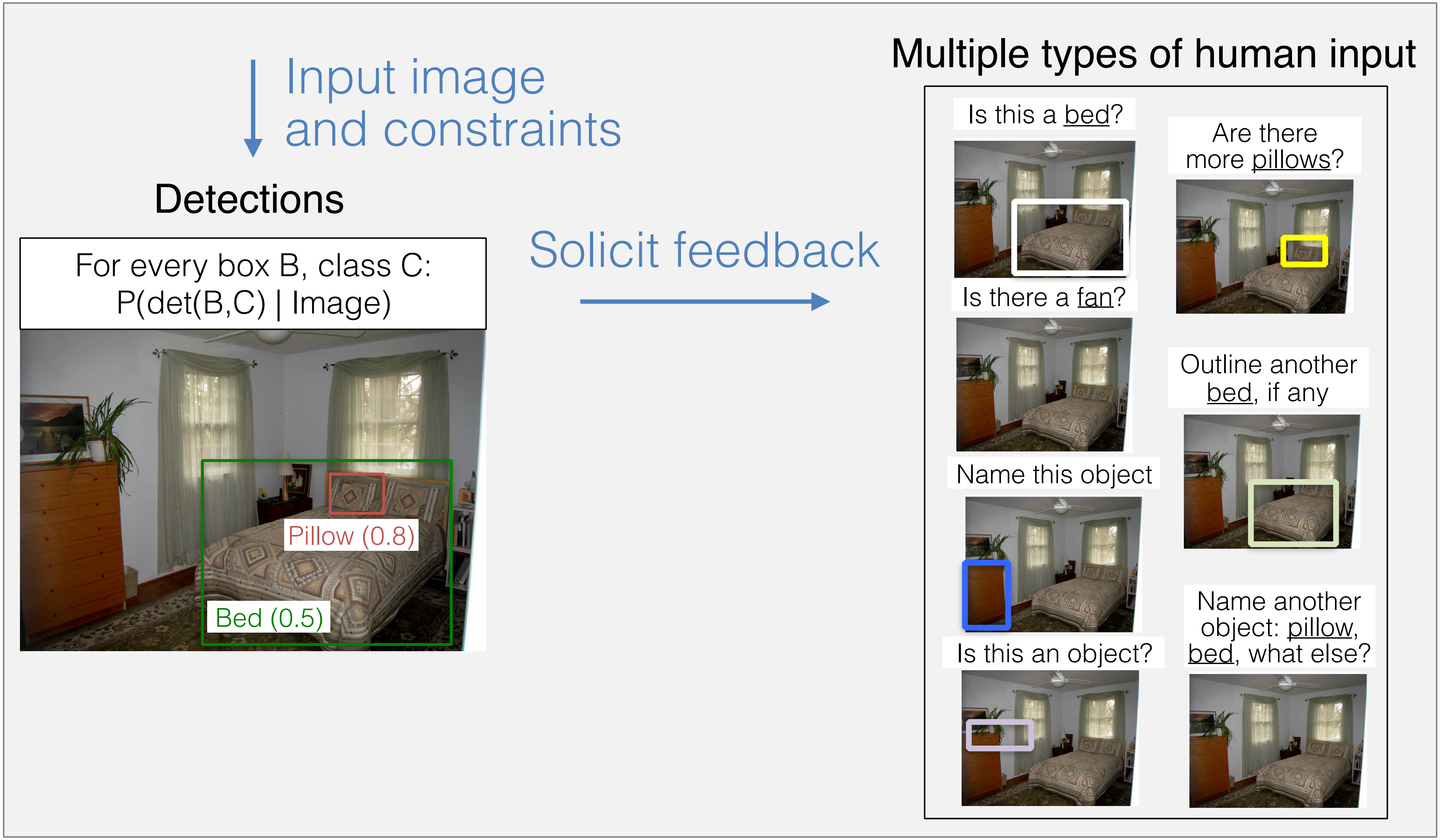
Detections

For every box B, class C:
P(det(B,C) | Image)



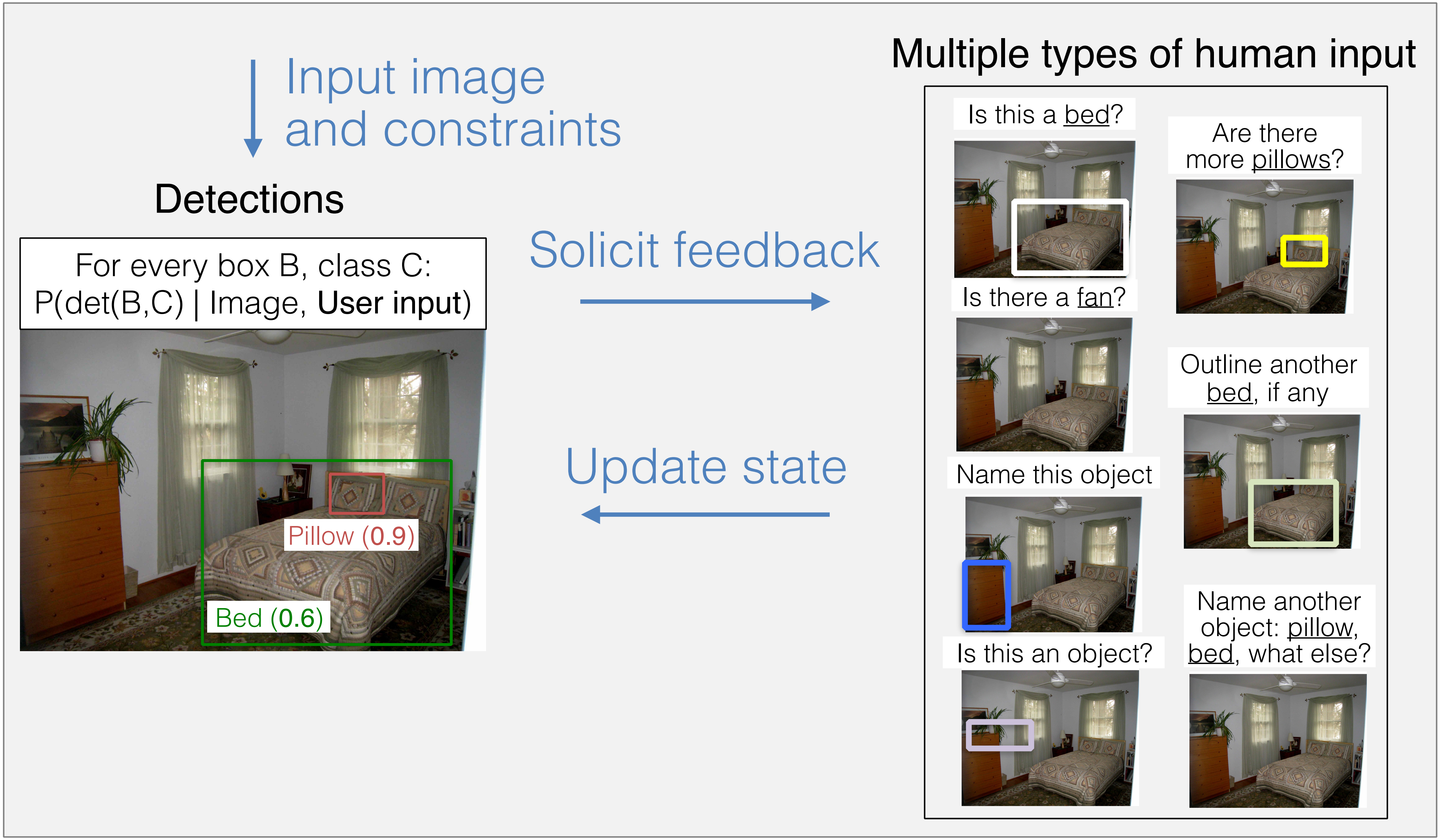
O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Human-machine collaboration for object annotation



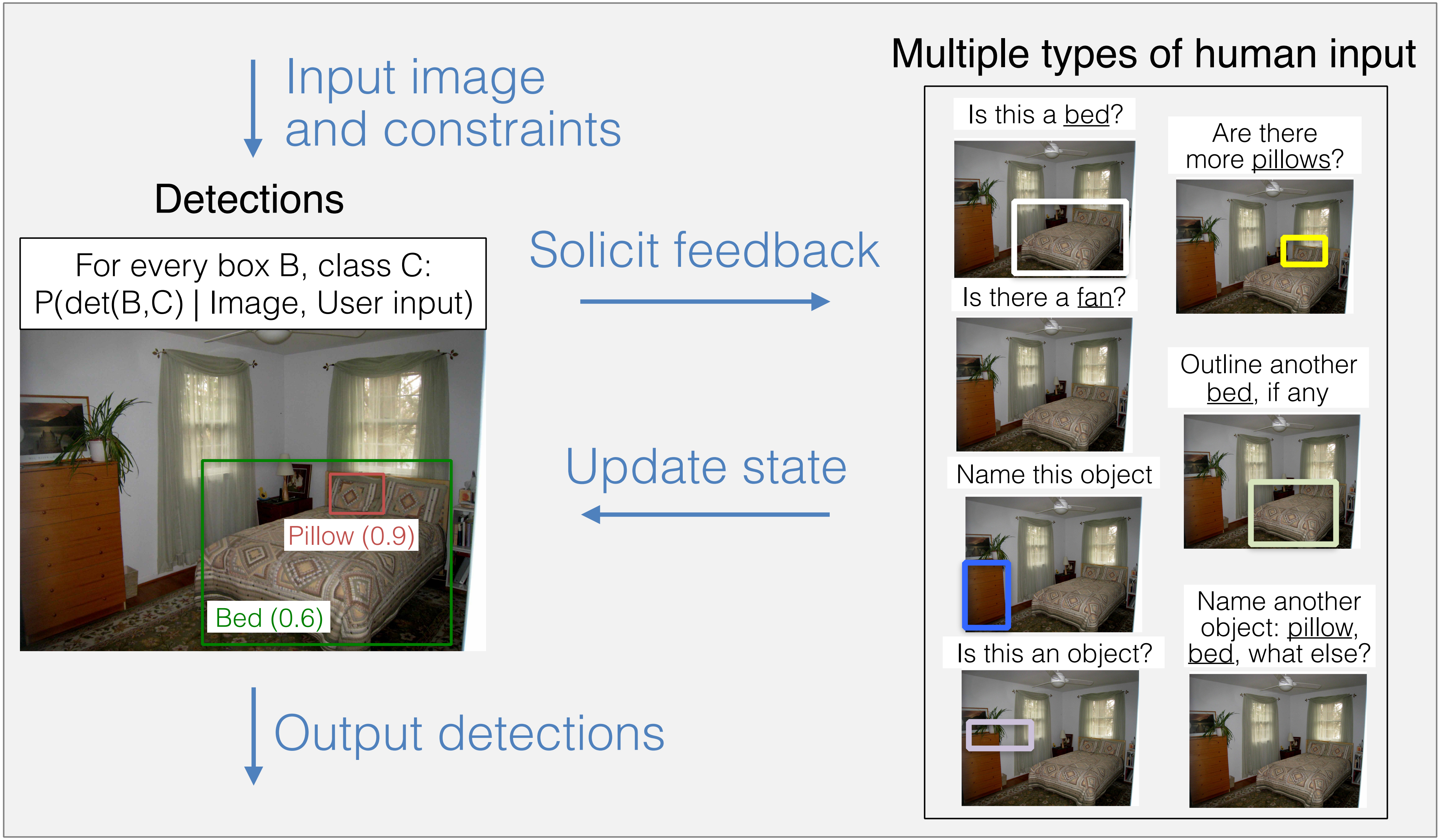
O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Human-machine collaboration for object annotation



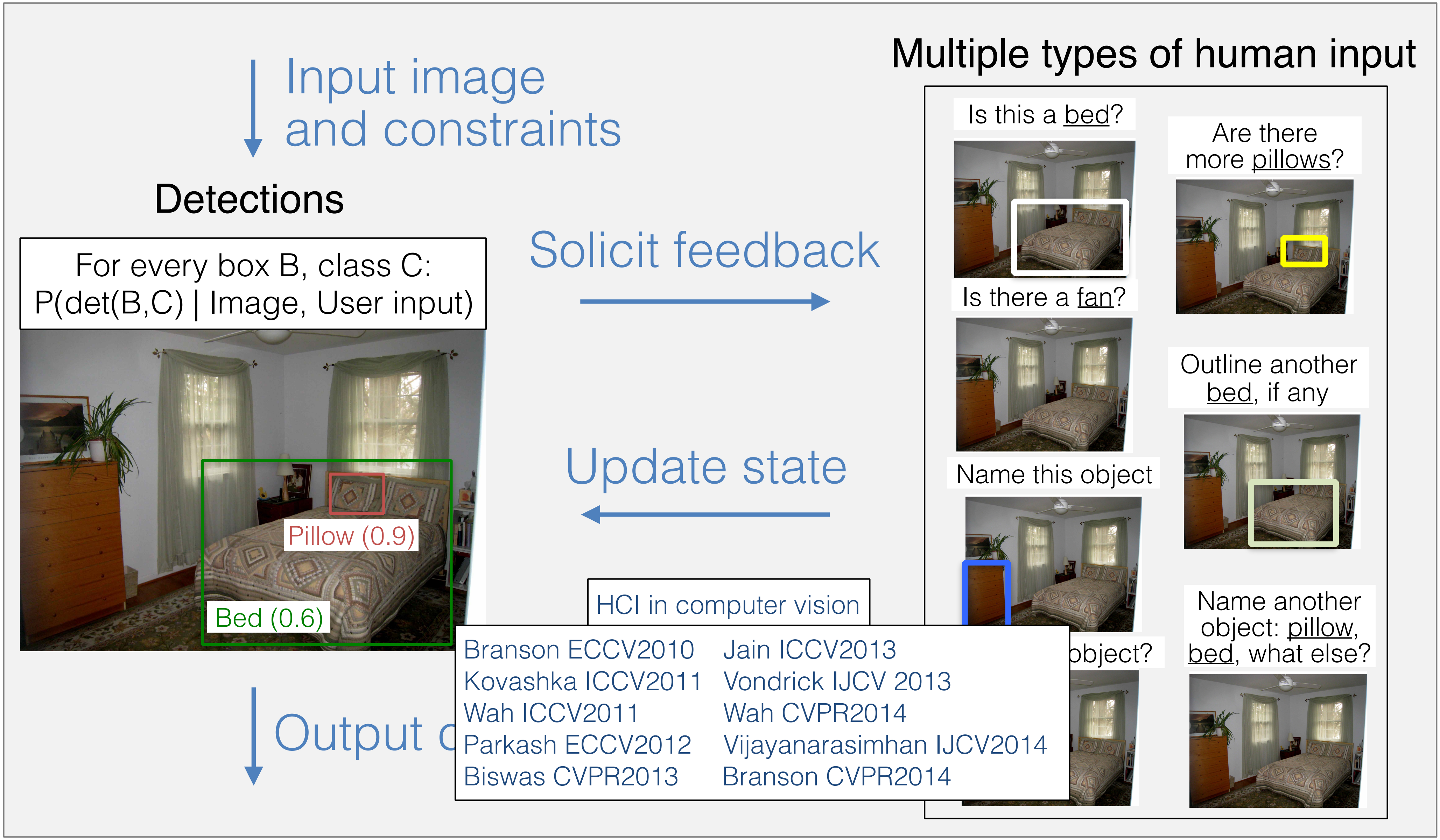
O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Human-machine collaboration
for object annotation



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Human-machine collaboration for object annotation



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Some qualitative results

Computer

Object Detection

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

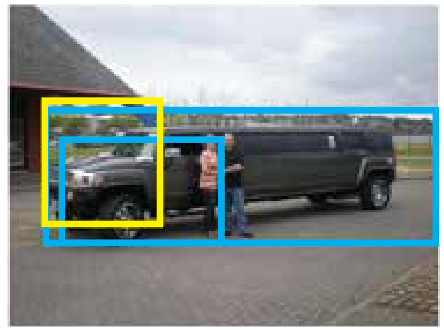
Some qualitative results

Computer

Object Detection

Computer

Verify-box: Is the yellow box tight around a car



Human

Answer: No

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

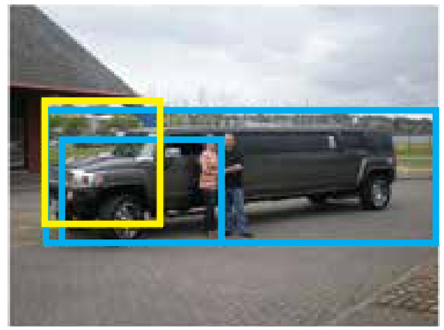
Some qualitative results

Computer

Object Detection

Computer

Verify-box: Is the yellow box tight around a car



Human

Answer: No

...

Computer

Draw-box: Draw a box around a person

Human

Answer: Yellow box below

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

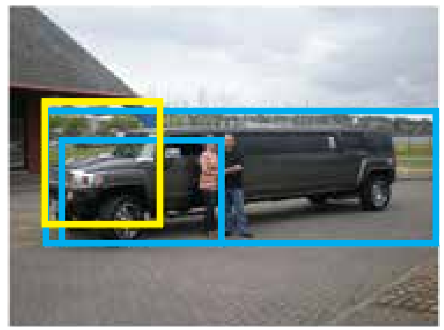
Some qualitative results

Computer

Object Detection

Computer

Verify-box: Is the yellow box tight around a car



Human

Answer: No

...

Computer

Final Labeling

Computer

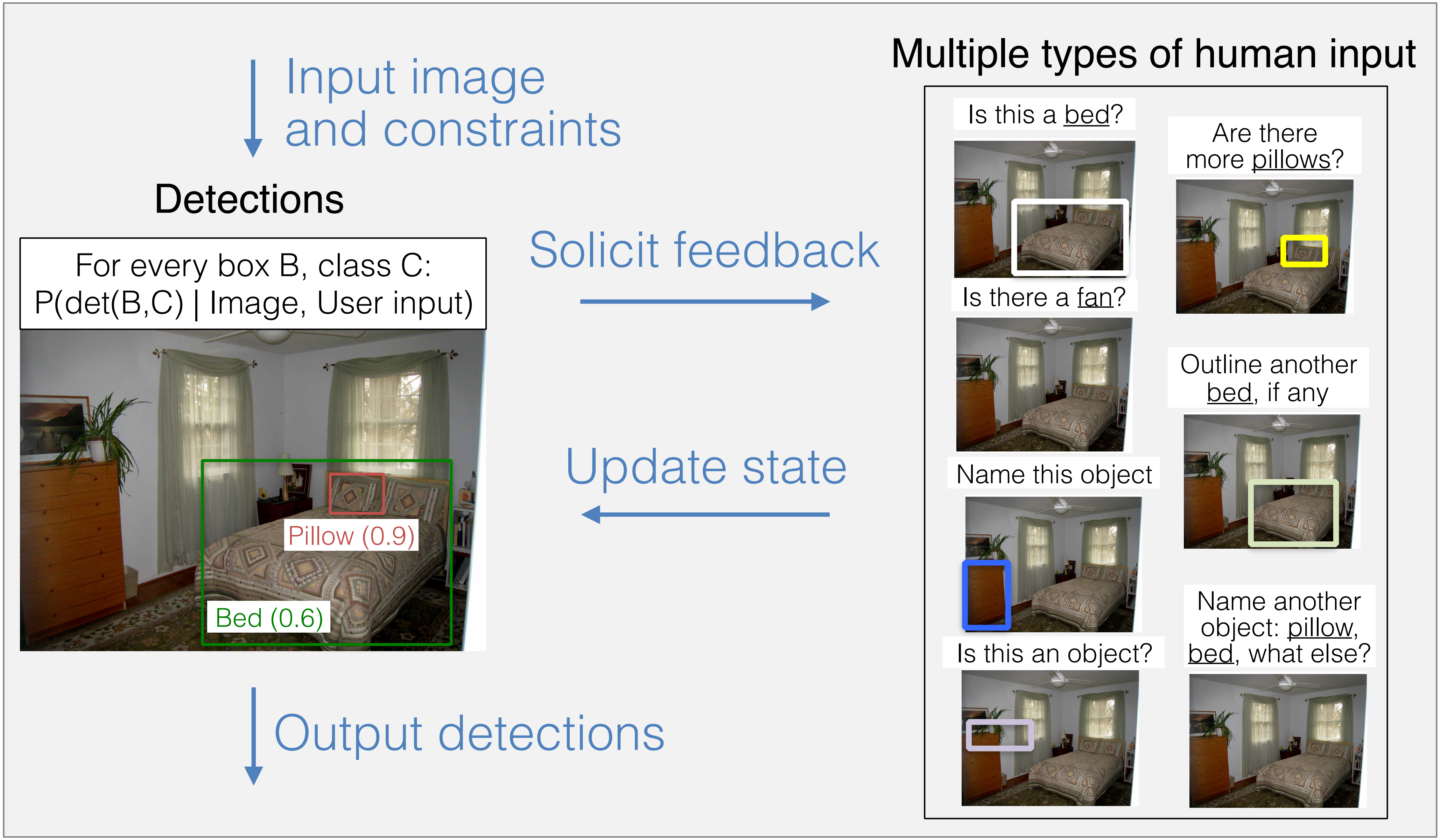
Draw-box: Draw a box around a person

Human

Answer: Yellow box below

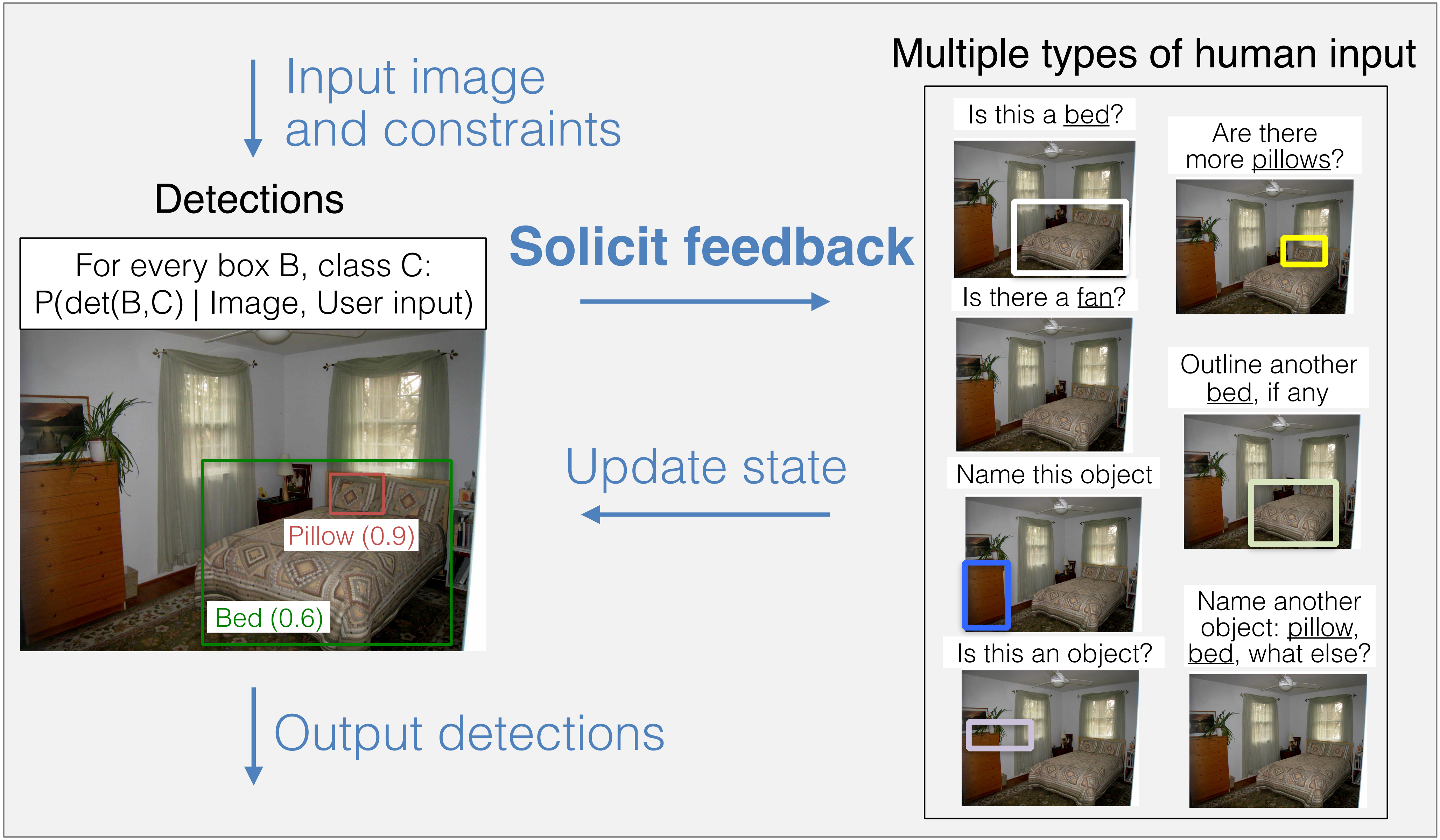
O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Human-machine collaboration for object annotation



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Human-machine collaboration for object annotation

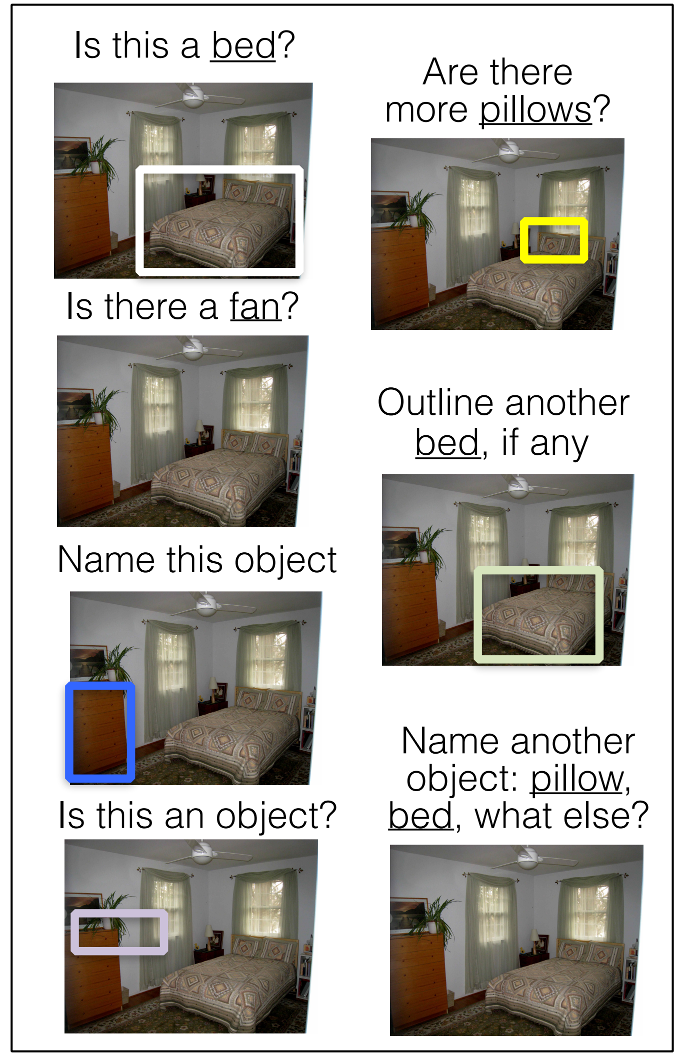


O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

What question to ask?

Current estimates

Decide which question to ask out of (infinitely) many options



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

What question to ask?

Current estimates

Decide which question to ask out of (infinitely) many options

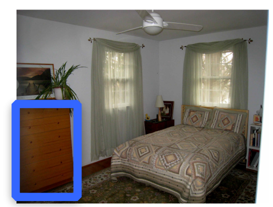
Is this a bed?

Are there more pillows?

Is there a fan?

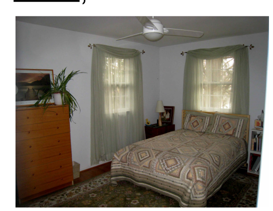
Outline another bed, if any

Name this object



Is this an object?

Name another object: pillow, bed, what else?



Update estimates depending on:

User answers (A)

or

User answers (B)

or

User answers (C)

or

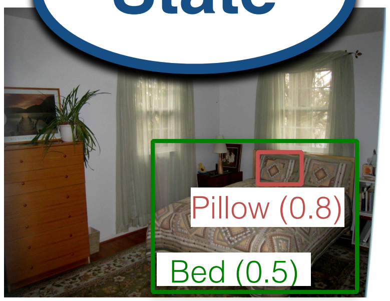
...

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

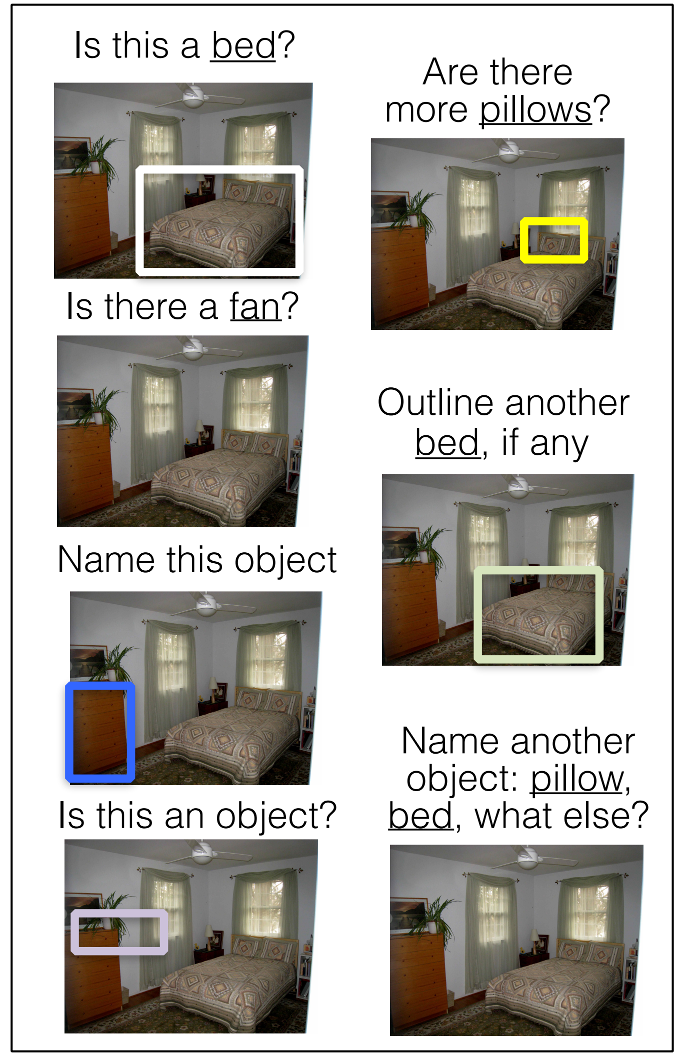
What question to ask?

Curr[illegible]mates

State



Decide which question to ask out of (infinitely) many options



Update estimates depending on:

User answers (A)

or

User answers (B)

or

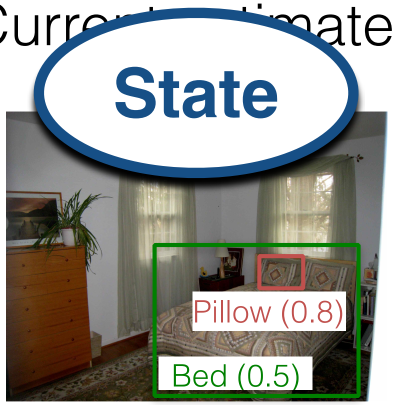
User answers (C)

or

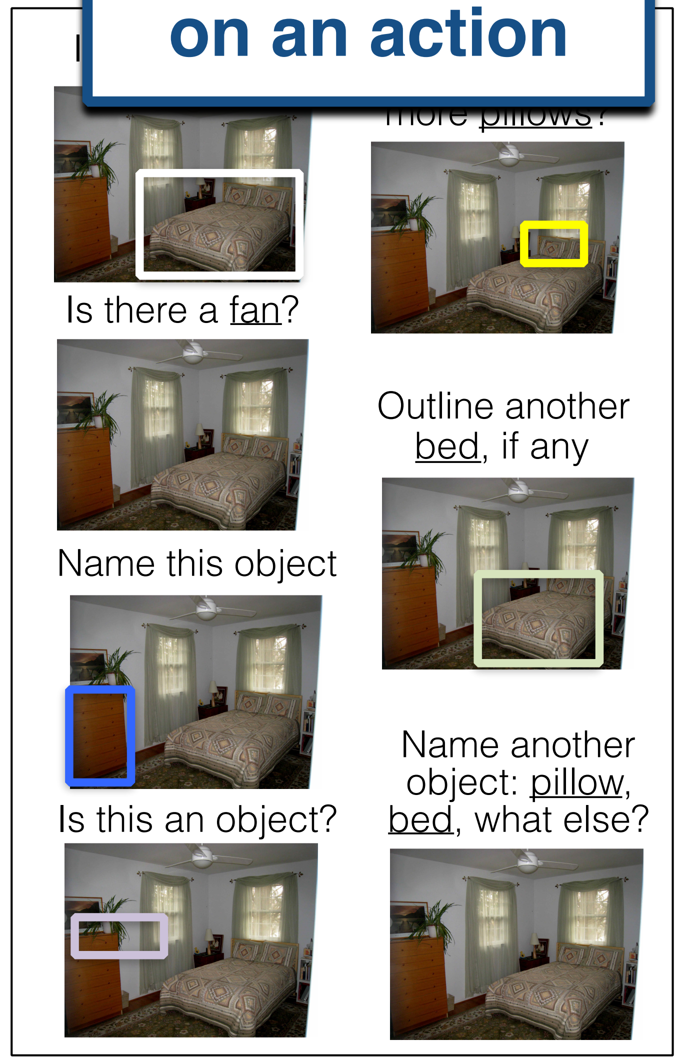
…

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

What question to ask?



Need to decide on an action



Update estimates depending on:

User answers (A)

or

User answers (B)

or

User answers (C)

or

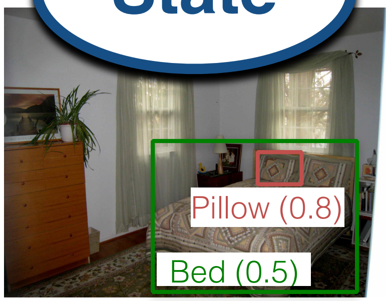
…

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

What question to ask?

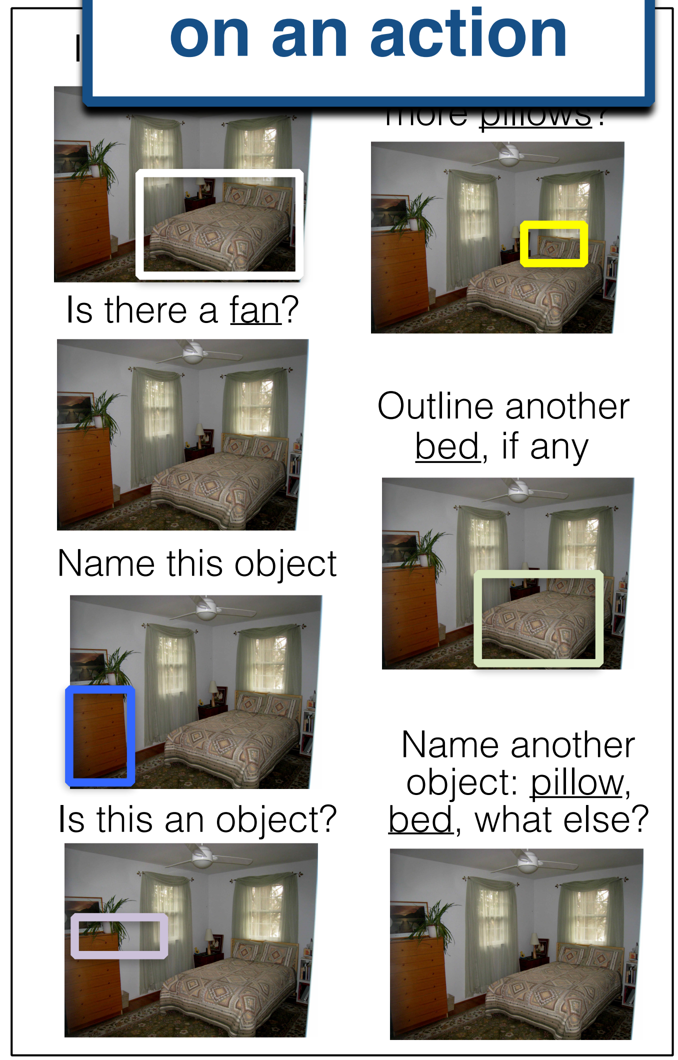
Curr imates

State



De k
out s

Need to decide on an action



Upd
de

User

Probabilistic transitions to new states

or

User answers (B)

or

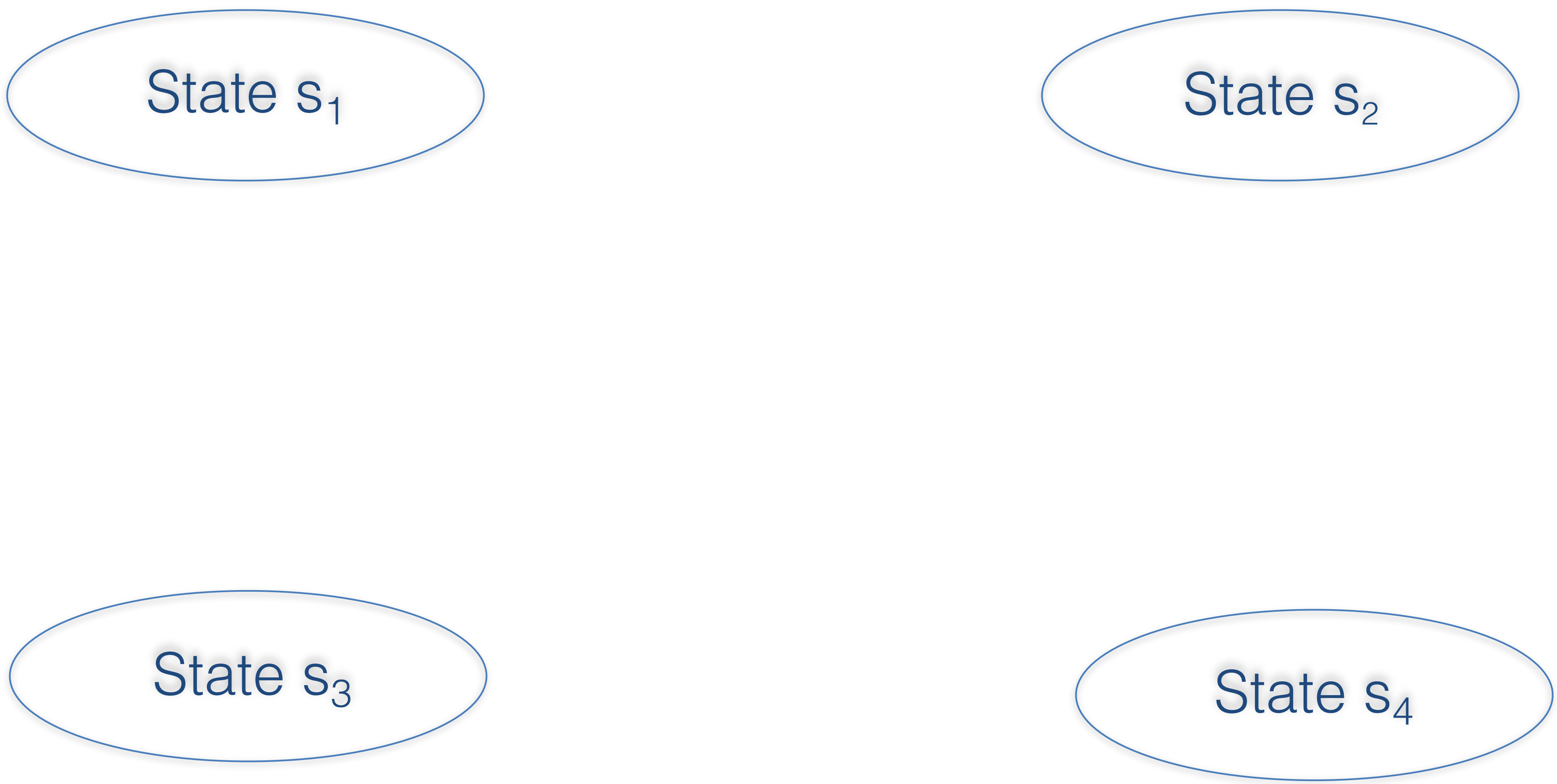
User answers (C)

or

...

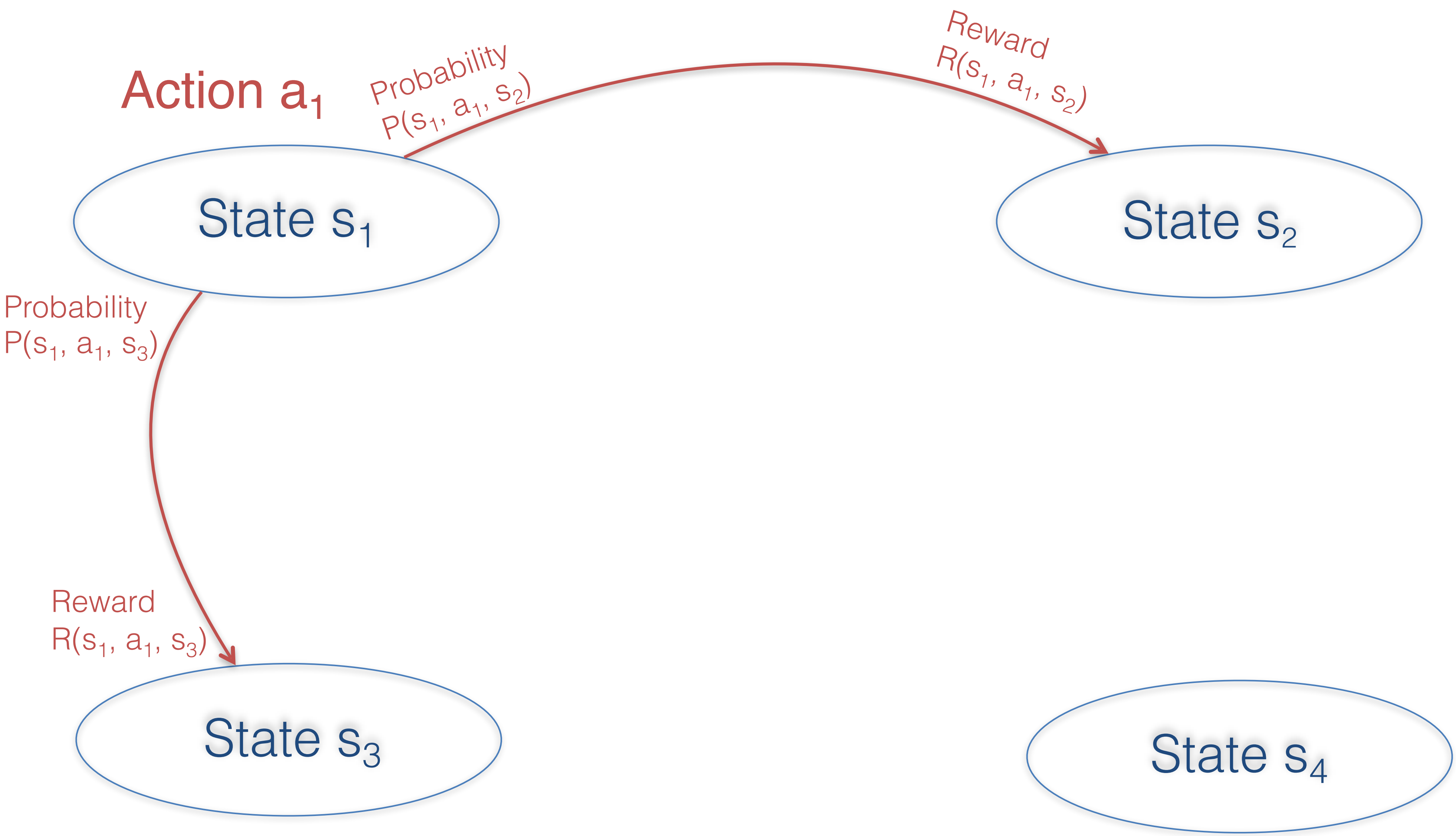
O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Model: Markov Decision Process (MDP)



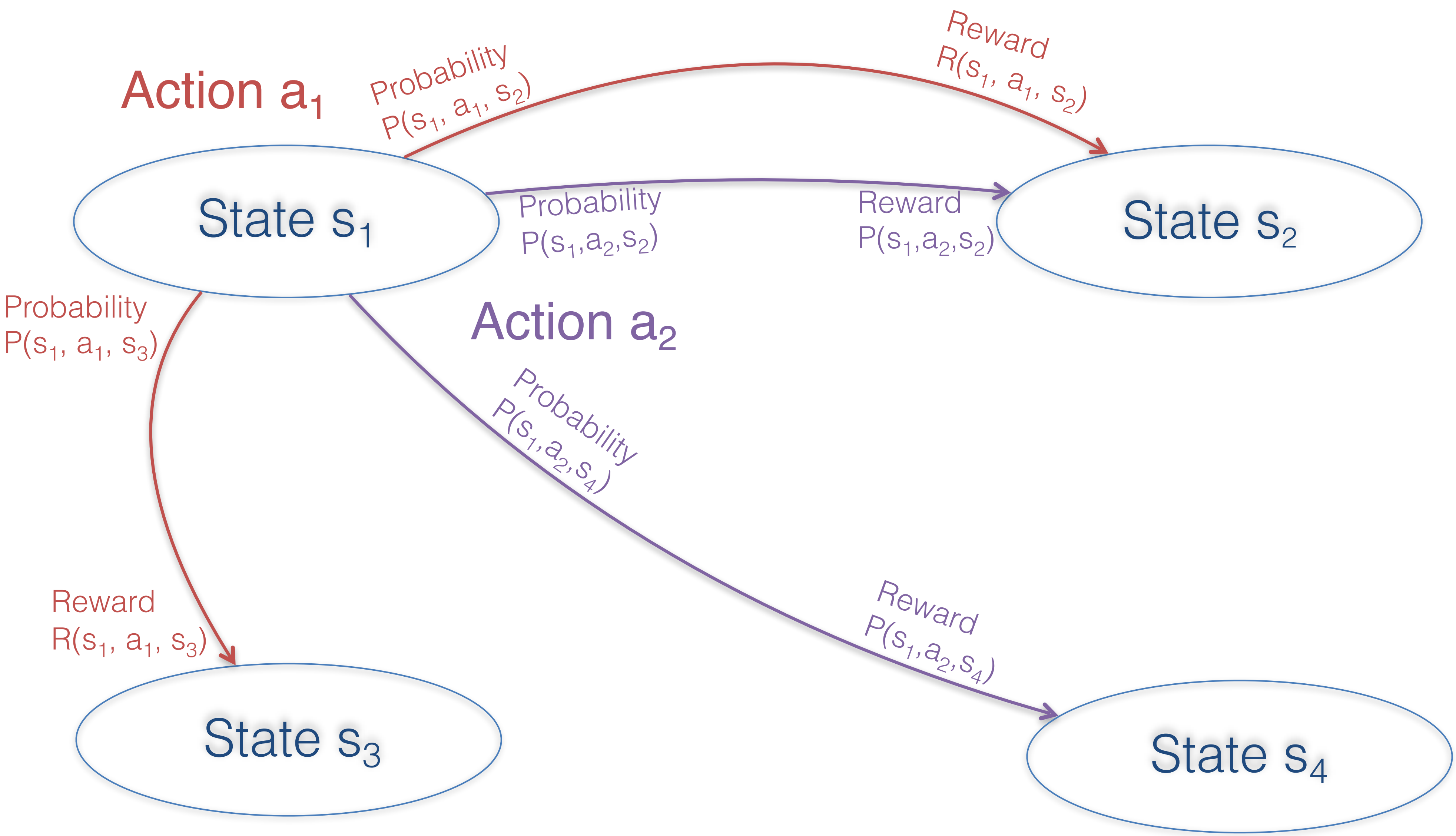
POMDP in vision Karayev CVPR2014, sensor placement Vaisenberg PMC2013, HCI Dai AAAI2010, Kamar AAMAS2012

Model: Markov Decision Process (MDP)



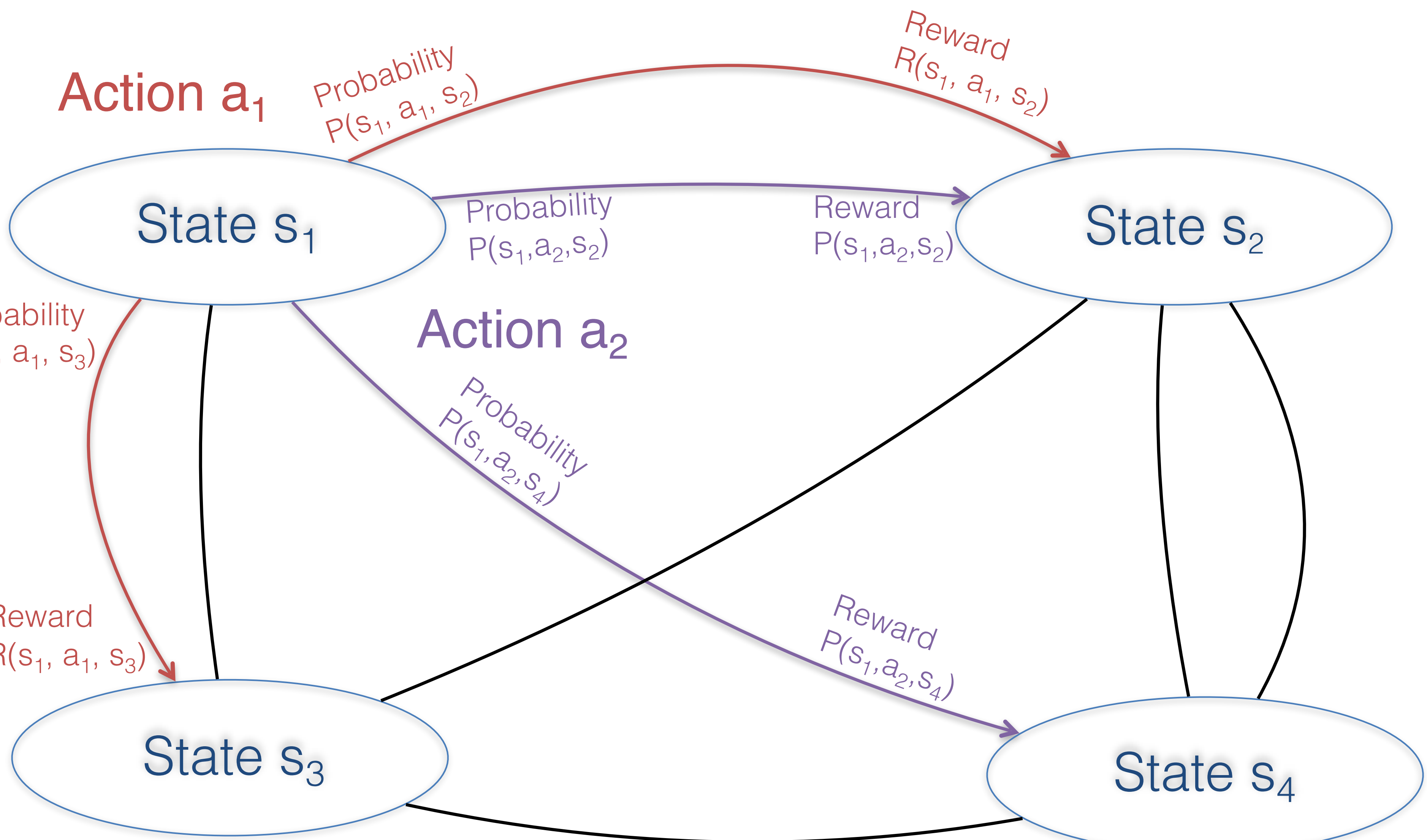
POMDP in vision Karayev CVPR2014, sensor placement Vaisenberg PMC2013, HCI Dai AAAI2010, Kamar AAMAS2012

Model: Markov Decision Process (MDP)



POMDP in vision Karayev CVPR2014, sensor placement Vaisenberg PMC2013, HCI Dai AAAI2010, Kamar AAMAS2012

Model: Markov Decision Process (MDP)

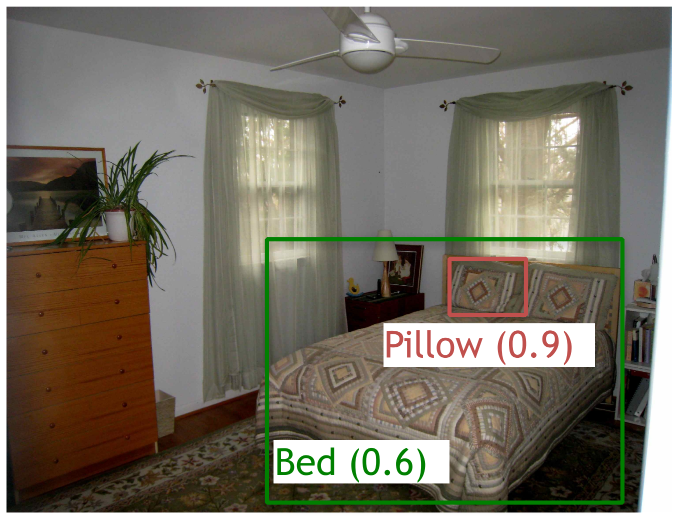


POMDP in vision Karayev CVPR2014, sensor placement Vaisenberg PMC2013, HCI Dai AAAI2010, Kamar AAMAS2012

Model: Markov Decision Process (MDP)

State: set of object detections, with probabilities

Computer+human



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Model: Markov Decision Process (MDP)

State: set of object detections, with probabilities

Action: a question to ask humans

1) Is there a fan?

Cost: 5.34 sec
Error rates: .13/.02

2) Is this a bed?

Cost: 5.89 sec
Error rates: .23/.07

3) Is this an object?

Cost: 5.71 sec
Error rates: .29/.04

4) Name this object.

Cost: 9.67 sec
Error rates: .25/.08/.06

5) Are there more pillows?

Cost: 7.57 sec
Error rates: .25/.26

6) Outline another bed, if any.

Cost: 10.21 sec
Error rates: .28/.16/.29

7) Name another object: pillow, bed, what else?

Cost: 9.46 sec
Error rates: .02/.12/.05

...

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Model: Markov Decision Process (MDP)

State: set of object detections, with probabilities

Action: a question to ask humans

Transition probability: probability distribution over user responses

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Model: Markov Decision Process (MDP)

State: set of object detections, with probabilities

Action: a question to ask humans

Transition probability: probability distribution over user responses

Reward: increase in estimated quality of labeling divided by the cost of actions

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Model: Markov Decision Process (MDP)

State: set of object detections, with probabilities

Action: a question to ask humans

Transition probability: probability distribution over user responses

Reward: increase in estimated quality of labeling divided by the cost of actions

Algorithm: 2-step lookahead search

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Model: Markov Decision Process (MDP)

State: set of object detections, with probabilities

Action: a question to ask humans

Transition probability: probability distribution over user responses

Reward: increase in estimated quality of labeling divided by the cost of actions

Algorithm: 2-step lookahead search

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Computing the transition probability

Given:

- An action/question A (e.g., “is there a fan in this image?”)
- Possible truths T_1, T_2, … (e.g., T_1 = “there is a fan”, T_2 = “there is no fan”)
- Image appearance I and all user responses so far U

Goal:

- Compute the probability of user answer u (e.g., u = user says “yes”)

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Computing the transition probability

Given:

- An action/question *A* (e.g., “is there a fan in this image?”)
- Possible truths T_1, T_2, … (e.g., T_1 = “there is a fan”, T_2 = “there is no fan”)
- Image appearance *I* and all user responses so far *U*

Goal:

- Compute the probability of user answer *u* (e.g., *u* = user says “yes”)

$$P(u|I,U) = \sum_i P(u|T_i, I, U) P(T_i|I,U)$$

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Computing the transition probability

Given:

- An action/question A (e.g., “is there a fan in this image?”)
- Possible truths T_1, T_2, … (e.g., T_1 = “there is a fan”, T_2 = “there is no fan”)
- Image appearance I and all user responses so far U

Goal:

- Compute the probability of user answer u (e.g., u = user says “yes”)

$$P(u|I,U) = \sum_i \underline{P(u|T_i,I,U)}P(T_i|I,U)$$

$$= \sum_i \underline{P(u|T_i)}P(T_i|I,U)$$

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Computing the transition probability

Given:

- An action/question *A* (e.g., “is there a fan in this image?”)
- Possible truths T_1, T_2, … (e.g., T_1 = “there is a fan”, T_2 = “there is no fan”)
- Image appearance *I* and all user responses so far *U*

Goal:

- Compute the probability of user answer *u* (e.g., *u* = user says “yes”)

$$P(u|I,U) = \sum_i \underline{P(u|T_i,I,U)}P(T_i|I,U)$$

Simplifying assumptions of [Branson ECCV10]: user’s answer is independent of (1) other users, and (2) image appearance

$$= \sum_i \underline{P(u|T_i)}P(T_i|I,U)$$

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Computing the transition probability

Given:

- An action/question A (e.g., "is there a fan in this image?")
- Possible truths T_1, T_2, … (e.g., T_1 = "there is a fan", T_2 = "there is no fan")
- Image appearance I and all user responses so far U

Goal:

- Compute the probability of user answer u (e.g., u = user says "yes")

$$P(u|I,U) = \sum_i \underline{P(u|T_i,I,U)}P(T_i|I,U)$$

Simplifying assumptions of [Branson ECCV10]: user's answer is independent of (1) other users, and (2) image appearance

$$= \sum_i \underbrace{P(u|T_i)}_{\text{Precomputed error rates}}P(T_i|I,U)$$

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Computing the transition probability

Given:

- An action/question *A* (e.g., "is there a fan in this image?")
- Possible truths T_1, T_2, … (e.g., T_1 = "there is a fan", T_2 = "there is no fan")
- Image appearance *I* and all user responses so far *U*

Goal:

- Compute the probability of user answer *u* (e.g., *u* = user says "yes")

$$P(u|I,U) = \sum_i \underbrace{P(u|T_i, I, U)}P(T_i|I,U)$$

Simplifying assumptions of [Branson ECCV10]: user's answer is independent of (1) other users, and (2) image appearance

$$= \sum_i \underbrace{P(u|T_i)}_{\text{Precomputed error rates}} \underbrace{P(T_i|I,U)}_{\text{Current estimate of the correct answer}}$$

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Computing the correct answer

Given:

- Image appearance *I* and all user responses so far *U*

Goal:

- Compute the probability of truth *T* (e.g., *T* = there is a fan in the image)

$$\underbrace{P(T|I,U)}_{\text{Current estimate of the correct answer}} = \underbrace{P(T|I)}_{\text{Computer vision model}} \prod_{k} \underbrace{P(u_k|T)}_{\text{Precomputed error rates}}$$

Number of users

Simplifying assumptions of [Branson ECCV10]: user's answer is independent of (1) other users, and (2) image appearance

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Computing the correct answer

Given:

- Image appearance *I* and all user responses so far *U*

Goal:

- Compute the probability of truth *T* (e.g., *T* = there is a fan in the image)

$$\underbrace{P(T|I,U)}_{\text{Current estimate of the correct answer}} = \underbrace{P(T|I)}_{\text{Computer vision model}} \prod_{k} \underbrace{P(u_k|T)}_{\text{Precomputed error rates}}$$

Number of users

Simplifying assumptions of [Branson ECCV10]: user's answer is independent of (1) other users, and (2) image appearance

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Multiple computer vision models

Computer+human

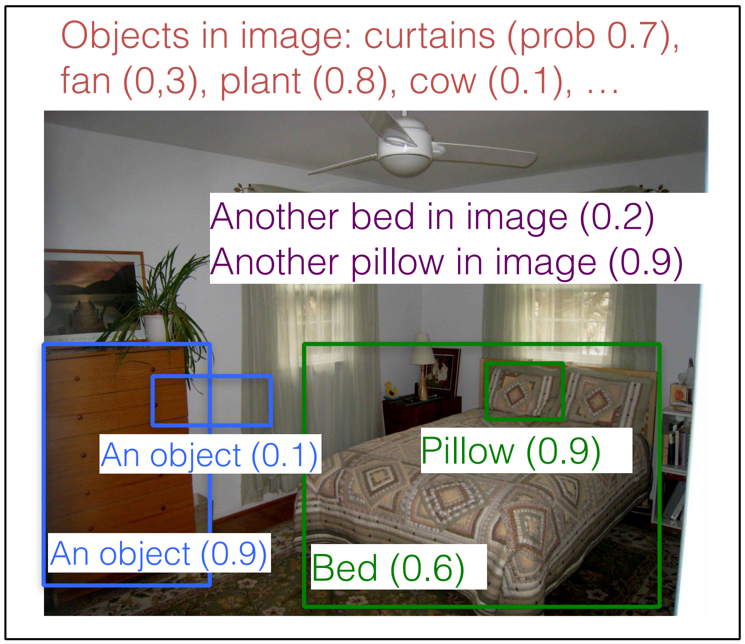


Image classifiers:

200-way CNN classifiers released with LSDA
Probabilities from Platt scaling
[Hoffman NIPS14, Yangqing Jia's Caffe, Platt99]

Object detectors:

200 object RCNN detectors + Platt scaling
[Girshick CVPR14, Yangqing Jia's Caffe, Platt99]

Probability of object in region:

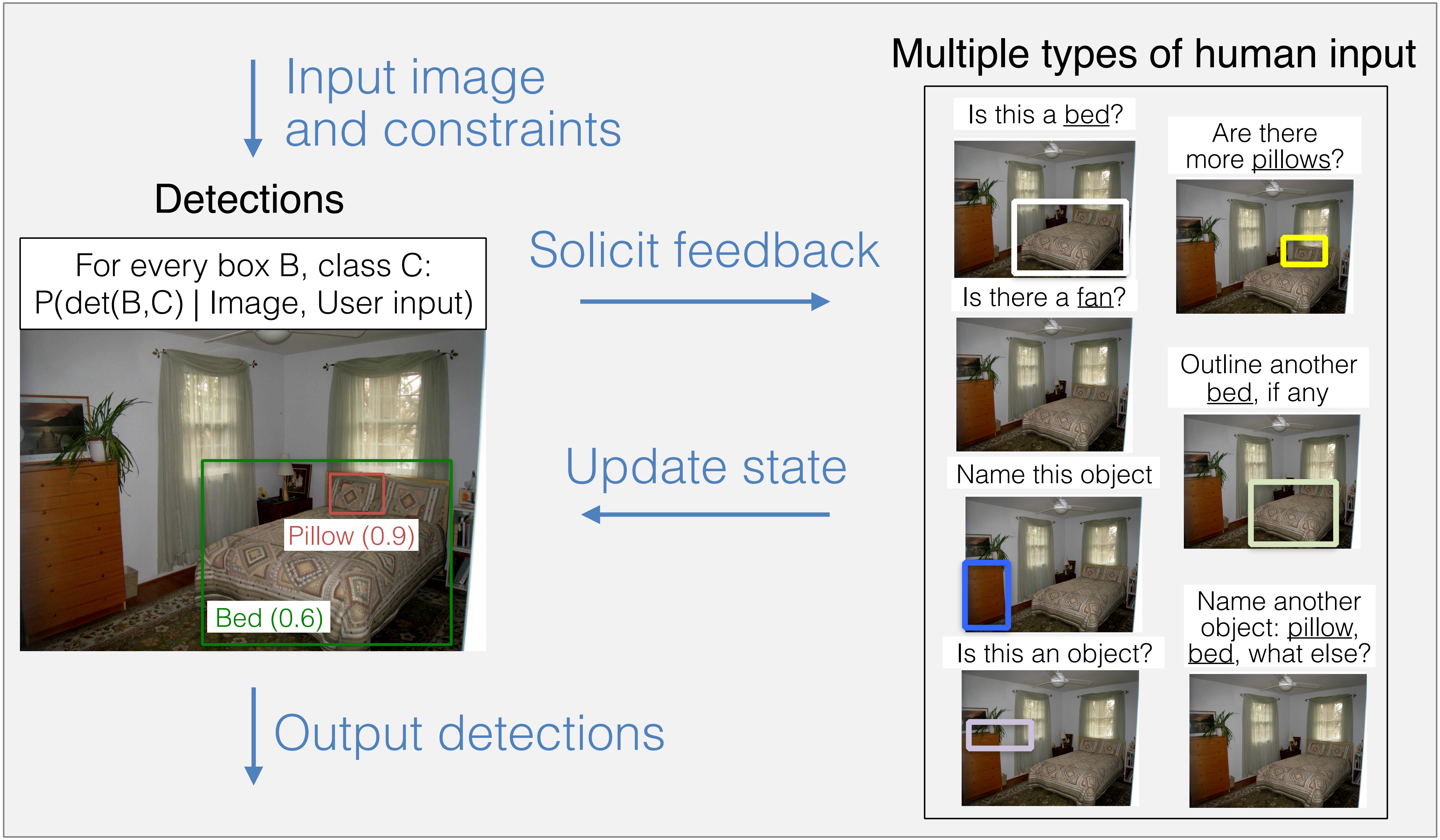
Objectness measure [Alexe PAMI2012]

Probability of another instance of same class, probability of another class in image:

Statistics from ILSVRC2014 val-DET data

O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

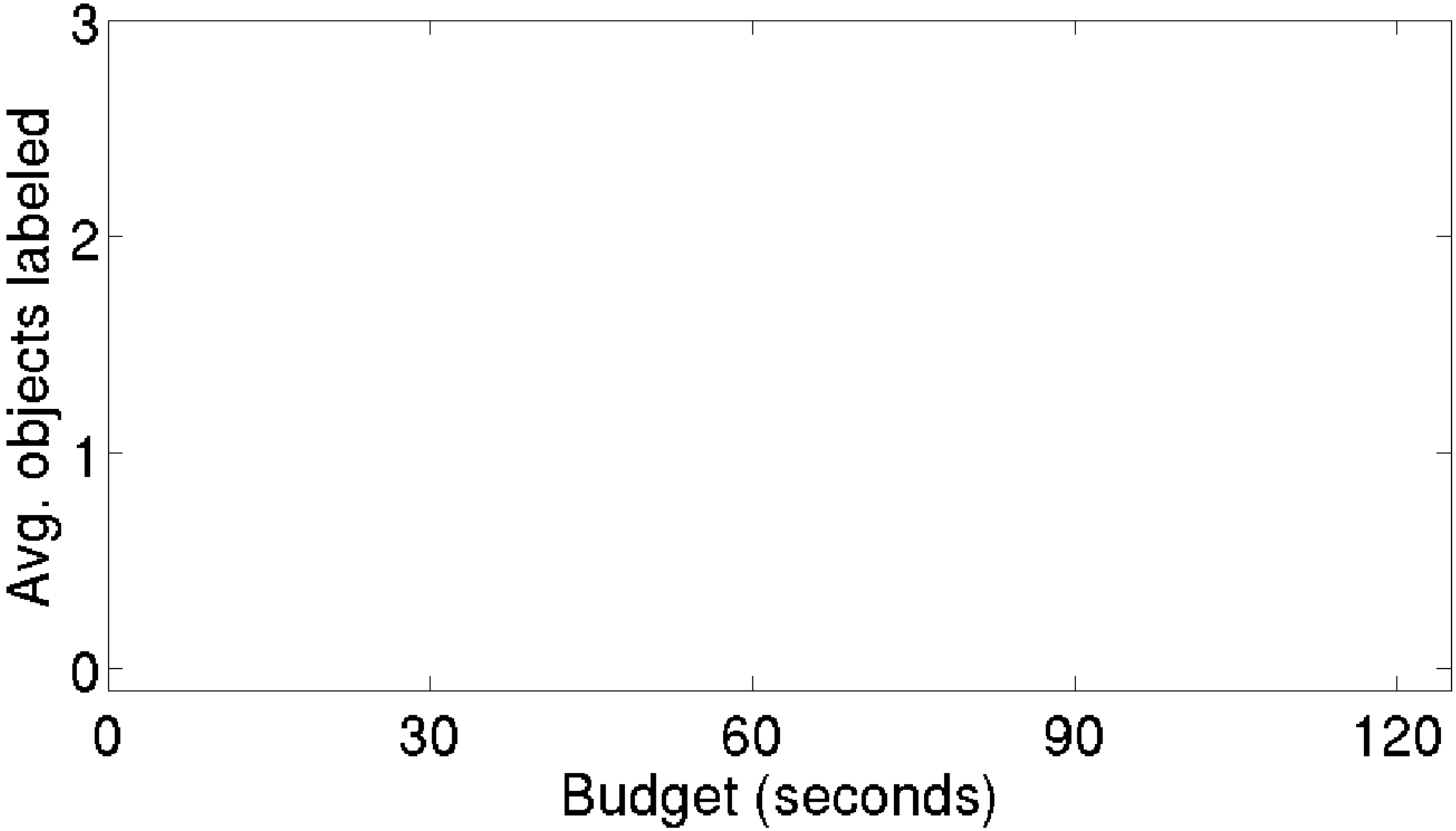
Human-machine collaboration for object annotation



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Results

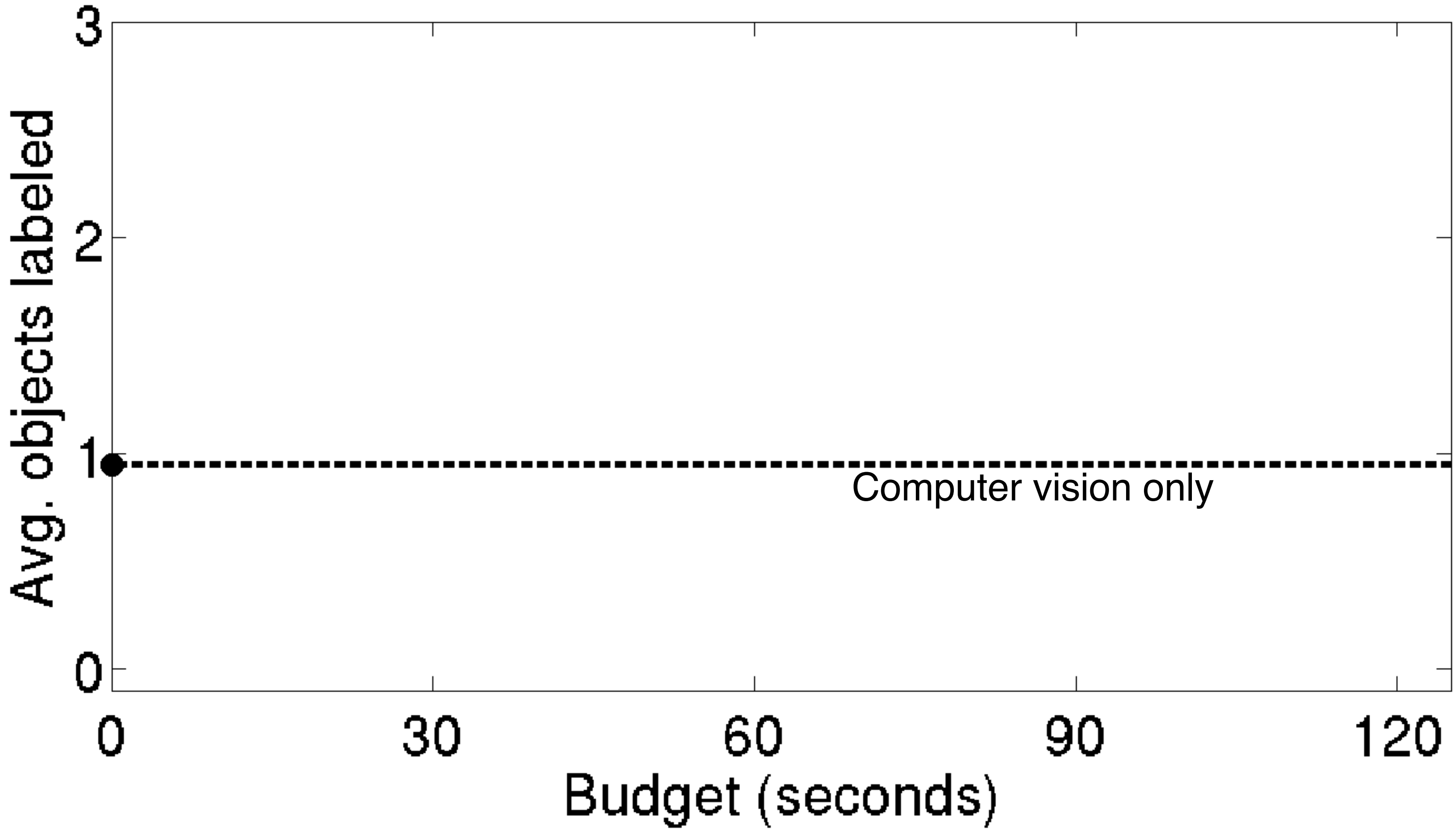
2K images of ILSVRC2014 detection val set with at least 4 object instances
Human error rates computed from AMT experiments
Annotation experiments in simulation



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Results

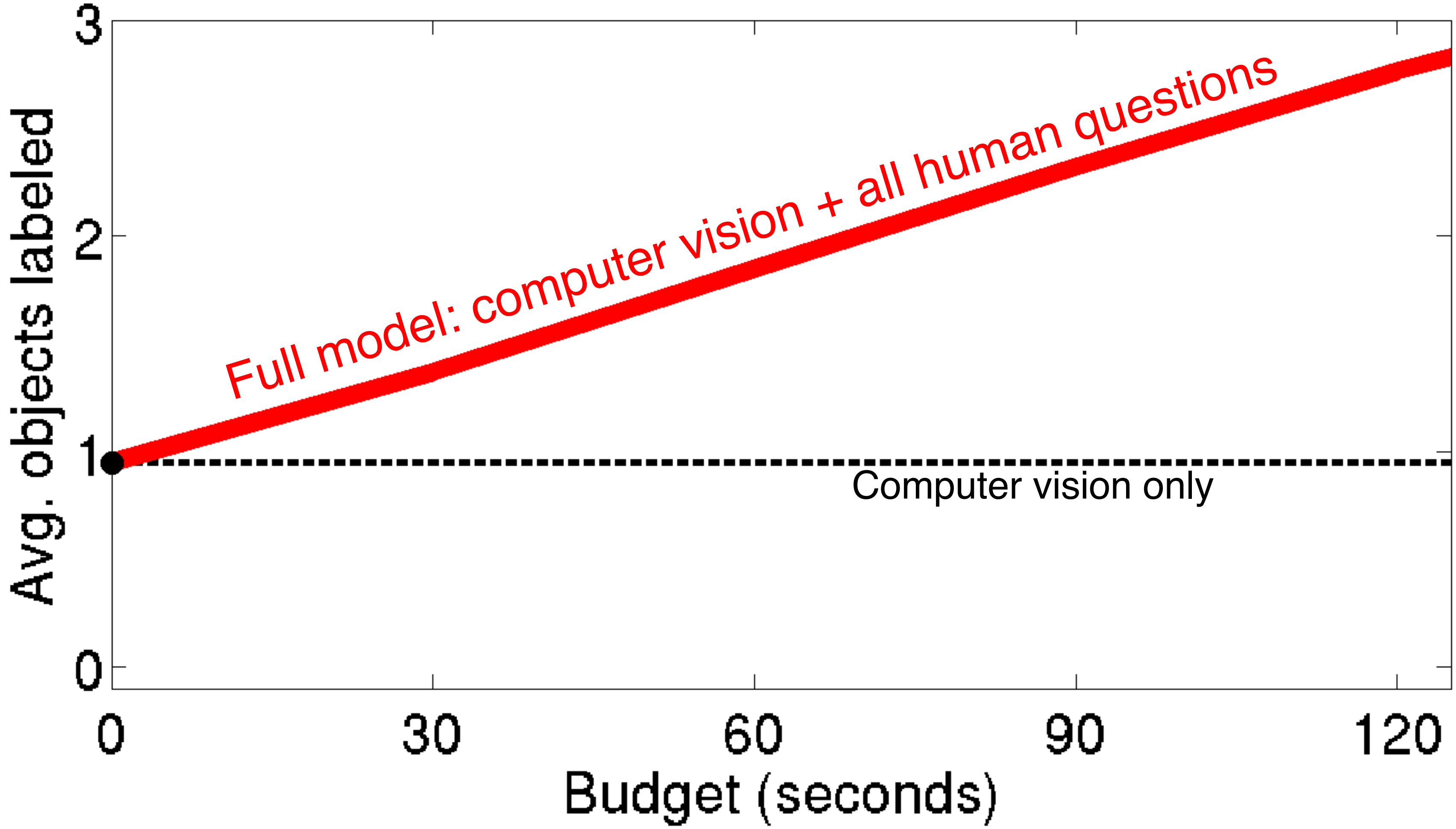
2K images of ILSVRC2014 detection val set with at least 4 object instances
Human error rates computed from AMT experiments
Annotation experiments in simulation



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Results

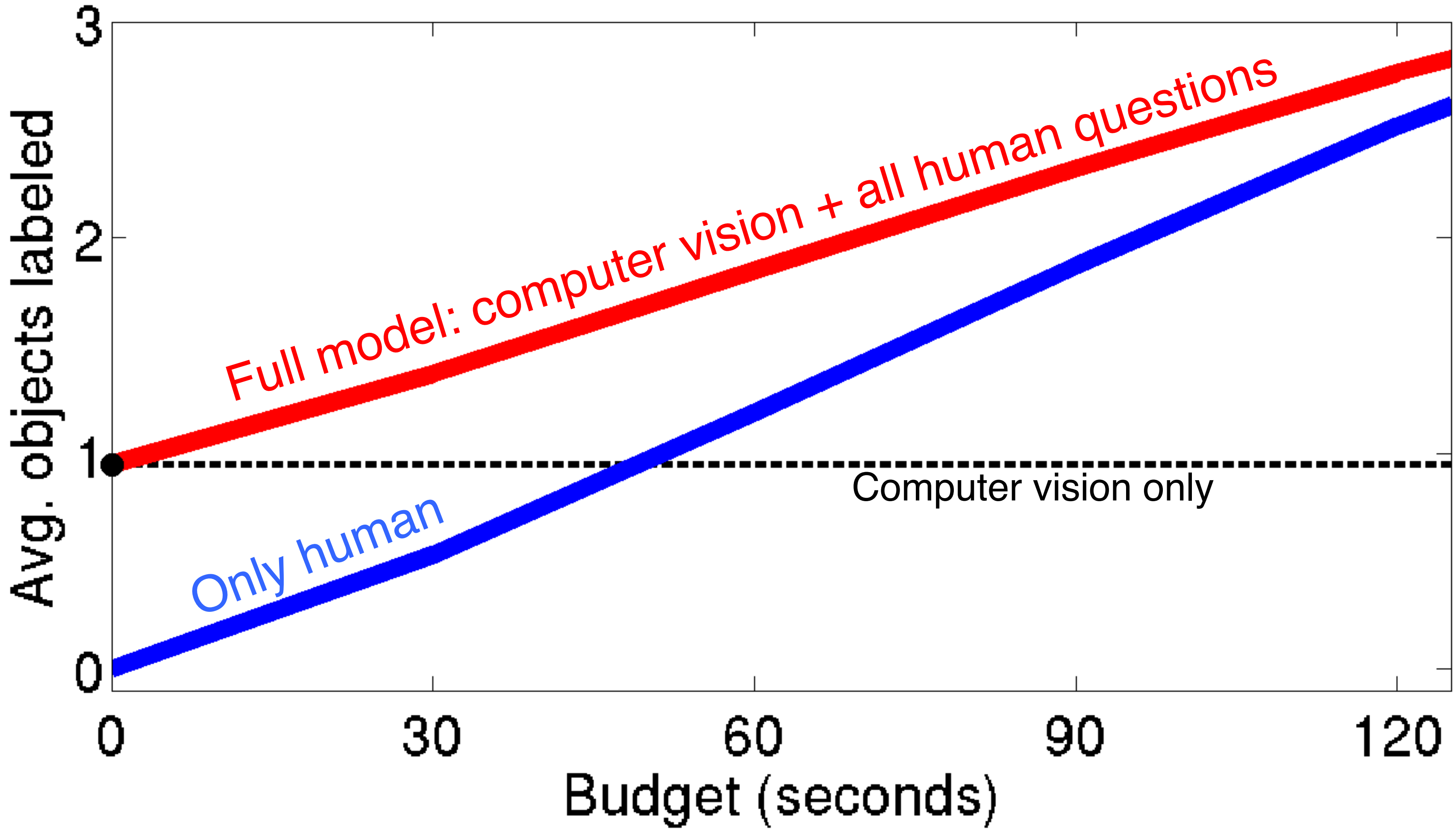
2K images of ILSVRC2014 detection val set with at least 4 object instances
Human error rates computed from AMT experiments
Annotation experiments in simulation



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Results

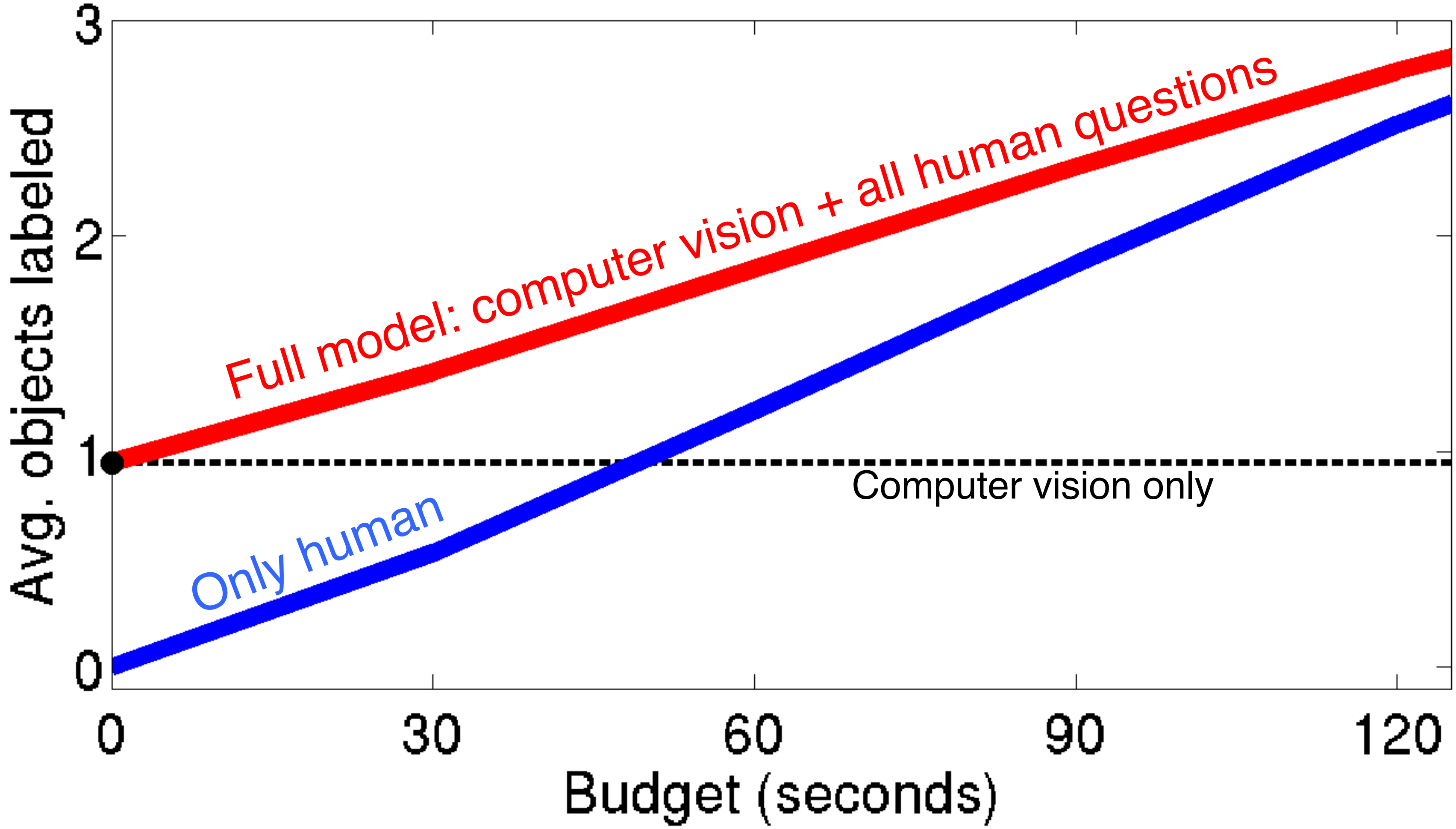
2K images of ILSVRC2014 detection val set with at least 4 object instances
Human error rates computed from AMT experiments
Annotation experiments in simulation



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Takeaways

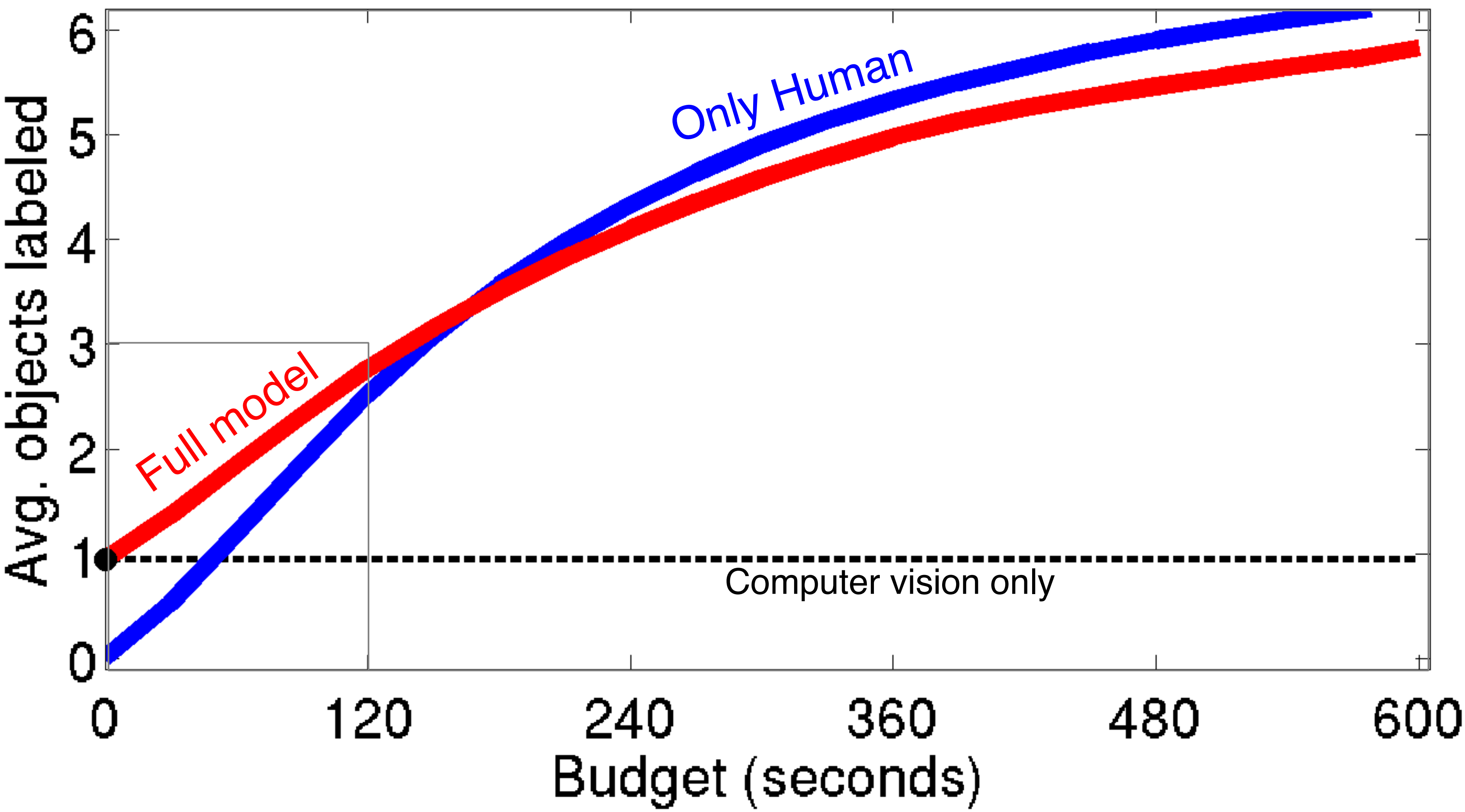
1) CV and humans are mutually beneficial



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Takeaways

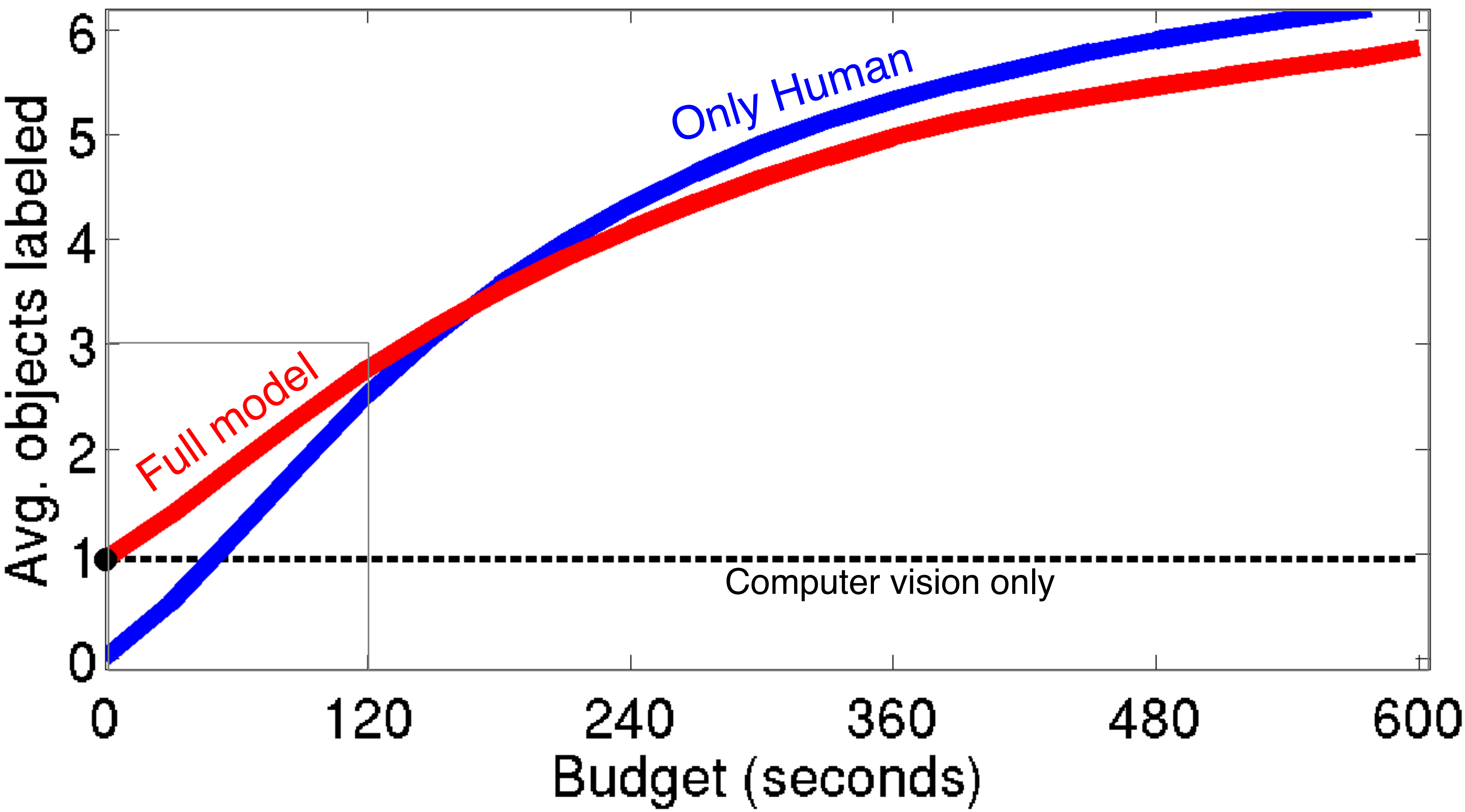
1) CV and humans are mutually beneficial



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Takeaways

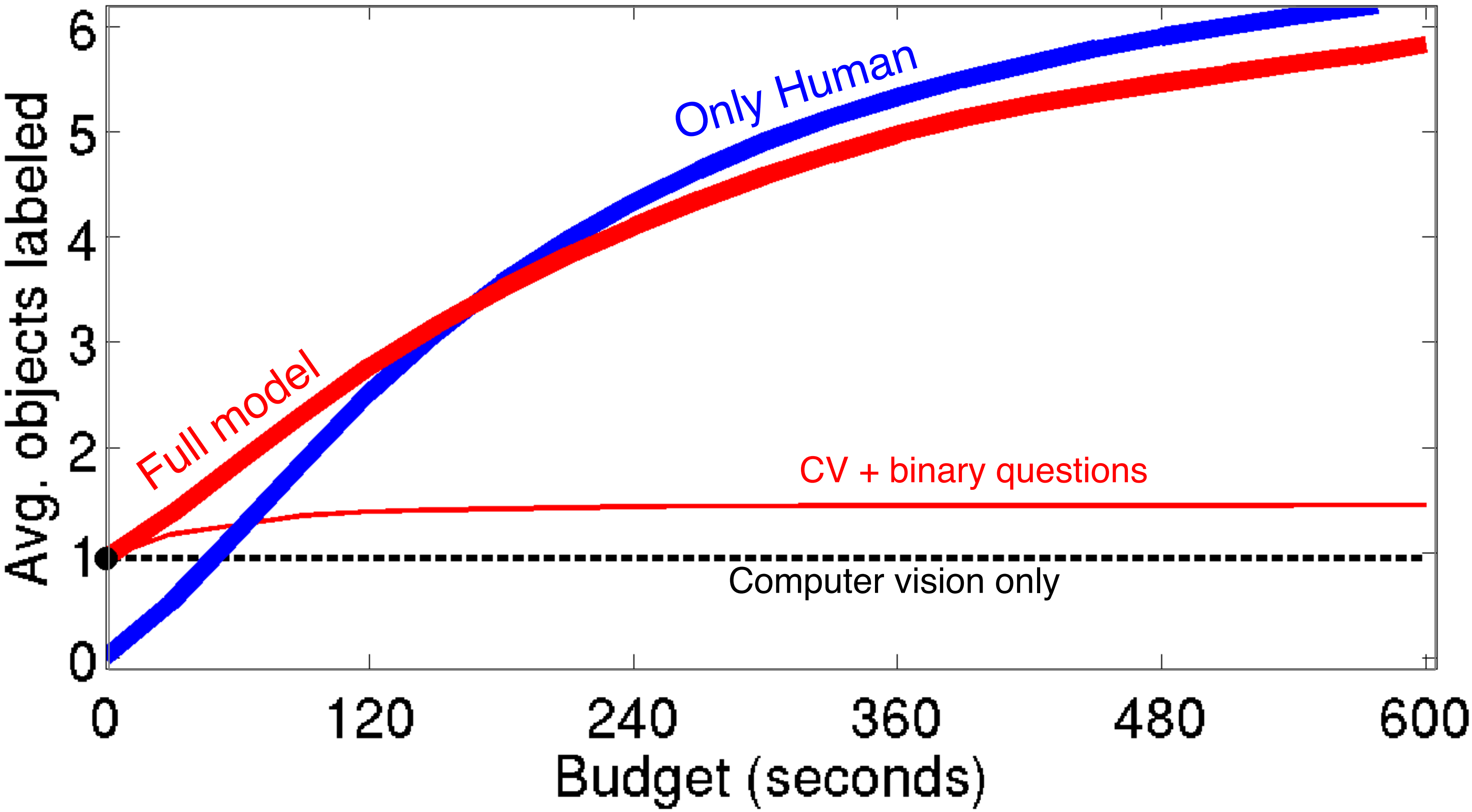
1) CV and humans are mutually beneficial
2) CV models are not perfectly calibrated



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Takeaways

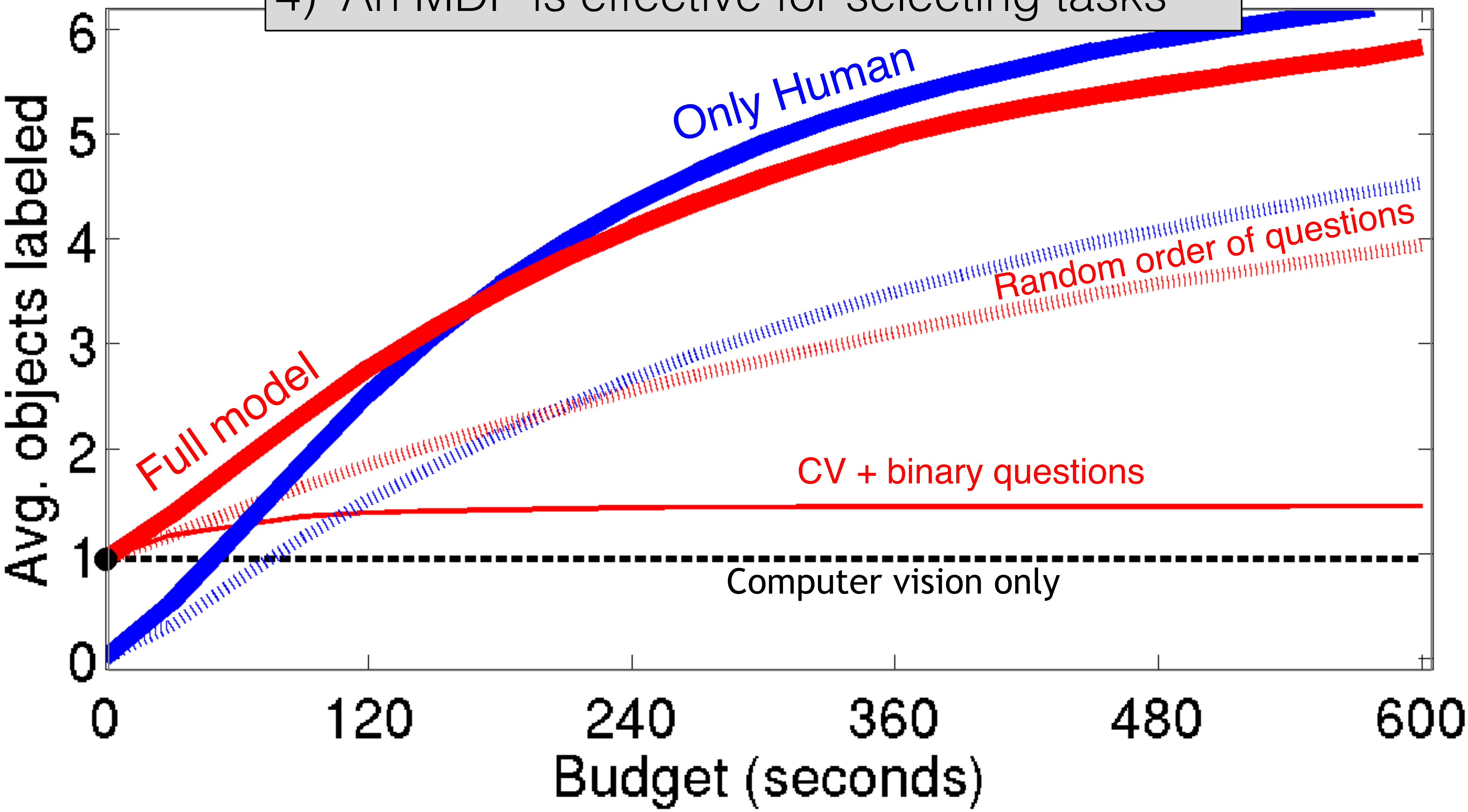
1) CV and humans are mutually beneficial
2) CV models are not perfectly calibrated
3) Complex human tasks are necessary



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Takeaways

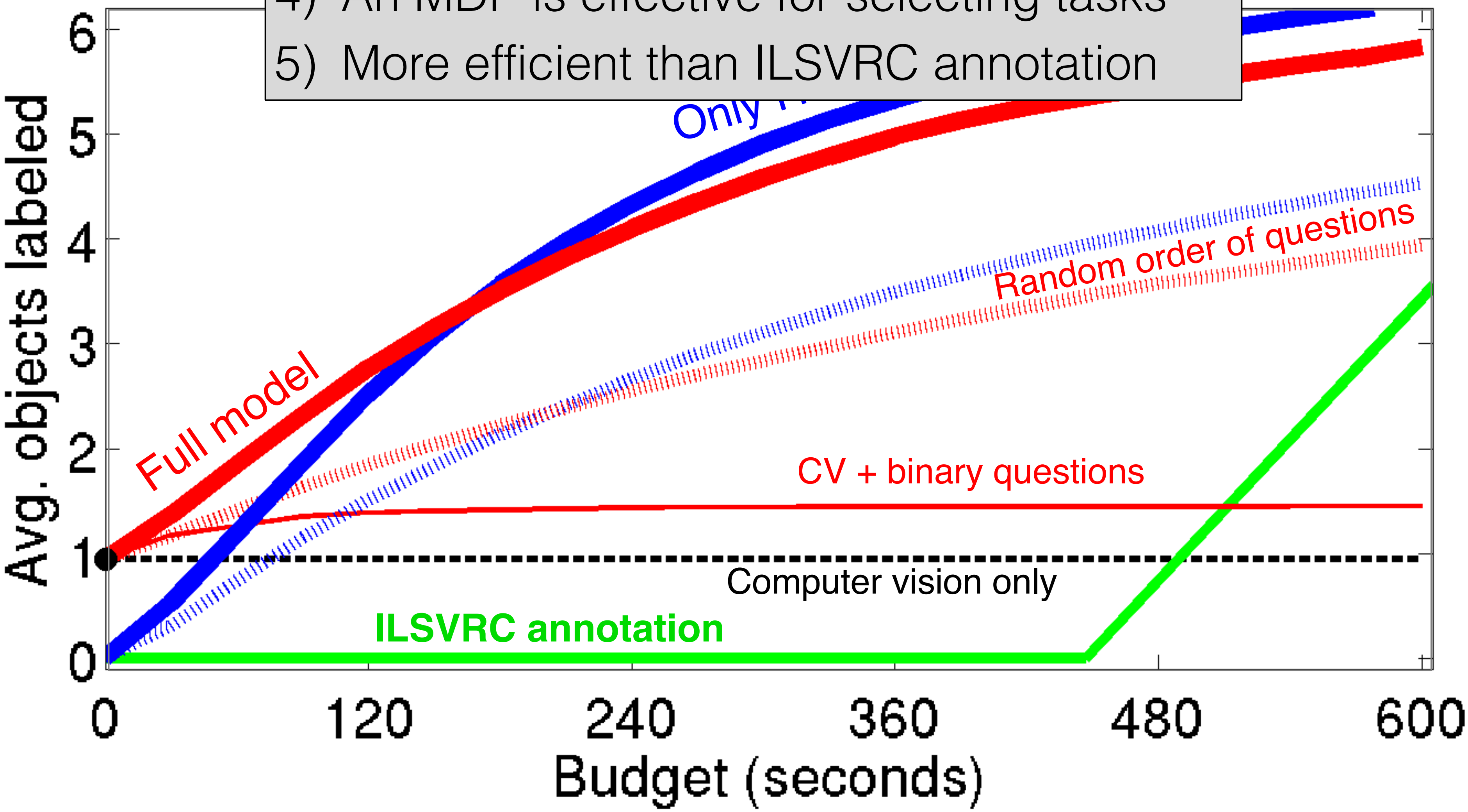
1) CV and humans are mutually beneficial
2) CV models are not perfectly calibrated
3) Complex human tasks are necessary
4) An MDP is effective for selecting tasks



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

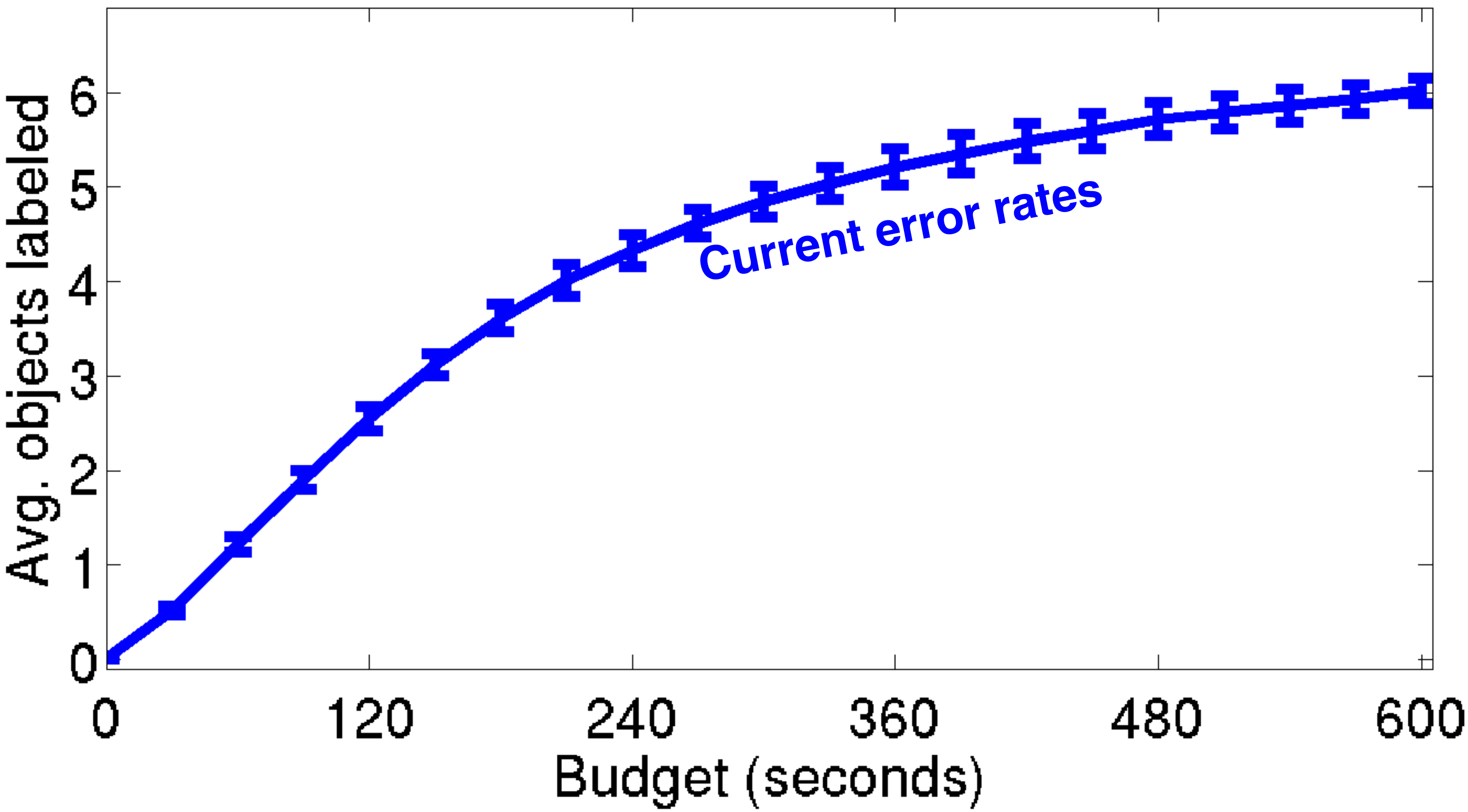
Takeaways

1) CV and humans are mutually beneficial
2) CV models are not perfectly calibrated
3) Complex human tasks are necessary
4) An MDP is effective for selecting tasks
5) More efficient than ILSVRC annotation



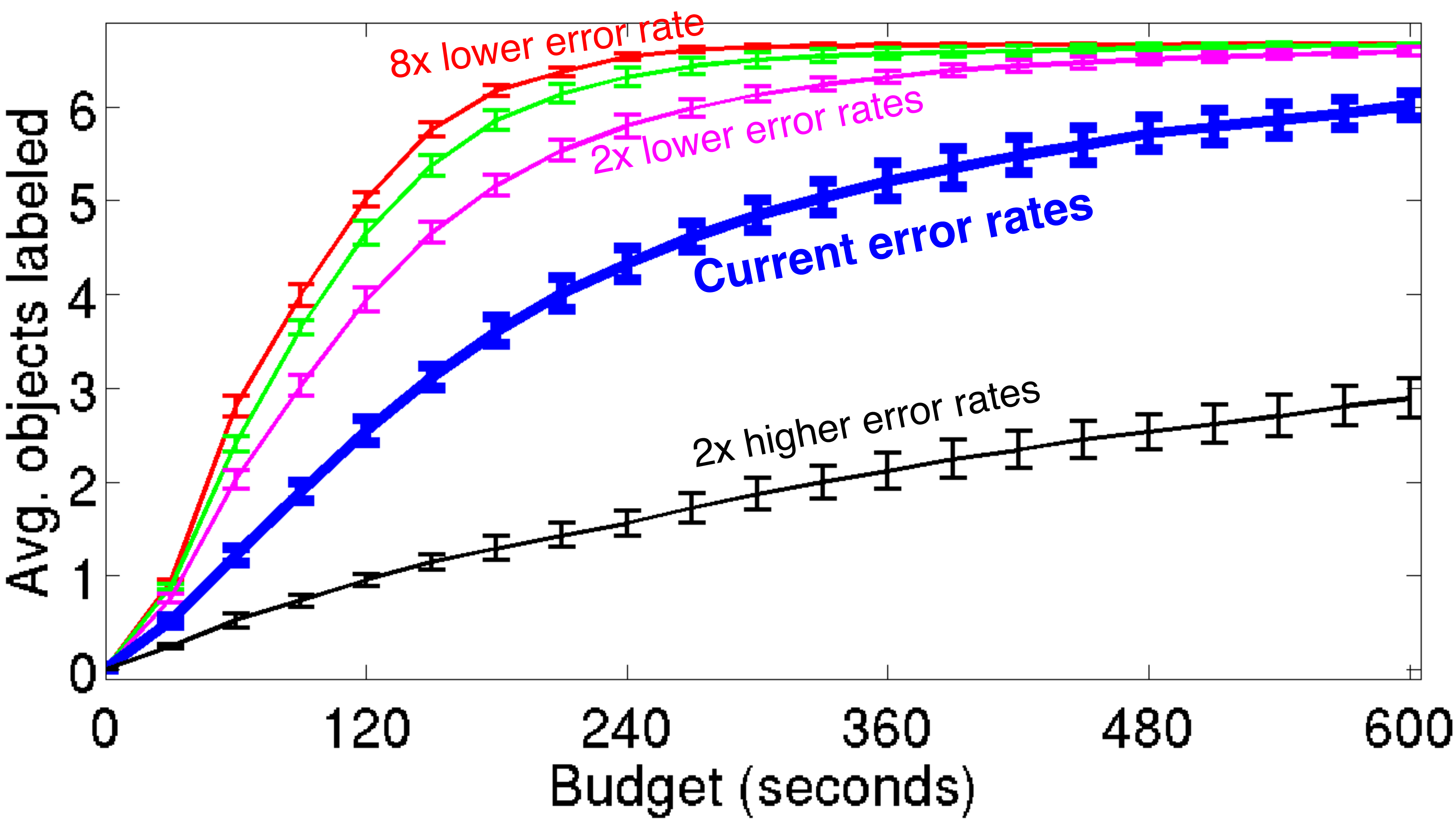
O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

What if humans were better?



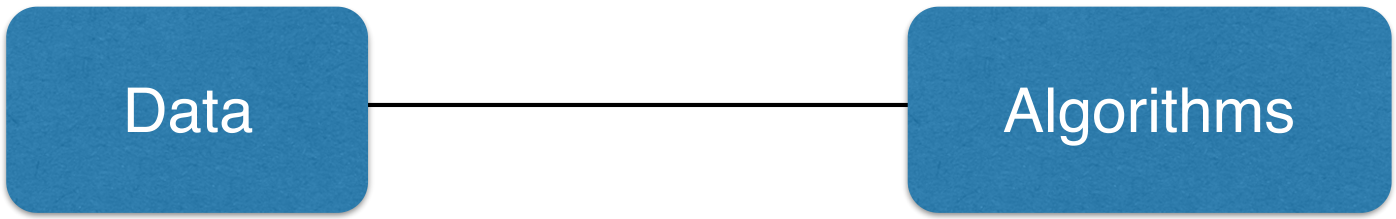
O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

What if humans were better?



O Russakovsky et al. Best of both worlds: human-machine collaboration for object annotation. CVPR 2015.

Scaling up object detection

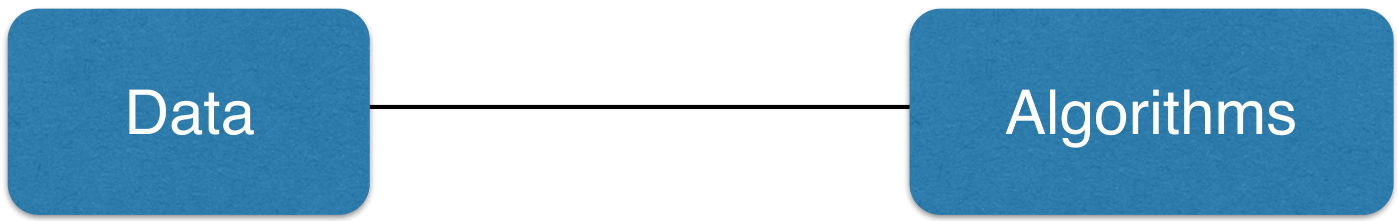


1) **Scaled up the data** by formulating data annotation as an optimization [CHI14, IJCV15]

2) **Developed algorithms** [CVPR10, ECCV12, CVPR15b] and **performed large-scale analysis** to gain insight into the state of the field [ICCV13, IJCV15]

3) Combine insights from both

Scaling up object detection



1) **Scaled up the data** by formulating data annotation as an optimization [CHI14, IJCV15]

2) **Developed algorithms** [CVPR10, ECCV12, CVPR15b] and **performed large-scale analysis** to gain insight into the state of the field [ICCV13, IJCV15]

3) Created a **principled framework** for image understanding using crowd engineering insights and state-of-the-art vision algorithms [CVPR15a]

Bird's-eye view of my research

A. Computer vision (& machine learning)

1. Object recognition: scale and analysis [**ICCV13**, **IJCV15**], accuracy [ICRA10, ECCV12 CVPR15b], efficiency [CVPR10], attributes [ECCVW10]
2. Holistic scene understanding: scene classification [UnderReviewA], semantic segmentation [UnderReviewB],
3. Video understanding: human action detection [TechReport15, UnderReviewC]

Bird's-eye view of my research

A. Computer vision (& machine learning)

1. Object recognition: scale and analysis [**ICCV13**, **IJCV15**], accuracy [ICRA10, ECCV12 CVPR15b], efficiency [CVPR10], attributes [ECCVW10]
2. Holistic scene understanding: scene classification [UnderReviewA], semantic segmentation [UnderReviewB],
3. Video understanding: human action detection [TechReport15, UnderReviewC]

B. Human-in-the-loop machine learning

Bird's-eye view of my research

A. Computer vision (& machine learning)

1. Object recognition: scale and analysis [**ICCV13**, **IJCV15**], accuracy [ICRA10, ECCV12, CVPR15b], efficiency [CVPR10], attributes [ECCVW10]
2. Holistic scene understanding: scene classification [UnderReviewA], semantic segmentation [UnderReviewB],
3. Video understanding: human action detection [TechReport15, UnderReviewC]

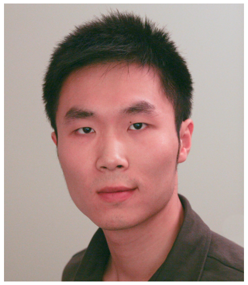
B. Human-in-the-loop machine learning

1. Teaching: crowd engineering [**CHI14, IJCV15**], tradeoff between annotation cost and accuracy [UnderReviewB]
2. Active learning
3. Practical human-and-CV collaborations [**CVPR15a**]

Acknowledgements

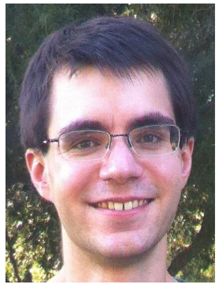
ILSVRC team

http://image-net.org/challenges/LSVRC ilsvrc@image-net.org



Jia Deng
(U. of Michigan)

Hao Su
(Stanford U.)

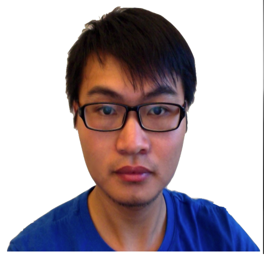


Jonathan Krause
(Stanford U.)

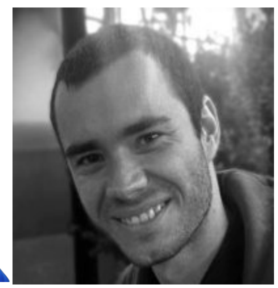
Sanjeev Satheesh
(Stanford U.)

Sean Ma
(Stanford U.)

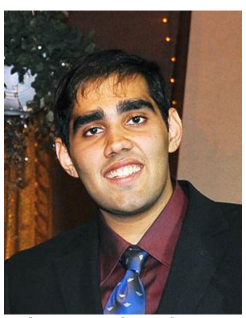
Zhiheng Huang
(Stanford U.)



Wei Liu
(UNC Chapel Hill)



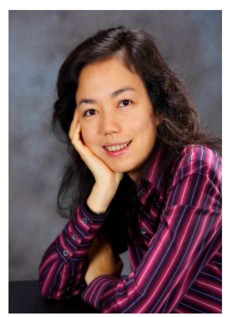
Andrej Karpathy
(Stanford U.)



Aditya Khosla
(MIT)

Michael Bernstein
(Stanford U.)

Alexander Berg
(UNC Chapel Hill)



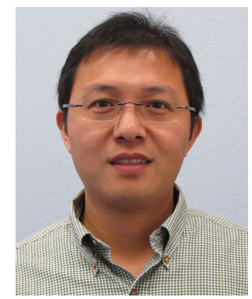
Fei-Fei Li
(Stanford U.)

Other co-authors

Amy Bearman
(Stanford U.)

Vittorio Ferrari
(U. of Edinburgh)

Jia Li
(Snapchat)



Yuanqing Lin
(NEC Labs)

Andrew Ng
(Stanford U.)

Kai Yu
(Baidu)

facebook Google NVIDIA

Questions?

http://cs.cmu.edu/~orussako olgarus@cmu.edu

Scaling up object detection

1) **Scaled up the data** by formulating data annotation as an optimization [CHI14, IJCV15]

2) **Developed algorithms** [CVPR10, ECCV12, CVPR15b] and **performed large-scale analysis** to gain insight into the state of the field [ICCV13, IJCV15]

3) Created a **principled framework** for image understanding using crowd engineering insights and state-of-the-art vision algorithms [CVPR15]

Bird's-eye view

A. Computer vision (& machine learning): Pixel-level image understanding [CVPR10, ECCVW10, ECCV12, ICCV13, CVPR15b, IJCV15, UnderReviewA, UnderReviewB] , video understanding, [TechReport15, UnderReviewC]

B. Human-in-the-loop machine learning: Crowd engineering [CHI14, IJCV15], tradeoff between human cost and accuracy [UnderReviewB], practical human-and-CV collaborations [CVPR15a]