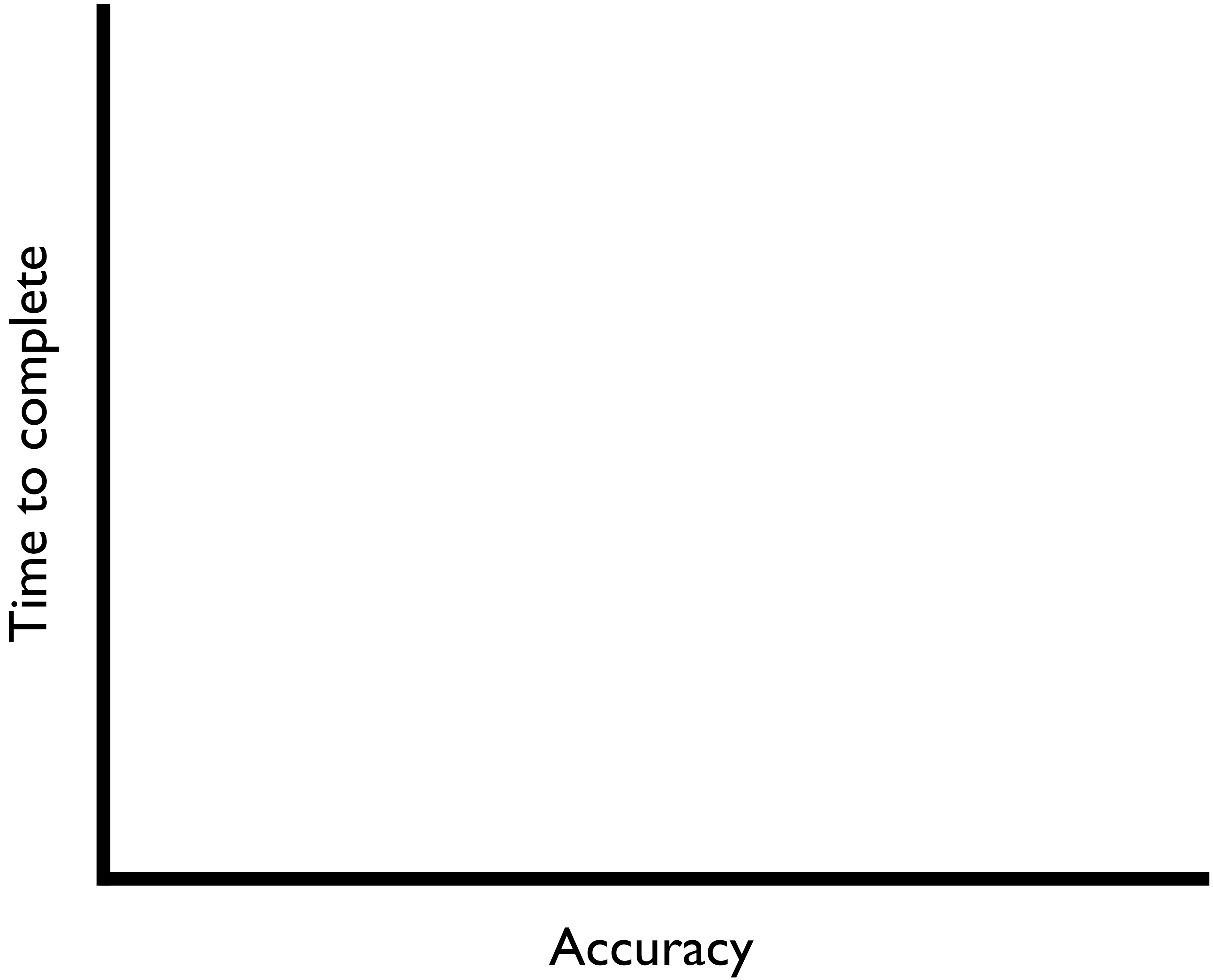
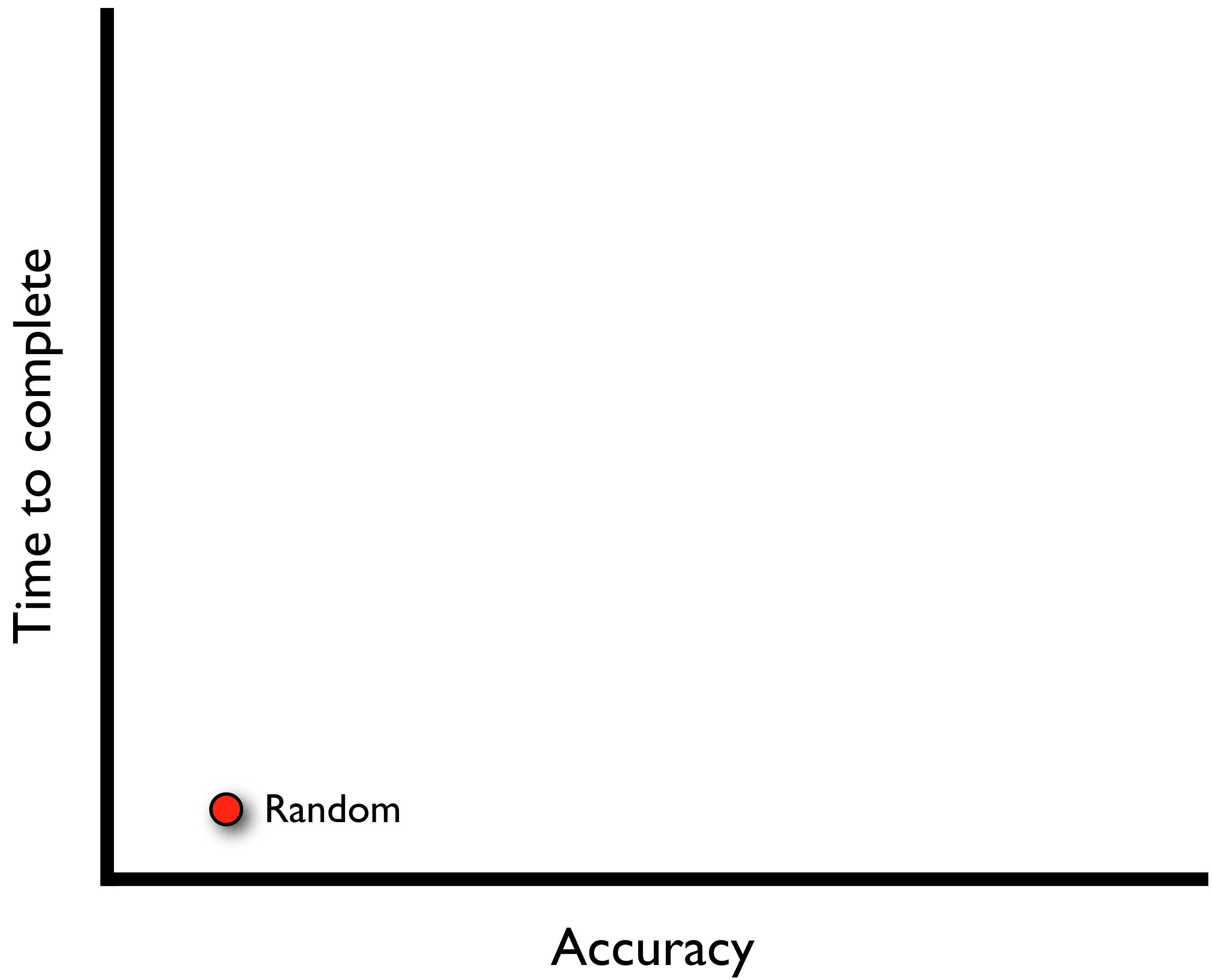
Image Spirit: Verbal Guided Image Parsing

M. M. Cheng et al.
Presented by Brian Dolhansky

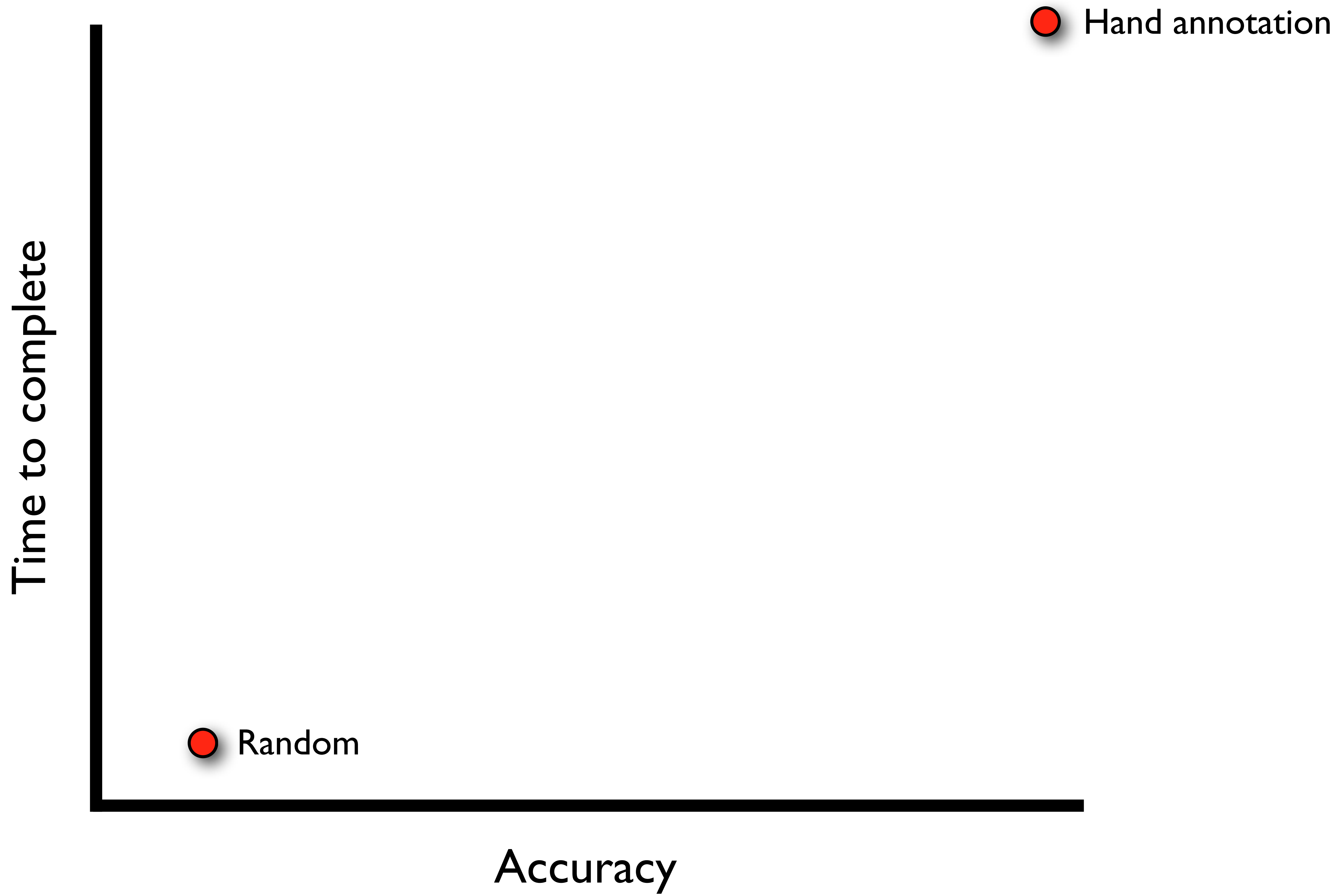
Completing a difficult task



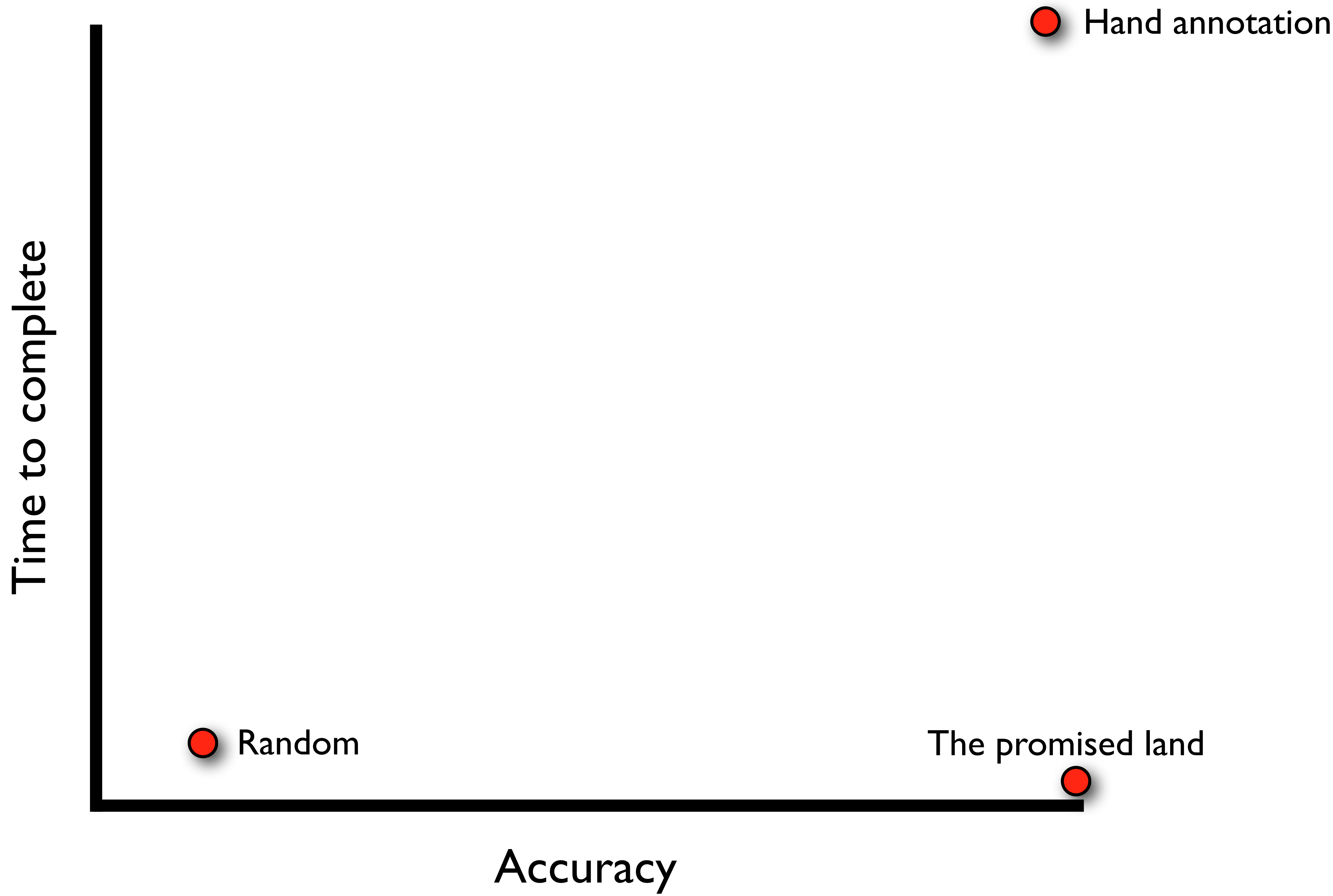
Completing a difficult task



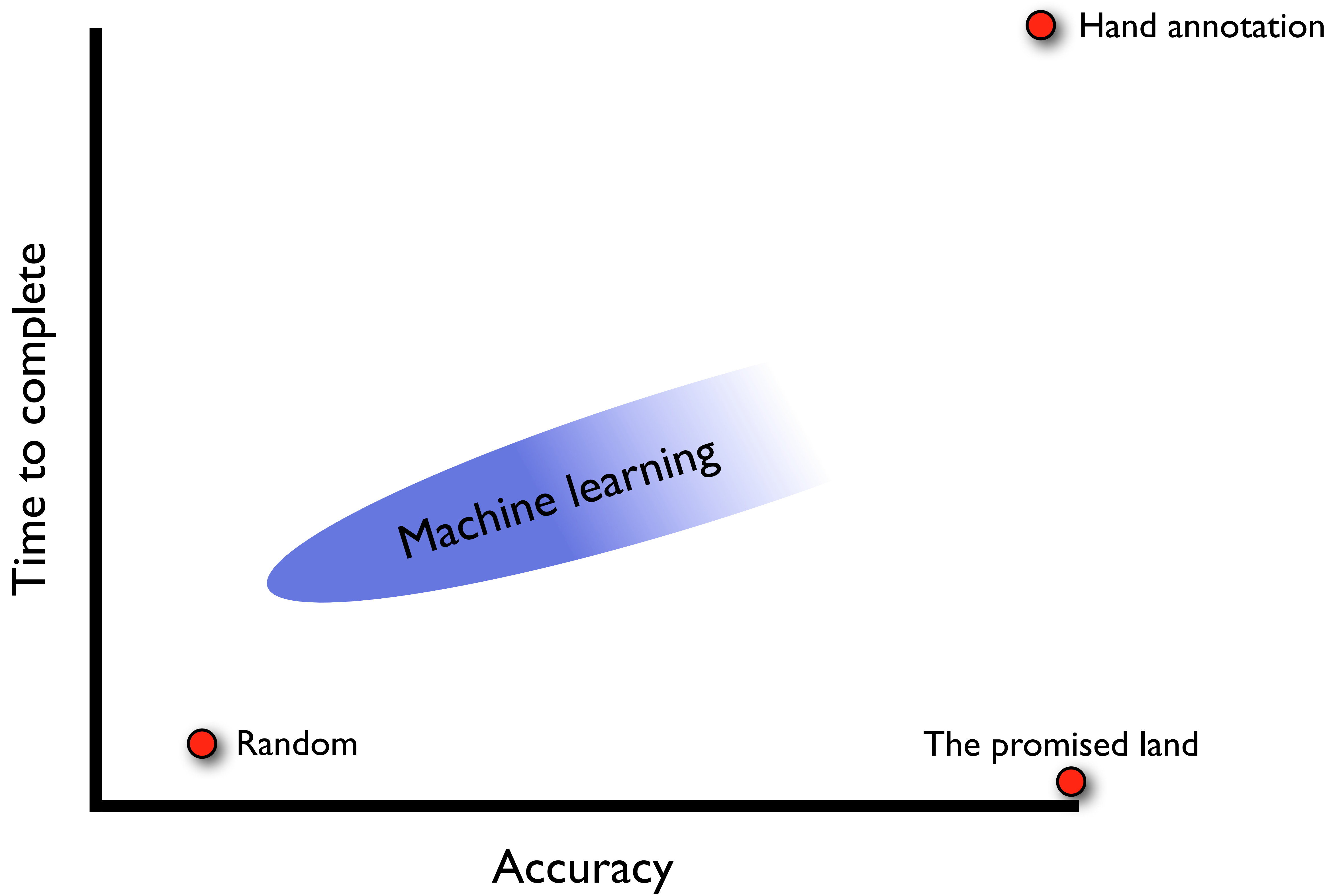
Completing a difficult task



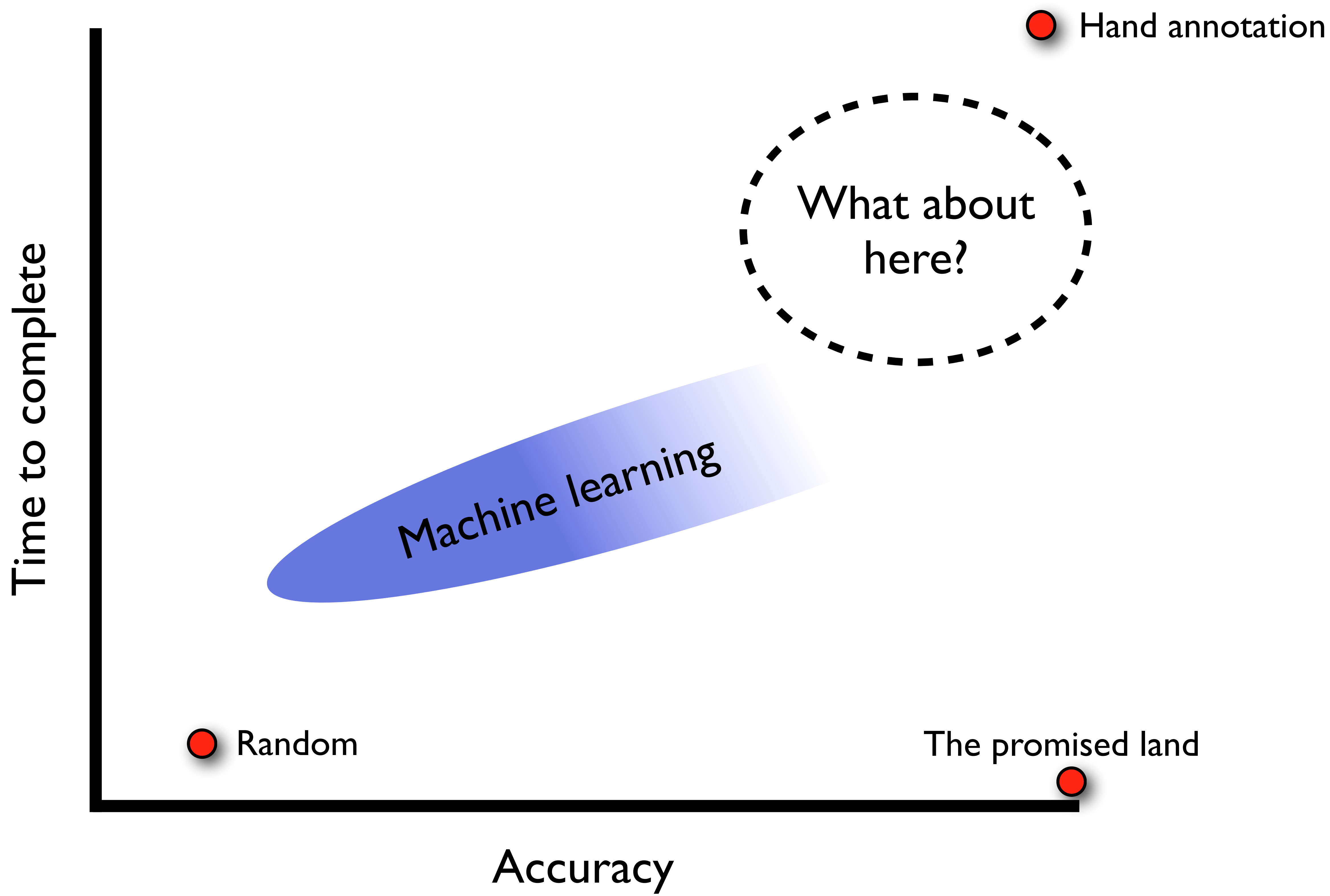
Completing a difficult task



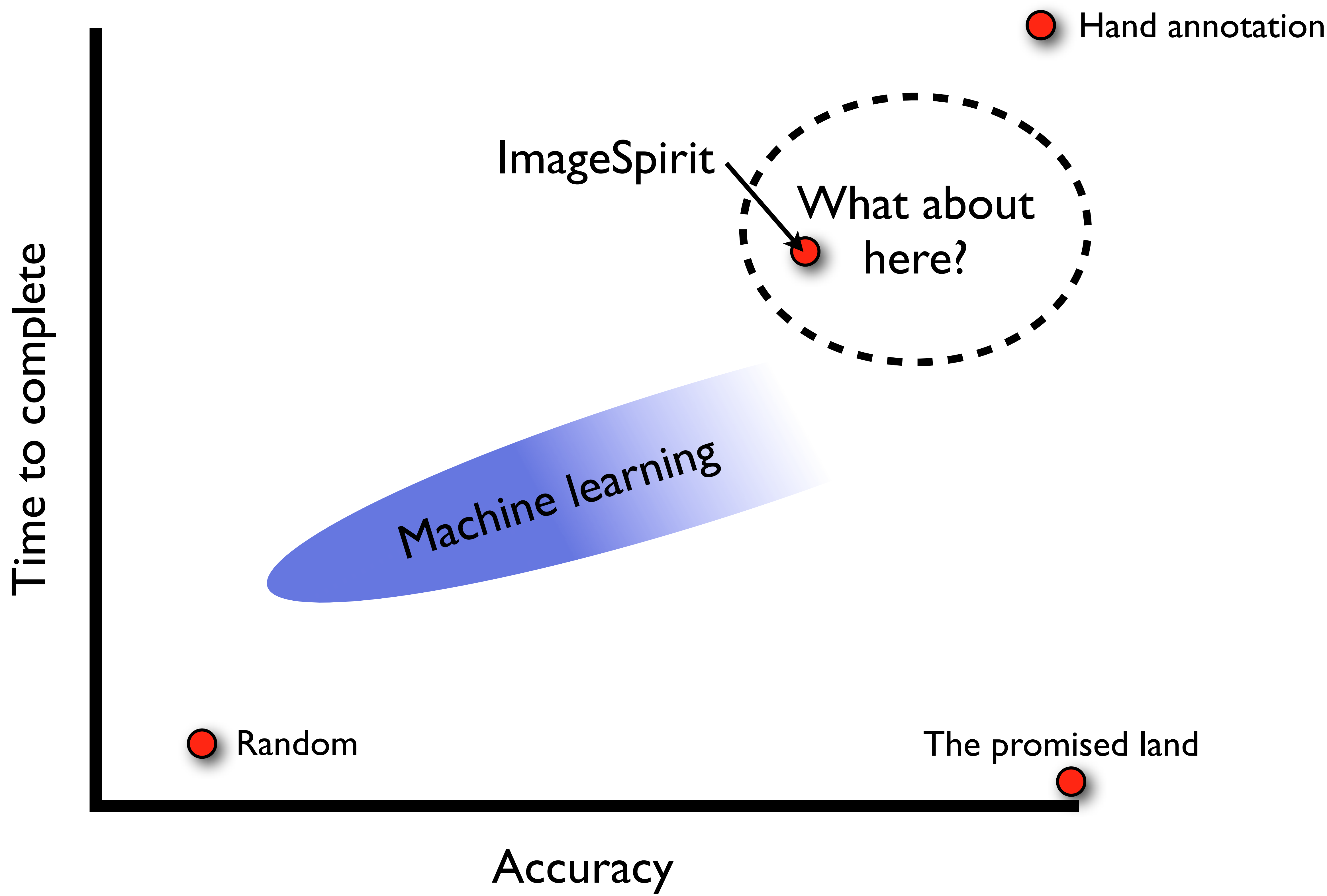
Completing a difficult task



Completing a difficult task



Completing a difficult task



ImageSpirit

(1) Segment an image with object/**attribute** labels...

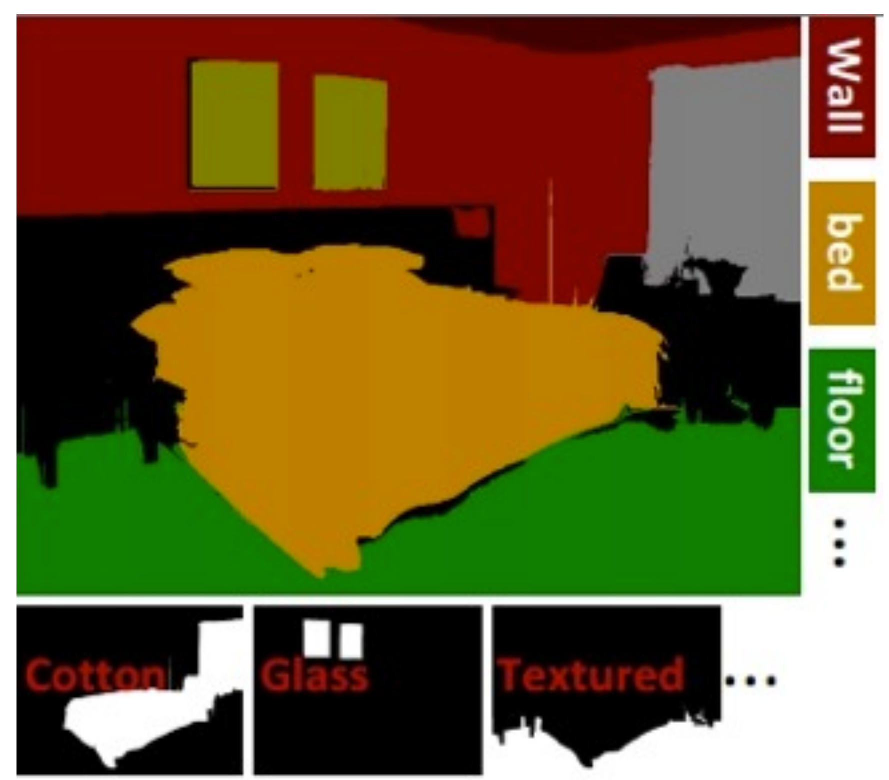
object: bed, *attributes*: {cotton, textured}

ImageSpirit

(1) Segment an image with object/**attribute** labels...

object: bed, *attributes*: {cotton, textured}

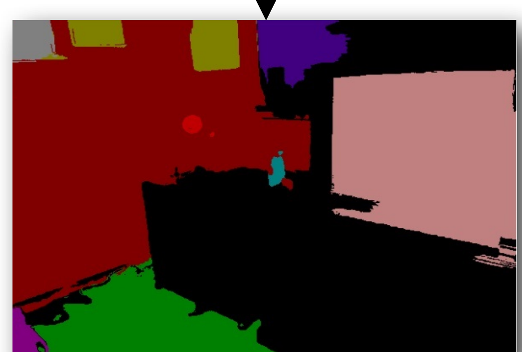
(2) ... and refine with verbal input



System Overview

Source image

Object/
Attribute
Model

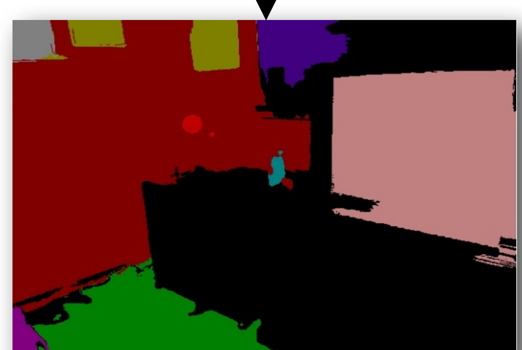


Initial per-pixel
segmentation

System Overview

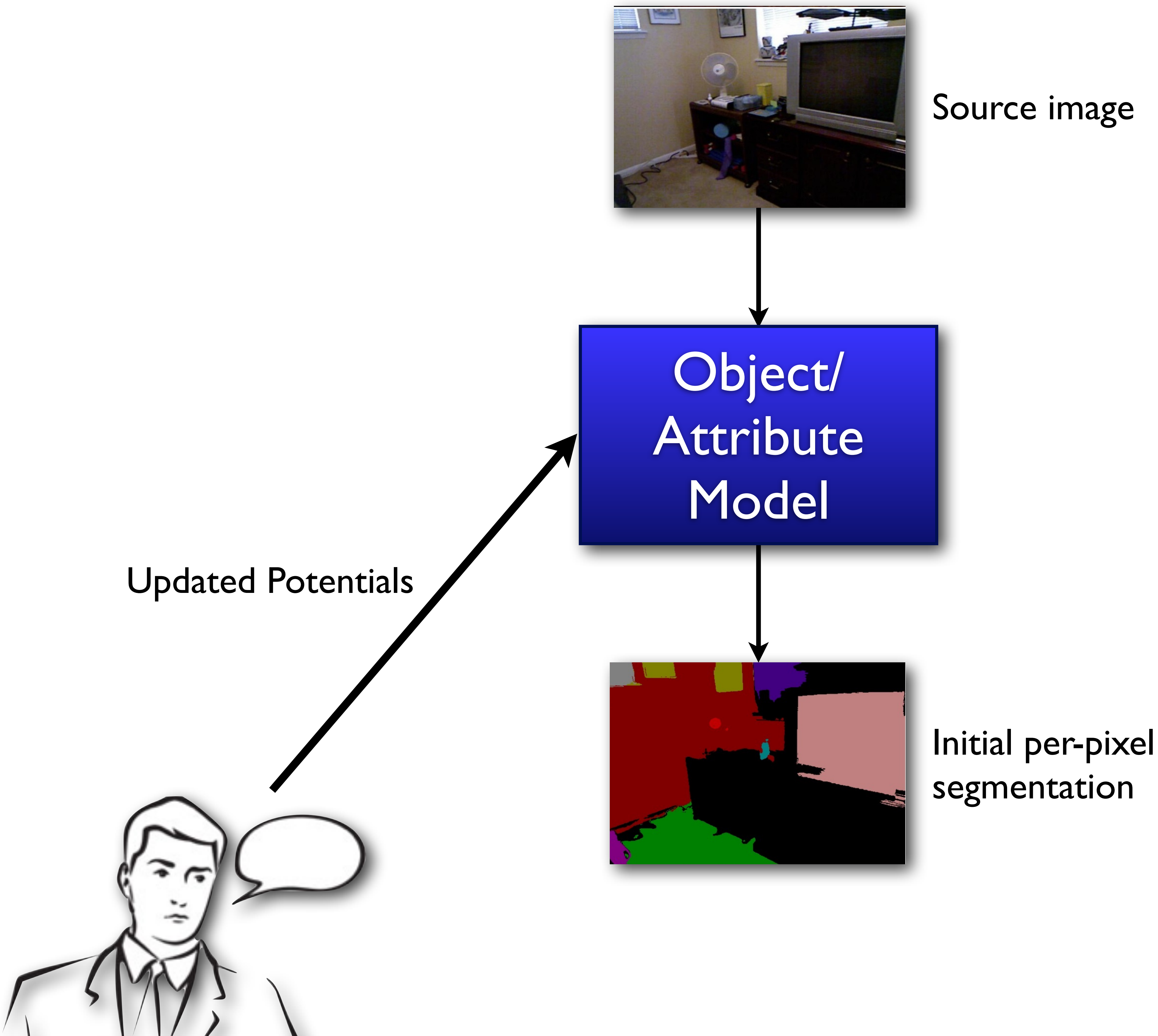
Source image

Object/
Attribute
Model



Initial per-pixel
segmentation

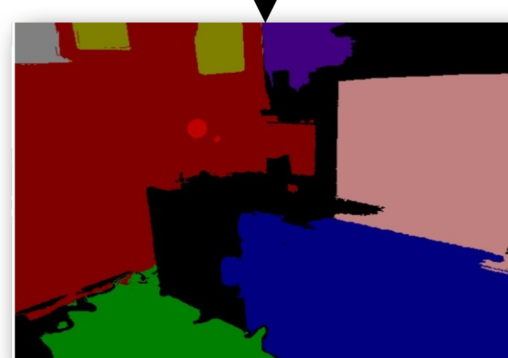
System Overview



System Overview

Source image

Object/
Attribute
Model++

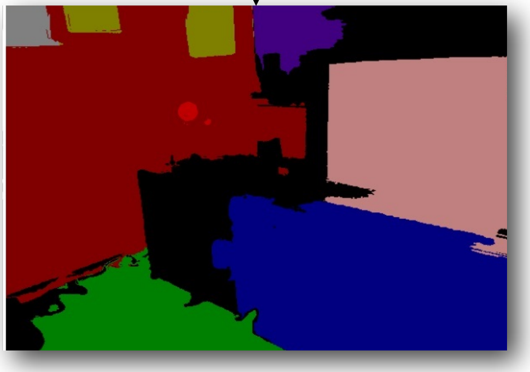


Refined segmentation

System Overview

Source image

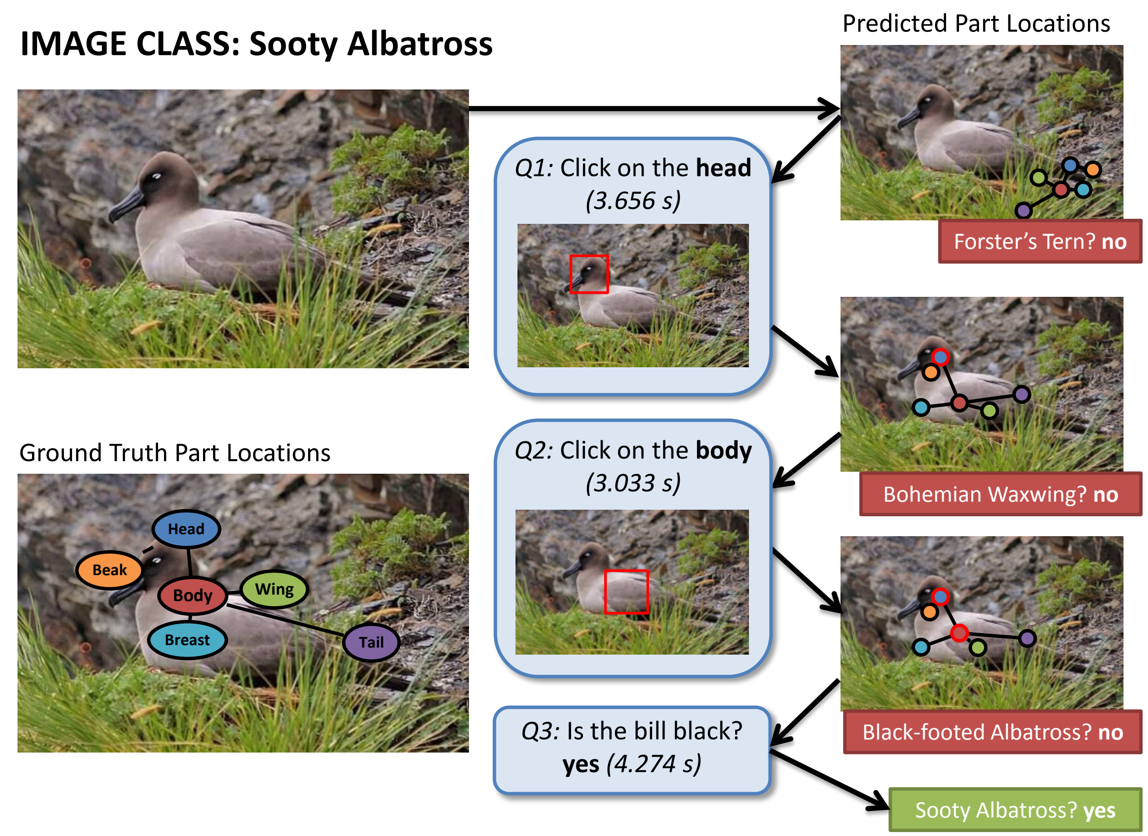
Object/
Attribute
Model++



Refined segmentation

Related Work - Humans in the Loop

Seeded Graph Cut for Image Segmentation*

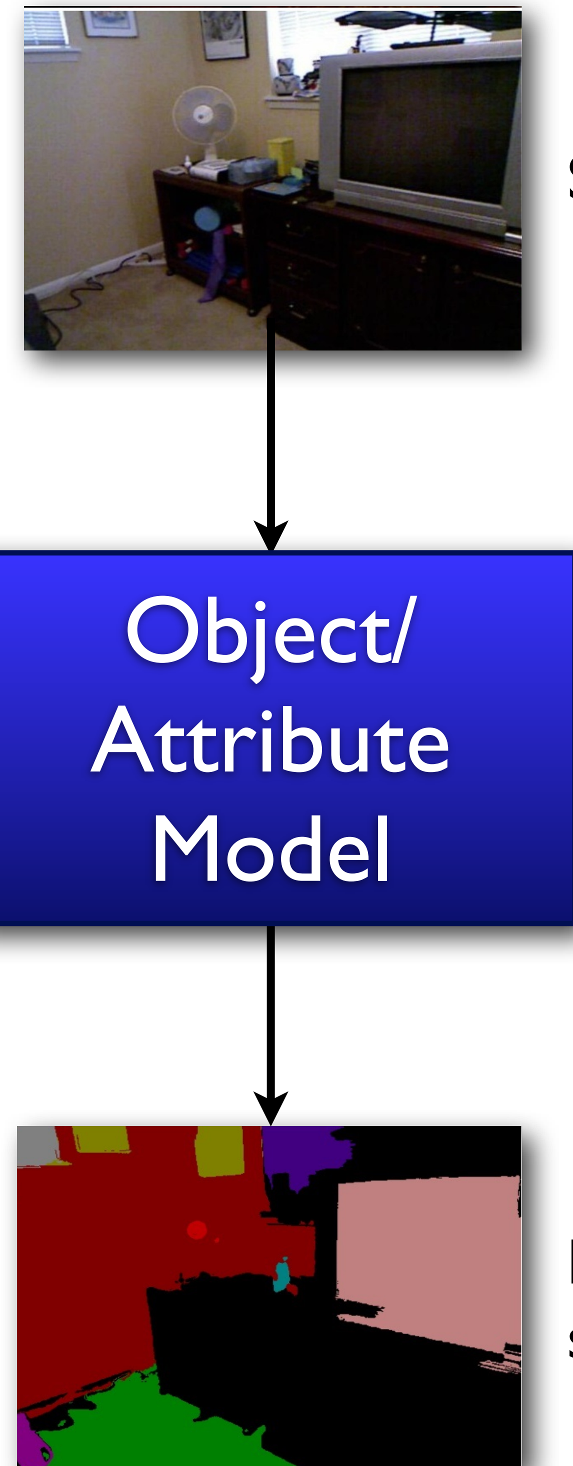


Verbal Cues for Object Recognition/Classification°

°Sinop et al. 2007

°Branson et al. 2010

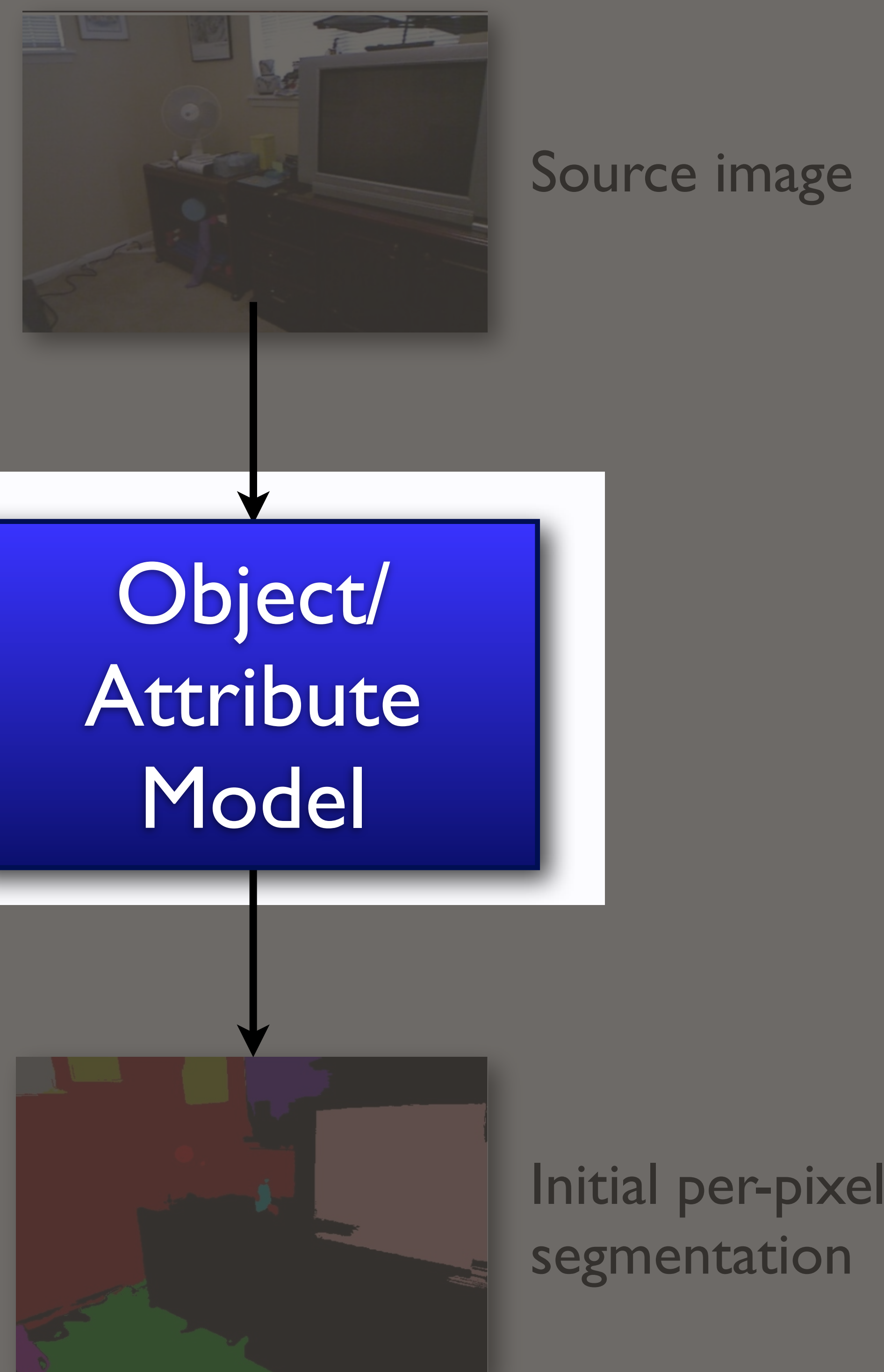
System Overview



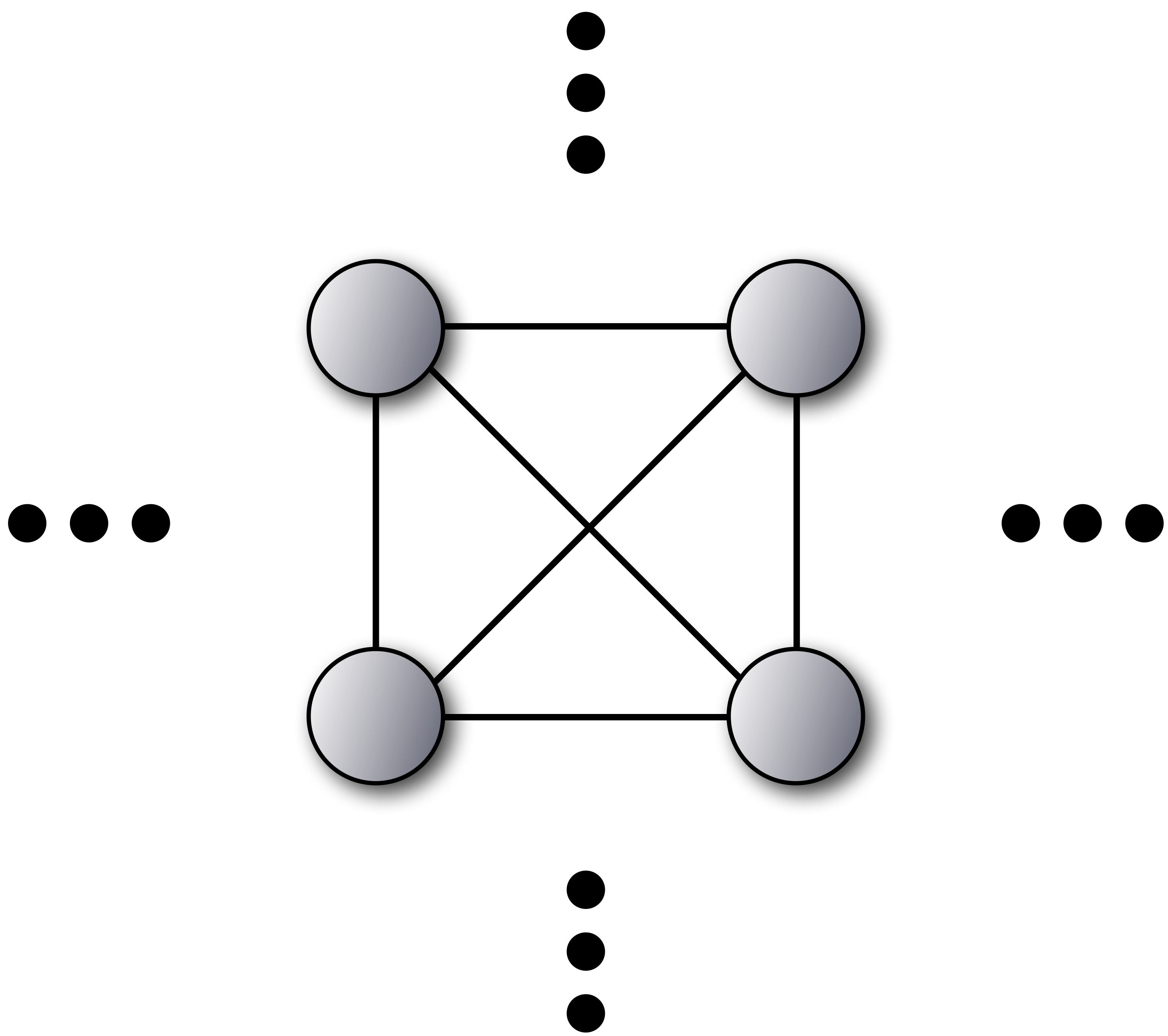
Source image

Initial per-pixel segmentation

System Overview



Fully-connected CRF



Krahenbuhl and Koltun, 2011

Fully-connected CRF

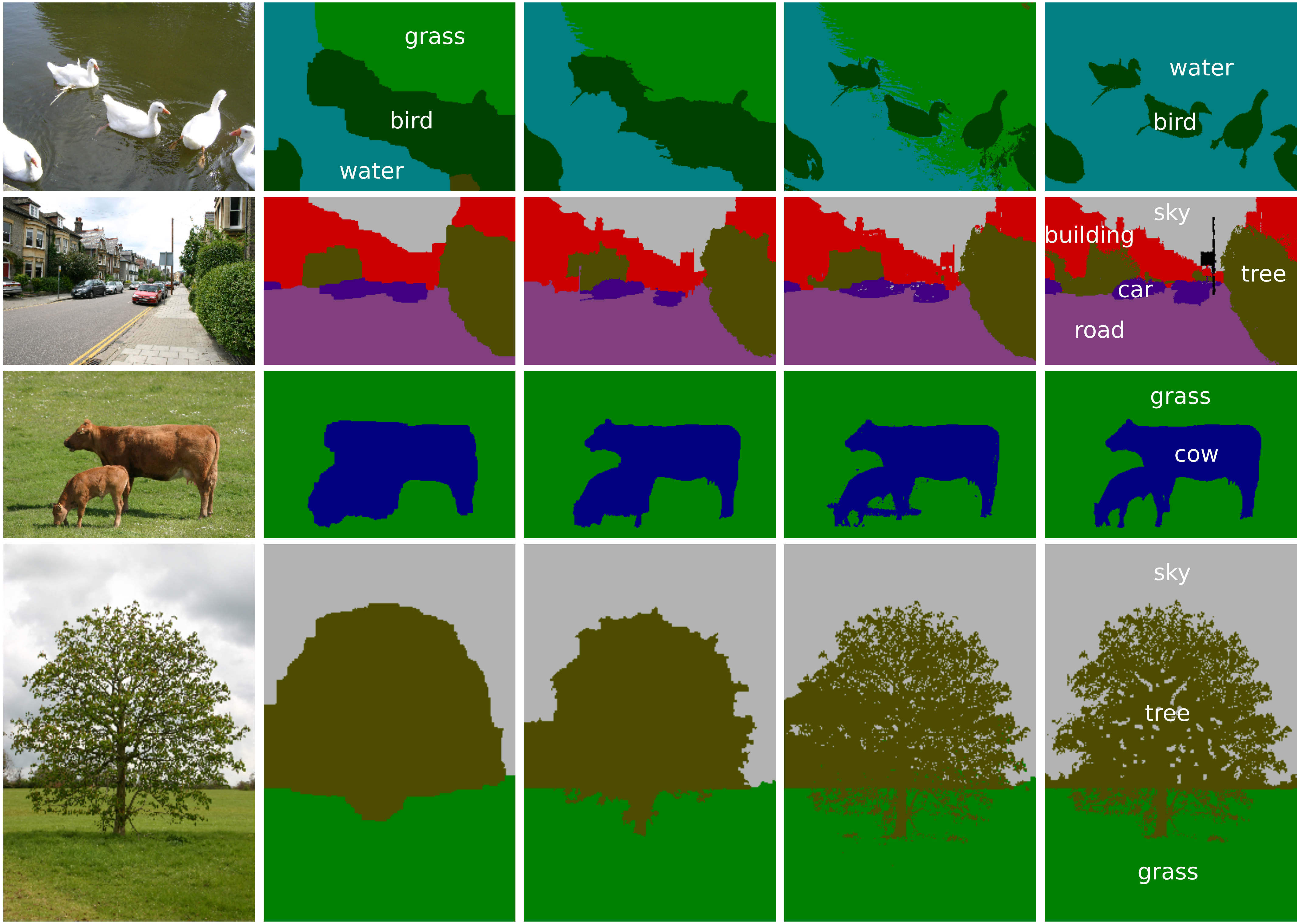


Image | Grid CRF | Robust P^n CRF | Our approach | Accurate ground truth

Krahenbuhl and Koltun, 2011

Object/Attribute Model

Per-pixel labels: $Z_i = (X_i, Y_i)$

Object label: $X_i \in \mathcal{O}$

e.g.: $x_i =$ cabinet $x_i =$ chair

Attribute label: $Y_i \in \mathcal{P}(\mathcal{A})$

e.g.: $y_i = \emptyset$ $y_i =$ {wood}

$y_i =$ {wood, painted, textured}

Joint configuration: $\mathbf{z} = \{Z_1 = z_1, Z_2 = z_2, ...Z_N = z_N\}$

Image data: $\mathbf{I} \in \mathbb{R}^3$

Object/Attribute Model

Fully-connected CRF decomposition:

$$E(\mathbf{z}) = \sum_i \psi(z_i) + \sum_{i<j} \psi(z_i, z_j)$$

Unary terms: enforce object/attribute assignments

Pairwise terms: enforce consistent labelings between nearby pixels

Object/Attribute Model

Fully-connected CRF decomposition:

$$E(\mathbf{z}) = \boxed{\sum_i \psi(z_i)} + \sum_{i<j} \psi(z_i, z_j)$$

Unary terms: enforce object/attribute assignments

Pairwise terms: enforce consistent labelings between nearby pixels

Unary Term

$$\psi_i(z_i) = \psi_i^{\mathcal{O}}(x_i) + \sum_a \psi_{i,a}^{\mathcal{A}}(y_{i,a}) + \sum_{o,a} \psi_{i,o,a}^{\mathcal{O},\mathcal{A}}(x_i, y_{i,a}) + \sum_{a \neq a'} \psi_{i,a,a'}^{\mathcal{A}}(y_{i,a}, y_{i,a'})$$

Pixel/object likelihood term: *

$$\psi_i^{\mathcal{O}}(x_i) = -\log(\Pr_{\mathcal{O}}(x_i | I_i))$$

Pixel/attribute likelihood term: *

$$\psi_{i,a}^{\mathcal{A}}(y_{i,a}) = -\log(\Pr_{\mathcal{A}}(y_{i,a} | I_i))$$

Object/attribute relationship term:

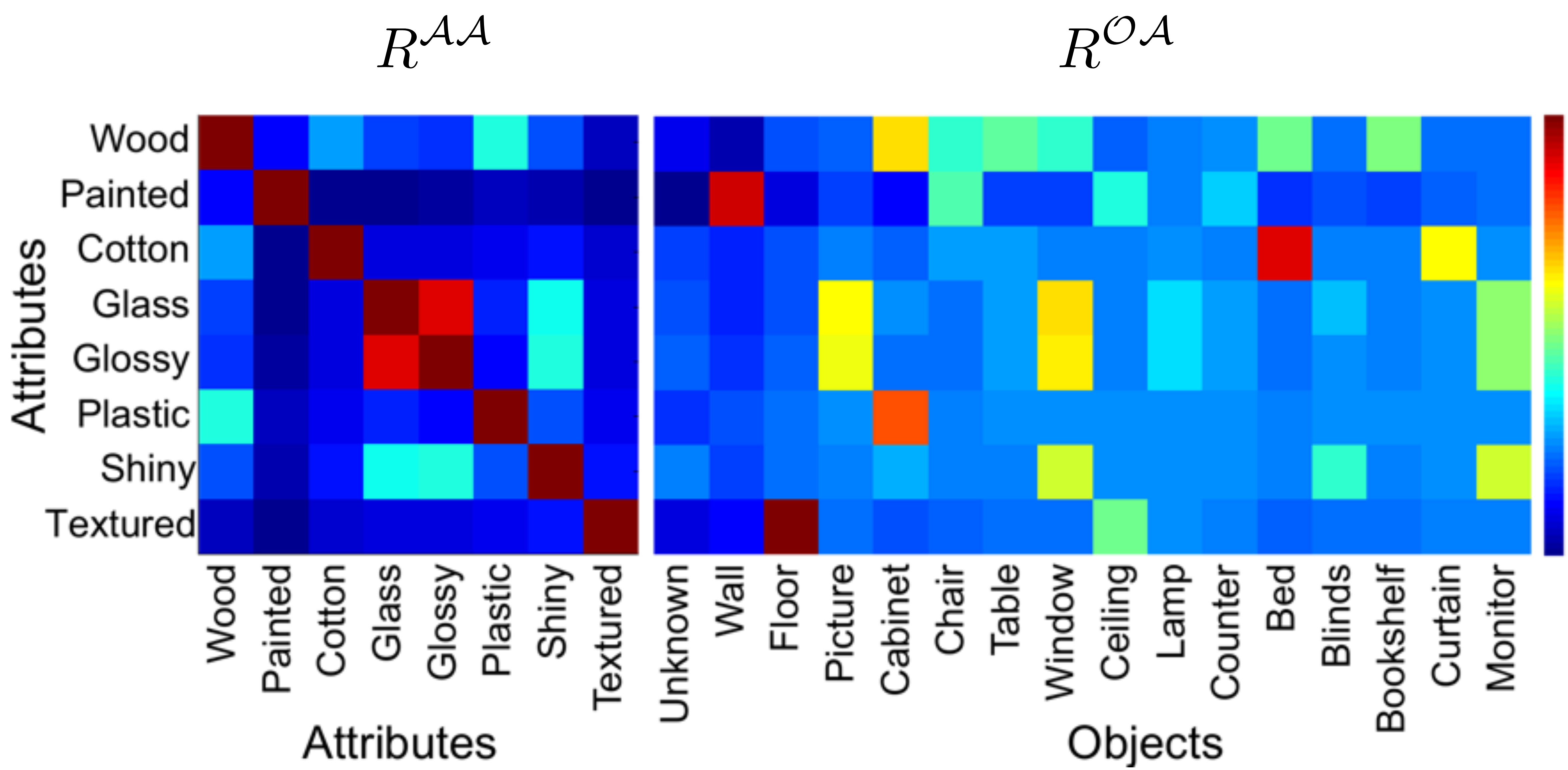
$$\psi_{i,o,a}^{\mathcal{O},\mathcal{A}}(x_i, y_{i,a}) = \mathbf{1}[\mathbf{1}[x_i = o] \neq y_{i,a}] \cdot \lambda_{\mathcal{O}\mathcal{A}} R^{\mathcal{O}\mathcal{A}}(o, a)$$

Attribute/attribute relationship term:

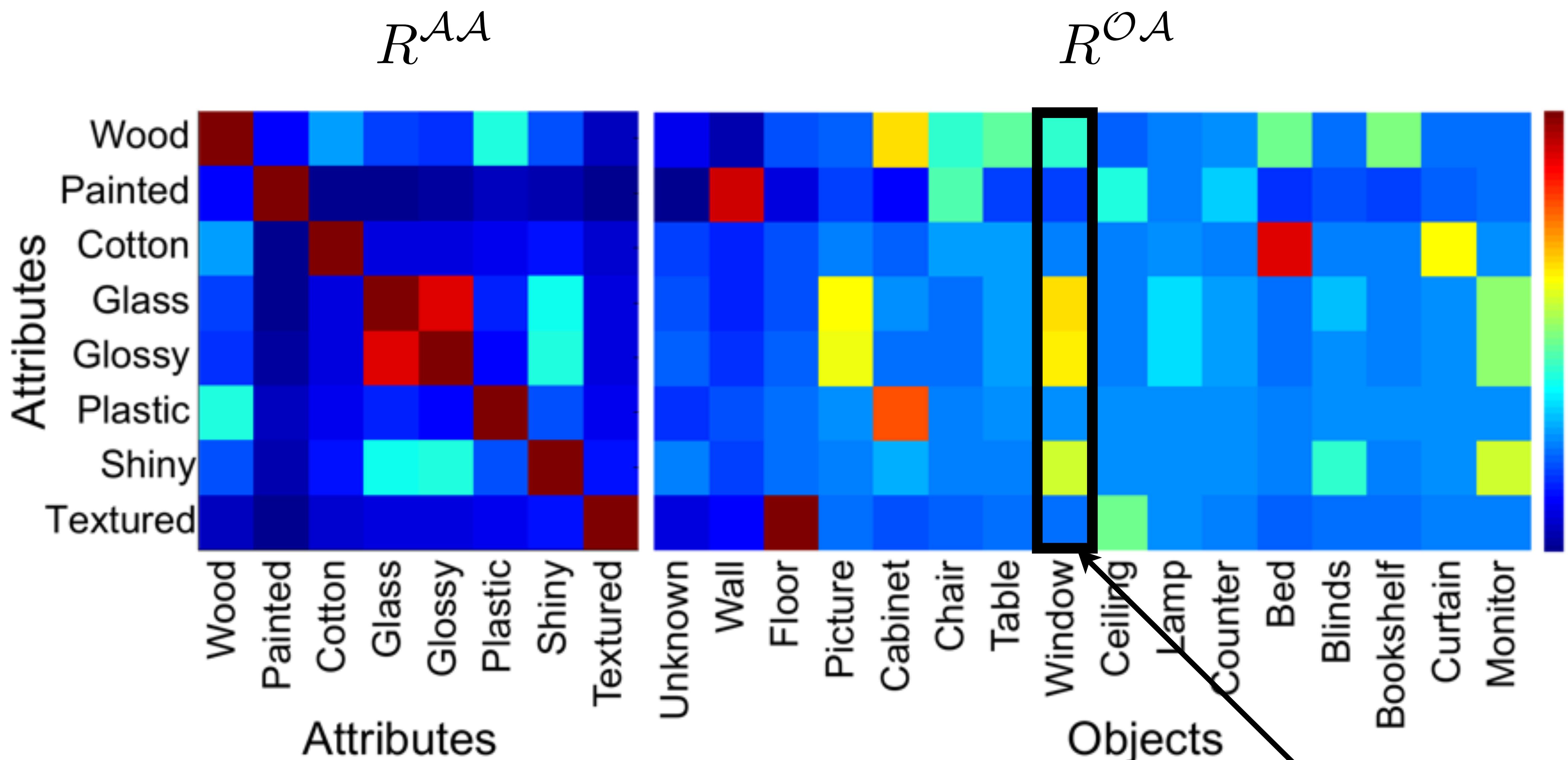
$$\psi_{i,a,a'}^{\mathcal{A}}(y_{i,a}, y_{i,a'}) = \mathbf{1}[y_{i,a} \neq y_{i,a'}] \cdot \lambda_{\mathcal{A}} R^{\mathcal{A}\mathcal{A}}(a, a')$$

* TextonBoost, Shotton et al., 2009

Relationships (co-occurrences)

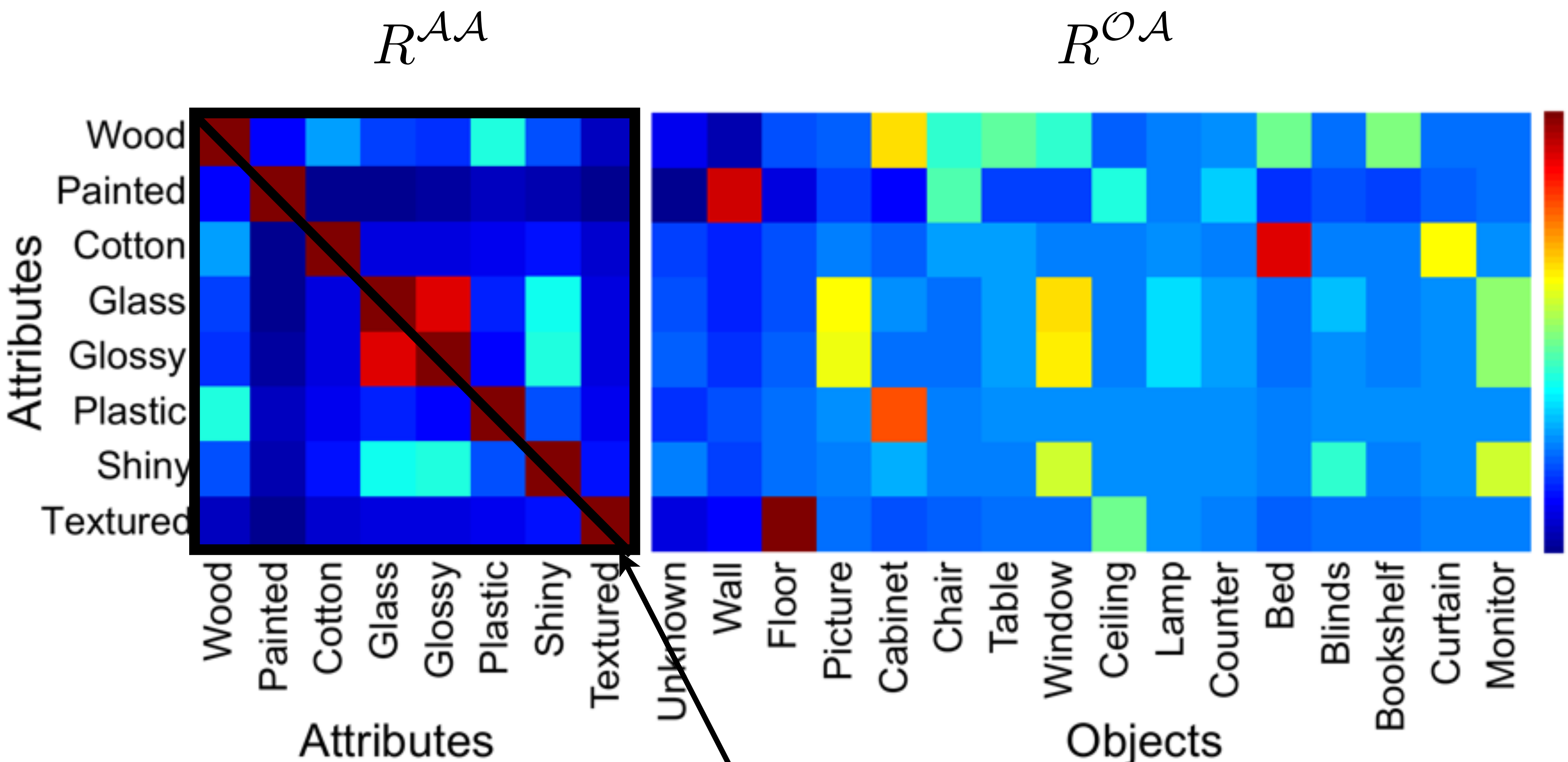


Relationships (co-occurrences)



$$\psi_{i,o,a}^{\mathcal{O},\mathcal{A}}(x_i, y_{i,a}) = \mathbf{1}[\mathbf{1}[x_i = o] \neq y_{i,a}] \cdot \lambda_{\mathcal{OA}} R^{\mathcal{OA}}(o, a)$$

Relationships (co-occurrences)



$$\psi^{\mathcal{A}}_{i,a,a'}(y_{i,a}, y_{i,a'}) = \mathbf{1}[y_{i,a} \neq y_{i,a'}] \cdot \lambda_{\mathcal{A}} R^{\mathcal{A}}(a, a')$$

Object/Attribute Model

Fully-connected CRF decomposition:

$$E(\mathbf{z}) = \sum_i \psi(z_i) + \boxed{\sum_{i<j} \psi(z_i, z_j)}$$

Unary terms: enforce object/attribute assignments

Pairwise terms: enforce consistent labelings between nearby pixels

Edge Term

$$\psi_{i,j}(z_i, z_j) = \psi^{\mathcal{O}}_{i,j}(x_i, x_j) + \sum_a \psi^{\mathcal{A}}_i(y_{i,a}, y_{j,a})$$

Neighboring object agreement term:

$$\psi^{\mathcal{O}}_{i,j}(x_i, x_j) = \mathbf{1}[x_i \neq x_j] \cdot g(i,j)$$

Neighboring attribute agreement term:

$$\psi^{\mathcal{A}}_i(y_{i,a}, y_{j,a}) = \mathbf{1}[y_{i,a} \neq y_{j,a}] \cdot g(i,j)$$

Gaussian similarity function: *

$$g(i,j) = \underbrace{w_1 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\mu^2} - \frac{|I_i - I_j|^2}{2\theta_\nu^2}\right)}_{\text{appearance kernel}} + \underbrace{w_2 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right)}_{\text{smoothness kernel}}$$

*Krahenbuhl and Koltun, 2011

Efficient joint inference

Minimize *E*($\mathbf{z}$) with mean field approximation of:

$$P \propto \exp(-E(\mathbf{z}))$$

using Q_i where:

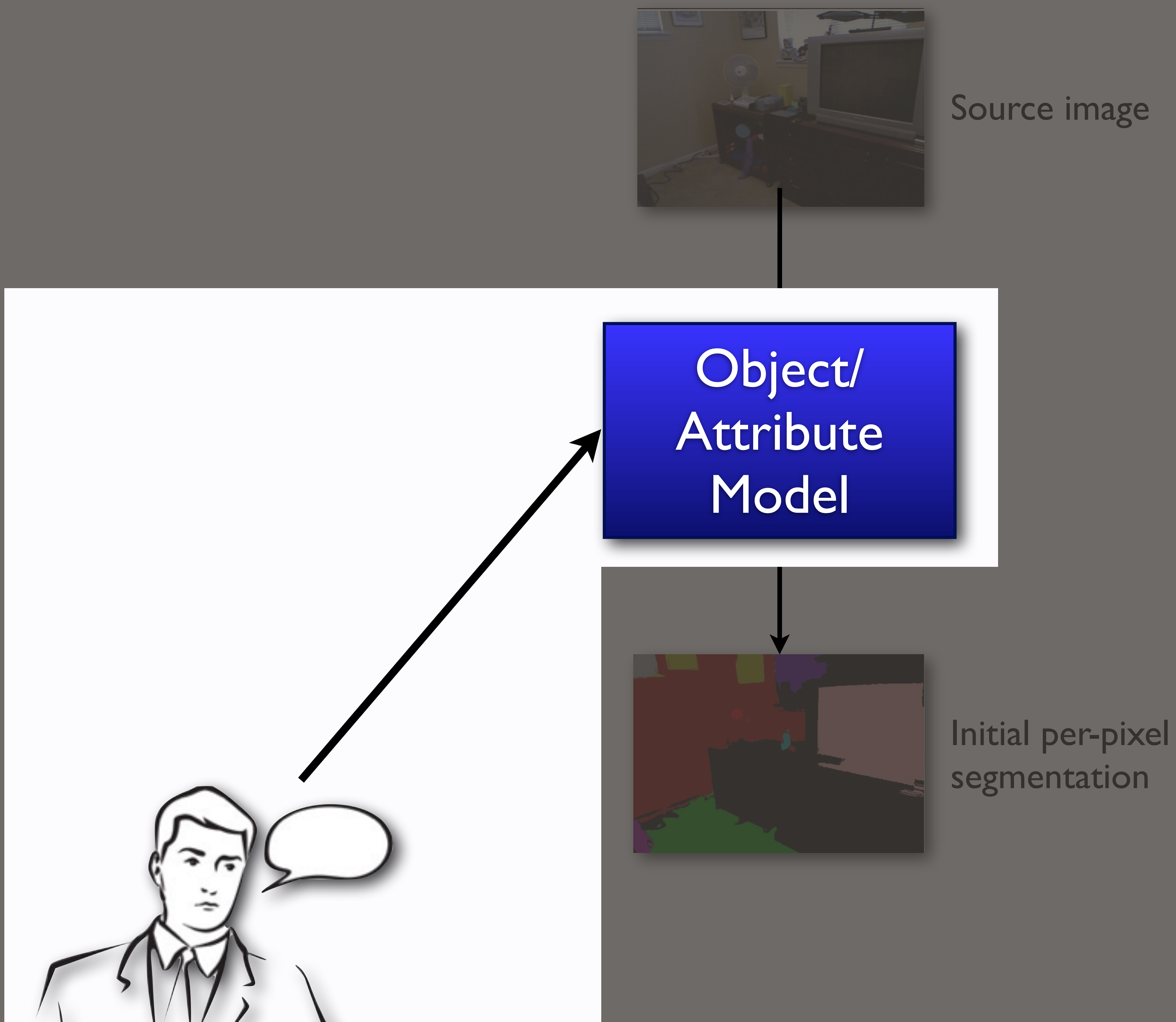
$$Q_i(z_i) = Q_i^{\mathcal{O}}(x_i) \prod_a Q_{i,a}^{\mathcal{A}}(y_i, a)$$

Given Gaussian pairwise costs, by using efficient filtering techniques,*
computing each Q_i is $O(n)$ instead of $O(n^2)$!

(Where n=#pixels)

*Krahenbuhl and Koltun, 2011

System Overview



Verbal attributes interaction

"Refine the white textured cotton bed in center-middle."

Verbal attributes interaction

"Refine the white textured cotton bed in center-middle."

(1) Update relationship matrices:

$$\tilde{R}^{\mathcal{O}\mathcal{A}}_{de} = \lambda_1 + \lambda_2 R^{\mathcal{O}\mathcal{A}}_{de}$$

$$\tilde{R}^{\mathcal{A}\mathcal{A}}_{ef} = \lambda_3 + \lambda_4 R^{\mathcal{A}\mathcal{A}}_{ef}$$

Verbal attributes interaction

"Refine the white textured $\boxed{\textit{cotton bed}}$ *in center-middle."*

(1) Update relationship matrices:

$$\tilde{R}_{de}^{\mathcal{O}\mathcal{A}} = \lambda_1 + \lambda_2 R_{de}^{\mathcal{O}\mathcal{A}}$$

$$\tilde{R}_{ef}^{\mathcal{A}\mathcal{A}} = \lambda_3 + \lambda_4 R_{ef}^{\mathcal{A}\mathcal{A}}$$

Verbal attributes interaction

"Refine the white $\boxed{\textit{textured cotton}}$ *bed in center-middle."*

(1) Update relationship matrices:

$$\tilde{R}_{de}^{\mathcal{O}\mathcal{A}} = \lambda_1 + \lambda_2 R_{de}^{\mathcal{O}\mathcal{A}}$$

$$\tilde{R}_{ef}^{\mathcal{A}\mathcal{A}} = \lambda_3 + \lambda_4 R_{ef}^{\mathcal{A}\mathcal{A}}$$

Verbal attributes interaction

"Refine the white textured cotton bed in center-middle."

(1) Update relationship matrices:

$$\tilde{R}^{\mathcal{O}\mathcal{A}}_{de} = \lambda_1 + \lambda_2 R^{\mathcal{O}\mathcal{A}}_{de}$$

$$\tilde{R}^{\mathcal{A}\mathcal{A}}_{ef} = \lambda_3 + \lambda_4 R^{\mathcal{A}\mathcal{A}}_{ef}$$

(2) Update unary potentials using response map $\mathbf{R}$

$$\tilde{\psi_i}^{\mathcal{O}}(x_i) = \psi_i^{\mathcal{O}}(x_i) - \frac{\lambda_5}{R(i)}, \text{ if } x_i = d$$

(a) source image (b) white $\mathbf{R}_c$ (c) center-middle $\mathbf{R}_s$

Verbal attributes interaction

"Refine the white textured cotton bed in center-middle."

(1) Update relationship matrices:

$$\tilde{R}^{\mathcal{OA}}_{de} = \lambda_1 + \lambda_2 R^{\mathcal{OA}}_{de}$$

$$\tilde{R}^{\mathcal{AA}}_{ef} = \lambda_3 + \lambda_4 R^{\mathcal{AA}}_{ef}$$

(2) Update unary potentials using response map **R**

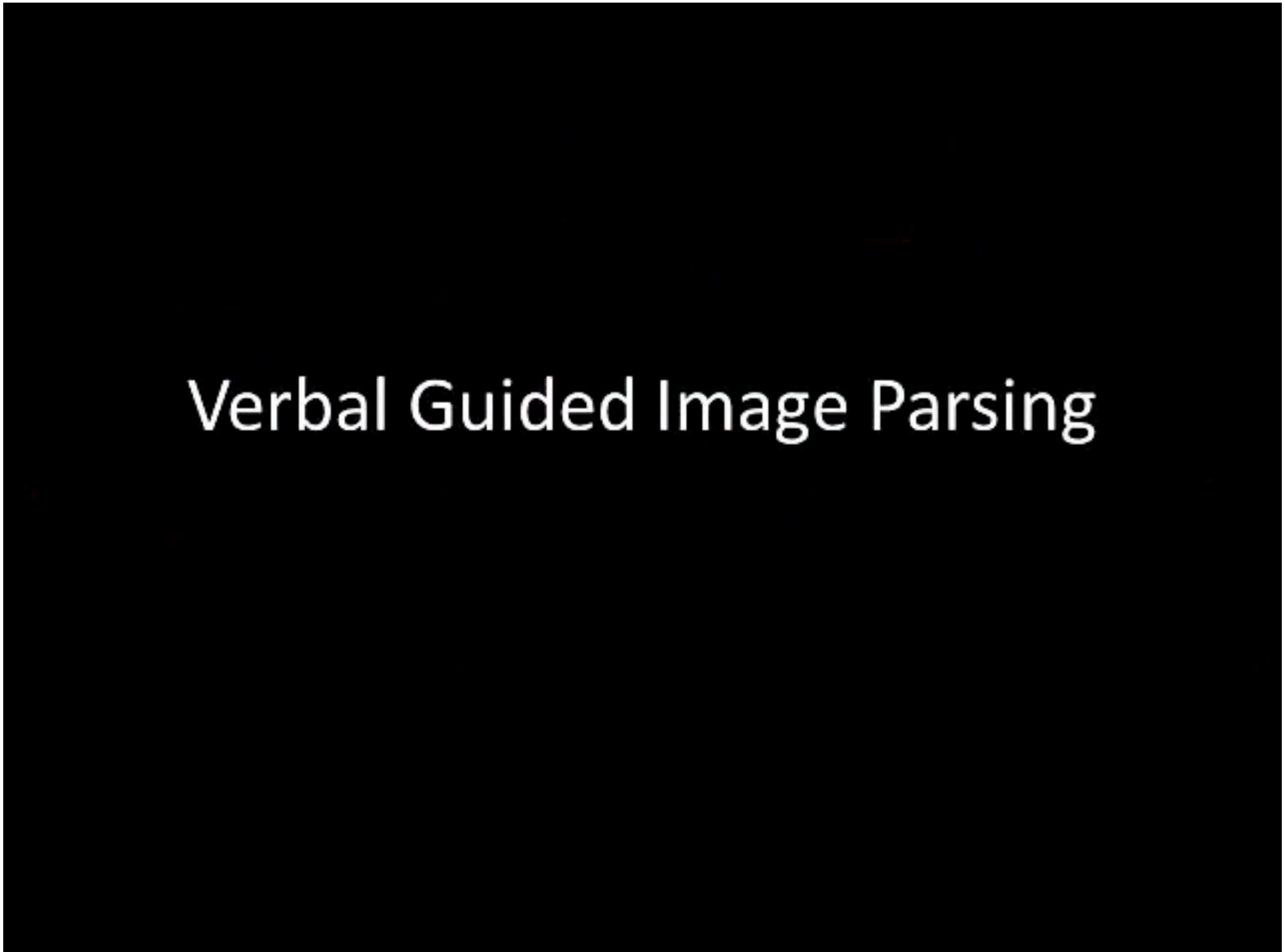
$$\tilde{\psi}_i^{\mathcal{O}}(x_i) = \psi_i^{\mathcal{O}}(x_i) - \frac{\lambda_5}{R(i)}, \text{ if } x_i = d$$

(a) source image (b) white $\mathbf{R}_c$ (c) center-middle $\mathbf{R}_s$

(3) Re-run inference to obtain refined segmentation

Implementation examples

Implementation examples



Results

Table II. Quantitative results on aNYU dataset.

Methods	H-CRF	DenseCRF	Our-auto	Our-inter
Label accuracy	51.0%	50.7%	56.9%	- -
Inference time	13.2s	0.13s	0.54s	0.21s
Has attributes	NO	NO	YES	YES

Table III. Evaluation for verbal guided segmentation.

Methods	DenseCRF	Our-auto	Our-inter
Label accuracy	52.1%	56.2%	80.6%

Results

Table IV. Comparison of different interaction modality.

Interaction modality	Verbal	Mouse	Verbal + Mouse
Average interaction time (s)	6.9	28.1	11.7
Average label accuracy (%)	80.1	98.1	98.3
Average user preference (%)	12.5	17.5	70.0

*Population consists of graduate CS students

Questions / Discussion