The Revolution will not be supervised!

Alexei Efros

UC Berkeley

Object Understanding for Interaction Workshop
Santiago, Chile (Dec 17, 2015)

Neruda, The Book of Questions

Tell me, is the rose naked
or is that her only dress?

Why do trees conceal
the splendor of their roots?

Who hears the regrets
of the thieving automobile?

Is there anything in the world sadder
than a train standing in the rain?

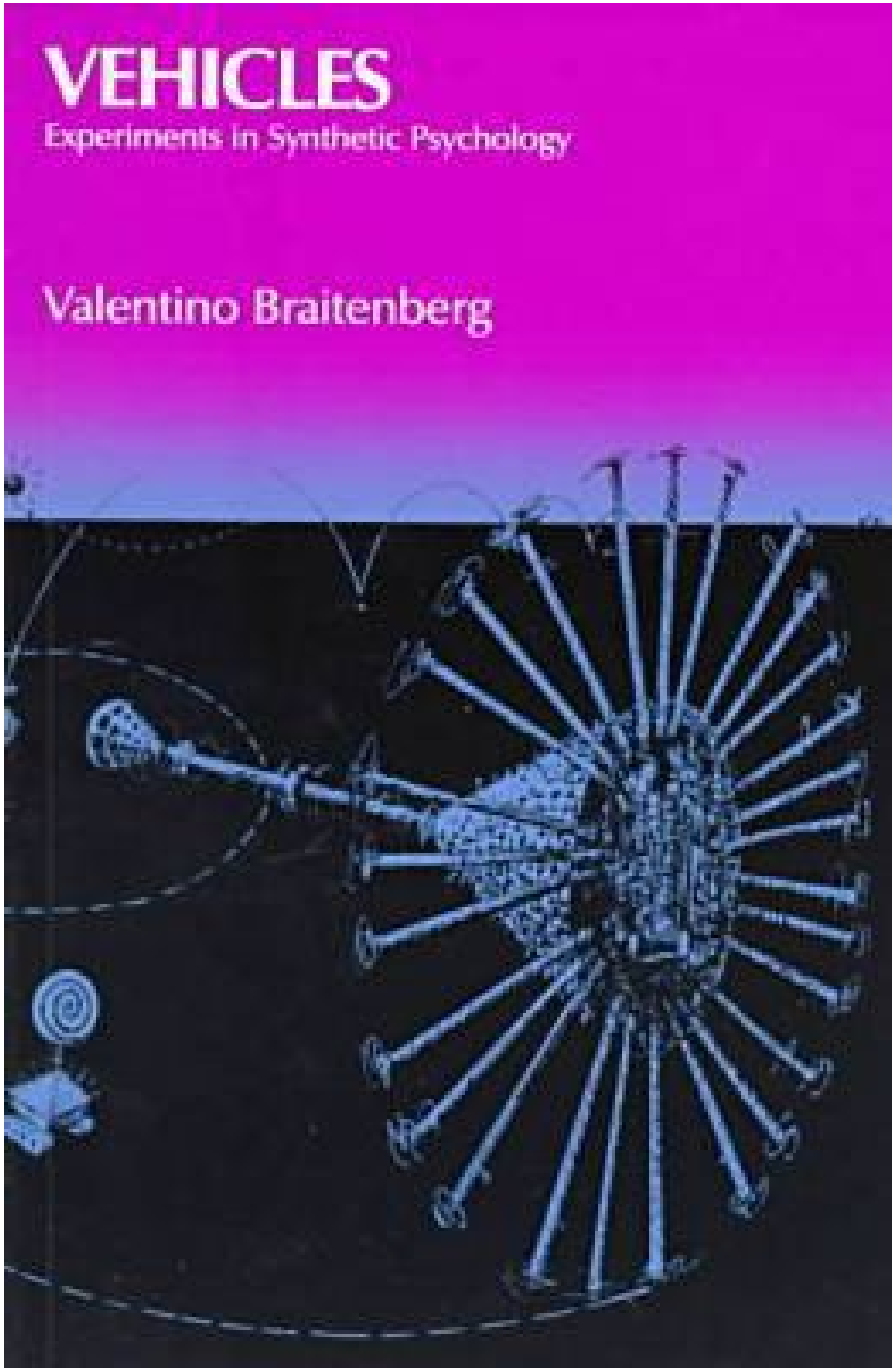
Why do we have vision?

- “To see what is where by looking”
  - Aristotle, Marr, etc
- .
- .
- .
- .
- “To make babies who make babies, etc”
  - Darwin, Dawkins, etc.

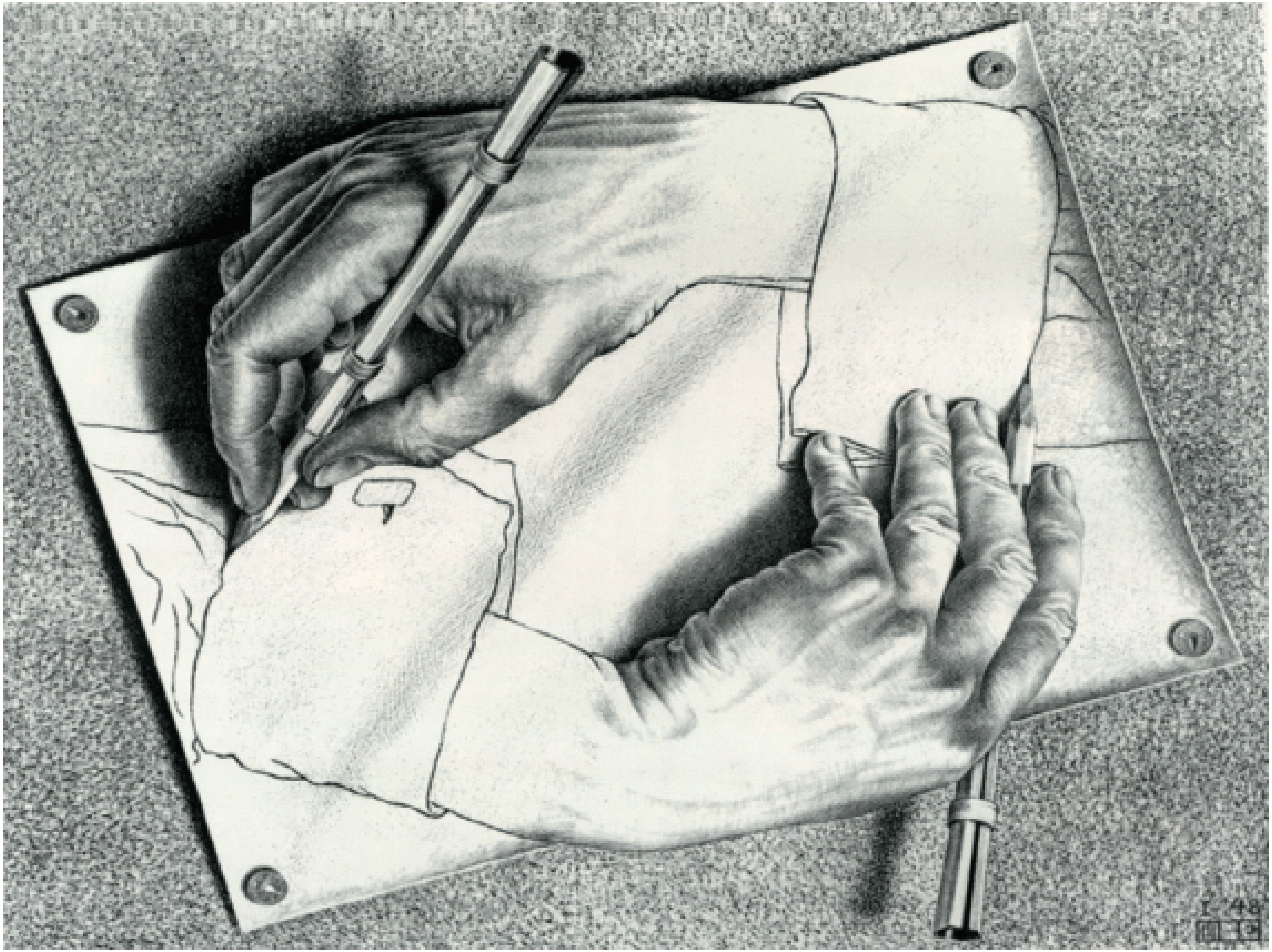
A more immediate purpose

- We want to focus on what's important **to us**
  - Learn a distance that will only care about our long-term goal (e.g. reproduction)
- But we don't want to wait and see if our grand-kids get into Harvard
- Therefore: supervision!
  - But how did it boot-strap? Who was the first Teacher / ImageNet Labeller?

The organism vs. the environment



Visual data as its own supervision



Drawing Hands, M.C. Escher, 1948

The world as supervision

1. Try to predict some aspect of the world that we interact with / have effect on:
   - What’s gonna happen next?
   - What’s to my left?
   - What can I touch?
   - What will make a sound?
   - Etc.
2. Try it out and see if you are right
3. Learn!

Big Questions

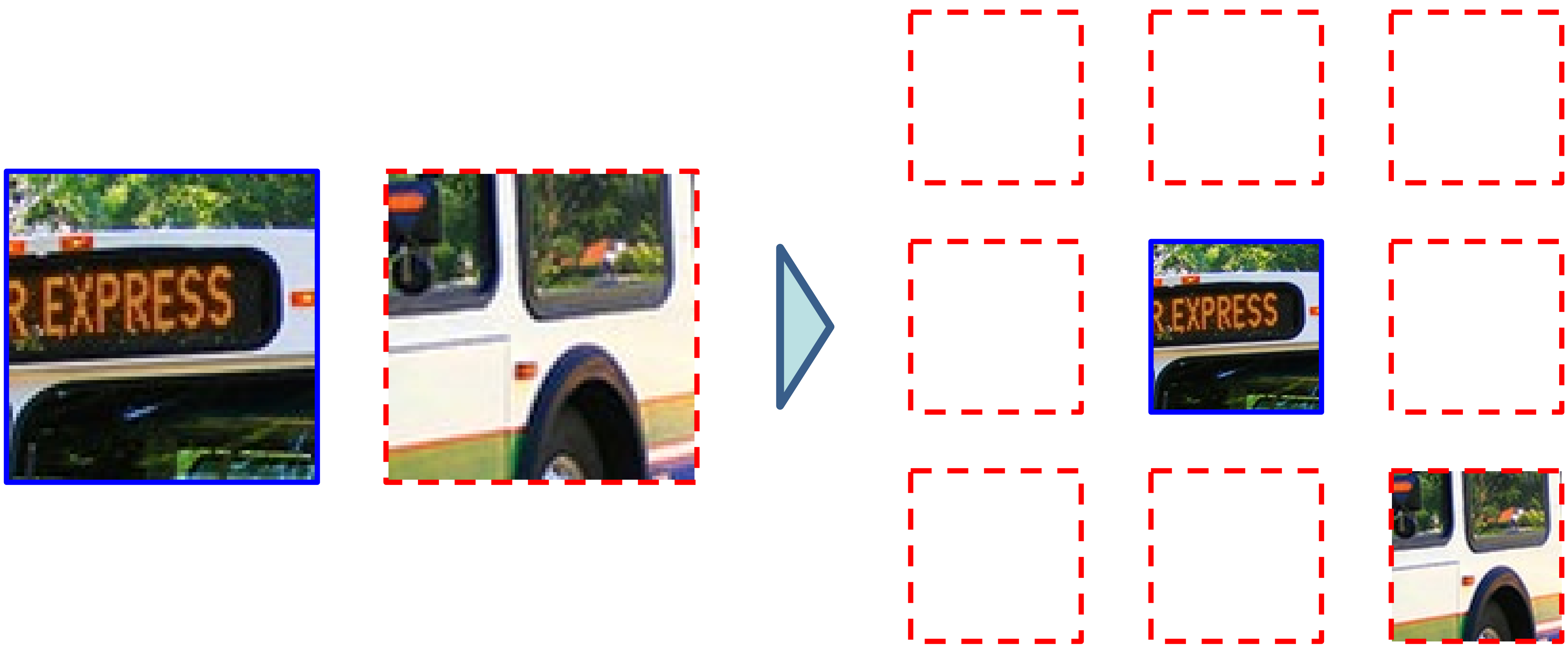
- What do we predict?
  - Generative: pixels, features
  - Discriminative: easily-testable predicates
  - Hybrids:
    - Hypothesis testing (e.g. Doerch, ECCV 2014)
    - Adversarial training

- When do we give up?

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Abhinav Gupta, Alexei Efros

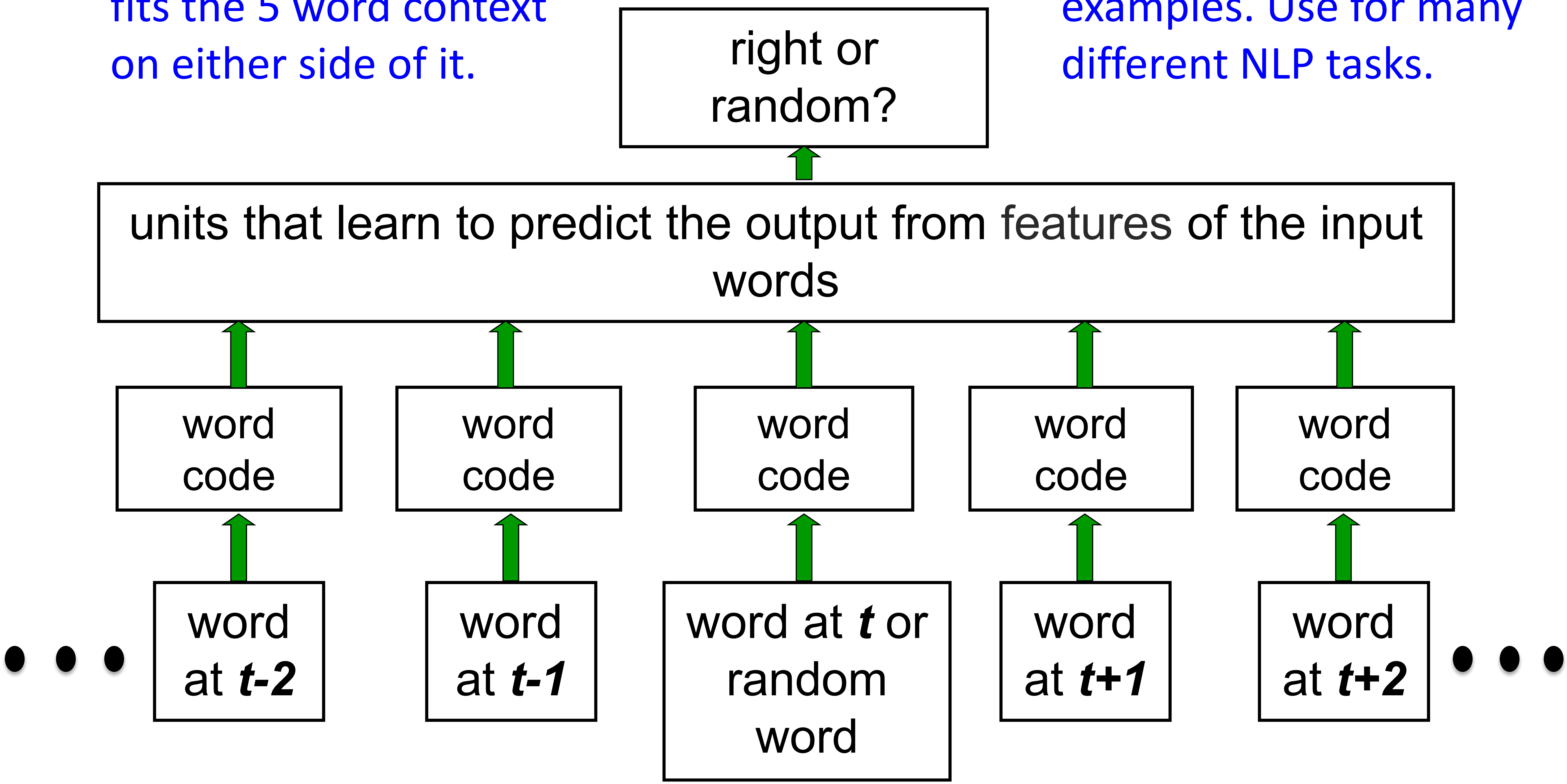
ICCV 2015


R EXPRESS
R EXPRESS

A simple way to learn feature vectors for words
(Collobert and Weston, 2008)

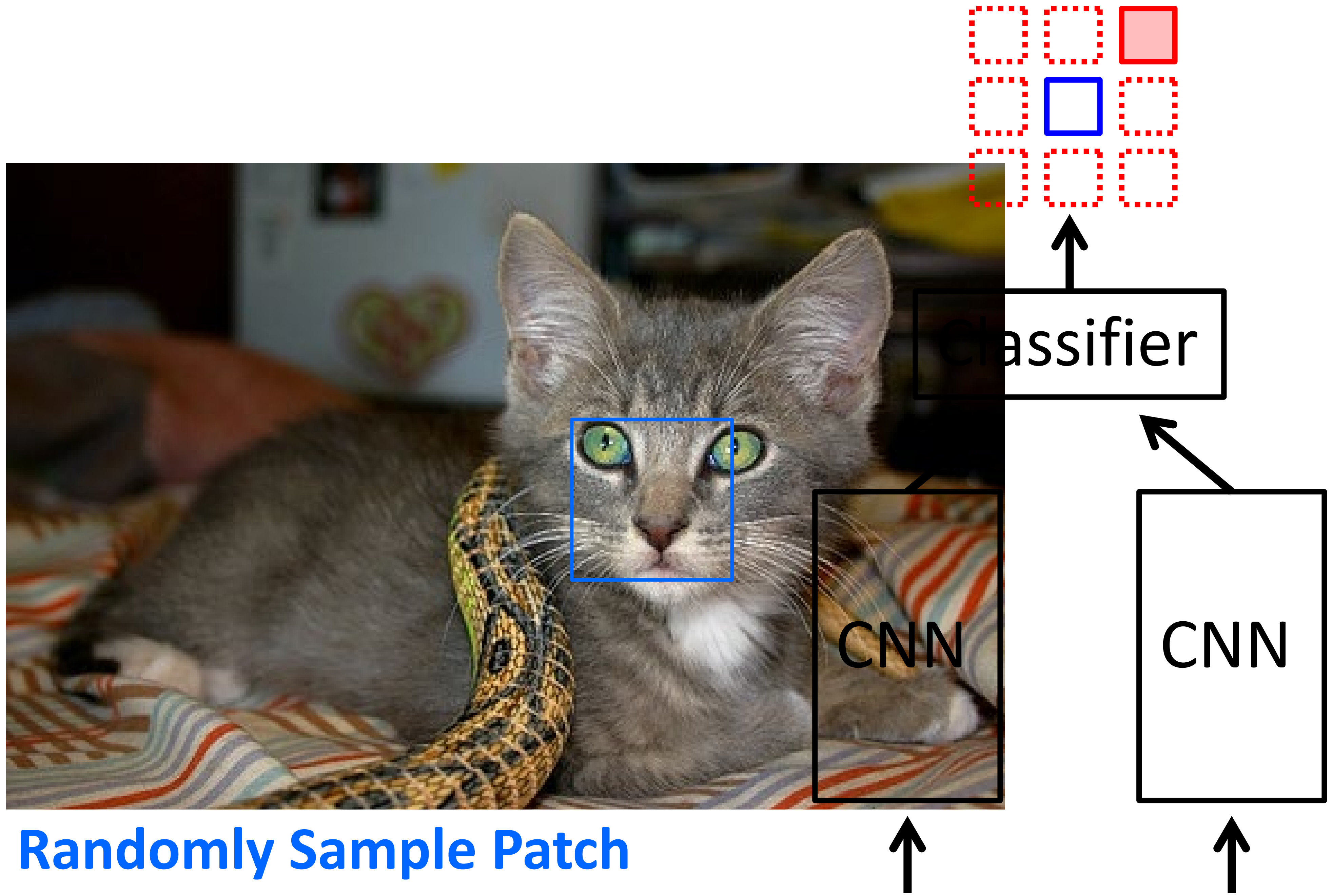
Learn to judge if a word fits the 5 word context on either side of it.

Train on ~600 million examples. Use for many different NLP tasks.



Slide from Geoff Hinton

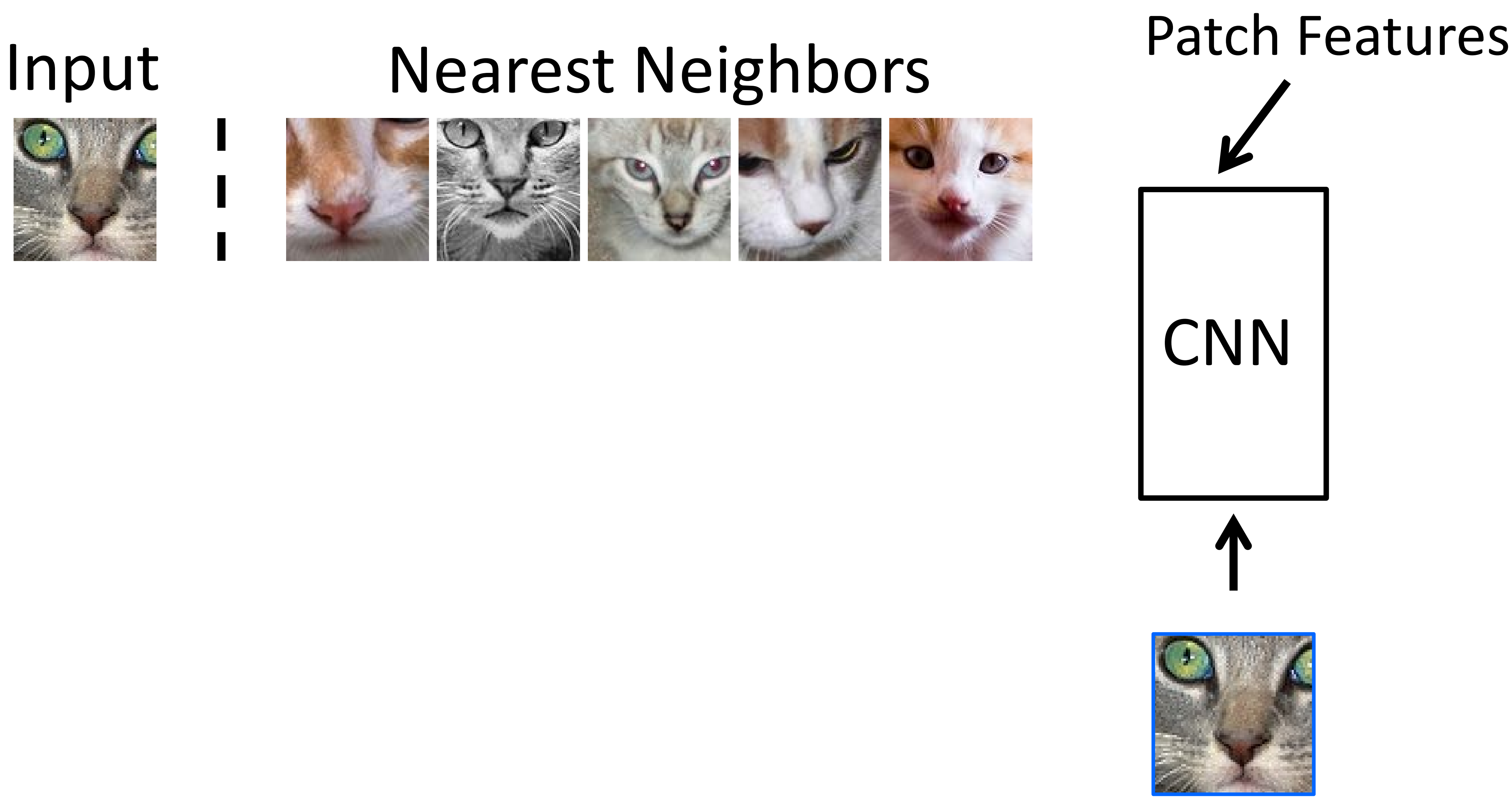
Unlabeled training image



Randomly Sample Patch

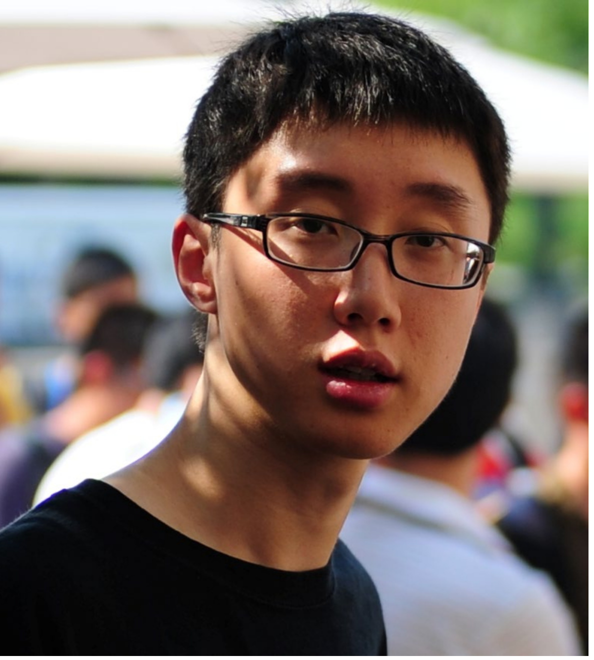
Sample Second Patch

Train Deep Net to recover relative position


Input
Nearest Neighbors
Patch Features
CNN

The Gelato Bet

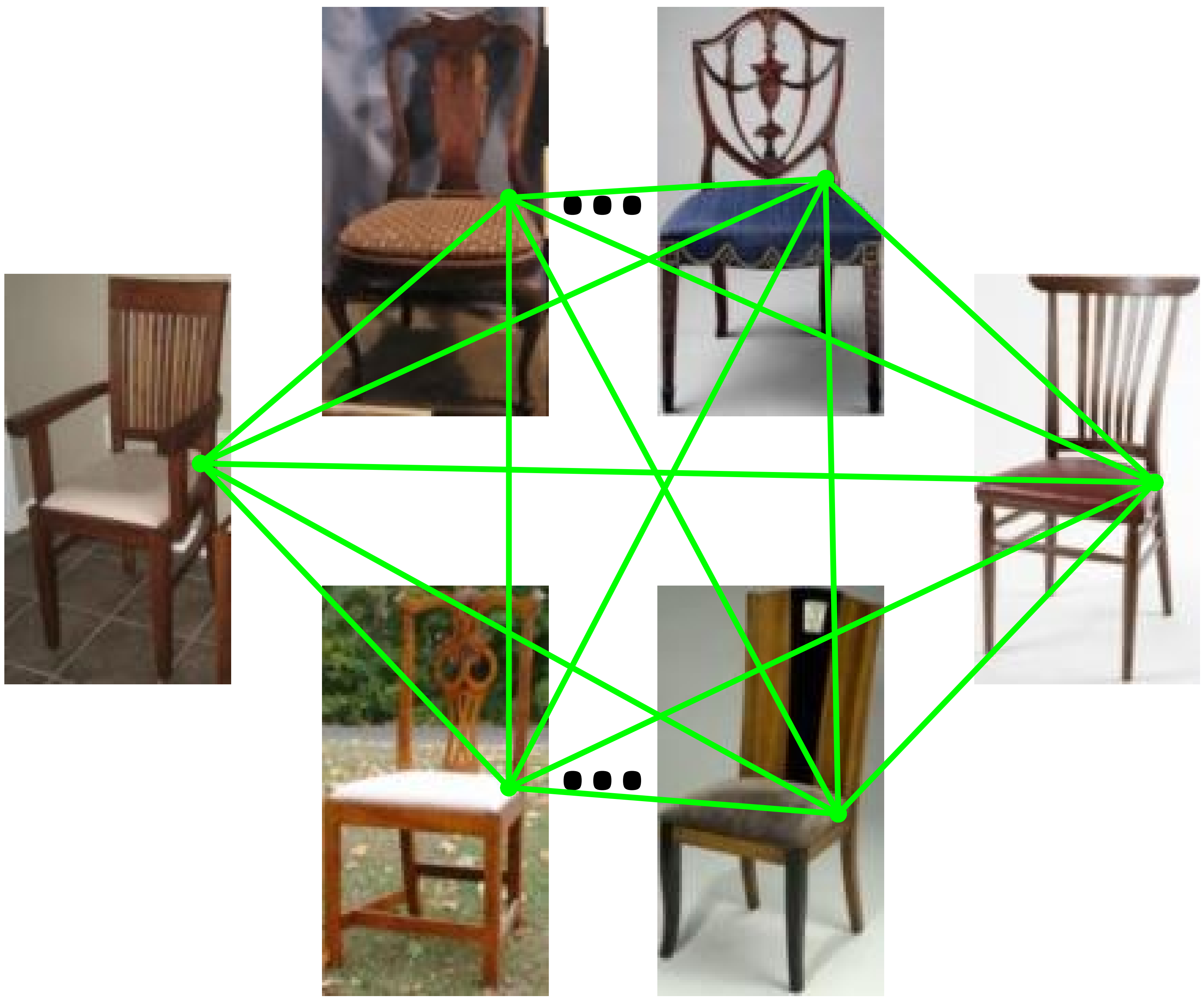
"If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, Mr. Malik promises to buy Mr. Efros one (1) gelato (two scoops: one chocolate, one vanilla)."



Learning Dense Correspondences via 3D-guided Cycle Consistency

Tinghui Zhou, Qixing Huang (TTIC), Mathieu Aubry, Philipp Krähenbühl, Alyosha Efros

In Submission

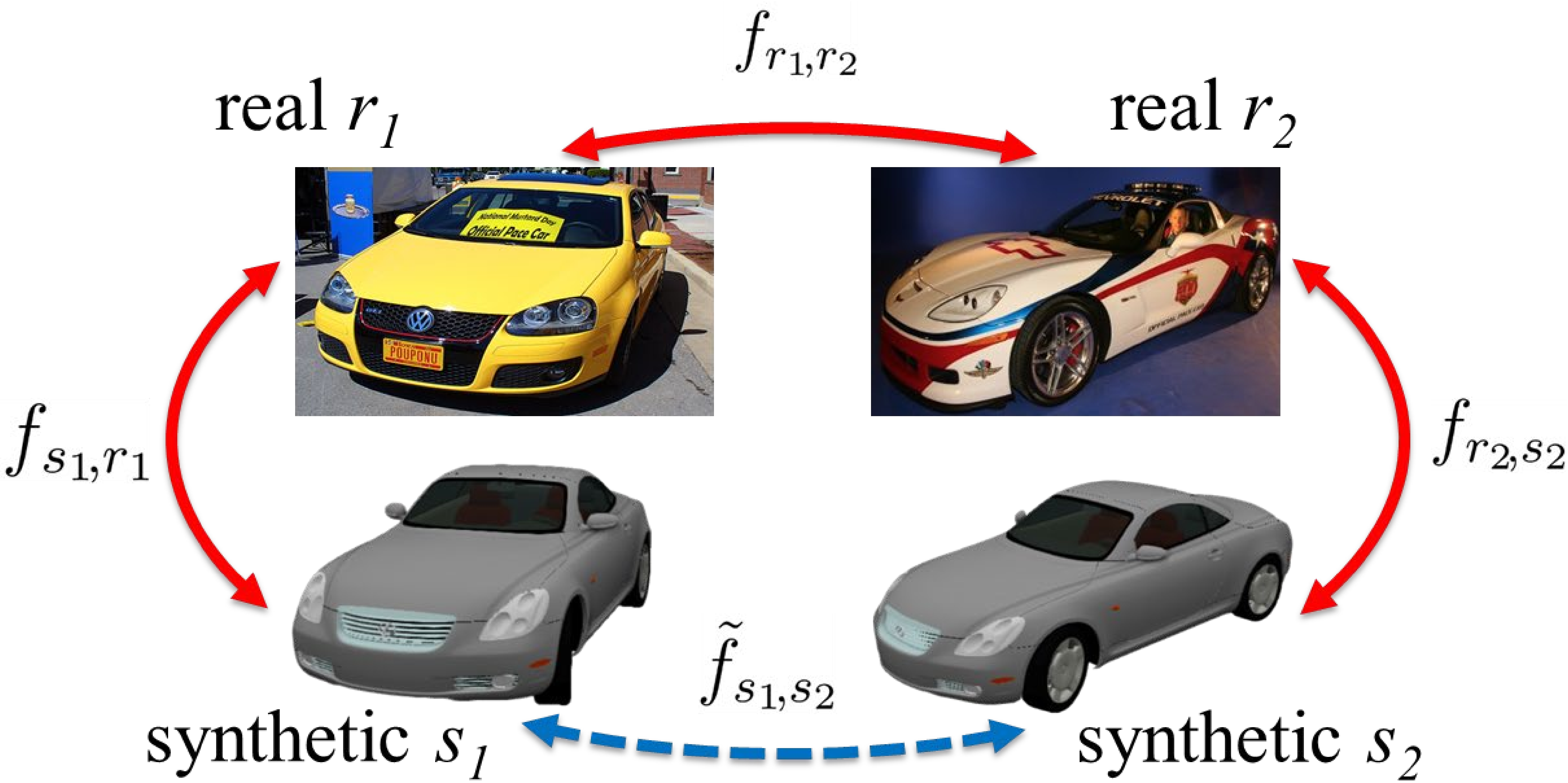


CNN for Dense Correspondence

- CNN works great when direct supervision is available (classification, detection, etc.)
- No direct supervision is available for dense, cross-instance matching
- **Key idea**: consistency as supervisory signal

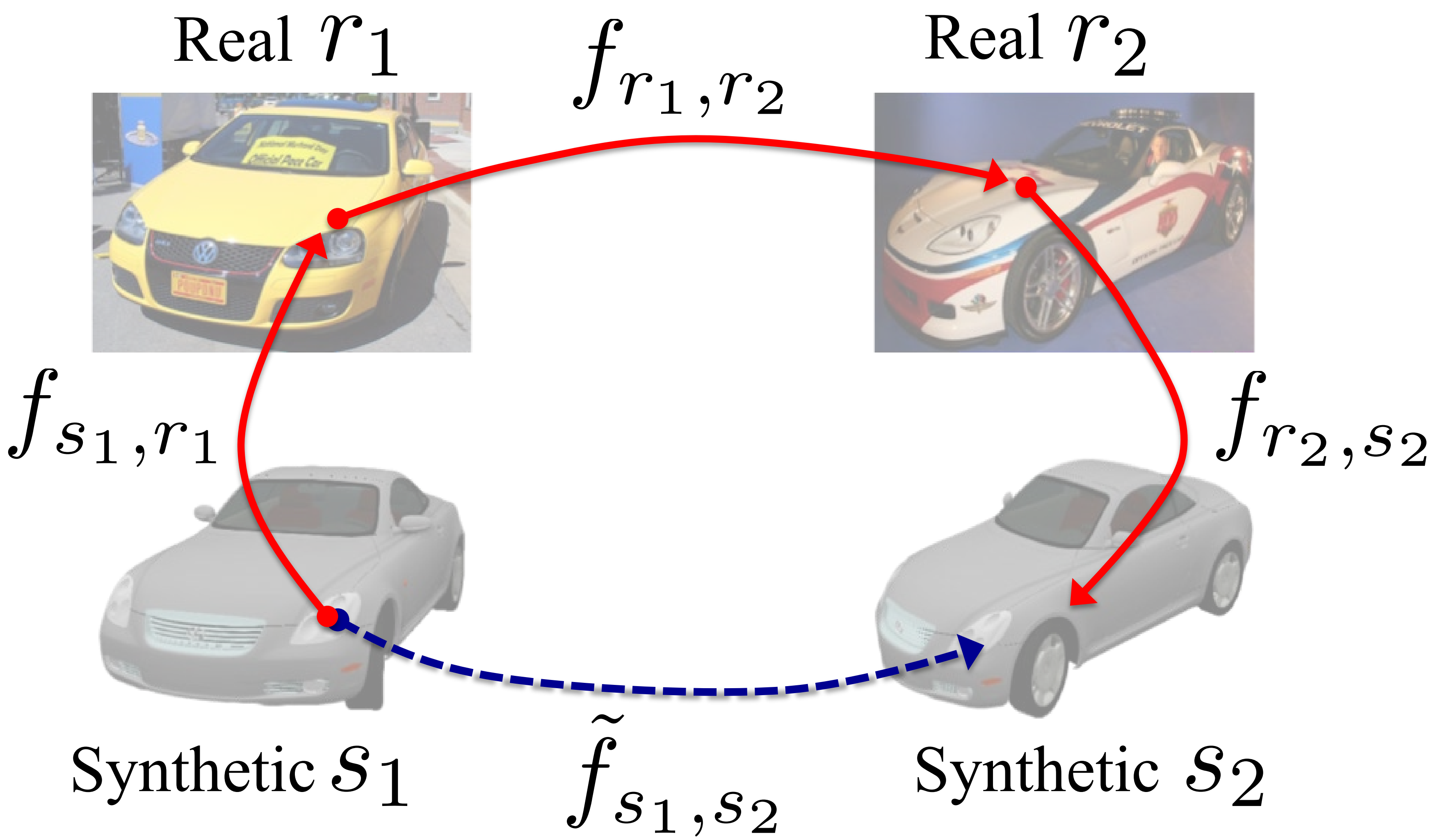
- But how to prevent trivial / diverging solutions?

3D-guided Cycle Consistency Training



$$\tilde{f}_{s_1,s_2} = f_{s_1,r_1} \circ f_{r_1,r_2} \circ f_{r_2,s_2}$$

Minimize Inconsistency

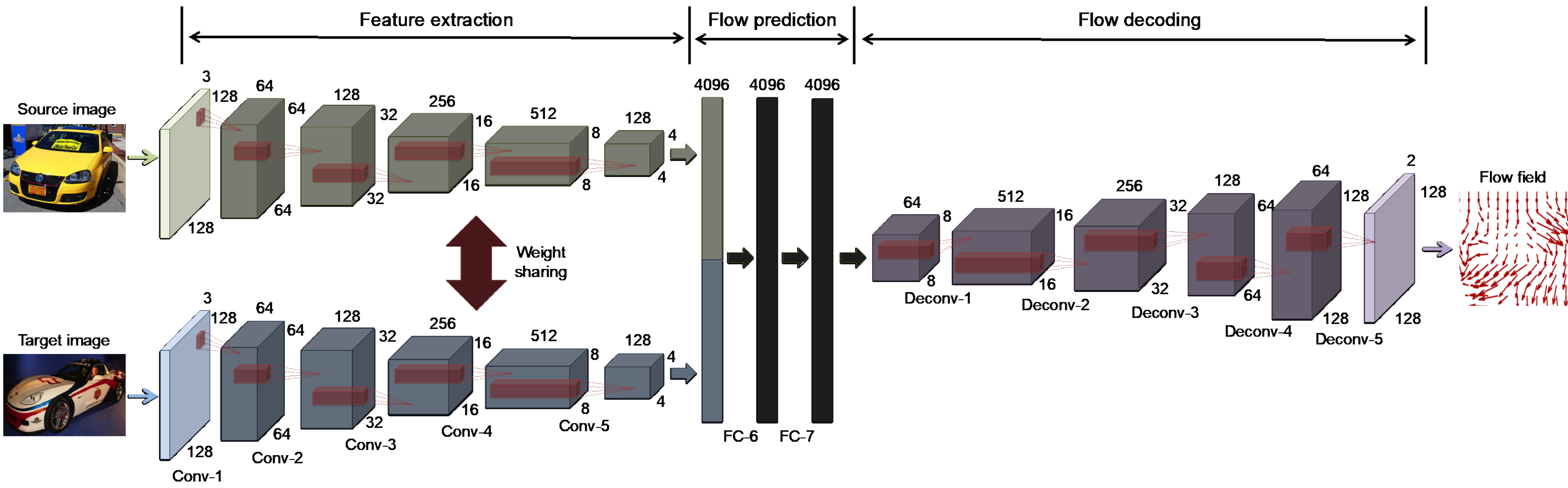


$$\underset{\theta}{\text{minimize}} \sum_{<s_1,s_2,r_1,r_2>} \rho\left(\tilde{f}_{s_1,s_2} - f_{s_1,r_1} \circ f_{r_1,r_2} \circ f_{r_2,s_2}\right)$$

A single flow prediction network f parameterized by θ

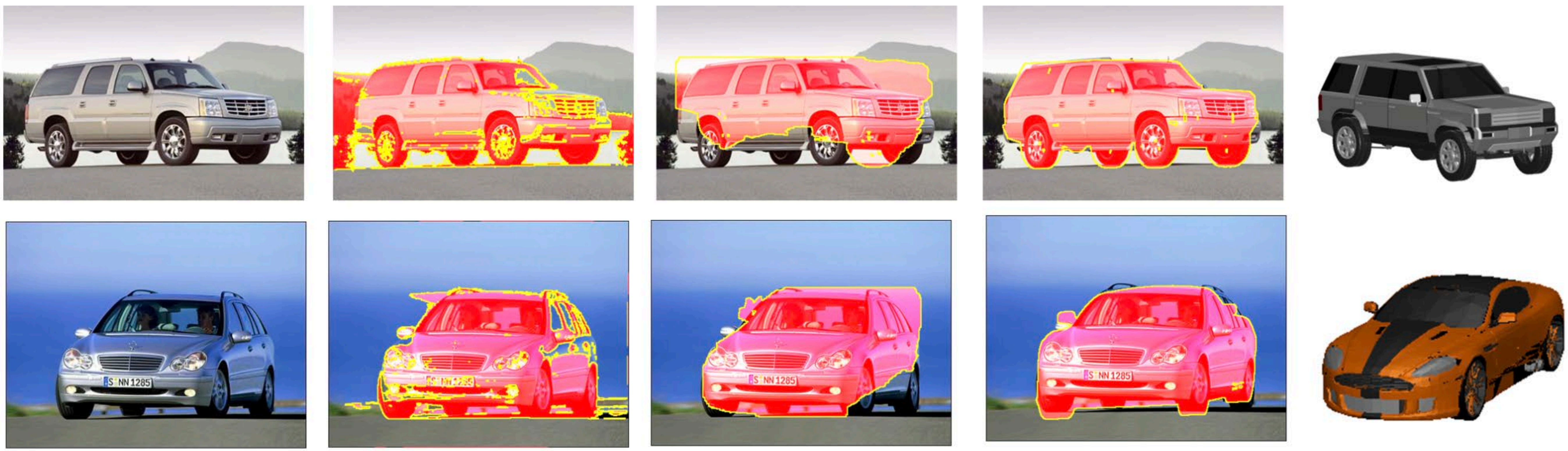
Flow Network

Shared by synthetic-to-real, real-to-real, real-to-synthetic



Application: Segmentation Transfer

Query	Dense CRF	SIFTflow Transfer	Our Flow Transfer	Retrieved Shape



Application: Segmentation Transfer

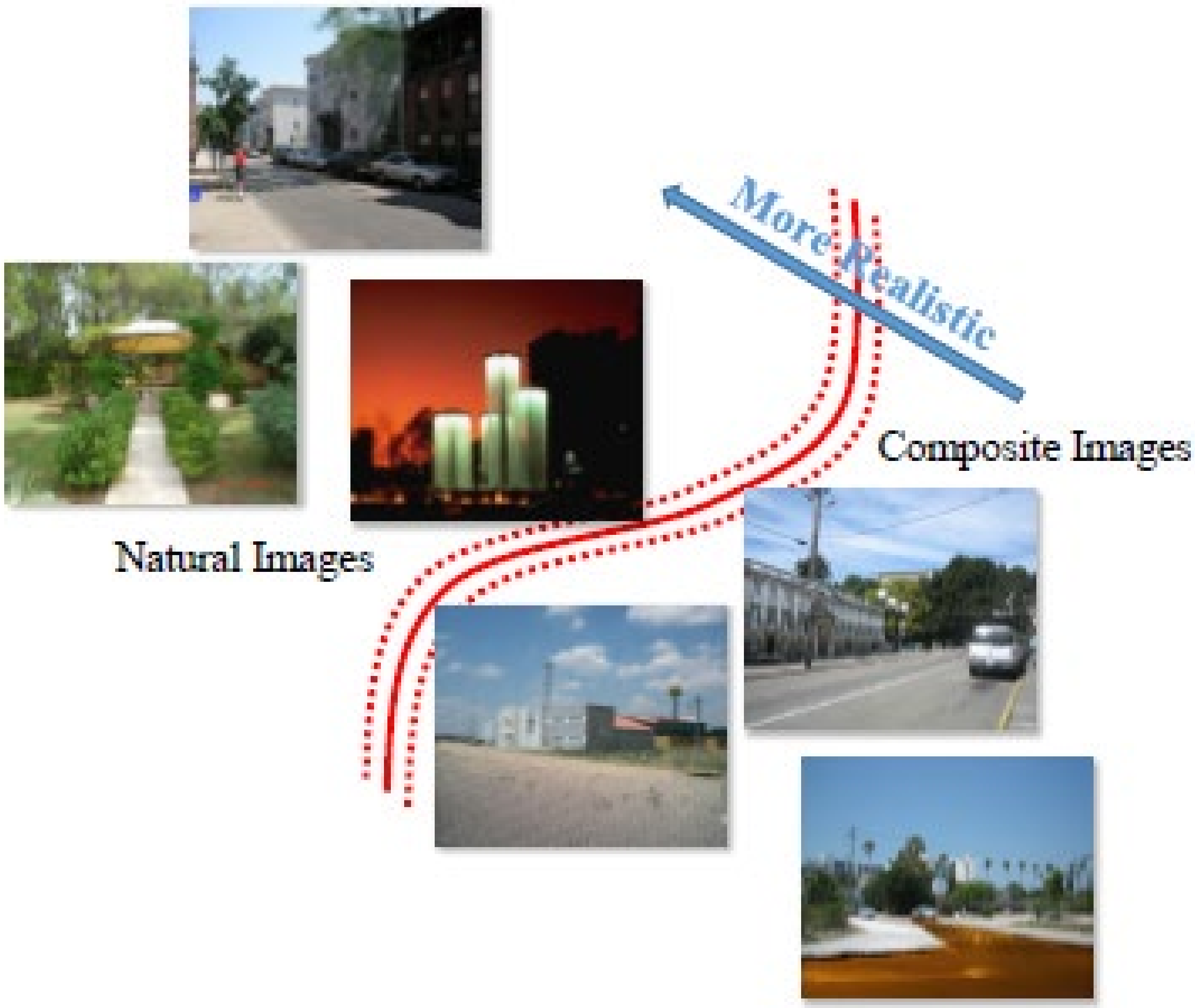
Query | Dense CRF | SIFTflow Transfer | Our Flow Transfer | Retrieved Shape

Learning Perception of Realism for Object Composition

Jun-Yan Zhu[1] *Philipp Krähenbühl*[1]
Eli Shechtman[2] *Alexei A. Efros*[1]

UC Berkeley[1] *Adobe*[2]

Supervision By Image Realism



Ranking Image Composites

Most realistic composites

Least realistic composites

Color Adjustment Results

Original

Recolored

A bit like

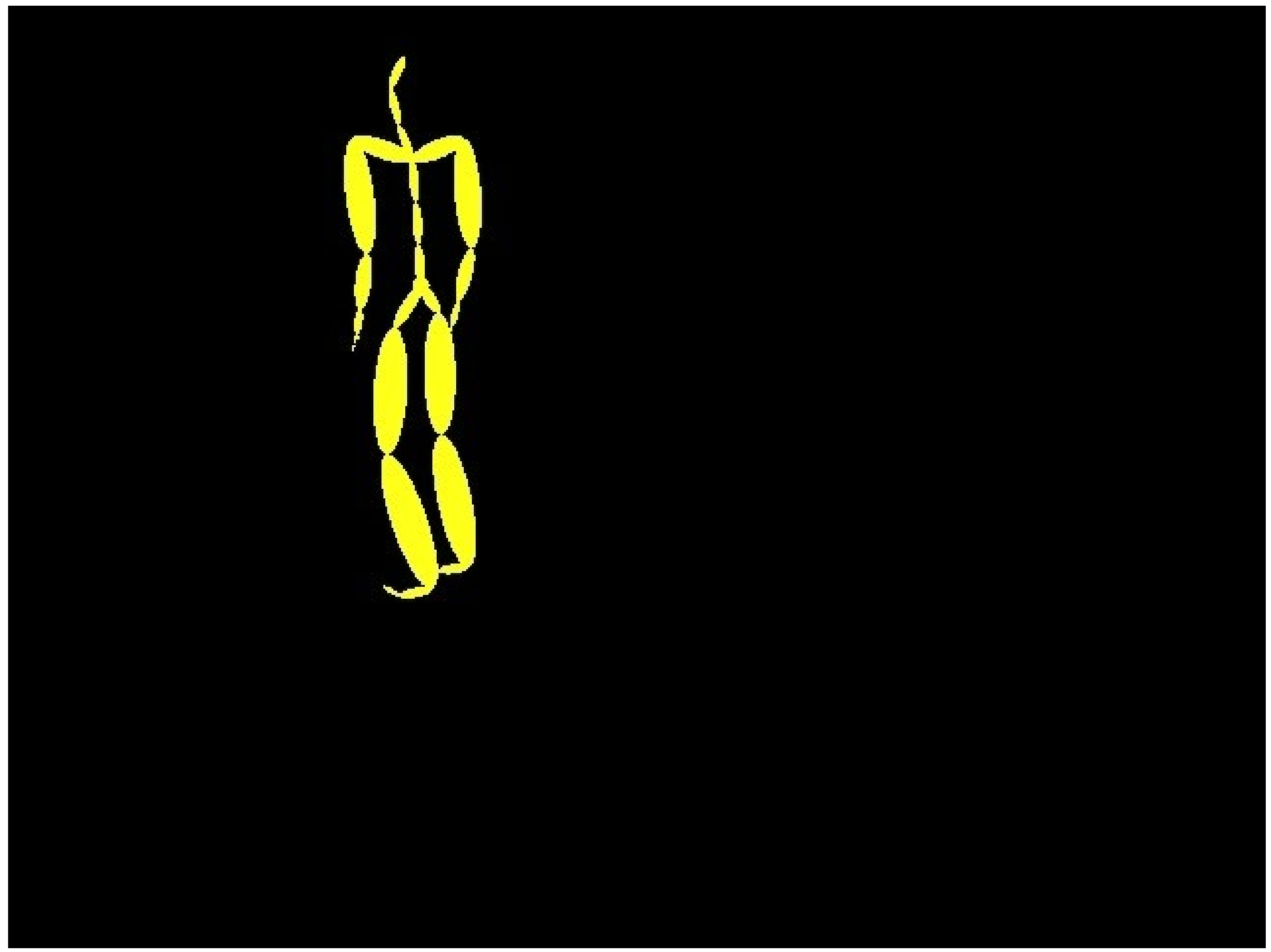
watching the paint dry…

People Watching: Human Actions as a Cue for Single-View Geometry

David Fouhey, Vincent Delaitre,
Abhinav Gupta, Alexei Efros, Ivan Laptev, Josef Sivic

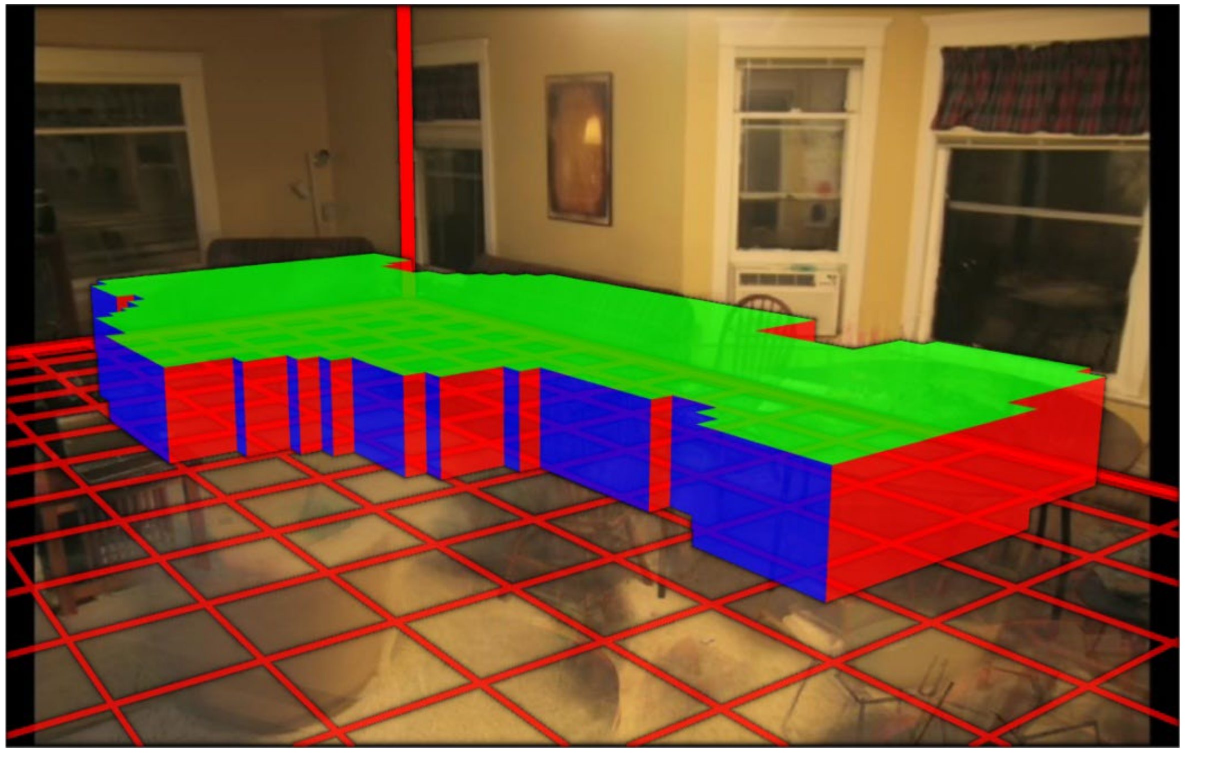
ECCV 2012

Human as Sensor



Humans as Active Sensors

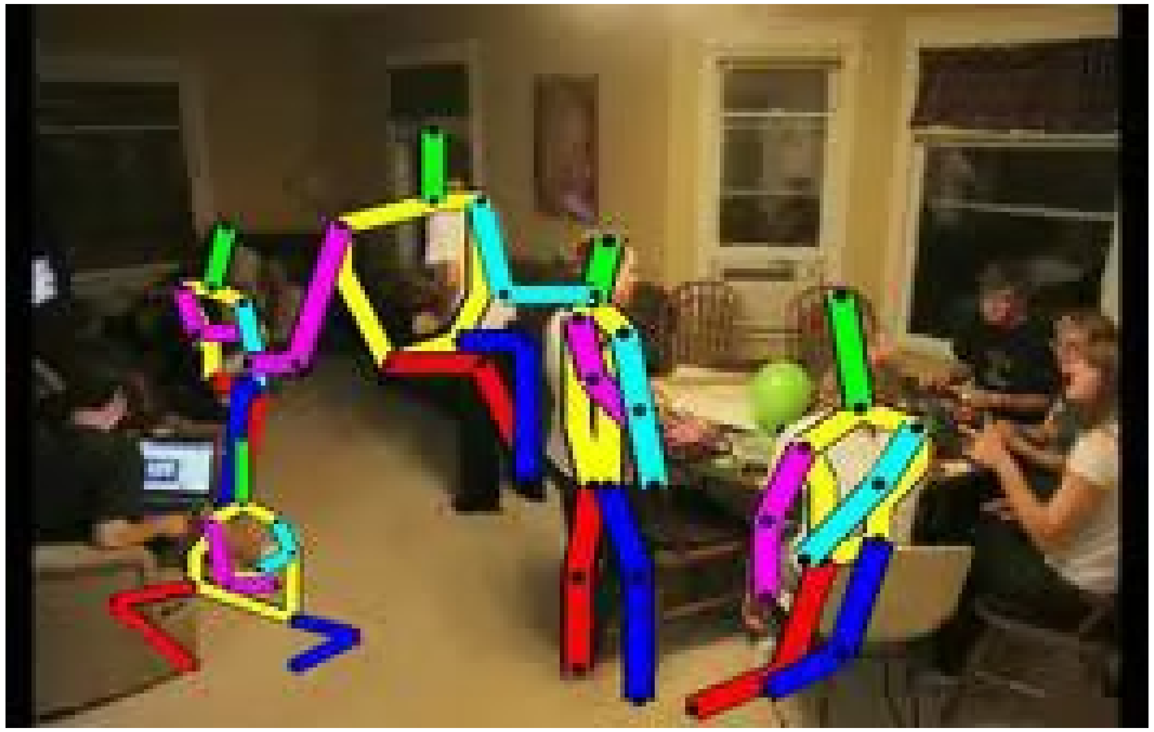
Input:
Timelapse



Output:
3D Understanding

Our Approach

Timelapse



Pose Detections

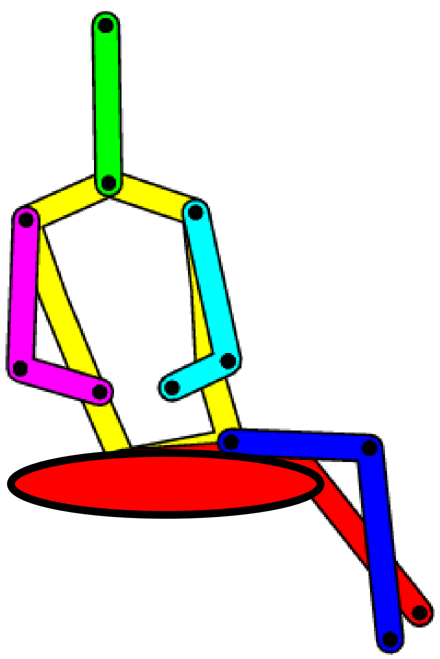
Our Approach

Timelapse

Pose Detections

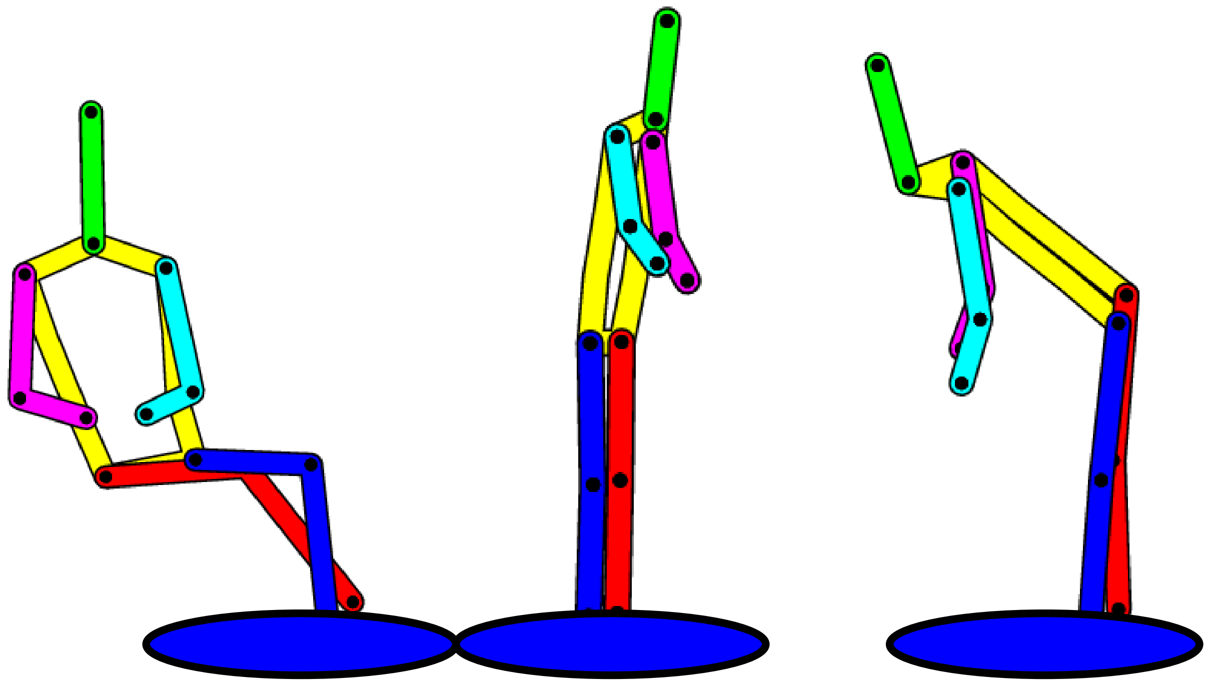
Estimate Functional Regions from Poses

From Poses to Functional Regions



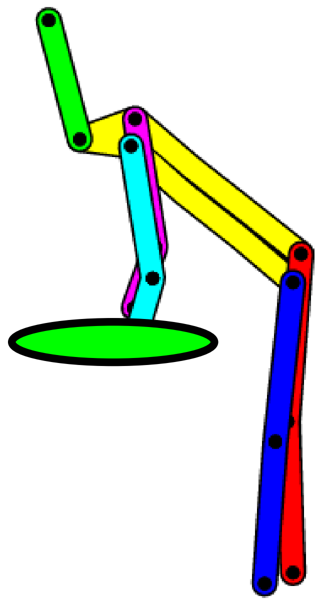
Sittable Regions at Pelvic Joint

From Poses to Functional Regions



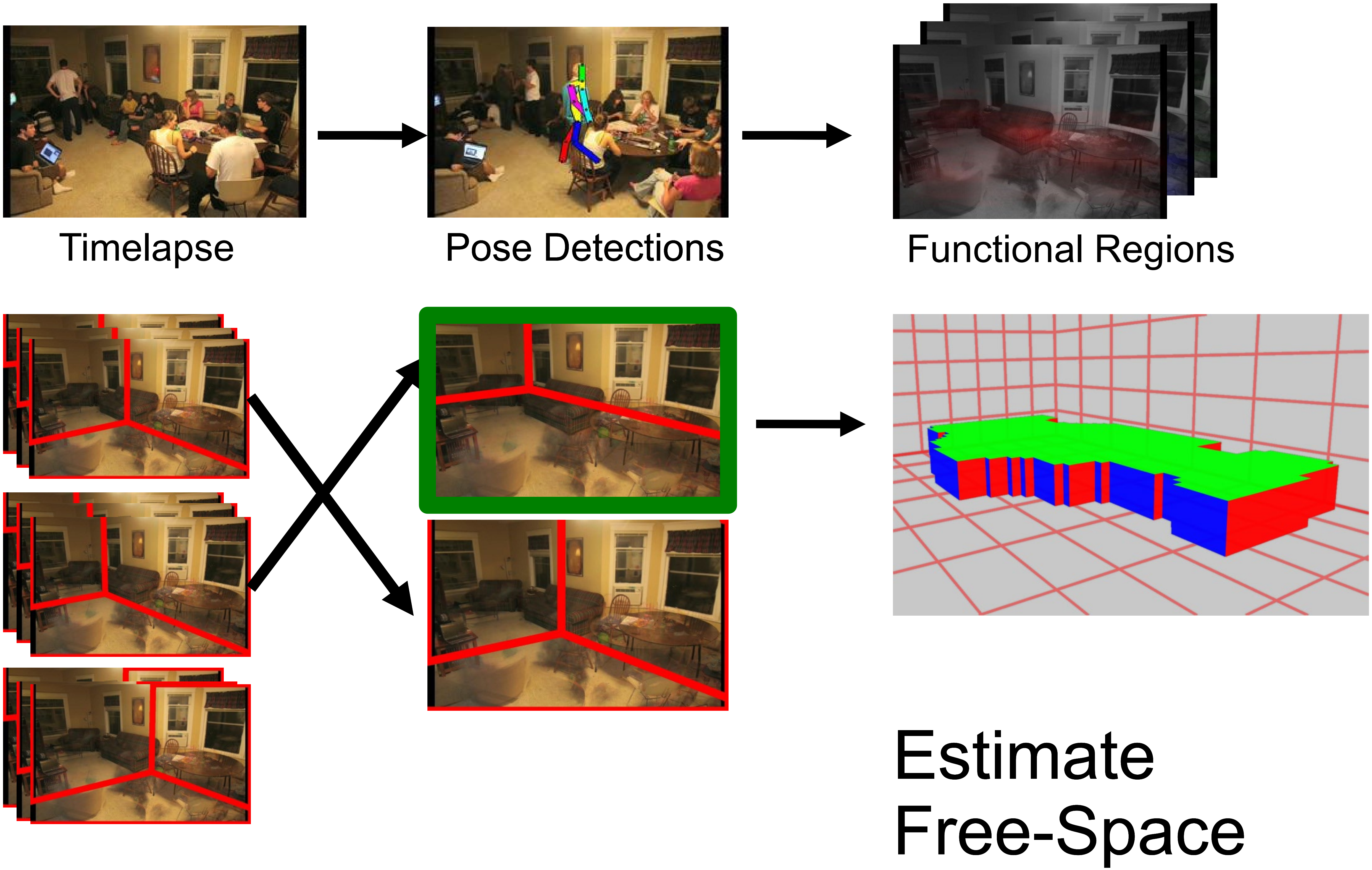
Walkable Regions at Feet

Affordance Constraints

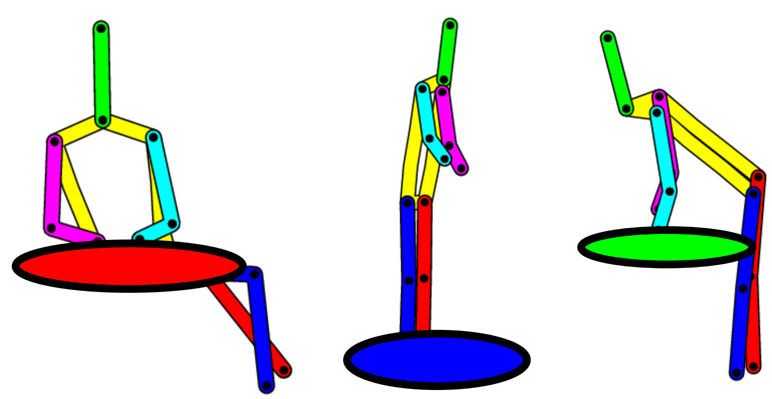


Reachable Regions at Hands

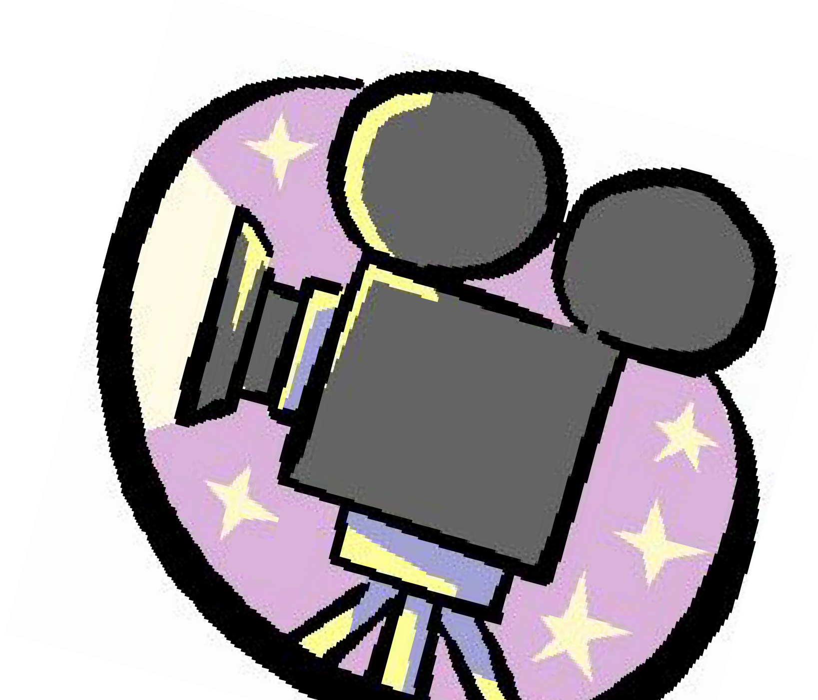
Our Approach



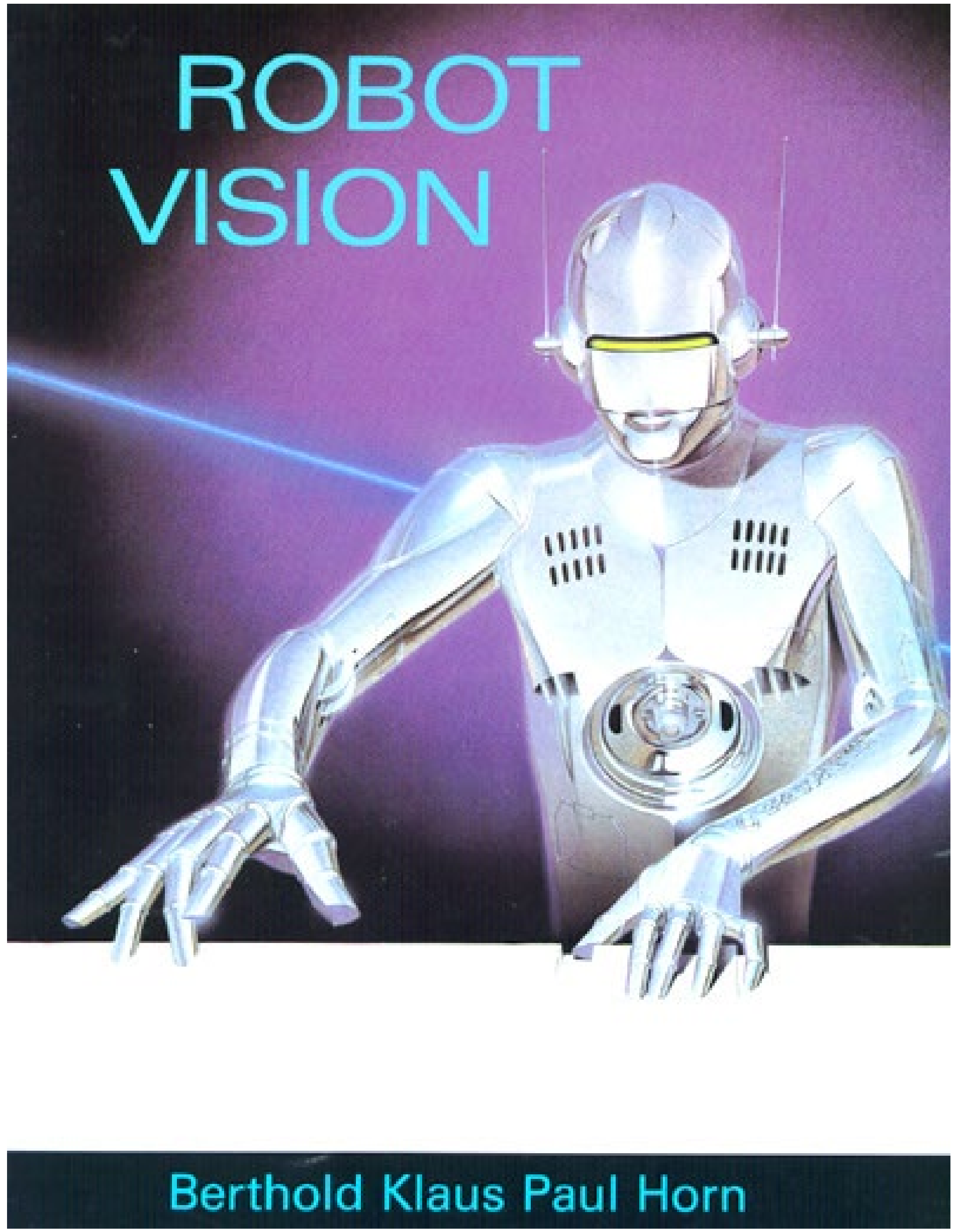
Sample Result



e.g. **24/7 Poke-bot Farm**



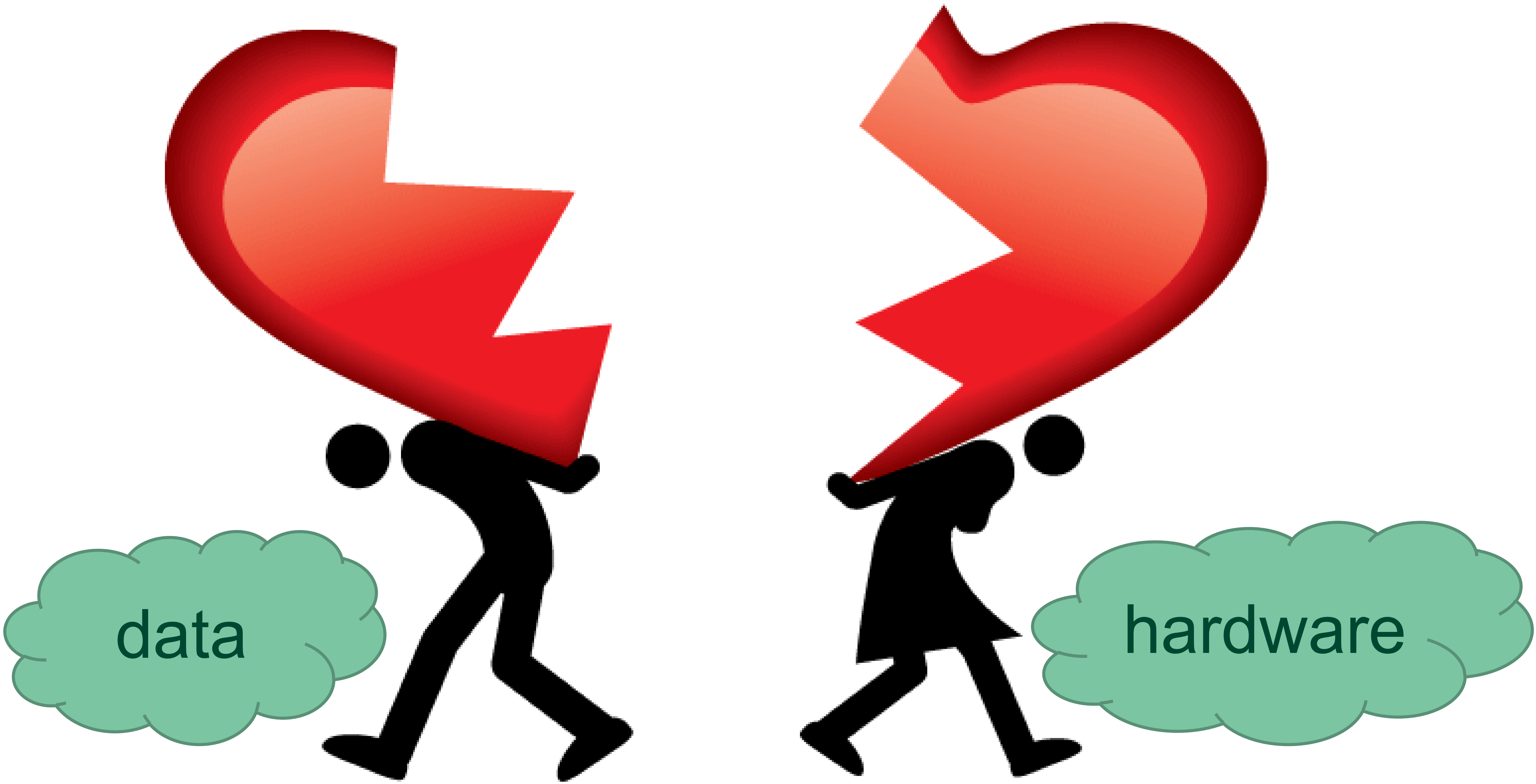
Robotics and Vision used to be “an item”



Horn, 1986

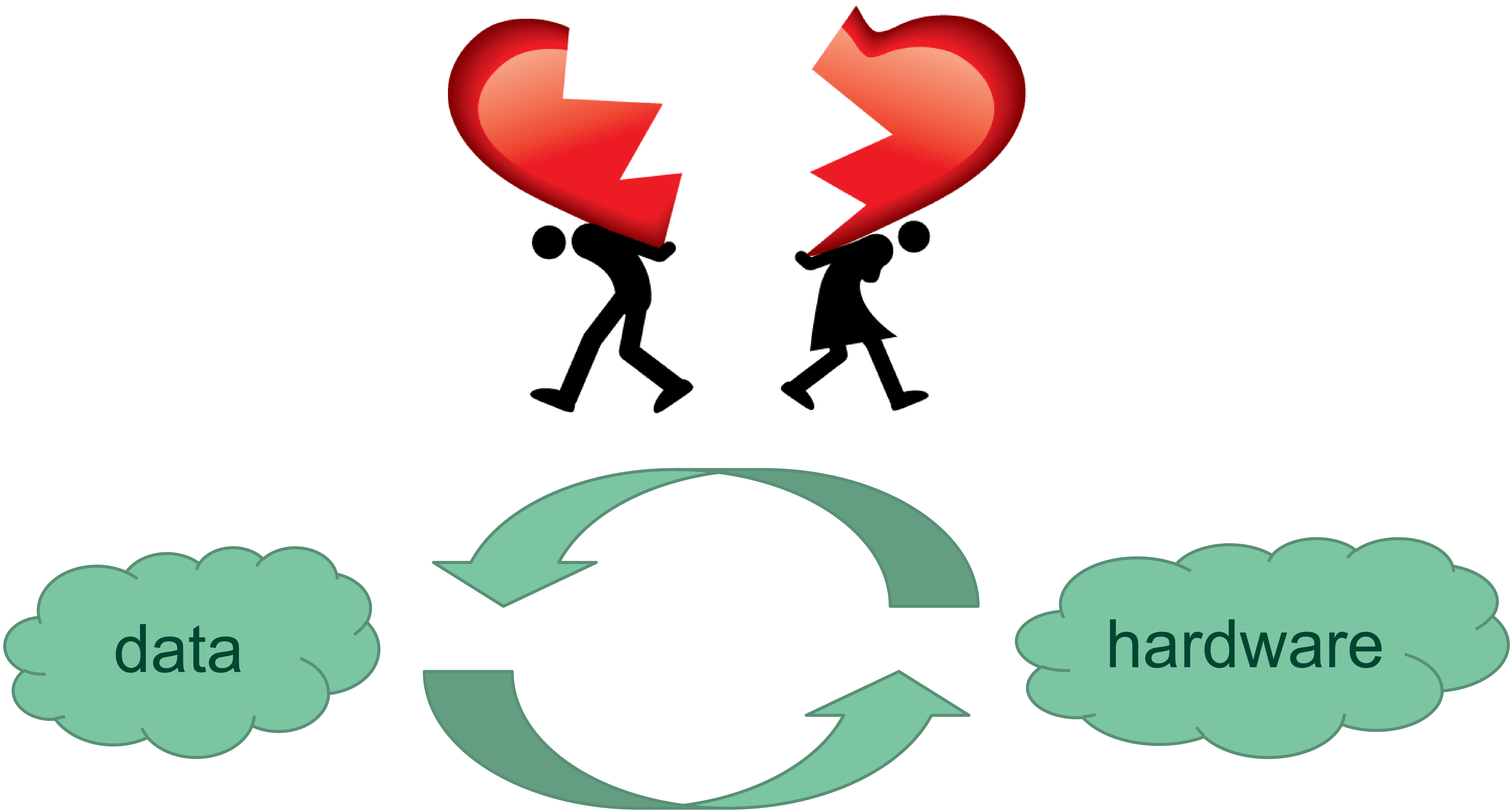
seduced by…

`

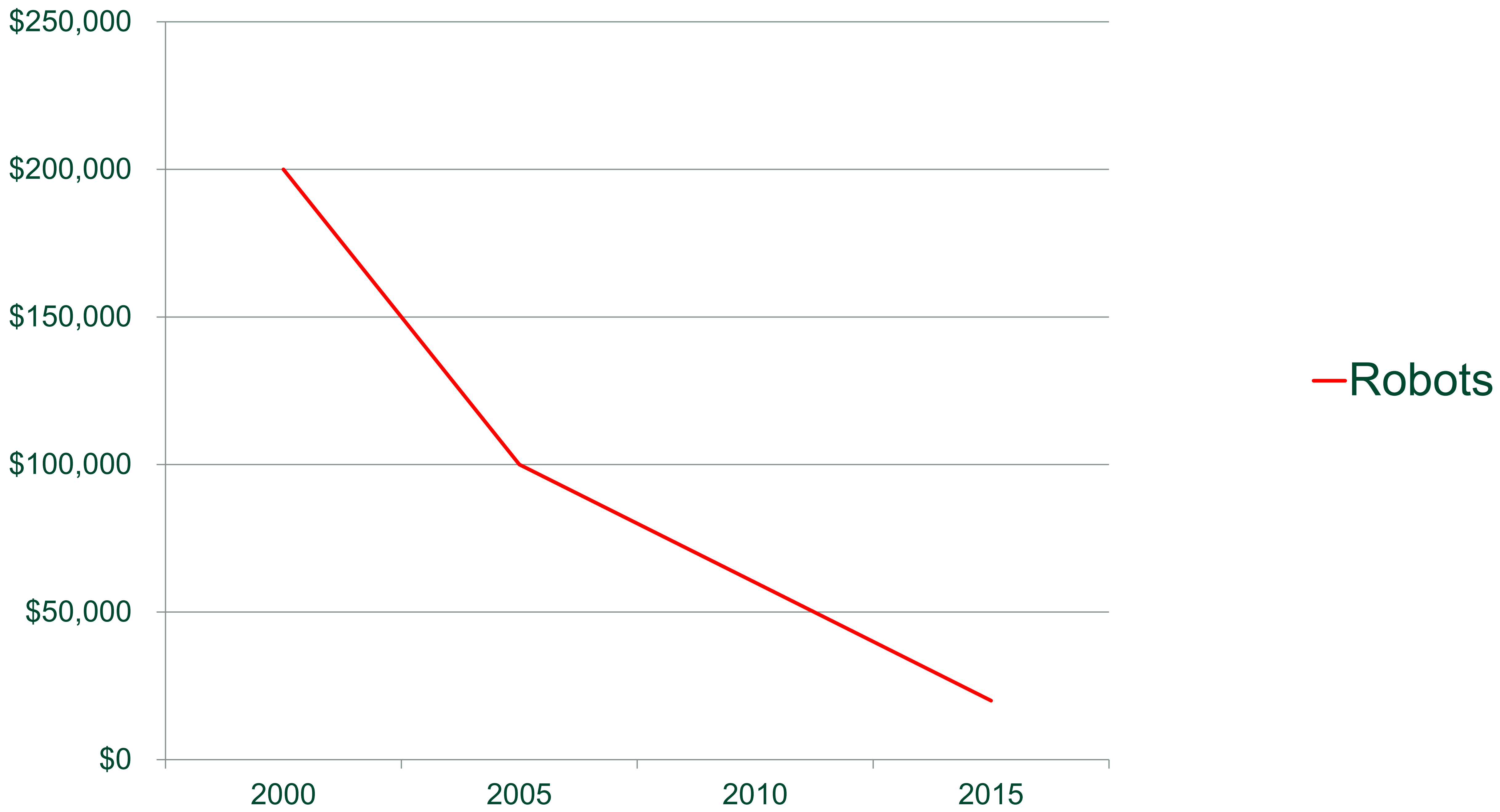


but actually…

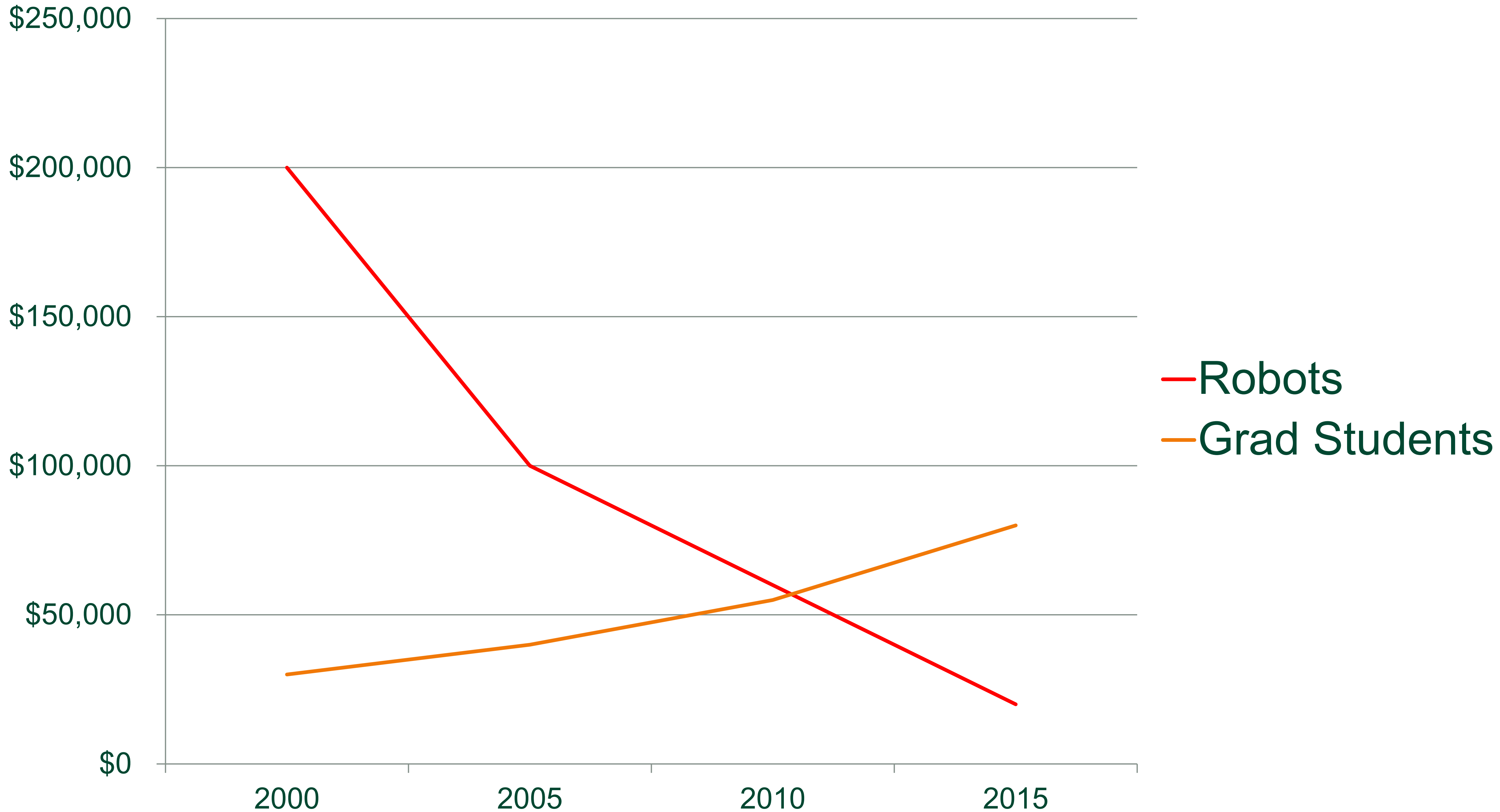
`



Time to re-evaluate?



Time to re-evaluate?



Why do we have vision?

"Why, why, why…the question the pigs ask the wolf, that the wolf asks the hunter, that the hunter asks God–and the answer, as it comes at midnight, after all the other, patient parental answers (…), the final exhausted midnight-in-the-lamplight answer (…) is the only answer there is, the Bible's answer to Job: because that's the way the puppet master chose to do it, because that's the way the guy who works the puppets chose to have it done."

-- Adam Gopnik, *Paris to the Moon*