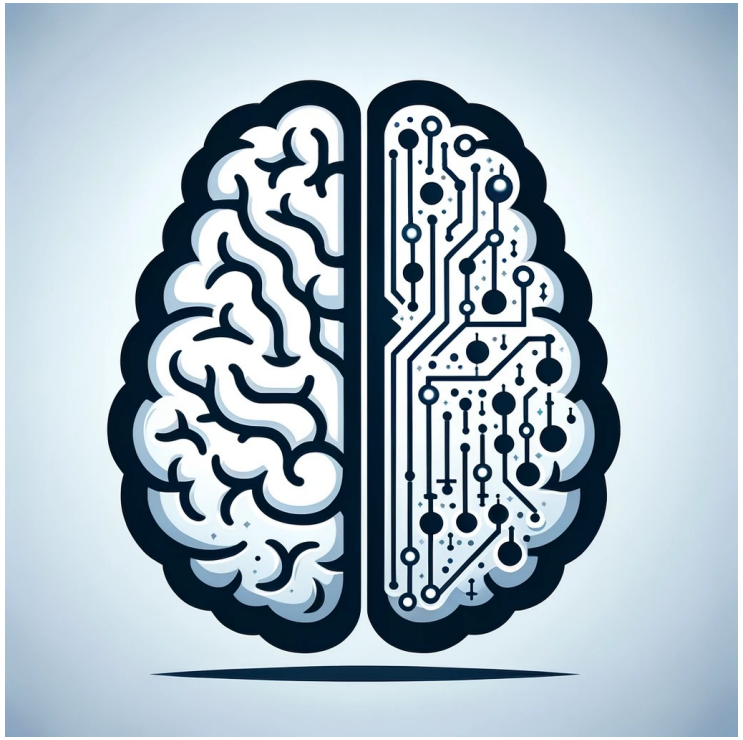
EN 601.473/601.673: Cognitive Artificial Intelligence (CogAI)

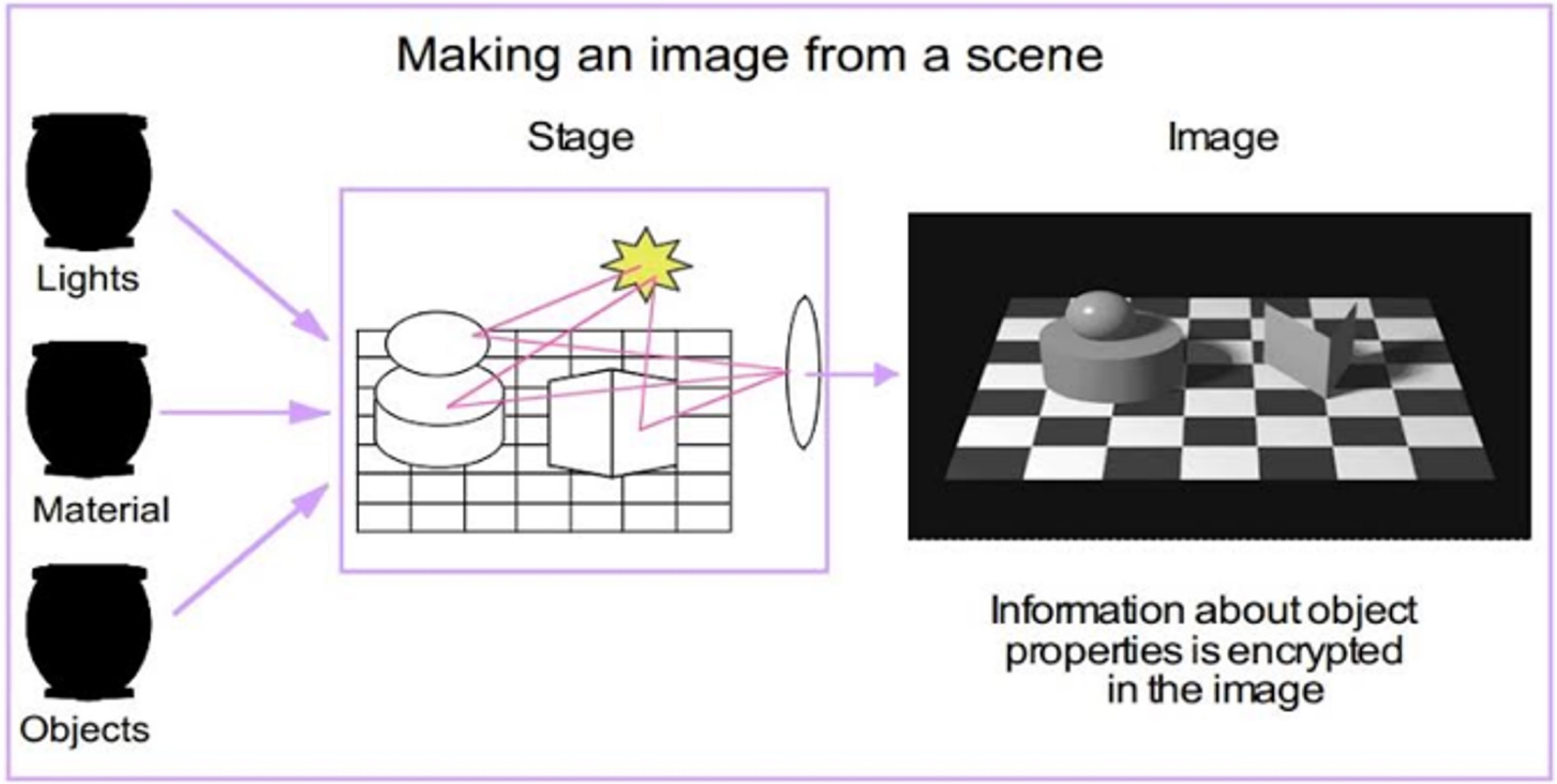


Lecture 16: Physical scene understanding

Tianmin Shu

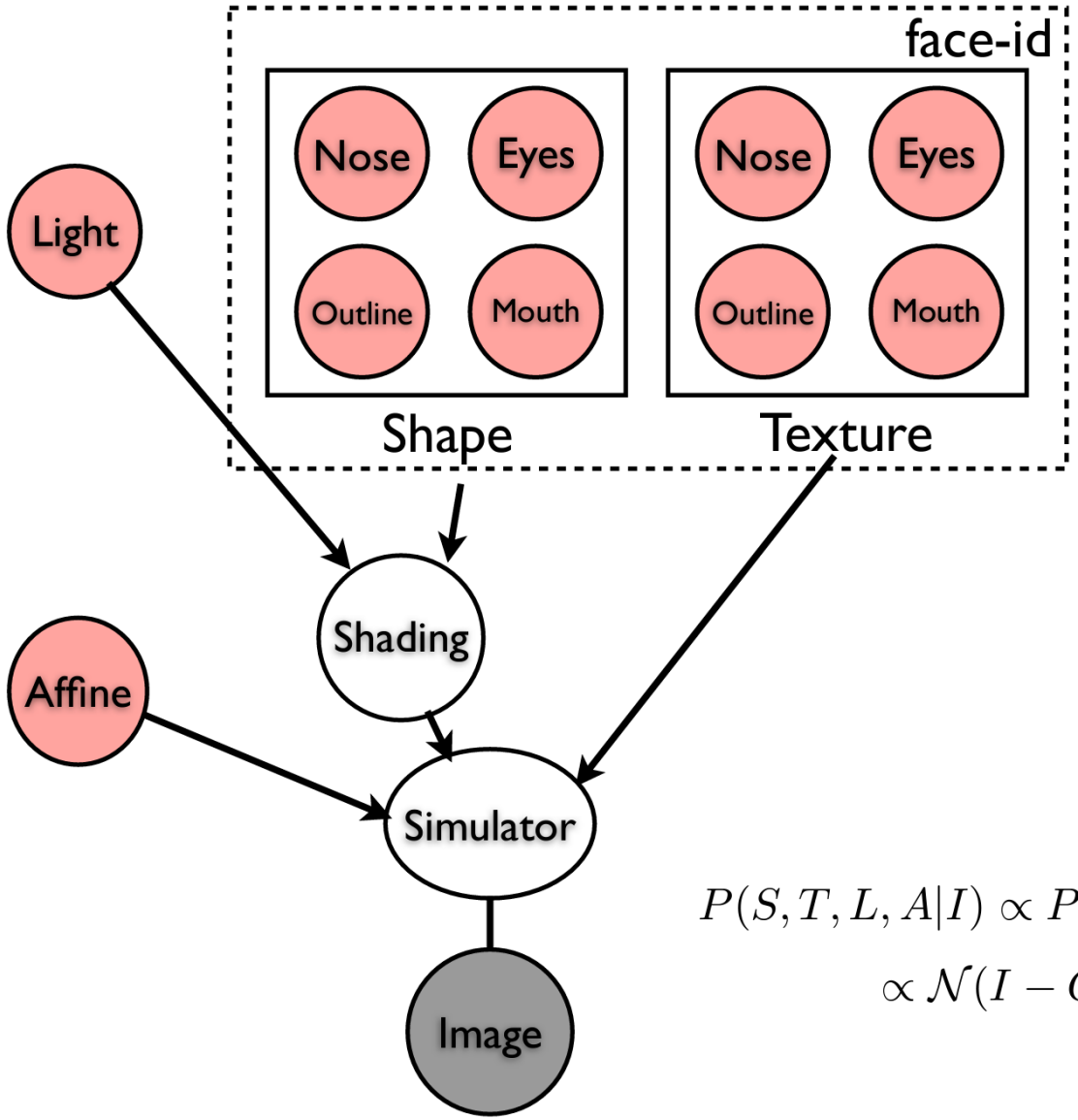
Vision as inverse graphics

3D scene reconstruction



[Kersten, NeurIPS 1998 Tutorial on Computational Vision]

Vision as inverse graphics



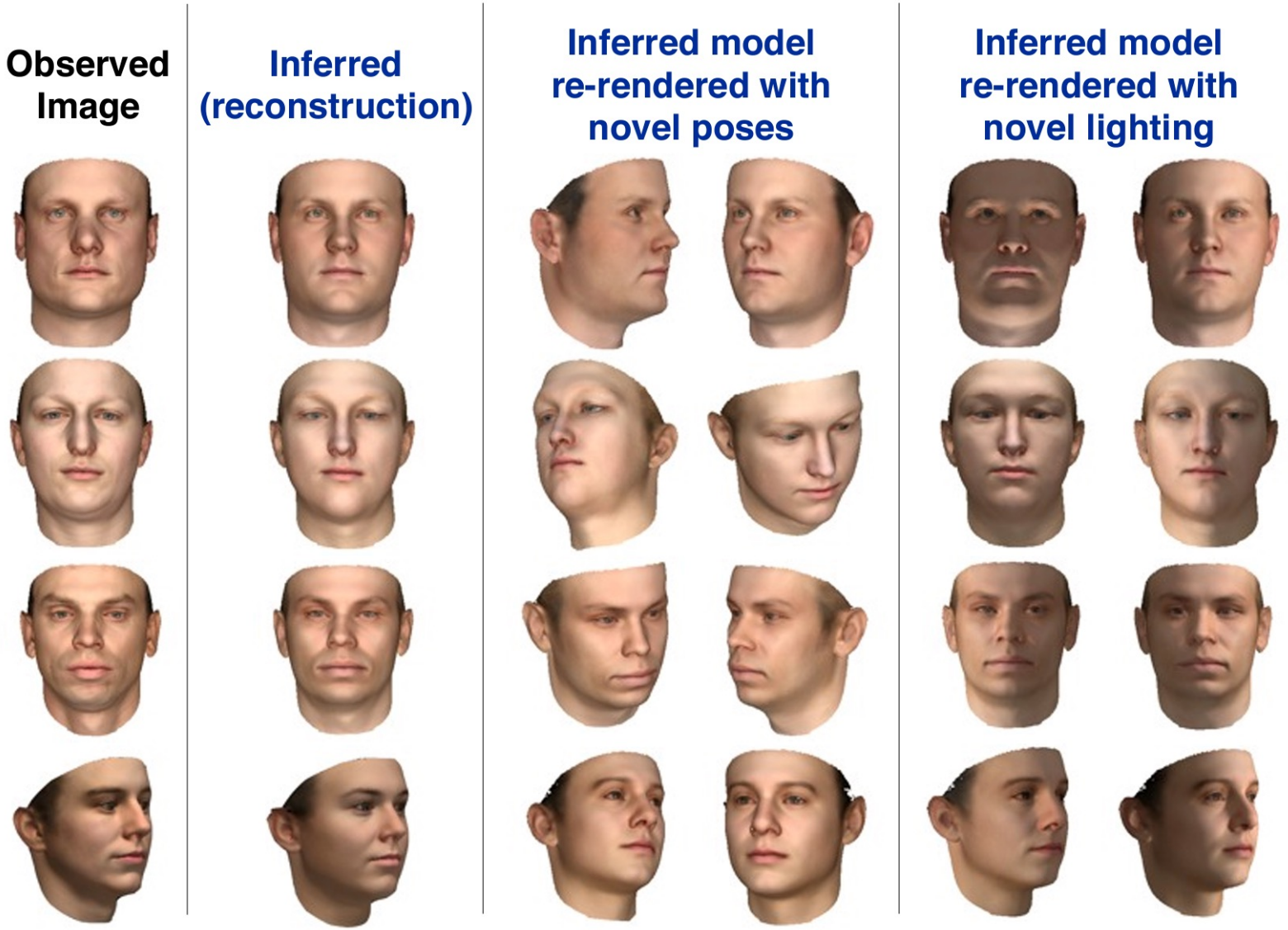
3D face reconstruction

Inference Problem:

$$P(S,T,L,A|I) \propto P(I|S,T,L,A)P(L)P(S)P(T)P(A)$$
$$\propto \mathcal{N}(I-O;0,0.1)P(L)P(A)\prod_i P(S_i)P(T_i)$$

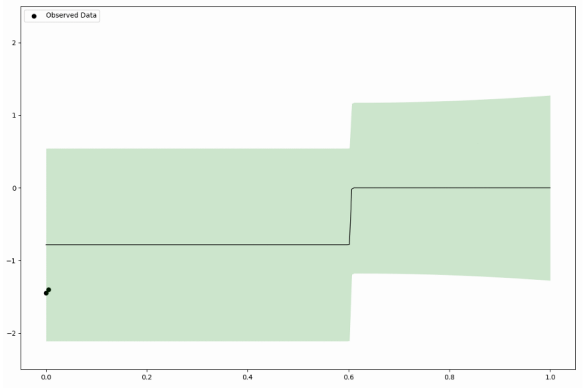
[From Tejas Kulkarni]

Probabilistic 3D face reconstruction



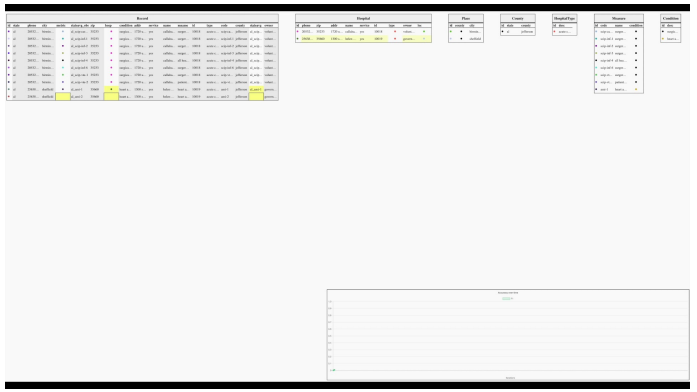
Applications where probabilistic programming outperforms SOTA machine learning

Automated data modeling



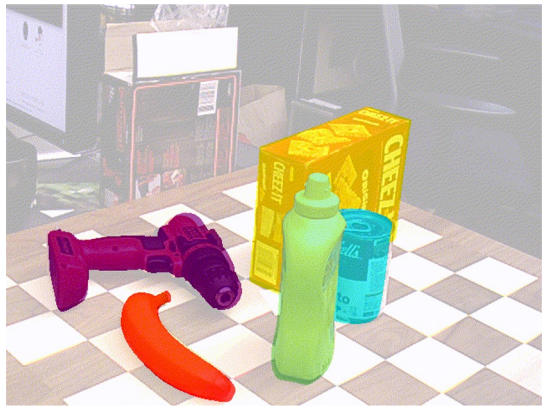
Saad, Cusumano-Towner, Schaechtle, Rinard, Mansinghka (POPL 2019)
Saad and Mansinghka (UAI 2021)
Schaechtle, Freer, Shelby, Saad, and Mansinghka (AutoML 2022)

Common-sense data cleaning



Lew, Agrawal, Sontag, and Mansinghka (AISTATS 2021)

3D scene perception



Gothoskar et al. (NeurIPS 2021)

Vision as inverse graphics

***Picture*: A Probabilistic Programming Language for Scene Perception**

Tejas D Kulkarni
MIT
tejask@mit.edu

Pushmeet Kohli
Microsoft Research
pkohli@microsoft.com

Joshua B Tenenbaum
MIT
jbt@mit.edu

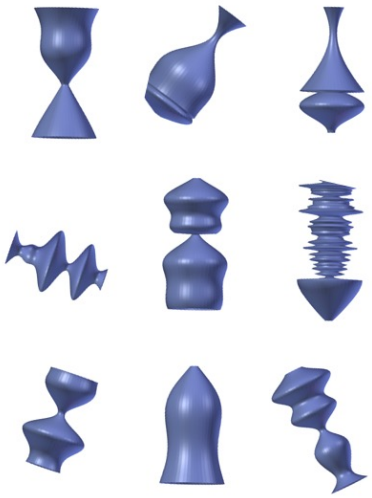
Vikash Mansinghka
MIT
vkm@mit.edu

CVPR 2015 (Best paper honorable mention)

p(h, I): a probabilistic program using graphics rendering

Random samples I_R drawn from example probabilistic programs

3D Face program



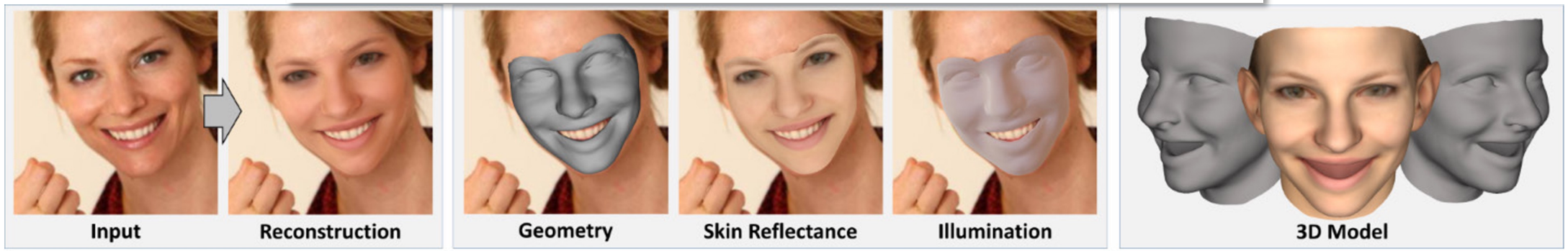
3D object program

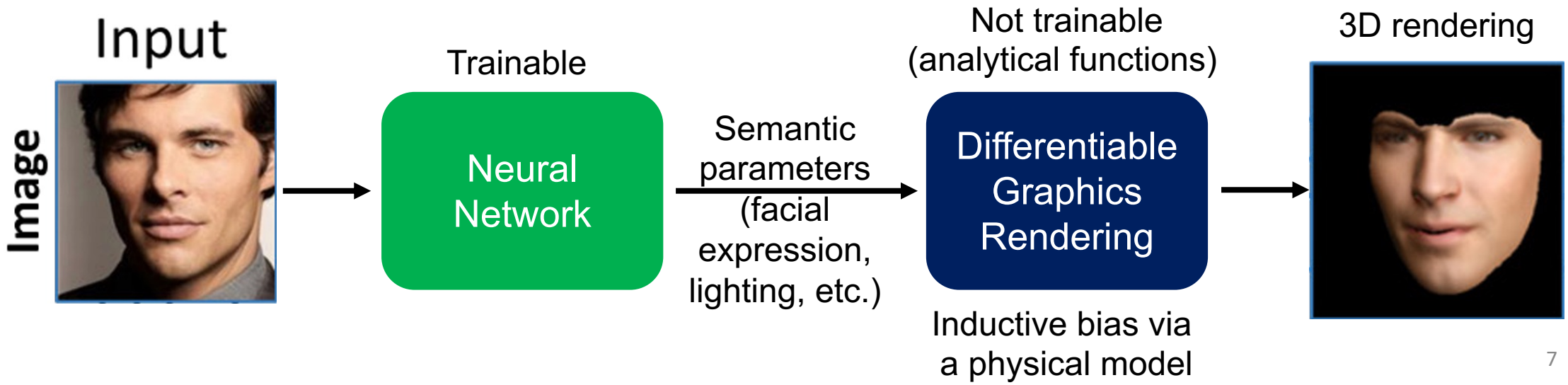
3D human-pose program

Differentiable 3D rendering

MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction

Ayush Tewari[1] Michael Zollhöfer[1] Hyeongwoo Kim[1] Pablo Garrido[1]
Florian Bernard[1,2] Patrick Pérez[3] Christian Theobalt[1]

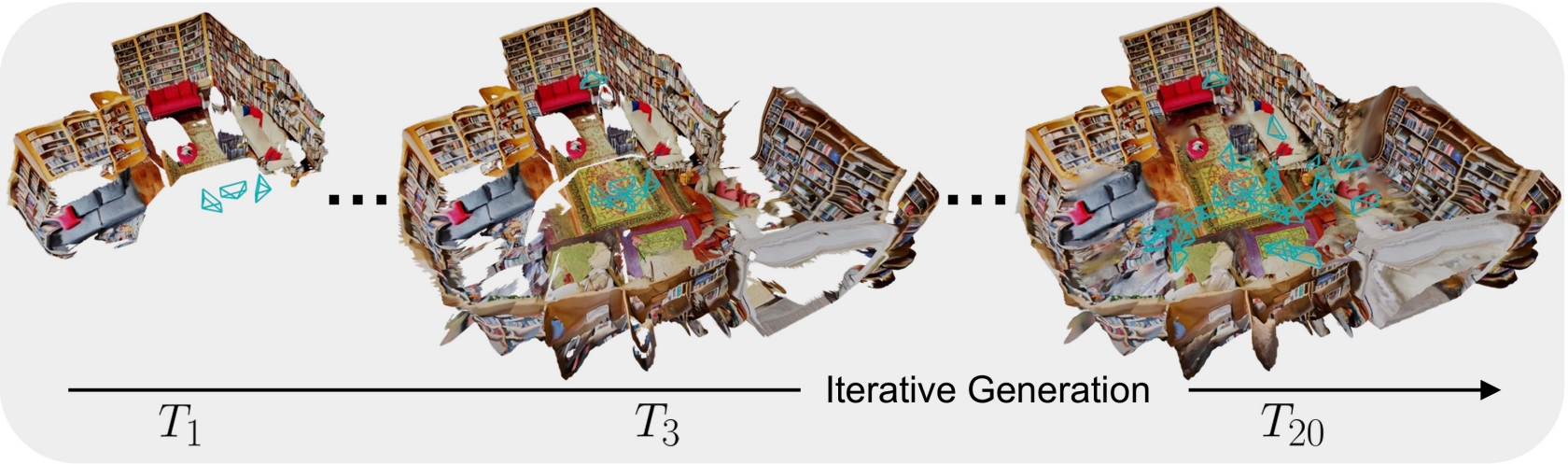




Neural 3D rendering

Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models

Lukas Höllein[1*] Ang Cao[2*] Andrew Owens[2] Justin Johnson[2] Matthias Nießner[1]
[1]Technical University of Munich [2]University of Michigan



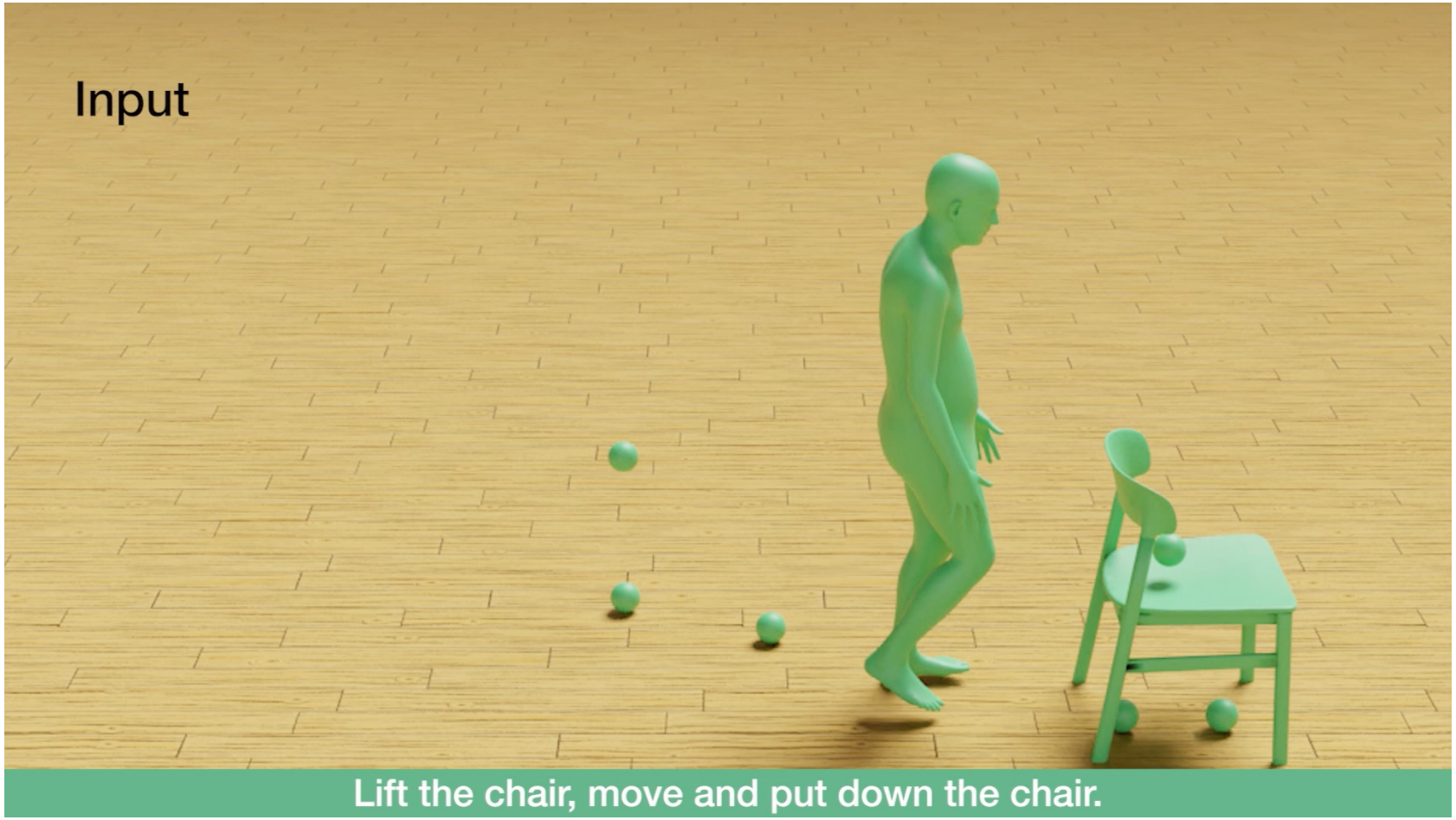
(a) 3D Mesh Generation from Text

(b) Rendered Image + Mesh

Neural 3D rendering

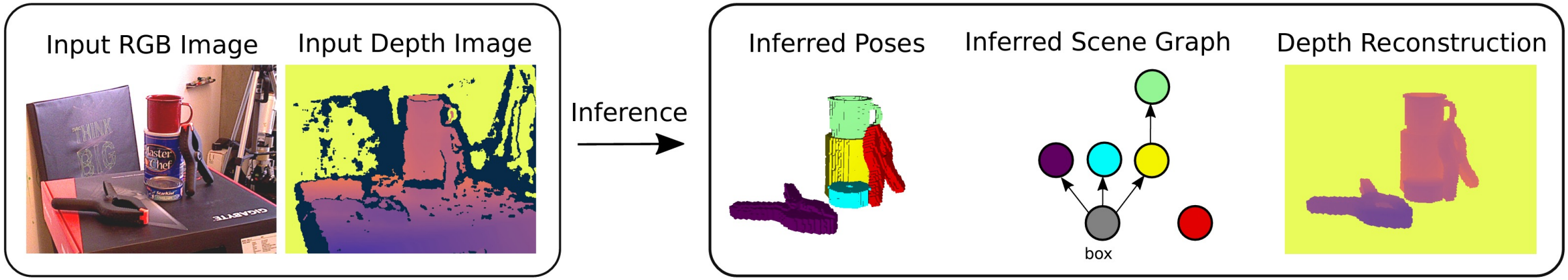
Controllable Human-Object Interaction Synthesis

Jiaman Li[1], Alexander Clegg[2], Roozbeh Mottaghi[2], Jiajun Wu[1], Xavier Puig[2†], C. Karen Liu[1†]
[1]Stanford University, [2]FAIR, Meta

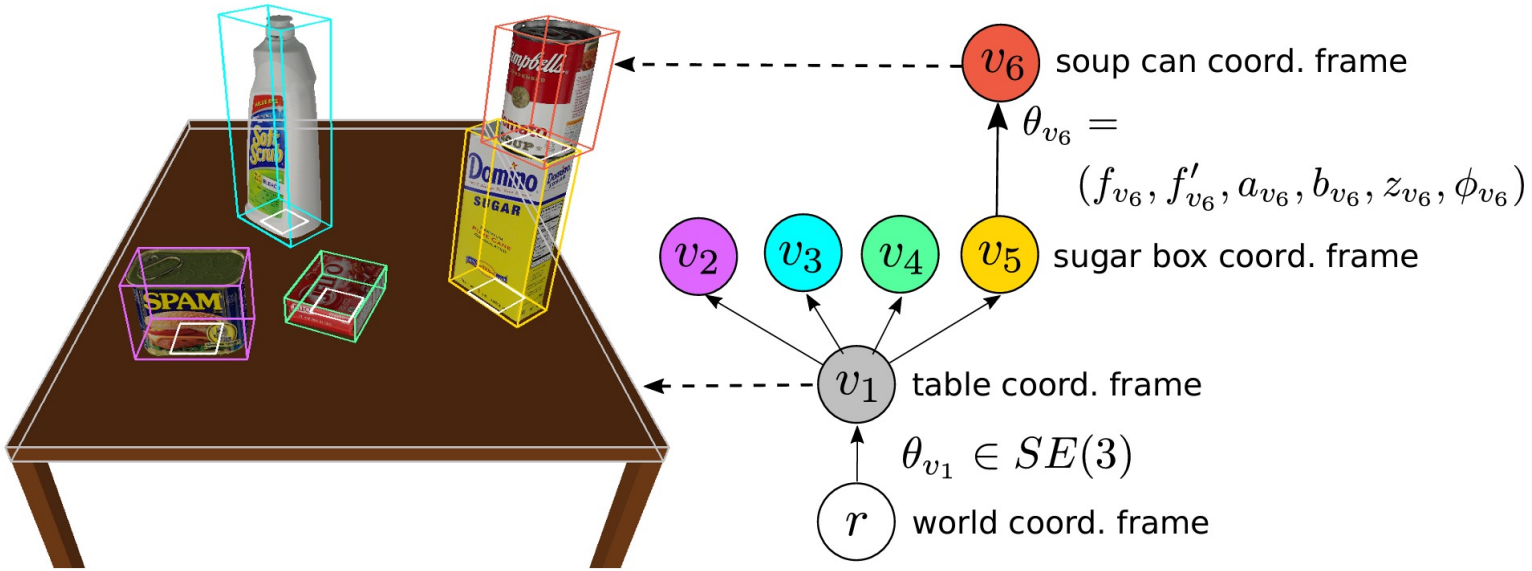


3D scene perception in Gen

Gothoskar et al. (2021)



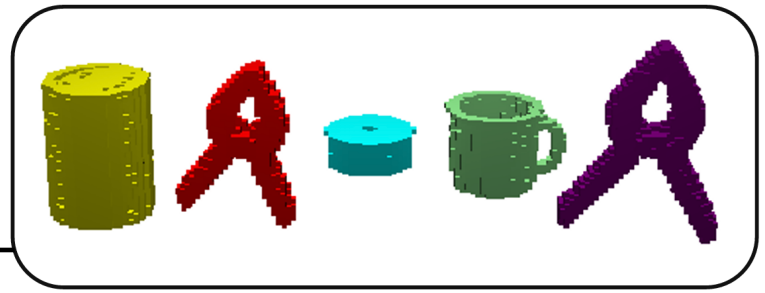
Hierarchical scene graph

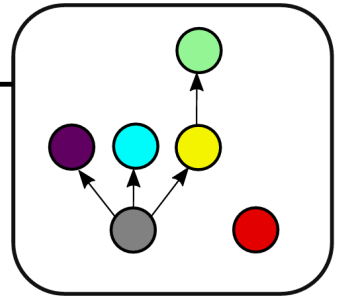


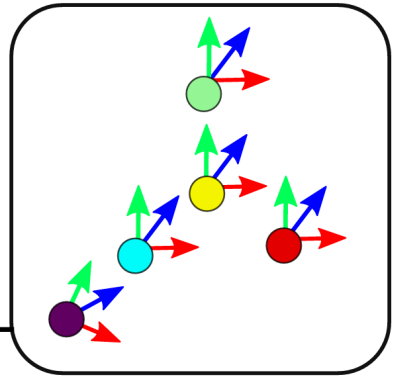
3DP3: 3D Scene Perception via Probabilistic Programming model

Generative model in Gen

procedure 3DP3-GENERATIVE-MODEL(M,N)
  parameters: M (= num object types),
    N (= num objects in scene)
  for $i \in \{1, \ldots, M\}$ **do**
    $\mathbf{O}^{(i)} \sim$ VOXEL-SHAPE-PRIOR(i)
  end for
  $\mathbf{c} \sim$ OBJECT-CLASS-ASSIGNMENTS-PRIOR(N, M)
  $G := (V, E) \sim$ GRAPH-STRUCTURE-PRIOR(N)
  for $v \in$ TOPO-SORT(V, E) where $(u, v) \in E$ **do**
    if $u = r$ **then**
      $\theta_v \sim$ UNIFORM-6DOF-POSE-PRIOR()
      $\mathbf{x}_v \leftarrow \theta_v$
    else
      $\theta_v \sim$ CONTACT-PRIOR($G, v, \mathbf{O}^{(1:M)}, \mathbf{c}$)
      $\mathbf{x}_v \leftarrow \mathbf{x}_u \cdot$ COMPUTE-RELATIVE-POSE($\theta_v, \mathbf{O}^{(1:M)}, \mathbf{c}$)
    end if
  end for
  $\tilde{\mathbf{Y}} \leftarrow$ UNPROJECT(DEPTH-RENDER($\mathbf{x}, \mathbf{O}^{(1:M)}, \mathbf{c}$))
  $\mathbf{Y} \sim$ NOISE-MODEL($\tilde{\mathbf{Y}}$)
end procedure

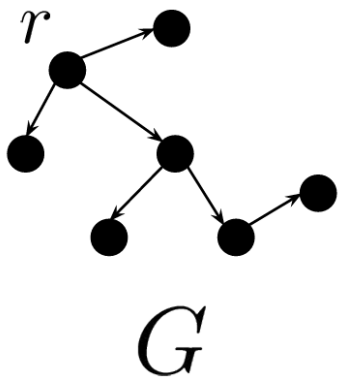






Inference: RJMCMC

Sever-graft move for reversible transition between graphs



Results

DenseFusion (an end-to-end method)

3DP3 with initialized with DenseFusion

Summary

- Generative models of the world, p(h, X)
- Can be implemented via
  - Probabilistic programs
  - Simulators
  - Differentiable generative models (analytical models or neural models)
  - Or a combination of all above options
- Probabilistic programs allow sampling in infinity and complex hypothesis space
- Neural amortized inference can be efficient and robust
  - Efficient: high quality proposals that narrow down the hypothesis space or produce a good initial guess
  - Robust: estimate uncertainty by sampling & evaluating multiple hypotheses, generative models allow reasoning in unfamiliar / corner cases (OOD generalization)

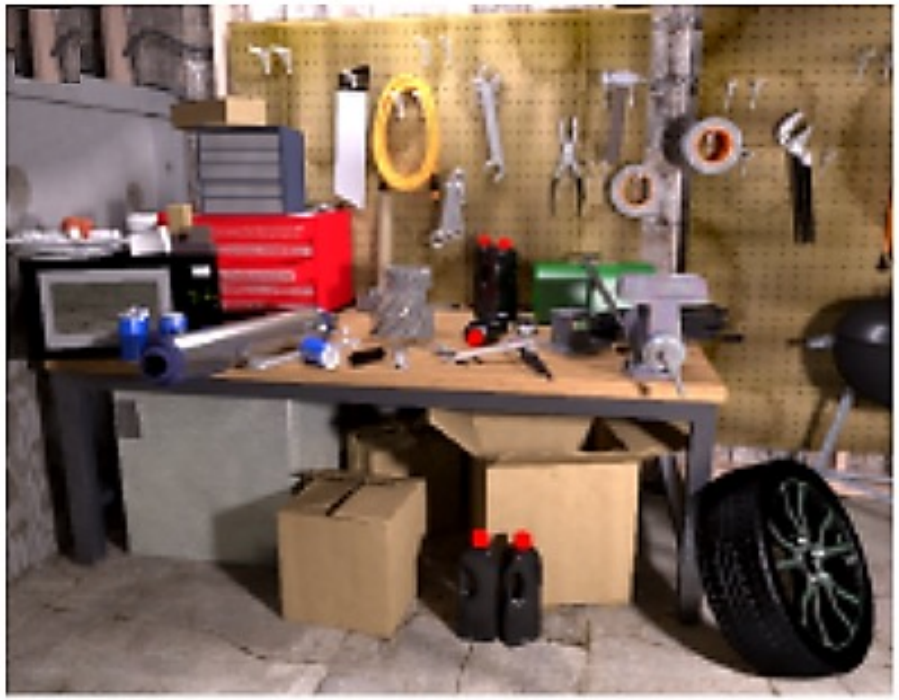
Physical and social reasoning

- Common sense scene understanding, intuitive theories
- Physical reasoning
- Social reasoning
- Implemented using the CogAI toolbox

Physical and social reasoning

- Common sense scene understanding, intuitive theories
- Physical reasoning
- Social reasoning

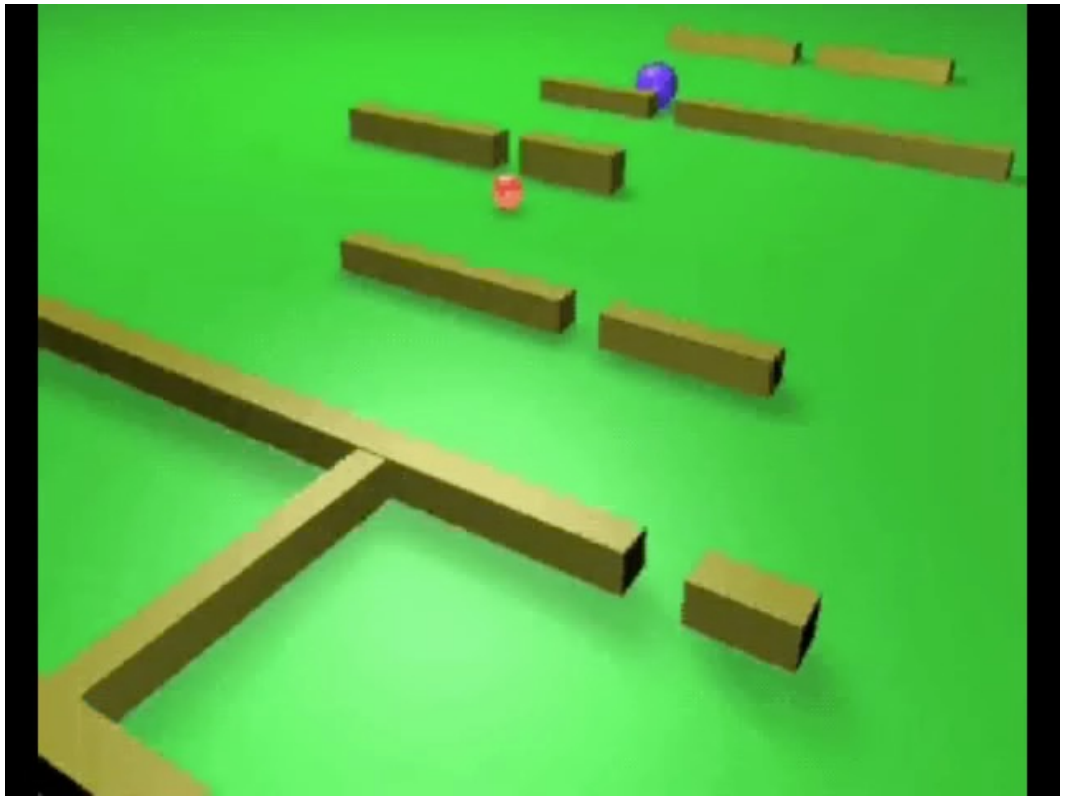
Commonsense scene understanding



The roots of common sense

The roots of common sense

Understanding dynamic events with commonsense theories



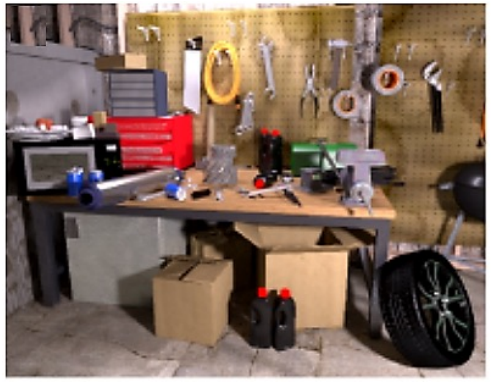
(Southgate and Csibra)

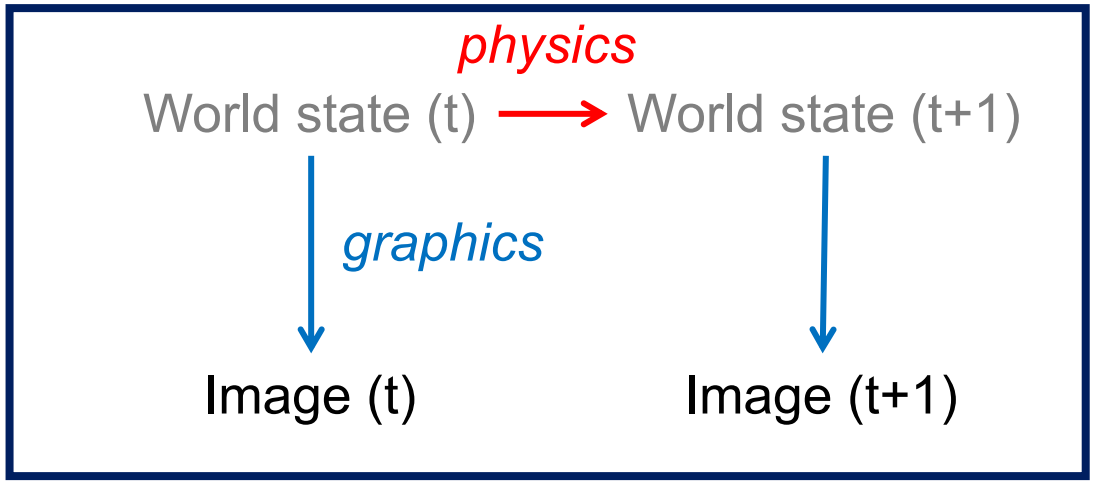
(Heider and Simmel)

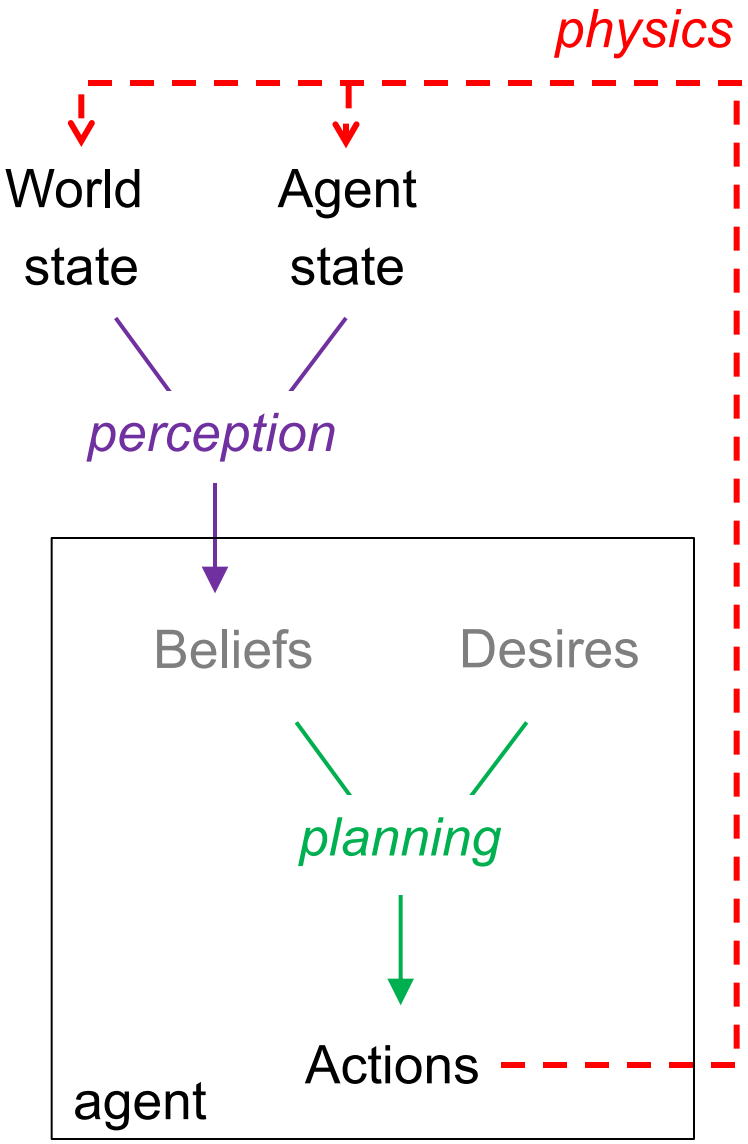
Intuitive physics: objects, forces and masses
Intuitive psychology: beliefs and desires
Intuitive sociology: us and them
Intuitive morality: good and bad

Reverse-engineering common sense

Probabilistic programs as generative models of intuitive physics, intuitive psychology







Physical and social reasoning

- Common sense scene understanding, intuitive theories

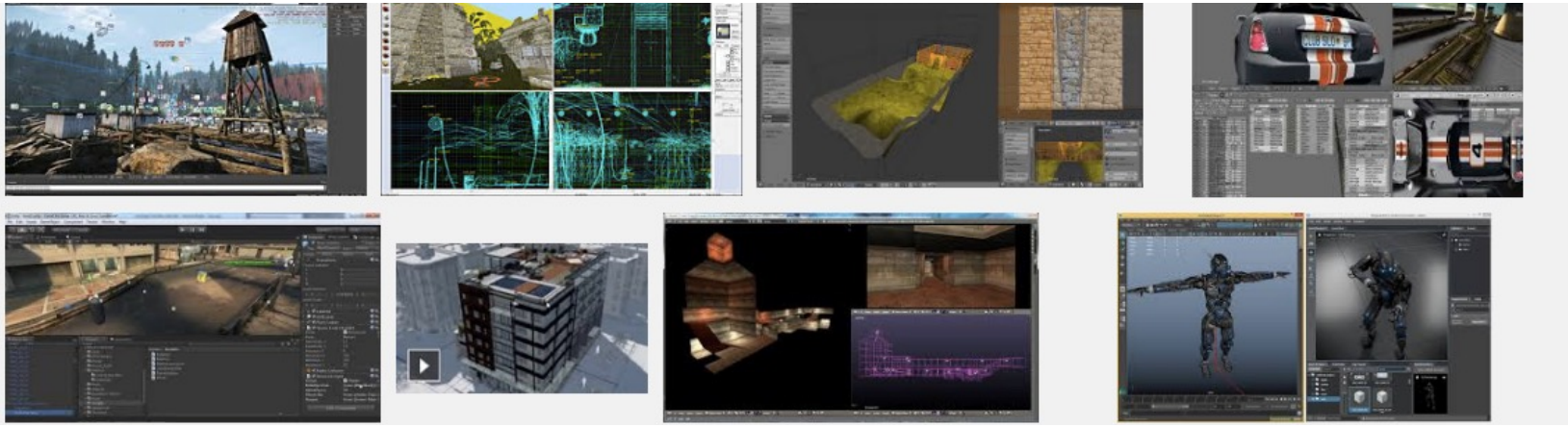
- Physical reasoning

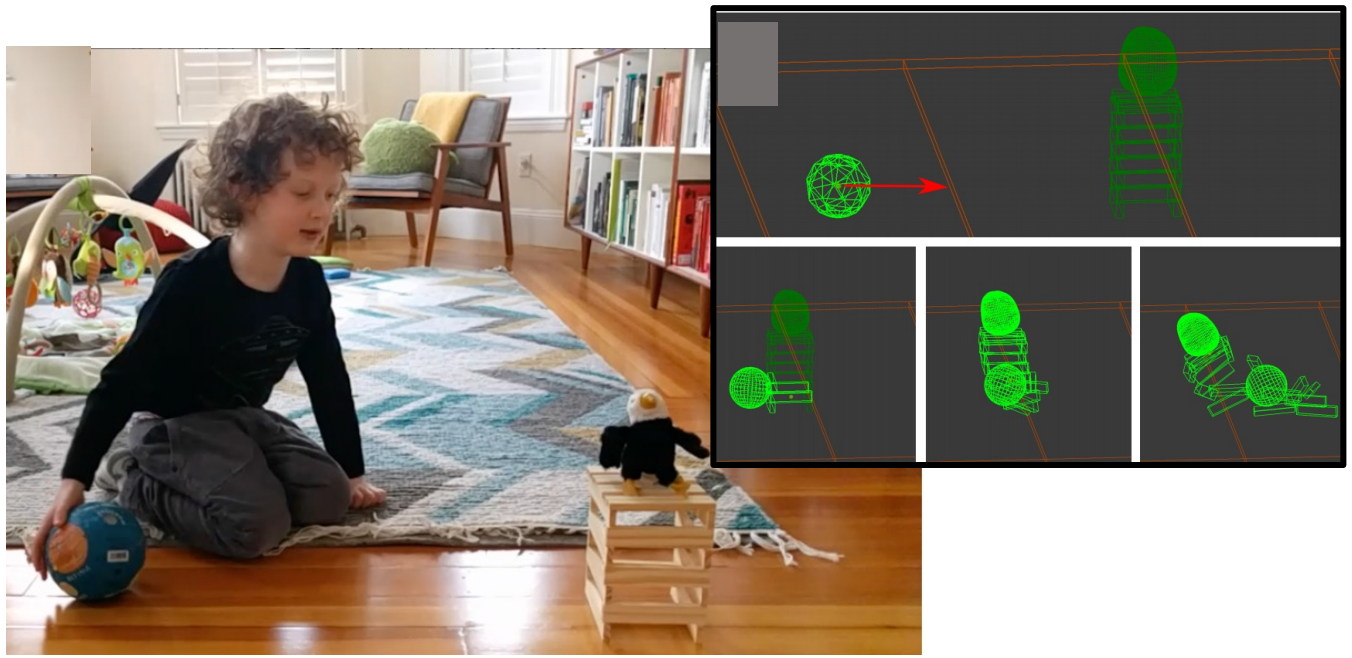
- Social reasoning

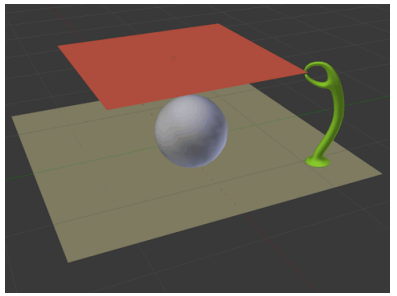
Generative models based on simulation in video game-style representations of the world

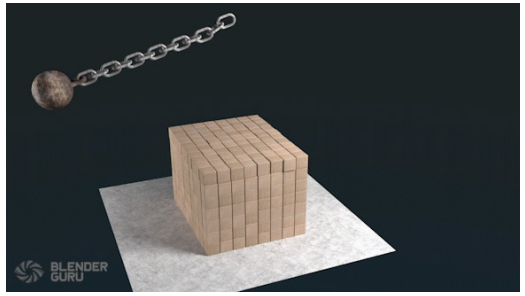
"The game engine in your head":

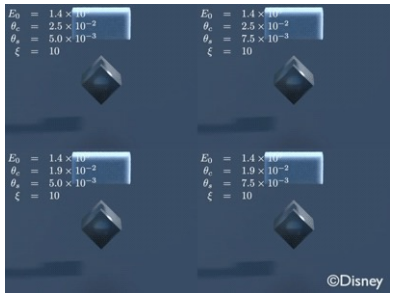
- Very fast
- Approximate programs for simulating graphics, physics, planning…











The intuitive physics engine

Physics

… World state (t-1) → World state (t) → World state (t+1) …

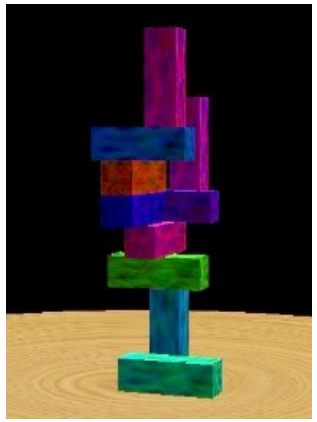
Vision

Graphics

Image (t-1)

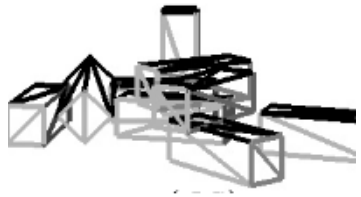
Image (t)

Image (t+1)



(Battaglia et al., *PNAS* 2013; Hamrick et al., *Cognition* 2016)

The intuitive physics engine



… World state (t-1) → World state (t) → World state (t+1) …

Image (t-1)

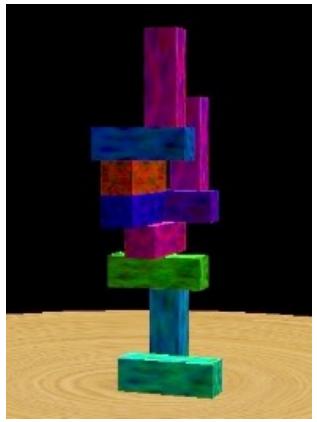
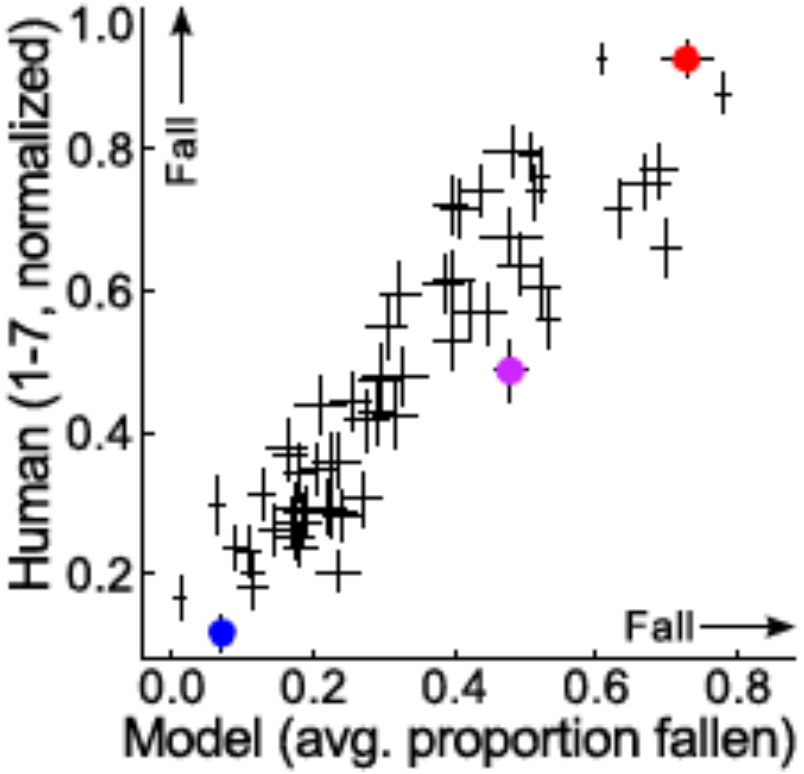


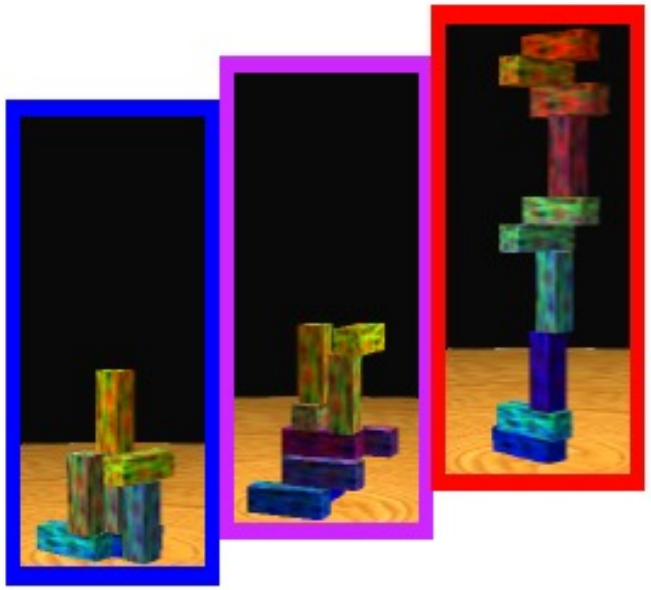
Image (t)

Image (t+1)

(Battaglia et al., *PNAS* 2013; Hamrick et al., *Cognition* 2016)

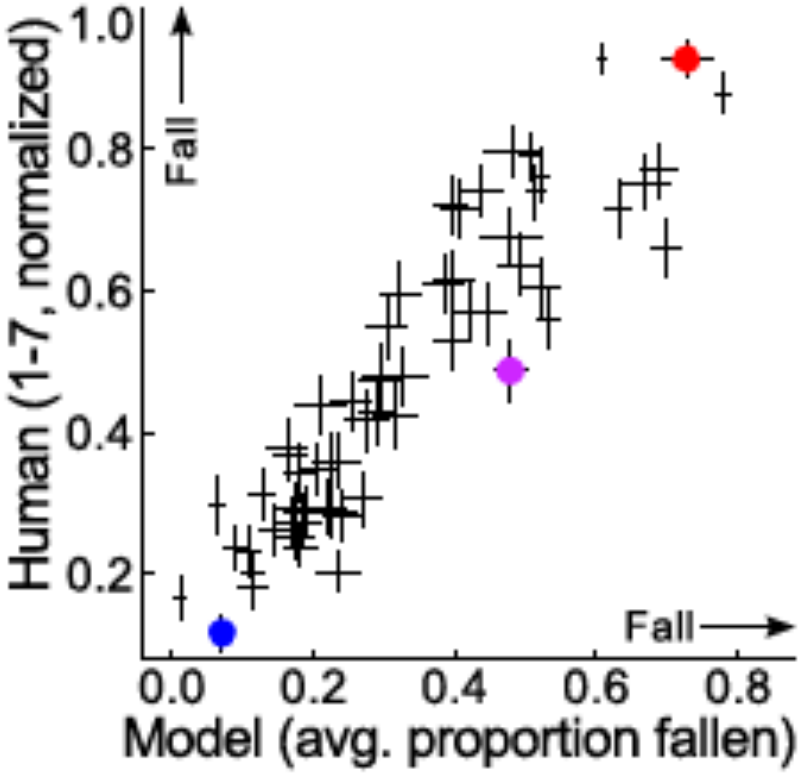
The intuitive physics engine

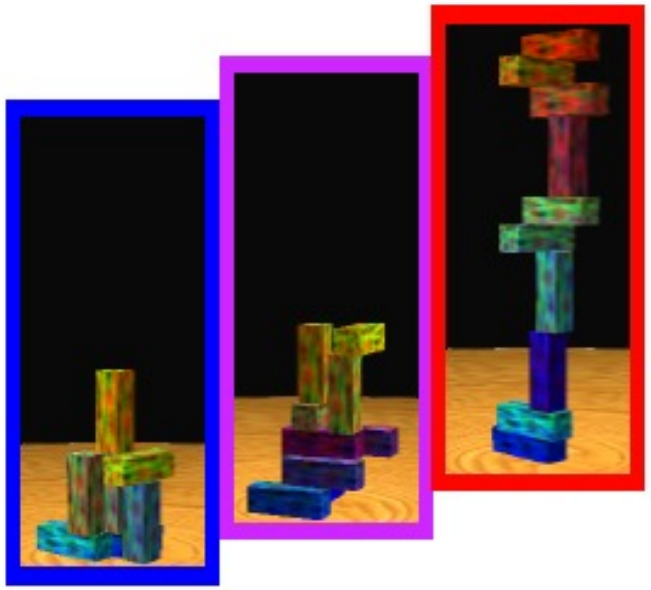




(Battaglia et al., *PNAS* 2013; Hamrick et al., *Cognition* 2016)

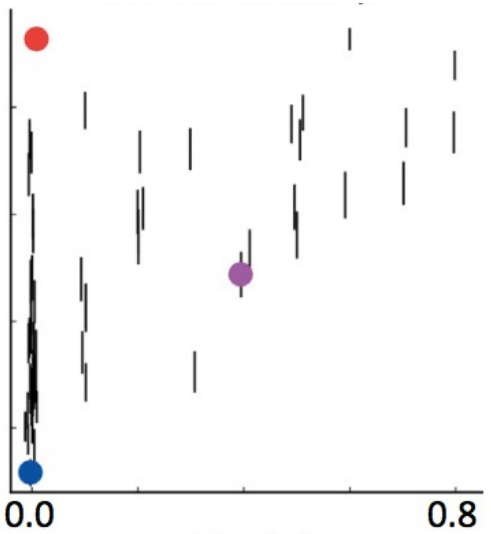
The intuitive physics engine





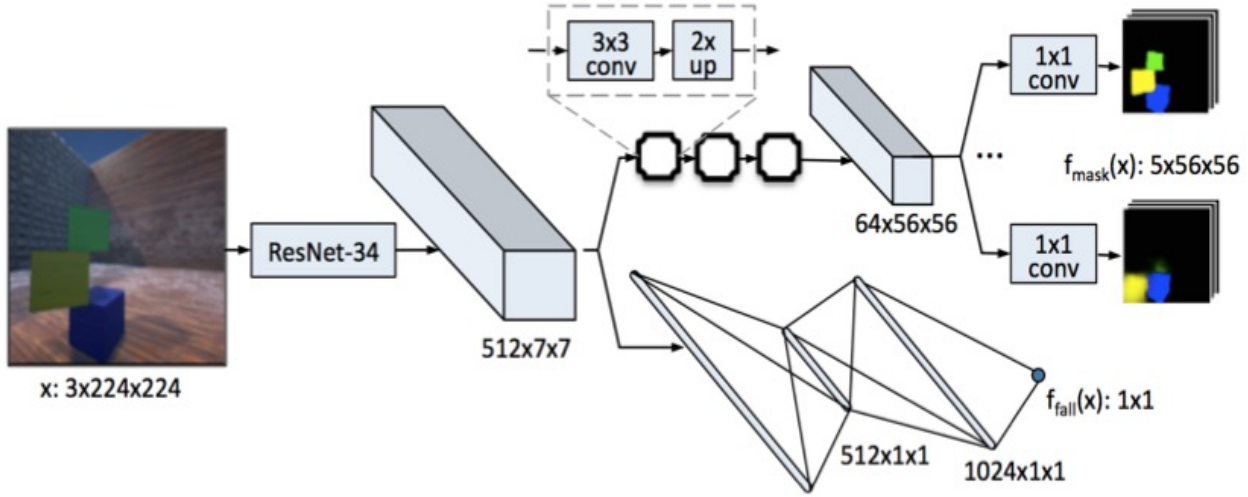
Mental simulations are *probabilistic* and *approximate.*

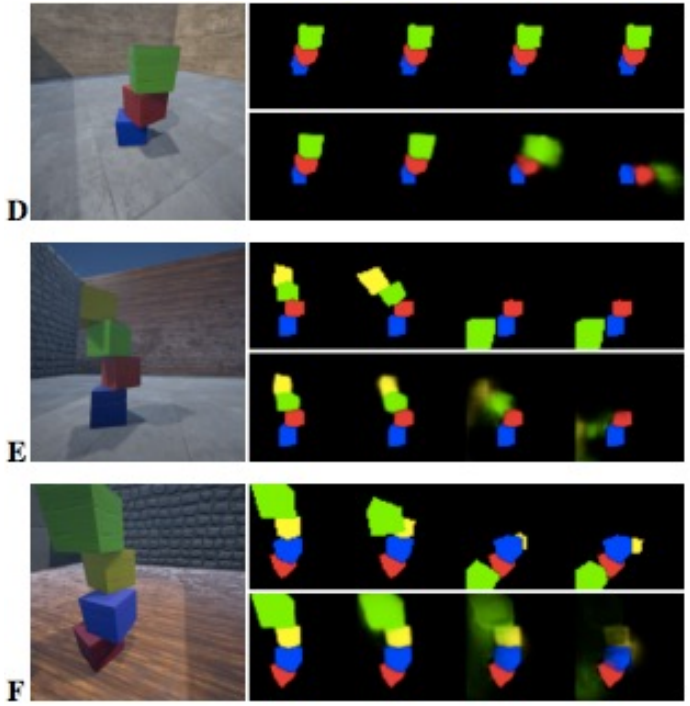
Baseline comparison:
ground truth physics (no uncertainty)



(Battaglia et al., *PNAS* 2013; Hamrick et al., *Cognition* 2016)

An alternative to simulation-based modeling?





Can we treat intuitive physics as a pattern recognition task?

PhysNet (Facebook AI; Lerer et al 2016)

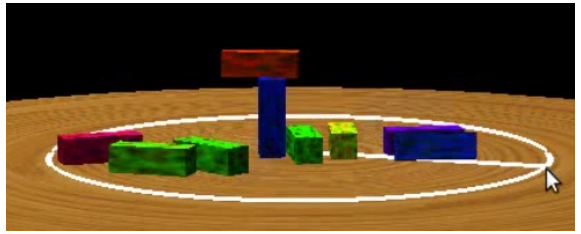
Requires much more training than people get (200K for 2-4 cubes), and doesn't generalize in all the ways that people do.

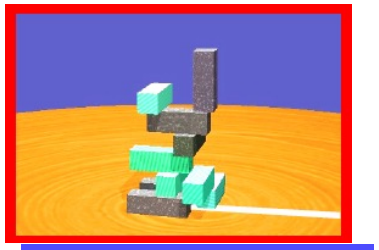
Without explicit representations of objects and their interactions, probably not compositional enough to capture underlying causal structure.

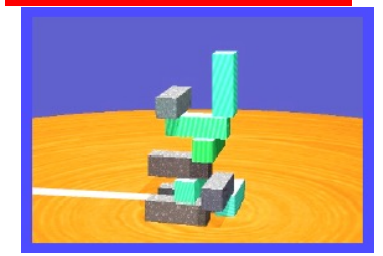
The intuitive physics engine

Will this stack of blocks fall?

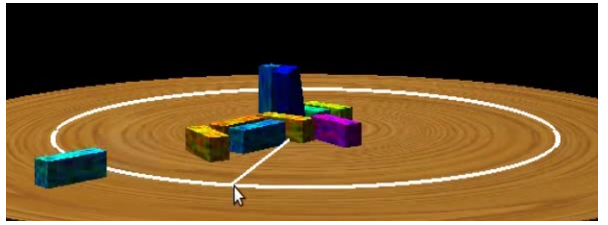
Which way will they fall?



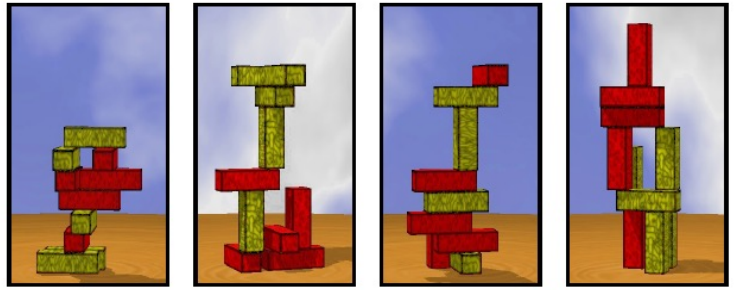




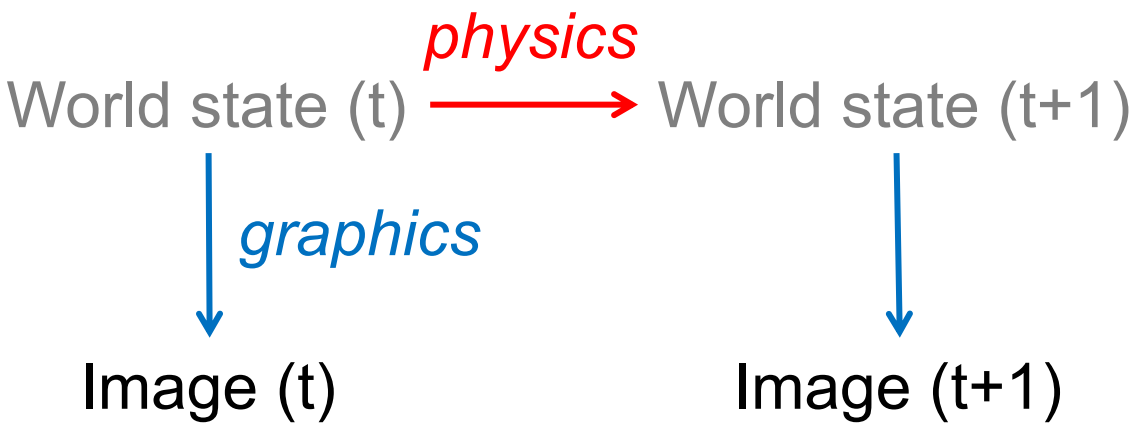
How far will they fall?



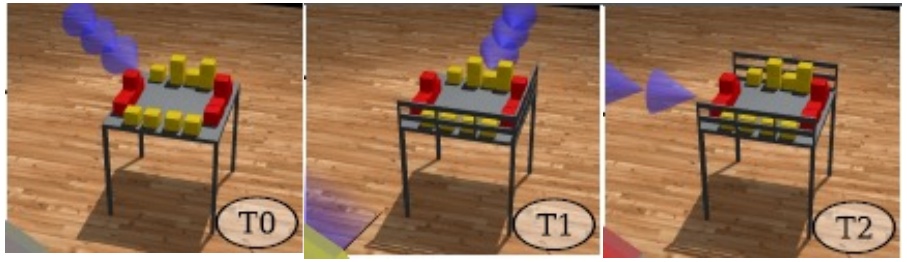
Is red or yellow heavier?

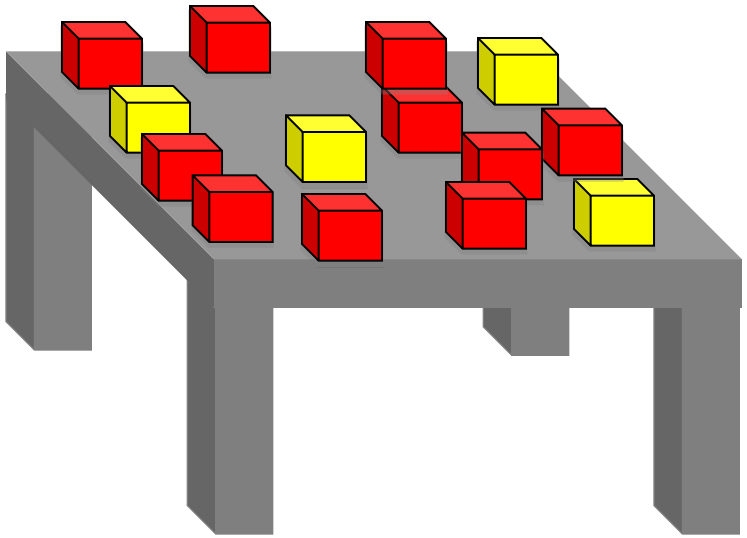


What if grey is much heavier than green?

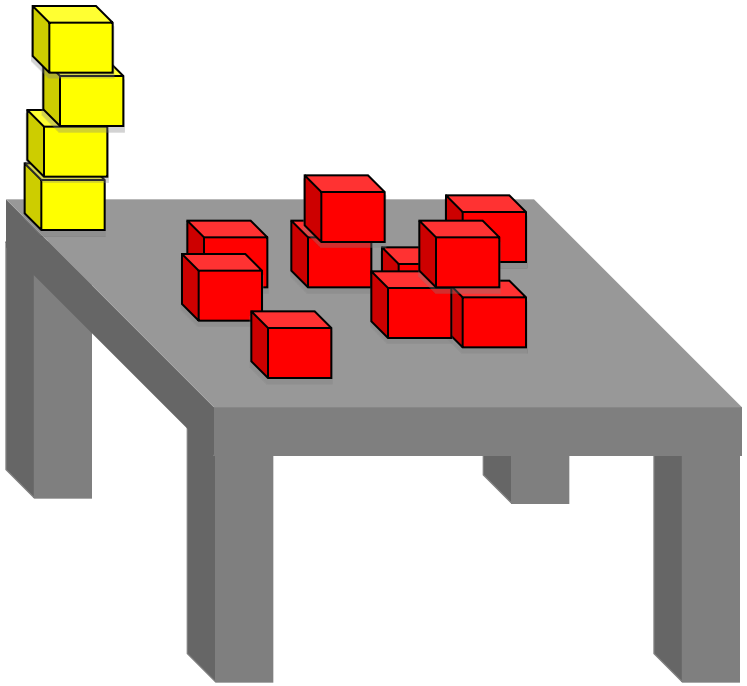


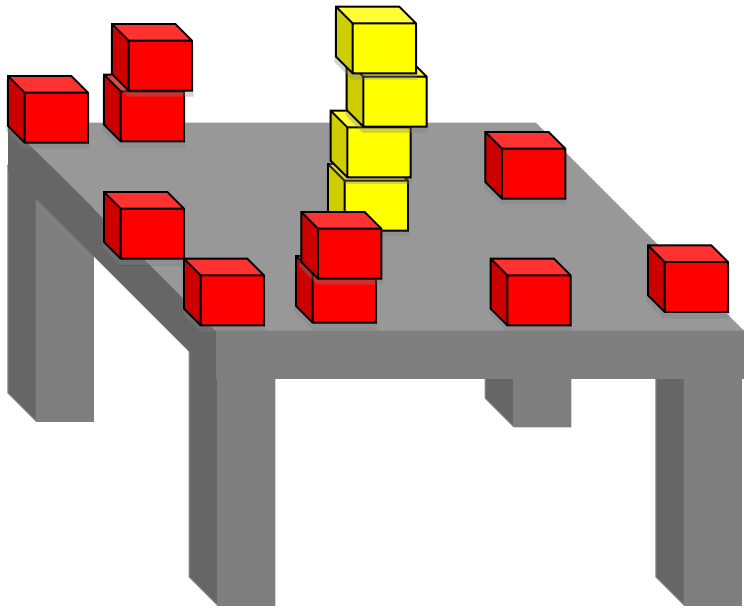
What will happen if you bump the table …?

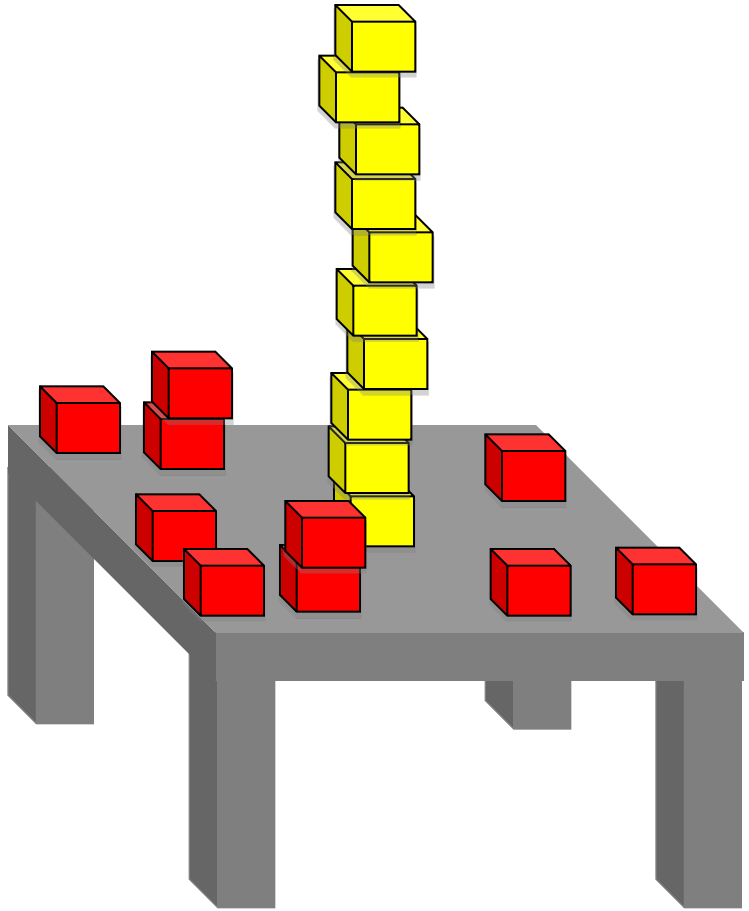


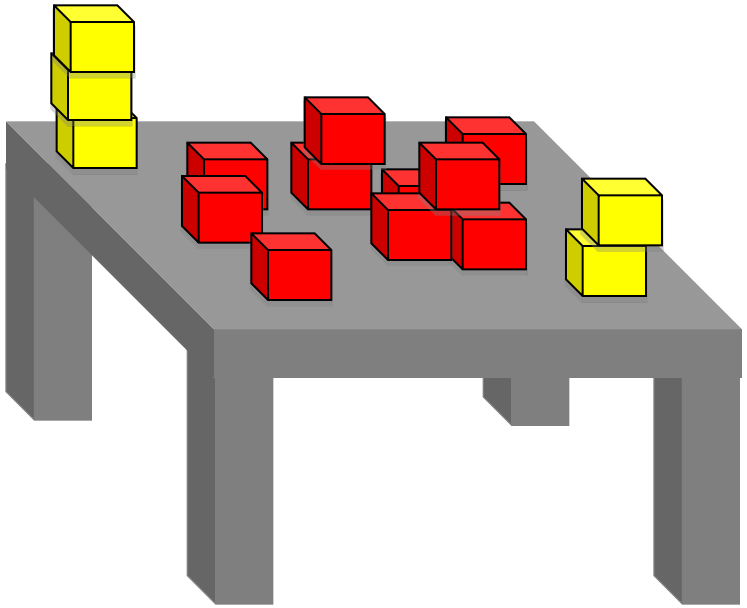


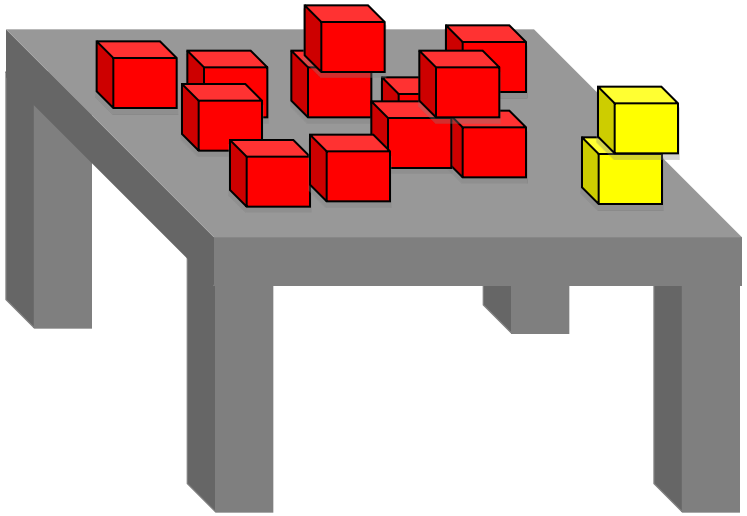
What if the table is bumped hard enough to knock some of the blocks onto the floor, is it more likely to be red blocks or yellow blocks?

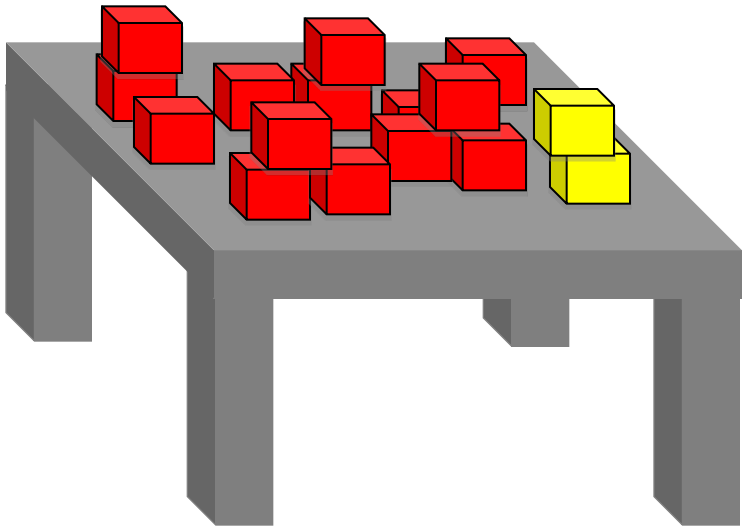


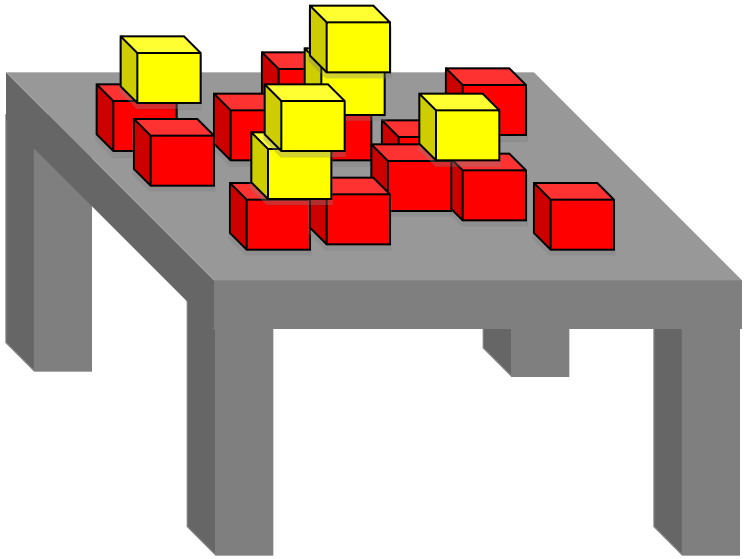




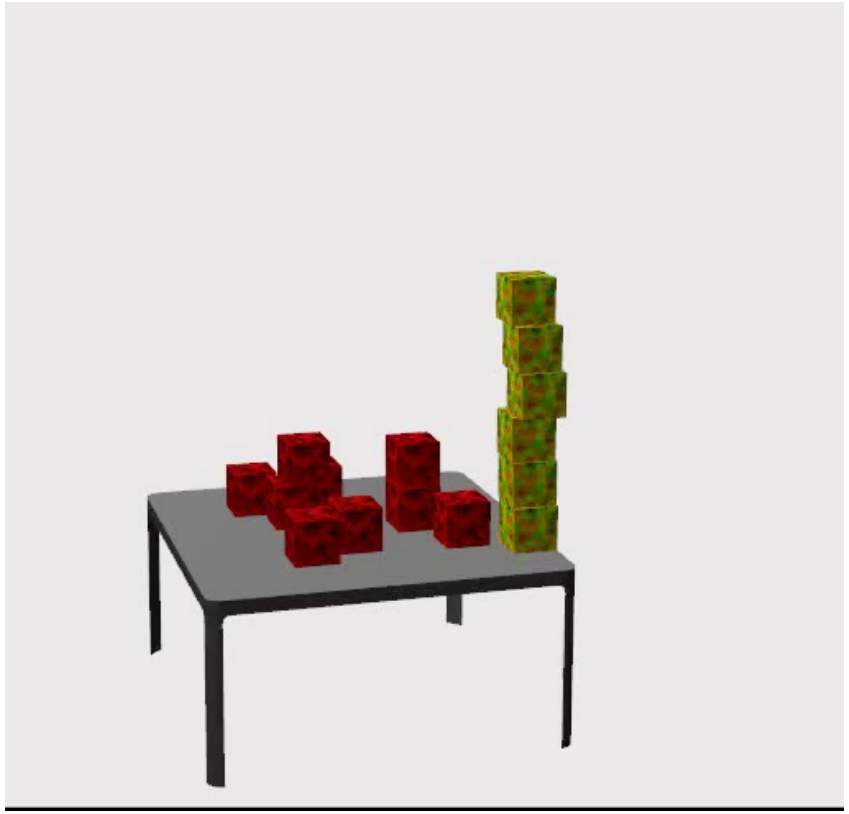




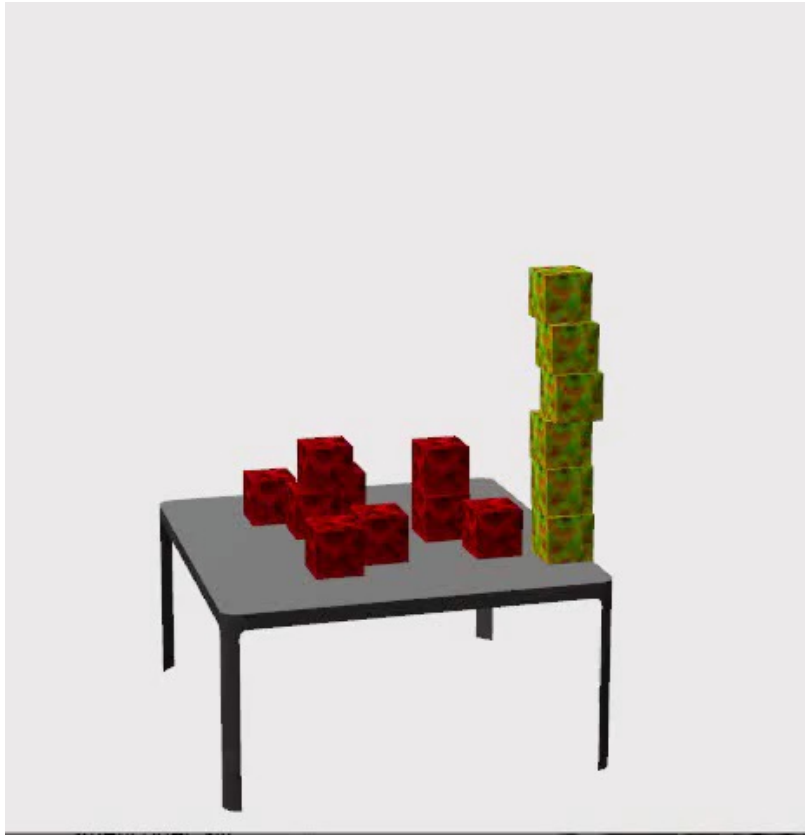




Prediction by simulation



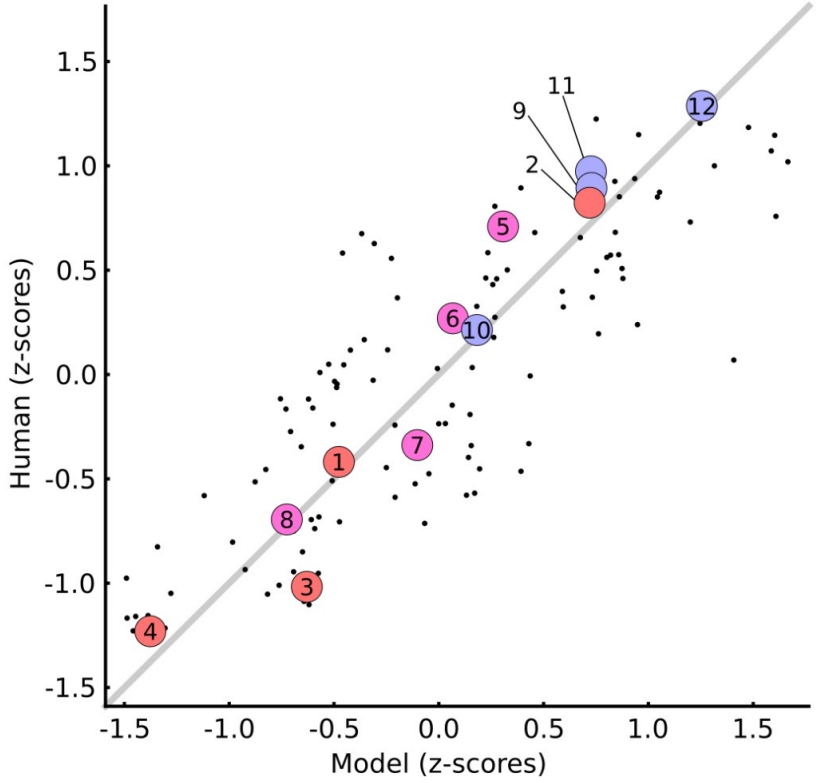
Smaller force



Larger force

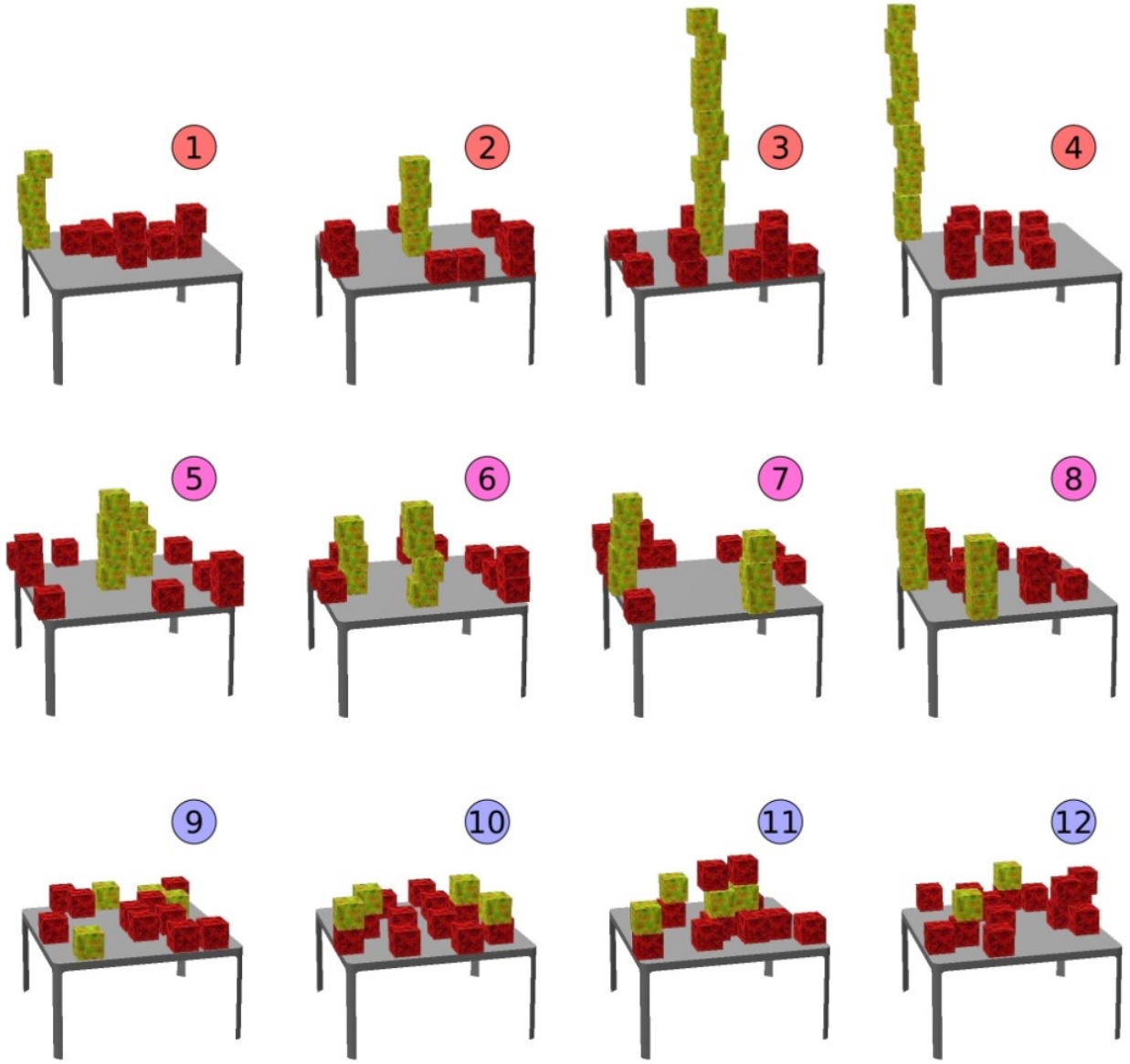
What will happen if…?

… you bump the table hard enough to knock some blocks onto the floor? Will you knock off more red, or yellow?



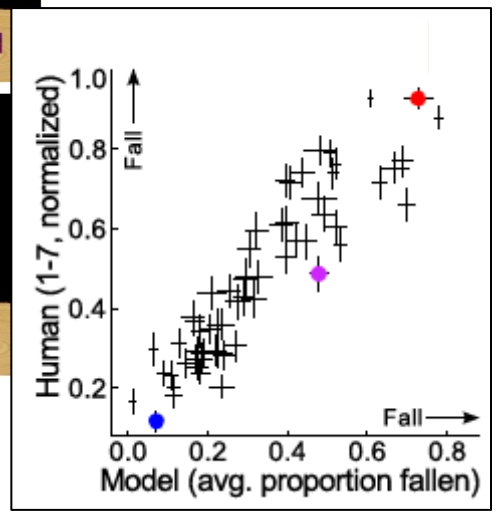
100% yellow

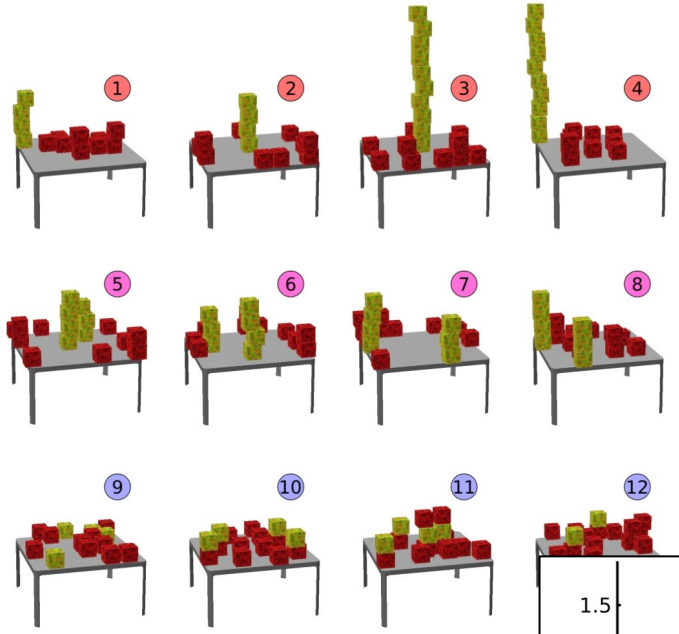
100% red



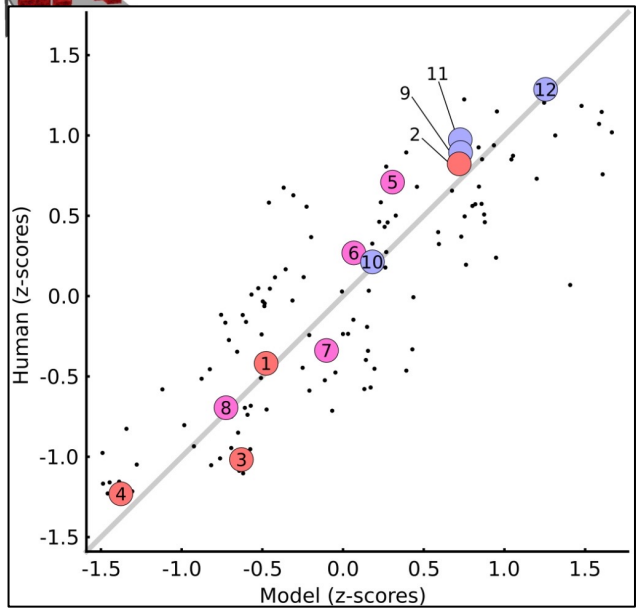
The intuitive physics engine

Will this stack of blocks fall?





What if you bump the table hard enough to knock some blocks onto the floor? Will you knock off more red, or yellow?



physics

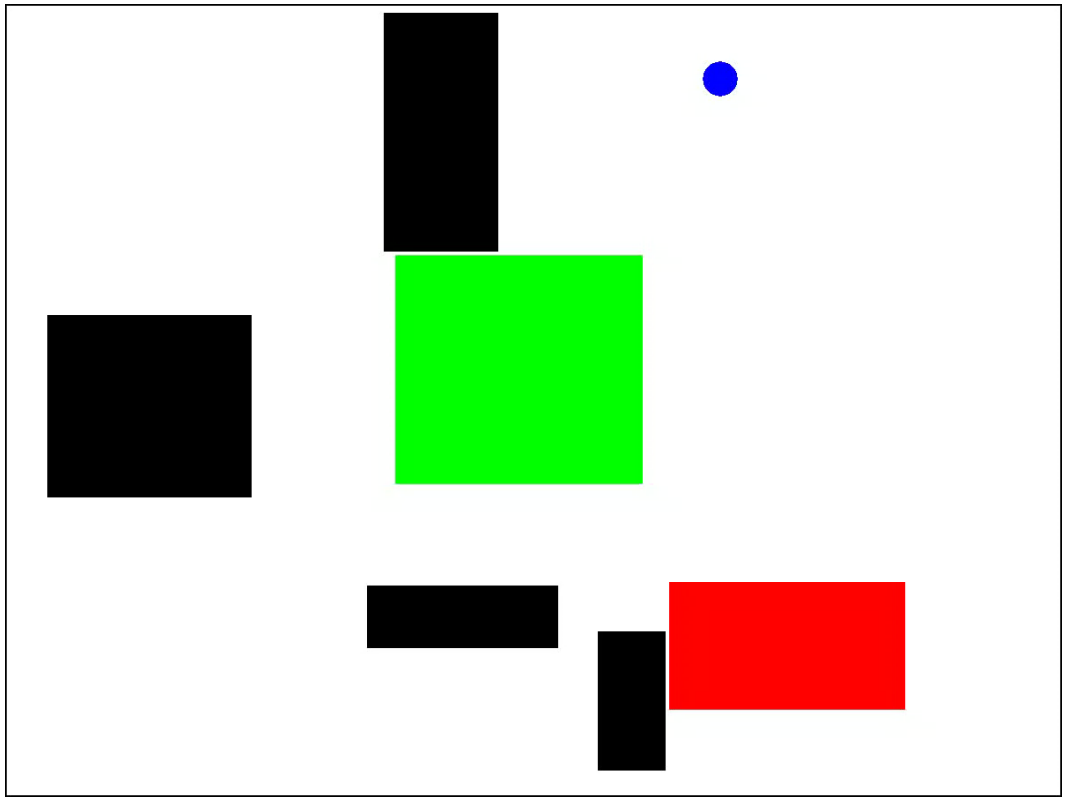
World state (t) → World state (t+1)

graphics

Image (t) → Image (t+1)

Dynamics in intuitive physics

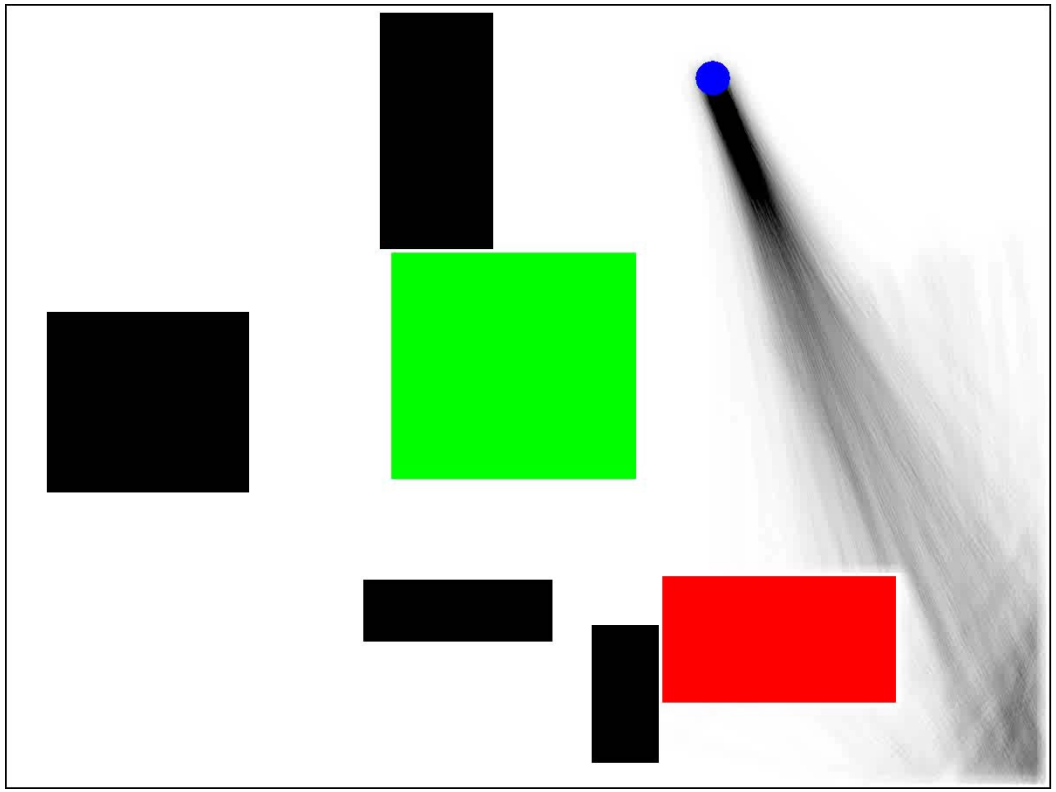
Ooooh or **Aaaah** ?

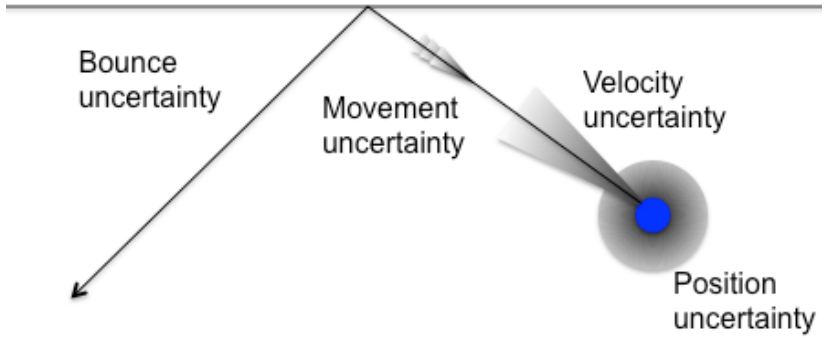


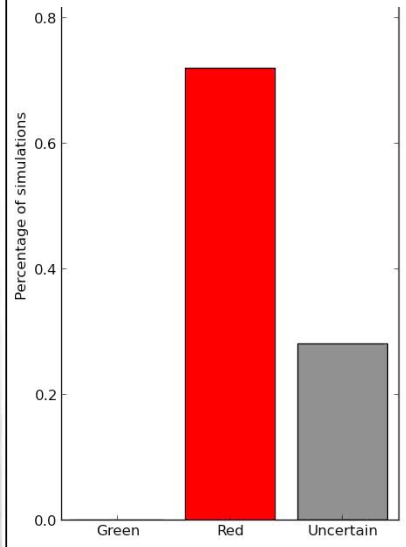
(Smith, Dechter, Tenenbaum, Vul, Cog Sci 2013)

Dynamics in intuitive physics

Ooooh or **Aaaah** ?

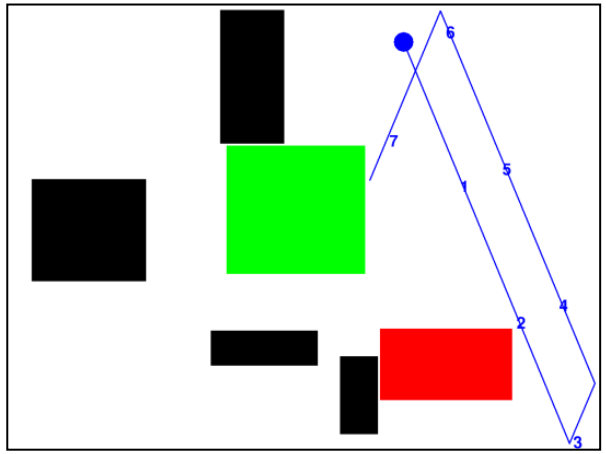


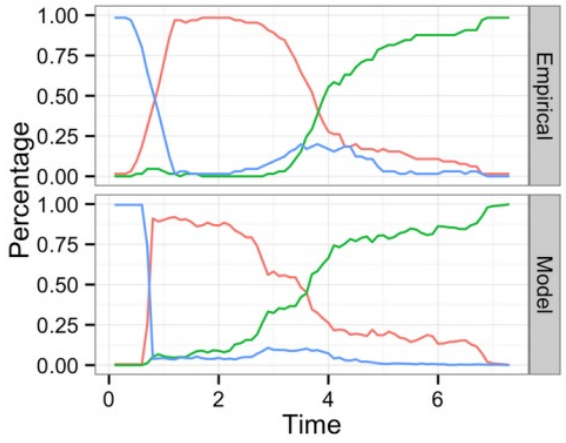


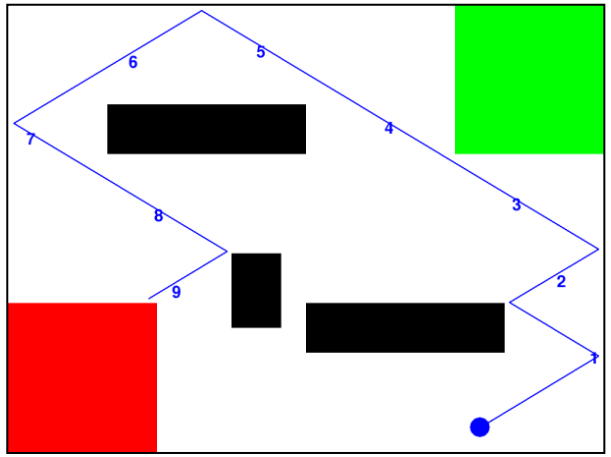


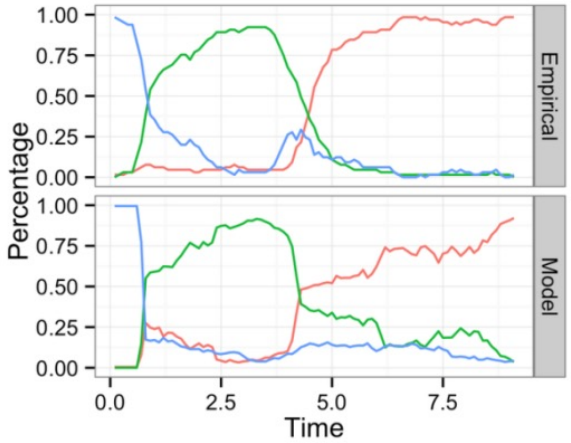
(Smith, Dechter, Tenenbaum, Vul, Cog Sci 2013)

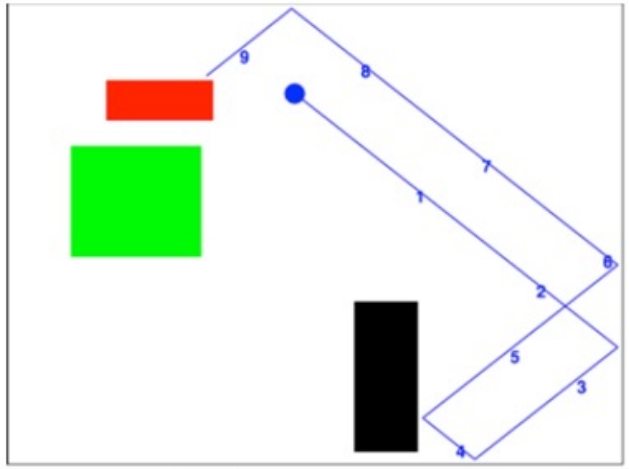
Dynamics in intuitive physics

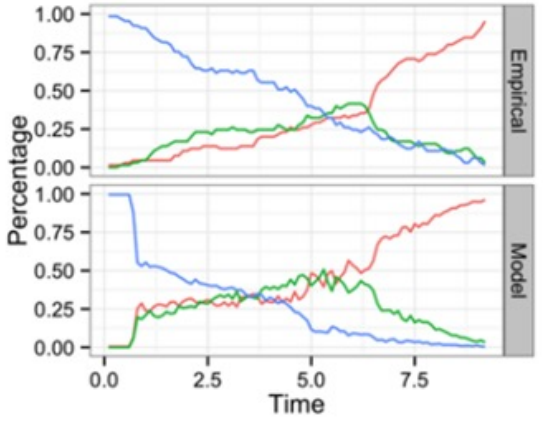


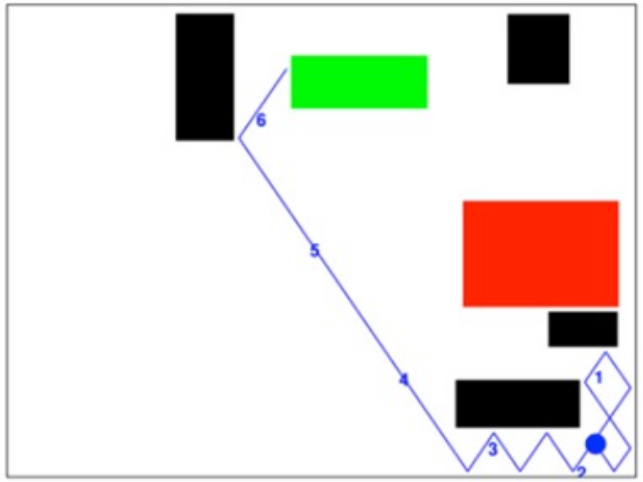


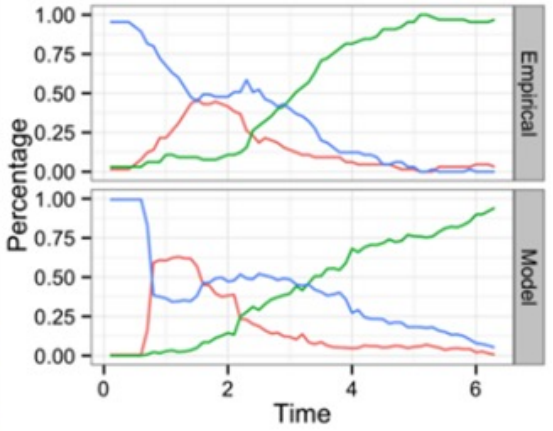








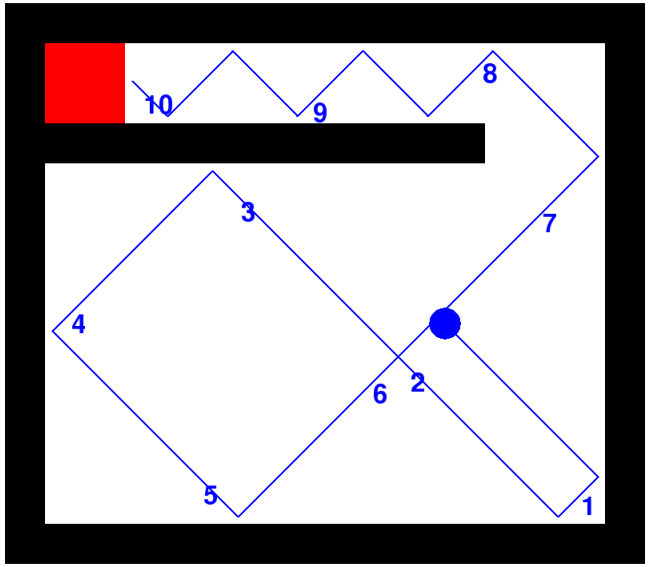


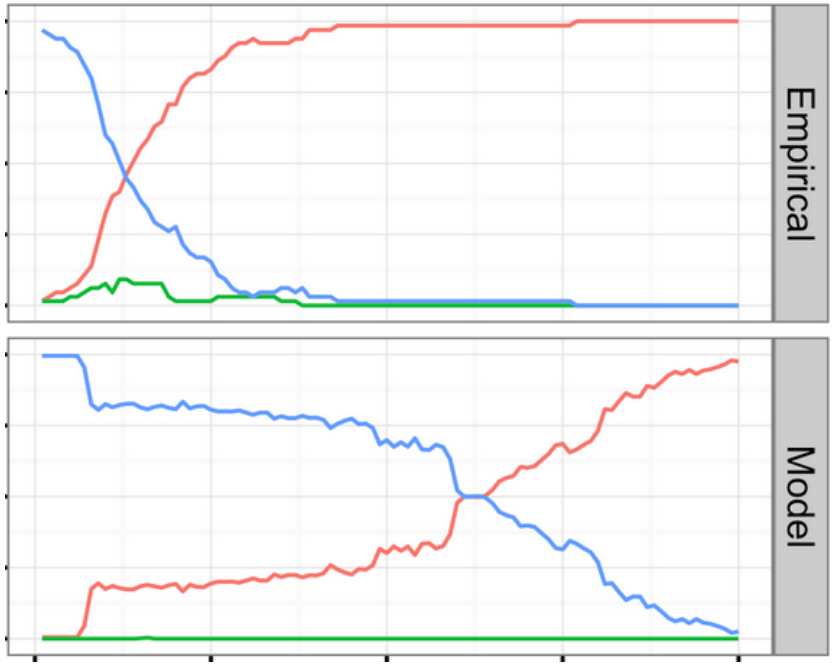


(Smith, Dechter, Tenenbaum, Vul, Cog Sci 2013)

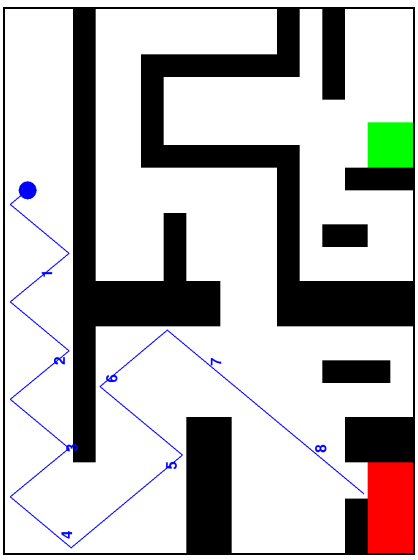
Qualitative reasoning too?

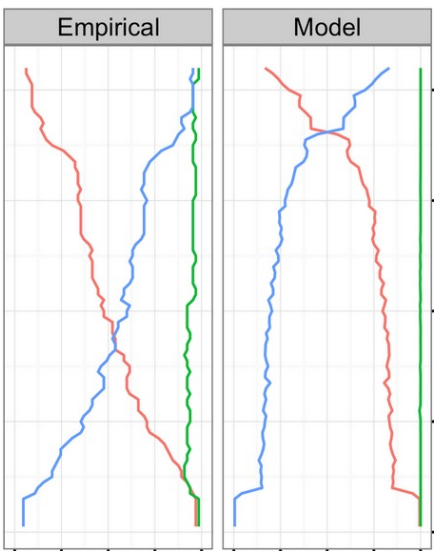
Or fast coarse simulation?

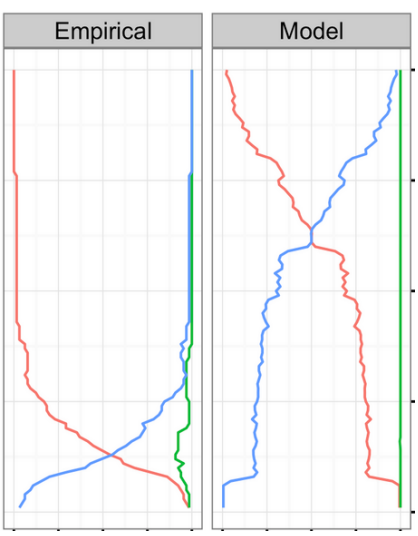


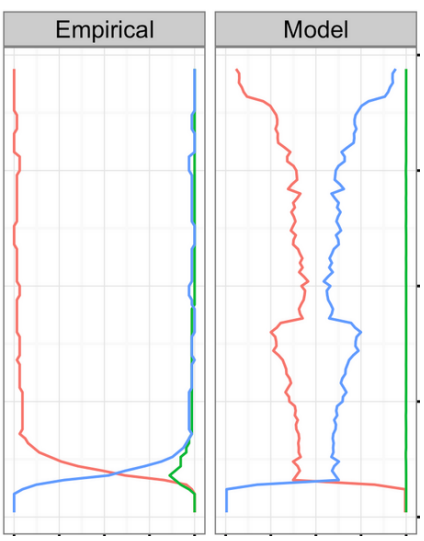


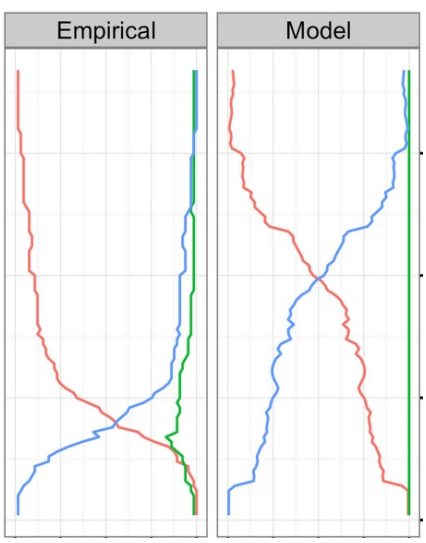
(Smith, Dechter, Tenenbaum, Vul, Cog Sci 2013)

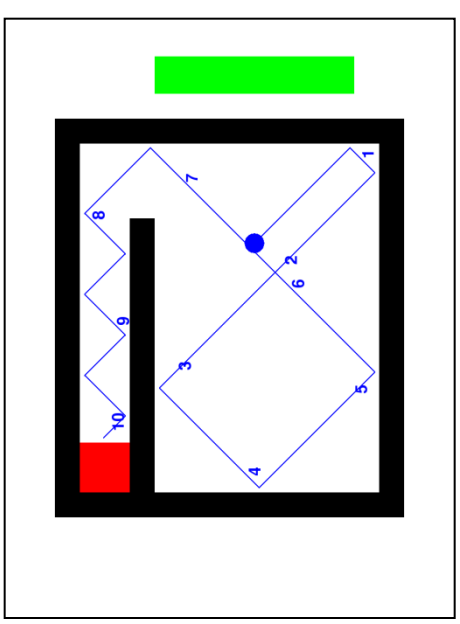

1
2
3
4
5
6
7
8

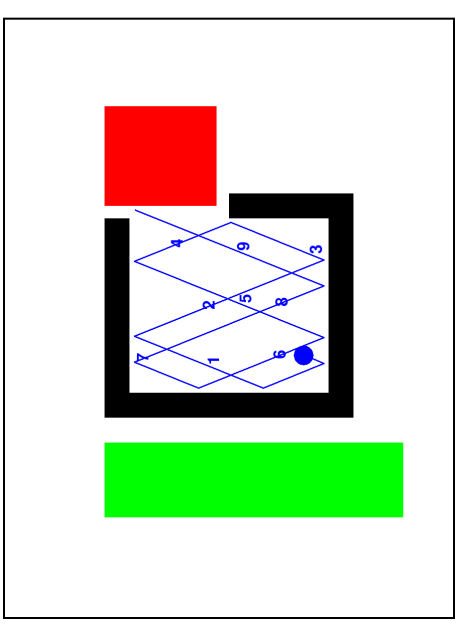

Empirical
Model

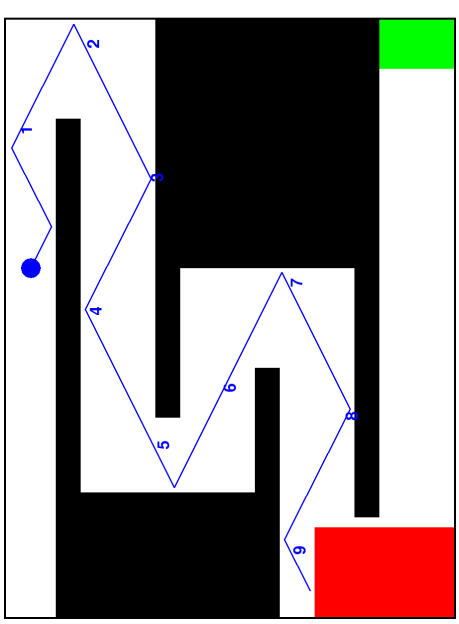

Empirical
Model


Empirical
Model

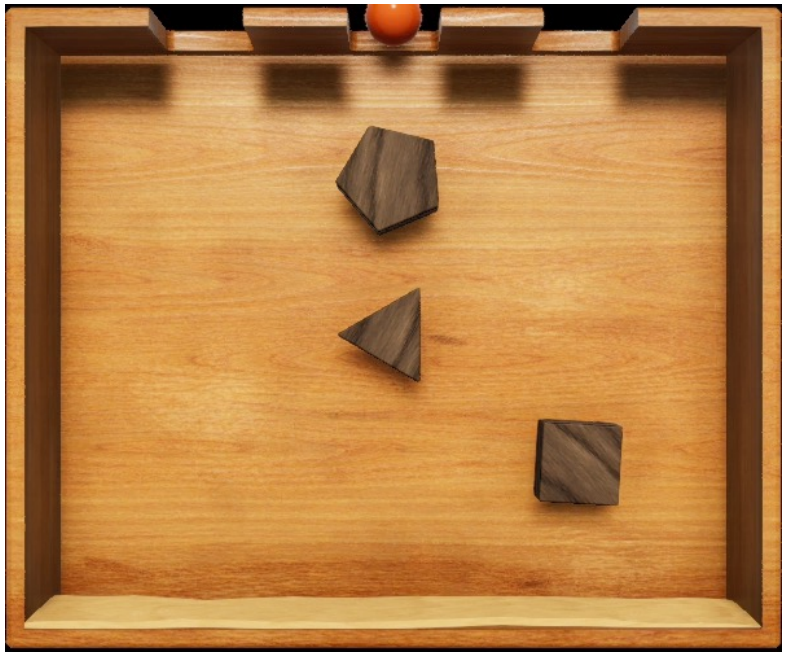

Empirical
Model


1
2
3
4
5
6
7
8
9
10


1
2
3
4
5
6
7
8
9

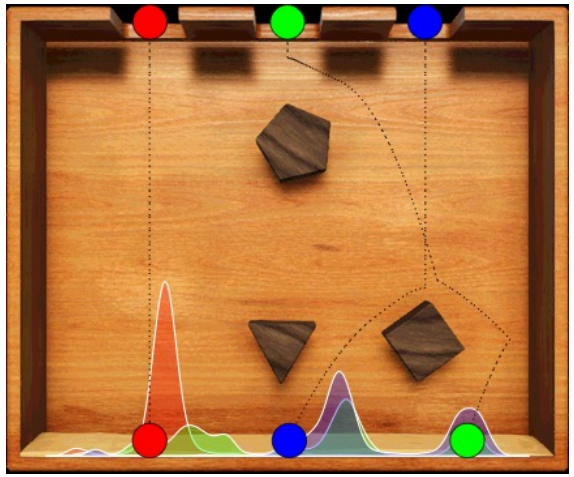

1
2
3
4
5
6
7
8
9

Multimodal understanding -- Plinko physics

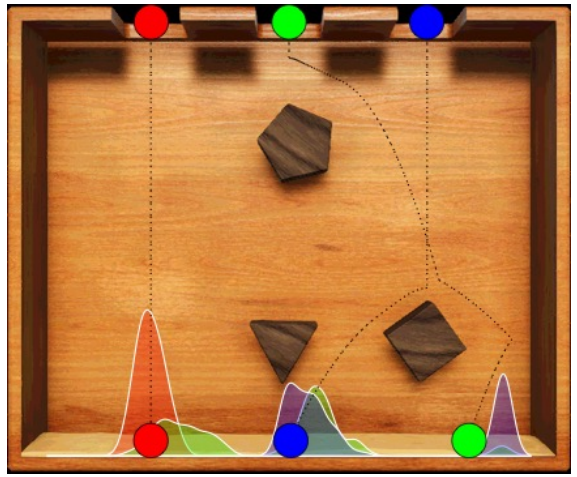


(Gerstenberg et al., 2021)

Prediction: where will the ball land?

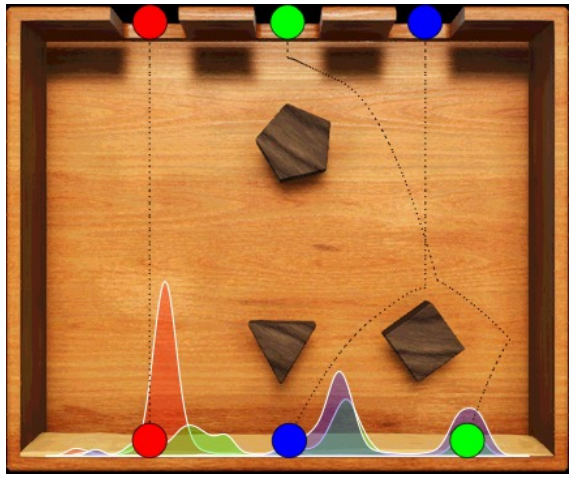


people

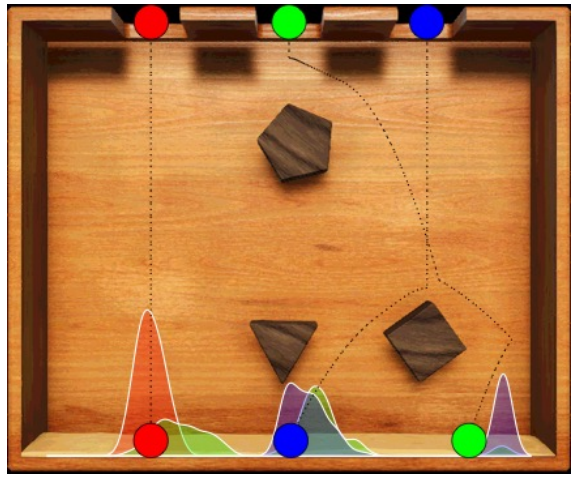


model

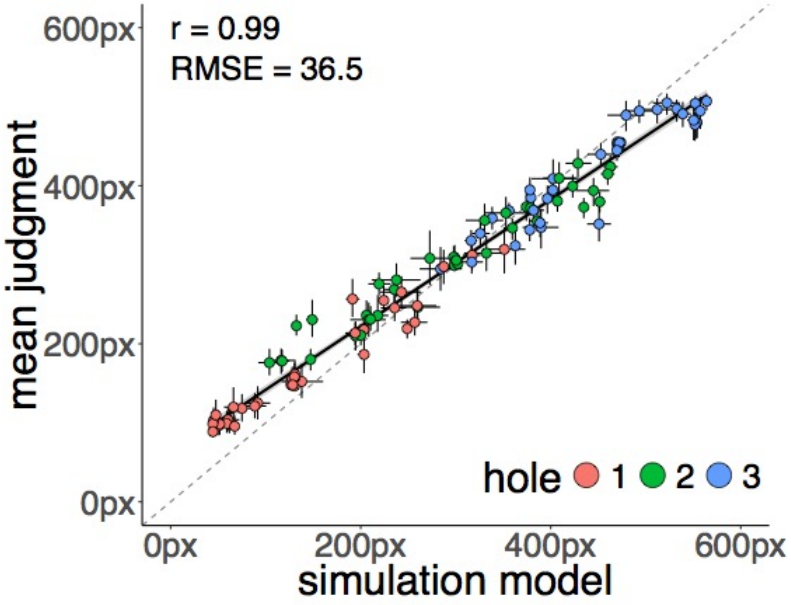
Prediction: where will the ball land?



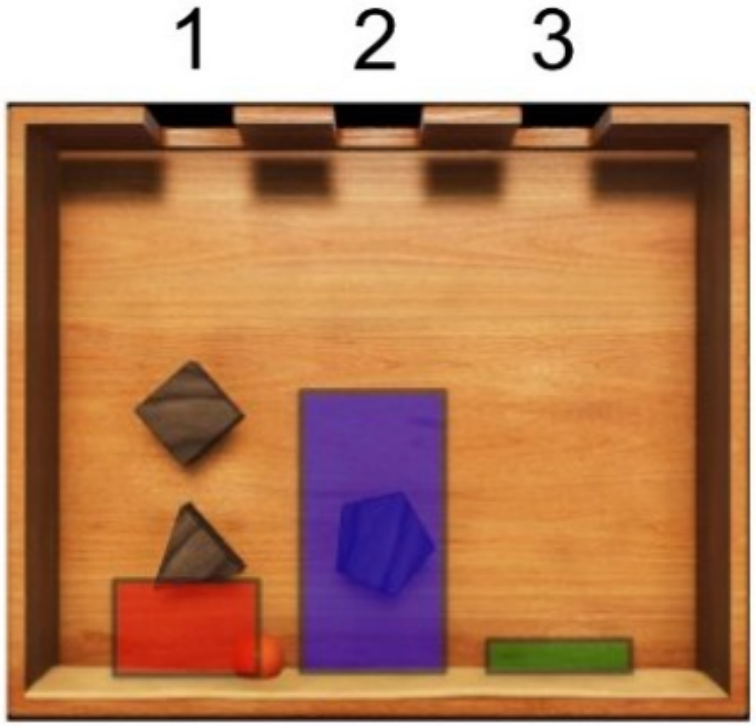
people

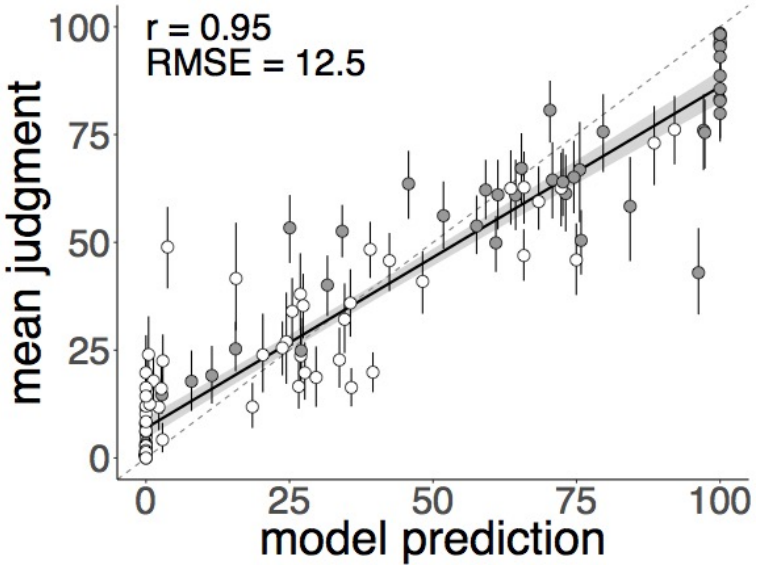


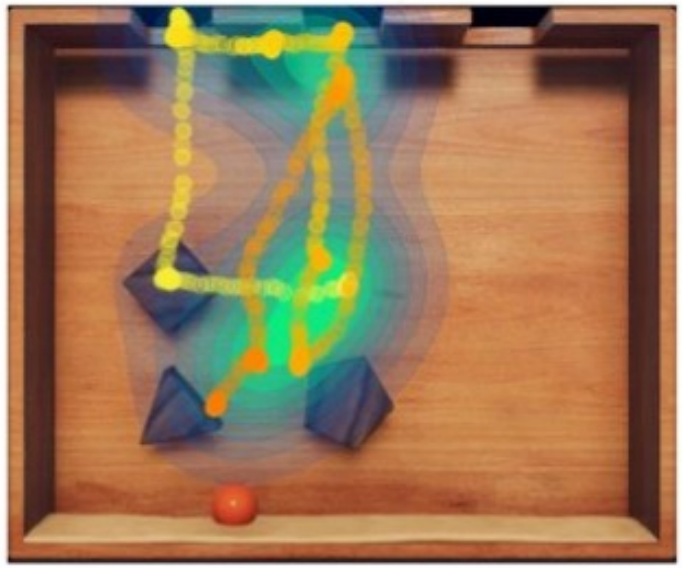
model



Inference: where was the ball dropped?



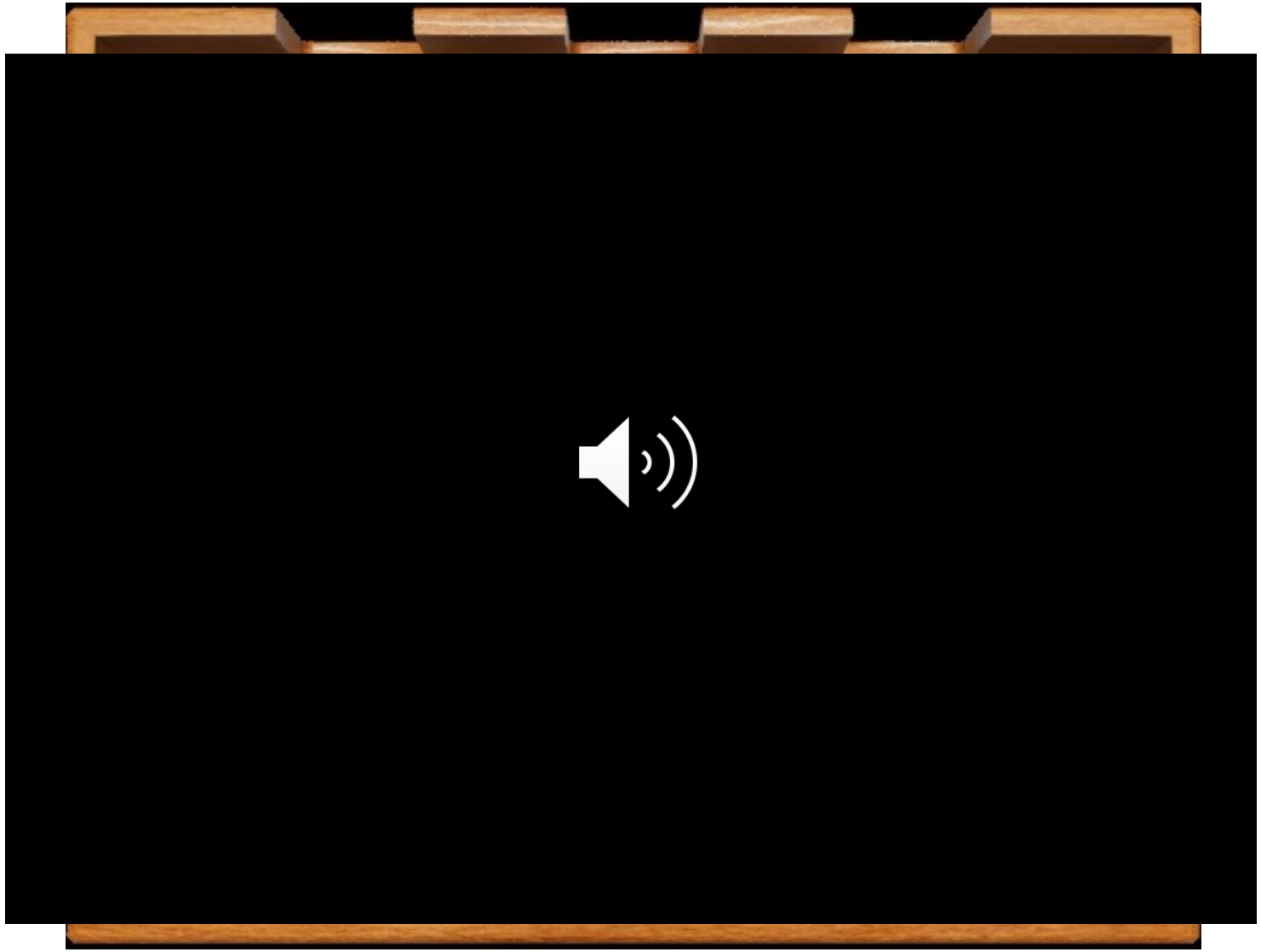




Eye tracking

Inference: where was the ball dropped?

Vision and sound



Inference: where was the ball dropped?

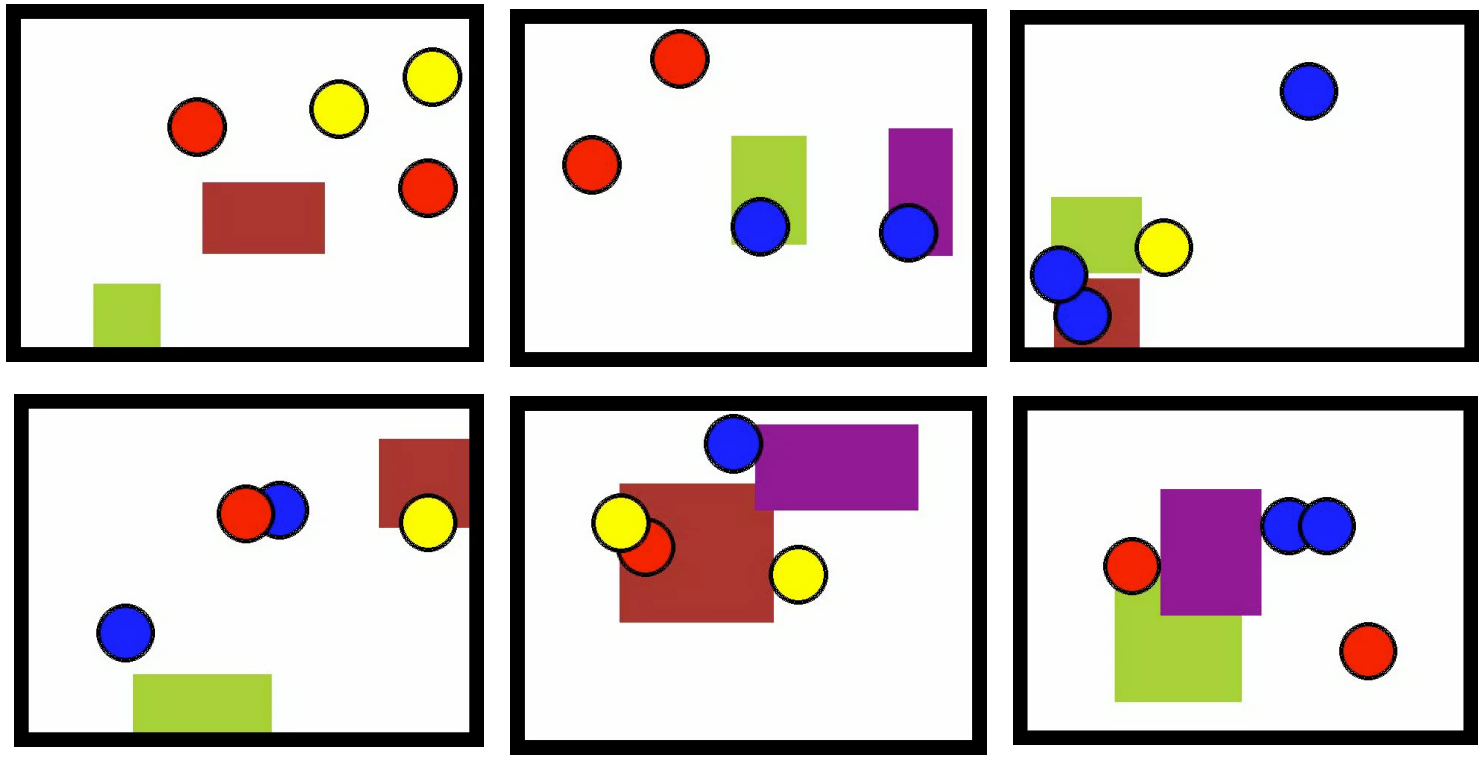
Vision and sound

Inference: where was the ball dropped?

Vision and sound

Rapid learning of physical properties and laws

Baby air hockey table

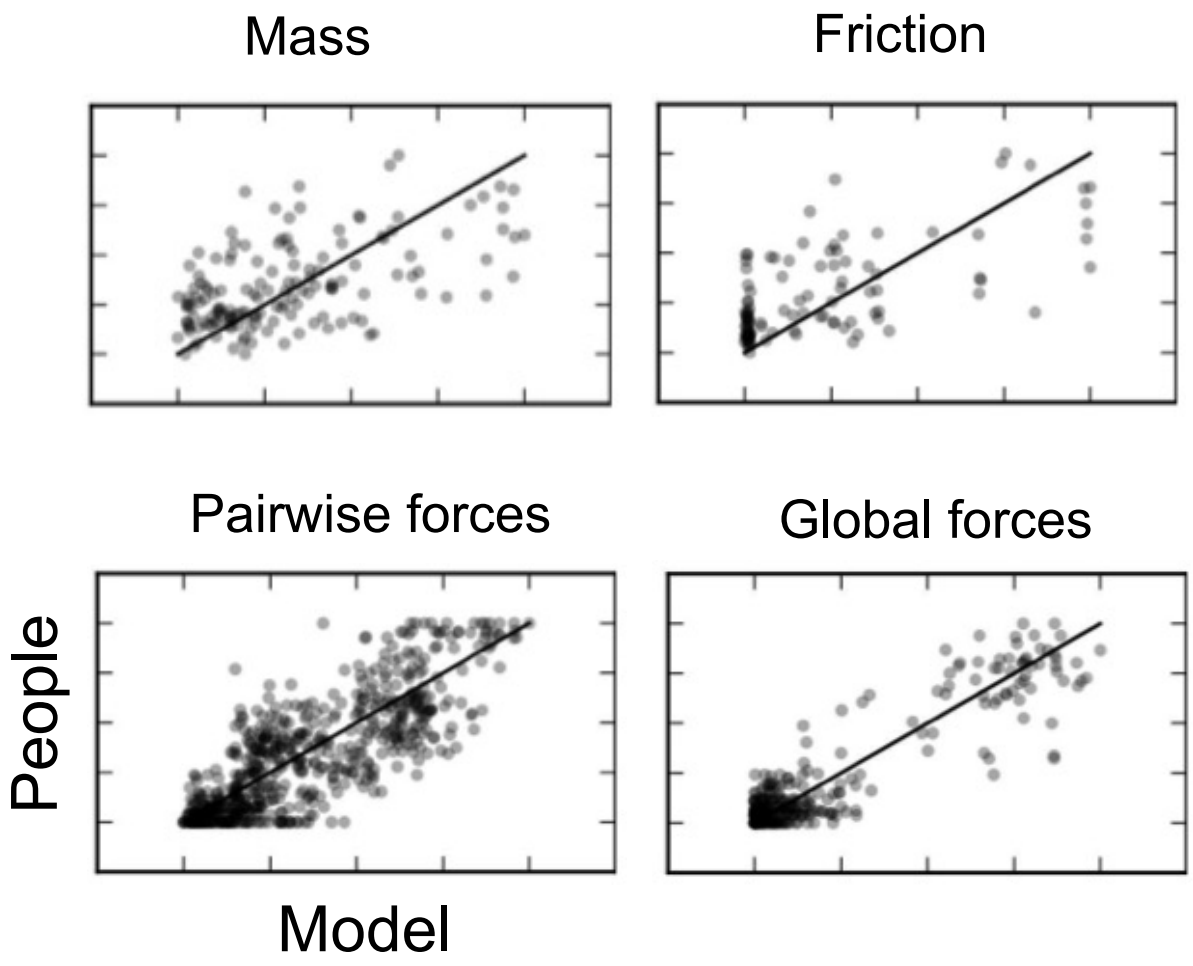


Inferring properties:
e.g., mass, charge, friction, elasticity, resistance…

Learning laws:
e.g., presence of forces and their shape, existence of hidden objects, kinds of substances …

(Ullman, Stuhlmuller, Goodman, Tenenbaum, *Cognitive Psychology* 2018)

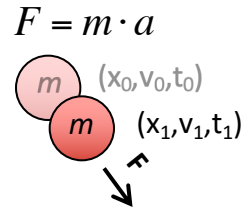
Rapid learning of physical properties and laws



Metatheory

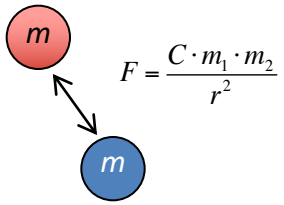
Objects

Inertial dynamics



Theories

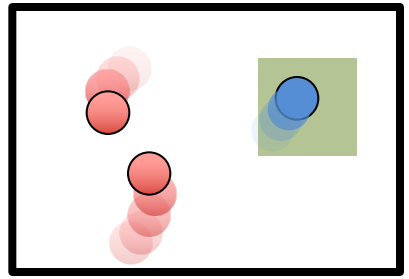
Different forces
Coupling
Global



Different masses

Different frictions

Scenes



(Ullman, Stuhlmuller, Goodman, Tenenbaum, *Cognitive Psychology* 2018)