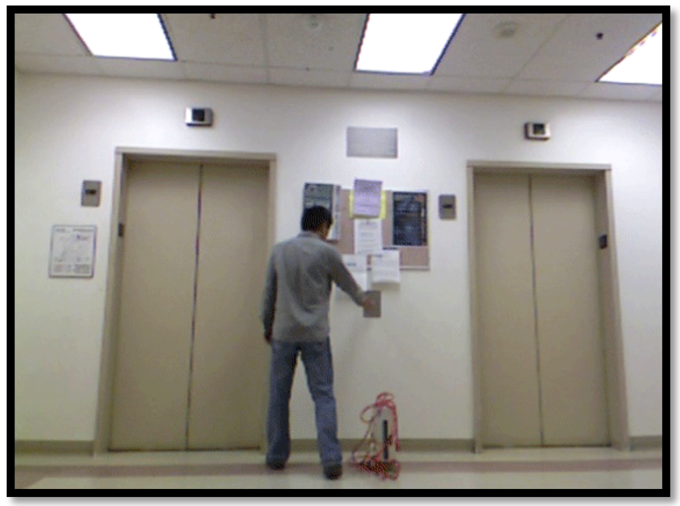
Learning and Inferring Perceptual Causality from Video

Amy Fire
UCLA

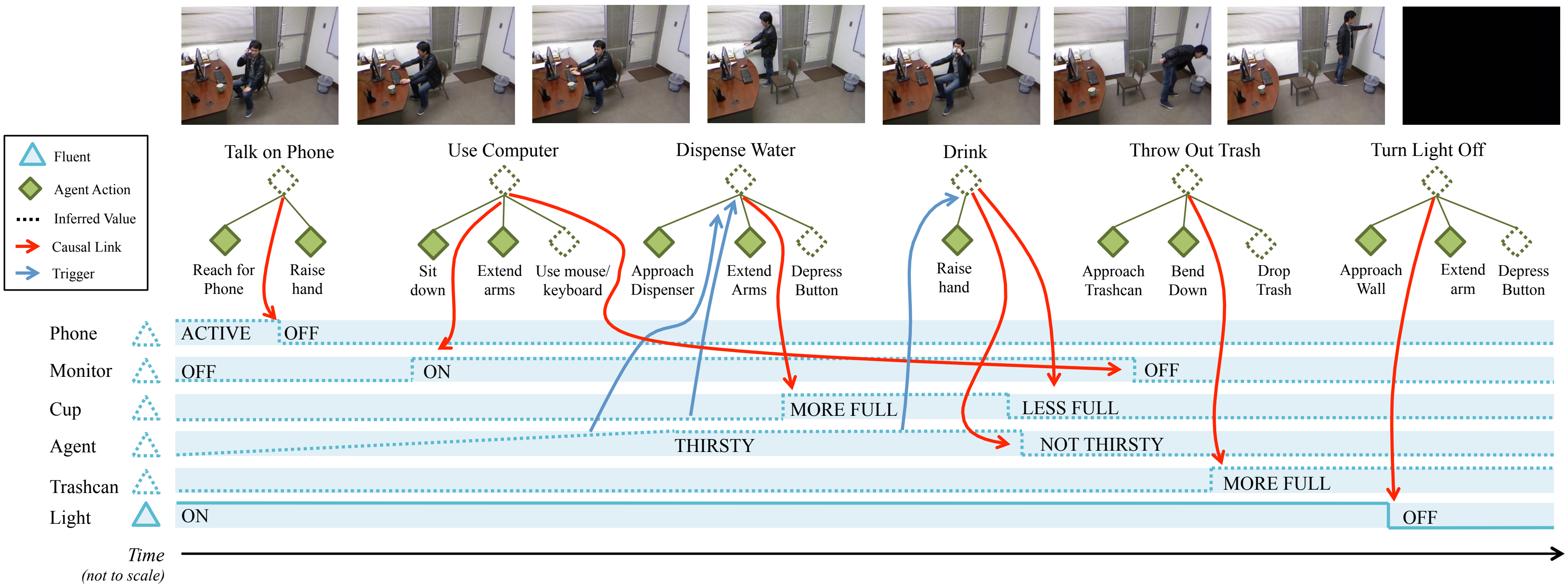
Overview

1. **Learn** the underlying causal relationships

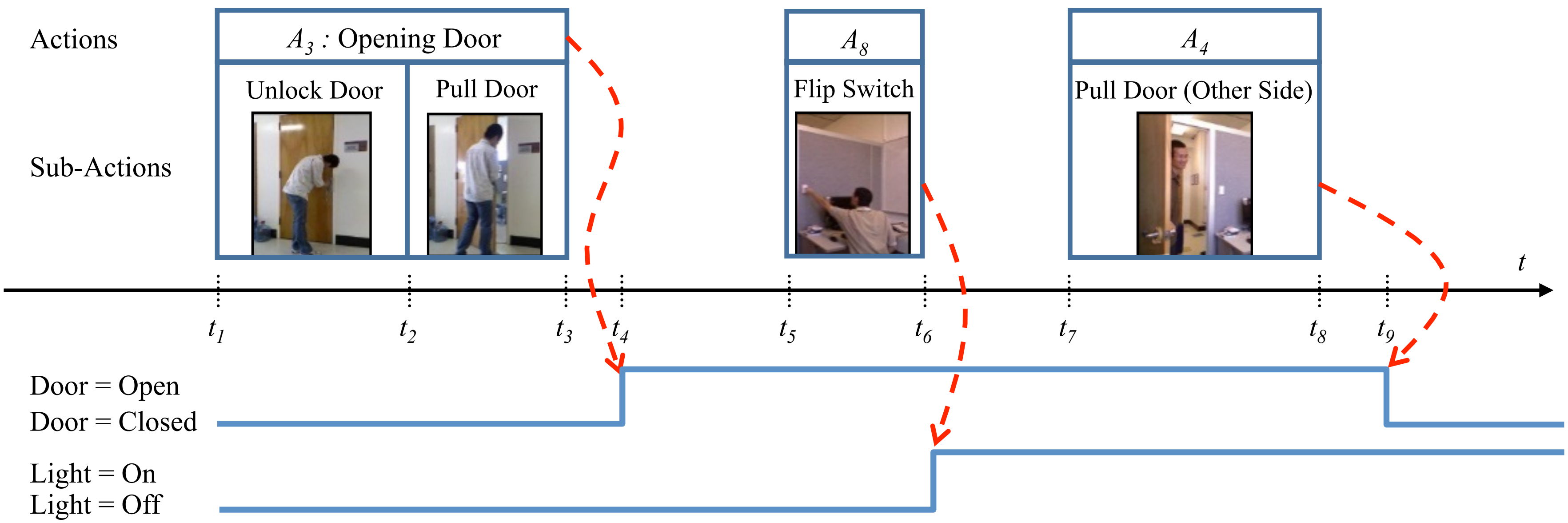


2. R**epresent** the causal relationships
3. I**nfer** instances from video.
   – Explain why (and why not) events happened.
   – Fill in gaps from ST explanations.

Example Causal Inference

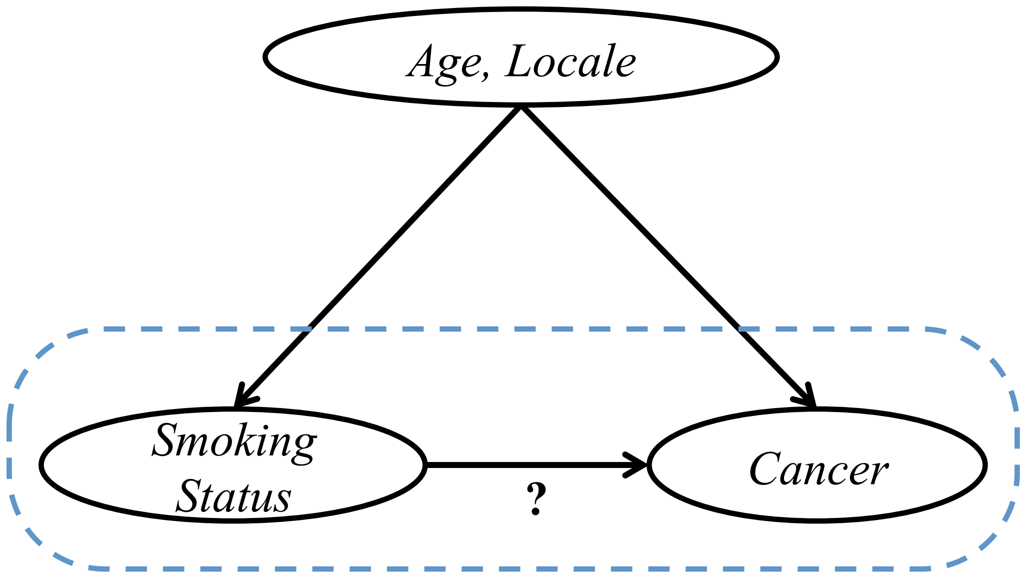


Causal Connections



WHAT HAS BEEN DONE

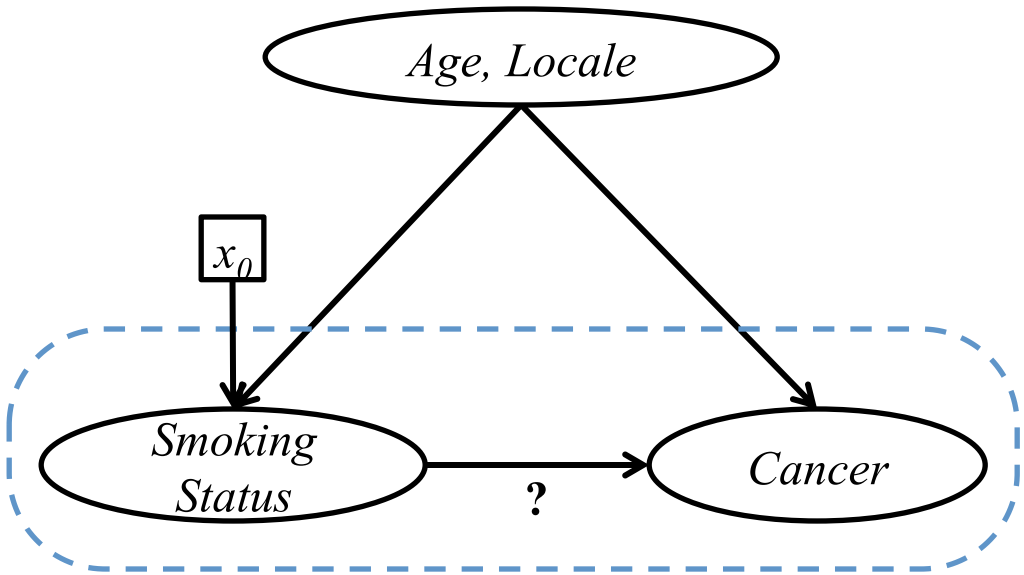
Causality



$$P\left(\text{Cancer} \mid \text{Smoking Status}\right)$$

- D. Rubin, “The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials,”

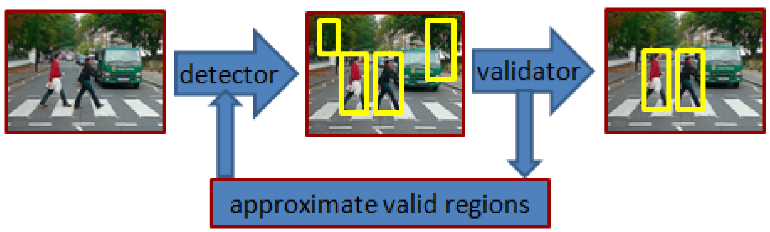
Causal Diagrams



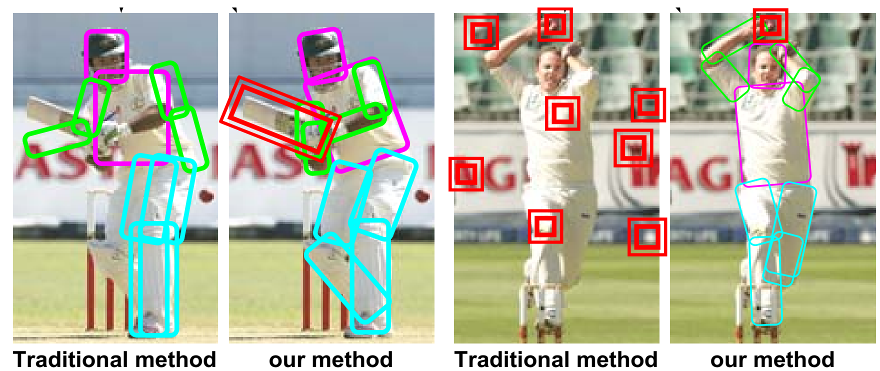
$$P\left(\text{Cancer} \mid do\left(\text{Smoking Status}\right)\right)$$

- Pearl – Causality 2000. Reasoning through constraint satisfaction.
- Mueller – Commonsense Reasoning 20. Reasoning through 1st order logic.

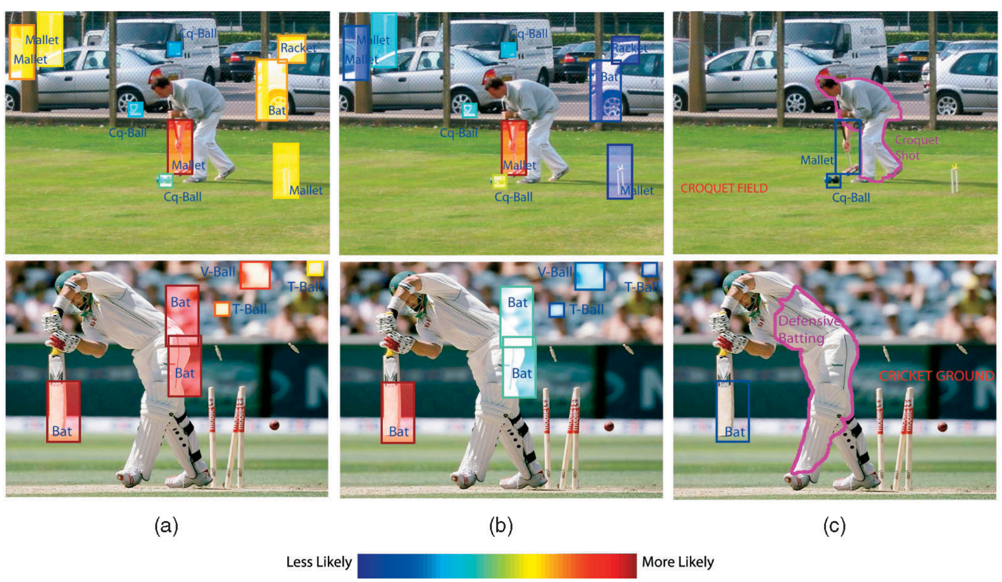
Context in Vision Research



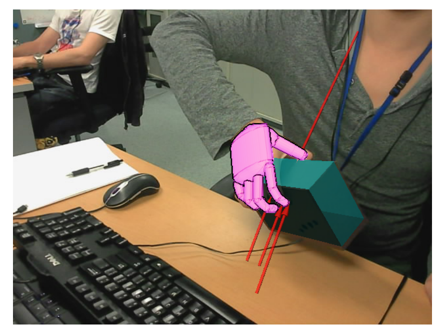
Saberian, et al. 2014. Using Context to
Improve Cascaded Pedestrian Detection



Yao, Fei-Fei 2010. Modeling Mutual Context of Object and
Human Pose in Human-Object Interaction Activities

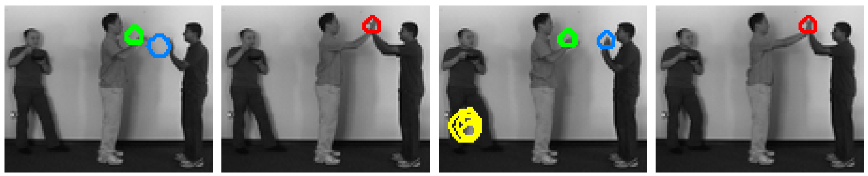


Gupta, et al. 2009. Observing Human-Object Interactions: Using
Spatial and Functional Compatibility for Recognition

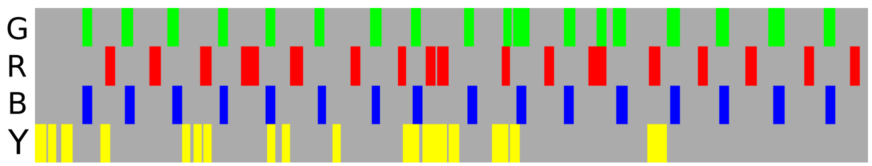


Pham et al., 2015 Towards force sensing
from vision: Observing hand-object
interactions to infer manipulation forces

Causality in Vision Research



(a) Frame 77 (b) Frame 85 (c) Frame 257 (d) Frame 266



(e) Co-occurrence Matrix of Point-Processes

Prabhakar, et al. 2010. Temporal Causality for the Analysis of Visual Events

CASE^{E}: A Hierarchical Event Representation for the Analysis of Videos

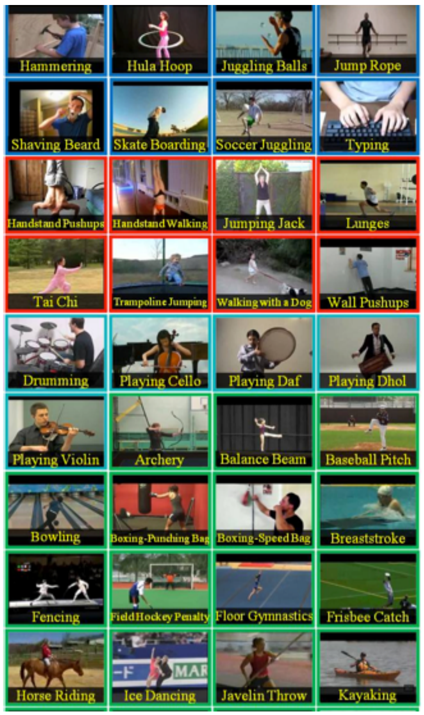
"Caravaggio pulled the chair therefore Michelangelo fell down."

[**PRED**: pull, **AG**: Caravaggio, **OBJ**: chair, **CAUSE**:
[**PRED**: fall, **D**: Michelangelo, **FAC**: down]]

Taylor, et al. 2015. Causal Video Object Segmentation from Persistence of Occlusions

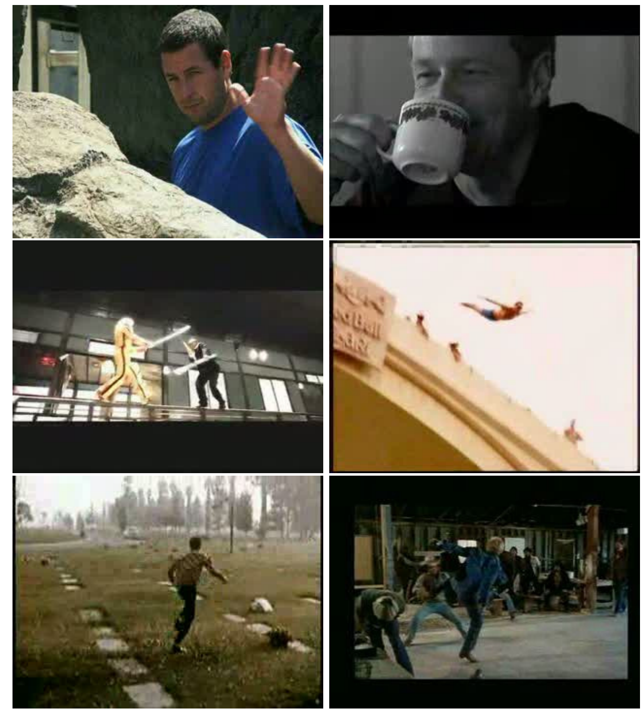
No Benchmarks for Causality in Vision

UCF-101



Olympic Sports

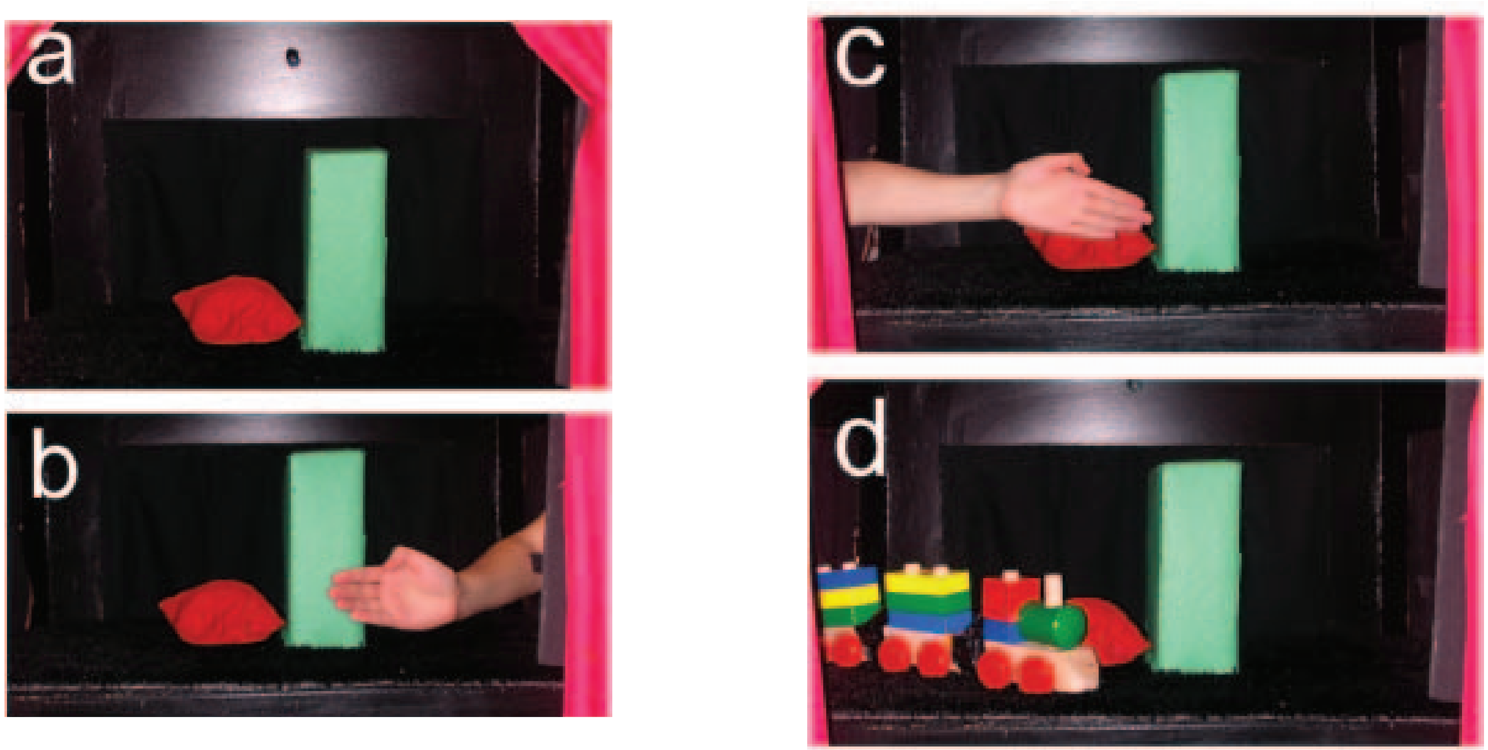
HMDB-51. Human Motion recognition



UT-Interaction Data. High-level human interactions

ACCESSING CAUSALITY IN VISION THROUGH COGNITIVE SCIENCE

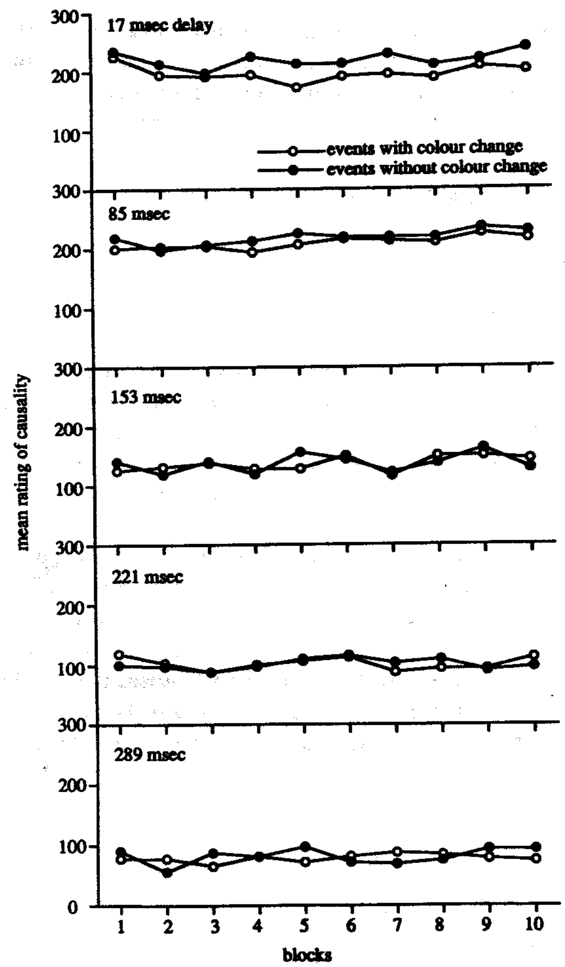
Perceptual Causality: Cognitive Science Agents Cause through Actions



Secret Agents : Inferences About Hidden Causes by 10- and 12-Month-Old Infants
R. Saxe, J.B. Tenenbaum and S. Carey

Heuristic 1: Action -> Effect

Perceptual Causality
Time between Cause and Effect is Short



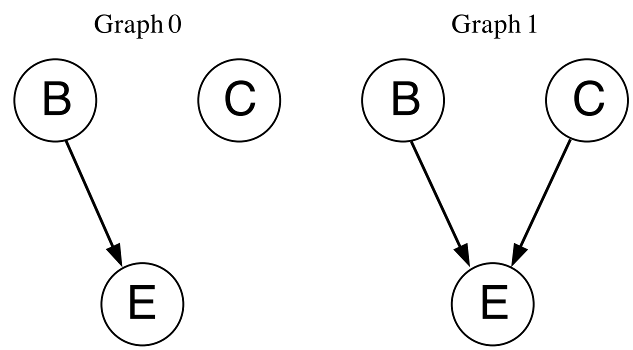
Schlottmann and Shanks. 1992. Evidence for a distance between judged and perceived causality

Heuristic 2: 0 < Time(Effect) – Time(Action) < δ

Perceptual Causality: Causality is Learned through Correlation

Contingency table representation used in elemental causal induction

	Effect present (e^+)	Effect absent (e^-)
Cause present (c^+)	$N(e^+, c^+)$	$N(e^-, c^+)$
Cause absent (c^-)	$N(e^+, c^-)$	$N(e^-, c^-)$



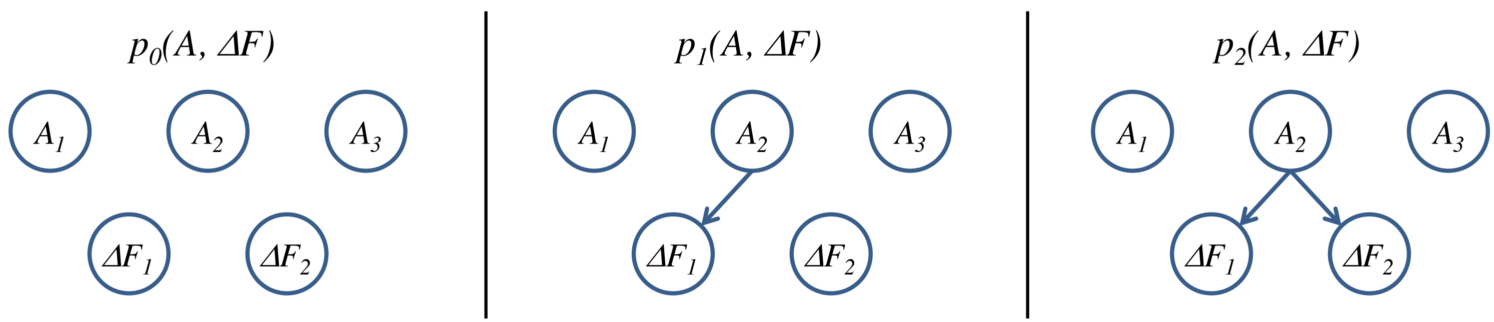
Griffiths and Tenenbaum. 2005. Structure and Strength in Causal Induction

Heuristic 3: Co-occurrence measures strength of perceptual causal relationships

LEARNING CAUSAL RELATIONSHIPS

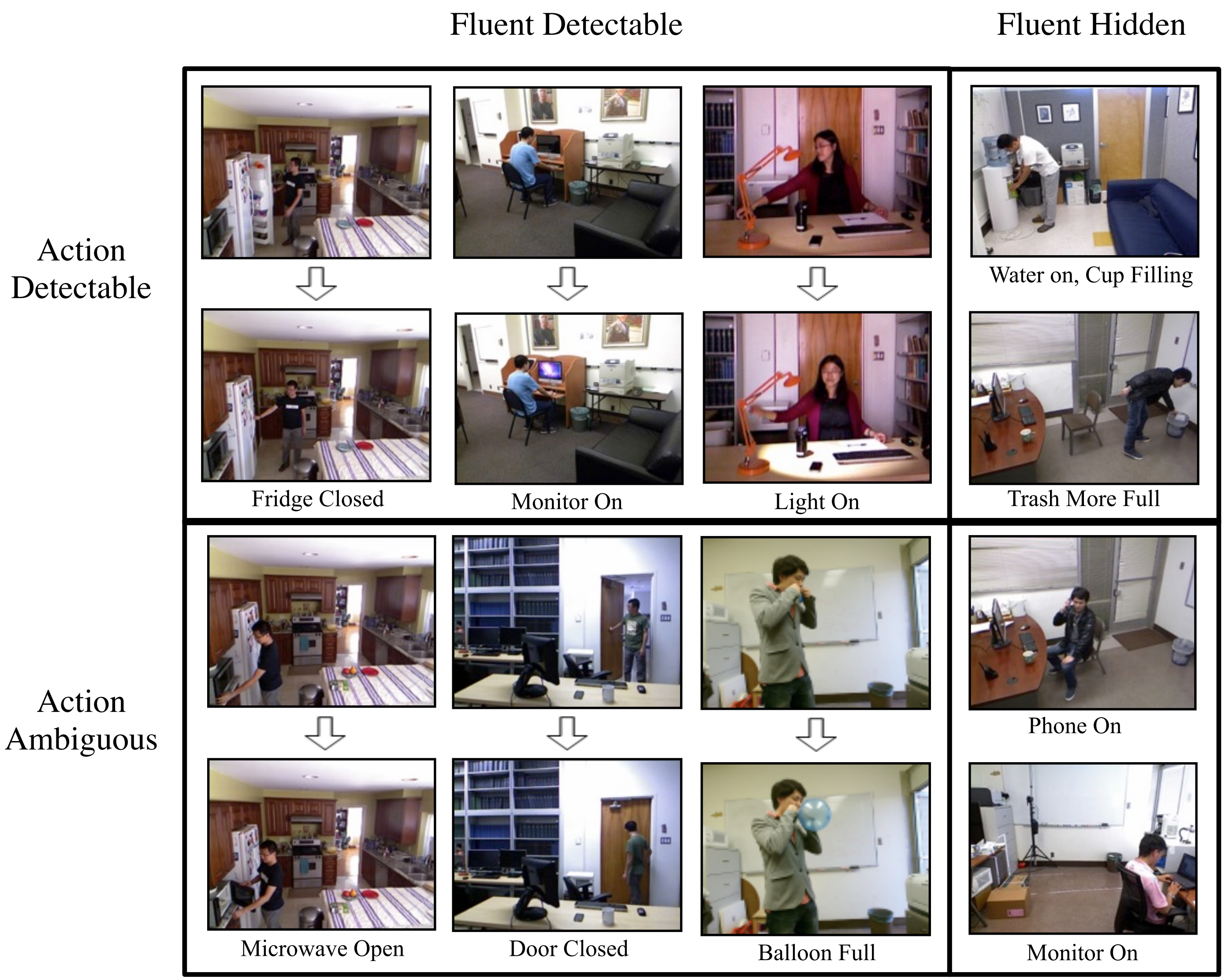
Assumptions for Learning

- Detections (and hierarchies) are sufficient
  - No hidden actions
  - No confounders
- Causal faithfulness
- The Heuristics
  - Heuristic 1: Action ➔ Effect
  - Heuristic 2: 0 < Time(Effect) – Time(Action) < δ
  - Heuristic 3: Co-occurrence measures strength of perceptual causal relationships



The Effects: Fluents
(Time-Varying Statuses)

$$\{\Delta F\}$$



Mueller – Commonsense Reasoning 2006

Causal Relations

$$\Omega_{CR} = \Omega_A \times \{\Delta F\}$$

$$\begin{Bmatrix} A_1 \\ \boxed{A_2} \\ A_3 \\ A_4 \\ A_5 \end{Bmatrix} \times$$

	F(t+1) = Open	***F(t+1) = Closed***
F(t) = Open	$O \to O$	$O \to C$
F(t) = Closed	$\boxed{C \to O}$	$C \to C$

Causal Relations

$$\Omega_{CR} = \Omega_A \times \{\Delta F\}$$

$$\begin{Bmatrix} A_1 \\ \boxed{A_2} \\ A_3 \\ A_4 \\ A_5 \end{Bmatrix} \times$$

	F(t+1) = Open	***F(t+1) = Closed***
F(t) = Open	$O \to O$	$O \to C$
F(t) = Closed	$\boxed{C \to O}$	$C \to C$

cr :

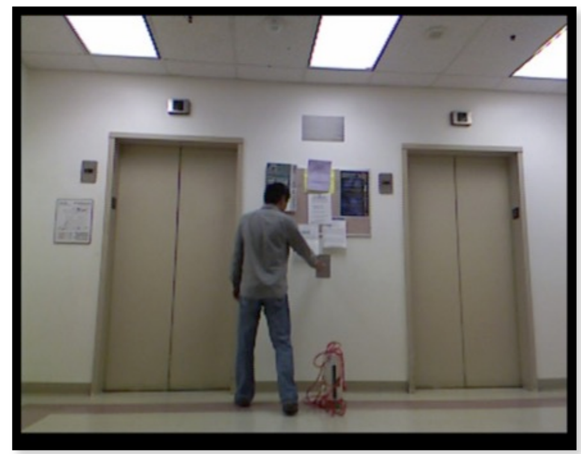
	¬Action	Action
¬Effect	c_0	c_1
Effect	c_2	c_3

$$\mathbf{cr} = (c_0, c_1, c_2, c_3)$$

Statistics on Relations: Histogram

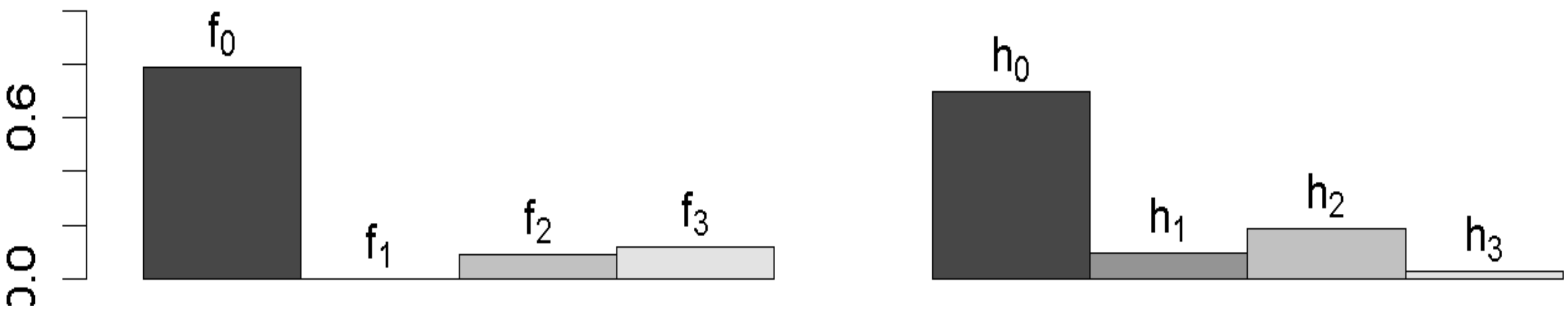
$$RF(\mathbf{cr}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{cr}(\mathbf{v}_i)$$

ΔF	A	cr	Current Model	Observed Data
0	0	$\mathbf{cr}_0$	h_0	f_0
0	1	$\mathbf{cr}_1$	h_1	f_1
1	0	$\mathbf{cr}_2$	h_2	f_2
1	1	$\mathbf{cr}_3$	h_3	f_3



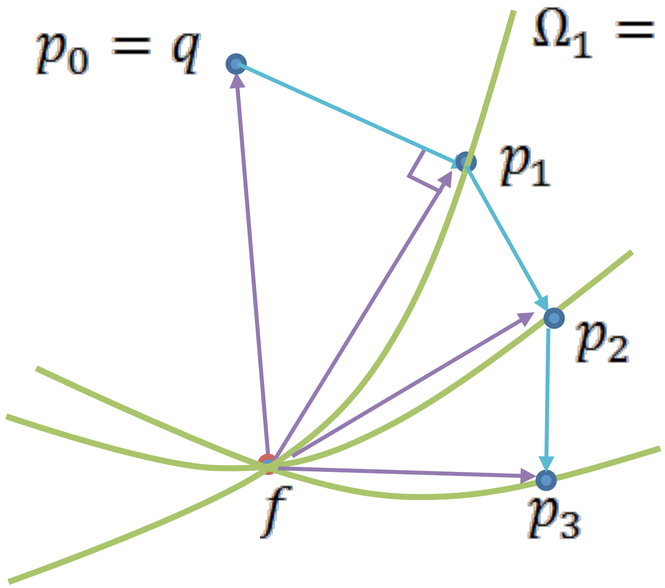
Causing vs. Non-Causing Actions

Causing Action



Non-Causing Action

Information projection



$$\Omega_1 = \{\, p\colon\ E_p[\mathrm{h}_1(\mathrm{I})] = E_f[\mathrm{h}_1(\mathrm{I})] \,\}$$

$$\Omega_2 = \{\, p\colon\ E_p[\mathrm{h}_2(\mathrm{I})] = E_f[\mathrm{h}_2(\mathrm{I})] \,\}$$

$$KL(f \mid p) = KL(f \mid p_+) + KL(p_+ \mid p)$$

DellaPietra, DellaPietra,Lafferty, 97
Zhu, Wu, Mumford, 97

Adding a Causal Relation to the Model

- Model Pursuit

$$p_0 \to p_1 \to \ldots \to \boxed{p \to p_+} \to \ldots \to p_k \approx f$$

(On ST-AOG)

$$p_+(pg) = \frac{1}{z_+} p(pg) \exp\left(-\langle \lambda_+, \mathbf{cr}_+ \rangle\right)$$

- Part 1: Find parameters
  - Model formed by min $KL\ (p_+ \,||\, p)$, matching statistics

$$E_{p_+}\left(\mathbf{cr}_+\right) = E_f\left(\mathbf{cr}_+\right)$$

- Part 2: Pursue **cr**. max $KL\ (p_+ \,||\, p)$

DellaPietra, DellaPietra,Lafferty, 97
Zhu, Wu, Mumford, 97

Proposition 1: Model Parameters

- Suppose $\mathbf{f}$ denotes the frequencies of $\mathbf{cr}_+$ as observed, and $\mathbf{h}$ denotes the expected frequencies from the probability model, p.
- If $p_+ = \min KL\ (p_+ \,||\, p)$, then p_+ is of the form

$$p_+(pg) = \frac{1}{z_+} p(pg) \exp\left(-\langle \lambda_+, \mathbf{cr}_+ \rangle\right)$$

and

$$\lambda_{+,i} = \log\left(\frac{h_i}{h_0} \cdot \frac{f_0}{f_i}\right)$$

for $i = 0,\ \ldots,\ 3$.

Prop 2: Selecting a Causal Relation

- Suppose $\mathbf{cr}$, $\mathbf{f}$, $\mathbf{h}$, p, and p_+ are as denoted before.

- Suppose further that $\mathbf{cr}_+$ is selected to provide the maximum reduction in KL-divergence,

$$\mathbf{cr}_+ = \underset{\mathbf{cr}}{\operatorname{argmax}} \left(KL\left(f \parallel p \right) - KL\left(f \parallel p_+ \right) \right)$$

- Then

$$\mathbf{cr}_+ = \underset{\mathbf{cr}}{\operatorname{argmax}}\, KL\left(p_+ \parallel p \right) = \underset{\mathbf{cr}}{\operatorname{argmax}}\, KL\left(\mathbf{f} \parallel \mathbf{h} \right)$$

Selection from ST-AOG

- Suppose parent **Or-node** A has children $A_1, \ldots, A_n$, with A_i as the true cause. Then

$$\mathrm{KL}\left(\mathbf{f}_A \parallel \mathbf{h}_A\right) \leq \mathrm{KL}\left(\mathbf{f}_{A_i} \parallel \mathbf{h}_{A_i}\right)$$

and

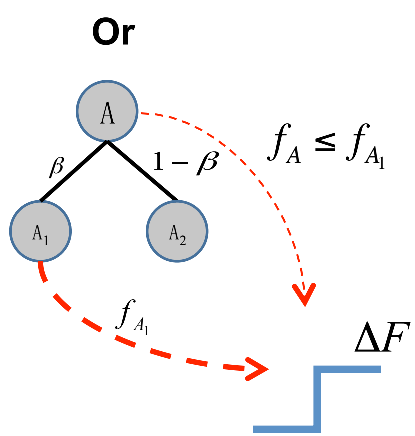
$$\mathbf{cr}_{A_i} = \underset{\mathbf{cr}_A, \mathbf{cr}_{A_1}, \ldots, \mathbf{cr}_{A_n}}{\operatorname{argmax}} KL\left(\mathbf{f} \parallel \mathbf{h}\right)$$

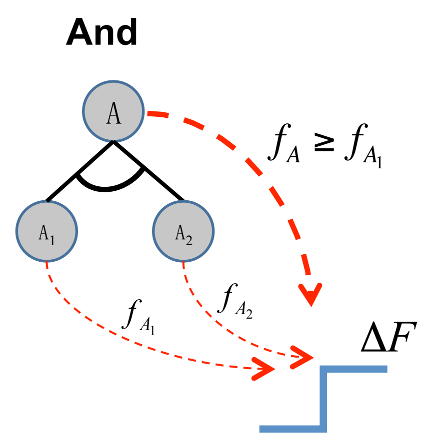
- Suppose parent **And-node** A has children $A_1, \ldots, A_n$, with A_i as the true cause. Then

$$\mathrm{KL}\left(\mathbf{f}_A \parallel \mathbf{h}_A\right) \geq \mathrm{KL}\left(\mathbf{f}_{A_i} \parallel \mathbf{h}_{A_i}\right)$$

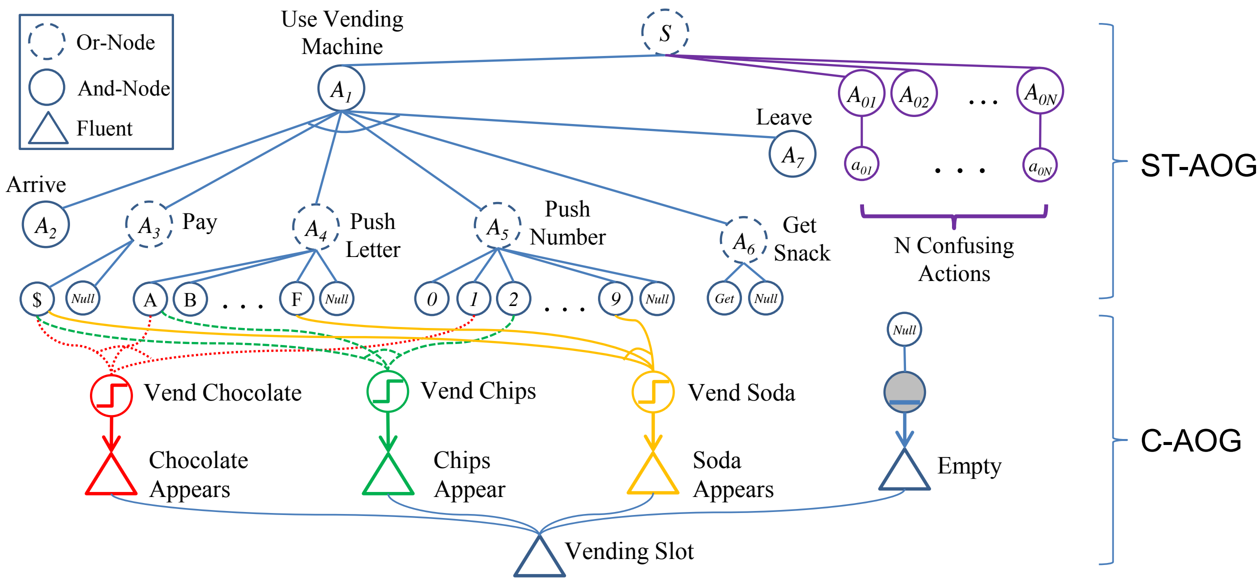
and

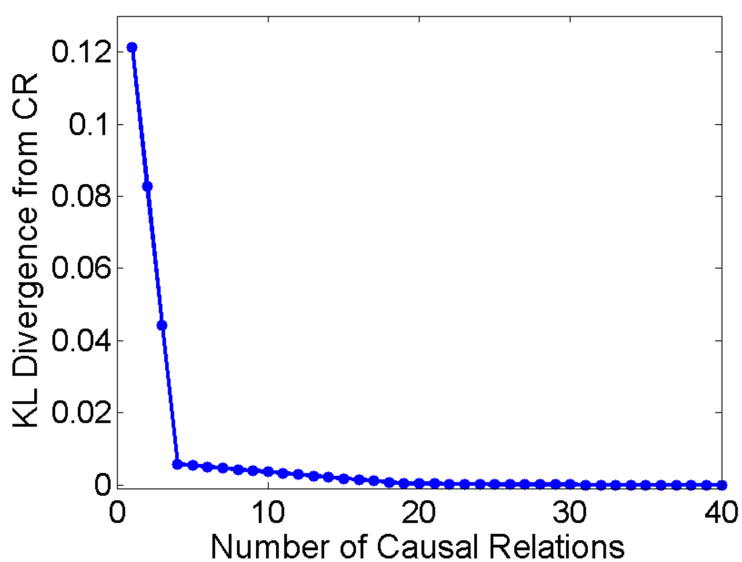
$$\mathbf{cr}_{A} = \underset{\mathbf{cr}_A, \mathbf{cr}_{A_1}, \ldots, \mathbf{cr}_{A_n}}{\operatorname{argmax}} KL\left(\mathbf{f} \parallel \mathbf{h}\right)$$

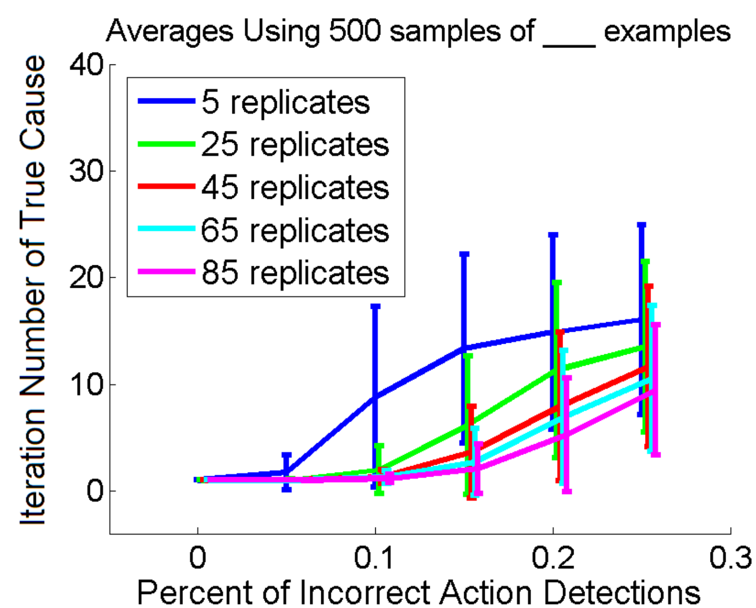




Vending Machine Simluation



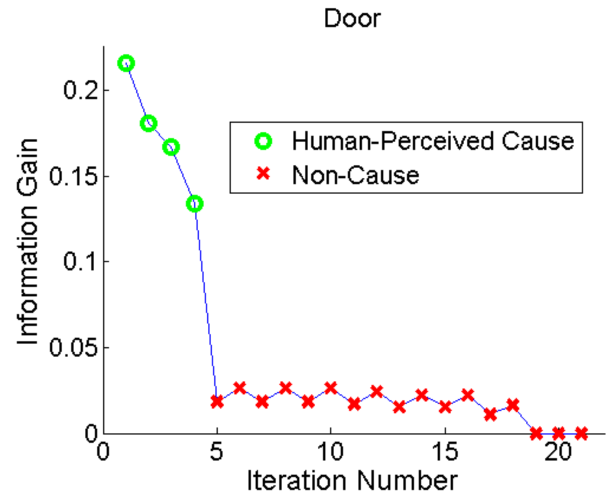




Office Experiment

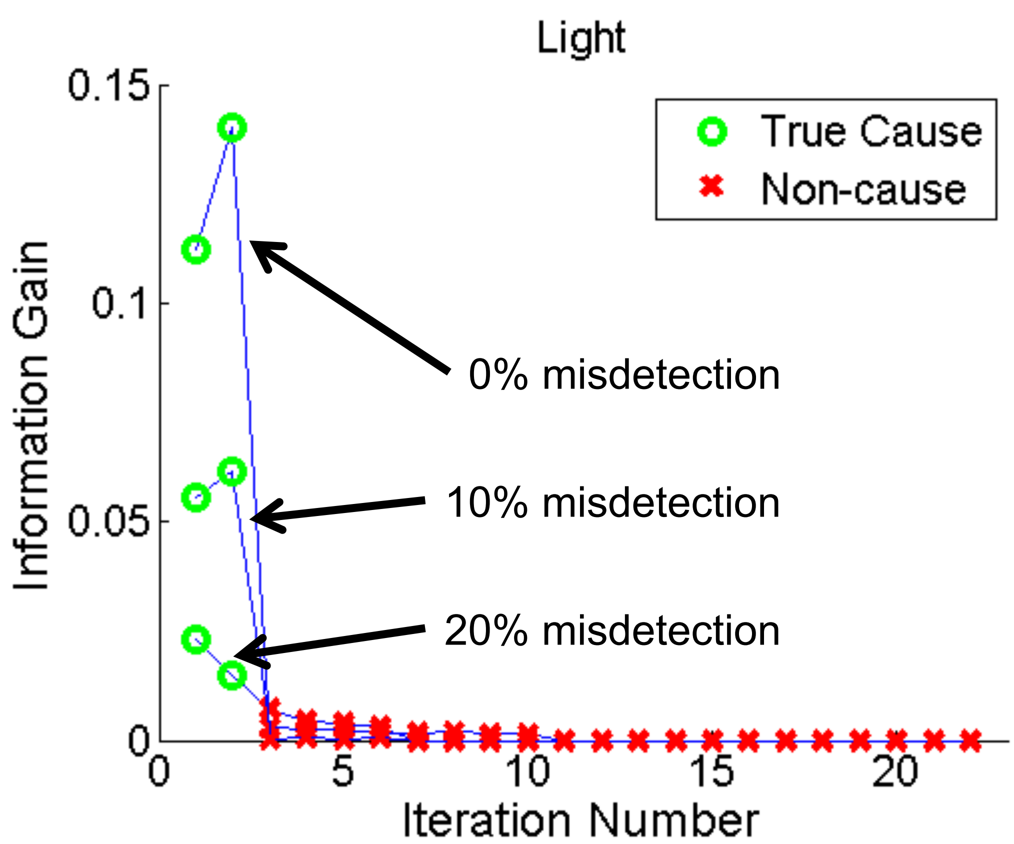
- 5 Scenes
  - Office
  - 3 Door ways (key lock, passcode lock, non-locking)
  - Elevator
- Actions happen 10-20 times; 19 types of low-level actions

Information Gains for the Door



	$C{\to}O$	$O{\to}C$	$O{\to}C$	$C{\to}O$	$O{\to}C$	$C{\to}O$	$O{\to}C$	$C{\to}O$	$O{\to}C$	$C{\to}O$	$O{\to}C$	$C{\to}O$	$O{\to}C$
	A_3	A_4	A_2	A_1	A_6	A_6	A_7	A_7	A_8	A_8	A_{10}	A_{10}	A_5
$k=1$	**0.2161**	0.1812	0.1668	0.1344	0.0185	0.0185	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155
$k=2$	0.0000	**0.1812**	0.1668	0.1344	0.0185	0.0185	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155
$k=3$	0.0000	0.0000	**0.1668**	0.1344	0.0185	0.0185	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155
$k=4$	0.0000	0.0000	0.0000	**0.1344**	0.0185	0.0185	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155
$k=5$	0.0000	0.0000	0.0000	0.0000	**0.0185**	**0.0185**	**0.0185**	**0.0185**	0.0185	0.0185	0.0170	0.0170	0.0155
$k=6$	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0264**	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155
$k=7$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0185**	**0.0185**	0.0185	0.0185	0.0170	0.0170	0.0155
$k=8$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0264**	0.0185	0.0185	0.0170	0.0170	0.0155
$k=9$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0185**	**0.0185**	0.0170	0.0170	0.0155
$k=10$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0264**	0.0170	0.0170	0.0155
$k=11$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0170**	**0.0170**	0.0155
$k=12$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0244**	0.0155
$k=13$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0155**

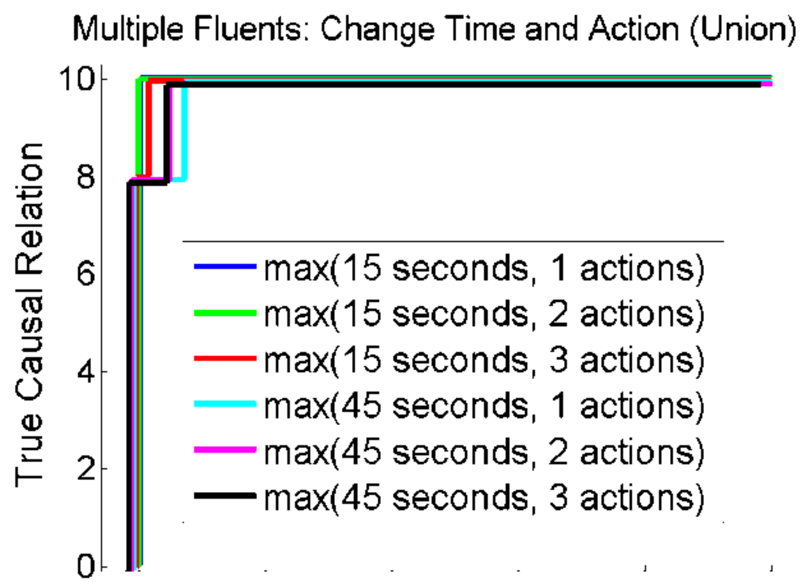
Increasing Misdetections
(Simulation)

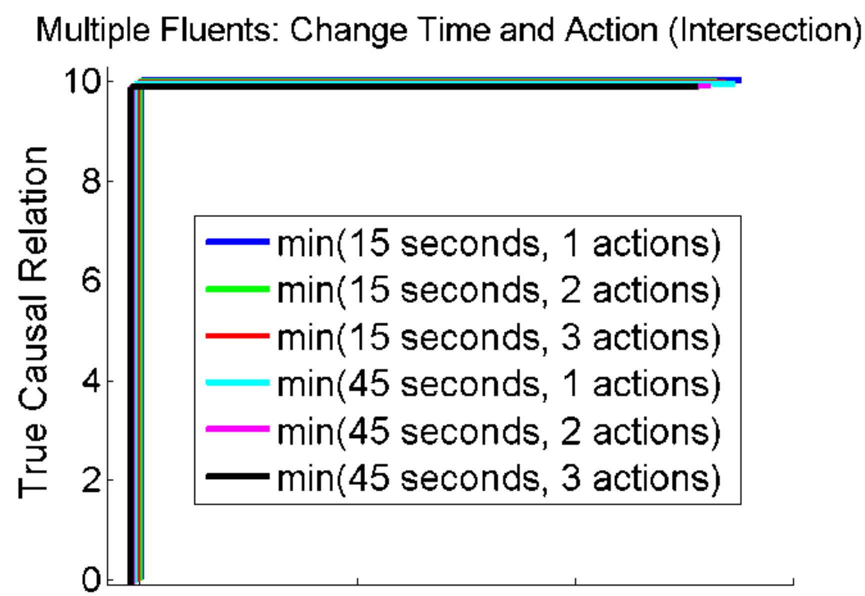


This includes increased false alarms and false negatives

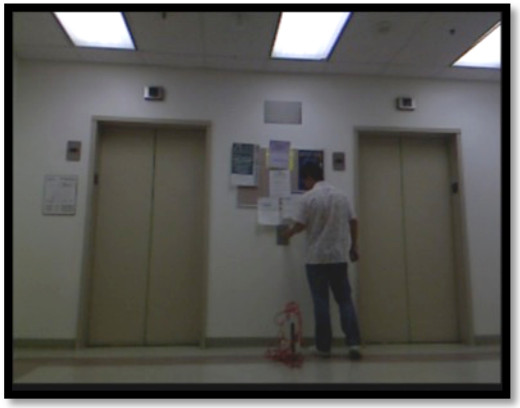
Preparing Video Clips: Latent Time

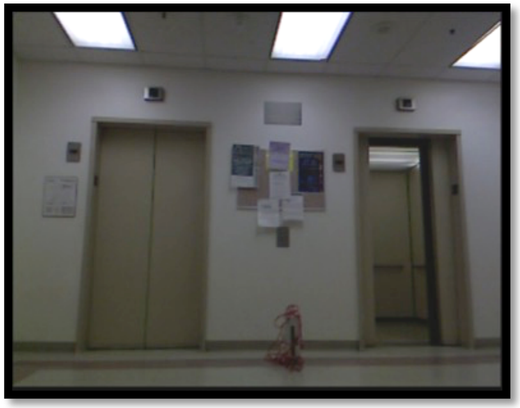
- 3 fluents, 10 true causes, 66 potential causal relations
- Actions happens 8-10 times





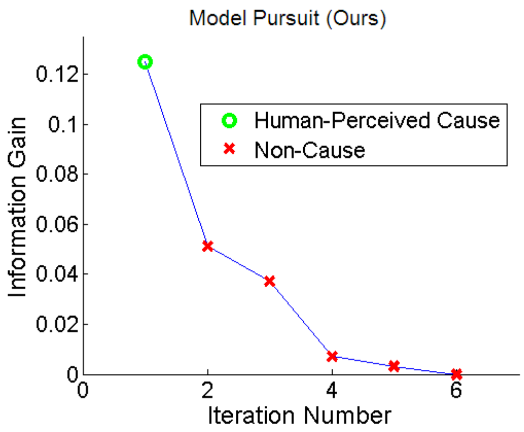
Delayed Effects: Performance vs. TE

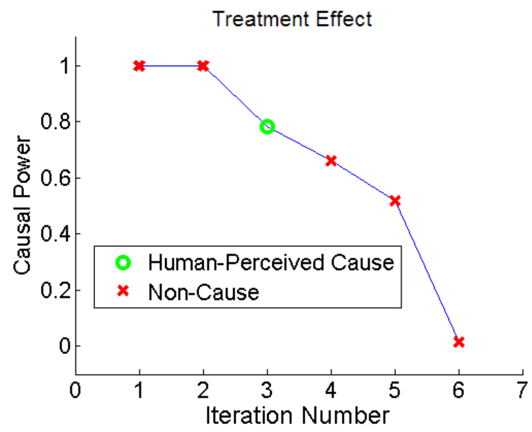




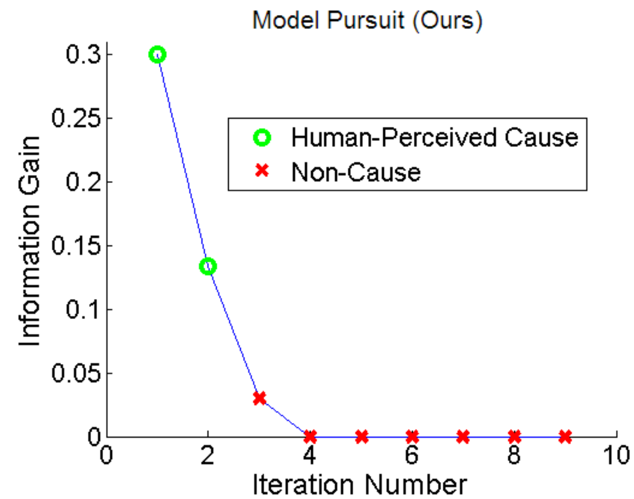
- TE looks at the marginal:

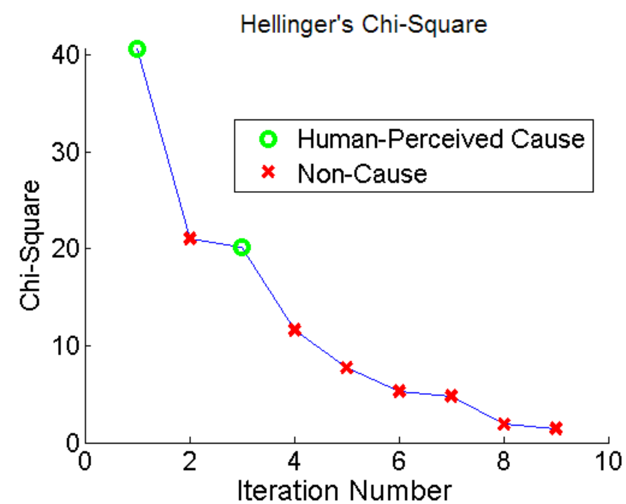
  $TE = P(\Delta F | A) - P(\Delta F | \neg A).$



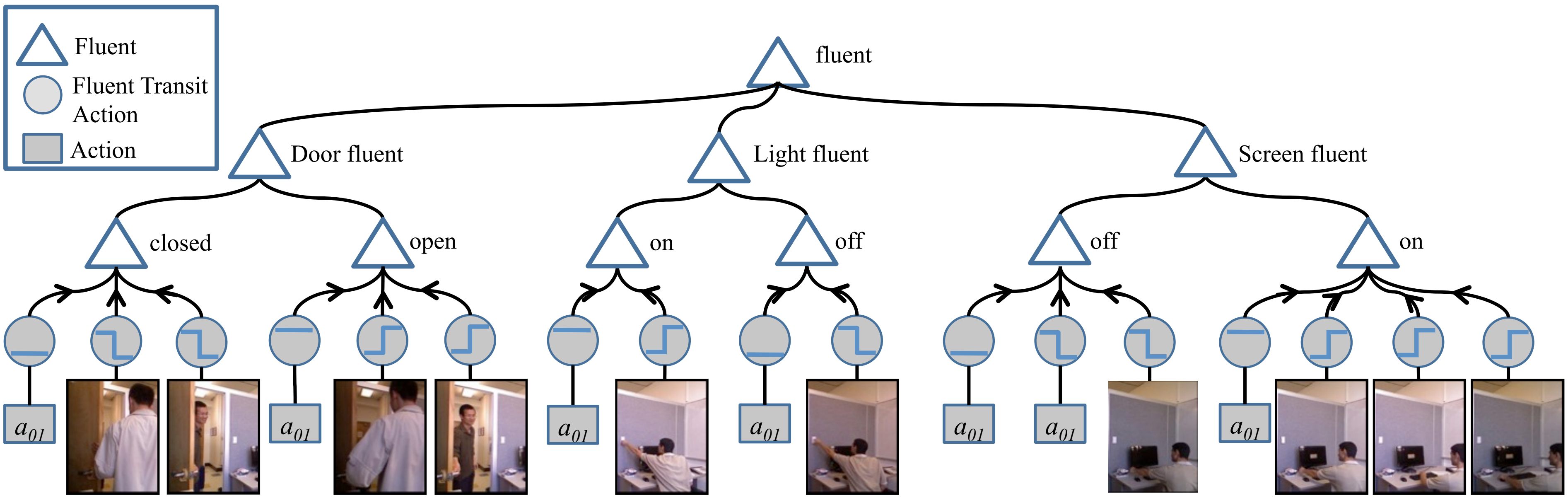


Hierarchical:
Performance vs. Hellinger χ^2





C-AOG



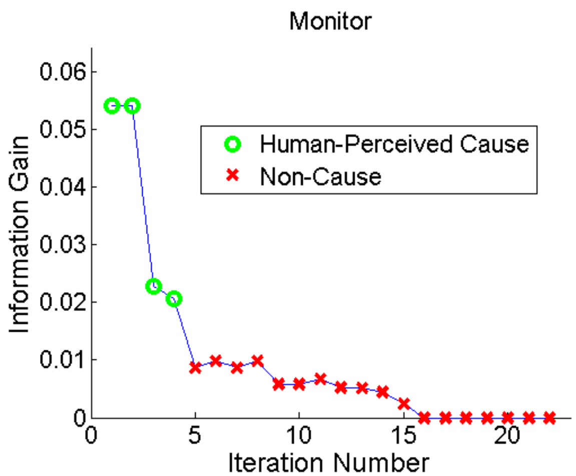
$$P_C(pg \mid V) = P_{STC}(pg \mid F, V) \propto \exp(-E_C(pg \mid V))$$

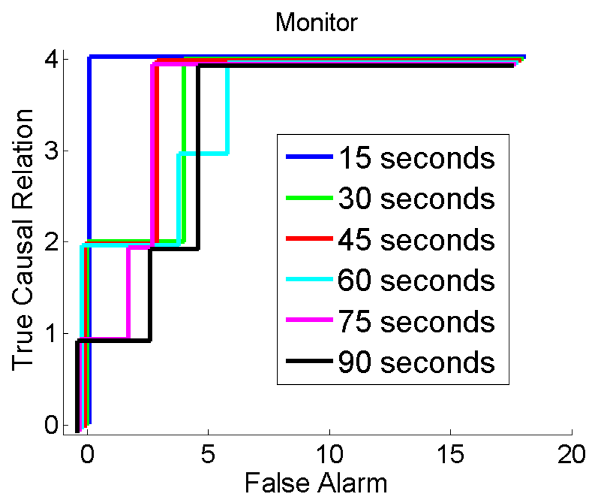
$$E_C(pg \mid V) = E_{ST}(pg \mid V) + \sum_{a \in CR(pg)} \lambda_a(w(a))$$

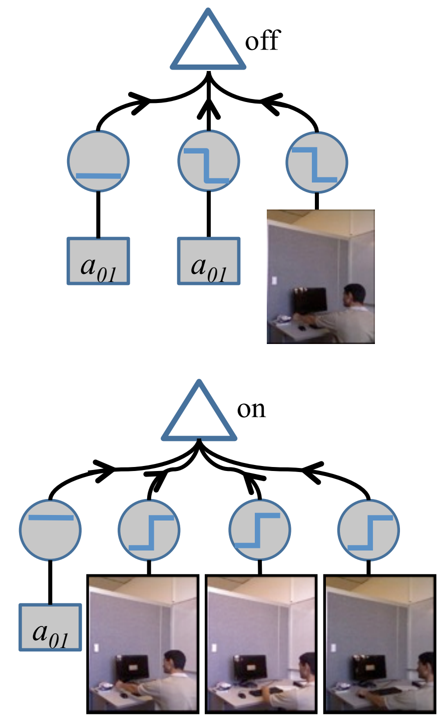
Hard Example: The Monitor

(Hidden Variables)

- Power button – turns power off and on
- Moving mouse or touching keyboard wakes screen if powered

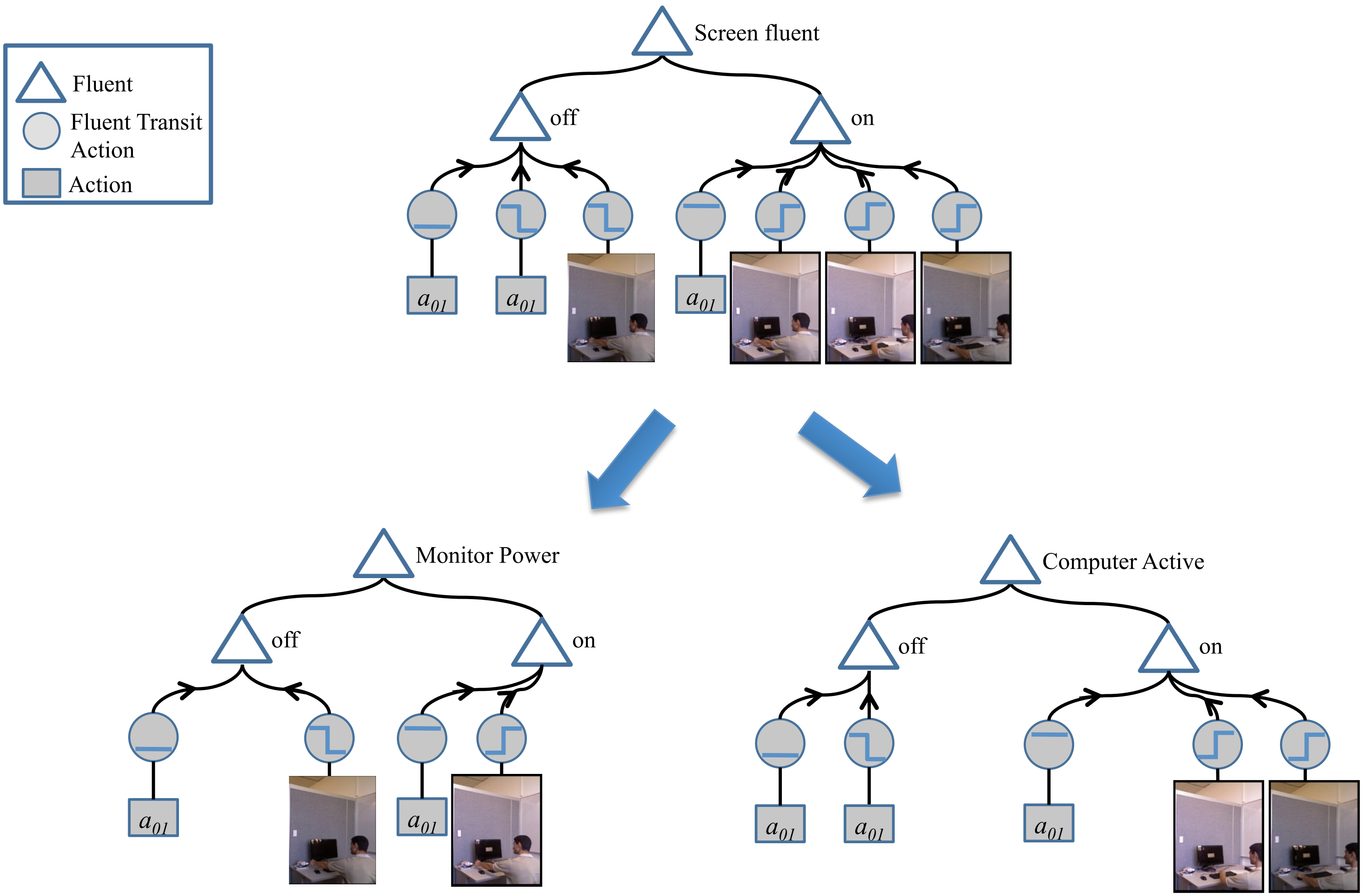






- TE and χ^2 are low for this example, reflecting the difficulty

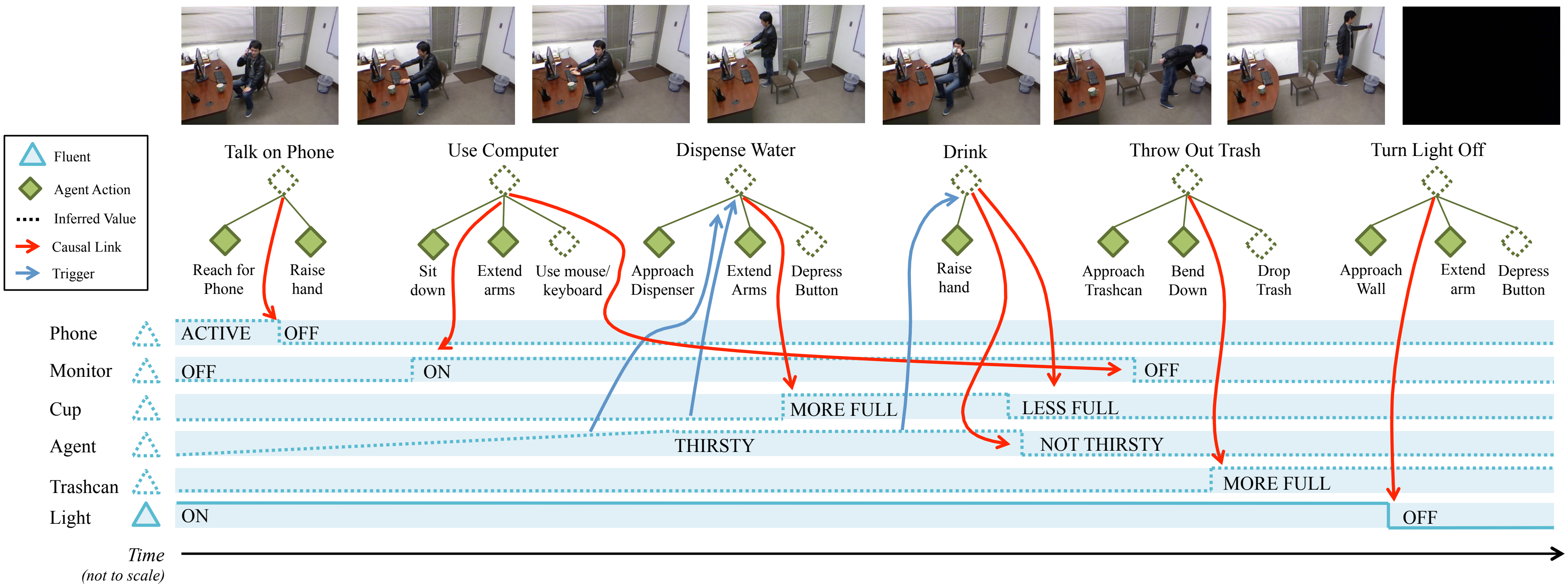
Monitor: What Happened



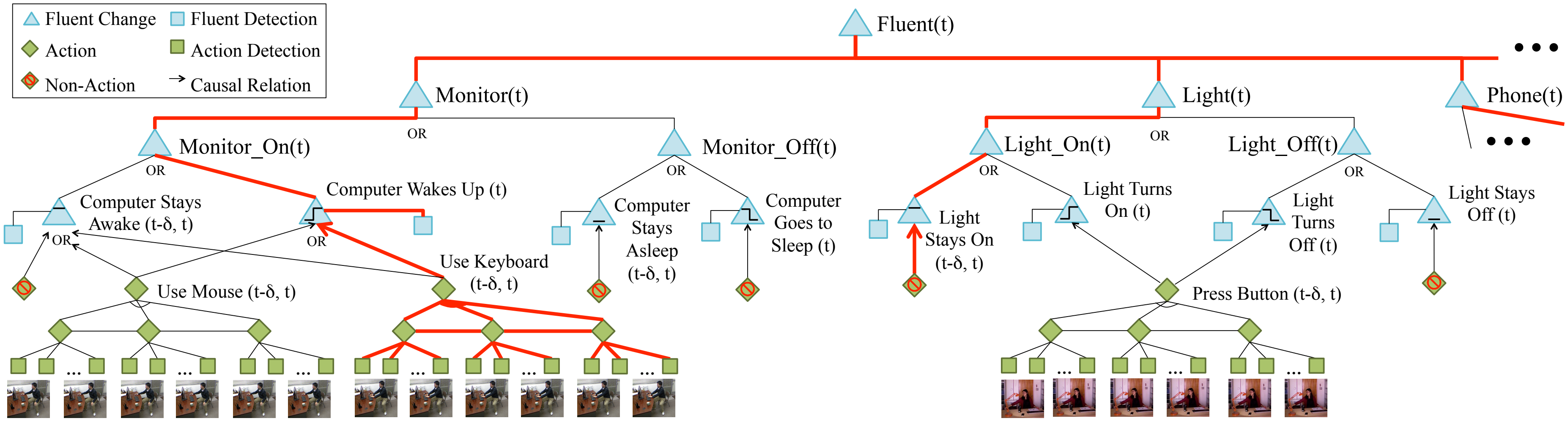
From Real Data

REASONING OVER TIME

Recall: Example Causal Inference



Grounding on Detectors



- Terminal Leaves
  - Represent features for detection
- Temporal Relations
  - Links connect nodes with temporal relationships

Parse Graph and Energy

$$P(pg_t|V[t-\delta,t]) \propto P(pg_t;\Theta) \prod_{l \in L(pg_t)} P(l|pg_t)$$

- Or-nodes

$$\mathcal{E}(O) = \max_{v \in ch(O)} (\mathcal{E}(v) + \langle \Theta_v, \lambda_v \rangle)$$

- And-nodes

$$\mathcal{E}(A) = \sum_{v \in ch(A)} \mathcal{E}(v|A)$$

- Temporal Relations

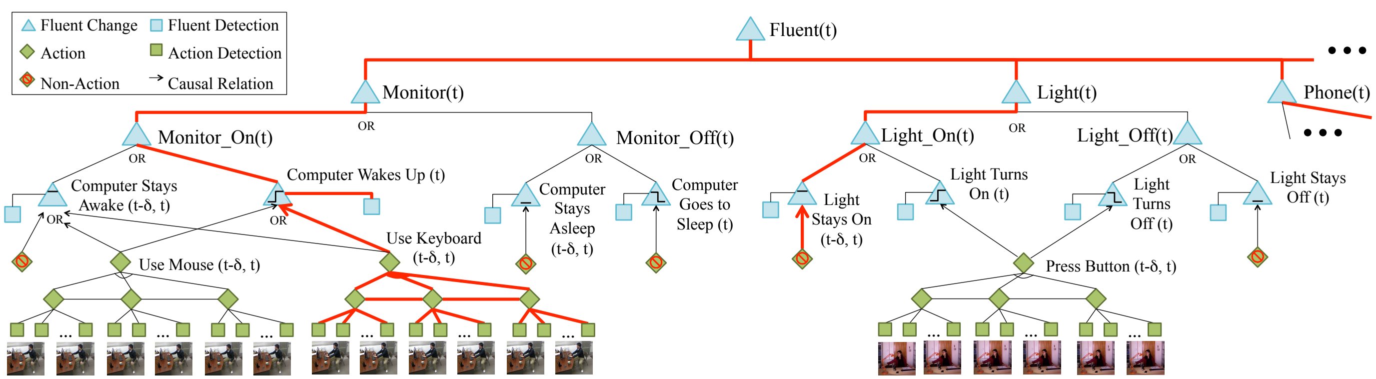
$$\tilde{v} = v_{i_1}, \ldots, v_{i_k} \qquad \mathcal{E}(R) = \psi_{\tilde{v}}(\tilde{v})$$

- Terminal Leaves

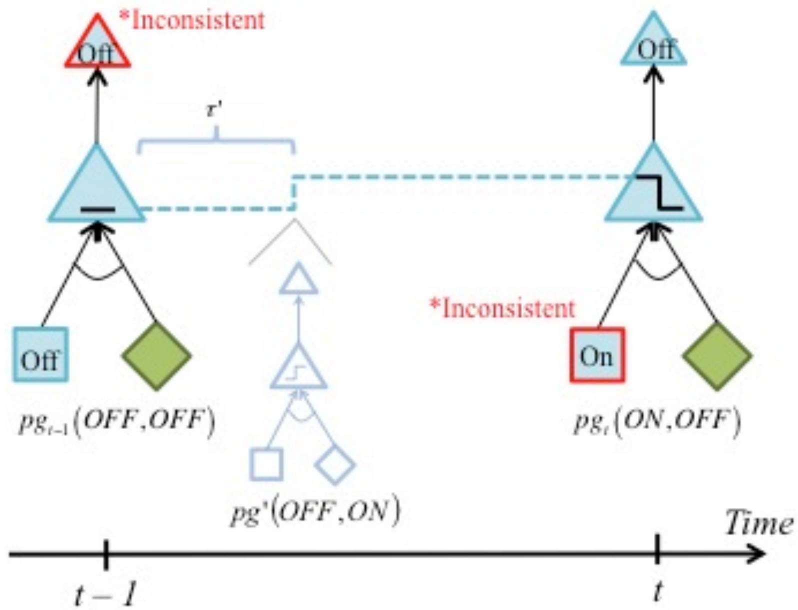
$$\mathcal{E}(l_F|F)$$

$$\mathcal{E}(l_A|A)$$

$$\mathcal{E}(pg_t|V[t-\delta,t]) = \sum_{l_F \in L_F(pg)} \mathcal{E}(l_F|F) + \sum_{l_A \in L_A(pg)} \mathcal{E}(l_A|A) + \sum_{\tilde{v} \in R} \psi_{\tilde{v}}(\tilde{v}) + \sum_{v \in O(pg)} \langle \Theta_v, \lambda_v \rangle$$

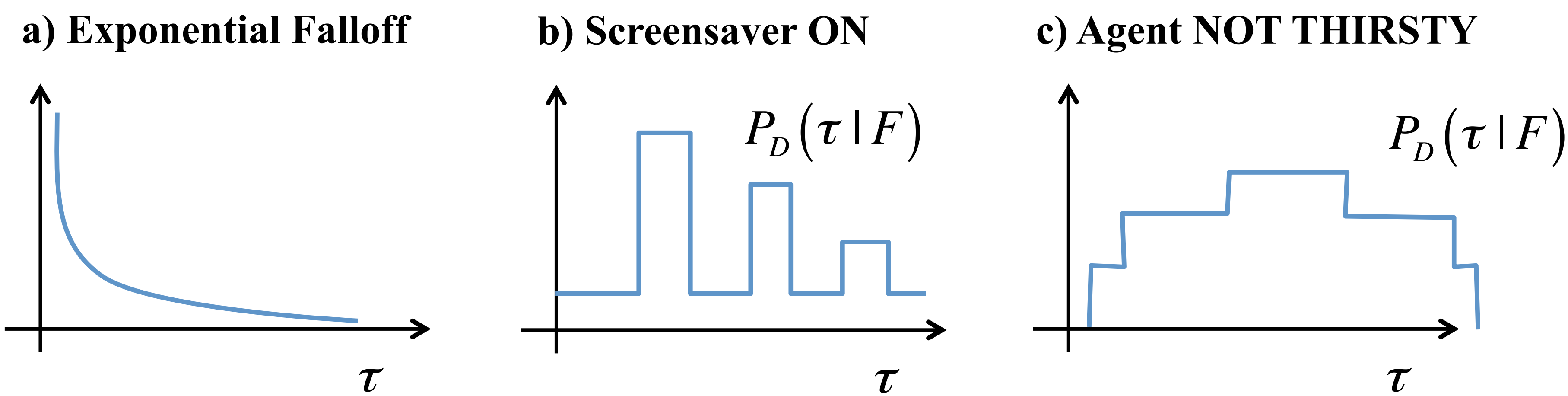


Issue Over Time: Consistency

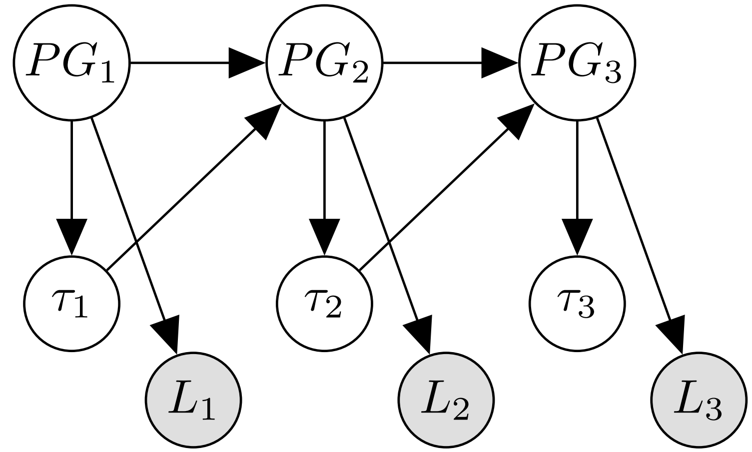


$$P(pg_t|pg_{t-1}) = \begin{cases} 0, \text{ if } pg_{t-1},\ pg_t \text{ inconsistent} \\ 1, \text{ otherwise.} \end{cases}$$

Issue Over Time: Non-Markovian Duration



Hidden Semi-Markov Model



$$\underbrace{pg_1, ...pg_1}_{\tau_1}, \underbrace{pg_2, ...pg_2}_{\tau_2}, \underbrace{pg_3, ...pg_3}_{\tau_3}$$

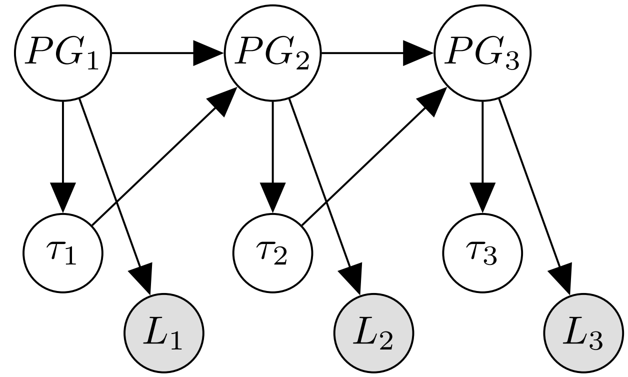
$$P(PG_t = pg | PG_{t-1} = pg', \tau_{t-1} = d) = \begin{cases} \delta(pg, pg'), \text{ if } d > 0 \\ \qquad \text{(remain in same state)} \\ P(pg|pg'), \text{ if } d = 0 \\ \qquad \text{(transition per Eq. 5.6)} \end{cases}$$

$$P(\tau_t = d' | PG_t = pg) = \begin{cases} \delta(d', d-1), \text{ if } d > 0 \\ \qquad \text{(decrement)} \\ P(\tau|F), \text{ if } d = 0 \\ \qquad \text{(per Sec. 5.2.2).} \end{cases}$$

Murphy, 2002. HSMM

Viterbi Algorithm

$$\mathbf{PG}^*, \tilde{\tau}^* = \underset{\mathbf{PG}, \tilde{\tau}}{\operatorname{argmax}} P(\mathbf{PG}, \tilde{\tau} | V)$$



- Viterbi equation

$$\begin{aligned} V_t(pg, \tau) &\triangleq \max_{pg', \tau'} P\left(PG_t = pg, \tau_t = \tau, PG_{t-1} = pg', \tau_{t-1} = \tau', L_{1:t} = l_{1:t}\right) \\ &= P(l_{t-\tau+1:t} | pg) \max_{pg', \tau'} P(pg, | pg') P(\tau | F) V_{t-\tau}(pg', \tau'). \end{aligned}$$

- Remove $\boldsymbol{\tau}$ from state space

$$V_t(pg) = \max_{\tau} \left[P\left(l_{t-\tau+1:t} | pg\right) P\left(\tau | F\right) \max_{pg'} P\left(pg | pg'\right) V_{t-\tau}(pg') \right]$$

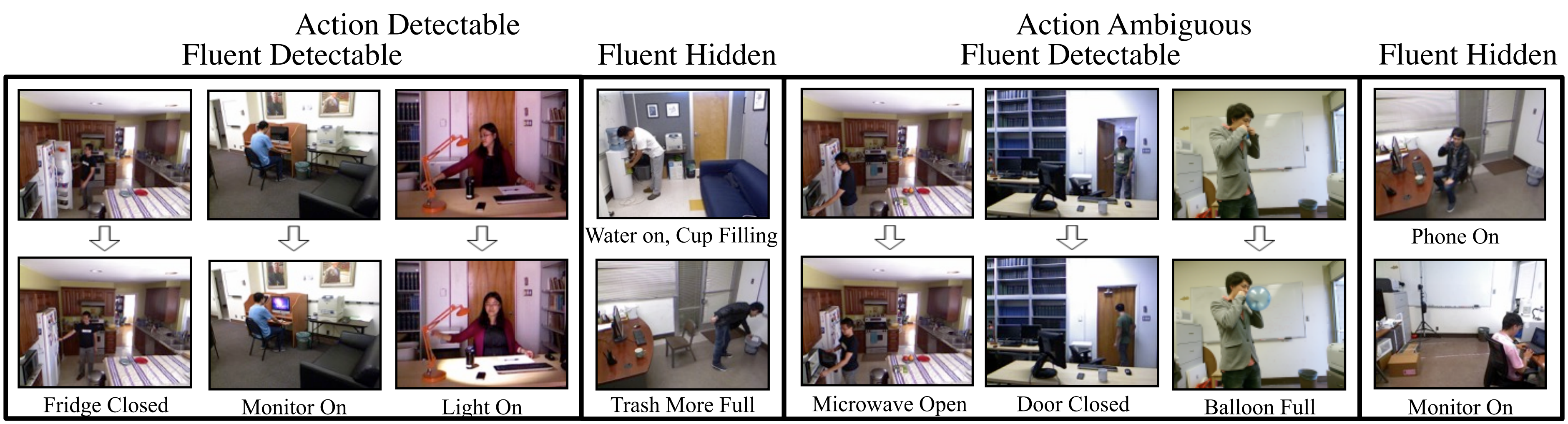
- Complexity $O(T \cdot |PG|^2 \cdot |\tau|)$
  - Precompute $P\left(l_{t-\tau+1:t} | pg\right)$

Murphy, 2002. HSMM

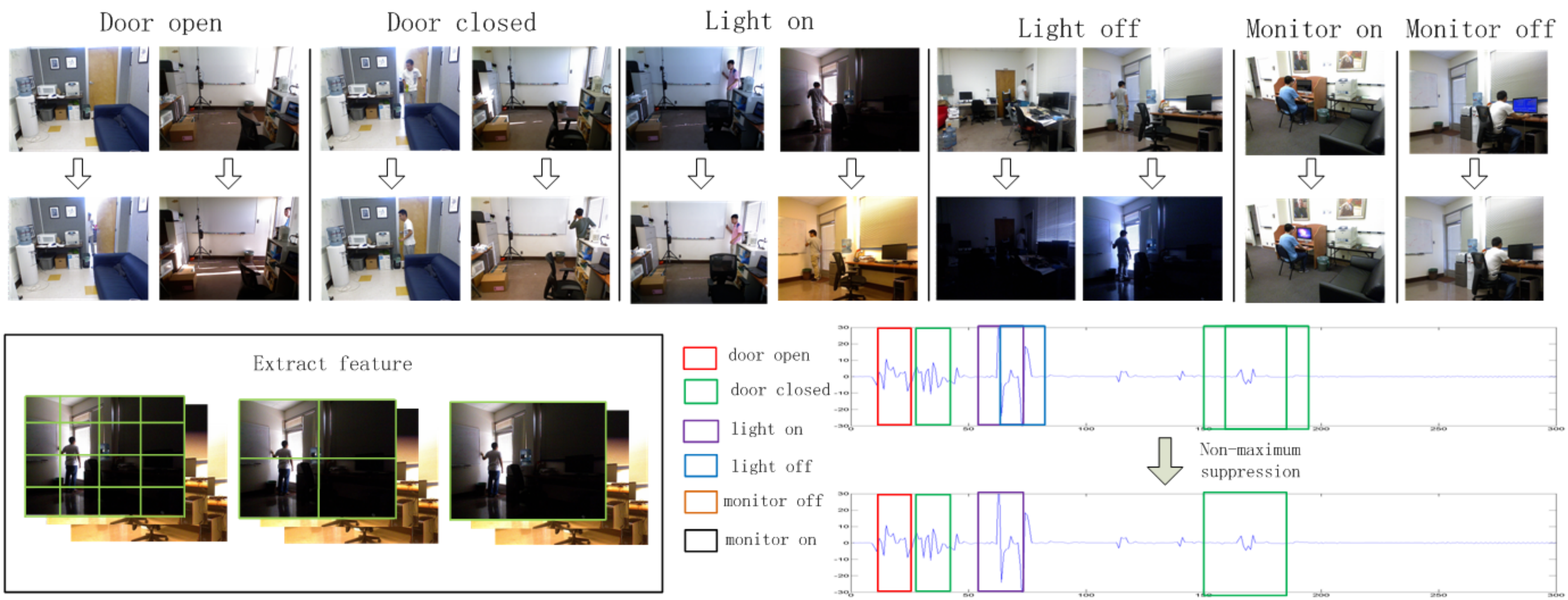
New Causal Video Data

- 4D Data
- nFrames/clip ~ 300
- Training
  - 3-10 of each

Object	Fluent Values	Causing Actions	nScenes	nClips	nFrames
door	open/closed	open door, close door	4	50	10611
light	on/off	turn light on/off	4	34	16631
screen	on/off	use computer	4	179	56632
phone	active/off	use phone	5	68	30847
cup	more/less/same	fill cup, drink	3	48	16564
thirst	thirsty/not	drink	3	48	16564
waterstream	on/off	fill cup	3	40	14061
trash	more/less/same	throw trash out	4	11	2586
microwave	open/closed, running/not	open door, close door turn on	1	3	4245
balloon	full/empty	blow up balloon	1	3	664
fridge	open/closed	open door, close door	1	2	2751
blackboard	written on/ clear	write on board, erase	1	2	5205
faucet	on/off	turn faucet on/off	1	2	3013



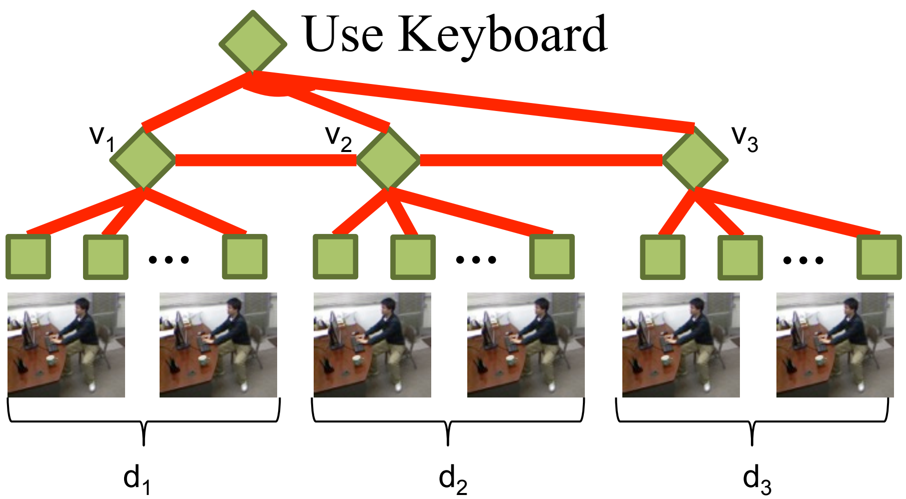
Detecting Fluent Changes



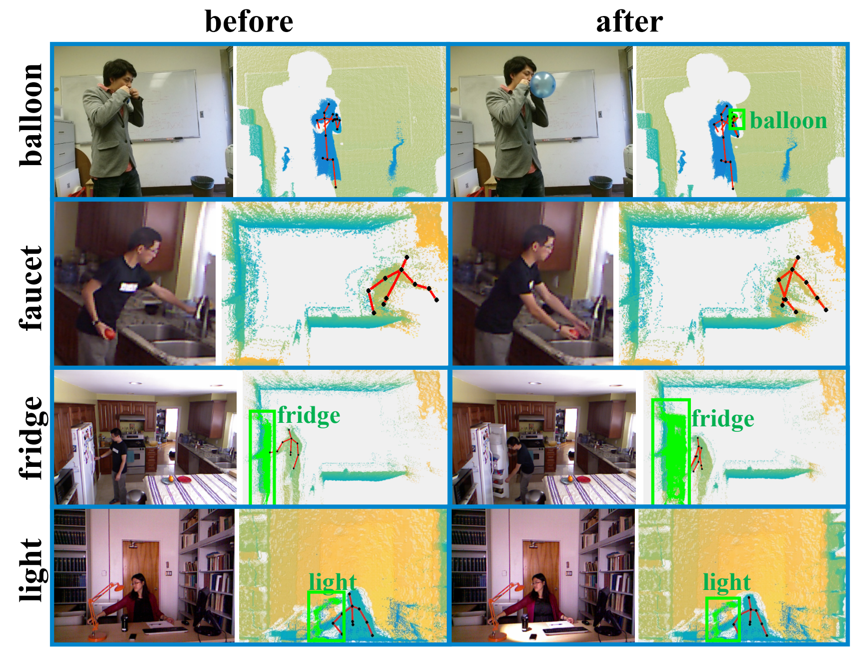
- 3-level spatial pyramid
- GentleBoost
- Non-max suppression

$$\mathcal{E}(l_F|F)$$

Detecting Actions



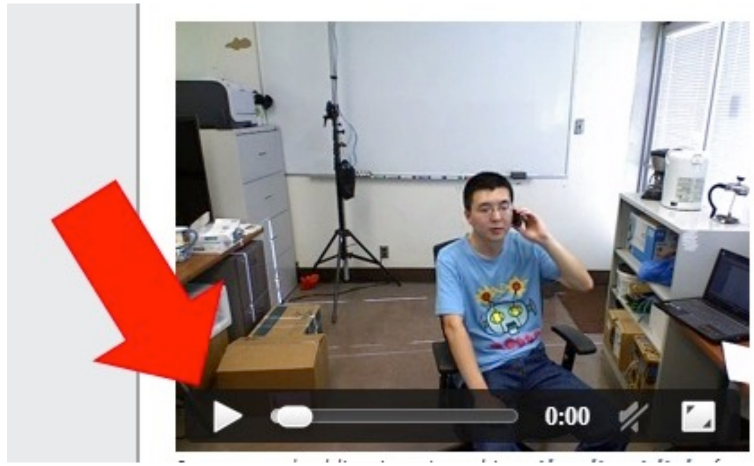
$$\psi(\tilde{v}) = P(v_n | v_{n-1}, d_{n-1})$$

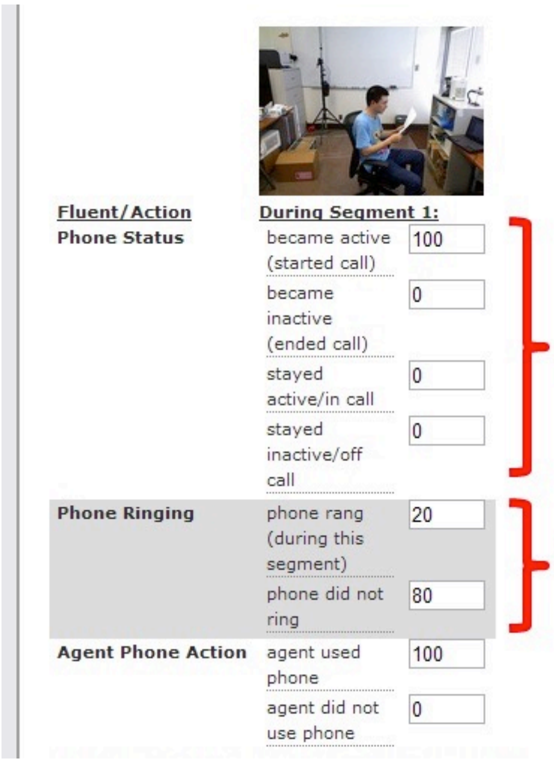


- Beam search k = 1,000,000
- Sliding Window: 50, 100, 150 frames
- Input to Causal Grammar: $\mathcal{E}(l_A|A)$
- Detection Baseline: Non-max suppression

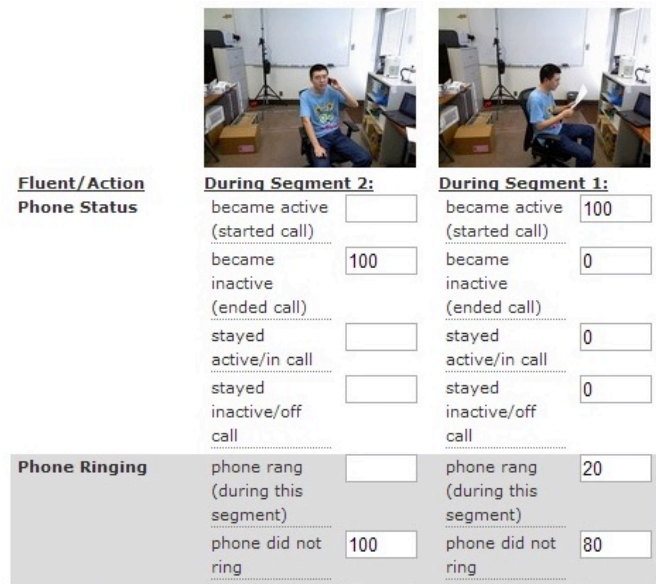
Wei et al., Modeling 4D Human-Object Interactions for Event and Object Recognition

Human Annotation



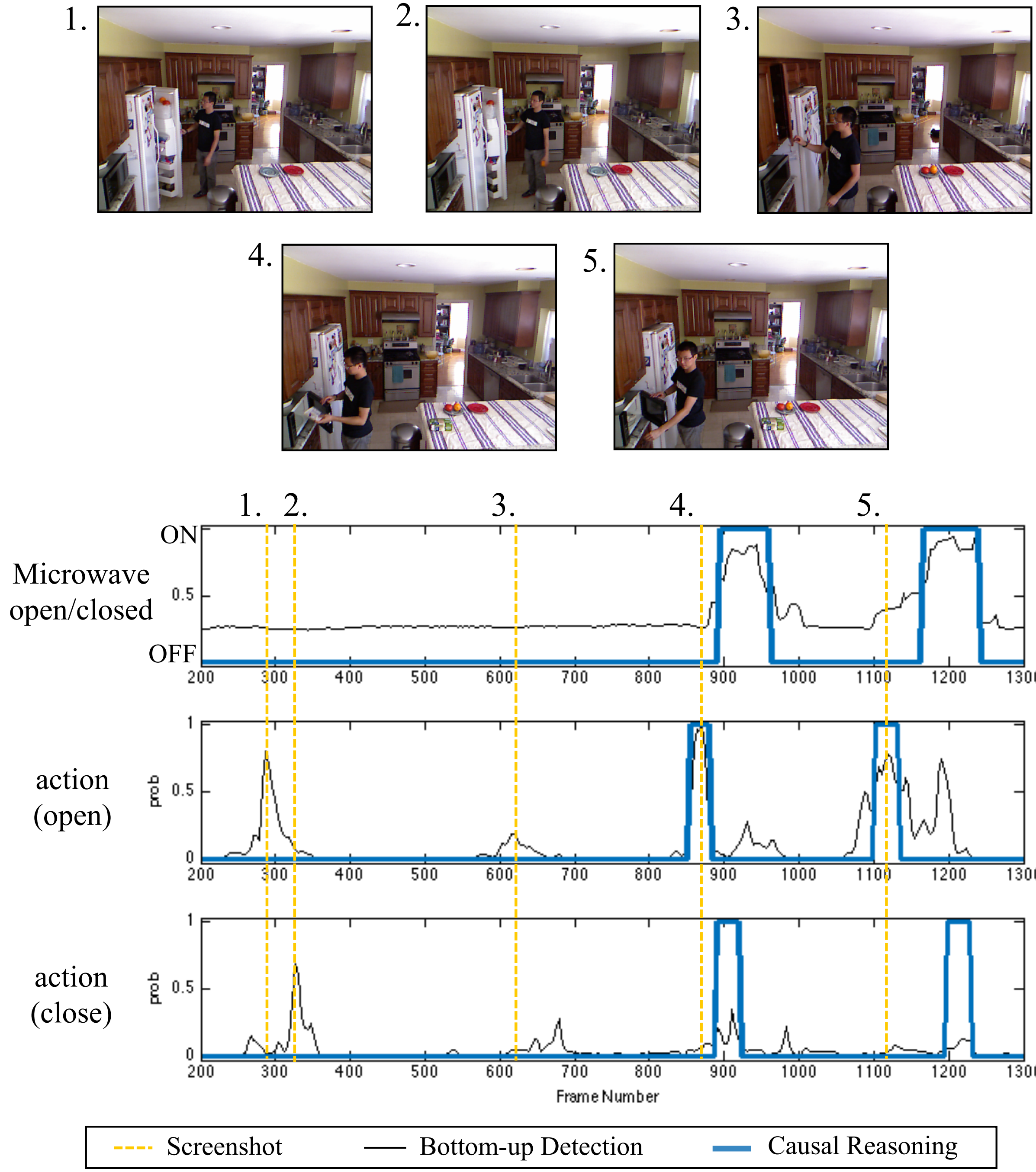


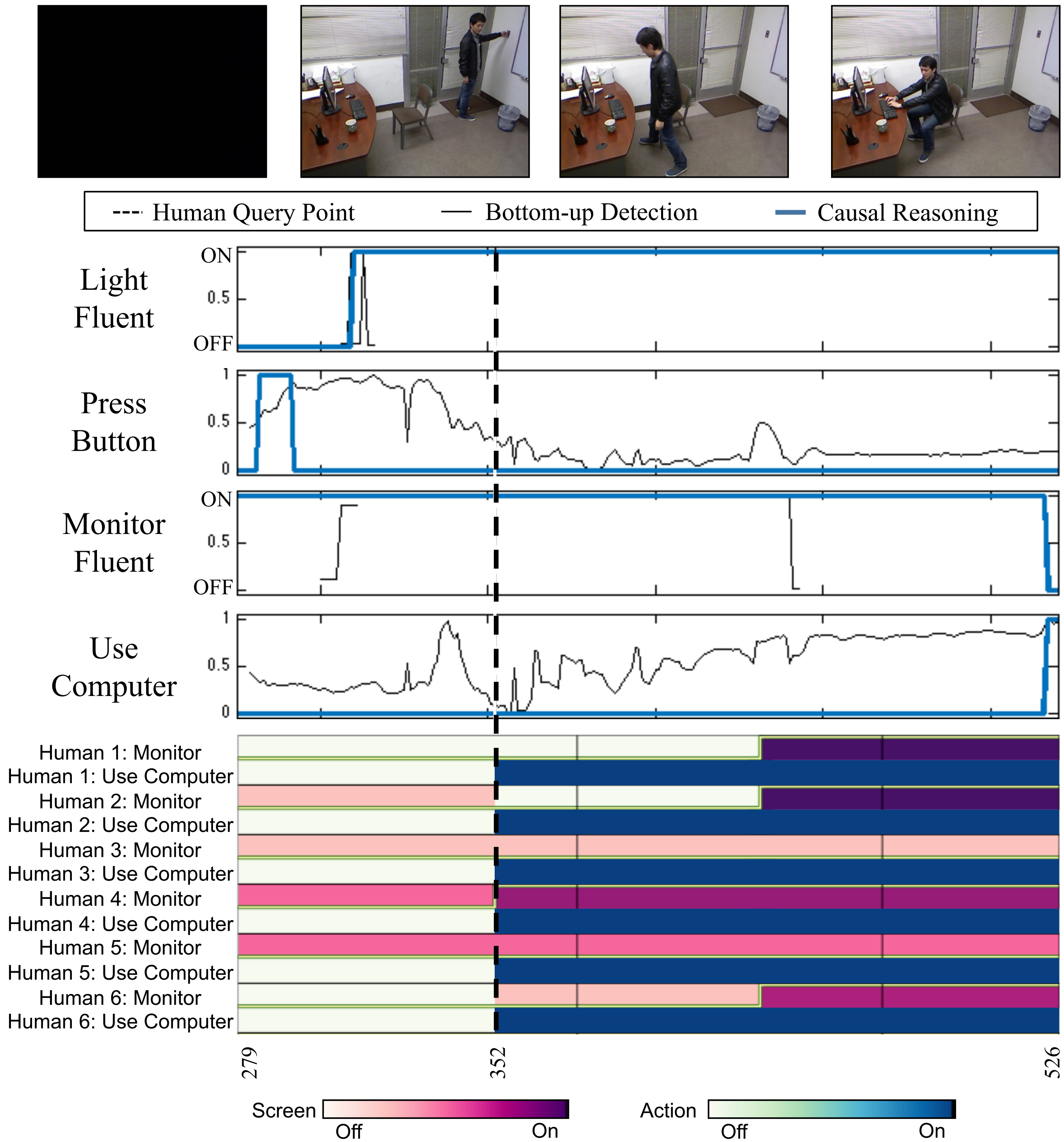
Each block must sum to 100



Edit earlier responses if needed

- **Evaluation**
  - Hit: Exactly match the nearest human
  - Ground truth positive: Human awarded more than 50 to a single answer


1.
2.
3.
4.
5.
Microwave open/closed
ON
OFF
0.5
action (open)
action (close)
prob
0
1
200
300
400
500
600
700
800
900
1000
1100
1200
1300
Frame Number
Screenshot
Bottom-up Detection
Causal Reasoning



Human Query Point
Bottom-up Detection
Causal Reasoning
Light Fluent
ON
0.5
OFF
Press Button
1
0.5
0
Monitor Fluent
ON
0.5
OFF
Use Computer
1
0.5
0
Human 1: Monitor
Human 1: Use Computer
Human 2: Monitor
Human 2: Use Computer
Human 3: Monitor
Human 3: Use Computer
Human 4: Monitor
Human 4: Use Computer
Human 5: Monitor
Human 5: Use Computer
Human 6: Monitor
Human 6: Use Computer
279
352
526
Screen
Off
On
Action
Off
On

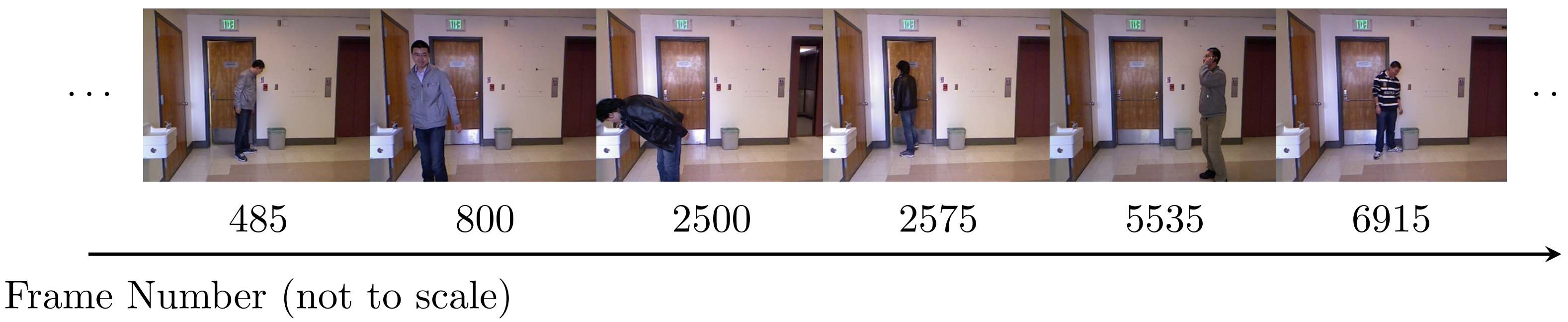
Hit Rates, PR

		trash	door	cup	light	screen	thirst	phone	waterstream	Average
Action	Noise	0.10	0.00	N/A	0.00	0.12	0.03	0.00	0.00	0.04
	Detection	0.62	0.45	N/A	0.57	0.61	0.41	0.33	0.38	0.48
	Causal	**0.87**	**0.58**	N/A	**0.80**	**0.67**	**0.76**	**0.40**	**0.88**	**0.71**
Fluent	Noise	*0.00*	0.00	*0.00*	0.00	0.25	*0.08*	*0.00*	*0.00*	0.04
	Detection	*0.00*	0.42	*0.00*	0.43	0.17	*0.11*	*0.00*	*0.00*	0.14
	Causal	***0.77***	**0.53**	***0.62***	**0.61**	**0.74**	***0.57***	***0.19***	***0.81***	**0.61**

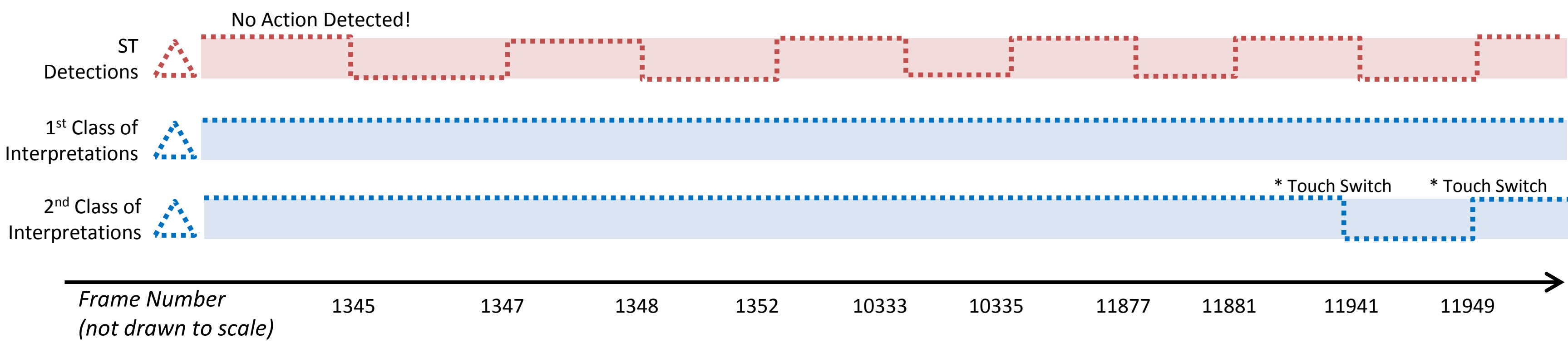
- Noise
  - All responses equally likely
- Causal Grammar
  - AP = 0.63, AR = 0.69
- Detections
  - AP = 0.29, AR = 0.31

1) Causal grammar wins!
2) Non-zero noise
3) Mismatch on hidden fluents: Detection, noise (thirst)
4) Hidden fluents improve actions through the prior
5) Fluent detections compete with action detections

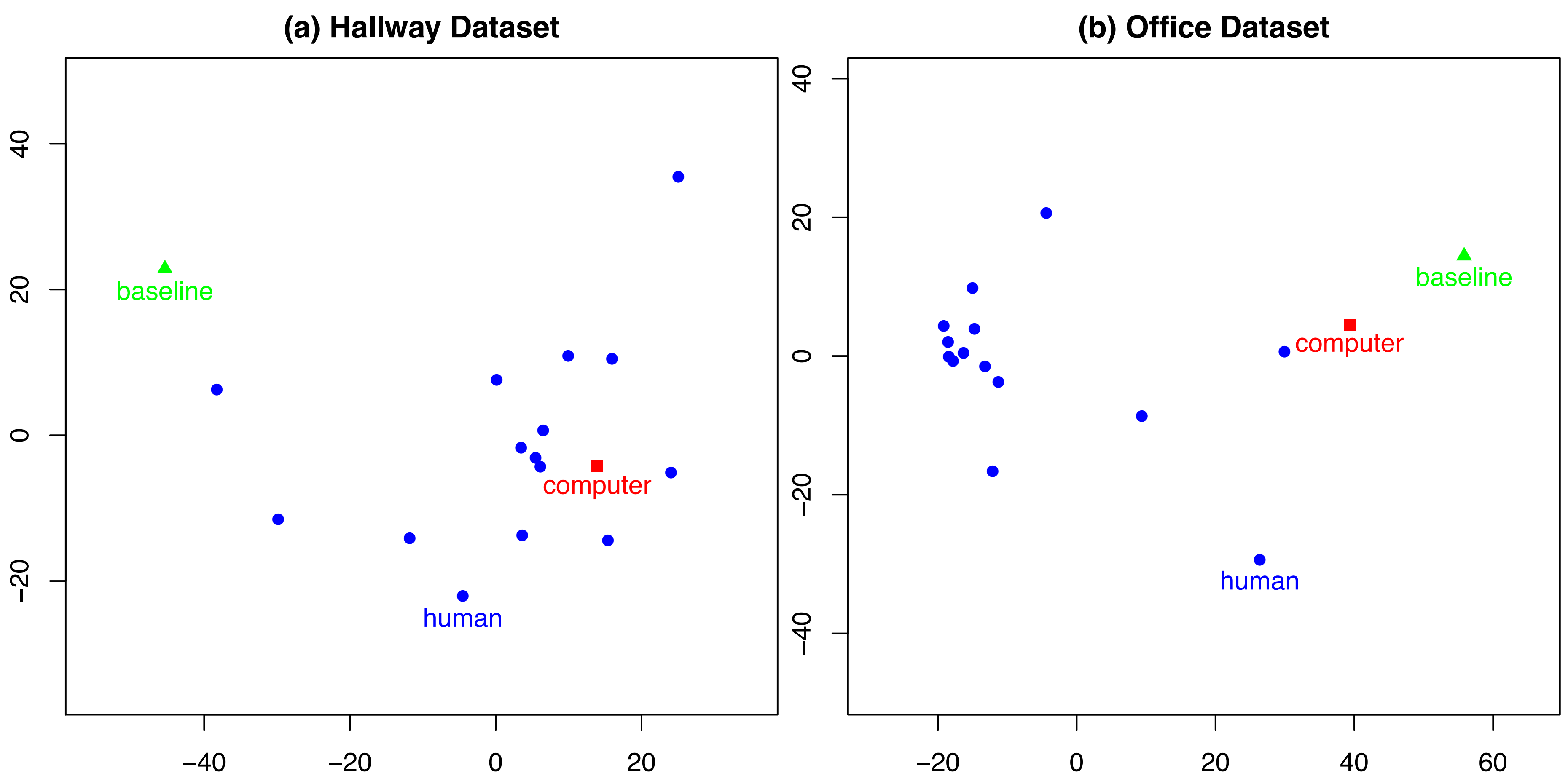
Experiment 2: Human Variability



- Correcting Misdetection



MDS plots



(a) Hallway Dataset
baseline
computer
human
40
20
0
−20
−40
−20
0
20
(b) Office Dataset
baseline
computer
human
40
20
0
−20
−40
−20
0
20
40
60

Summary

- Learning
  - Learning causal relations in an unsupervised way, linking fluent changes to their causing actions
- Representation
  - Provided representation for causal knowledge consistent with current And-Or Graph representations: The Causal And-Or Graph
- Inference
  - Through the extended C-AOG, provided framework for reasoning
  - Modeling perceptual causality may not be a true representation of the world, but it is useful.

Future Work

- Integrate learning with learning in other domains (spatial, temporal)
- Explore learning hidden variables
  - Explore temporal lag
  - Confounding
- Expand reasoning
  - We put a prior on why things happen
  - We need a prior on why they don't
  - More on intents/goals
  - More complicated scenarios
- Other paradigms for learning
  - Lasso: Constrain the lambdas
  - Bayesian prior
  - Online learning/dynamic experimental design
    - Handle new "surprising" information
    - Measure variability/uncertainty in our solutions for when we don't have ground truth
    - Learning: Selection analysis

Thank you!

Any Questions?