Temporal AND-OR Graph

for representation and recognition of Events, Actions, Motions

Song-Chun Zhu, Sinisa Todorovic, and Ales Leonardis

At CVPR, Providence, Rhode Island
June 16, 2012

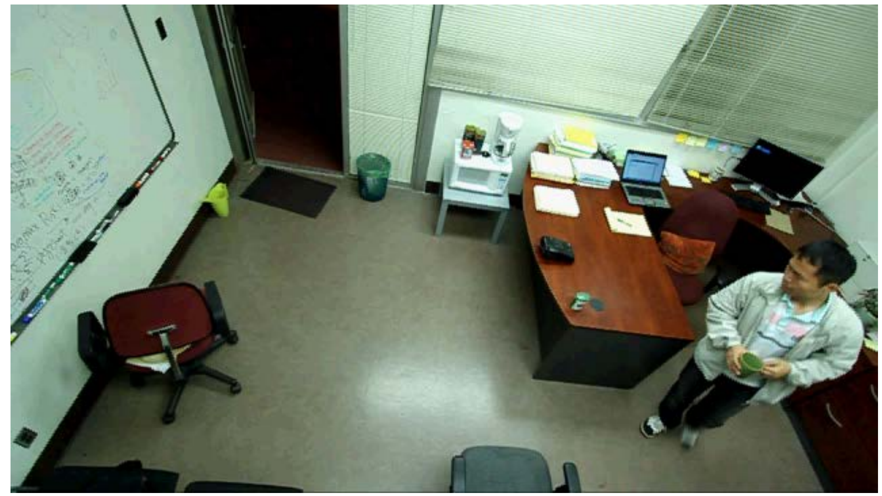
Goal: Recognize events in daily scenes

- For example, an office.

Ref. Pei, Si and Zhu, ICCV2011.

Challenges

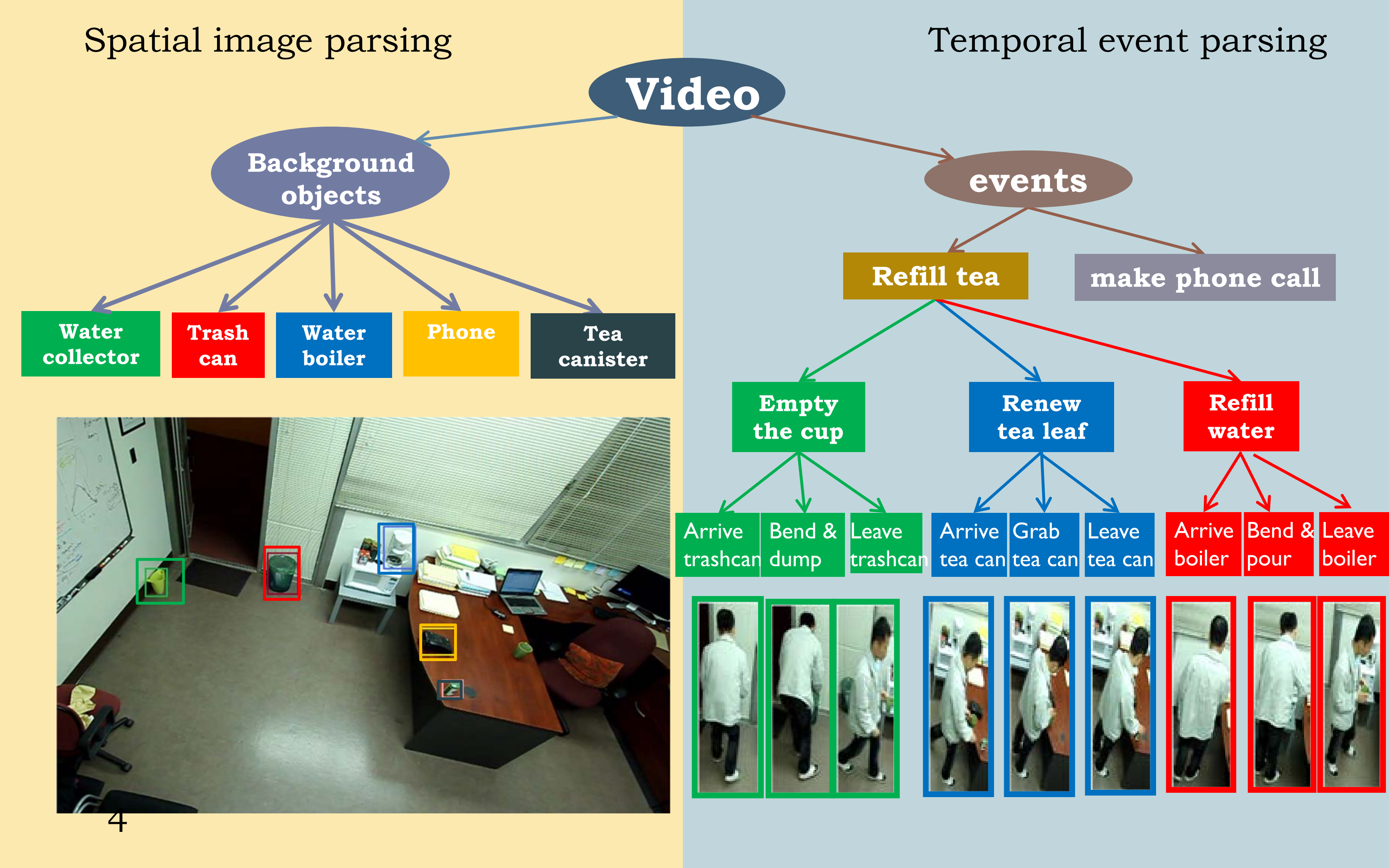
1. Events happen over an extended time period
    - Variant time-span
    - Could be interrupted
    - Multiple routes
    - Intention and prediction



2. Actions are hard to recognize
    - Subtle and similar
    - No salient motion/pose at most of the time
    - Contextual objects -- key!!

Use laptop | Read book | Dump water | Use microwave

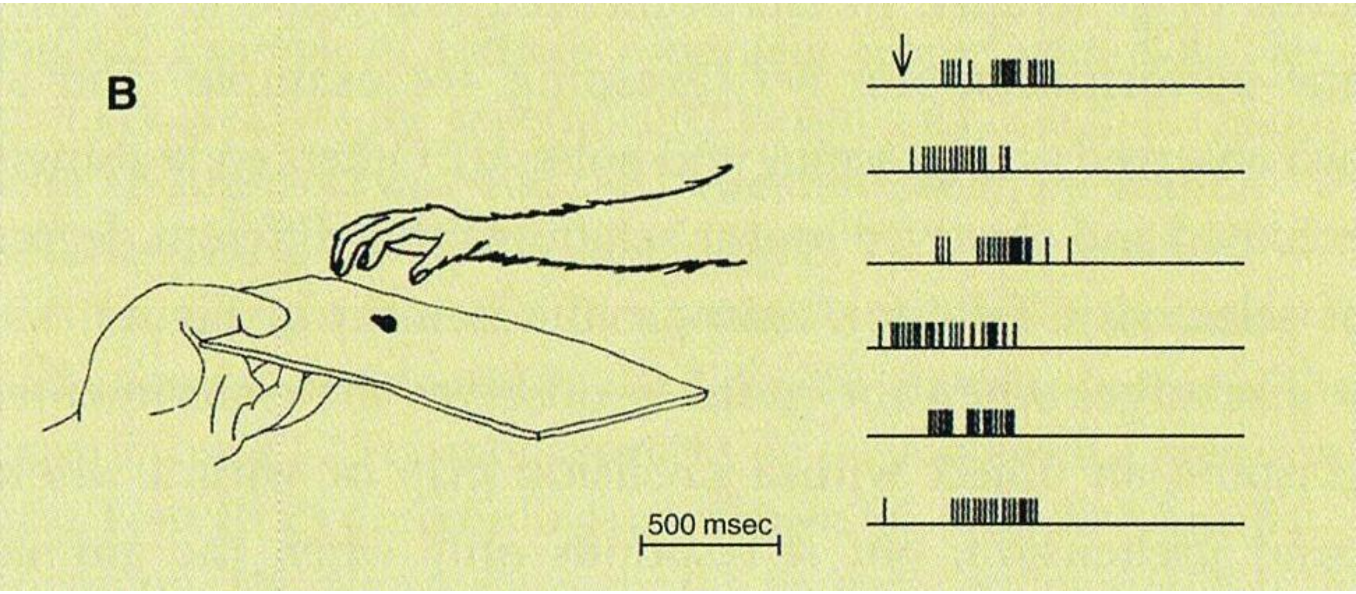
Overview of our approach



Scene parsing

How to define actions and events?

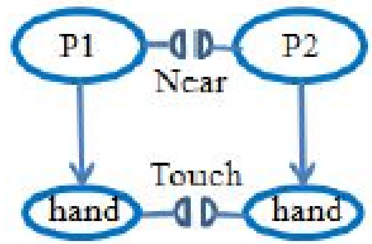
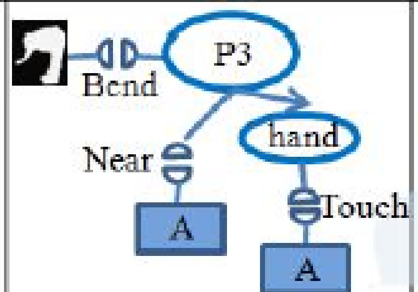
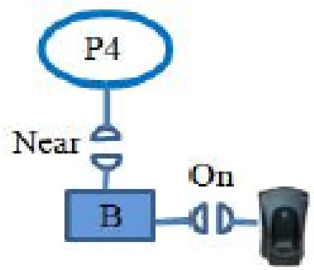
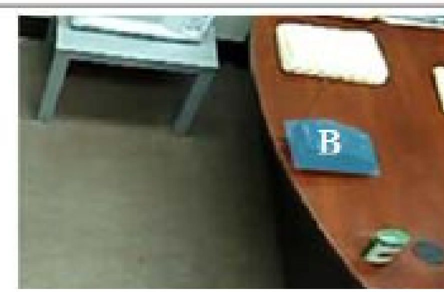
Some neurons in the pre-motor area encode actions



Mirror neurons firing when performing action or seeing other people performing the action

Gallese et al 96, Rizzolatti et a. 96

Actions = Spatiotemporal relations between body parts and objects in the scene

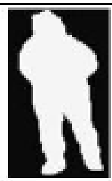
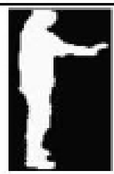
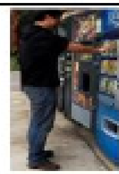
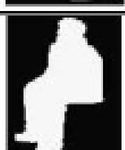
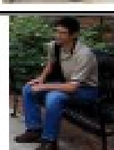
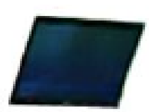
Atomic Actions	Fluents	Symbols: Foreground	Symbols: Background	Examples
Shake Hands(P1,P2)	Near(P1,P2) And Touch (P1.hand, P2.hand)			
Use Dispenser(P3)	Bend(P3) and Near(P3, A) And Touch(P3.hand, A)			
Pick up Phone(P4)	Touch(P4, B) And On(B)			

Some of the learned atomic actions by pursuing the co-occurrence of relations.

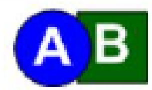
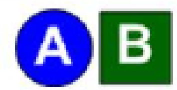
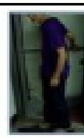
M.T. Pei, Z.Z. Si, B.Yao, and S.C. Zhu ,"Video Event Parsing and Learning with Goal and Intent Prediction," 2012

Actions = Spatiotemporal relations between body parts and objects in the scene

Unary Relations

Status of person	Symbols	Examples	Status of objects	Examples
Stand(P1)			On (phone)	
Stretch(P1)			Off (phone)	
Bend (P1)			On(screen)	
Sit (P2)			Off(screen)	

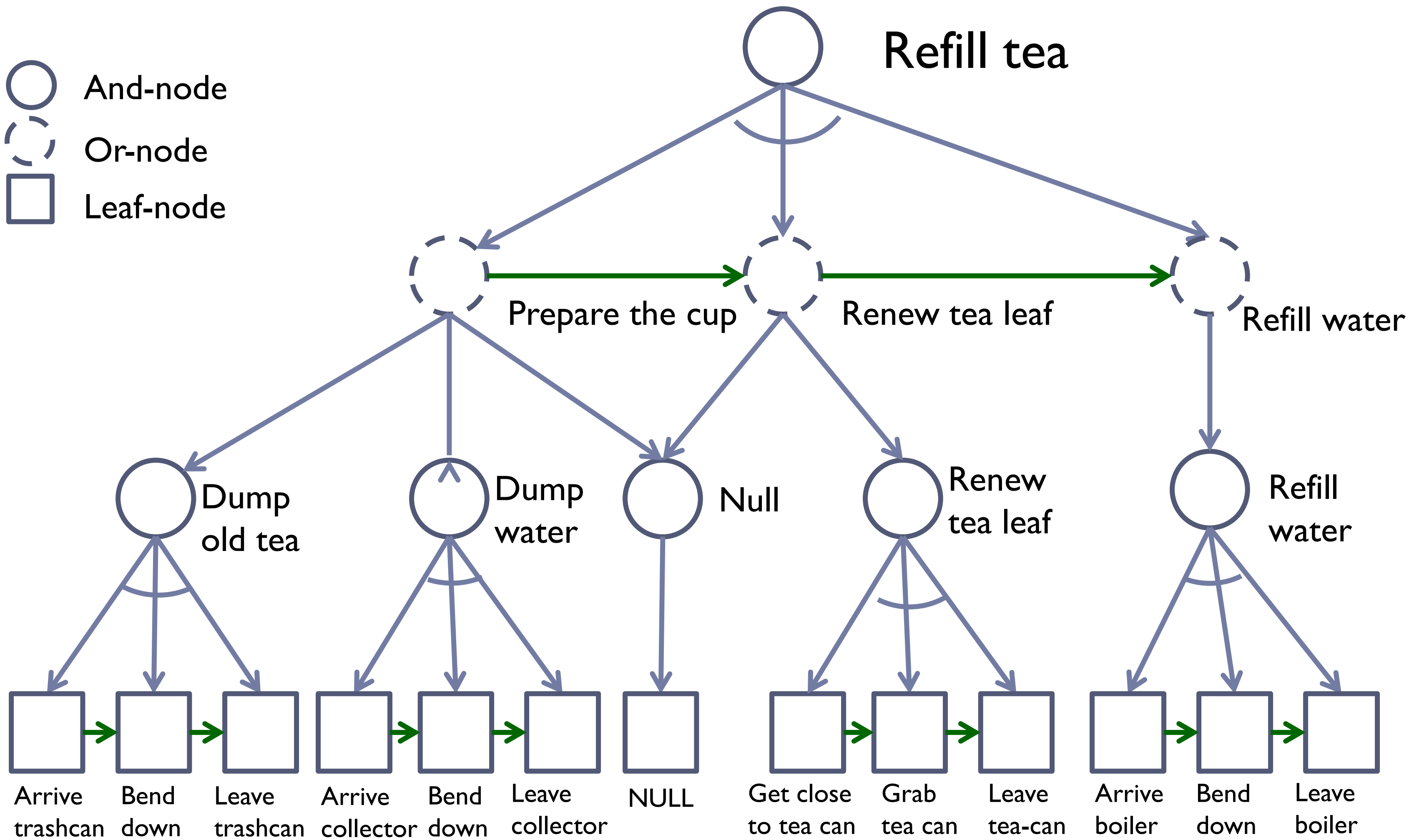
Binary Relations

Binary Fluent (A,B)	Touch (A,B)	Near (A,B)	Occlude (A,B)	In(A,B)
Symbols				
Examples				

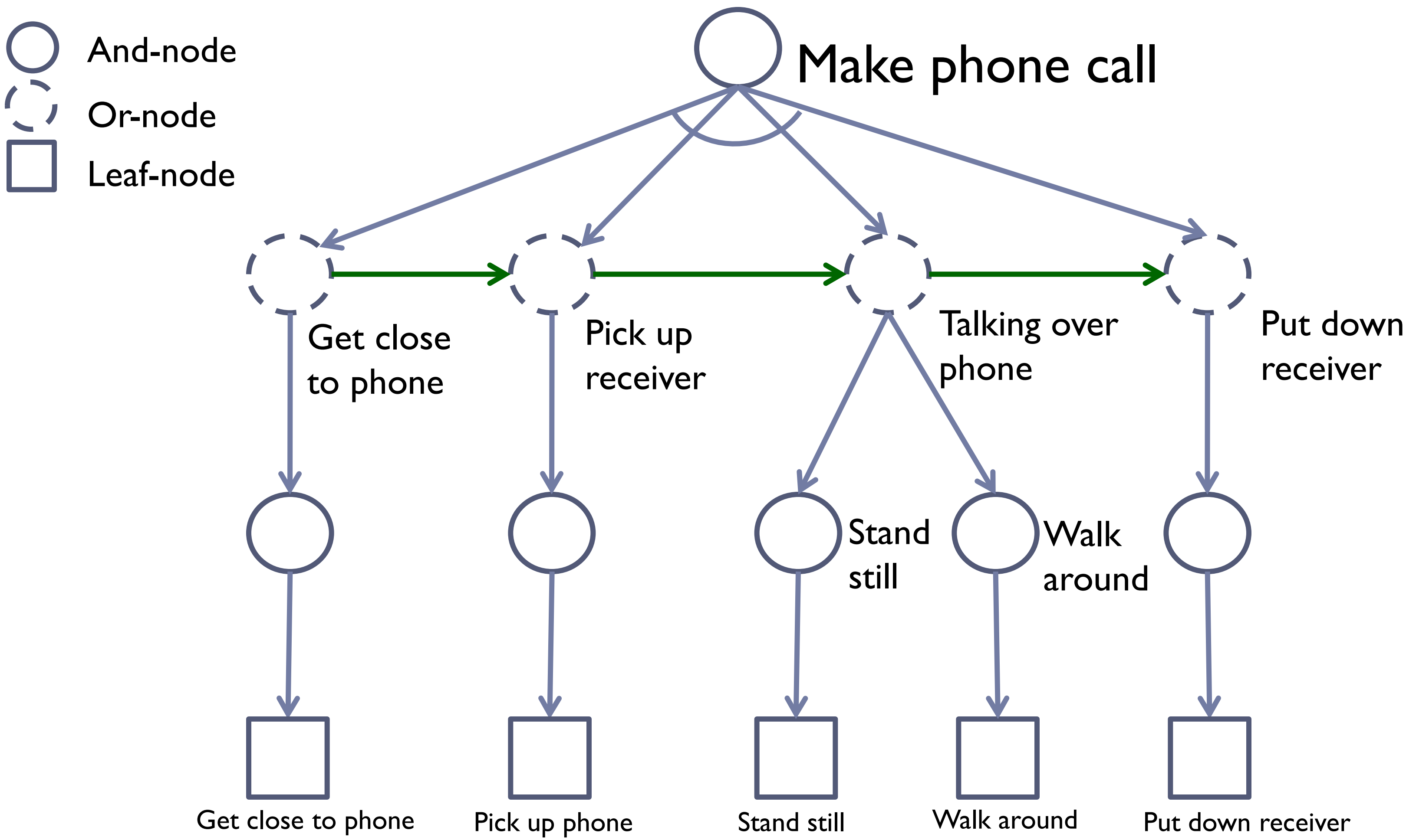
High-order relations:

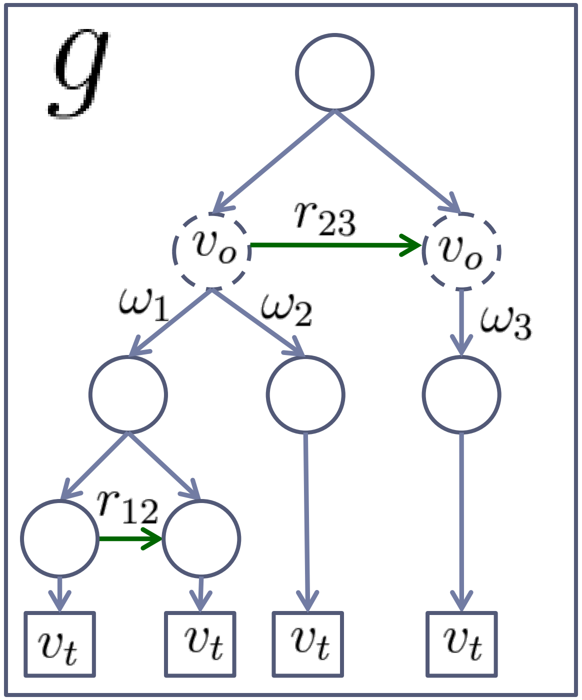
- E.g., surrounded by

Event as temporal And-Or-Graph



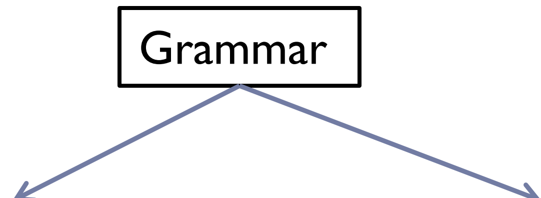
Event as temporal And-Or-Graph



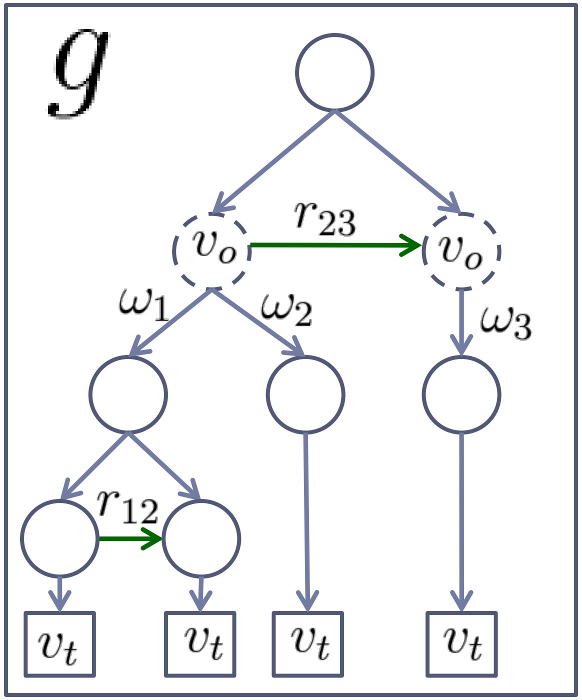


Formulation

$$p(g) = \frac{1}{Z} exp\{score(g)\}$$



$$score(g) = \underbrace{\sum_{v_t \in T(g)} \lambda_{v_t} \alpha(v_t)}_{\text{Data term}} + \underbrace{\sum_{v \in V_o(g)} \lambda_v \omega(v)}_{\text{Or node Frequency term}} + \underbrace{\sum_{(i,j) \in E(g)} \lambda_{ij} r_{ij}(v_i, v_j)}_{\text{Temporal Relations}}$$



Formulation

$$p(g) = \frac{1}{Z} exp\{score(g)\}$$

Grammar

Data term

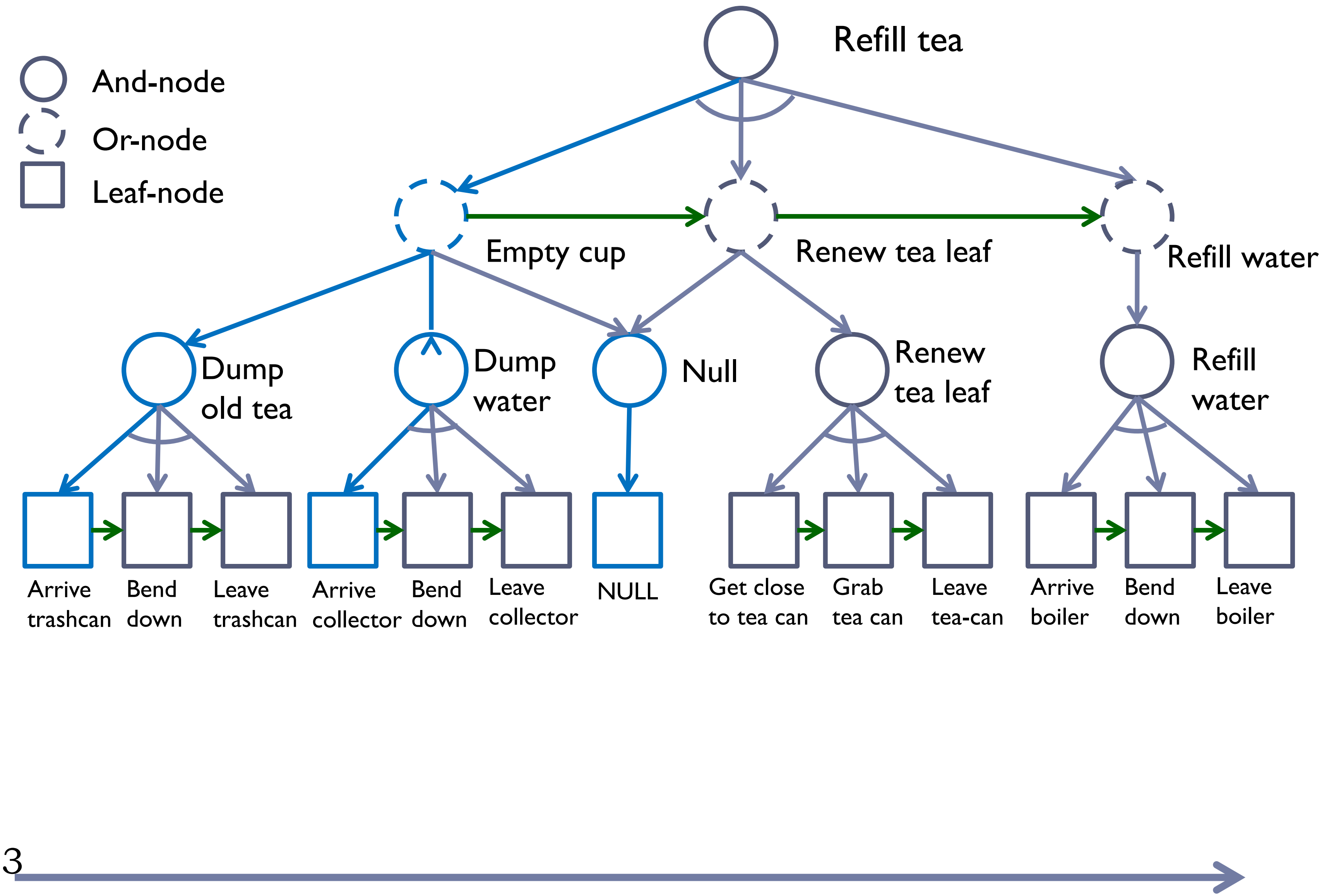
Or node Frequency

Temporal Relations

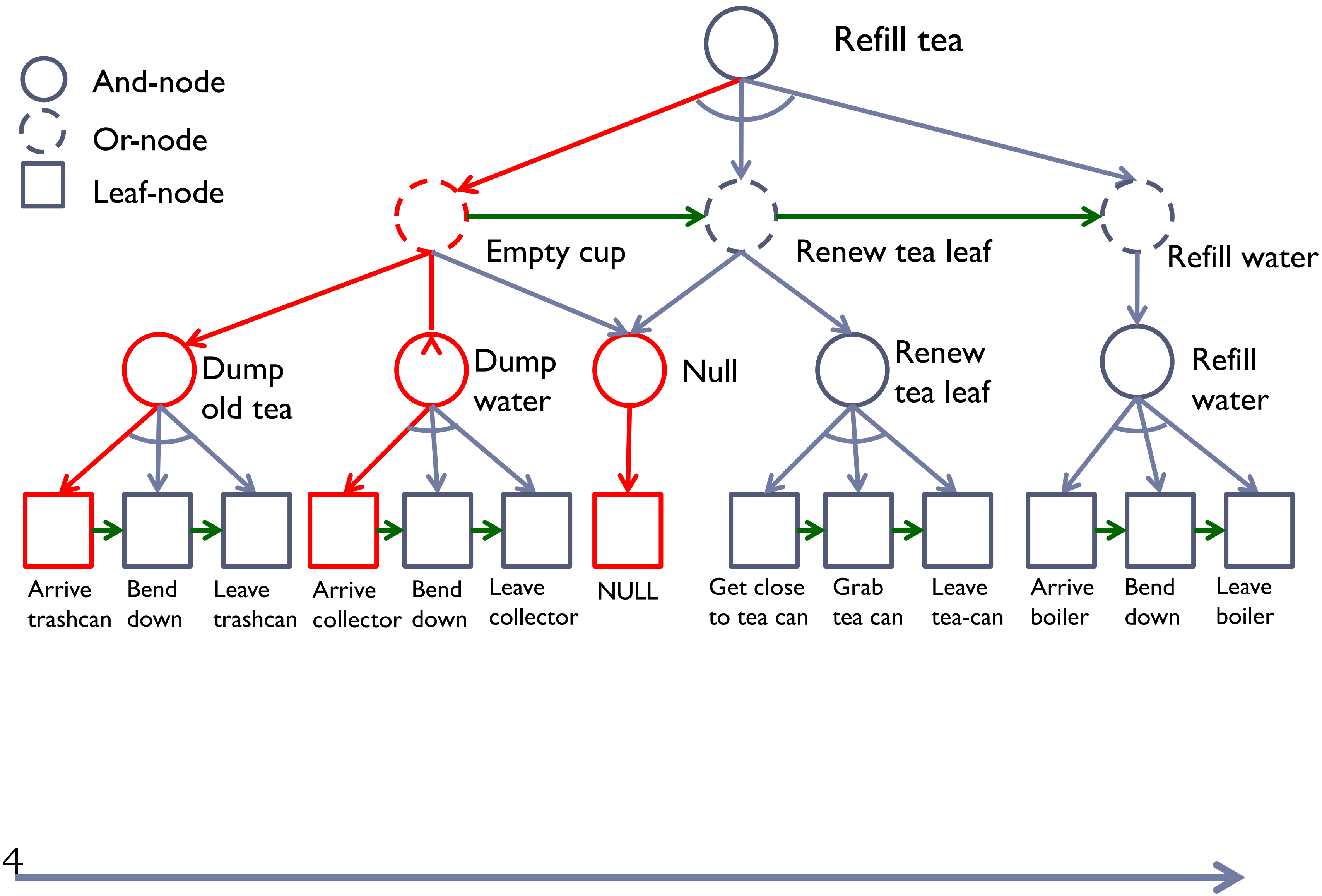
$$score(g) = \sum_{v_t \in T(g)} \lambda_{v_t} \alpha(v_t) + \sum_{v \in V_o(g)} \lambda_v \omega(v) + \sum_{(i,j) \in E(g)} \lambda_{ij} r_{ij}(v_i, v_j)$$

$$\alpha(v_t) = \sum_{i \in \mathcal{F}} \beta_i, h_i(v_t) - dist(P_{person}, P_{obj})$$

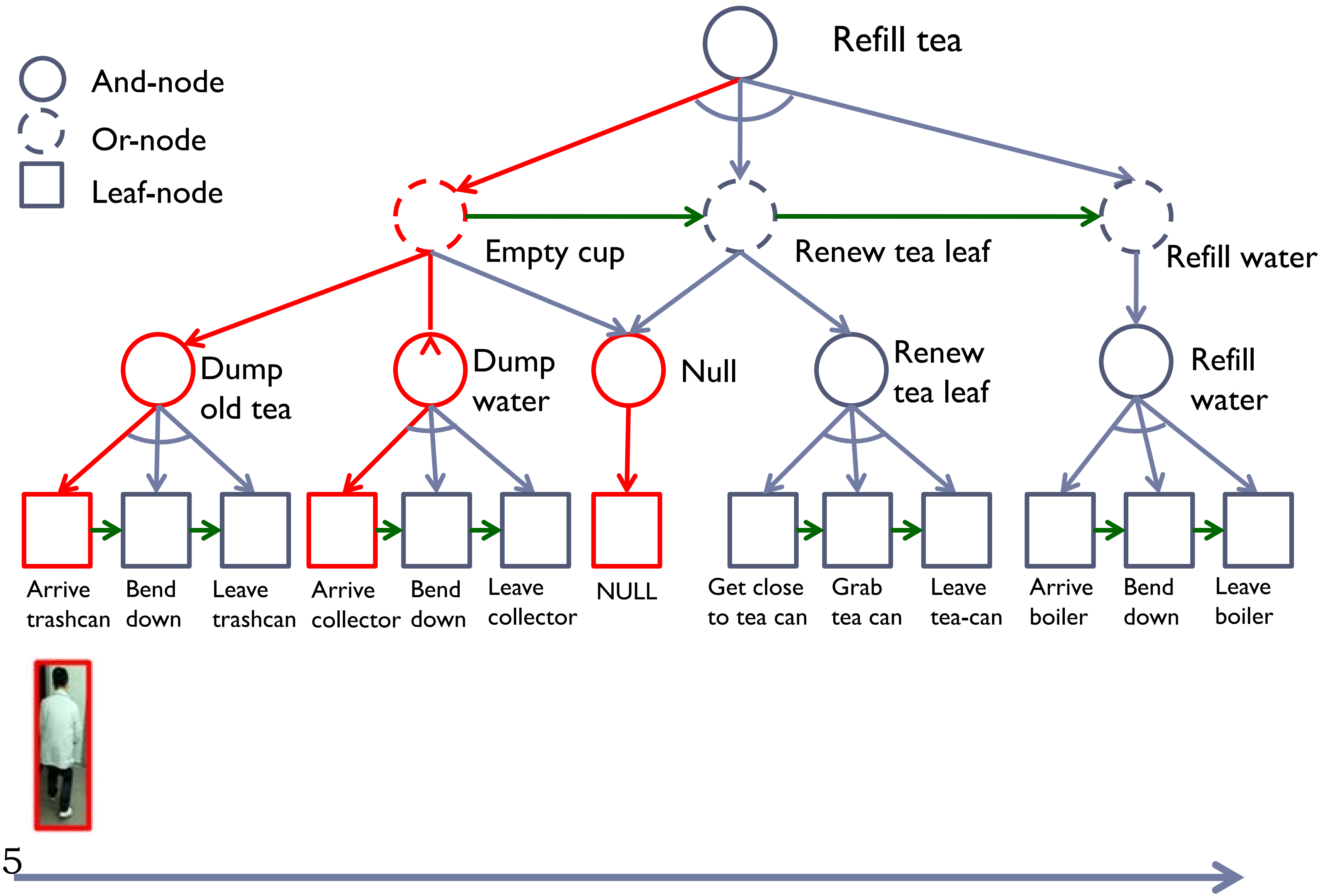
Parsing process (Earley Parser [Earley 1970])



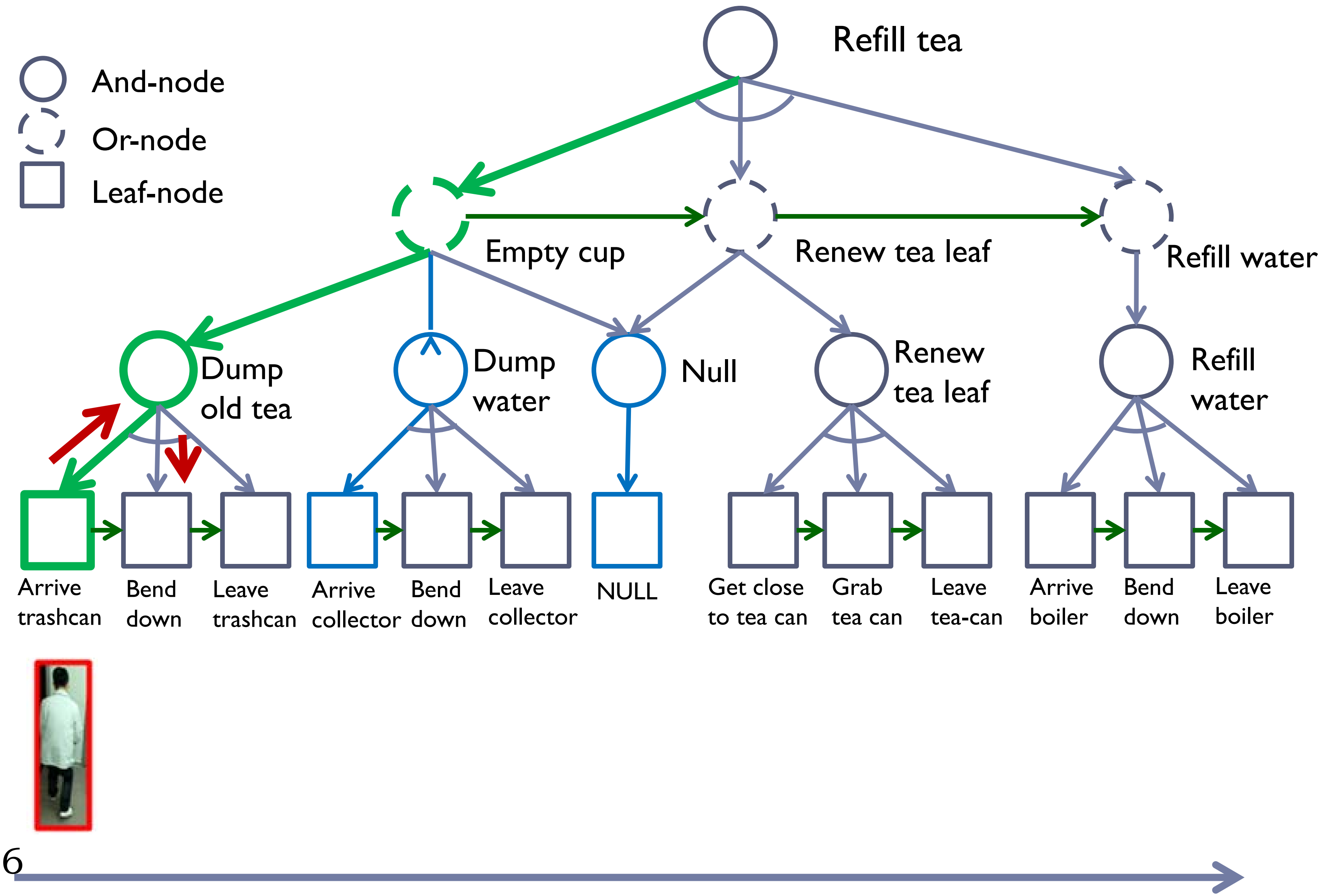
Parsing process (Earley Parser [Earley 1970])



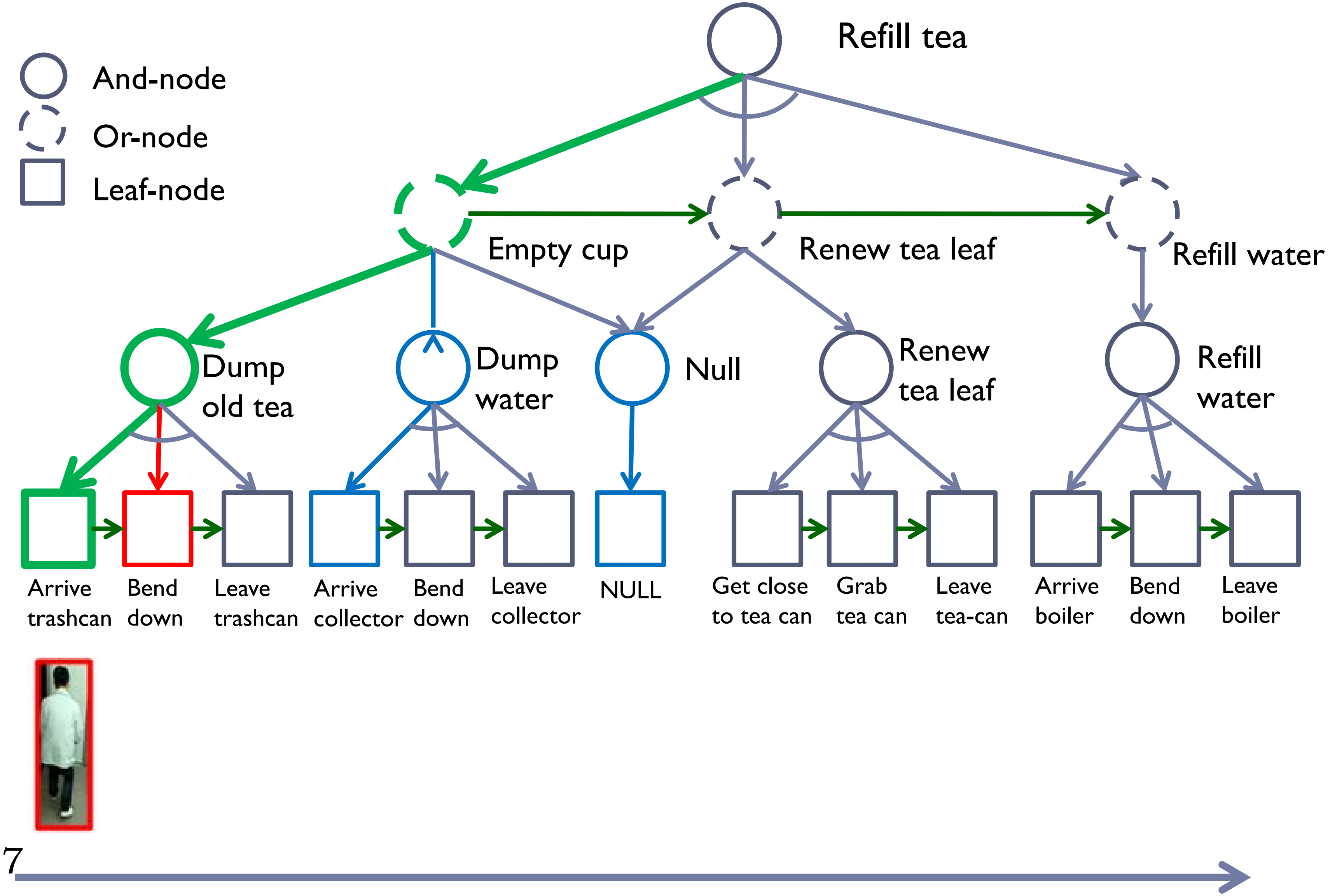
Parsing process (Earley Parser [Earley 1970])



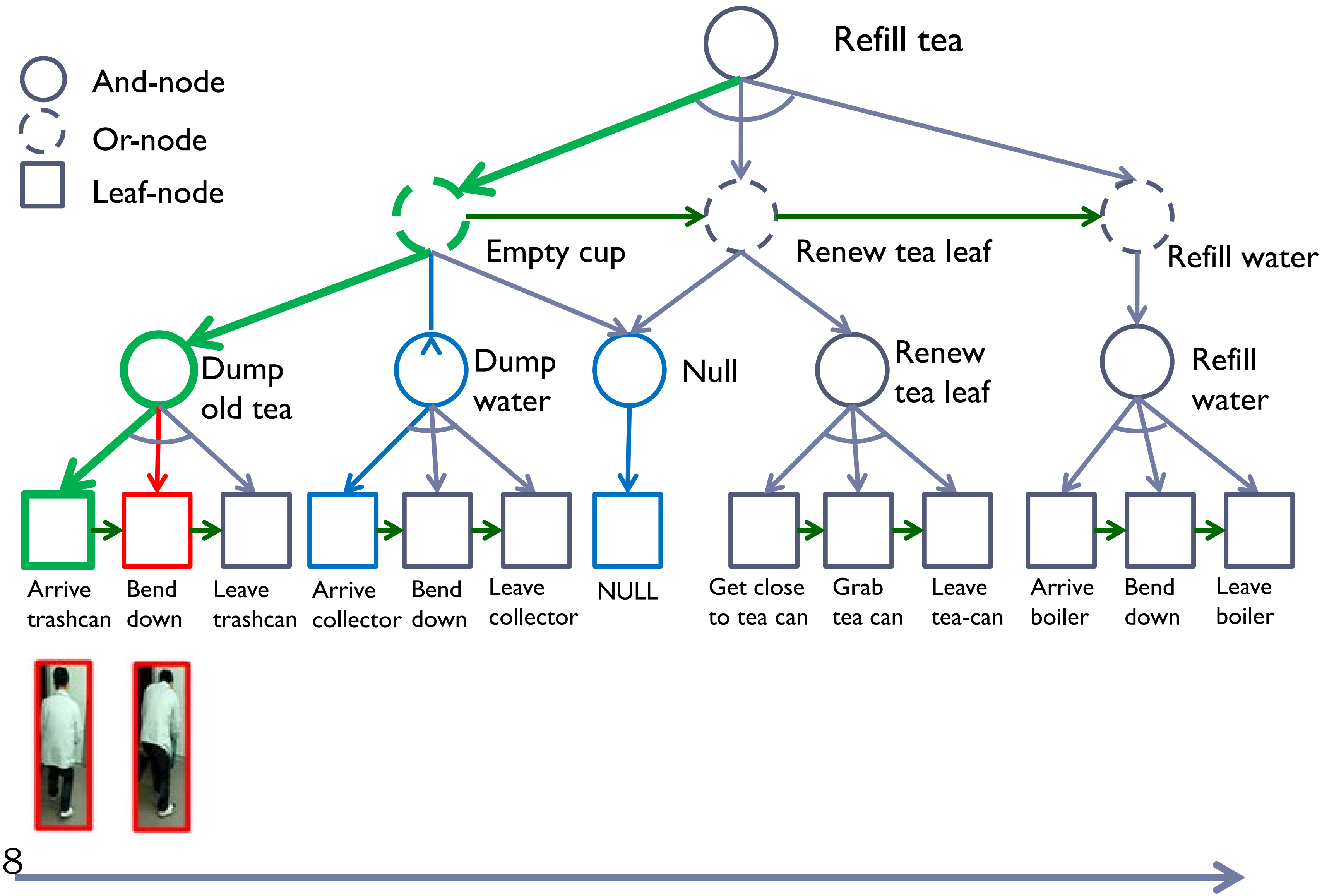
Parsing process (Earley Parser [Earley 1970])



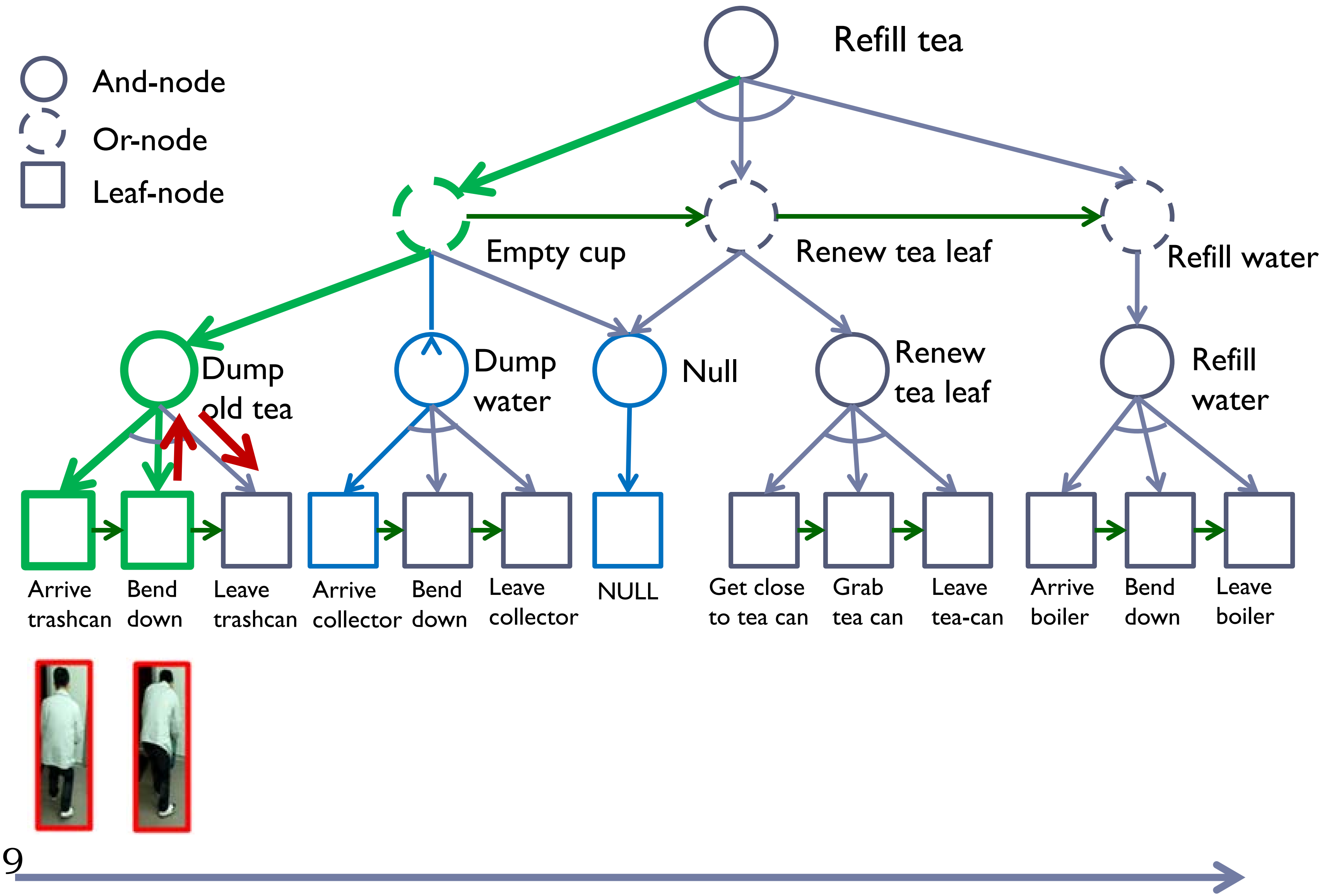
Parsing process (Earley Parser [Earley 1970])



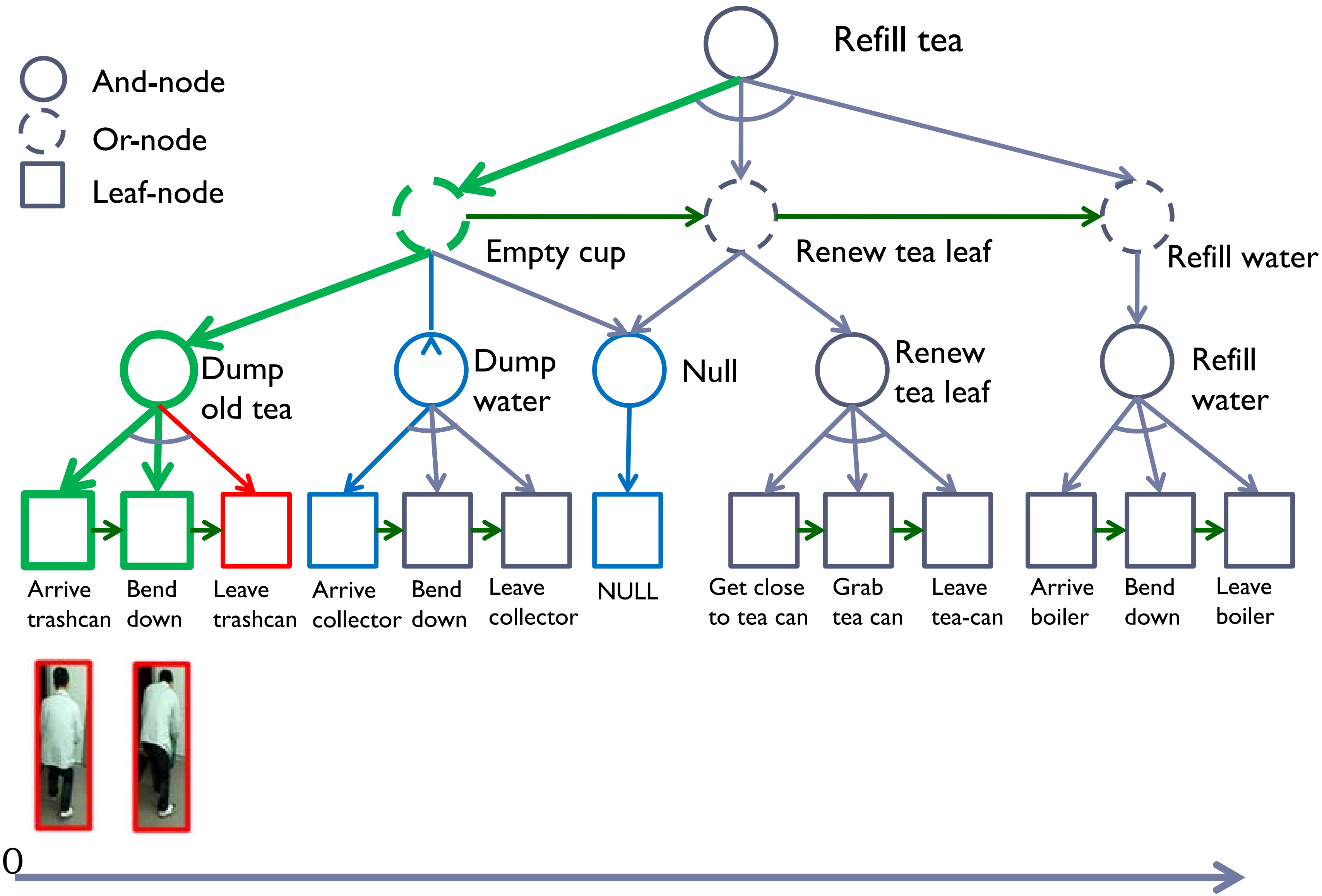
Parsing process (Earley Parser [Earley 1970])



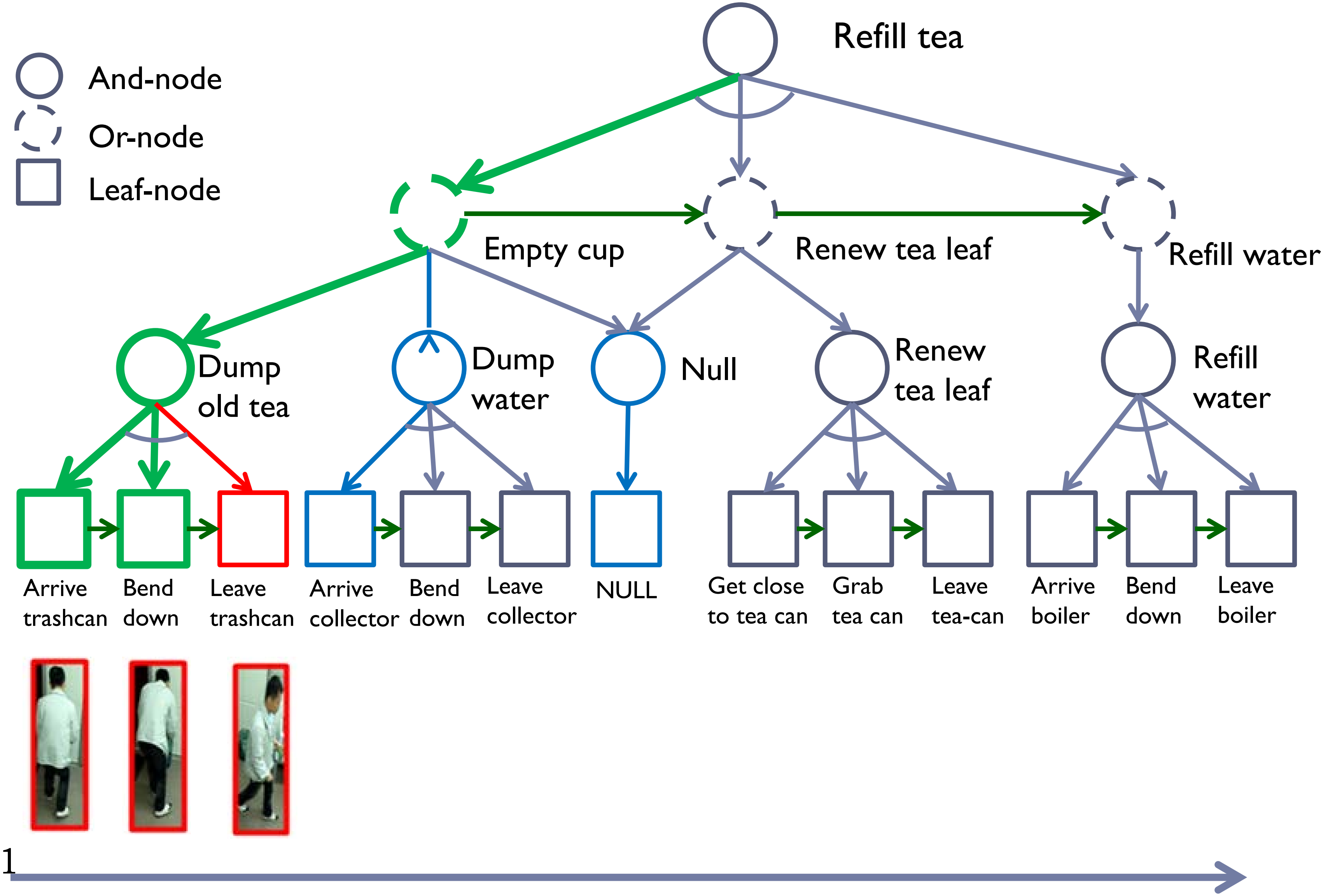
Parsing process (Earley Parser [Earley 1970])



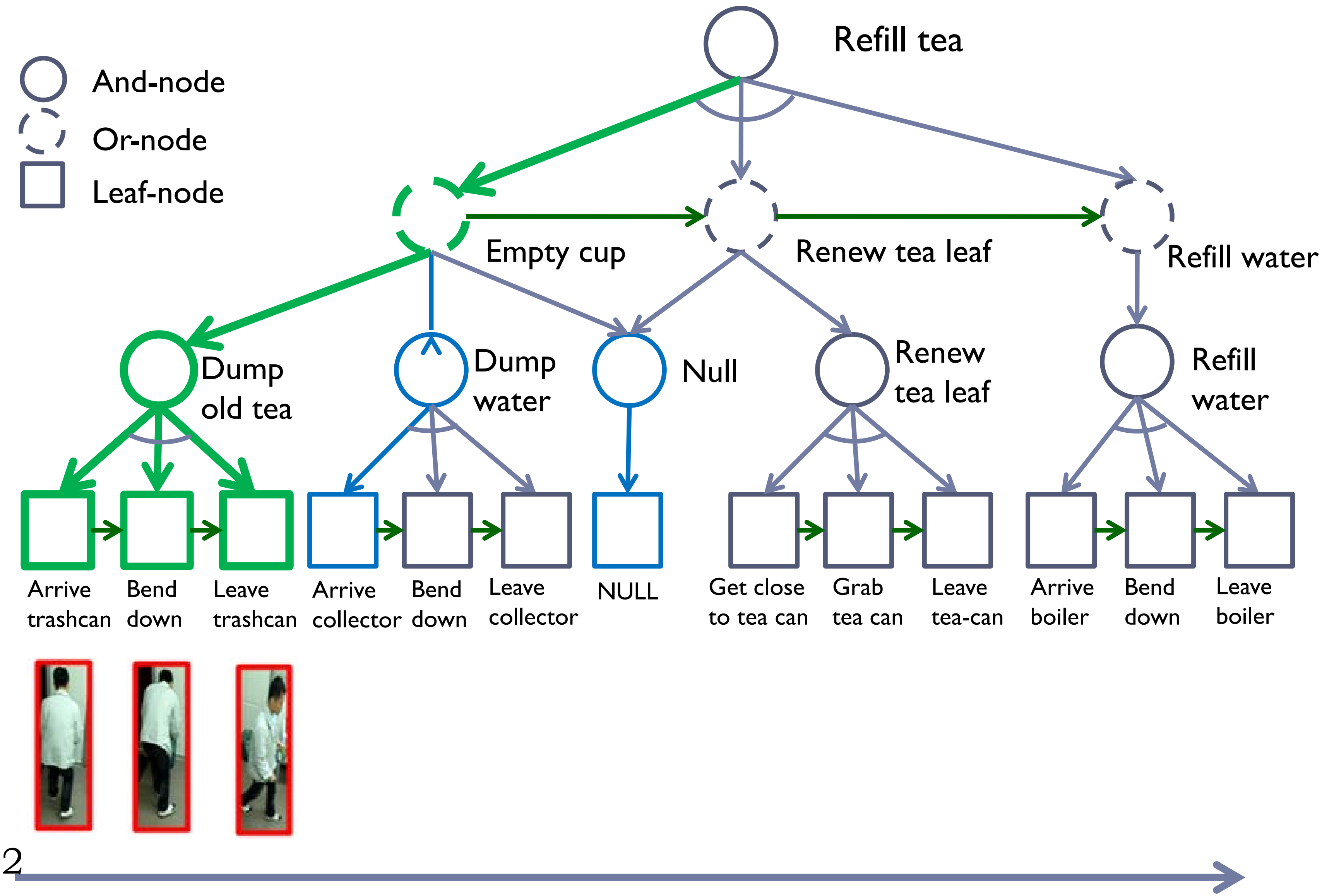
Parsing process (Earley Parser [Earley 1970])



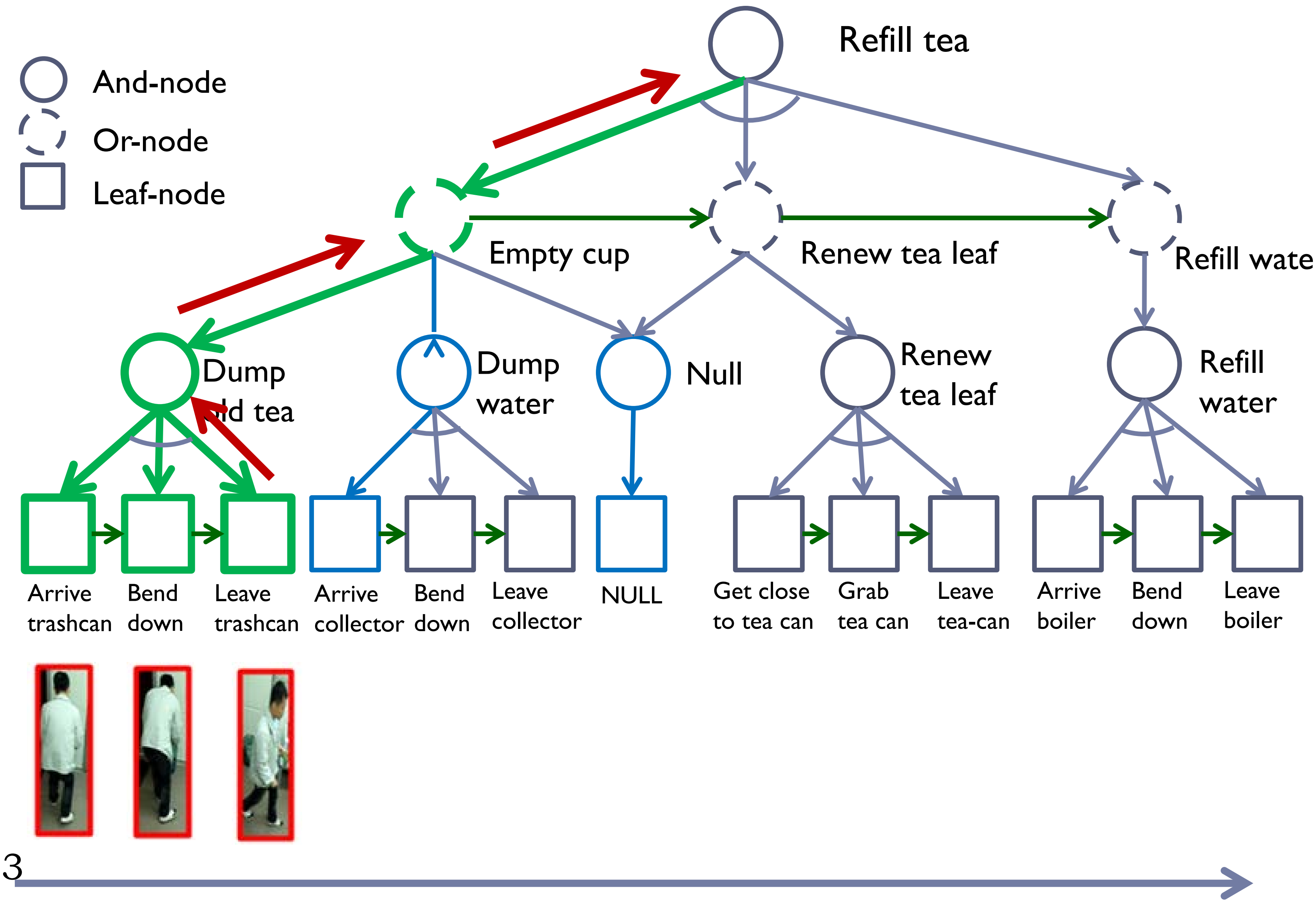
Parsing process (Earley Parser [Earley 1970])



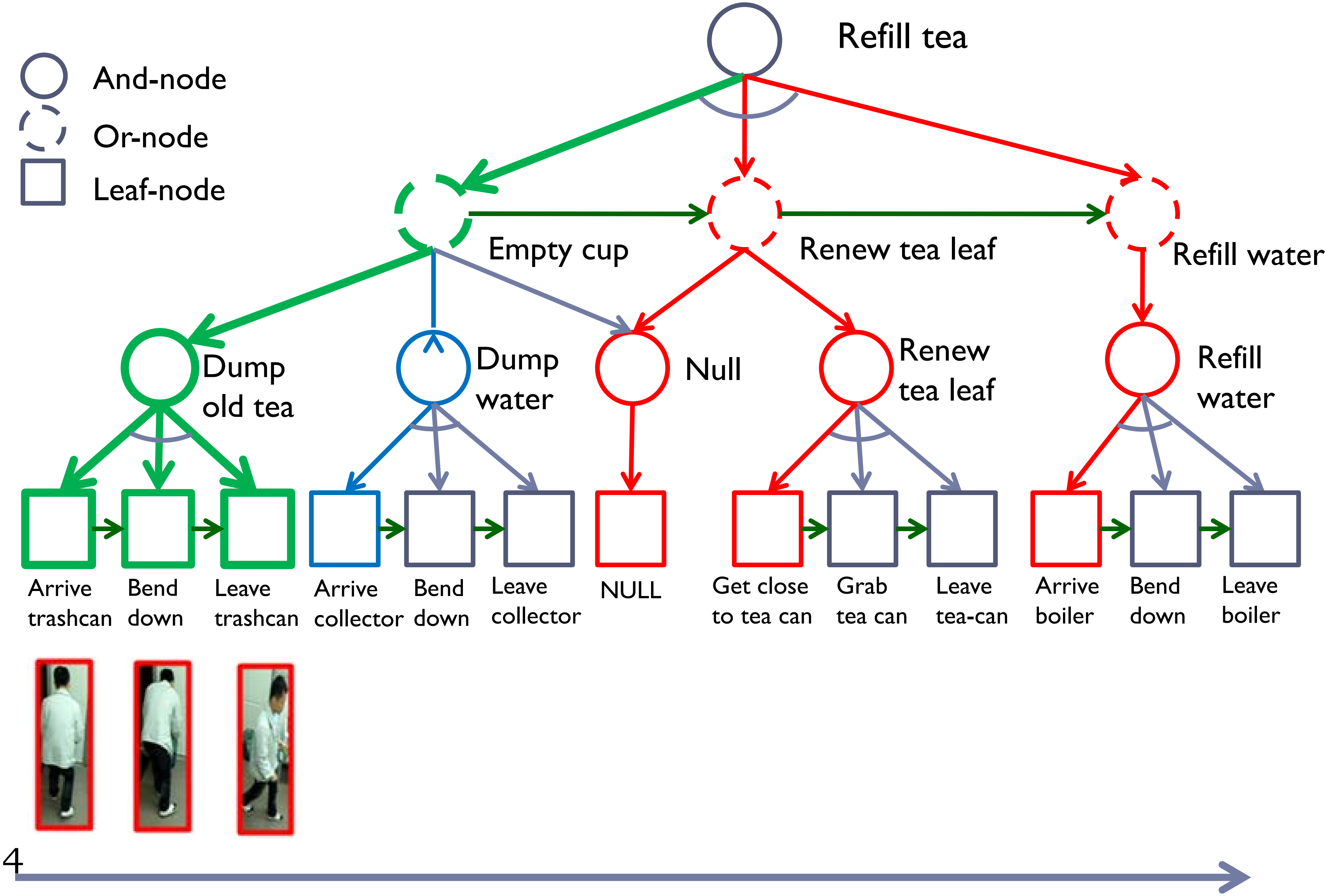
Parsing process (Earley Parser [Earley 1970])



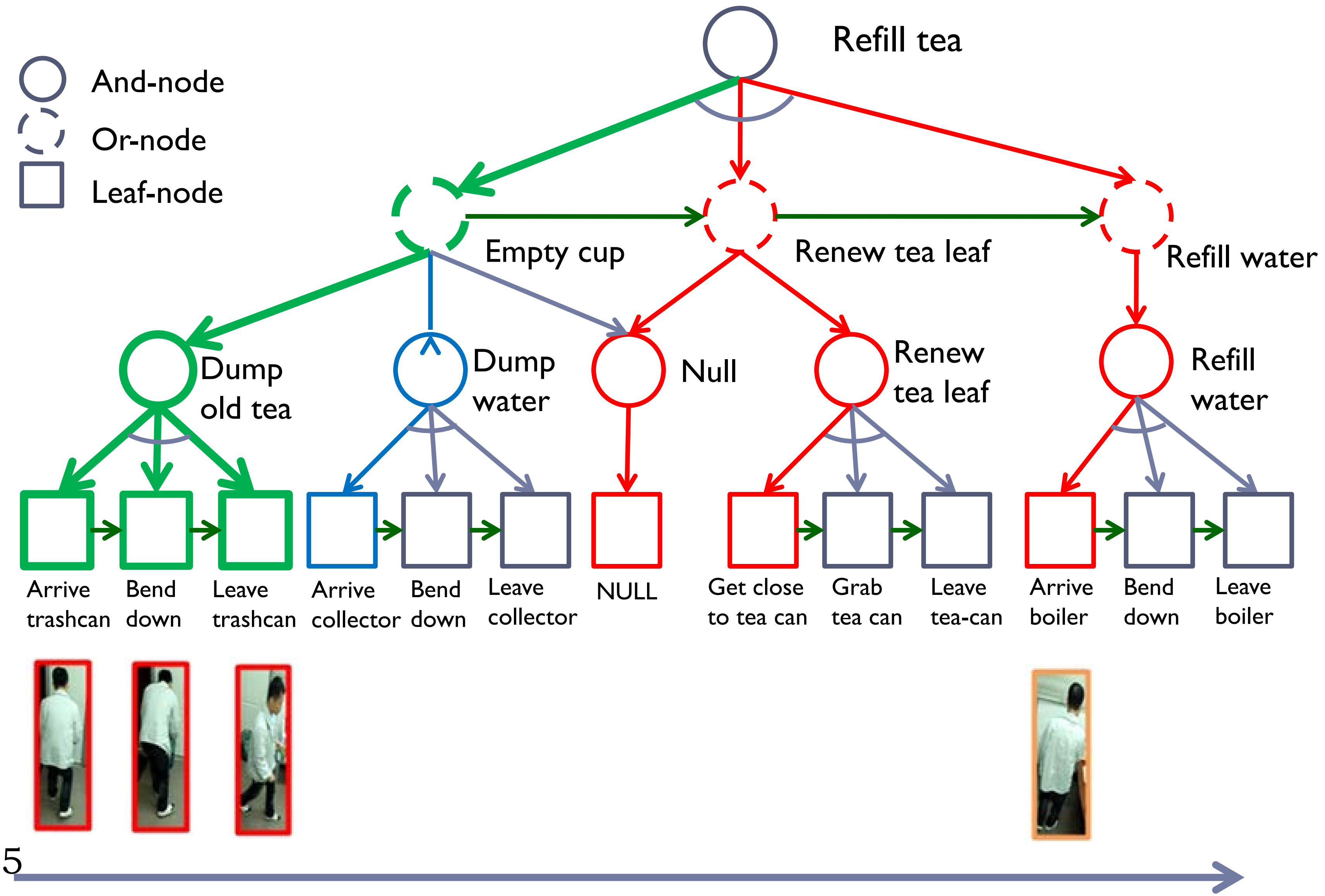
Parsing process (Earley Parser [Earley 1970])



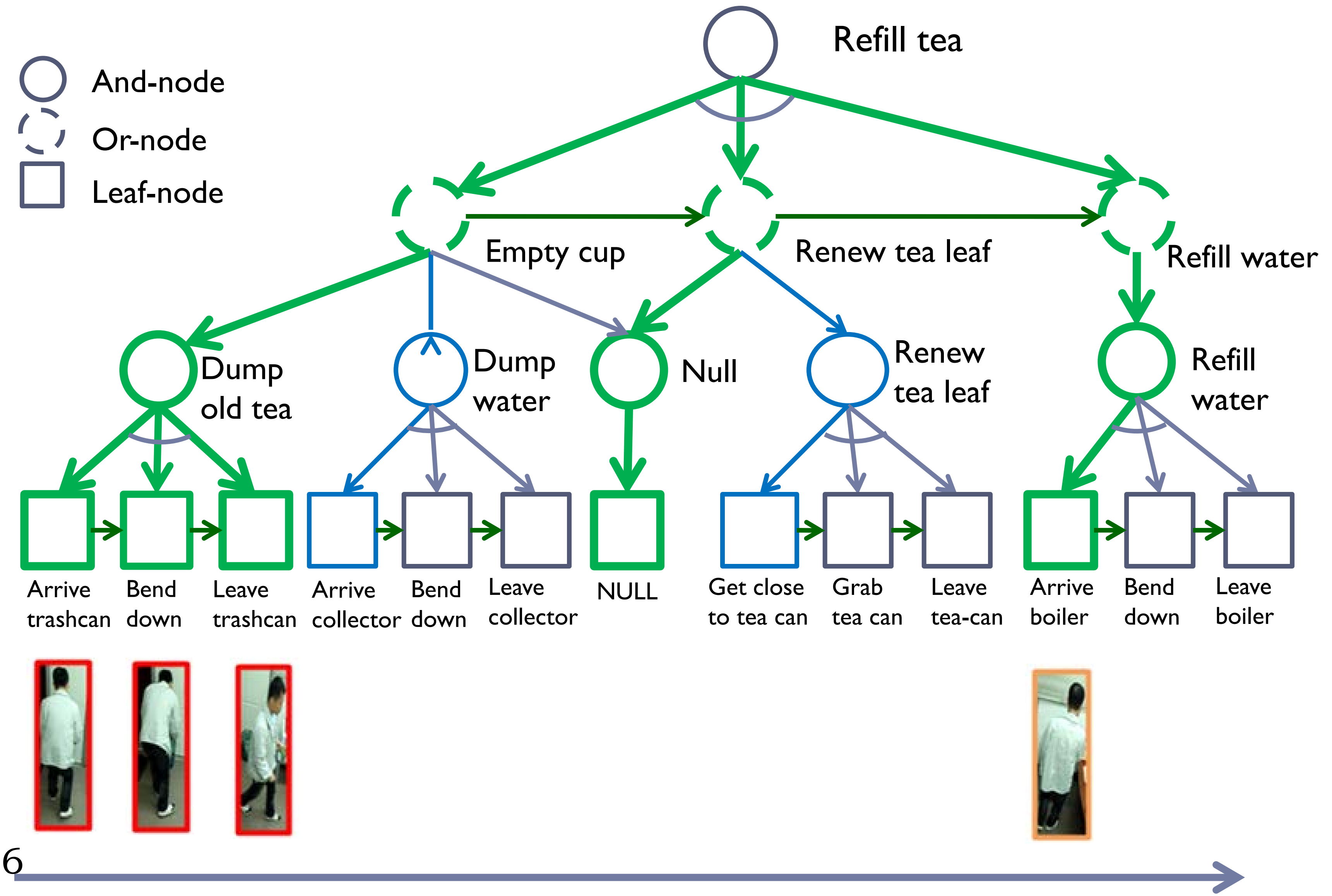
Parsing process (Earley Parser [Earley 1970])



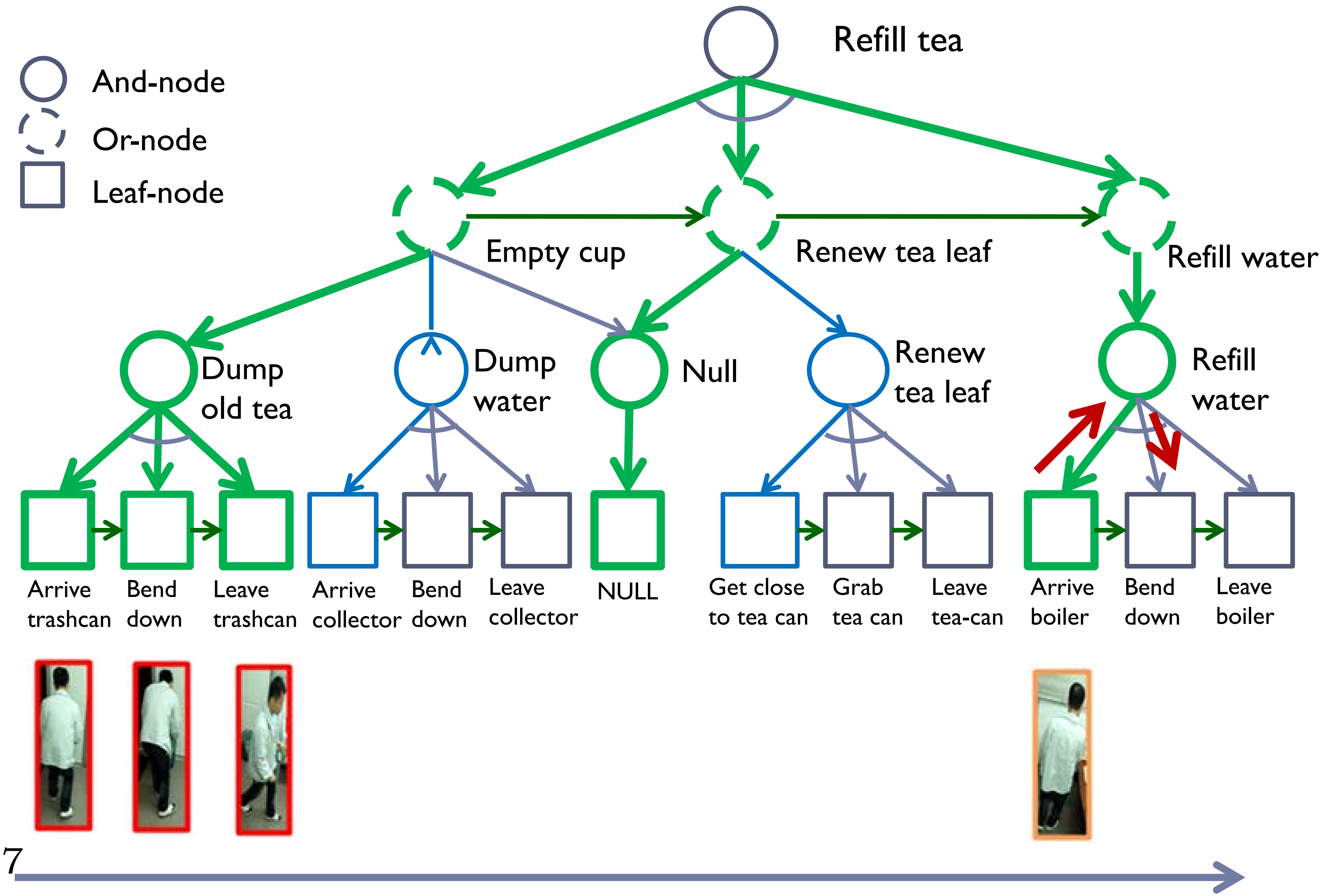
Parsing process (Earley Parser [Earley 1970])



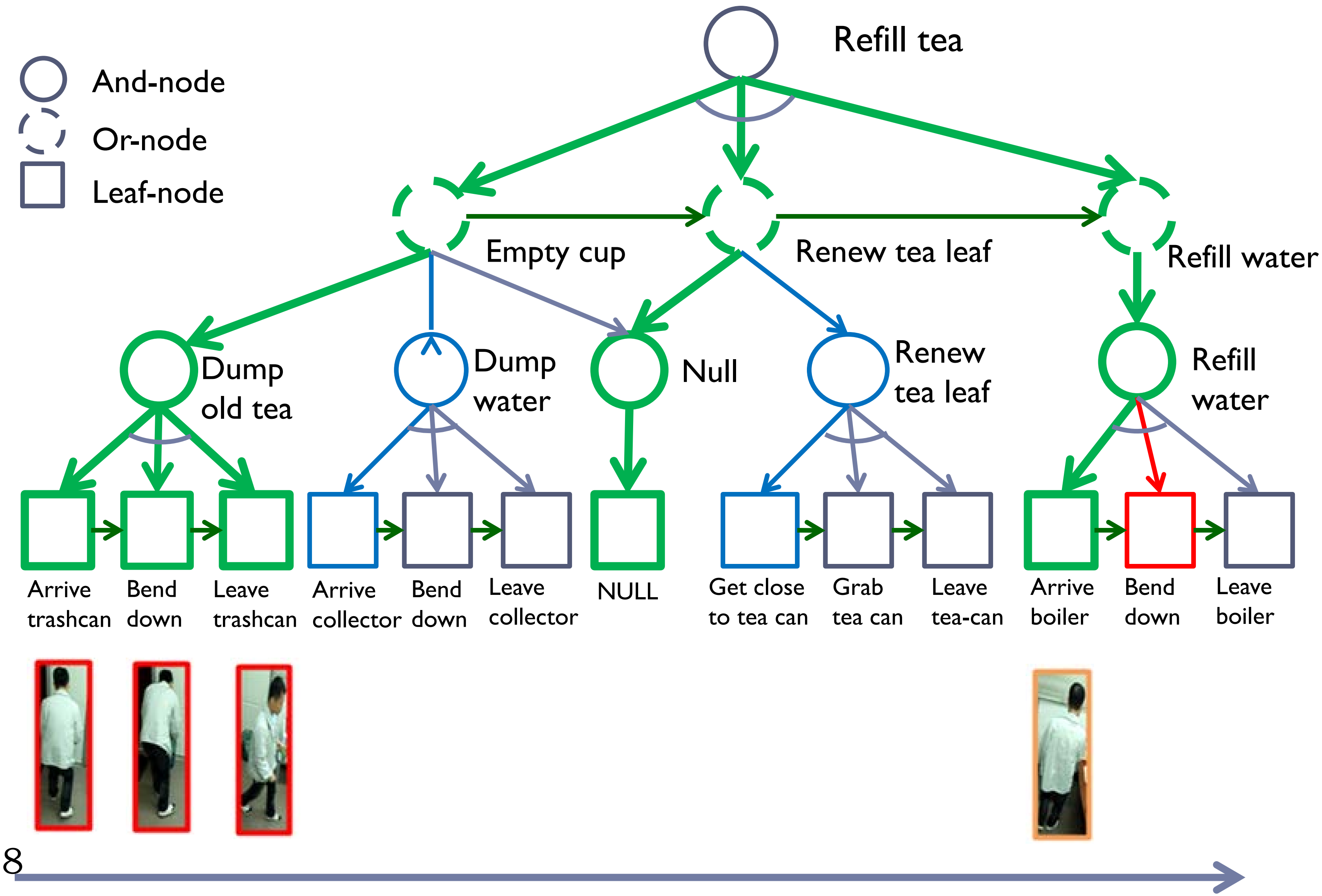
Parsing process (Earley Parser [Earley 1970])



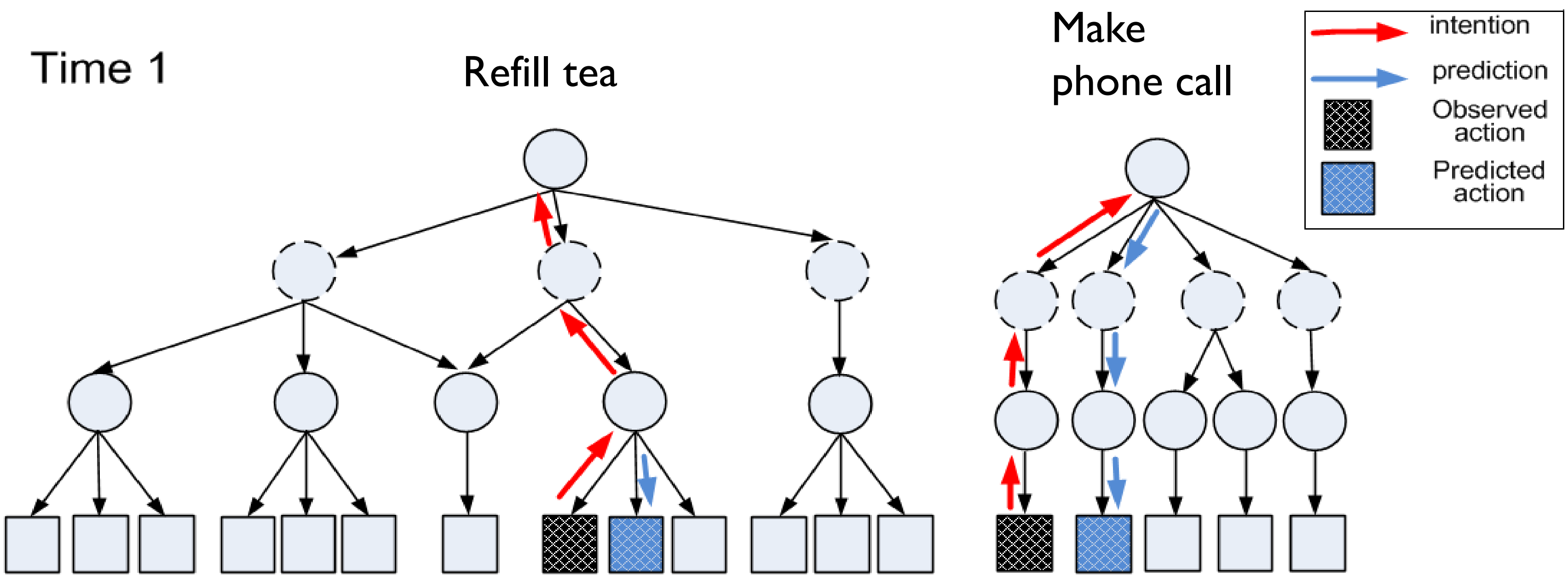
Parsing process (Earley Parser [Earley 1970])



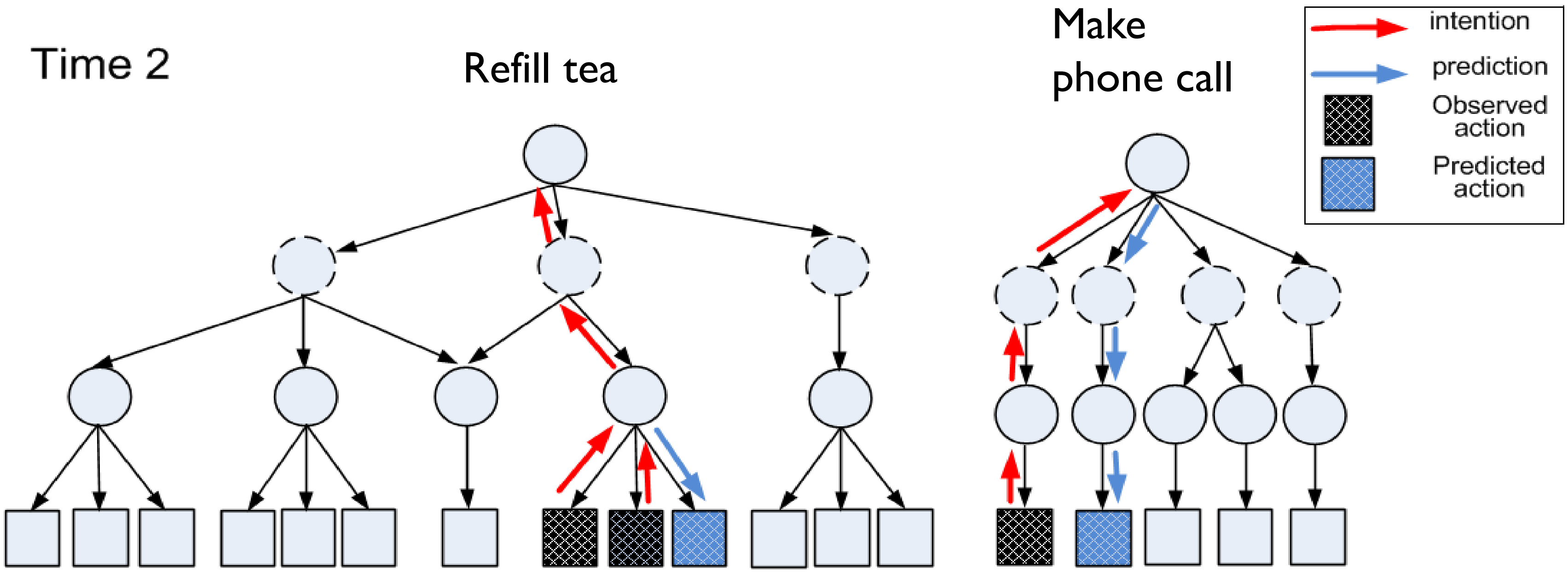
Parsing: A modified Earley parser [Earley 1970]

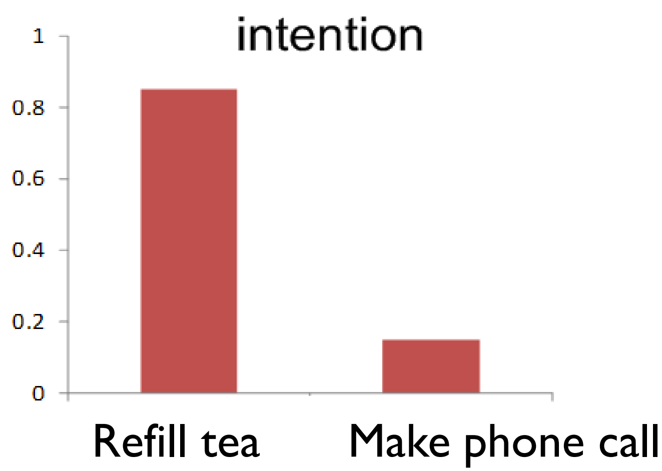


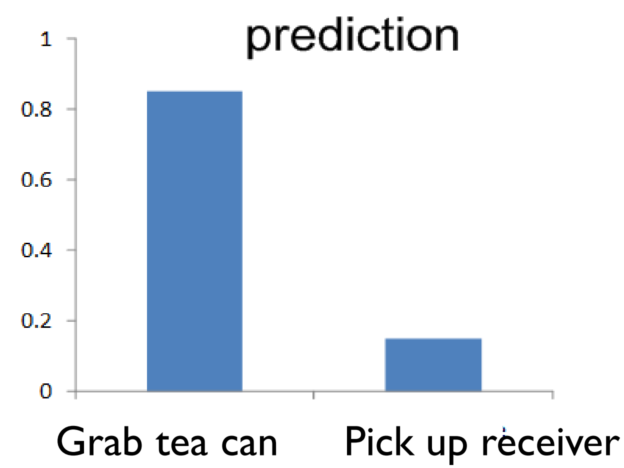
Intention and prediction



Intention and prediction





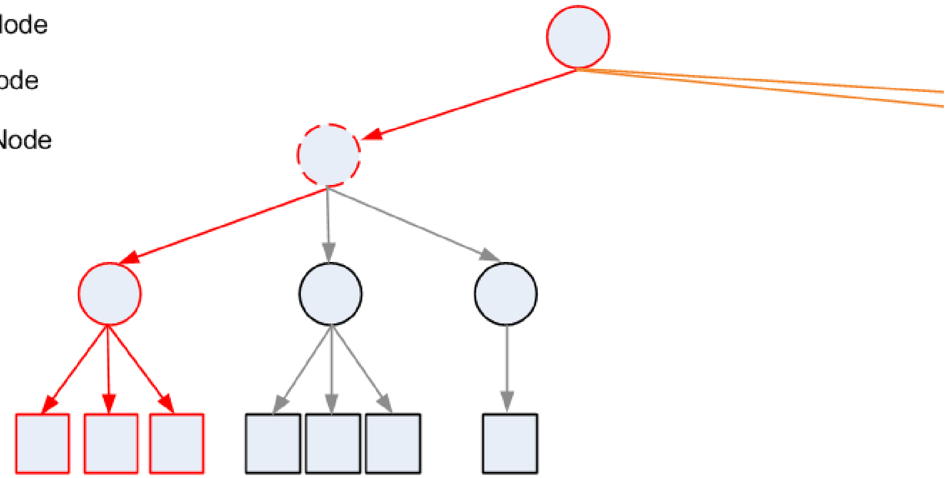


Handle event interruption

And-Node

Or-Node

Leaf-Node



First Partial parse tree of take water

Parse tree of take a phone

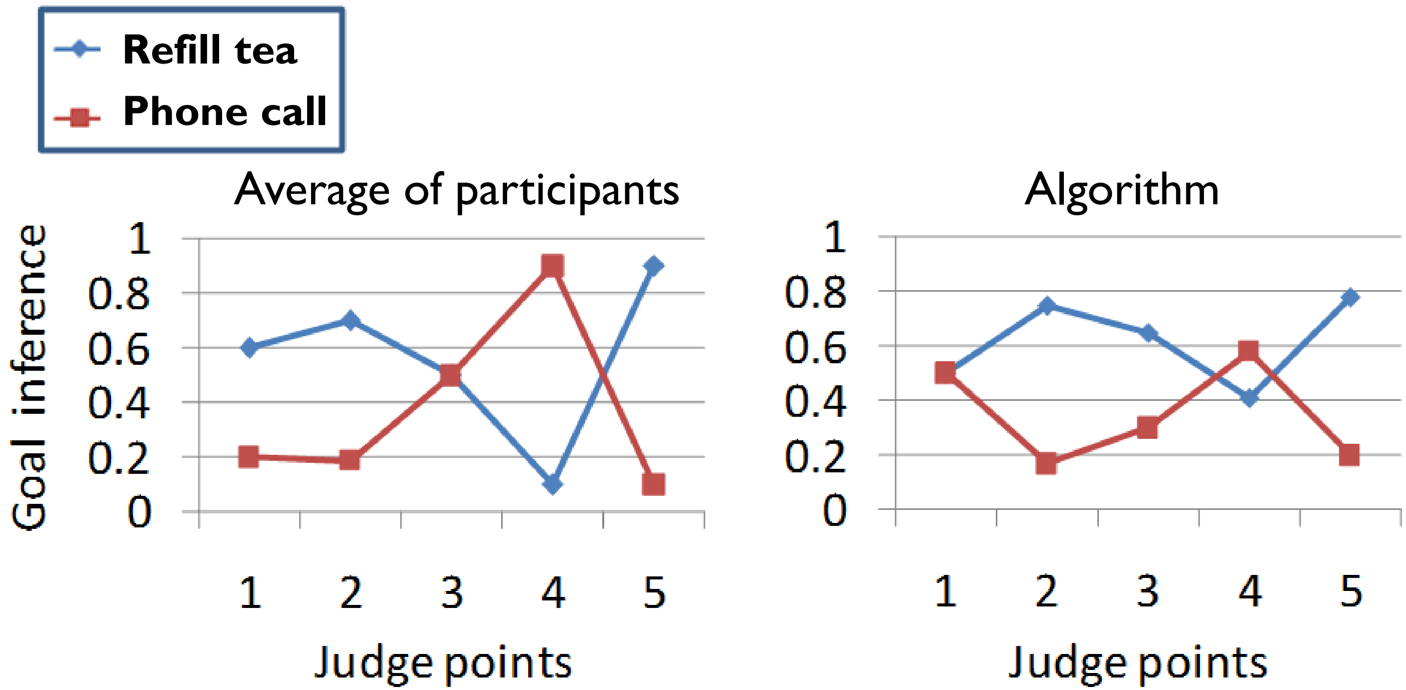
Second Partial parse tree of take water

Observed Data

Demo


And-node
Or-node
Leaf-node
Make a call
Take water
Intention
Prediction
1
0.8
0.6
0.4
0.2
0
take water
make a call
arrive trash can
arrive basin
arrive tea box
arrive wd
arrive phone

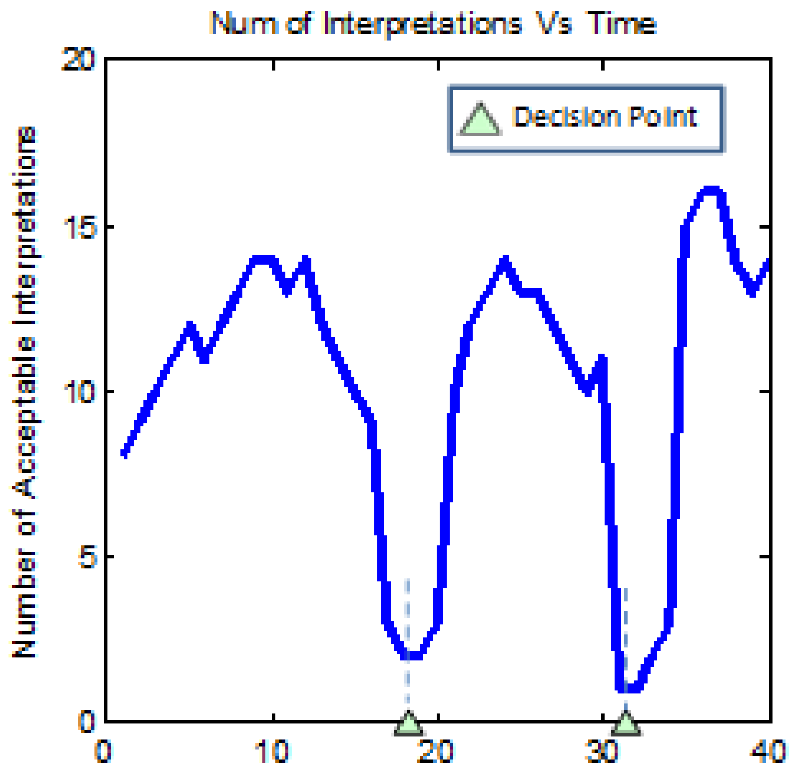
Comparison with human prediction



M.T. Pei, Z.Z. Si, B. Yao, and S.C. Zhu, "Video Event Parsing and Learning with Goal and Intent Prediction," 2012

related work: [Baker, Saxe and Tenenbaum 2009]

Computation complexity of parsing



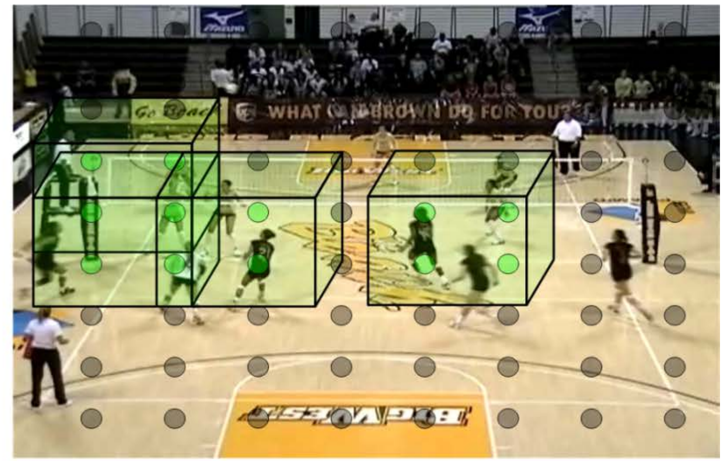
Pickup phone

Reach water boiler

- Initially the number of interpretations above a threshold grows rapidly over time.
- At certain decisive moments, i.e. when informative actions are observed, large number of unlikely interpretation drops below the threshold and hence is pruned.

Weakly Supervised Learning of Temporal AND-OR Graph

Given Input Video



Classify & Localize

Stochastic activity has a random number of:

- actors,
- activity parts,
- spatiotemporal configurations

Amer&Todorovic CVPR12

Examples: Activities with Stochastic Structure

Temporal AND-OR Graph

- AND nodes = Particular space-time configurations
- OR nodes = Alternative configurations
- Terminal nodes = BoWs

Temporal AND-OR Graph

$$S(C) = 0.5(0.4x_1 P_1 + 0.2\overline{x_1}(1 - P_1) + \cdots$$

Posterior:

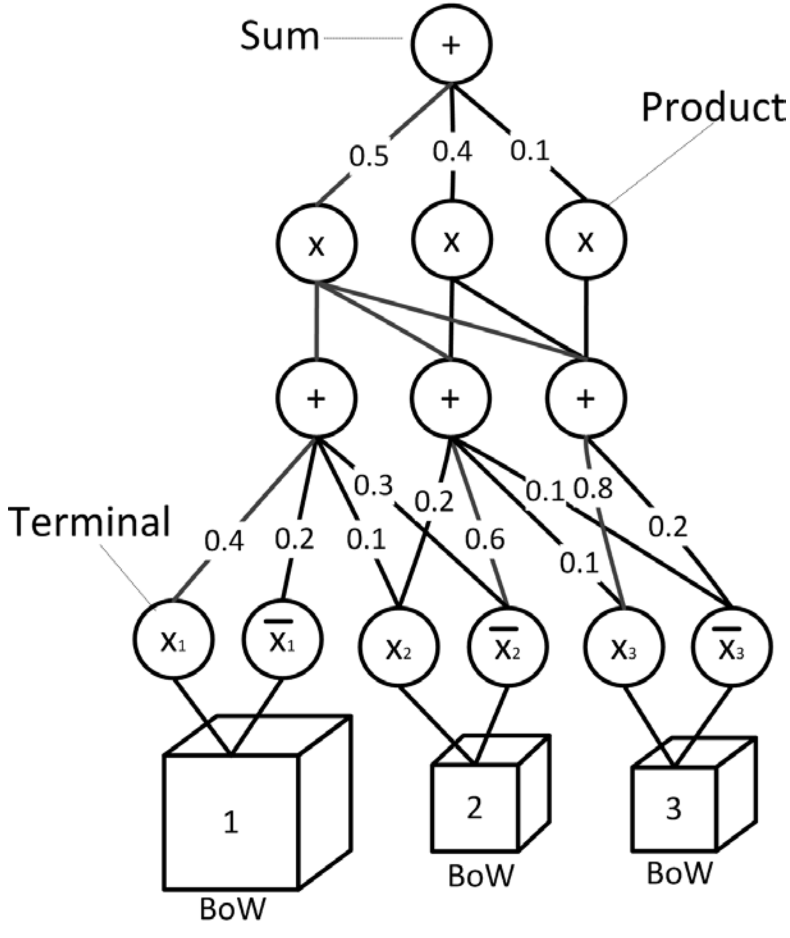
$$P(\boldsymbol{X}|C) = S(C)/S_{\boldsymbol{X}=\boldsymbol{1}}$$

OR nodes:

$$S_i(C) = \sum\nolimits_{j \in i^+} w_{ij} S_j(C)$$

AND nodes:

$$S_k(C) = \prod\nolimits_{l \in k^+} S_l(C)$$



Learning – Variational EM

1. Learn AND-OR graph structure parameters – W
2. Learn Counting Grid parameters – π

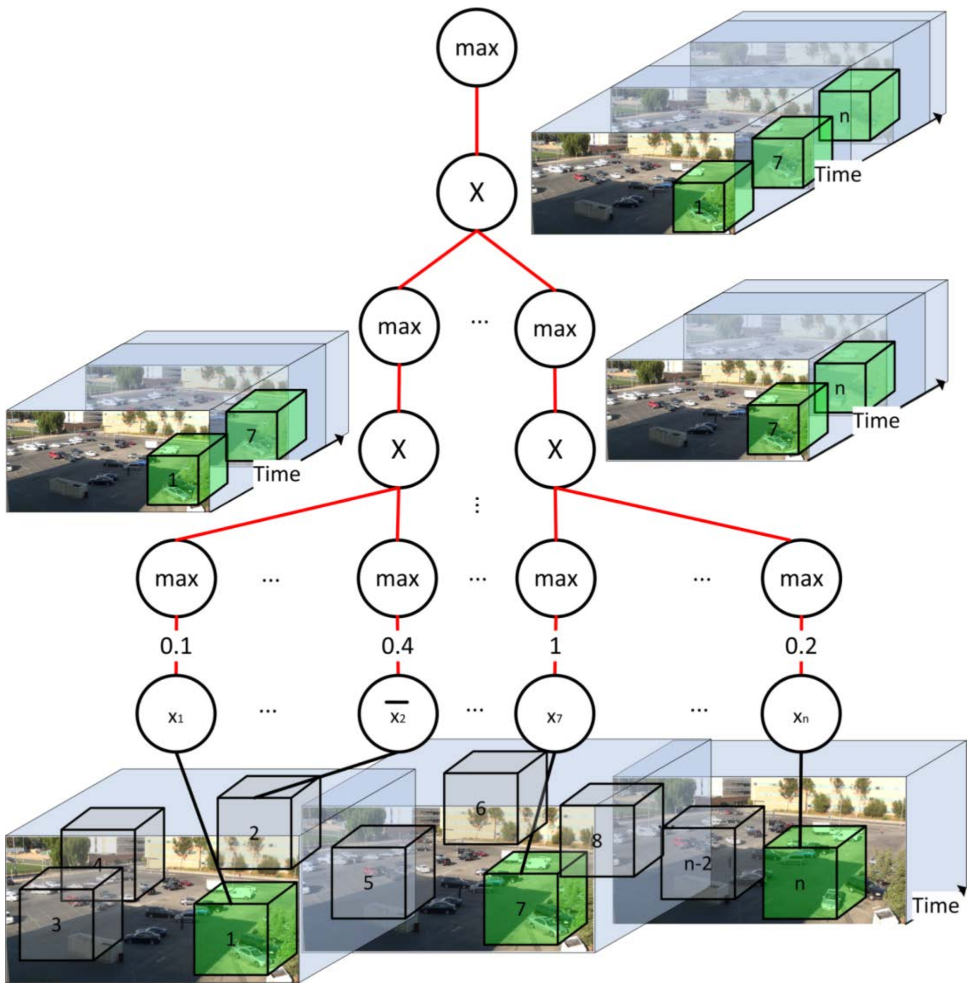
$$
\begin{aligned}
V \;=\; & \sum_t \Big[\sum_b Q_b \log[(w_{ib1} x_b^t - w_{ib2} \overline{x_b^t})/Q_b] \\
& + \sum_b Q_b \sum_z (c_{bz}^t + \theta_z - 1) \log\big[\sum_{u \in H_b} \pi_{uz} \big] \Big],
\end{aligned}
$$

$$
Q_b \propto \exp[\sum_{t,z} (w_{ib1} x_b^t - w_{ib2} \overline{x_b^t})(c_{bz}^t + \theta_z - 1) \log[\sum_{u \in H_b} \pi_{uz}]]
$$

Amer&Todorovic CVPR12

Bottom up/Top Down Most Probable Explanation

$$\text{MPE: } \hat{a} = \operatorname{argmax}_{a \in A} \hat{S}(C; a)$$



Results –Volleyball Dataset

Amer&Todorovic CVPR12

Results –Volleyball Dataset

Amer&Todorovic CVPR12