# Video Annotation and Tracking with Active Learning

Carl Vondrick        Deva Ramanan
UC Irvine

*To appear at NIPS 2011.*

# The era of big data

YouTube
Broadcast Yourself

vimeo

flickr™

facebook®

# Lots of annotated images

amazonmechanical turk
Artificial Artificial Intelligence

Sorokin and Forsyth. CVPR 2008.

LabelMe

Russell, et al. CVPR 2008.

IMAGENET

Deng, et al. CVPR 2009.

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

Everingham, et al. IJCV 2010.

ESP Game
Concentrate...

Ahn and Dabbish. CHI 2004.


Xiao, et al. CVPR 2010.

Torralba, et al. PAMI 2008.

# Where are the large, real world video datasets?

LabelMe video

Yuen, Russell, Liu, Torralba. ICCV 2009.

8 years of video uploaded every day to YouTube!

**A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video**

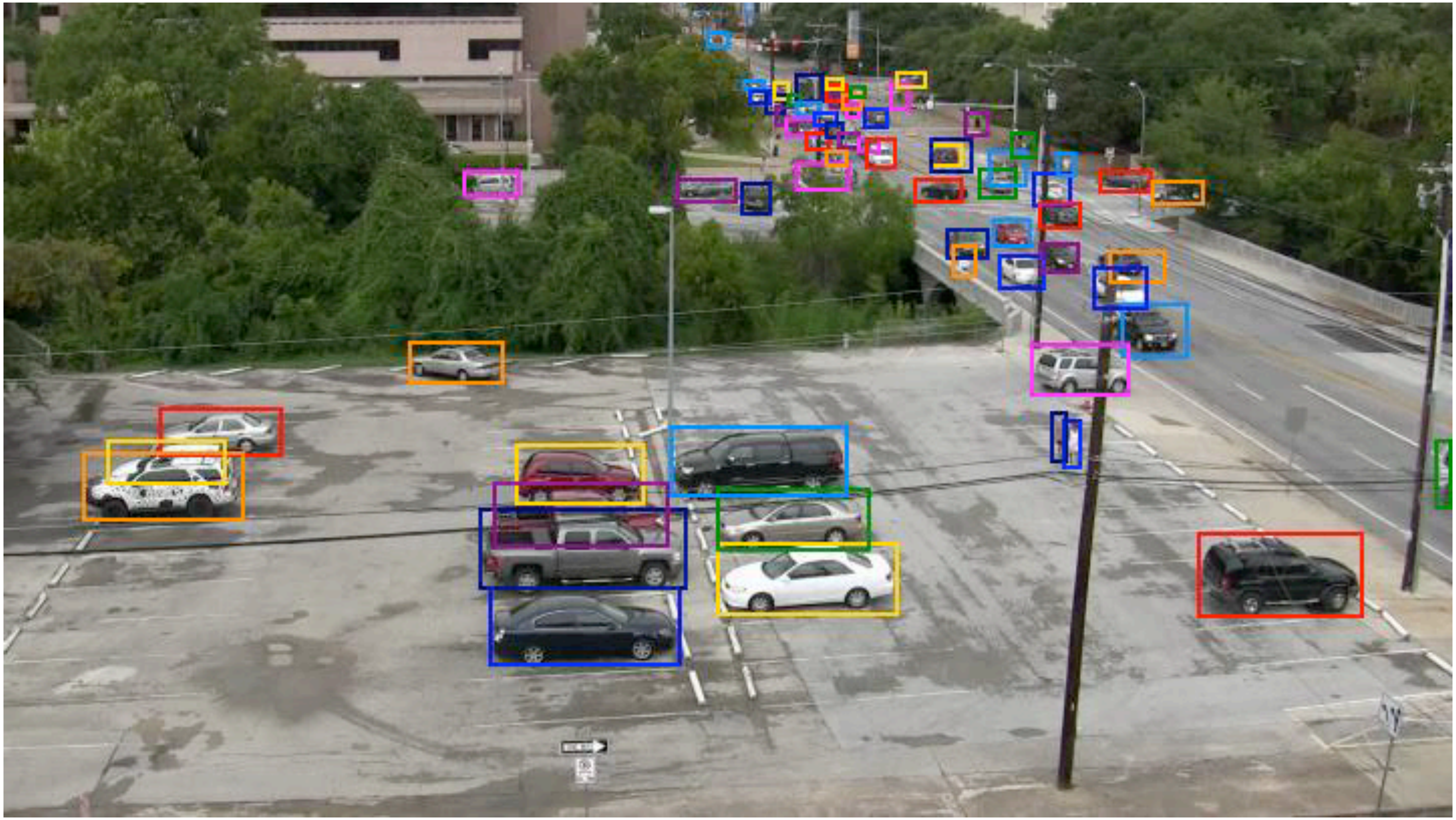
Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen,
Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears,
Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash,
Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, Mita Desai

{sangmin.oh, anthony.hoogs, amitha.perera, naresh.cuntoor}@kitware.com,
{cchen, jongtaeklee,saurajit, aggarwaljk}@mail.utexas.edu,{htlee, lsd}@umiacs.umd.edu,
{sweare, wangx16,qji}@ecse.rpi.edu, {kkreddy,shah}@eecs.ucf.edu, {jenny,torralba}@csail.mit.edu,
{cvondric,hpirsiav,dramanan}@ics.uci.edu, {bsong,anesdo, amitrc}@ee.ucr.edu, mita.desai@darpa.mil

# 24 authors!

just to build and evaluate a data set!

# Just your typical scene!



# What did it cost to annotate VIRAT?

Entire dataset is 27 hours

# What did it cost to annotate VIRAT?

## Entire dataset is 27 hours

$15,000

# What did it cost to annotate VIRAT?

Entire dataset is 27 hours

$15,000

8 months

# What did it cost to annotate VIRAT?

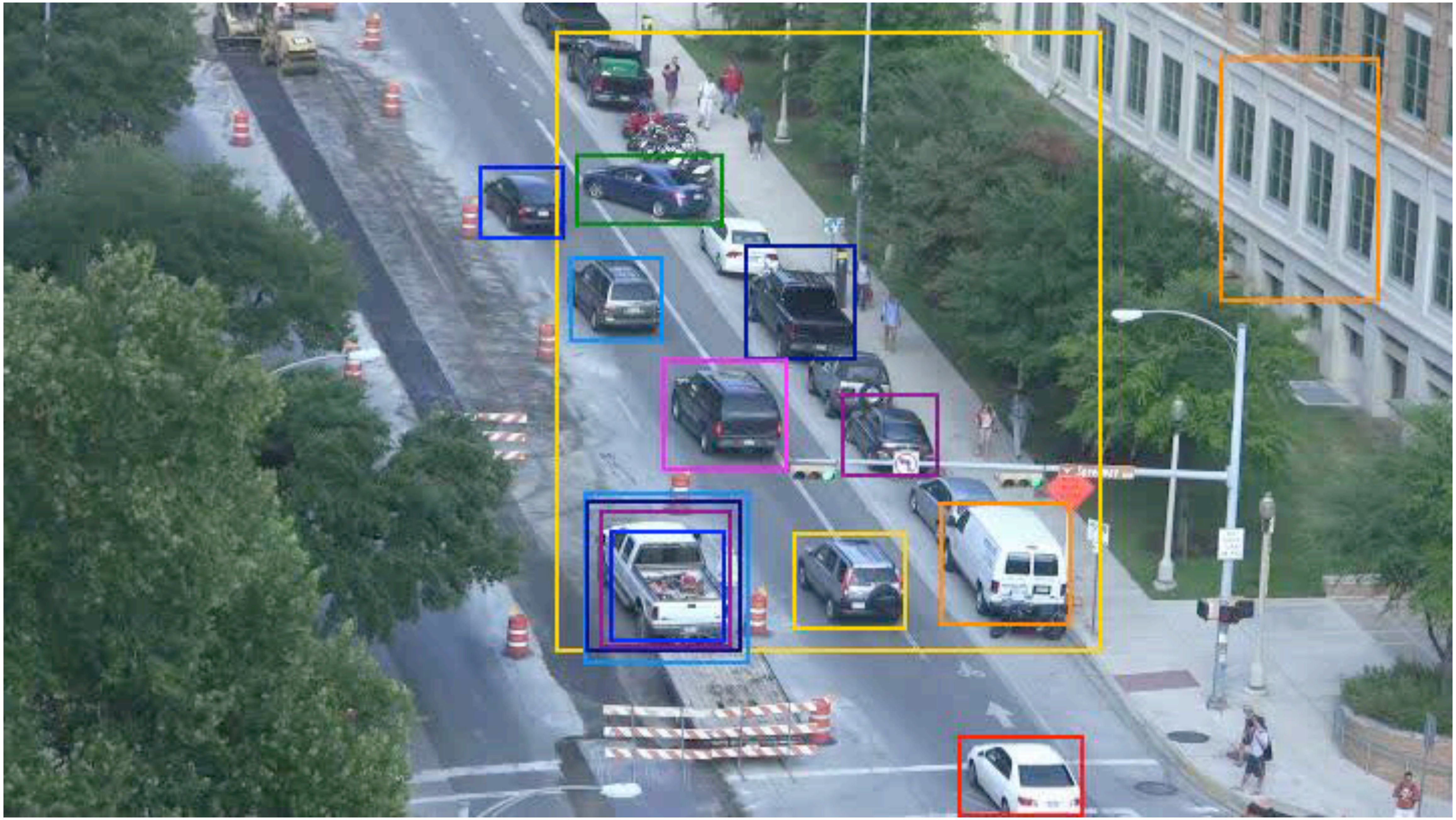
Entire dataset is 27 hours

this paper:

$15,000 ➔ $1,500

8 months ➔ 24 days

# When things go wrong...



When things go wrong...

"I would like to tell you that your Video annotation HIT is impossible... i just wasted 5 hrs for your stupid crap"
*--- Andrei, MTurk worker*

"I feel strongly about my 20 cents... I expect to paid in the next 24 hours or I WILL let the IRB know ASAP"
*--- quentin, student at an Ivy league university*

# Why is video annotation hard?

See our upcoming IJCV paper.

# Why is video annotation hard?

- **HCI**
  - How many objects should we annotate at once?
  - How do we visualize space and time?
  - How do we select key frames?

See our upcoming IJCV paper.

# Why is video annotation hard?

- **HCI**
  - How many objects should we annotate at once?
  - How do we visualize space and time?
  - How do we select key frames?
- **Crowdsourcing**
  - How do we split up work?
  - How do we do quality control?

See our upcoming IJCV paper.

# Why is video annotation hard?

- **HCI**
  - How many objects should we annotate at once?
  - How do we visualize space and time?
  - How do we select key frames?
- **Crowdsourcing**
  - How do we split up work?
  - How do we do quality control?
- **Economics**
  - How do we motivate users?
  - How do we minimize cost?

See our upcoming IJCV paper.

# Why is video annotation hard?

- **HCI**
  - How many objects should we annotate at once?
  - How do we visualize space and time?
  - How do we select key frames?
- **Crowdsourcing**
  - How do we split up work?
  - How do we do quality control?
- **Economics**
  - How do we motivate users?
  - How do we minimize cost?
- **Interpolation / Tracking**
  - How do we learn the appearance of an object?
  - How do we interpolate to minimize effort?

See our upcoming IJCV paper.

# Why is video annotation hard?

- **HCI**
  - How many objects should we annotate at once?
  - How do we visualize space and time?
  - How do we select key frames?
- **Crowdsourcing**
  - How do we split up work?
  - How do we do quality control?
- **Economics**
  - How do we motivate users?
  - How do we minimize cost?
- **Interpolation / Tracking**
  - How do we learn the appearance of an object?
  - How do we interpolate to minimize effort?

Today

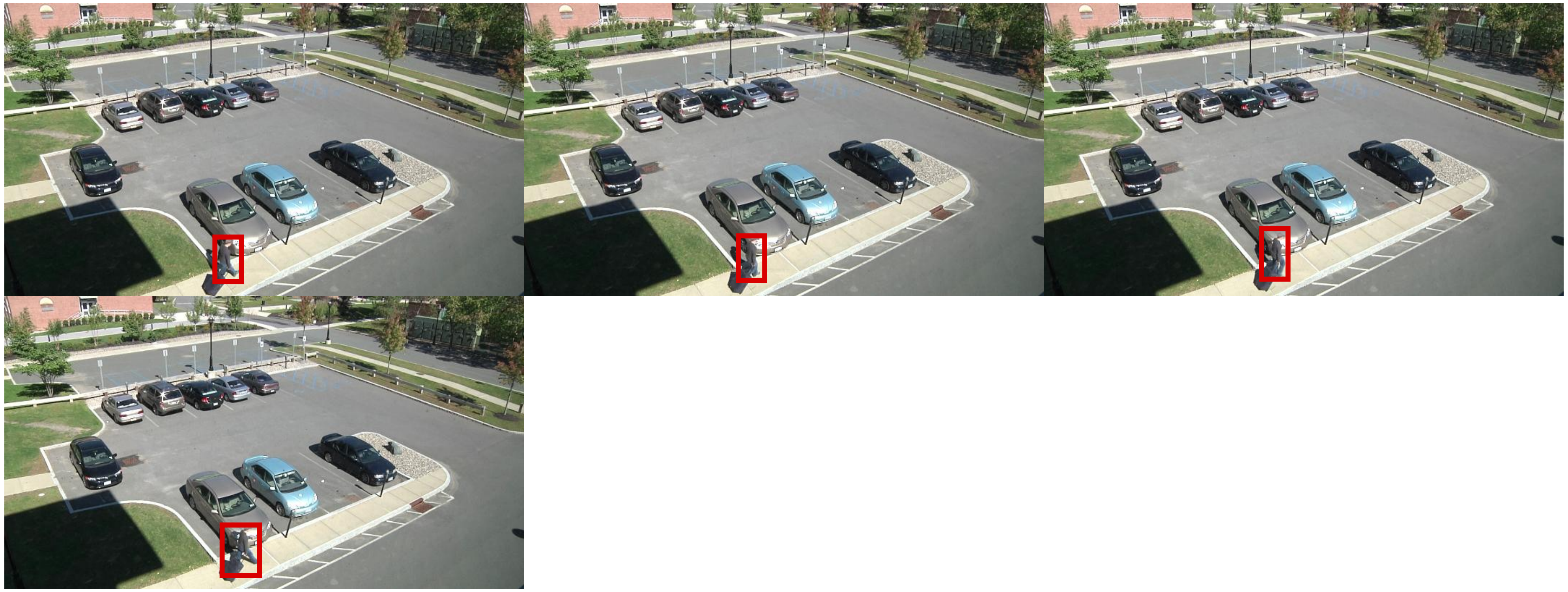
See our upcoming IJCV paper.

# Frame by Frame Labeling

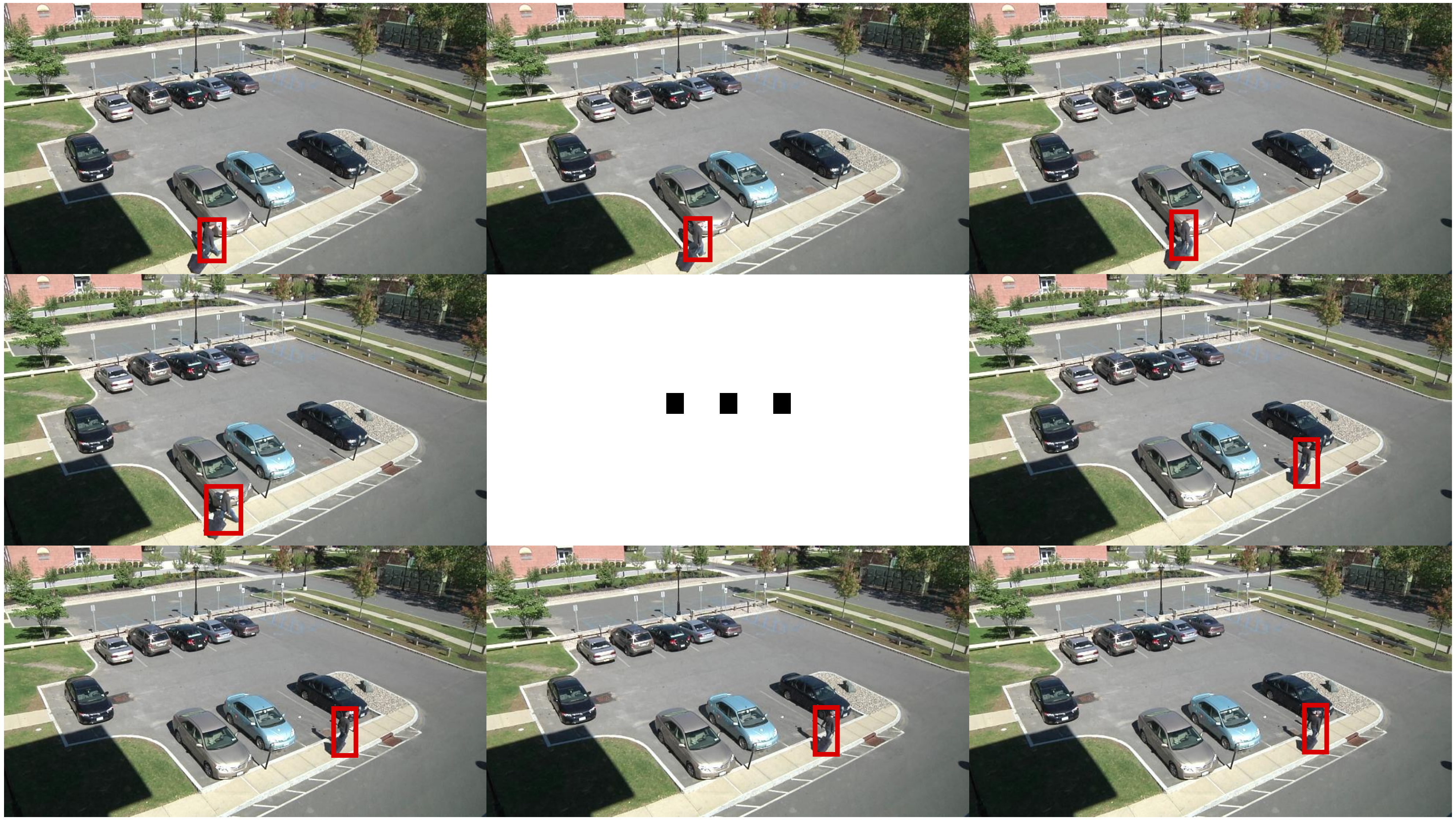
# Frame by Frame Labeling

# Frame by Frame Labeling

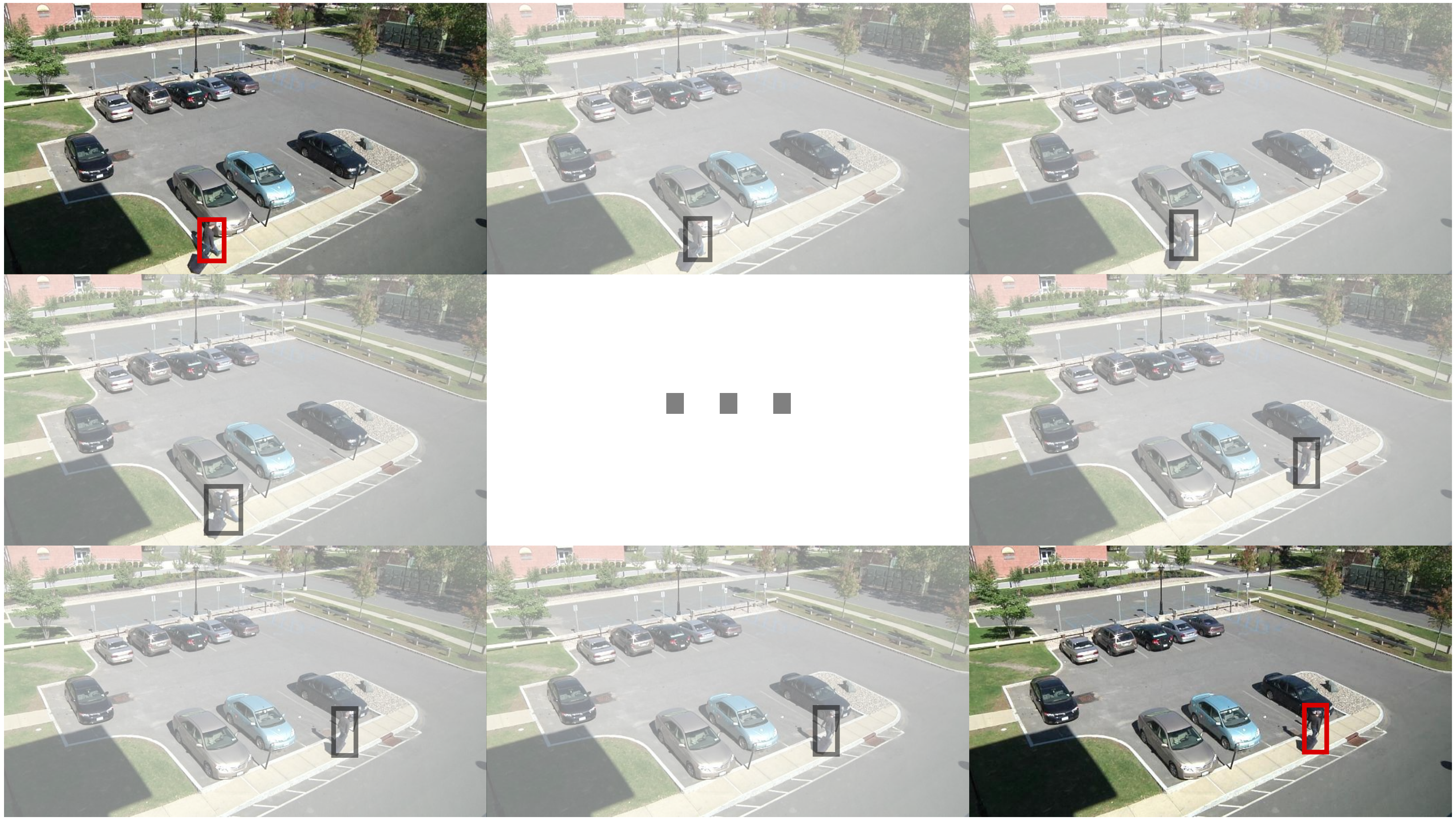
# Frame by Frame Labeling



# Frame by Frame Labeling



# Key Frame Annotation



video

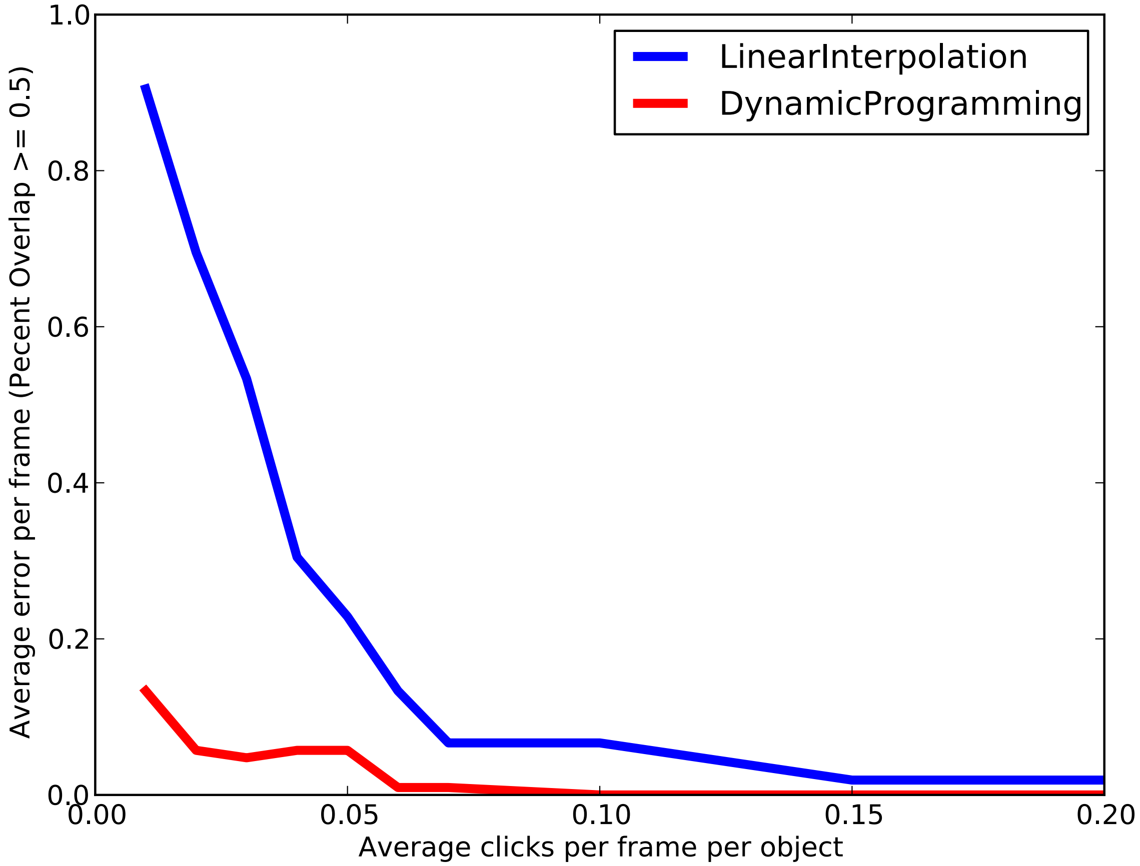
Yuen, Russell, Liu, Torralba. ICCV 2009.

# Tracking algorithms assist human annotators

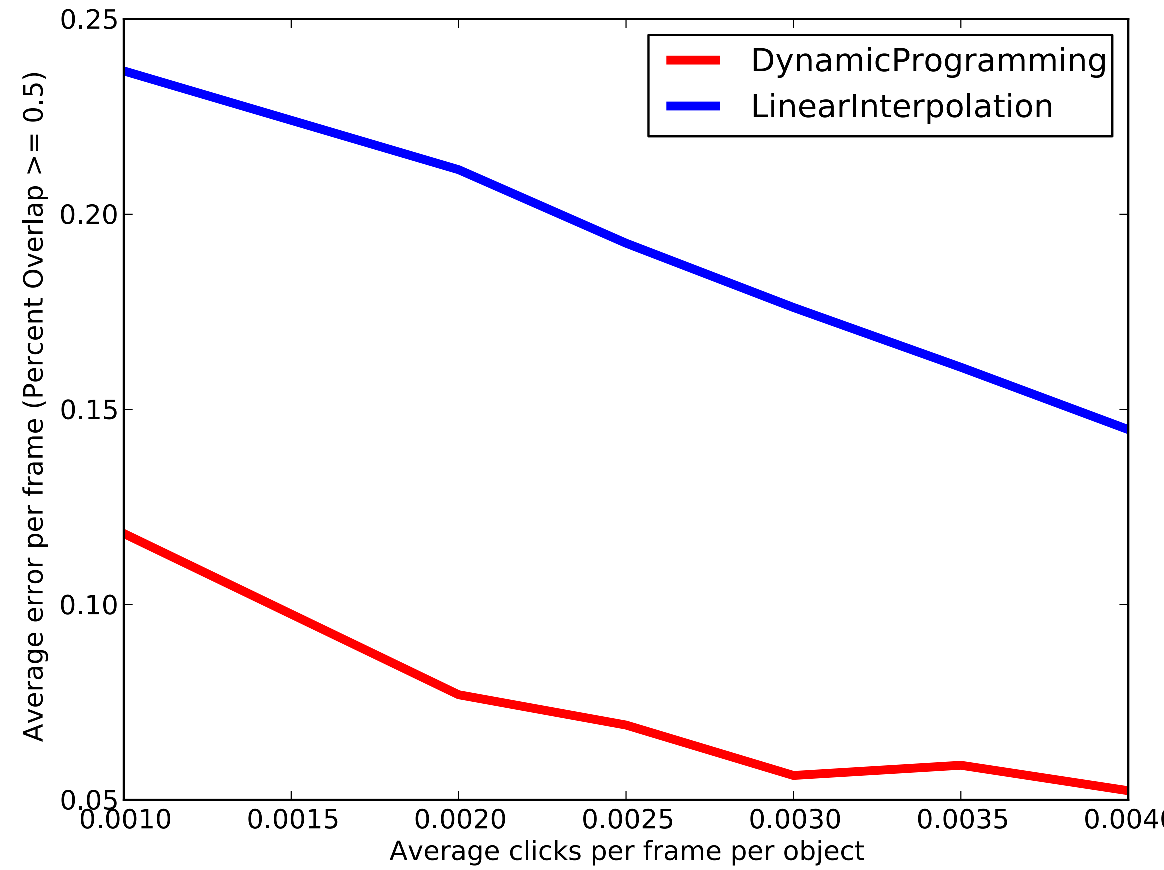
Vondrick, Ramanan, Patterson. ECCV 2010.

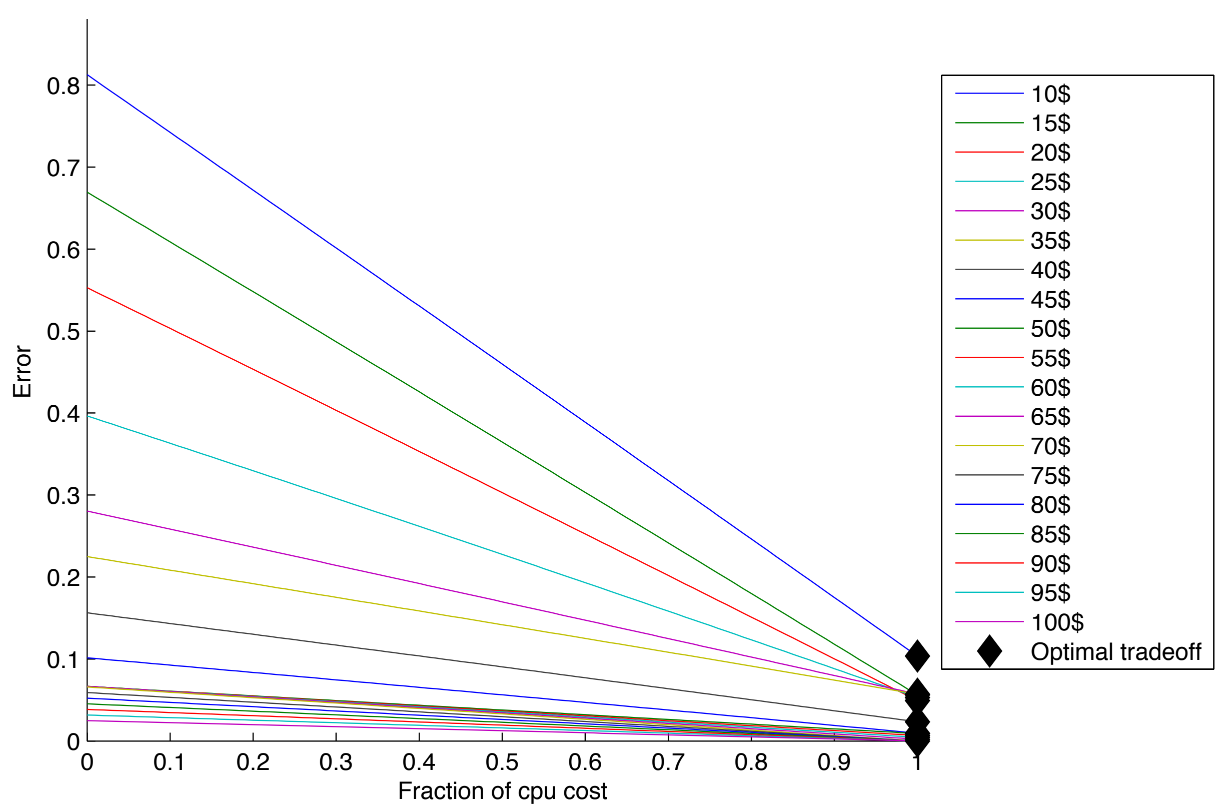
# Tracking algorithms are cost effective too

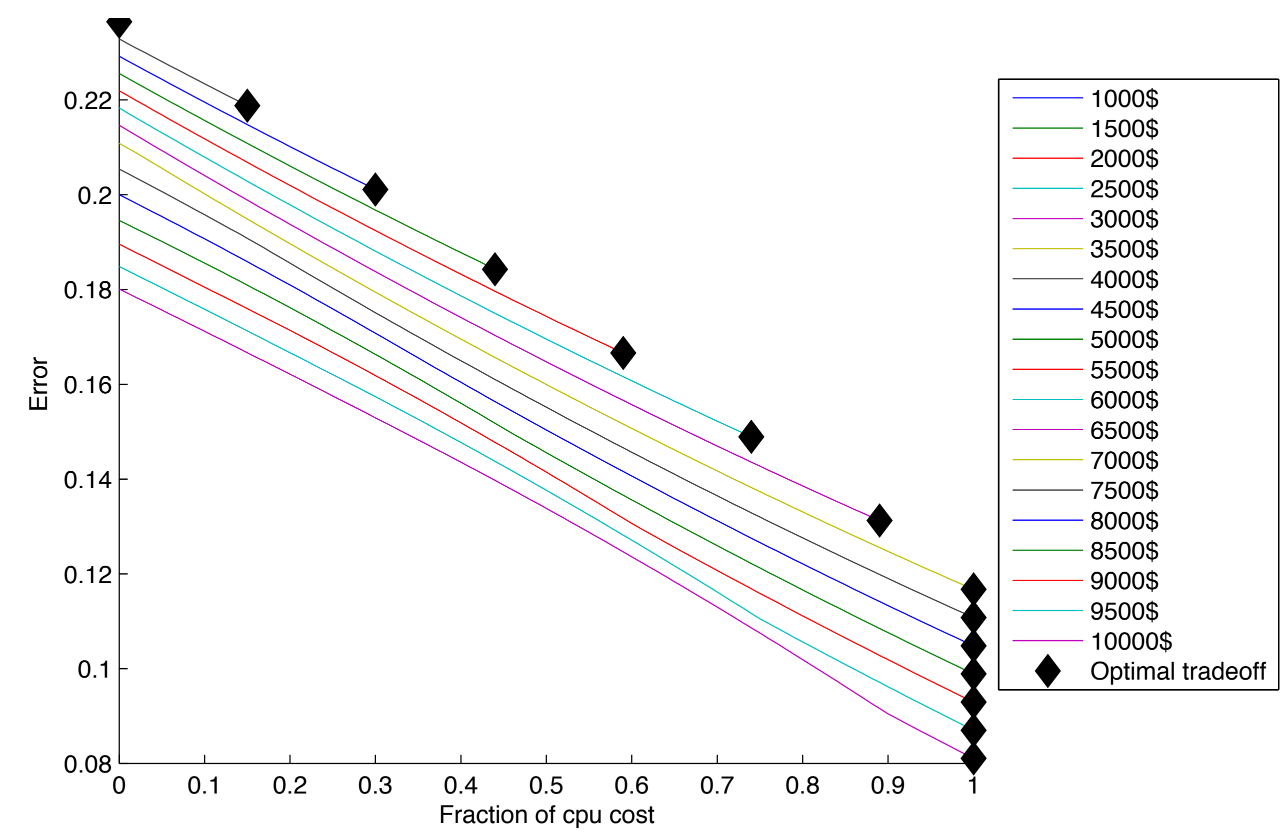
## Athletic Drills




## VIRAT Cars



Choice of key frames is crucial!

More clicks = Higher cost

# How do we pick key frames?

# How do we pick key frames?

**Fixed rate**: user annotates every T frames

- computer picks for user
- if motion is complex, T must be small
- seems to be wasteful
- dismissed by researchers

# How do we pick key frames?

**Fixed rate**: user annotates every T frames

- computer picks for user
- if motion is complex, T must be small
- seems to be wasteful
- dismissed by researchers

**User defined**: user annotates any frame he chooses

- user has complete freedom
- user can adjust T depending on complexity
- de-facto standard in video annotation

# Which is more efficient? Fixed or user defined?

* = user did fixed rate first

# Which is more efficient? Fixed or user defined?

| | Scripted | | | | Basketball | | | | VIRAT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Subject | User | Fixed | Ratio | Saved | User | Fixed | Ratio | Saved | User | Fixed | Ratio | Saved |
| A | 599 | **463** | 0.77 | 136 | 1,457 | **1,323** | 0.91 | 134 | **220** | 244 | 1.11 | -24 |
| B | 653 | **247** | 0.38 | 406 | 4,555 | **2,275** | 0.50 | 2,280 | **176** | **178** | 1.03 | -2 |
| C | 476 | **275** | 0.58 | 201 | 1,216 | **830** | 0.68 | 386 | 338 | **215** | 0.64 | 123 |
| D | 772 | **432** | 0.56 | 340 | 1,505 | **1,497** | 0.99 | 8 | 489 | **302** | 0.62 | 187 |
| *E | 605 | **371** | 0.61 | 234 | **935** | 1501 | 1.61 | -566 | 269 | **231** | 0.85 | 38 |
| *F | 654 | **472** | 0.72 | 182 | **1,672** | 1,858 | 1.11 | -186 | 372 | **326** | 0.87 | 46 |
| *G | 235 | **193** | 0.82 | 42 | **591** | 696 | 1.18 | -105 | 165 | **120** | 0.73 | 45 |
| *H | **312** | 331 | 1.06 | -19 | **656** | 748 | 1.14 | -92 | 172 | **164** | 0.95 | 8 |
| Mean | 538 | **348** | 0.66 | 190 | 1,573 | **1,341** | 0.96 | 232 | 275 | **223** | 0.83 | 53 |
| Signifiance | $\rho = 0.0350$ | | | | $\rho = 0.8576$ | | | | $\rho = 0.2618$ | | | |

Fixed rate key frames are faster!

* = user did fixed rate first

# Which is more efficient? Fixed or user defined?

| Subject | Scripted | | | | Basketball | | | | VIRAT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | User | Fixed | Ratio | Saved | User | Fixed | Ratio | Saved | User | Fixed | Ratio | Saved |
| A | 599 | **463** | 0.77 | 136 | 1,457 | **1,323** | 0.91 | 134 | **220** | 244 | 1.11 | -24 |
| B | 653 | **247** | 0.38 | 406 | 4,555 | **2,275** | 0.50 | 2,280 | **176** | **178** | 1.03 | -2 |
| C | 476 | **275** | 0.58 | 201 | 1,216 | **830** | 0.68 | 386 | 338 | **215** | 0.64 | 123 |
| D | 772 | **432** | 0.56 | 340 | 1,505 | **1,497** | 0.99 | 8 | 489 | **302** | 0.62 | 187 |
| *E | 605 | **371** | 0.61 | 234 | **935** | 1501 | 1.61 | -566 | 269 | **231** | 0.85 | 38 |
| *F | 654 | **472** | 0.72 | 182 | **1,672** | 1,858 | 1.11 | -186 | 372 | **326** | 0.87 | 46 |
| *G | 235 | **193** | 0.82 | 42 | **591** | 696 | 1.18 | -105 | 165 | **120** | 0.73 | 45 |
| *H | **312** | 331 | 1.06 | -19 | **656** | 748 | 1.14 | -92 | 172 | **164** | 0.95 | 8 |
| Mean | 538 | **348** | 0.66 | 190 | 1,573 | **1,341** | 0.96 | 232 | 275 | **223** | 0.83 | 53 |
| Significance | $\rho = 0.0350$ | | | | $\rho = 0.8576$ | | | | $\rho = 0.2618$ | | | |

34% saving

# Fixed rate key frames are faster!

* = user did fixed rate first

# Which is more efficient? Fixed or user defined?

| | Scripted | | | | Basketball | | | | VIRAT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Subject | User | Fixed | Ratio | Saved | User | Fixed | Ratio | Saved | User | Fixed | Ratio | Saved |
| A | 599 | **463** | 0.77 | 136 | 1,457 | **1,323** | 0.91 | 134 | **220** | 244 | 1.11 | -24 |
| B | 653 | **247** | 0.38 | 406 | 4,555 | **2,275** | 0.50 | 2,280 | **176** | **178** | 1.03 | -2 |
| C | 476 | **275** | 0.58 | 201 | 1,216 | **830** | 0.68 | 386 | 338 | **215** | 0.64 | 123 |
| D | 772 | **432** | 0.56 | 340 | 1,505 | **1,497** | 0.99 | 8 | 489 | **302** | 0.62 | 187 |
| *E | 605 | **371** | 0.61 | 234 | **935** | 1501 | 1.61 | -566 | 269 | **231** | 0.85 | 38 |
| *F | 654 | **472** | 0.72 | 182 | **1,672** | 1,858 | 1.11 | -186 | 372 | **326** | 0.87 | 46 |
| *G | 235 | **193** | 0.82 | 42 | **591** | 696 | 1.18 | -105 | 165 | **120** | 0.73 | 45 |
| *H | **312** | 331 | 1.06 | -19 | **656** | 748 | 1.14 | -92 | 172 | **164** | 0.95 | 8 |
| Mean | 538 | **348** | 0.66 | 190 | 1,573 | **1,341** | 0.96 | 232 | 275 | **223** | 0.83 | 53 |
| Significance | | $\boxed{\rho = 0.0350}$ | | | | $\rho = 0.8576$ | | | | $\rho = 0.2618$ | | |

Statistical significant

# Fixed rate key frames are faster!

* = user did fixed rate first

# Which is more efficient? Fixed or user defined?

2009 Marr prize winner

| Subject | Scripted User | Scripted Fixed | Scripted Ratio | Scripted Saved | Basketball User | Basketball Fixed | Basketball Ratio | Basketball Saved | VIRAT User | VIRAT Fixed | VIRAT Ratio | VIRAT Saved |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 599 | **463** | 0.77 | 136 | 1,457 | **1,323** | 0.91 | 134 | **220** | 244 | 1.11 | -24 |
| B | 653 | **247** | 0.38 | 406 | 4,555 | **2,275** | 0.50 | 2,280 | **176** | **178** | 1.03 | -2 |
| C | 476 | **275** | 0.58 | 201 | 1,216 | **830** | 0.68 | 386 | 338 | **215** | 0.64 | 123 |
| D | 772 | **432** | 0.56 | 340 | 1,505 | **1,497** | 0.99 | 8 | 489 | **302** | 0.62 | 187 |
| *E | 605 | **371** | 0.61 | 234 | **935** | 1501 | 1.61 | -566 | 269 | **231** | 0.85 | 38 |
| *F | 654 | **472** | 0.72 | 182 | **1,672** | 1,858 | 1.11 | -186 | 372 | **326** | 0.87 | 46 |
| *G | 235 | **193** | 0.82 | 42 | **591** | 696 | 1.18 | -105 | 165 | **120** | 0.73 | 45 |
| *H | **312** | 331 | 1.06 | -19 | **656** | 748 | 1.14 | -92 | 172 | **164** | 0.95 | 8 |
| Mean | 538 | **348** | 0.66 | 190 | 1,573 | **1,341** | 0.96 | 232 | 275 | **223** | 0.83 | 53 |
| Signifiance | $\rho = 0.0350$ | | | | $\rho = 0.8576$ | | | | $\rho = 0.2618$ | | | |

# Fixed rate key frames are faster!

* = user did fixed rate first

# Which is more efficient? Fixed or user defined?

Learning bias

| Subject | Scripted User | Scripted Fixed | Scripted Ratio | Scripted Saved | Basketball User | Basketball Fixed | Basketball Ratio | Basketball Saved | VIRAT User | VIRAT Fixed | VIRAT Ratio | VIRAT Saved |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 599 | **463** | 0.77 | 136 | 1,457 | **1,323** | 0.91 | 134 | **220** | 244 | 1.11 | -24 |
| B | 653 | **247** | 0.38 | 406 | 4,555 | **2,275** | 0.50 | 2,280 | **176** | **178** | 1.03 | -2 |
| C | 476 | **275** | 0.58 | 201 | 1,216 | **830** | 0.68 | 386 | 338 | **215** | 0.64 | 123 |
| D | 772 | **432** | 0.56 | 340 | 1,505 | **1,497** | 0.99 | 8 | 489 | **302** | 0.62 | 187 |
| *E | 605 | **371** | 0.61 | 234 | **935** | 1501 | 1.61 | -566 | 269 | **231** | 0.85 | 38 |
| *F | 654 | **472** | 0.72 | 182 | **1,672** | 1,858 | 1.11 | -186 | 372 | **326** | 0.87 | 46 |
| *G | 235 | **193** | 0.82 | 42 | **591** | 696 | 1.18 | -105 | 165 | **120** | 0.73 | 45 |
| *H | **312** | 331 | 1.06 | -19 | **656** | 748 | 1.14 | -92 | 172 | **164** | 0.95 | 8 |
| Mean | 538 | **348** | 0.66 | 190 | 1,573 | **1,341** | 0.96 | 232 | 275 | **223** | 0.83 | 53 |
| Signifiance | | $p = 0.0350$ | | | | $p = 0.8576$ | | | | $p = 0.2618$ | | |

# Fixed rate key frames are faster!

* = user did fixed rate first

# Which is more efficient? Fixed or user defined?

| | Scripted | | | | Basketball | | | | VIRAT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Subject | User | Fixed | Ratio | Saved | User | Fixed | Ratio | Saved | User | Fixed | Ratio | Saved |
| A | 599 | **463** | 0.77 | 136 | 1,457 | **1,323** | 0.91 | 134 | **220** | 244 | 1.11 | -24 |
| B | 653 | **247** | 0.38 | 406 | 4,555 | **2,275** | 0.50 | 2,280 | **176** | **178** | 1.03 | -2 |
| C | 476 | **275** | 0.58 | 201 | 1,216 | **830** | 0.68 | 386 | 338 | **215** | 0.64 | 123 |
| D | 772 | **432** | 0.56 | 340 | 1,505 | **1,497** | 0.99 | 8 | 489 | **302** | 0.62 | 187 |
| *E | 605 | **371** | 0.61 | 234 | **935** | 1501 | 1.61 | -566 | 269 | **231** | 0.85 | 38 |
| *F | 654 | **472** | 0.72 | 182 | **1,672** | 1,858 | 1.11 | -186 | 372 | **326** | 0.87 | 46 |
| *G | 235 | **193** | 0.82 | 42 | **591** | 696 | 1.18 | -105 | 165 | **120** | 0.73 | 45 |
| *H | **312** | 331 | 1.06 | -19 | **656** | 748 | 1.14 | -92 | 172 | **164** | 0.95 | 8 |
| Mean | 538 | **348** | 0.66 | 190 | 1,573 | **1,341** | 0.96 | 232 | 275 | **223** | 0.83 | 53 |
| Signifiance | $\rho = 0.0350$ | | | | $\rho = 0.8576$ | | | | $\rho = 0.2618$ | | | |

# Fixed rate key frames are faster!

* = user did fixed rate first

Fixed rate still wins under predictable, linear motion

Humans do not pick optimal key frames.

What frame should the user annotate next?

Use active learning.

# Large body of literature for active learning

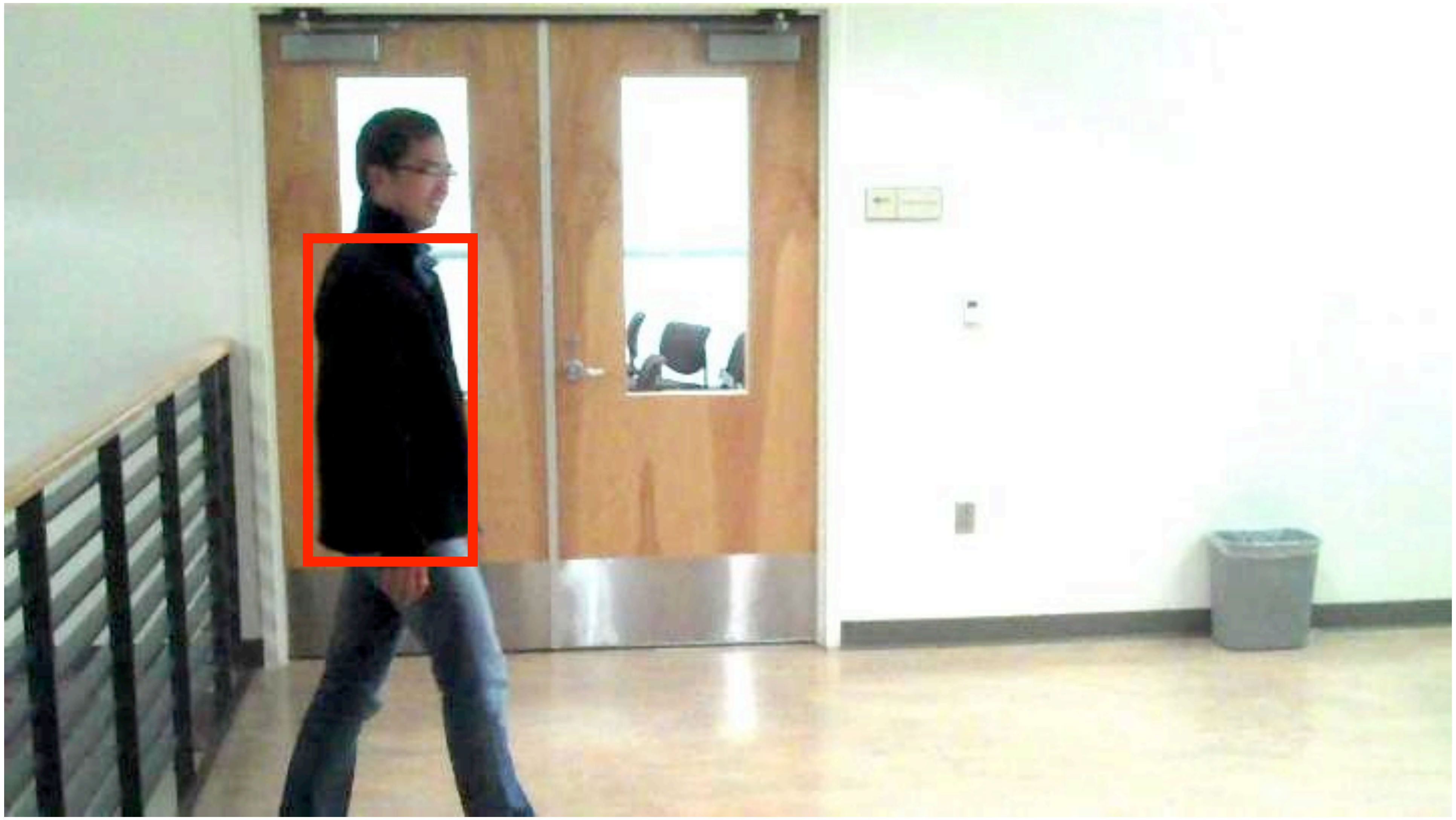
- Uncertainty sampling: query for least certain example
- Query-by-committee: most informative is example with most disagreement
- Expected model change: ask for example that would change the model the most
- Expected error reduction: ask for example that would reduce error the most
- Many more... see (Settles 2009) for survey

Why not use off-the-shelf active learning?

1. Video frames are structured, i.e. *non-i.i.d*

2. Active learning wants right *model*;
   we want right *labels*

# A Simple Tracker



Positives are the labeled boxes
Negatives are every other non-overlapping box

Extract color + HOG features from frames

Train linear SVM to discriminate:

$$w^* = \operatorname{argmin} \frac{1}{2} w \cdot w + C \sum_{n}^{N} \max(0, 1 - y_n w \cdot \phi_n(b_n))$$

appearance

motion model

$$E(b_{0:T}) = \sum_{t=0}^{T} U_t(b_t) + S(b_t, b_{t-1})$$

$$U_t(b_t) = \min\left(-w \cdot \phi_t(b_t), \alpha_1\right), \quad S(b_t, b_{t-1}) = \alpha_2 ||b_t - b_{t-1}||^2$$

Find least cost path:

$$b^*_{0:T} = \underset{b_{0:T}}{\operatorname{argmin}}\, E(b_{0:T}) \quad s.t. \quad b_t = b^i_t \quad \forall b^i_t \in \zeta$$

$$U_t(b_t) = \min\left(-w \cdot \phi_t(b_t), \alpha_1\right)$$

$$U_t(b_t) = \min\left(-w \cdot \phi_t(b_t), \alpha_1\right)$$

$$S(b_t, b_{t-1}) = \alpha_2 ||b_t - b_{t-1}||^2$$

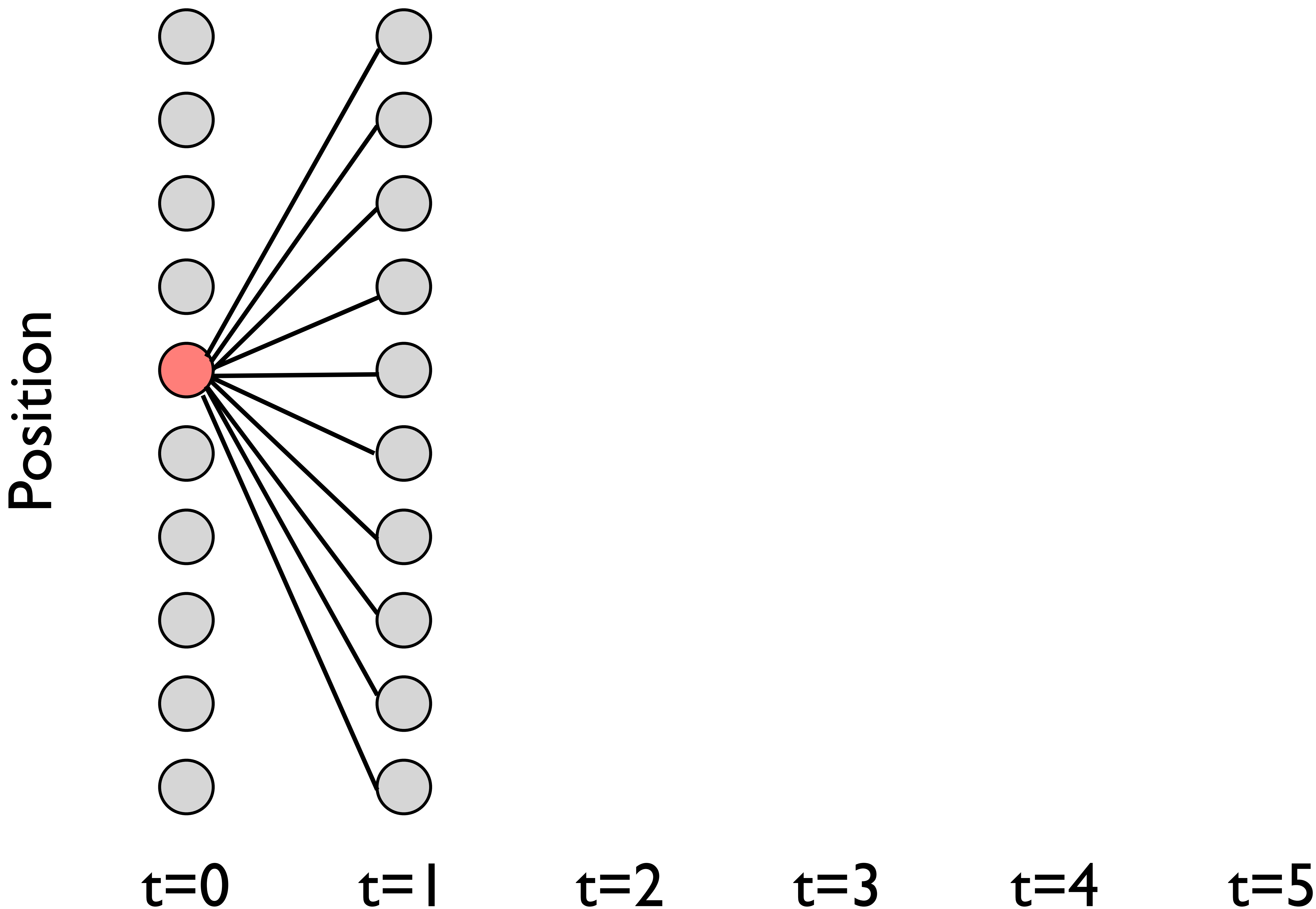
Solve the recursion with dynamic programming:

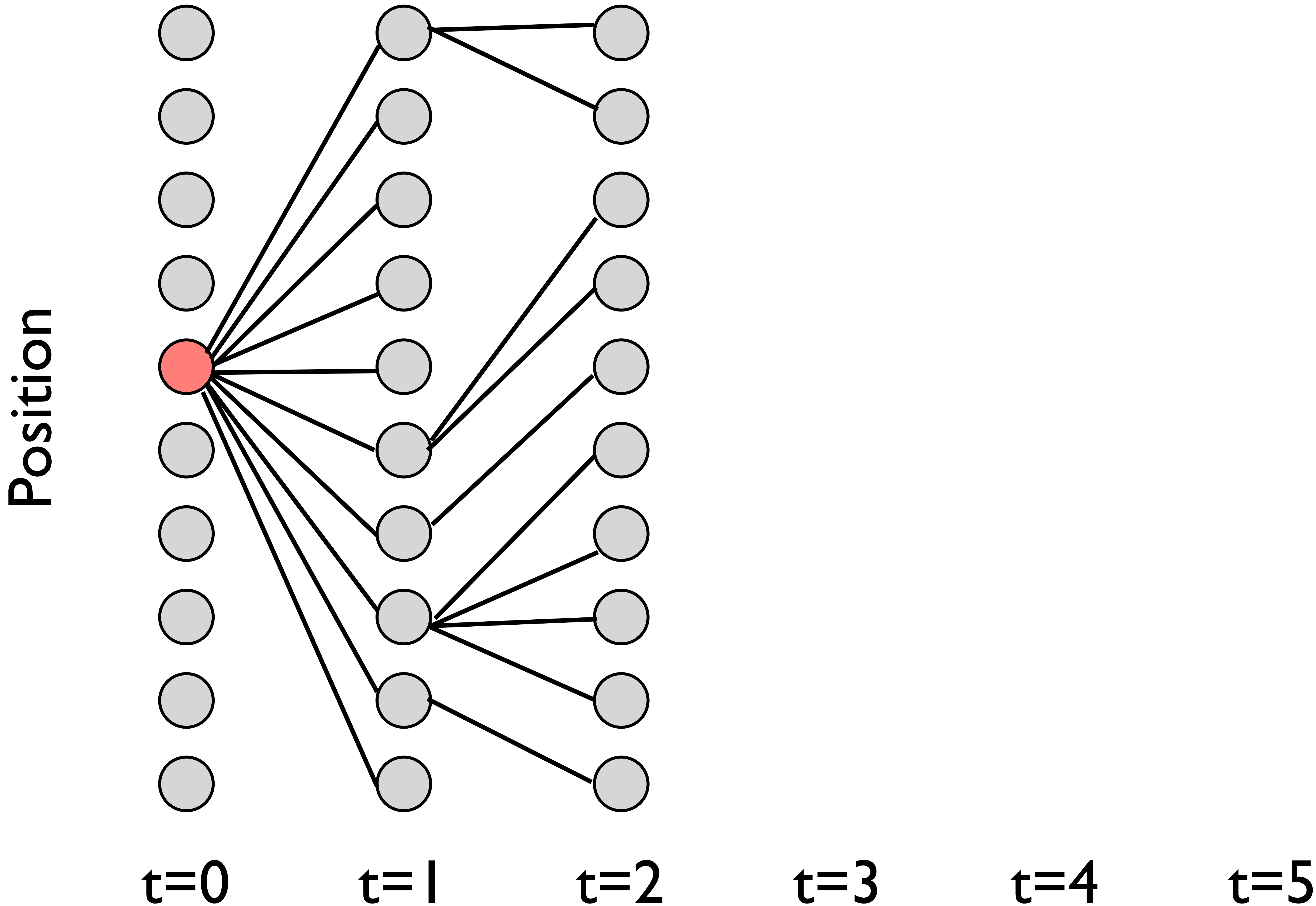
$$
\begin{aligned}
C_0^{\rightarrow}(b_0) &= U_0(b_0) \\
C_t^{\rightarrow}(b_t) &= U_t(b_t) + \min_{b_{t-1}} C_{t-1}^{\rightarrow}(b_{t-1}) + S(b_t, b_{t-1}) \\
\pi_t^{\rightarrow}(b_t) &= \operatorname*{argmin}_{b_{t-1}} C_{t-1}^{\rightarrow}(b_{t-1}) + S(b_t, b_{t-1})
\end{aligned}
$$

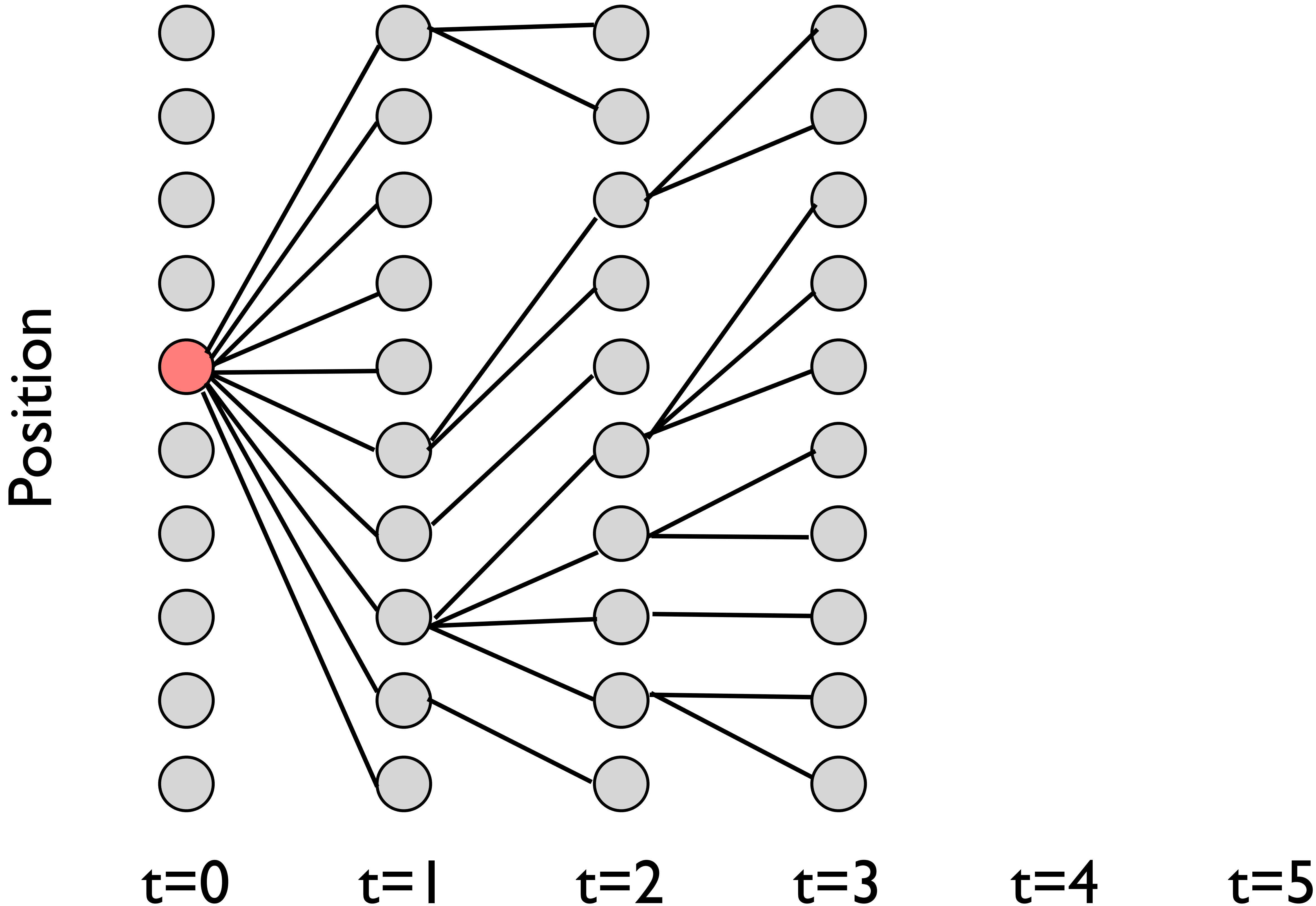
For K locations and T frames, can solve in O(TK)

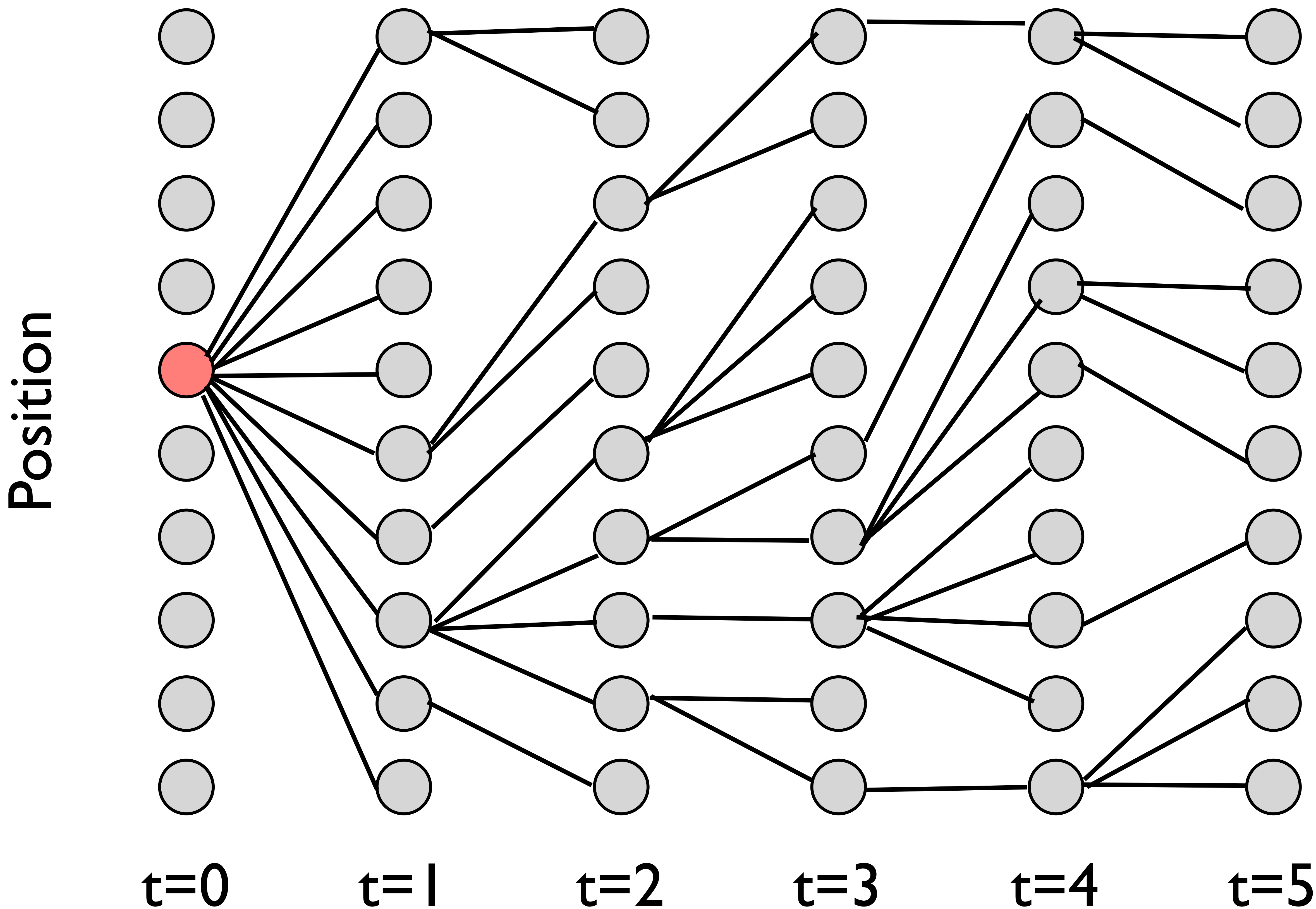
(with quadratic distance transform)


Position
t=0
t=1
t=2
t=3
t=4
t=5

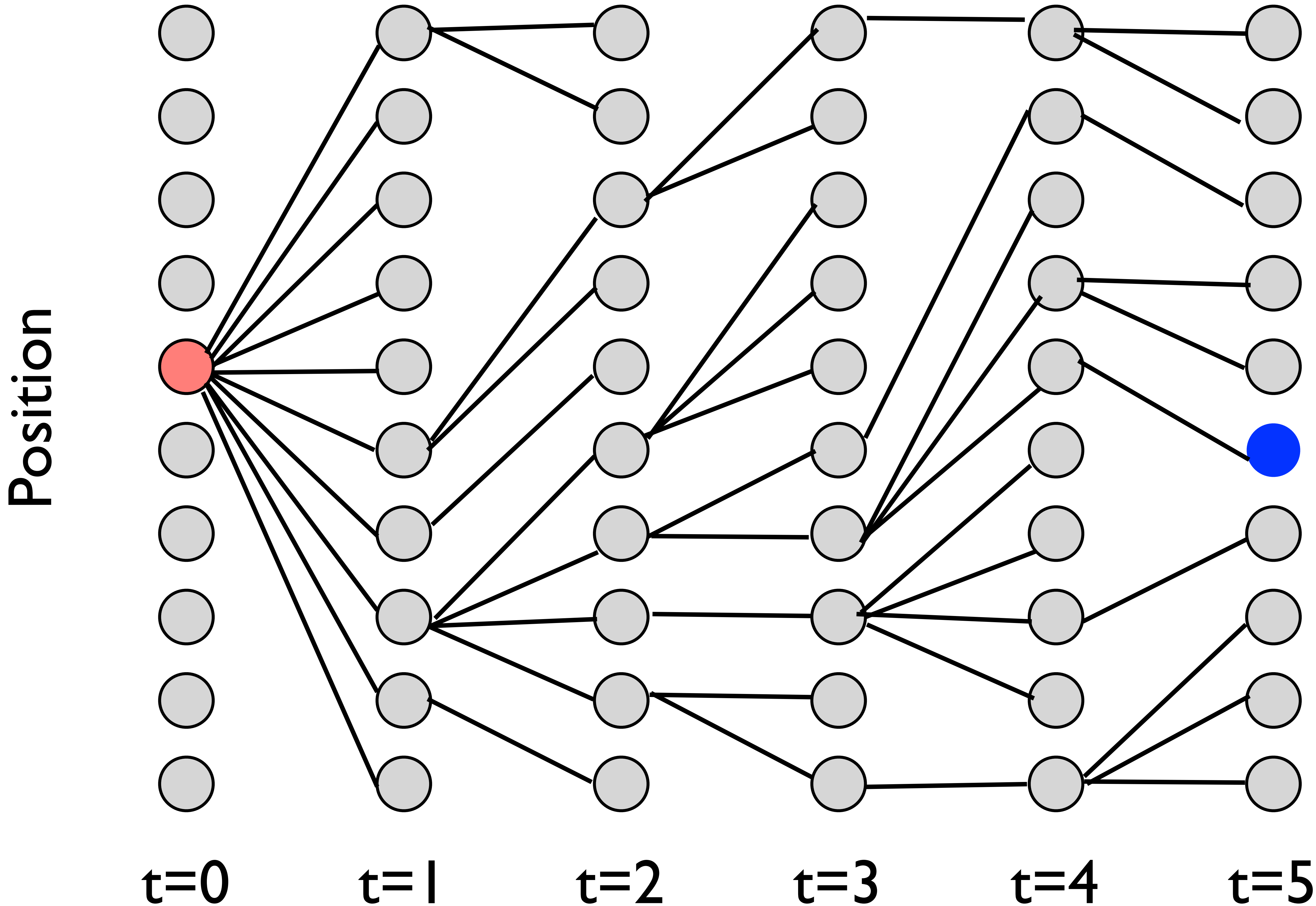

Position
t=0
t=1
t=2
t=3
t=4
t=5

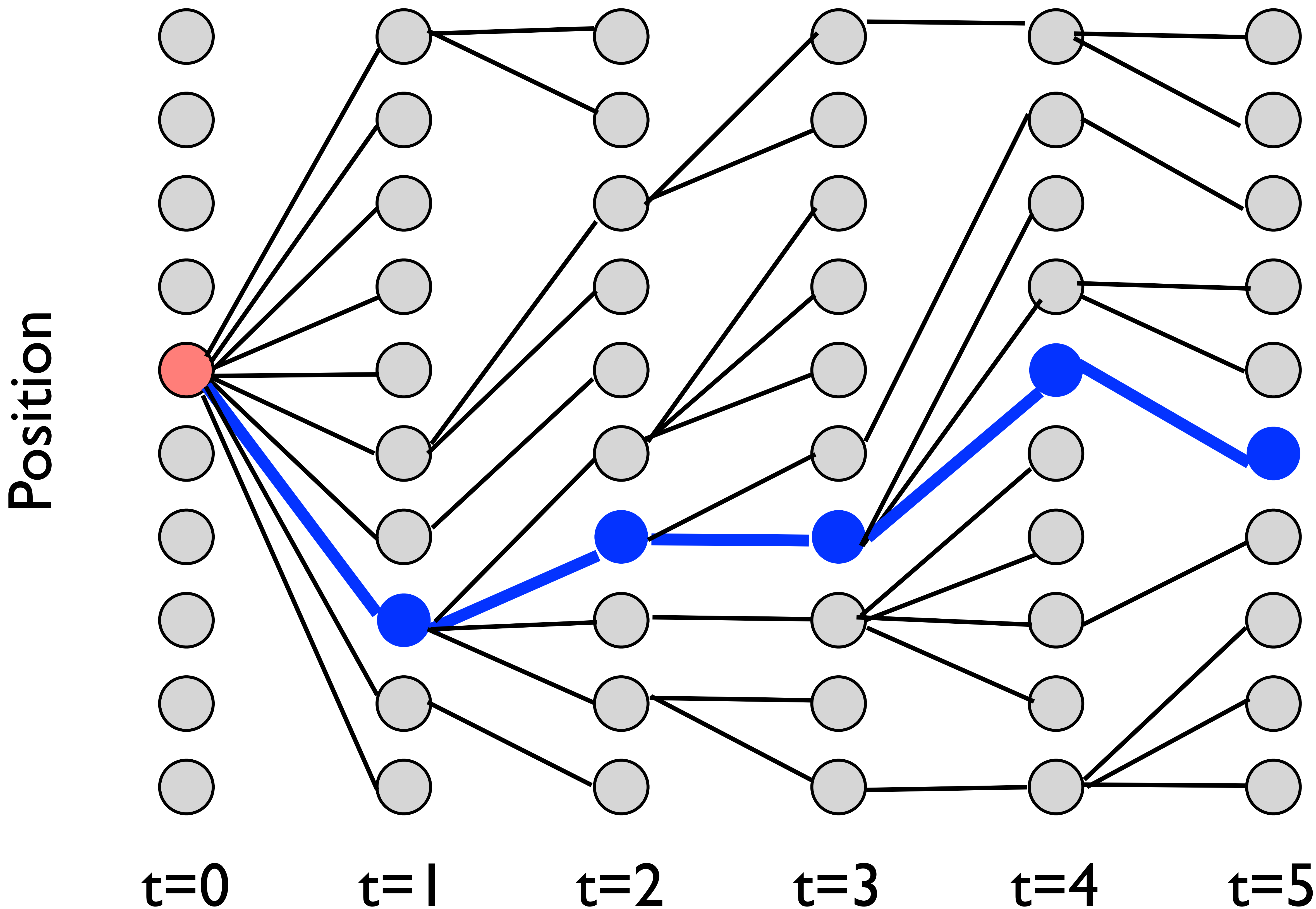

Position
t=0
t=1
t=2
t=3
t=4
t=5

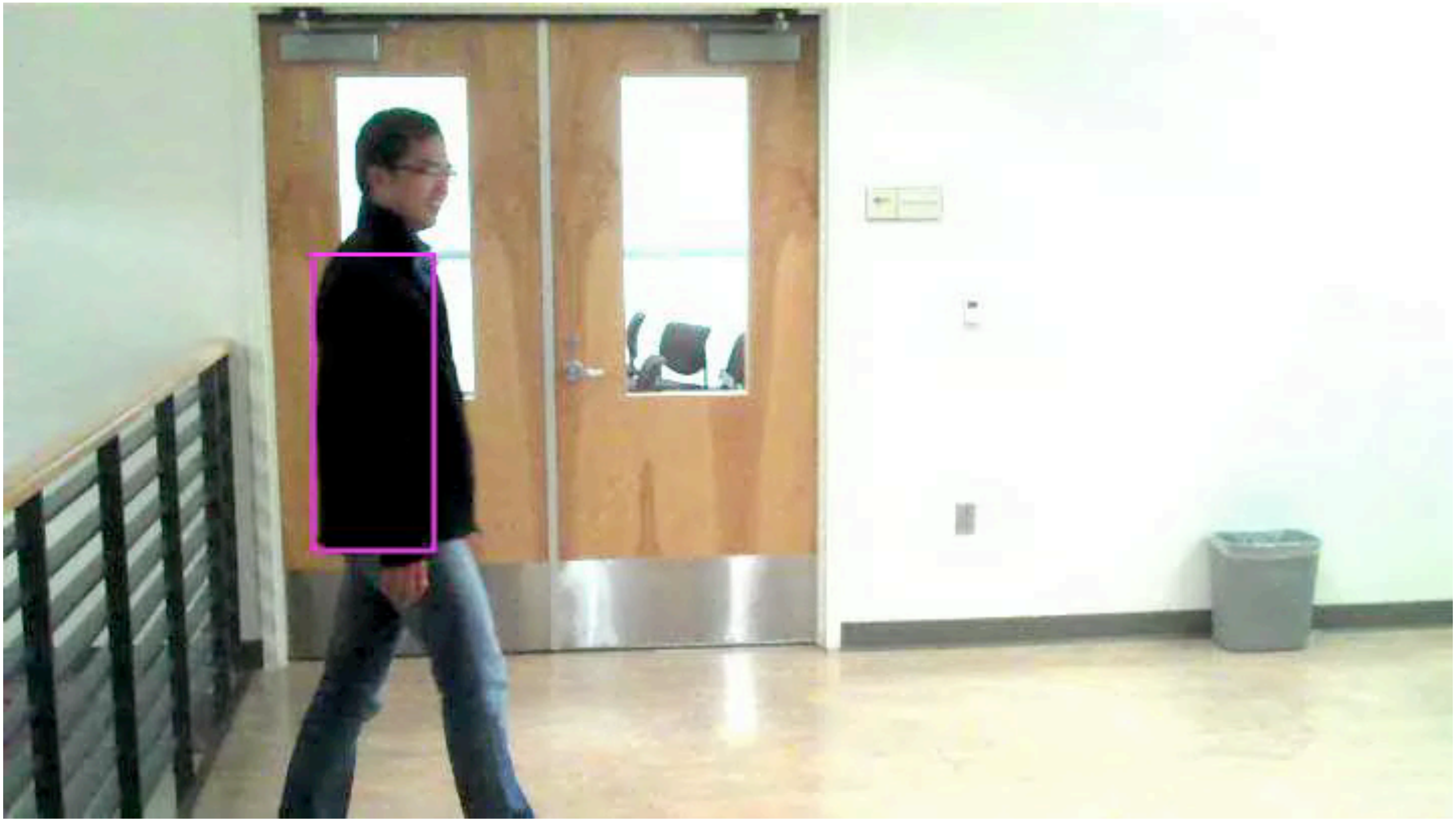

Position
t=0
t=1
t=2
t=3
t=4
t=5



Position
t=0
t=1
t=2
t=3
t=4
t=5


Position
t=0
t=1
t=2
t=3
t=4
t=5


Position
t=0
t=1
t=2
t=3
t=4
t=5



# Active Learning

Given previous annotations,
which frame should the user annotate next?

Optimal choice reduces the error the most:

$$t^{opt} = \underset{0 \le t \le T}{\operatorname{argmin}} \sum_{j=0}^{T} \operatorname{err}(b_j^{gt}, \operatorname{next}_j(b_t^{gt}))$$

# Maximum expected label change:

$$t^* = \operatorname*{argmax}_{0 \leq t \leq T} \sum_{i=0}^{K} P(b_t^i) \cdot \Delta I(b_t^i) \quad \text{where} \quad \Delta I(b_t^i) = \sum_{j=0}^{T} \text{err}(\text{curr}_j, \text{next}_j(b_t^i))$$

probability of
annotating at
location i

amount path
changes when
constrained by i

# Expected Label Change (ELC)
# vs
# Expected Gradient Length (EGL)

EGL asks for example that changes the model the most
ELC asks for frame that changes the label the most

EGL assumes i.i.d. for computational reasons
ECL assumes non-i.i.d.

Maximum expected label change:

$$t^* = \underset{0 \le t \le T}{\operatorname{argmax}} \sum_{i=0}^{K} P(b_t^i) \cdot \Delta I(b_t^i) \quad \text{where} \quad \Delta I(b_t^i) = \sum_{j=0}^{T} \operatorname{err}(\operatorname{curr}_j, \operatorname{next}_j(b_t^i))$$

# “What is probability user annotates here?”

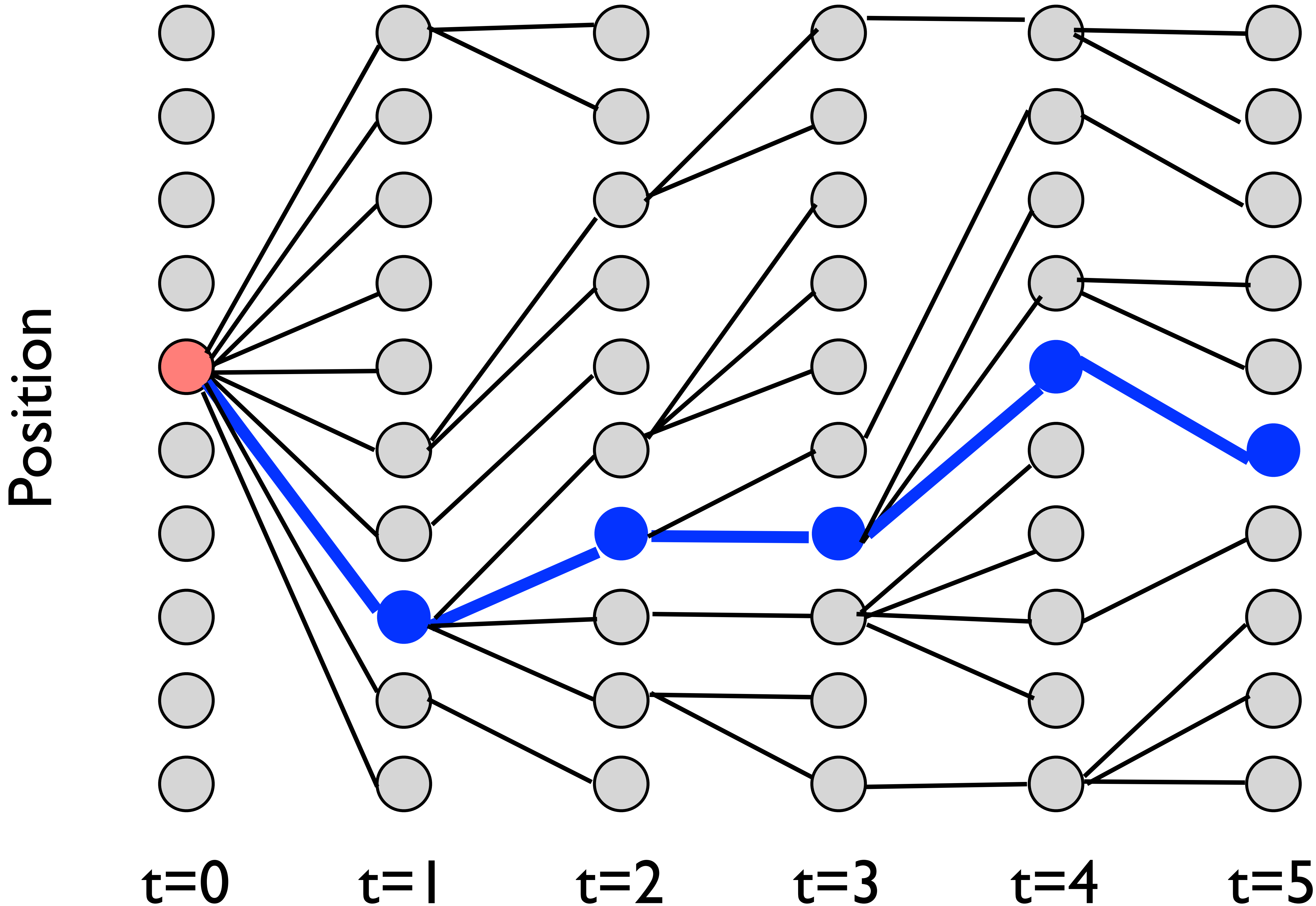
Gives path
constrained by b_t^i

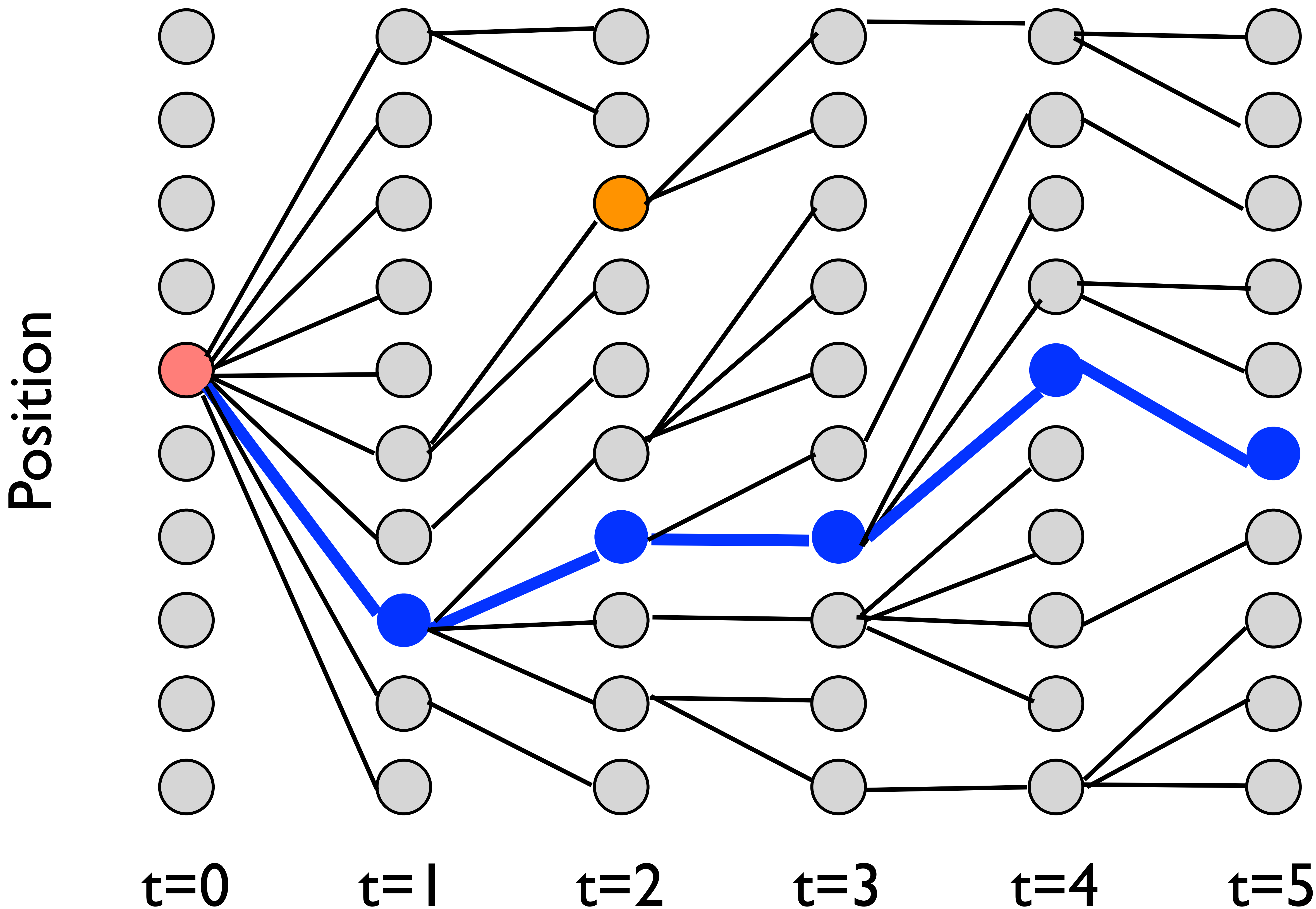
$$P(b_t^i) \propto \exp\left(\frac{-\Psi(b_t^i)}{\sigma^2}\right) \quad \text{where} \quad \Psi(b_t^i) = E\left(\text{next}_{0:T}(b_t^i)\right)$$

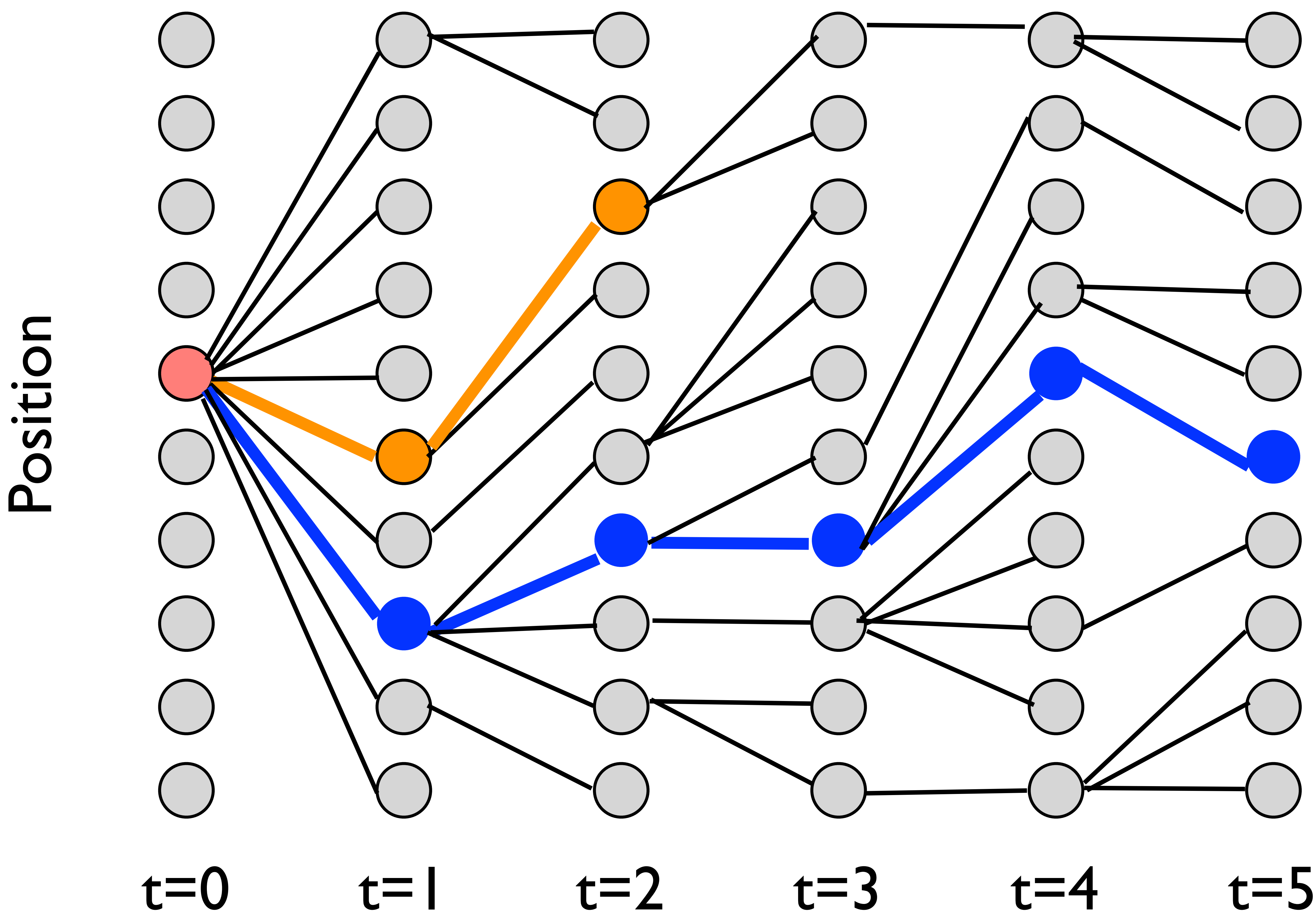
$$\Psi(b_t^i) = C_t^{\rightarrow}(b_t^i) + C_t^{\leftarrow}(b_t^i) - U(b_t^i)$$

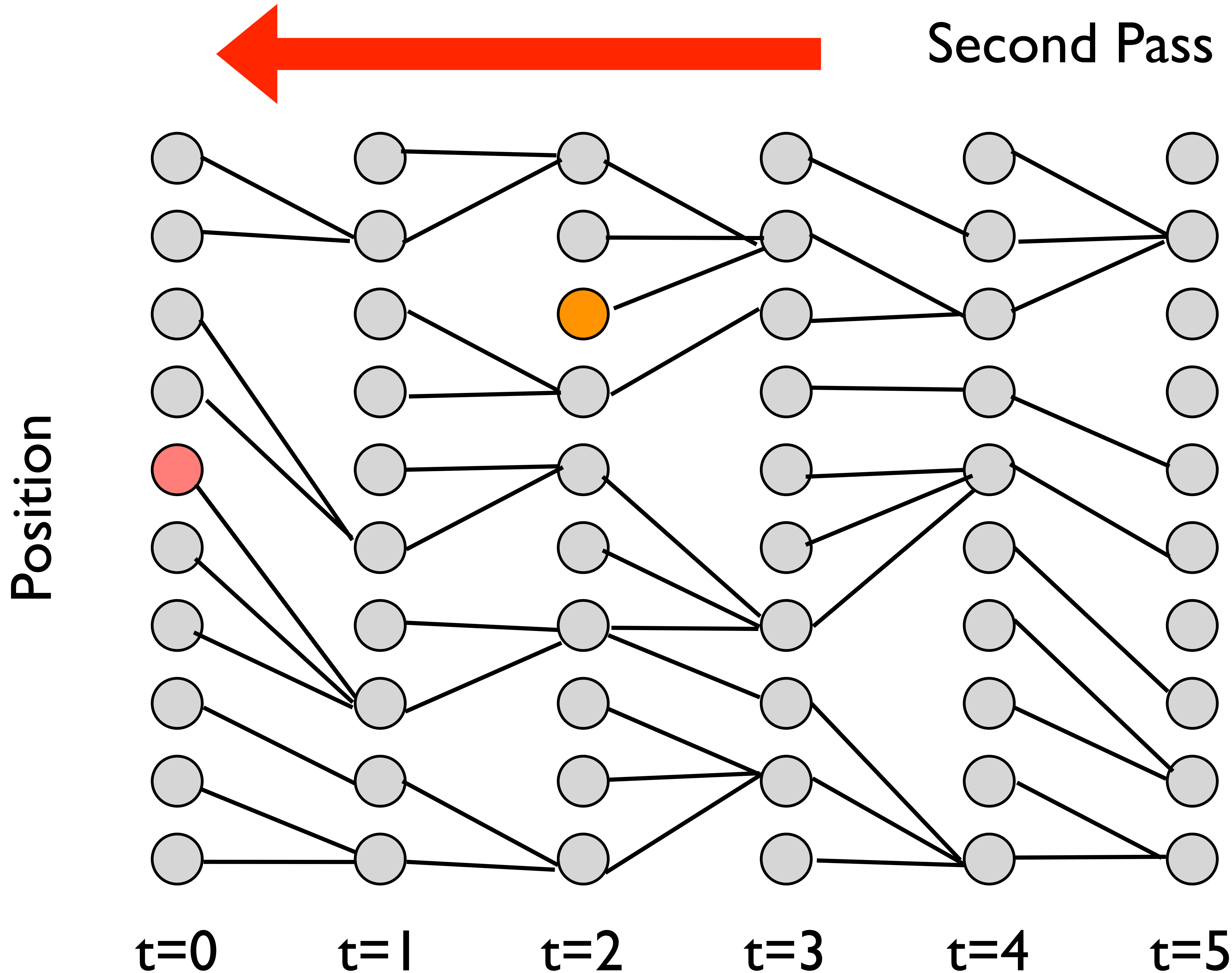
Score of path
constrained by b_t^i

Standard two-pass
algorithm to compute
min-marginals

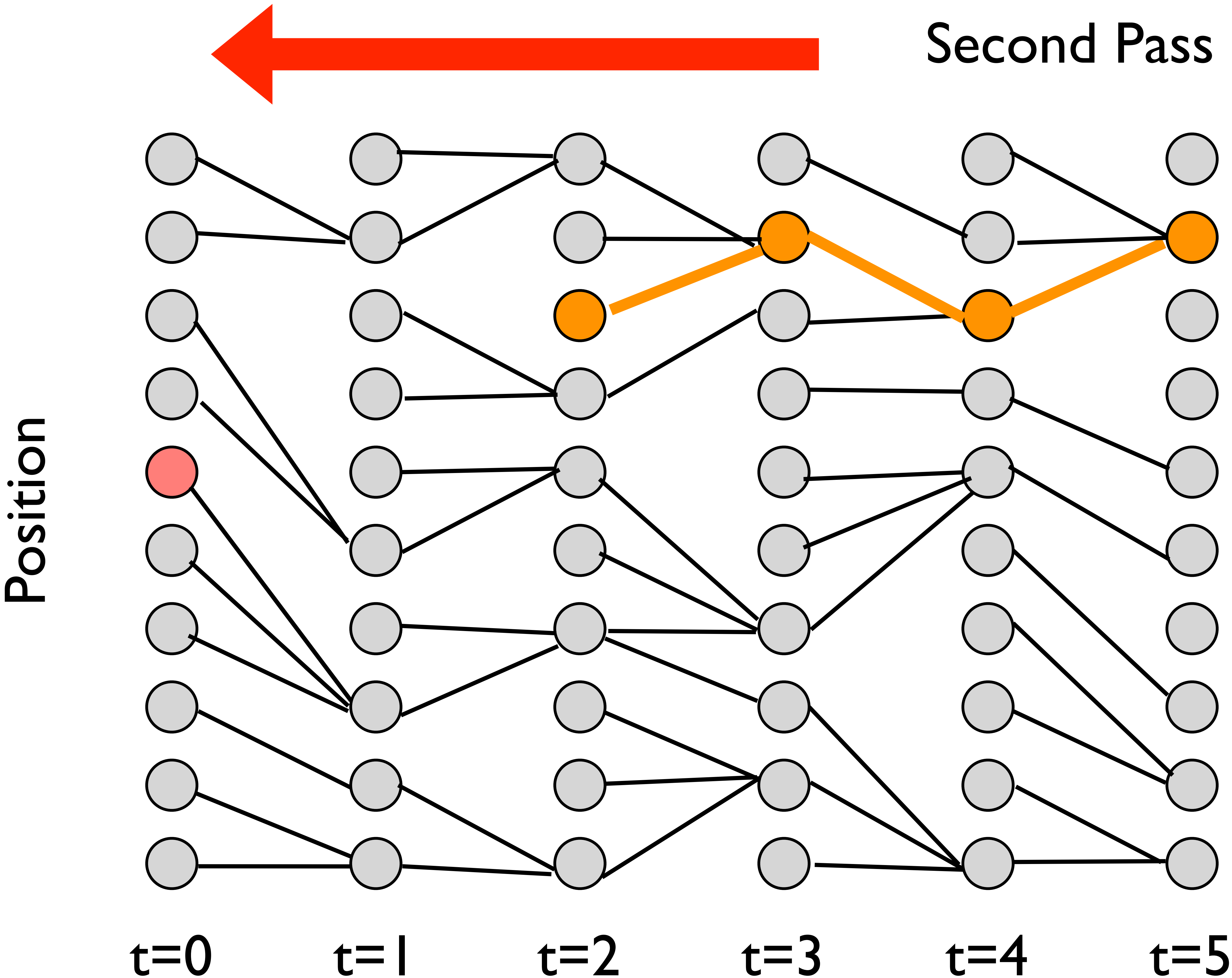

Position
t=0
t=1
t=2
t=3
t=4
t=5


Position
t=0
t=1
t=2
t=3
t=4
t=5

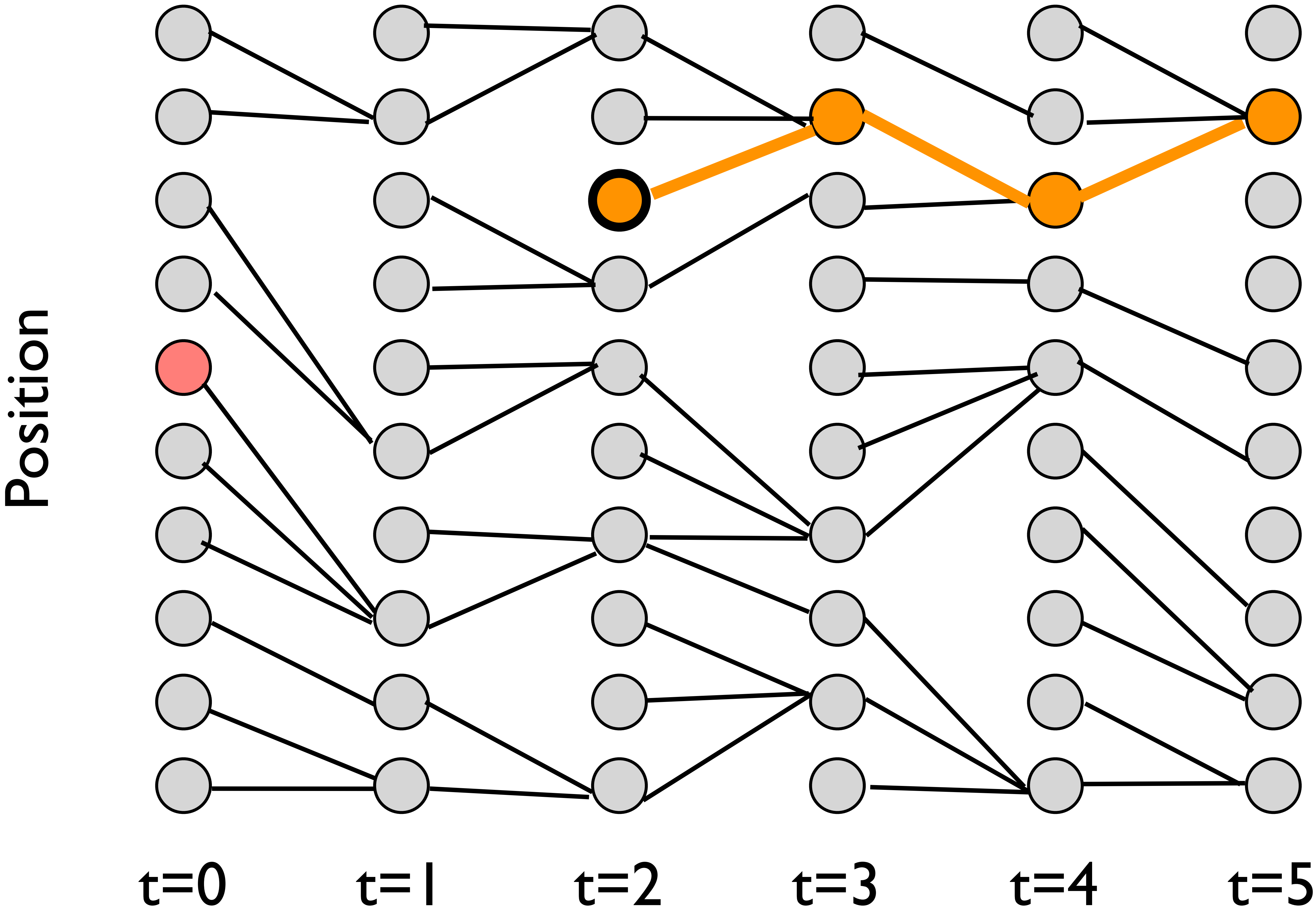

Position
t=0
t=1
t=2
t=3
t=4
t=5

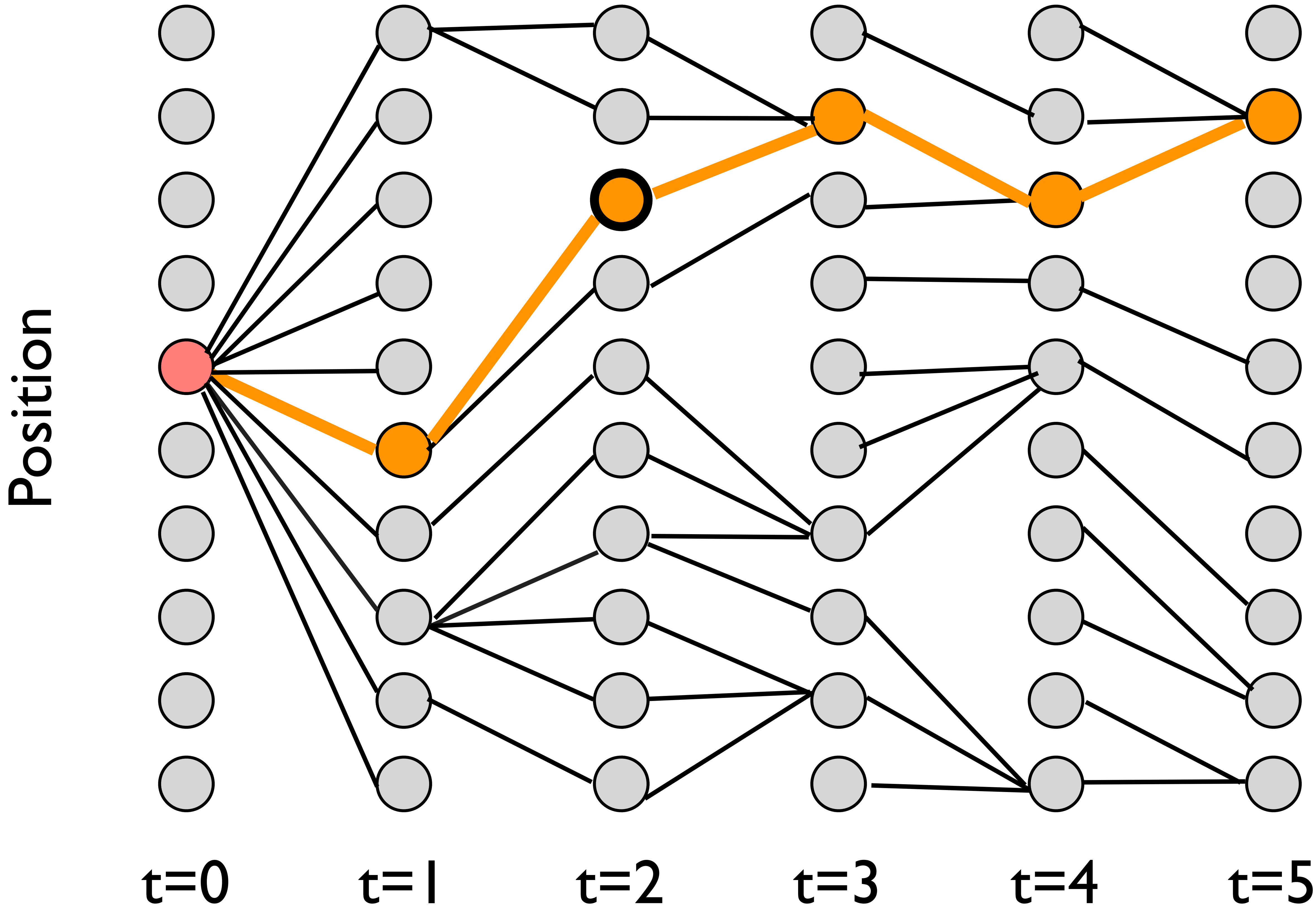


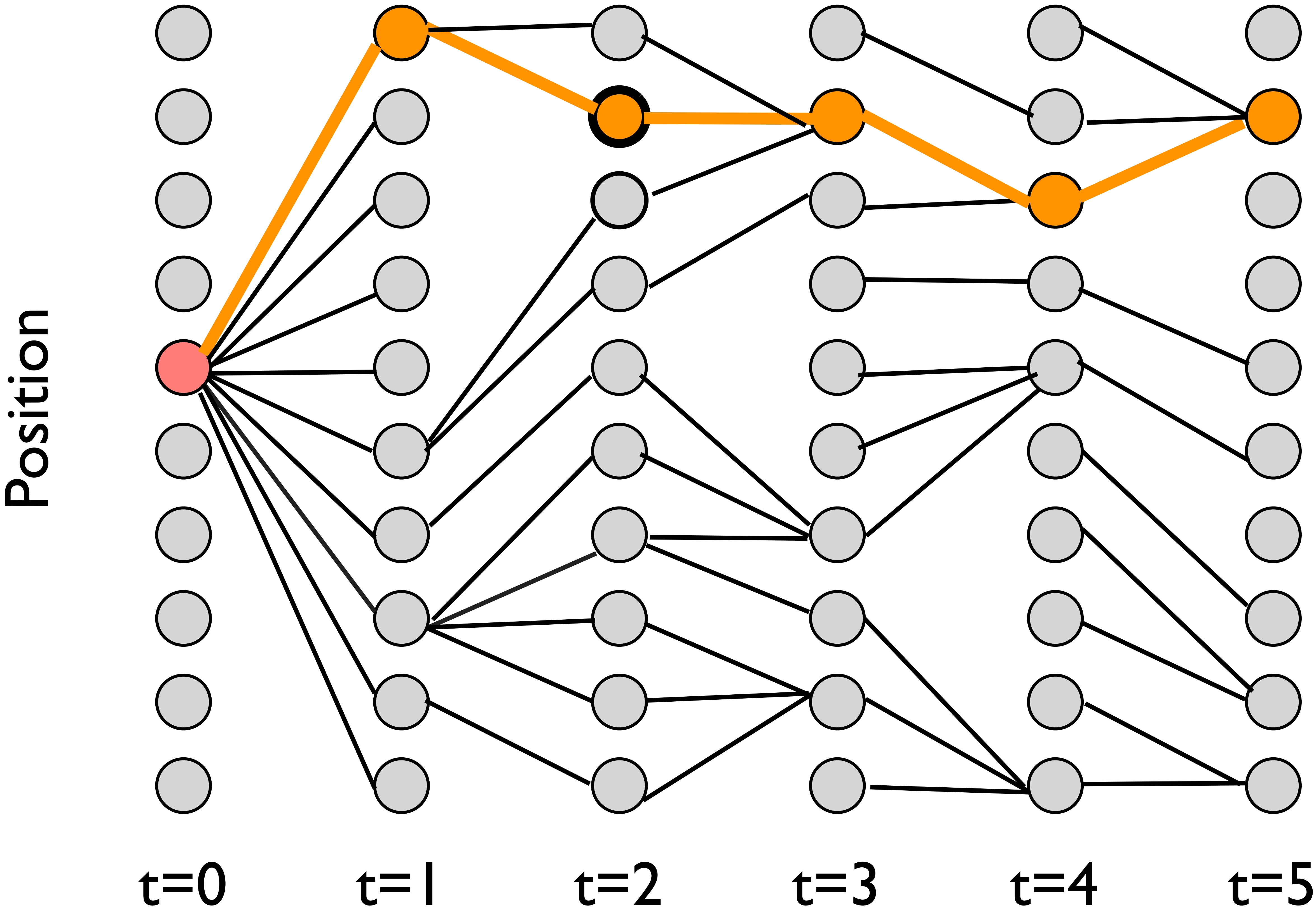
Second Pass
Position
t=0
t=1
t=2
t=3
t=4
t=5



Second Pass
Position
t=0
t=1
t=2
t=3
t=4
t=5


Position
t=0
t=1
t=2
t=3
t=4
t=5


Position
t=0
t=1
t=2
t=3
t=4
t=5


Position
t=0
t=1
t=2
t=3
t=4
t=5

Maximum expected label change:

$$t^* = \operatorname*{argmax}_{0 \le t \le T} \sum_{i=0}^{K} P(b_t^i) \cdot \Delta I(b_t^i) \quad \text{where} \quad \Delta I(b_t^i) = \sum_{j=0}^{T} \mathrm{err}(\mathrm{curr}_j, \mathrm{next}_j(b_t^i))$$

# “How much does label change from current estimate if this frame were annotated?”

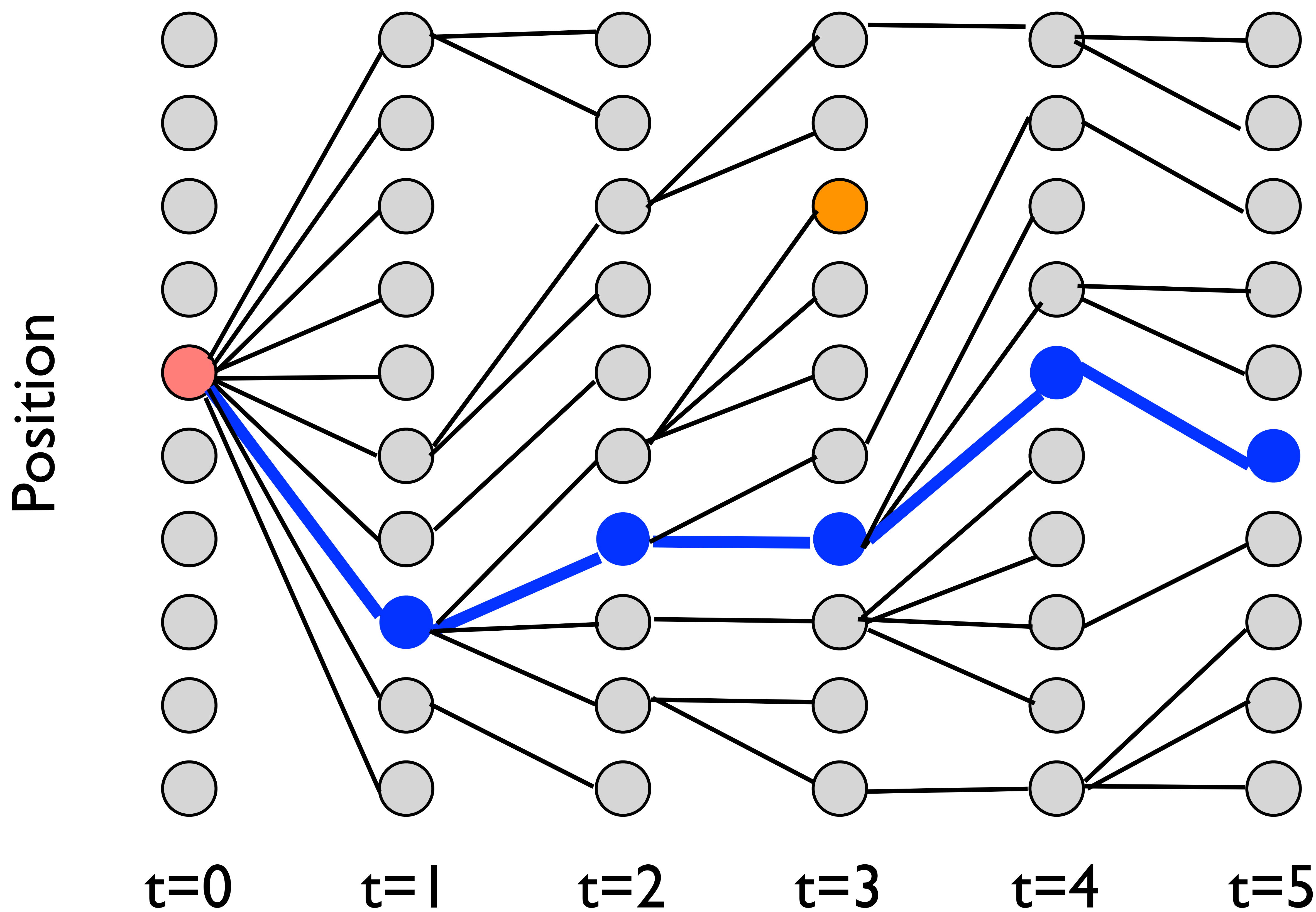
$$\Delta I(b_t^i) = \Theta_t^{\rightarrow}(b_t^i) + \Theta_t^{\leftarrow}(b_t^i) - \mathrm{err}(\mathrm{curr}_t, \mathrm{next}_t(b_t^i))$$

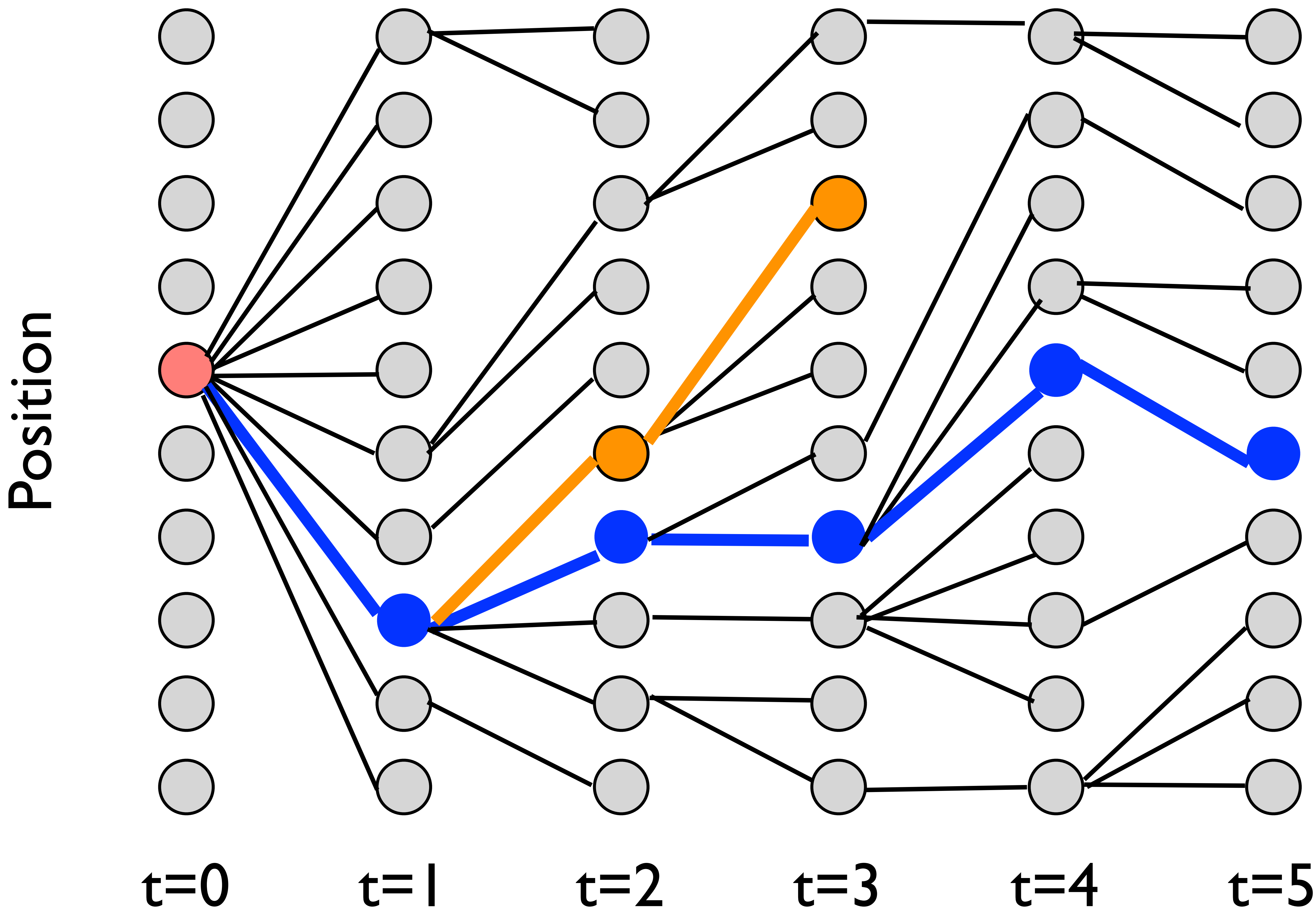
How much path changes had it gone through b_t^i

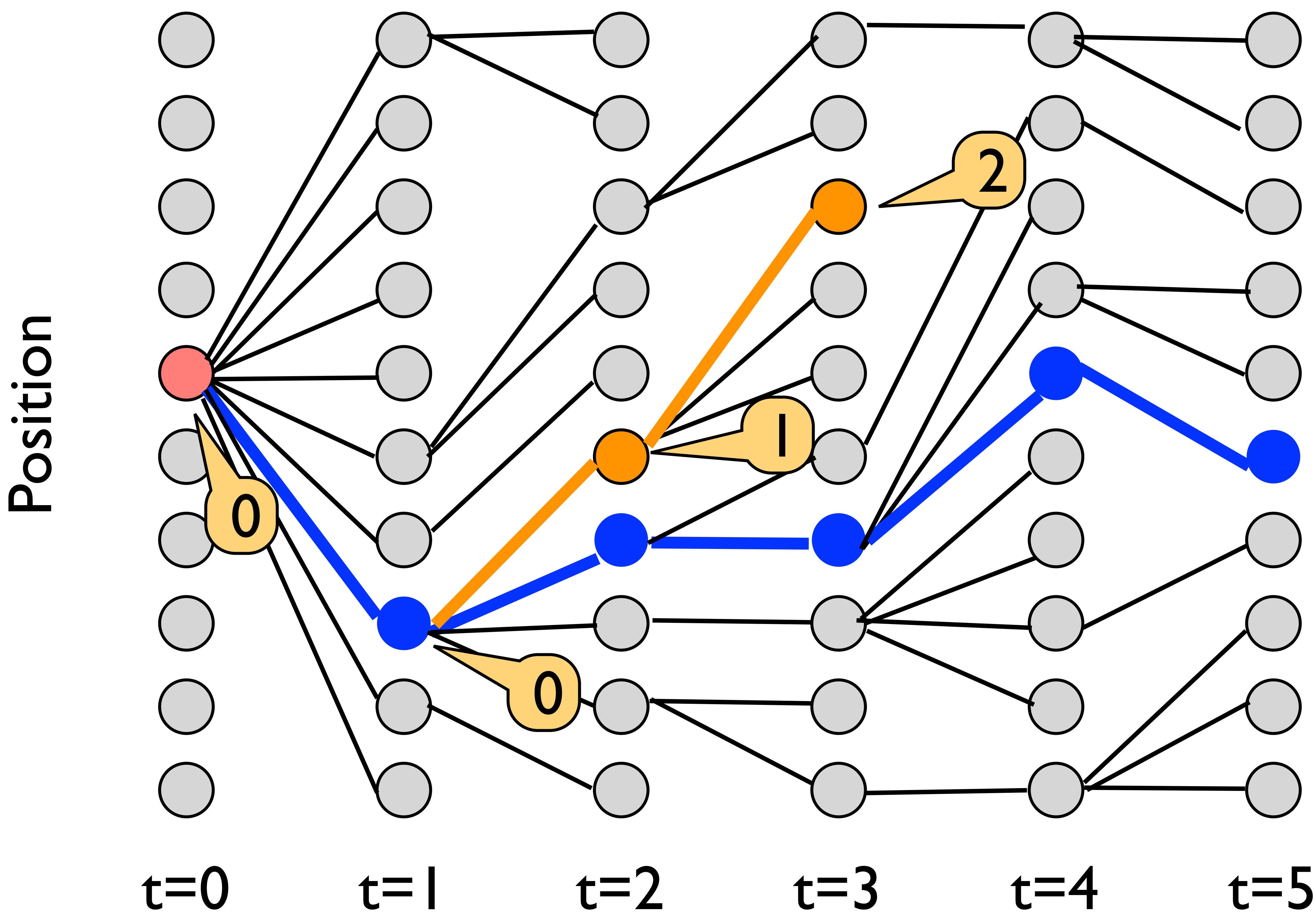
How much path changes up to this point

# Efficiently compute changes with DP:

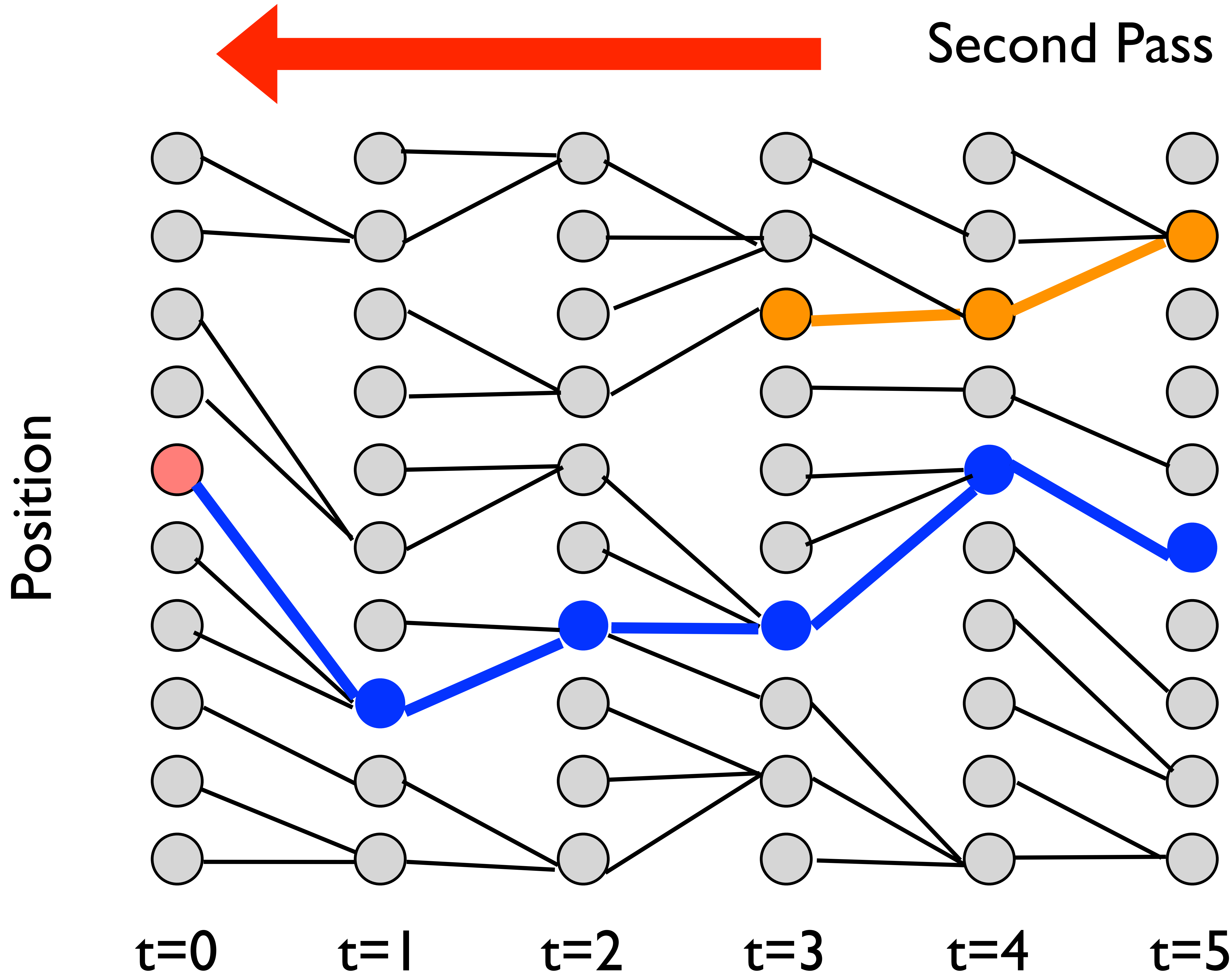
$$\Theta_0^{\rightarrow}(b_0) = \mathrm{err}(\mathrm{curr}_0, \mathrm{next}_0(b_0))$$
$$\Theta_t^{\rightarrow}(b_t) = \mathrm{err}(\mathrm{curr}_t, \mathrm{next}_t(b_t)) + \Theta_{t-1}^{\rightarrow}(\pi_t^{\rightarrow}(b_t))$$


Position
t=0
t=1
t=2
t=3
t=4
t=5

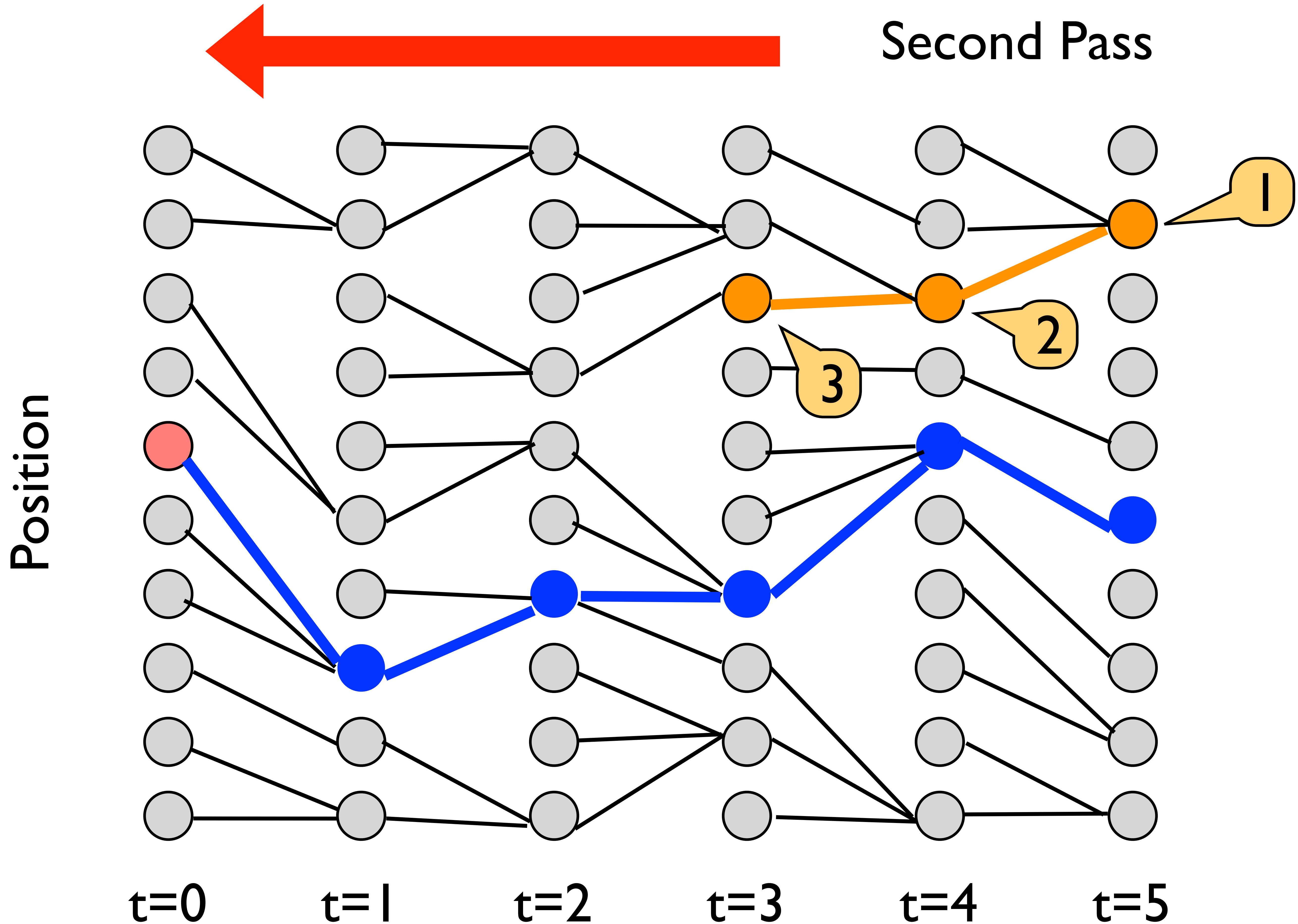

Position
t=0
t=1
t=2
t=3
t=4
t=5



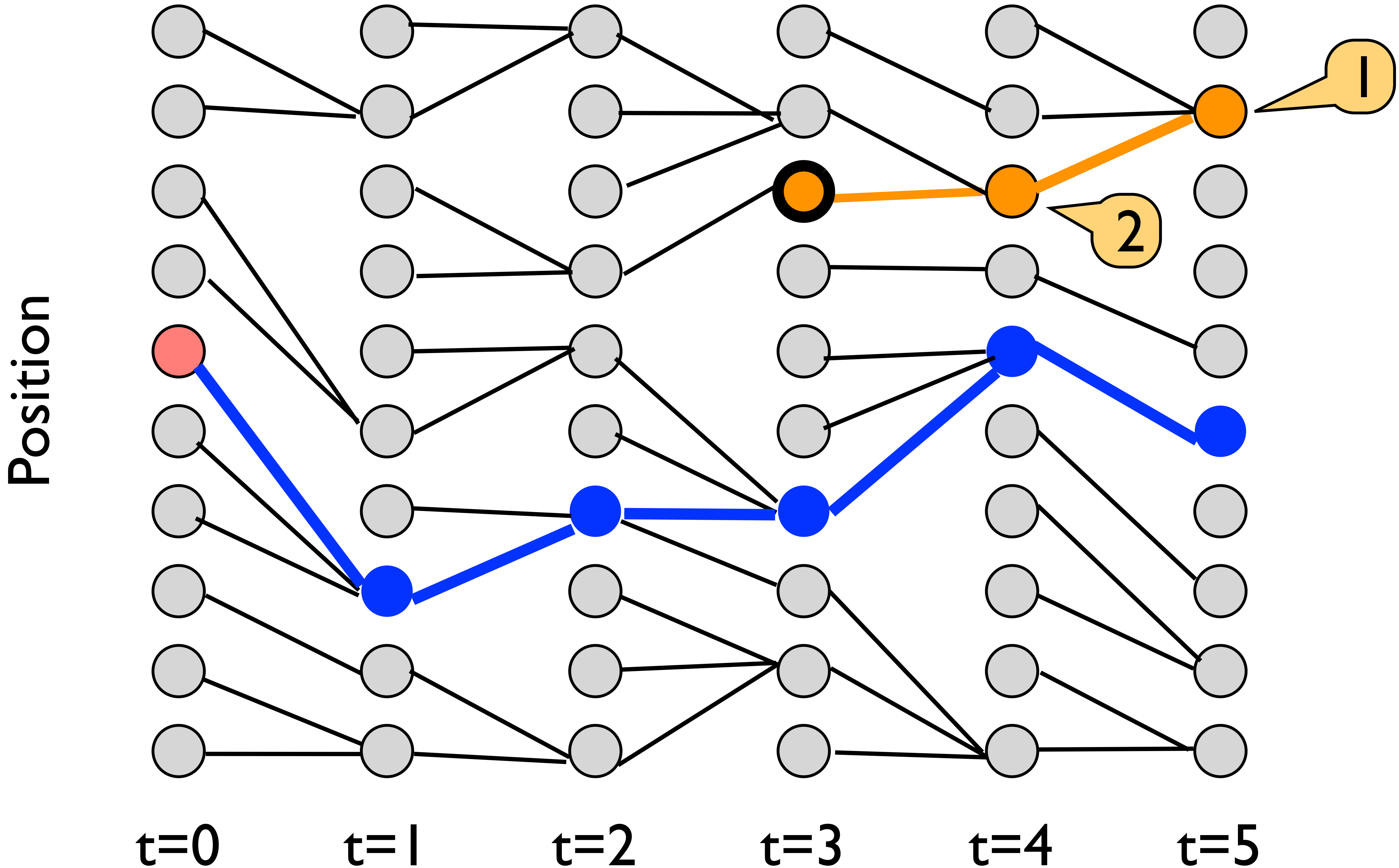
Position
0
0
1
2
t=0
t=1
t=2
t=3
t=4
t=5

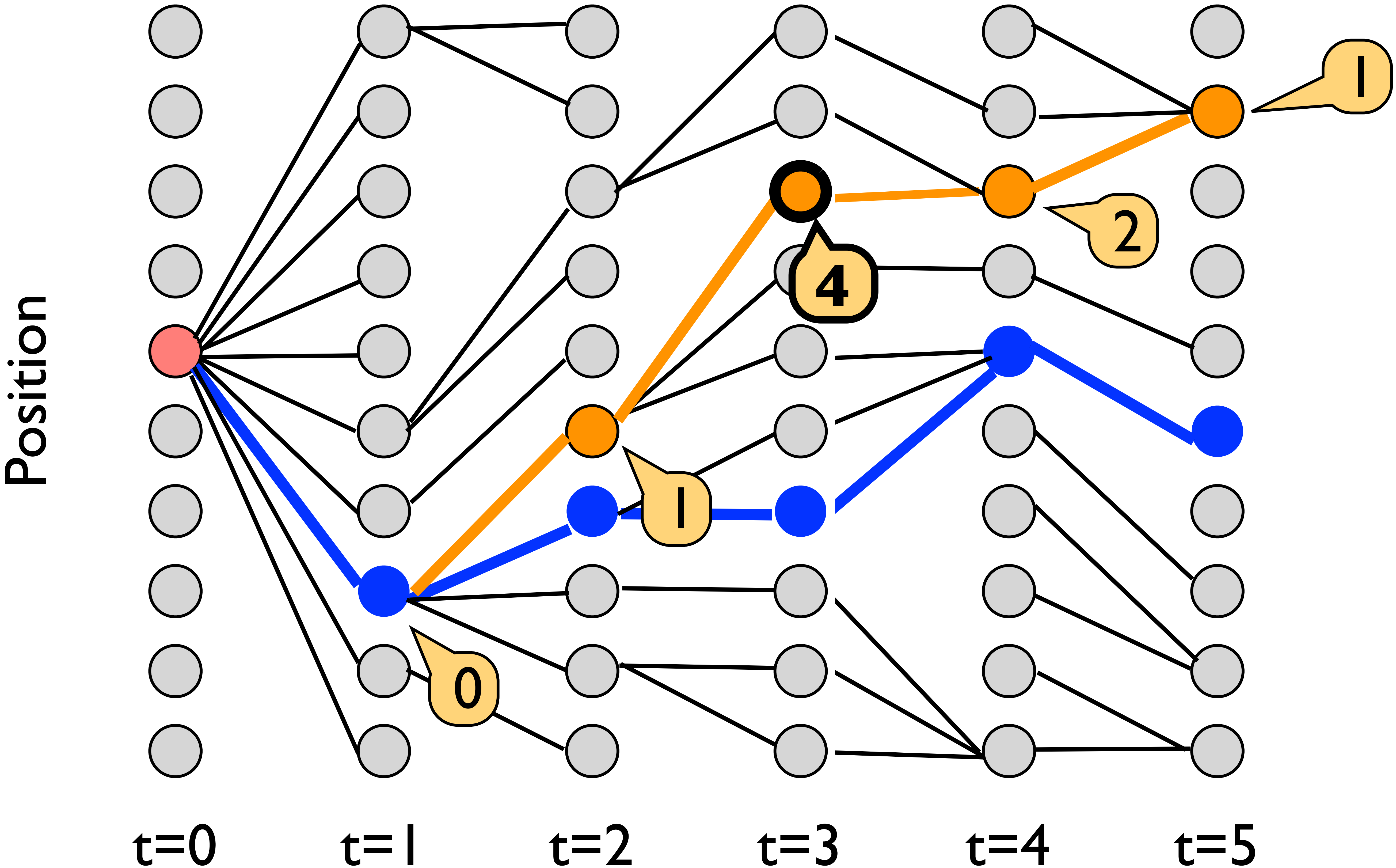


Second Pass
Position
t=0
t=1
t=2
t=3
t=4
t=5



Second Pass
Position
1
2
3
t=0
t=1
t=2
t=3
t=4
t=5


Position
1
2
t=0
t=1
t=2
t=3
t=4
t=5


Position
1
2
4
1
0
t=0
t=1
t=2
t=3
t=4
t=5

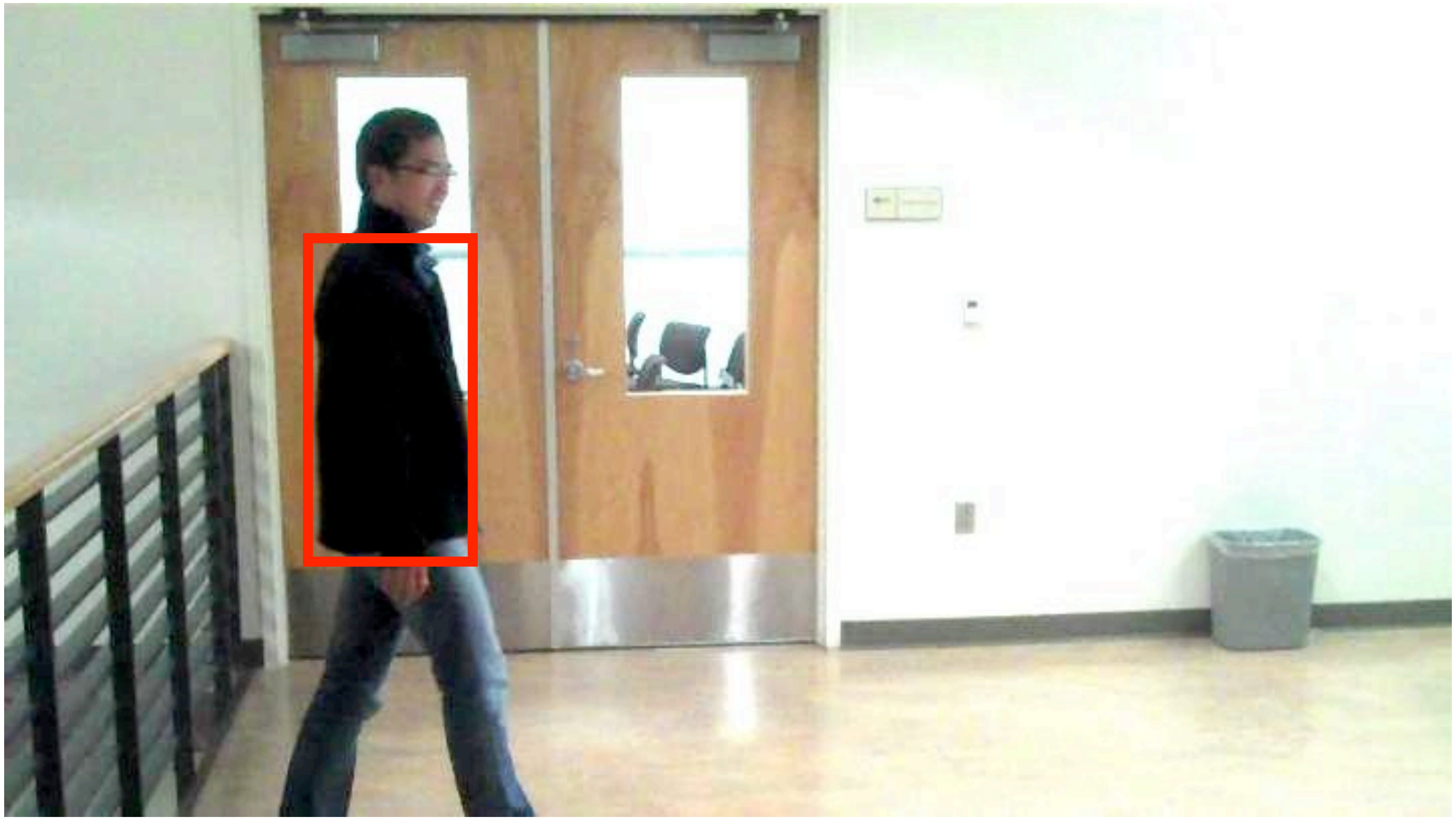
Maximum expected label change:

$$t^* = \operatorname*{argmax}_{0 \le t \le T} \sum_{i=0}^{K} P(b_t^i) \cdot \Delta I(b_t^i) \quad \text{where} \quad \Delta I(b_t^i) = \sum_{j=0}^{T} \operatorname{err}(\operatorname{curr}_j, \operatorname{next}_j(b_t^i))$$

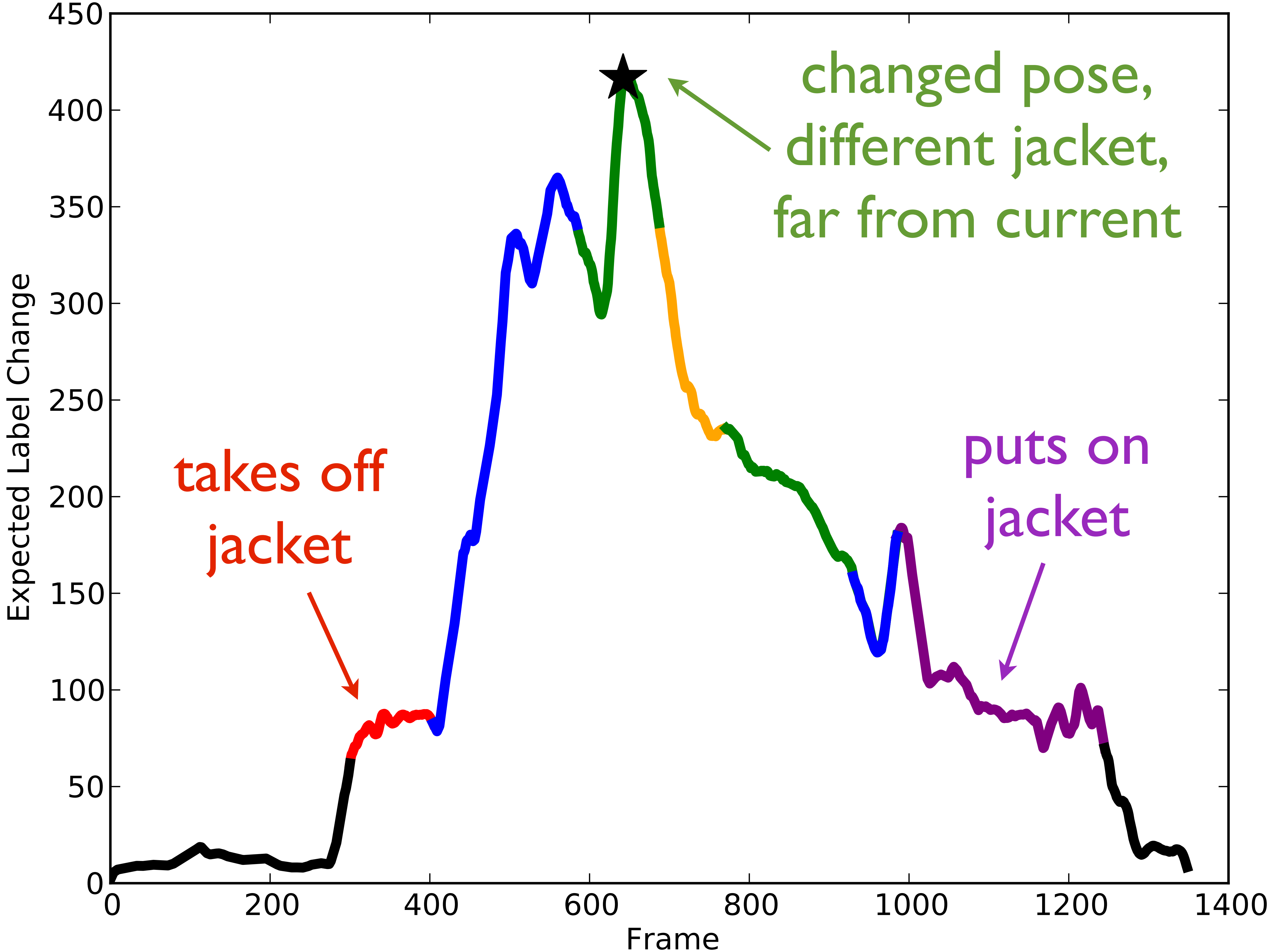
Stop requesting annotations when we don't expect a large label change:

$$\max_{0 \leq t \leq T} \sum_{i=0}^{K} P(b_t^i) \cdot \Delta I(b_t^i) < \text{tolerance}$$

In practice, budget expires first!

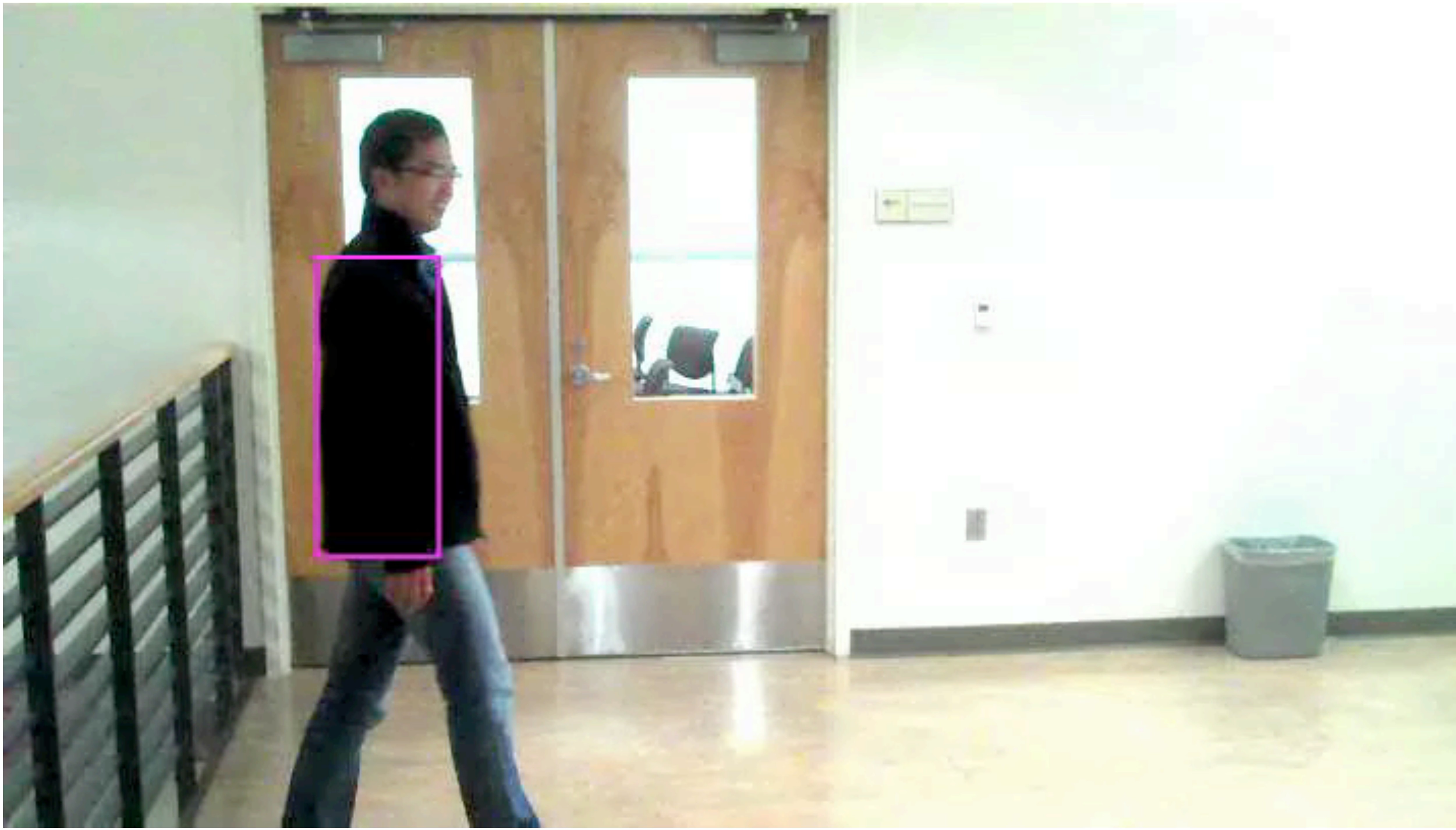


# First frame labeled only:

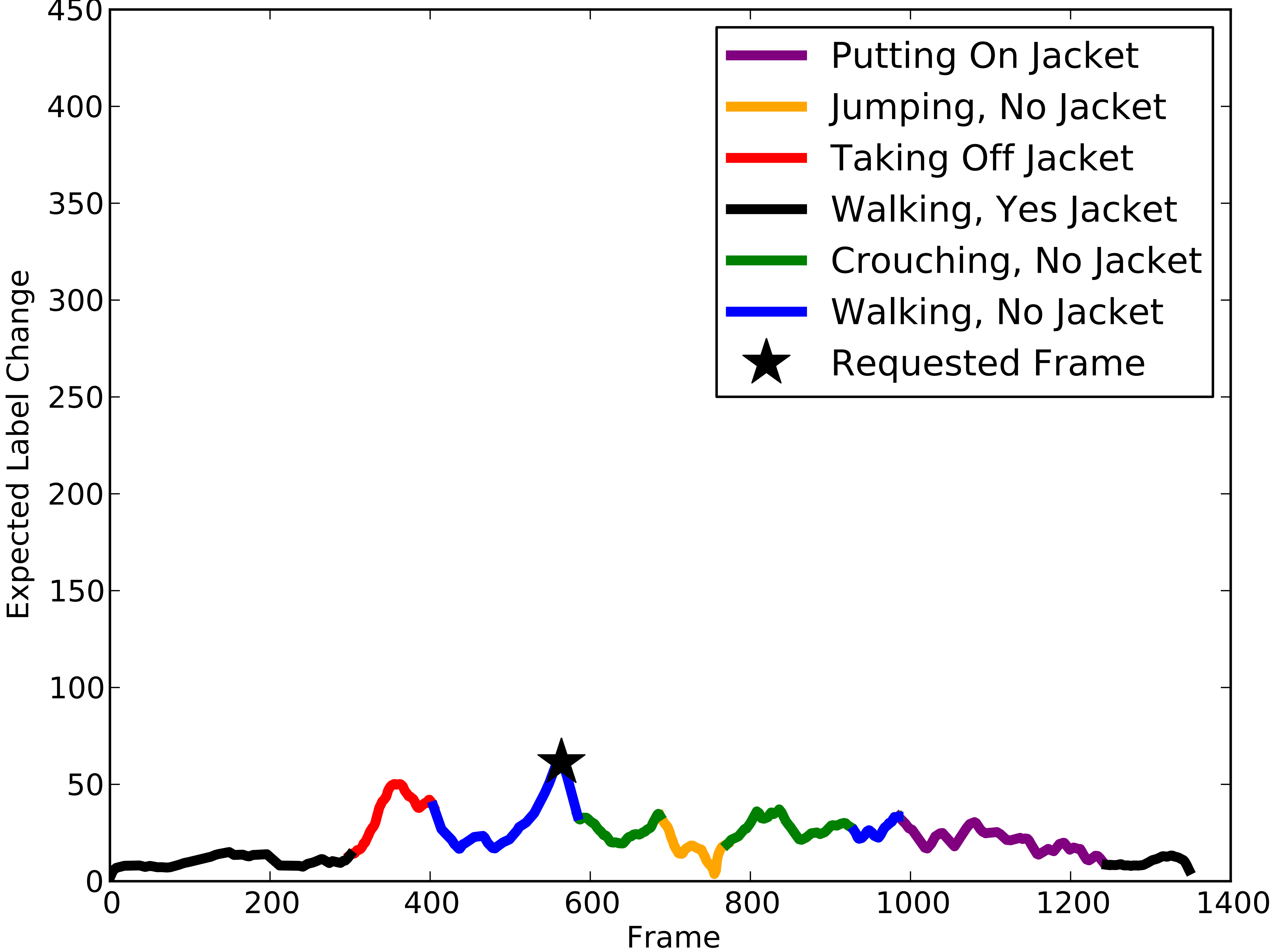


# Requested Frame

# Requested Frame

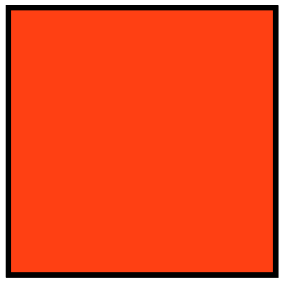


# with requested frame:

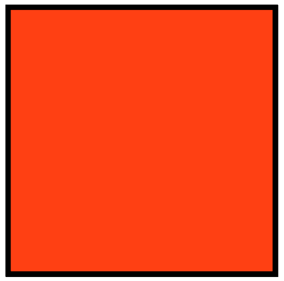


[redacted]

[redacted]

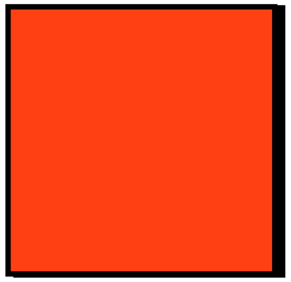


Pass Through



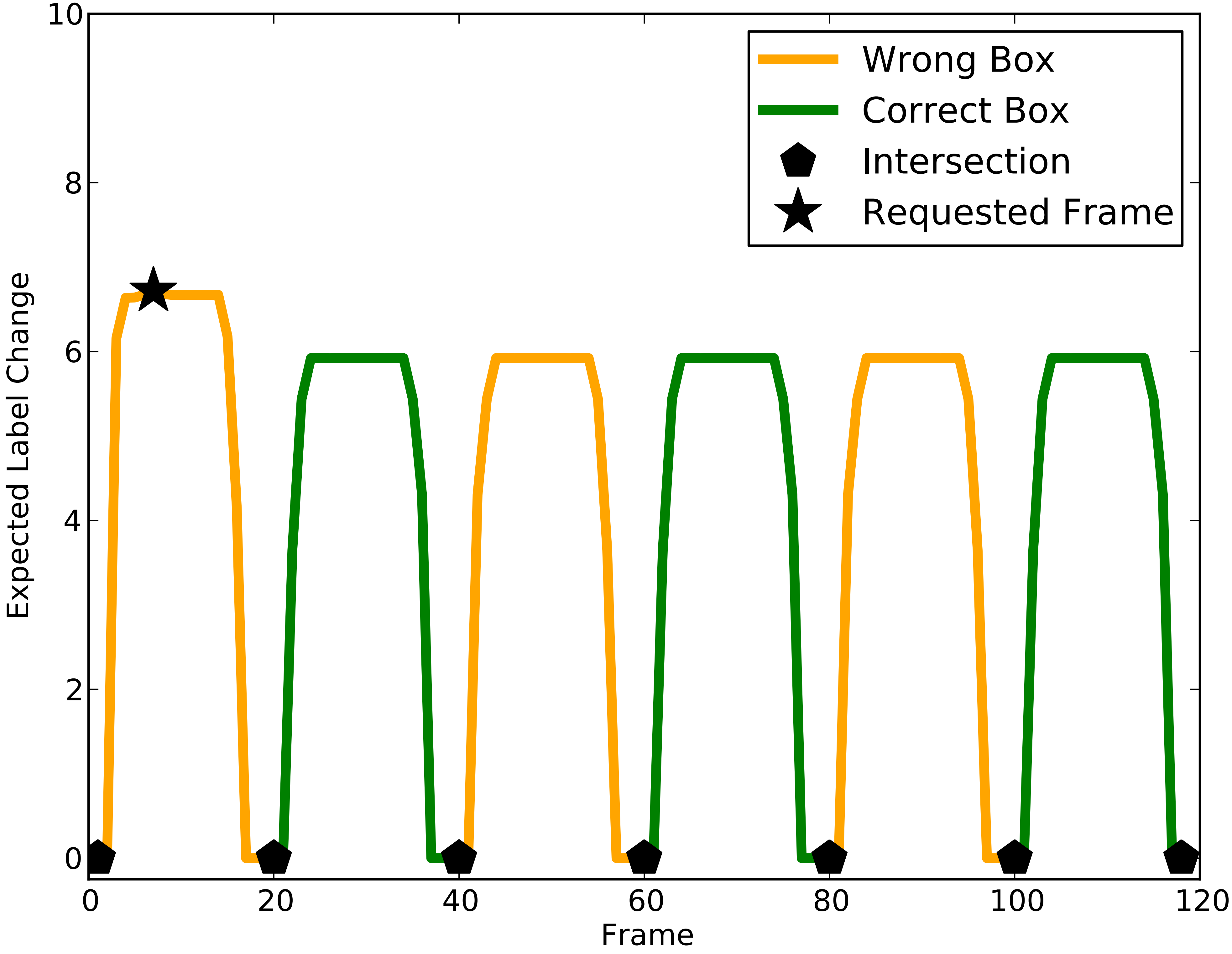
Pass Through

Bounce

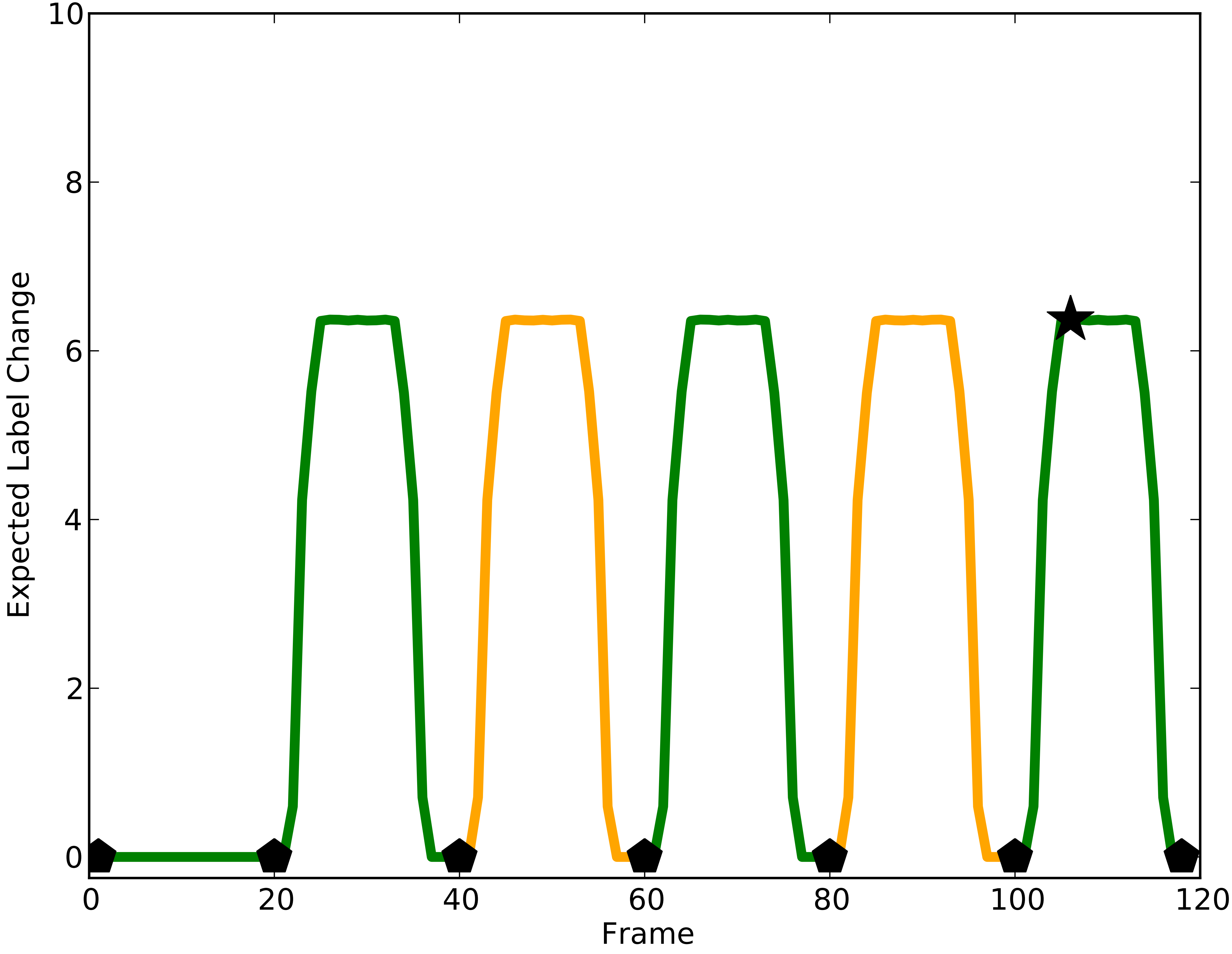


Bounce

Which one happened?



Wrong Box
Correct Box
Intersection
Requested Frame
Expected Label Change
Frame
0
2
4
6
8
10
0
20
40
60
80
100
120



Expected Label Change
Frame
0
2
4
6
8
10
0
20
40
60
80
100
120

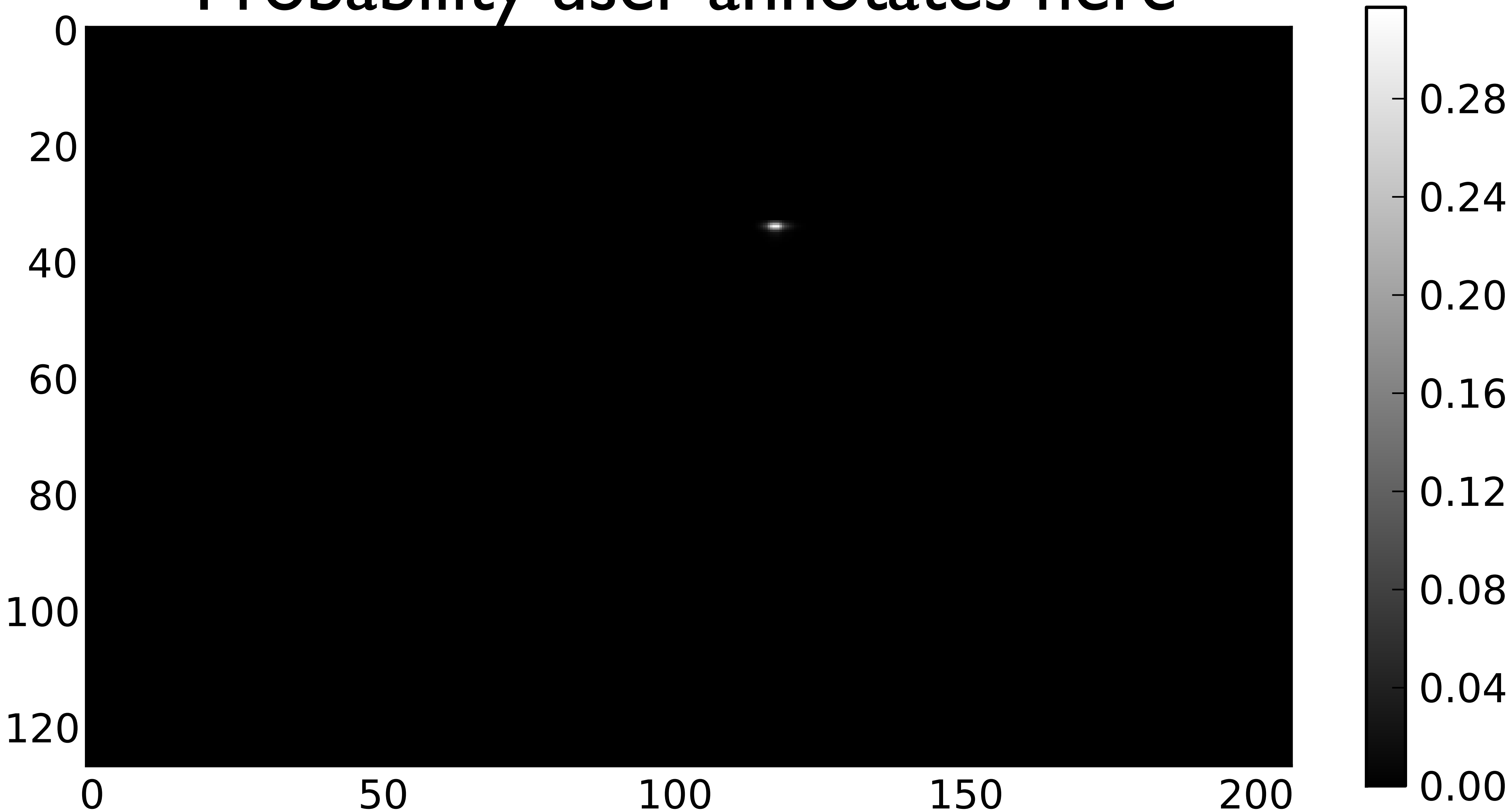
[redacted]

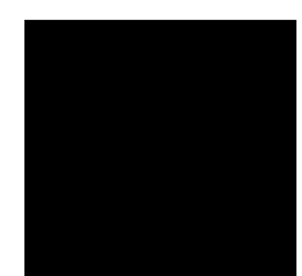
[redacted]

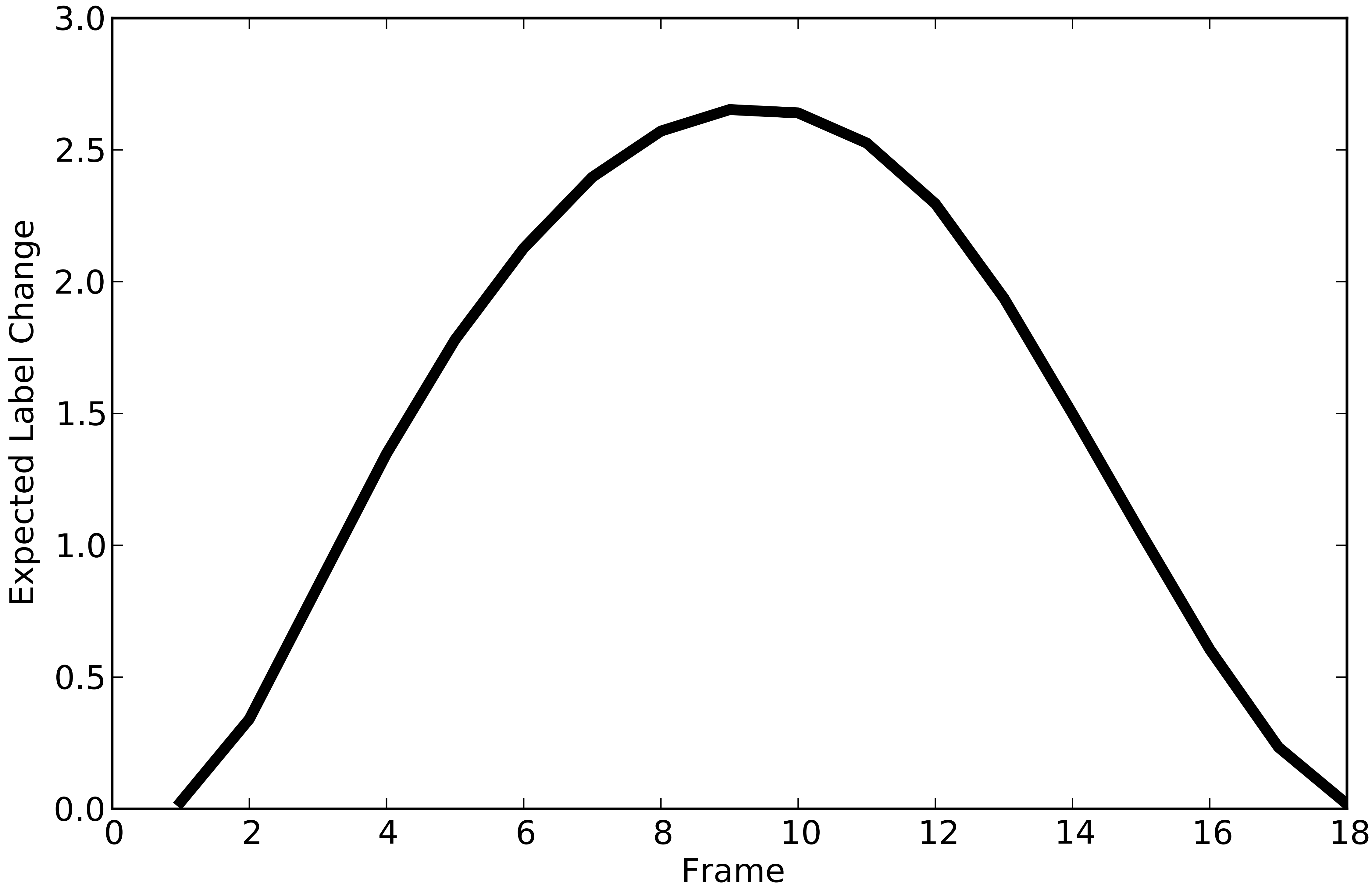
[redacted]

[redacted]

# Probability user annotates here

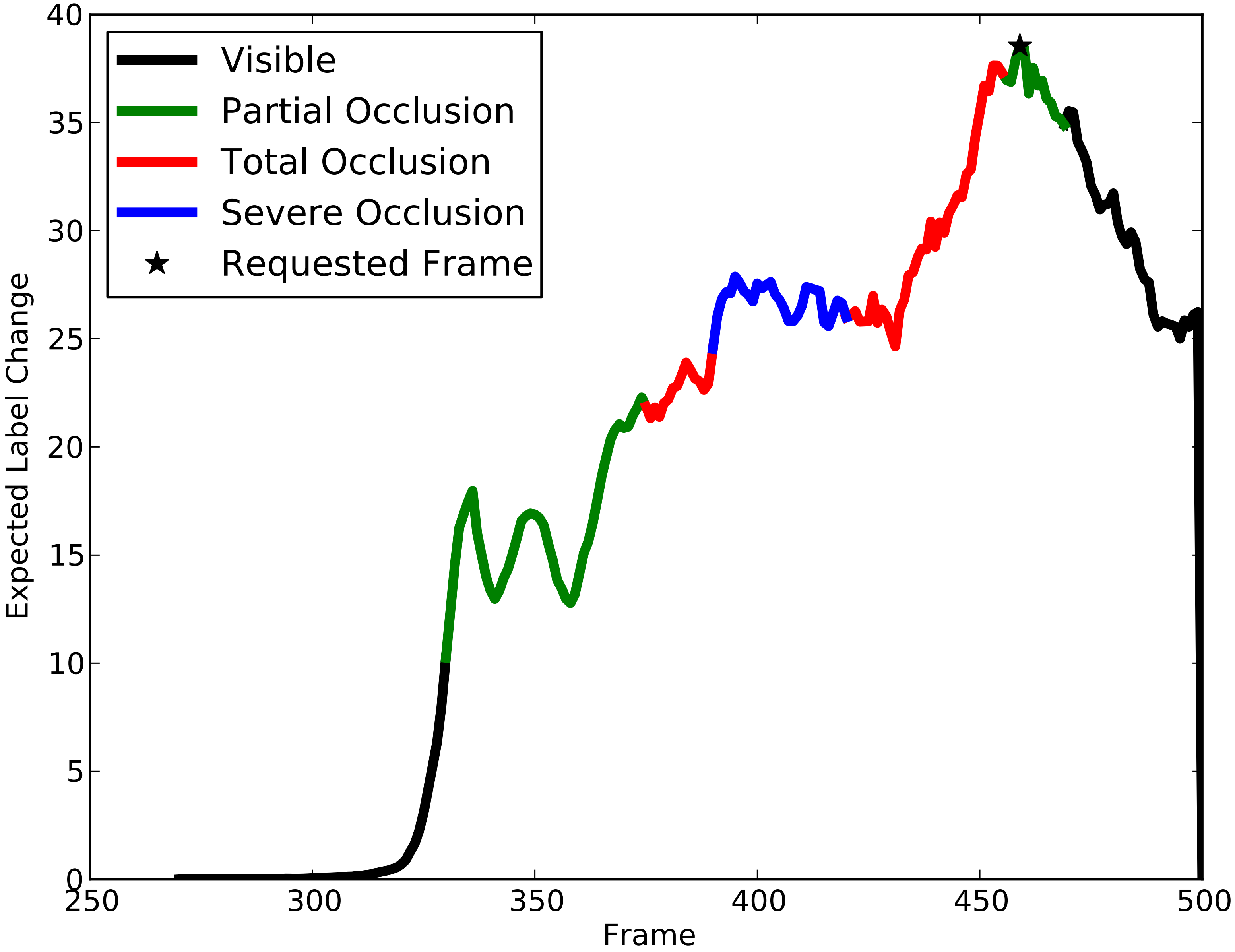


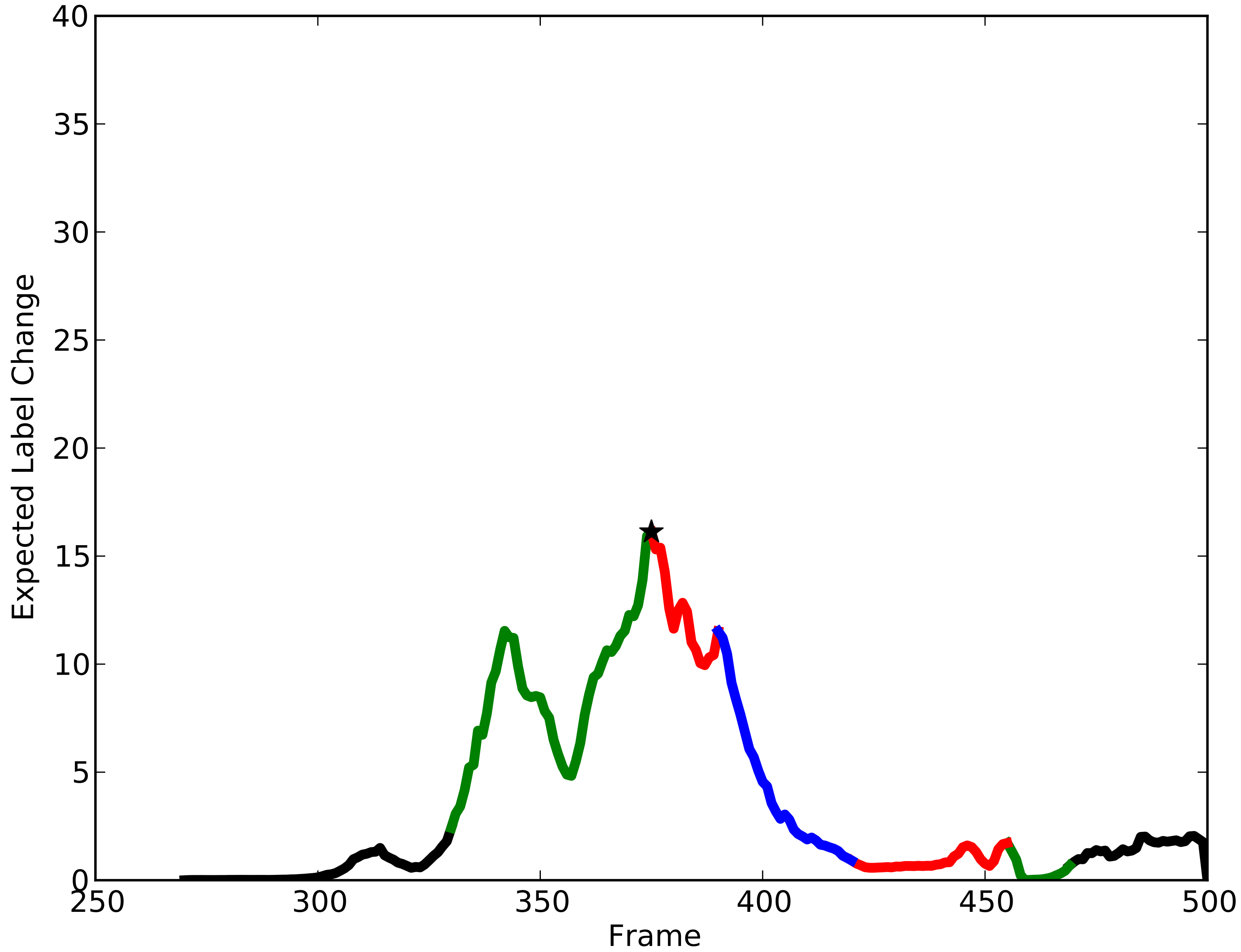


3.0
2.5
2.0
1.5
1.0
0.5
0.0
Expected Label Change
0
2
4
6
8
10
12
14
16
18
Frame

# Tracking Under Occlusion

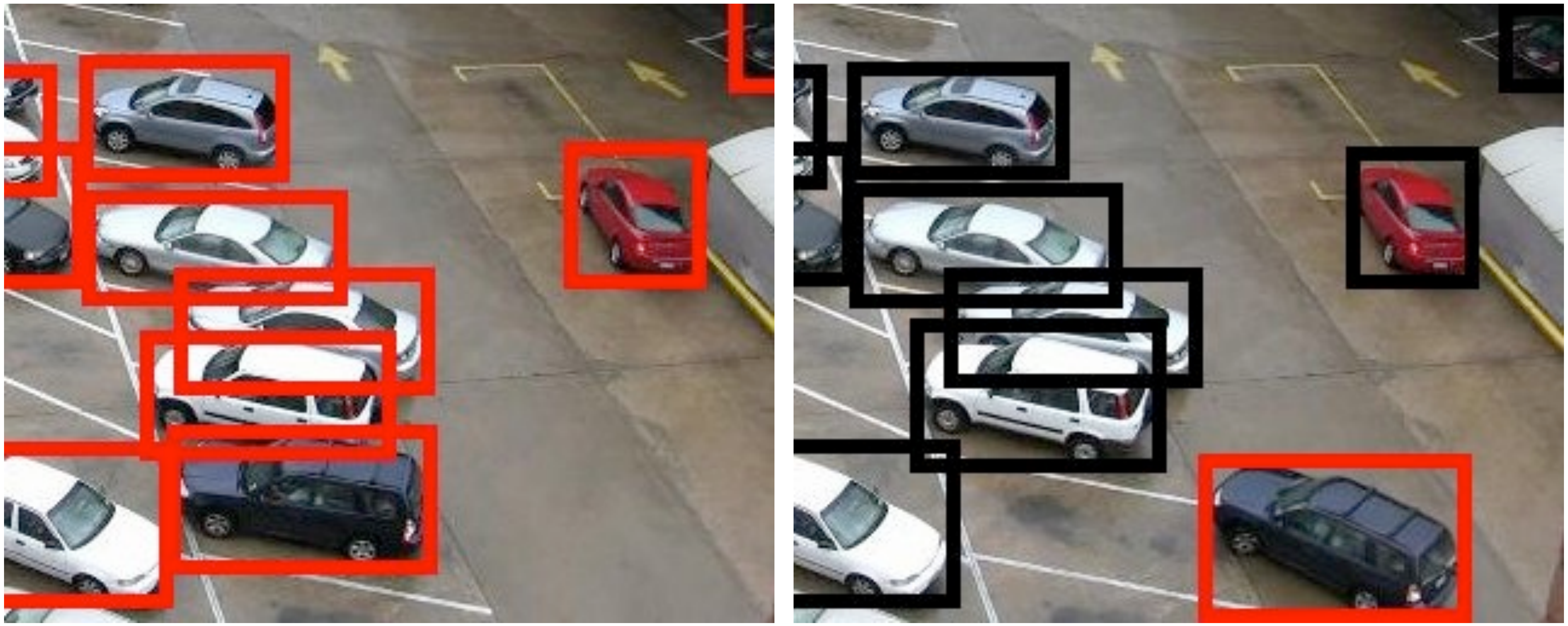



# Tracking Under Occlusion



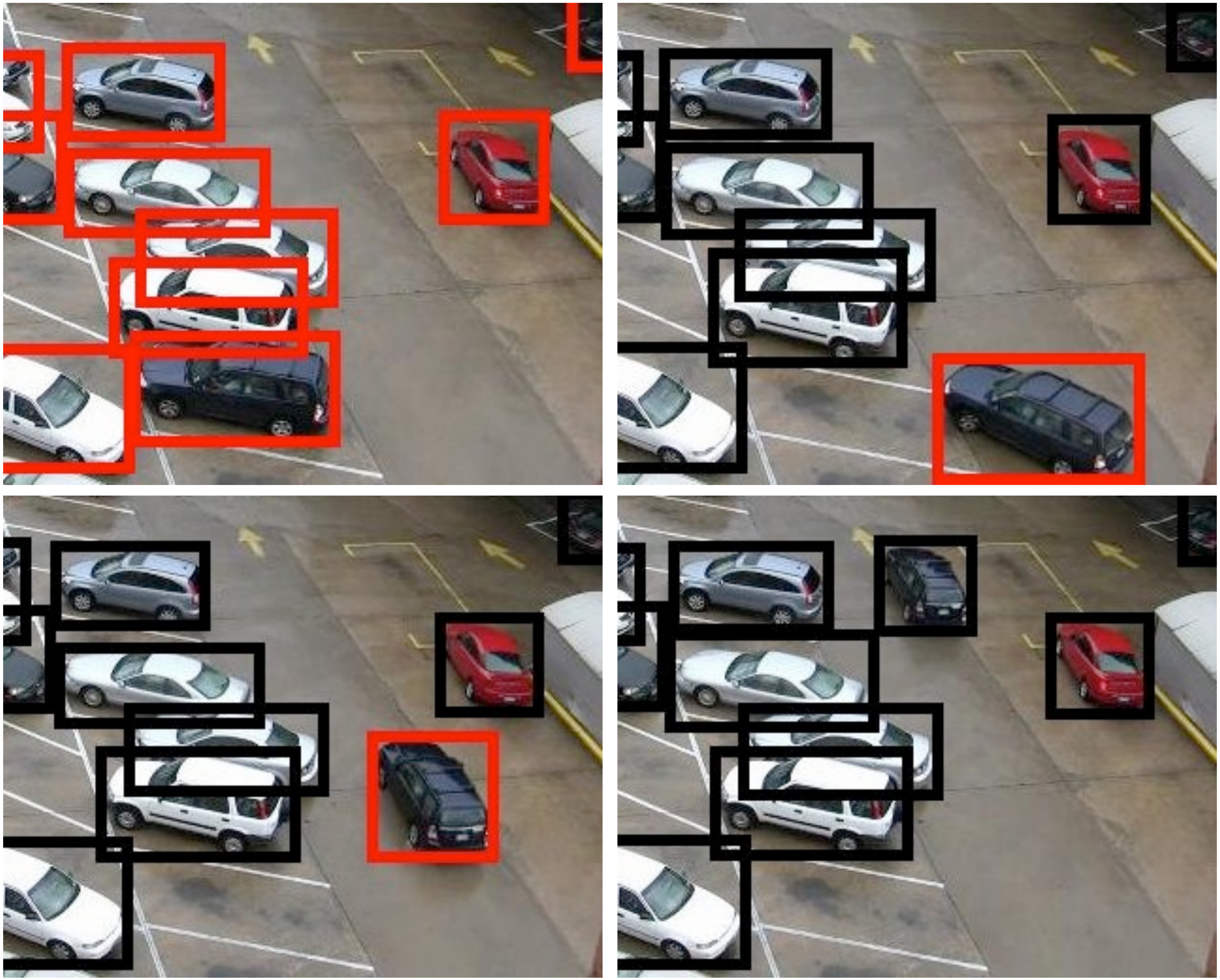

# Transferring Wasted Clicks

# Transferring Wasted Clicks



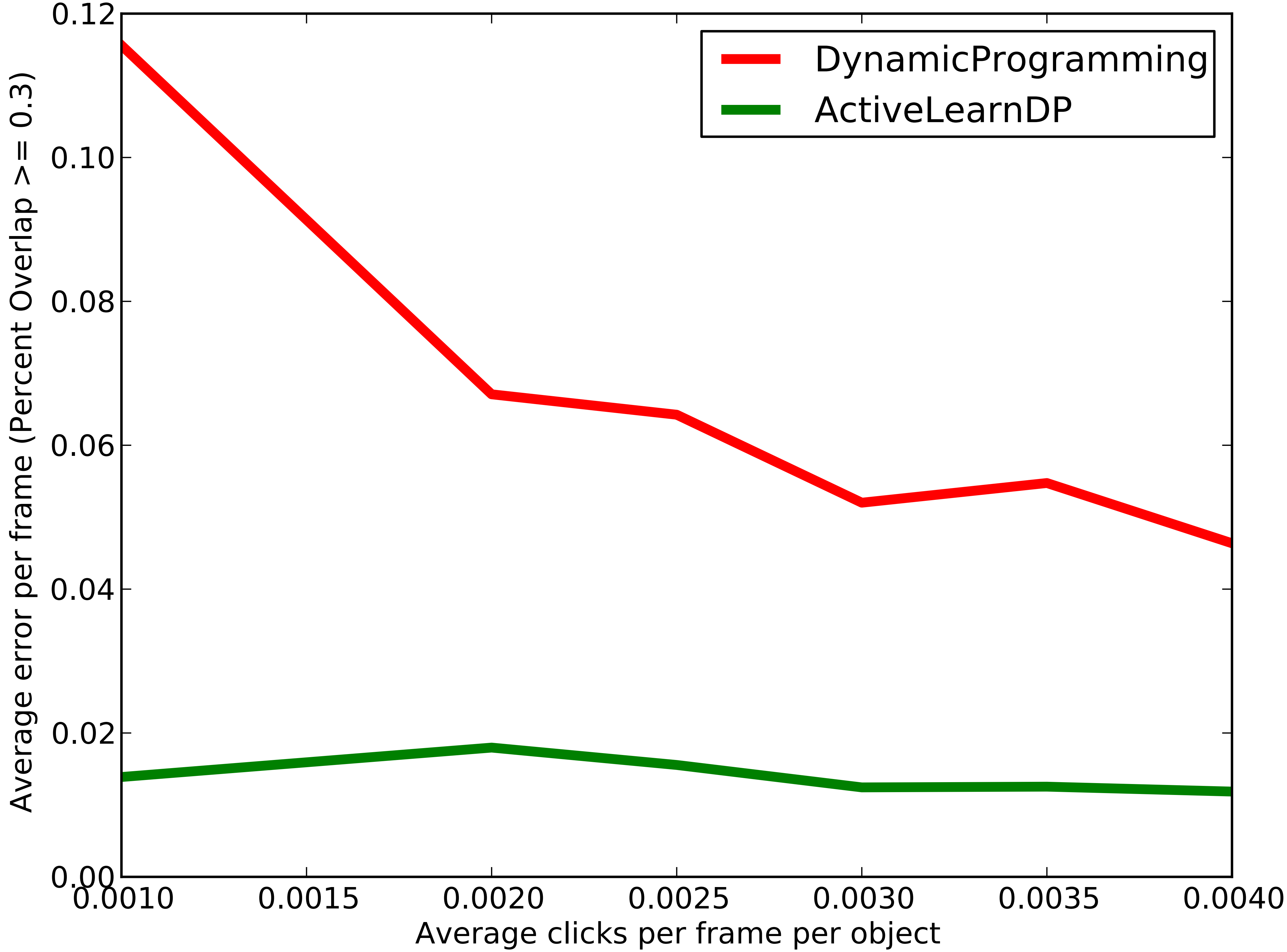
# Transferring Wasted Clicks

# Transferring Wasted Clicks



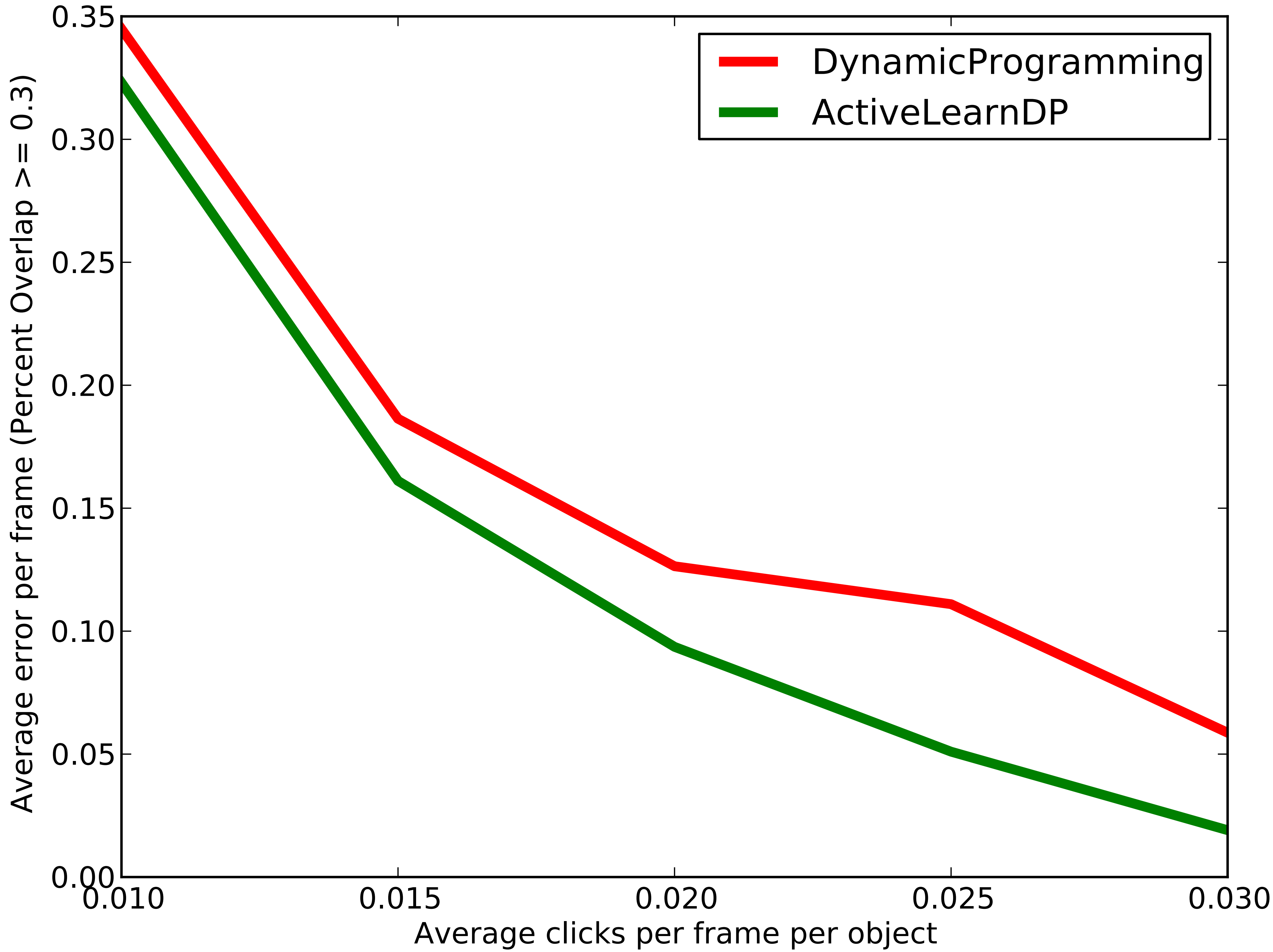
# Benchmark Evaluation: VIRAT

# Performance on VIRAT Cars



# Benchmark Evaluation: Basketball

# Performance on Basketball Players



# Summary

- Humans do not pick an optimal set of key frames
- Humans do not intuitively understand the behavior of *any* imperfect interpolation scheme

- Active learning with a tracker picks better key frames, reducing costs

# Future work

**Hypothesis:** The order of requested frames is crucial for the user experience.

How can we do far-sighted active learning for video annotation?

Huge impact:

| | | |
|---|---|---|
| $15,000 | → | $1,500 |
| 8 months | → | 24 days |

What would you buy with the extra

$13,500?

Thanks!