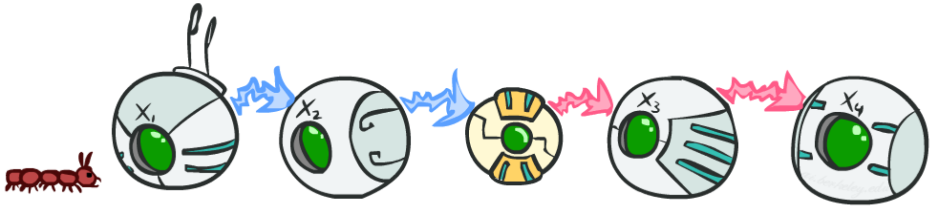
Markov Models, Particle Filters



CSE 473: Introduction to Artificial Intelligence

James Weichert (jpw@cs)
cs.uw.edu/473
acknowledgements

Welcome!
What's your favorite summertime activity?
sli.do **#cse473**

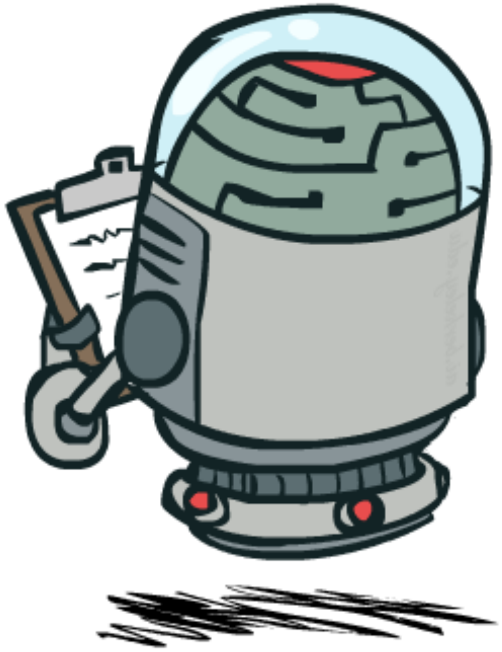
Administrivia

ANNOUNCEMENTS

- **Test 2** next Friday

ASSIGNMENTS

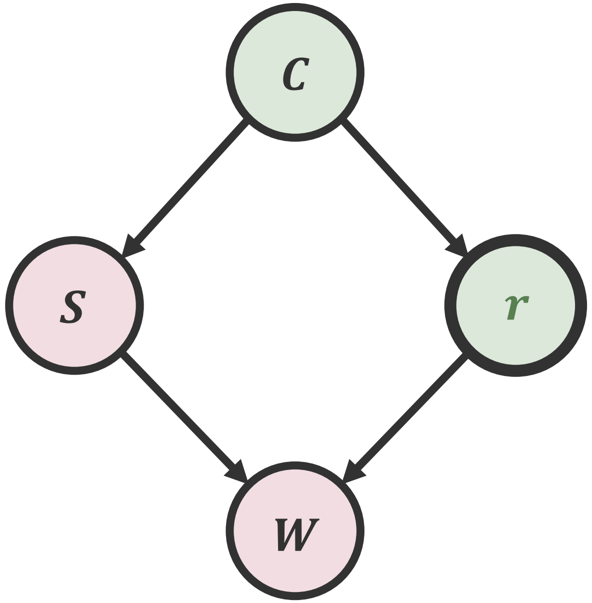
- **Project 3 due** tomorrow (7.30)
- **Homework 3 due** next Thursday (8.6)



Uncertainty and Time

CONTEXT

So far, we've focused on arbitrary arrangements of random variables



What about scenarios involving ***sequences*** *of observations?*

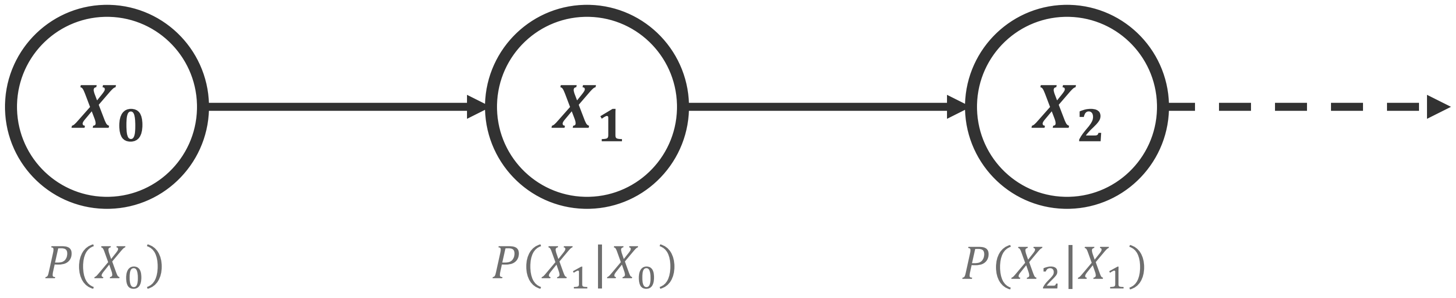
- Applications
  - Forecasting
  - Speech recognition
  - Robot localization
  - User attention
  - Medical monitoring

All involve reasoning over 'time'

Markov Models

DEFINITION

- Value of X at time t is a state (usually discrete, finite)

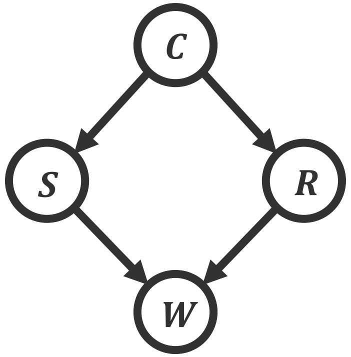


- Transition Probabilities $P(X_t|X_{t-1})$
  - With the Markov assumption: $P(X_t|X_{t-1}) = P(X_t|X_0, \ldots, X_{t-1})$
  - First-order Markov model implies X_{t+1} independent of $X_0, \ldots, X_{t-1}$ given X_t
  - **Stationarity**: transition probabilities do not change

- Joint Distribution: $P(X_0, \ldots, X_T) = P(X_0) \prod_t P(X_t|X_{t-1})$

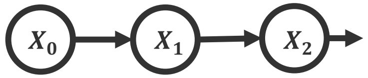
Bayes Nets and Markov Models

CONTEXT



Similarities

- Directed acyclic graphs
- Infer joint probabilities from conditional distributions
- Markov independence



Differences

- Markov models can involve infinitely many variables
- Repetition of transition model not part of BN syntax

QUESTIONS

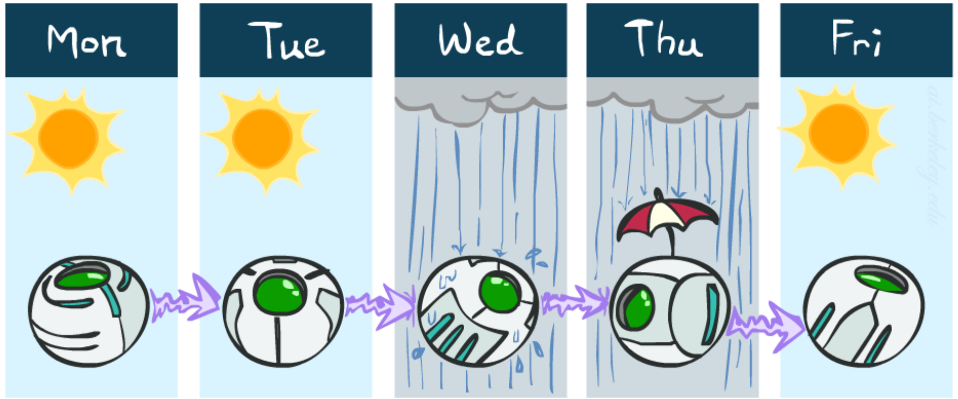
Questions?

live and on sli.do #cse473

Weather

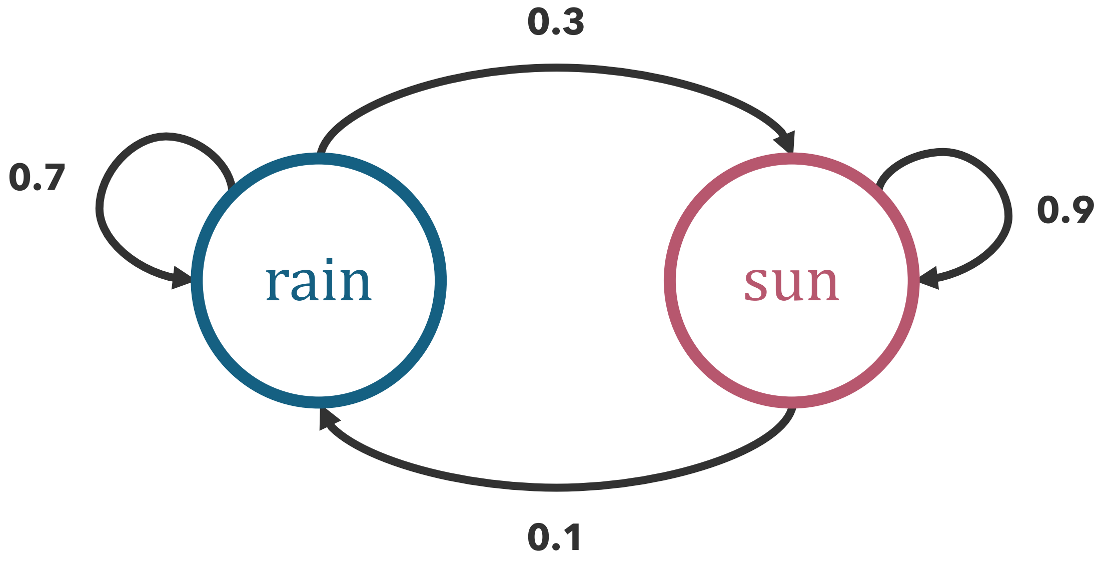
EXAMPLE

Goal: What's the weather like tomorrow?



- States: $\{\text{rain}, \text{sun}\}$
- Initial Distribution
  - $P(X_0 = \text{rain}) = 0.5$
  - $P(X_0 = \text{sun}) = 0.5$
- Transition Model $P(X_t | X_{t-1})$

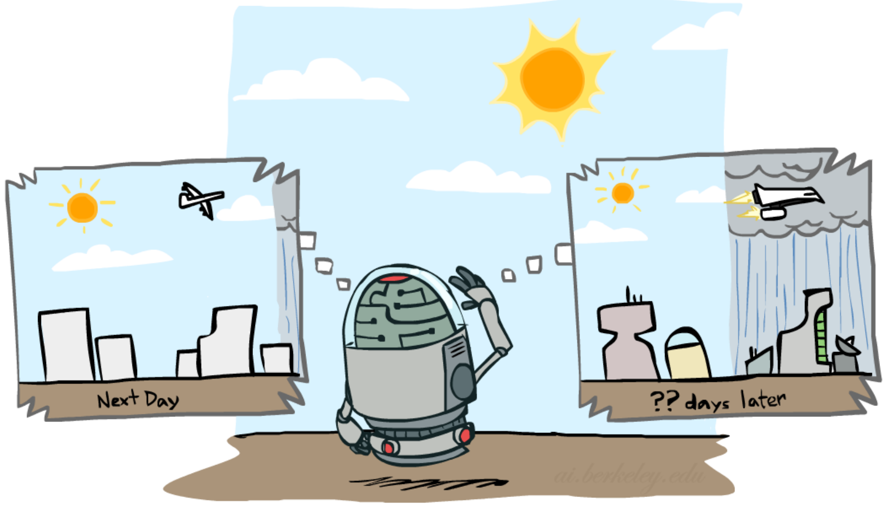
X_{t-1}	X_t	$P(X_t \vert X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	sun	0.7



Weather Prediction (1)

EXAMPLE

Goal: What's the weather at time 1?



- Initial Distribution
  - $P(X_0 = \text{rain}) = 0.5$
  - $P(X_0 = \text{sun}) = 0.5$

- Transition Model $P(X_t | X_{t-1})$

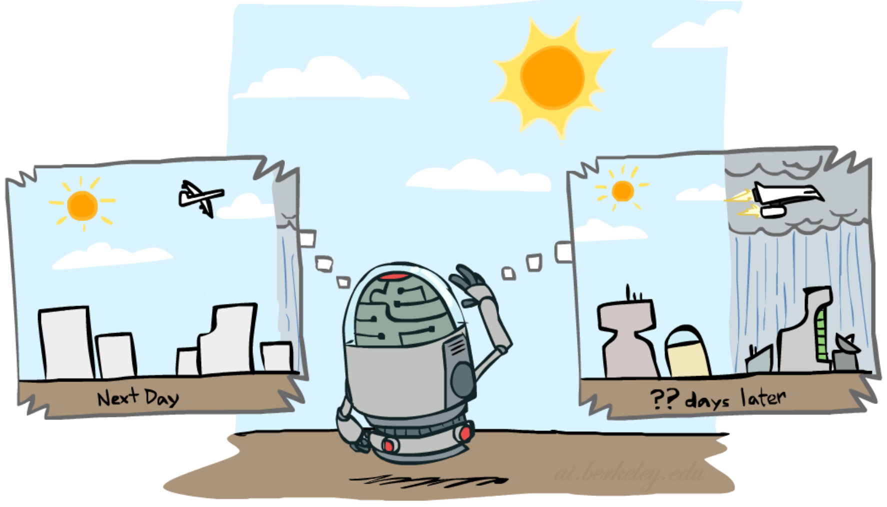
X_{t-1}	X_t	$P(X_t \| X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	sun	0.7

$$P(X_1) = \sum_{x_0} P(X_1, X_0 = x_0)$$

$$= 0.5 \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix} + 0.5 \begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix}$$

Weather Prediction (2)

EXAMPLE



Goal: What's the weather at time 2?

- Initial Distribution
  - $P(X_0 = \text{rain}) = 0.5$
  - $P(X_0 = \text{sun}) = 0.5$
- Transition Model $P(X_t | X_{t-1})$

X_{t-1}	X_t	$P(X_t \vert X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	sun	0.7

$$P(X_2) = \sum_{x_1} P(X_2, X_1 = x_1)$$

$$= 0.6 \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix} + 0.4 \begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix} = \begin{bmatrix} 0.66 \\ 0.34 \end{bmatrix}$$

Weather Prediction (3)

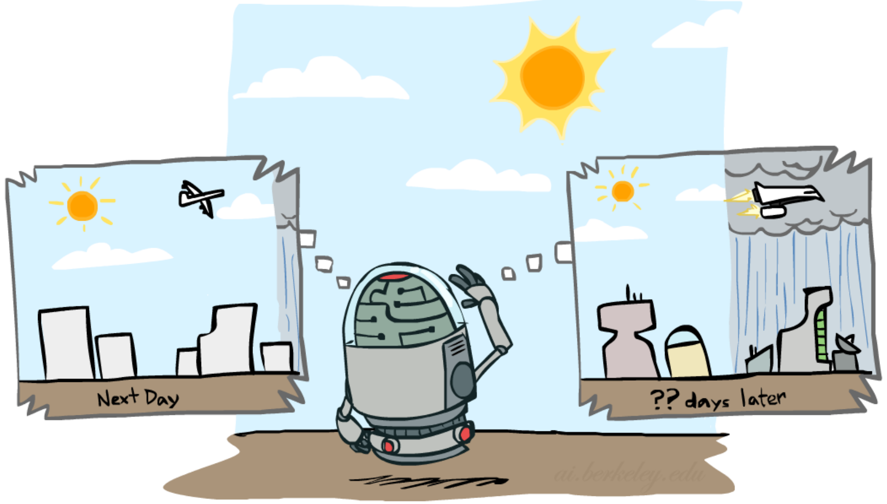
EXAMPLE

Goal: What's the weather at time 3?

- Initial Distribution
  - $P(X_0 = \text{rain}) = 0.5$
  - $P(X_0 = \text{sun}) = 0.5$

- Transition Model $P(X_t|X_{t-1})$

X_{t-1}	X_t	$P(X_t\|X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	sun	0.7



$$P(X_3) = \sum_{x_2} P(X_3, X_2 = x_2)$$

$$= 0.66 \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix} + 0.34 \begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix} = \begin{bmatrix} 0.696 \\ 0.304 \end{bmatrix}$$

Forward Algorithm

ALGORITHM

Iterate update step starting at $t = 0$:

$$P(X_t) = \sum_{x_{t-1}} P(X_t, X_{t-1} = x_{t-1}) = \sum_{x_{t-1}} \underbrace{\boxed{P(X_{t-1} = x_{t-1})}}_{\text{Prior Belief}} \underbrace{\boxed{P(X_t | X_{t-1} = x_{t-1})}}_{\text{Transition Model}}$$

N-Gram Models

EXAMPLE

Goal: Predict next word in sequence given training data

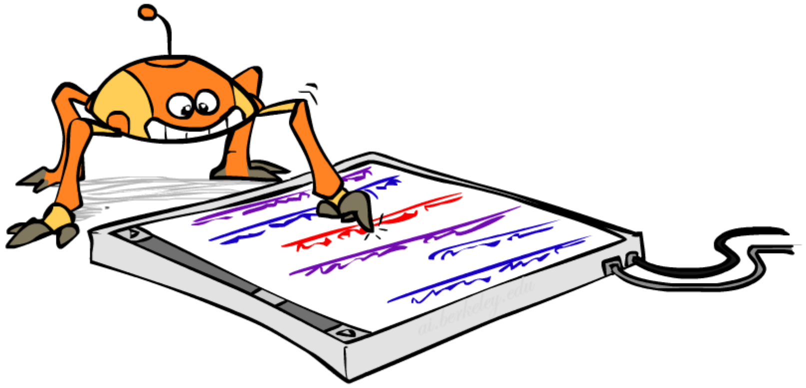
- States: word at position t in sentence
- Transition Model
  - **Unigram** (zero-order): $P(\text{Word}_t = i)$
    "logical are as are confusion a may right tries agent goal the was…"
  - **Bigram** (first-order): $P(\text{Word}_t = i | \text{Word}_{t-1} = j)$
    "systems are very similar computational approach would be represented…"
  - **Trigram** (second-order): $P(\text{Word}_t = i | \text{Word}_{t-1} = j, \text{Word}_{t-1} = k)$
    "planning and scheduling are integrated the success of naïve bayes model is…"

Web Browsing

EXAMPLE

Goal: Find stationary distribution over pages

- What fraction of time is spent in any given page?

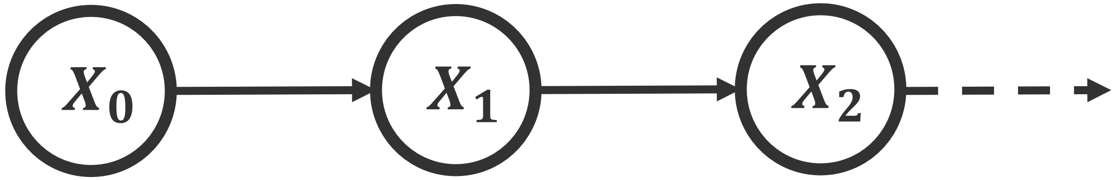


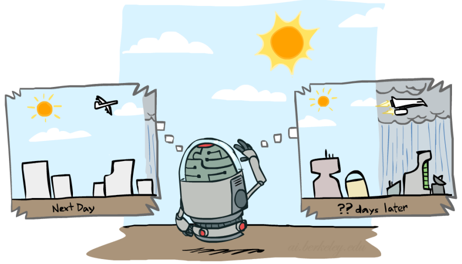
- States: URL visited at step t
- Transition Model
  - With probability p, choose an outgoing link at random
  - With probability $1 - p$, choose an arbitrary new page
- Application: Google Page Rank

Stationary Distributions

FOUNDATIONS

What is $lim_{t\to\infty} P(X_t)$?





X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	sun	0.7

The **stationary distribution** P_∞ satifies

$$P_\infty = P_{\infty+1} = T^\top \mathrm{P}_\infty$$

- Solve via matrix multiplication or system of equations

$$\begin{bmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{bmatrix} \begin{bmatrix} p \\ 1-p \end{bmatrix} = \begin{bmatrix} p \\ 1-p \end{bmatrix}$$

$$0.9p + 0.3(1-p) = p$$
$$0.1p + 0.7(1-p) = (1-p)$$

$$\Longrightarrow p = 0.75$$

QUESTIONS

Questions?

live and on sli.do #cse473

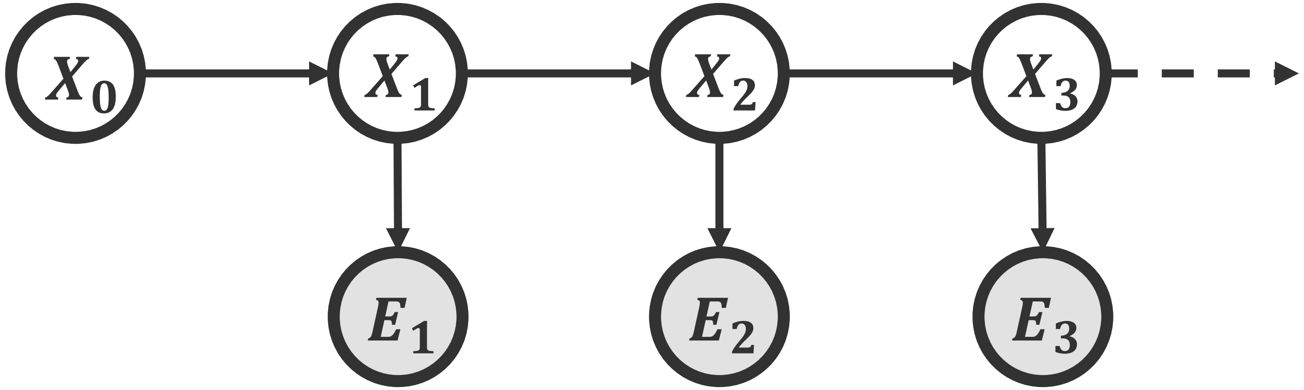
Partially Observable Models

Hidden Markov Models

FOUNDATIONS

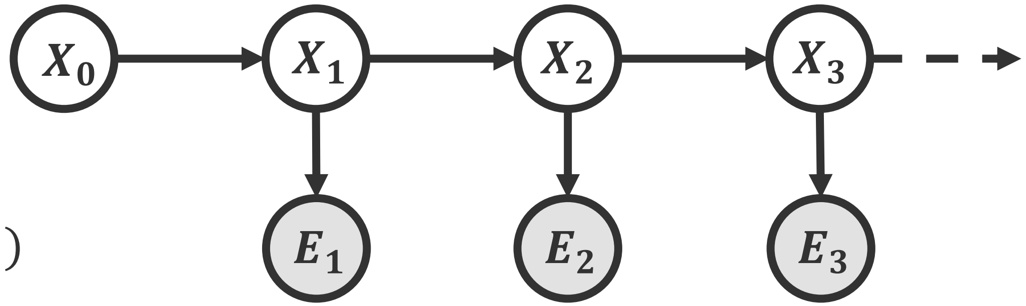
Hidden Markov Models (HMMs) describe problems where the (true) state is not observed directly

- Underlying Markov chain over states X_t
- Observe evidence E_t at each time step
- X_t is (usually) a discrete variable, E_t may be continuous and may consist of several variables



HMM Probability Model

DEFINITION



- Joint Distribution
  - For Markov Model: $P(X_0, \dots, X_T) = P(X_0)\prod_{t=1:T} P(X_t|X_{t-1})$
  - For Hidden Markov Model:

$$P(X_{0:T}, E_T) = P(X_0)\prod_{t=1:T} \underbrace{\boxed{P(X_t|X_{t-1})}}_{\text{Transition Model}} \underbrace{\boxed{P(E_t|X_t)}}_{\text{Likelihood}}$$

- Implications:
  - From Markov Models: future states are independent of past given present
  - Current evidence is independent of everything else given current state

Are evidence variables (absolutely) independent of each other?

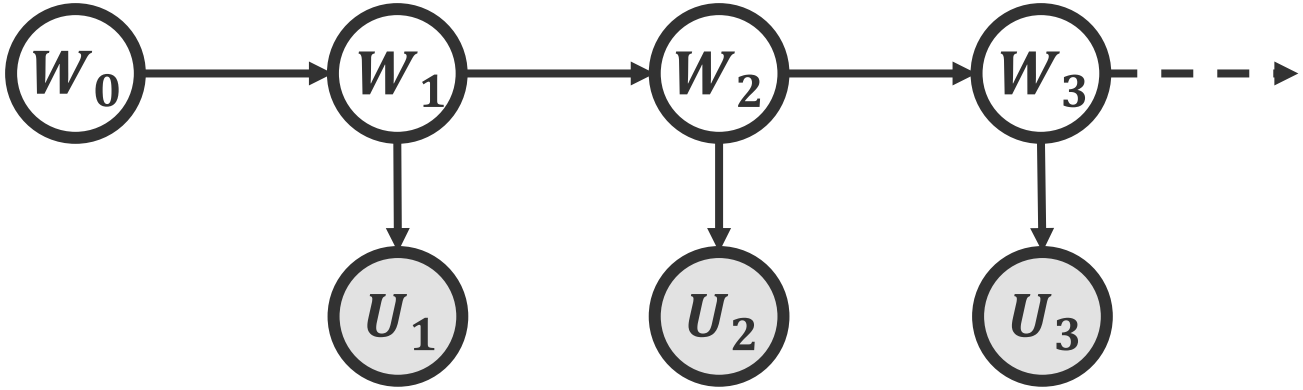
$$\boxed{X_{a:b} = X_a, X_{a+1}, \dots, X_b}$$

Weather Hidden Markov Model

EXAMPLE

- Hidden variable $W_t = \{\text{sun}, \text{rain}\}$

W_{t-1}	W_t	$P(W_t \mid W_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	sun	0.7



- Observations $U_t = \{\text{true}, \text{false}\}$

W_t	U_t	$P(U_t \mid W_t)$
sun	true	0.2
sun	false	0.8
rain	true	0.9
rain	false	0.1

Other HMM Examples

EXAMPLE

- Speech Recognition
  - Observations are acoustic signals (continuous)
  - States are specific positions in specific words
- Machine Translation
  - Observations are words
  - States are translation options
- Robot Tracking
  - Observations are range readings (continuous)
  - States are positions on a map (continuous)

Questions?

live and on sli.do #cse473

Inference Tasks

FOUNDATIONS

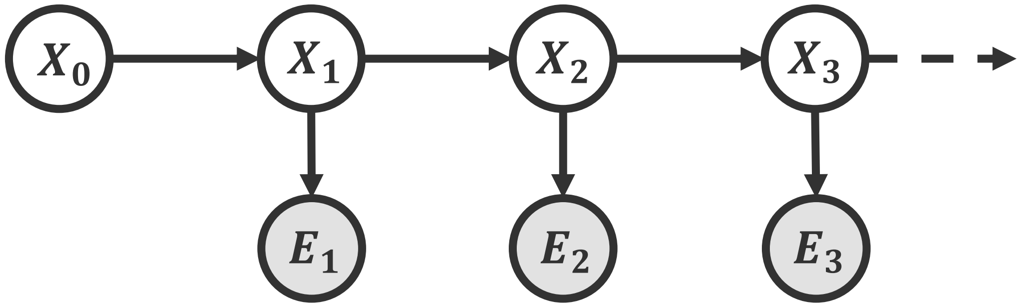
- **Filtering**: $P(X_t|e_{1:t})$
  - Belief state: input to decision process of a rational agent
- **Prediction**: $P(X_{t+k}|e_{1:t})$ for $k > 0$
  - Evaluation of possible action sequences (filtering without evidence)
- **Smoothing**: $P(X_k|e_{1:t})$ for $0 \leq k < t$
  - Better estimate of past states (essential for learning)
- **Most Likely Explanation**: $\text{argmax}_{x_{1:t}} P(x_{1:t}|e_{1:t})$
  - Speech recognition, decoding with a noisy channel

Filtering

DEFINITION

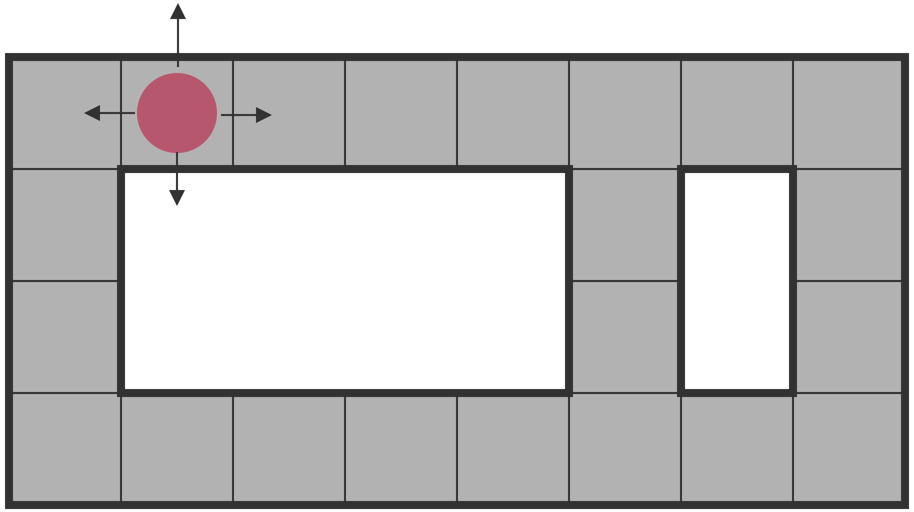
Filtering (or monitoring or state estimation) is the task of maintaining the distribution $f_{1:t} = P(X_t | e_{1:t})$

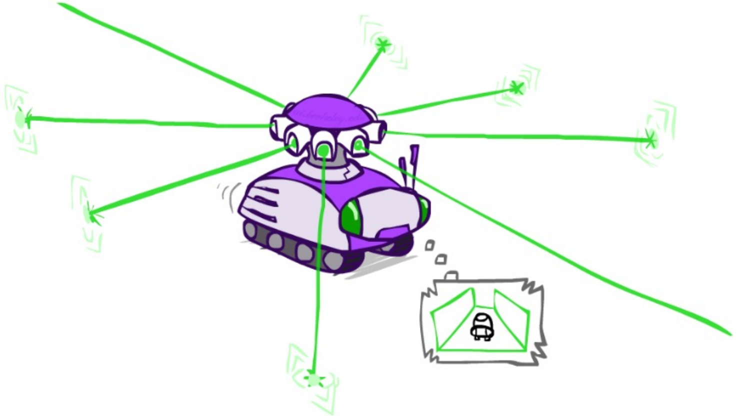
- Start with f_0 in an initial setting, usually uniform
- Fundamental task in many engineering and science domains



Robot Localization

EXAMPLE





Probability

0 1

Sensor Model
- Four bits for isWall in each direction
- Never more than 1 mistake

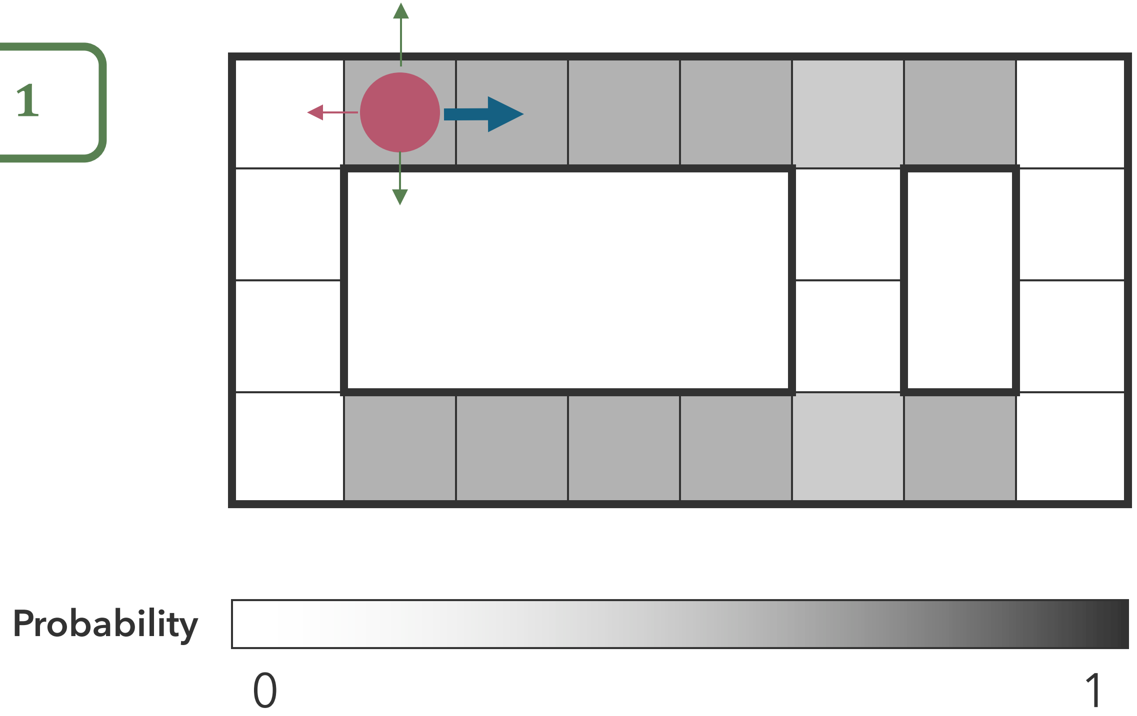
Transition Model
- Action may fail with small probability

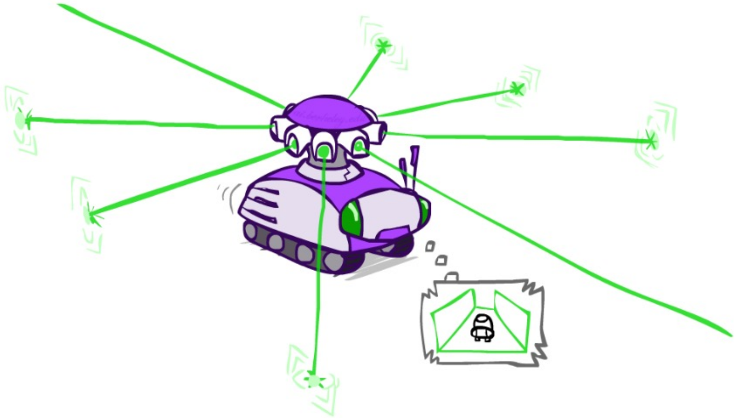
Example from Michael Pfeiffer

Robot Localization (1)

EXAMPLE

$t = 1$





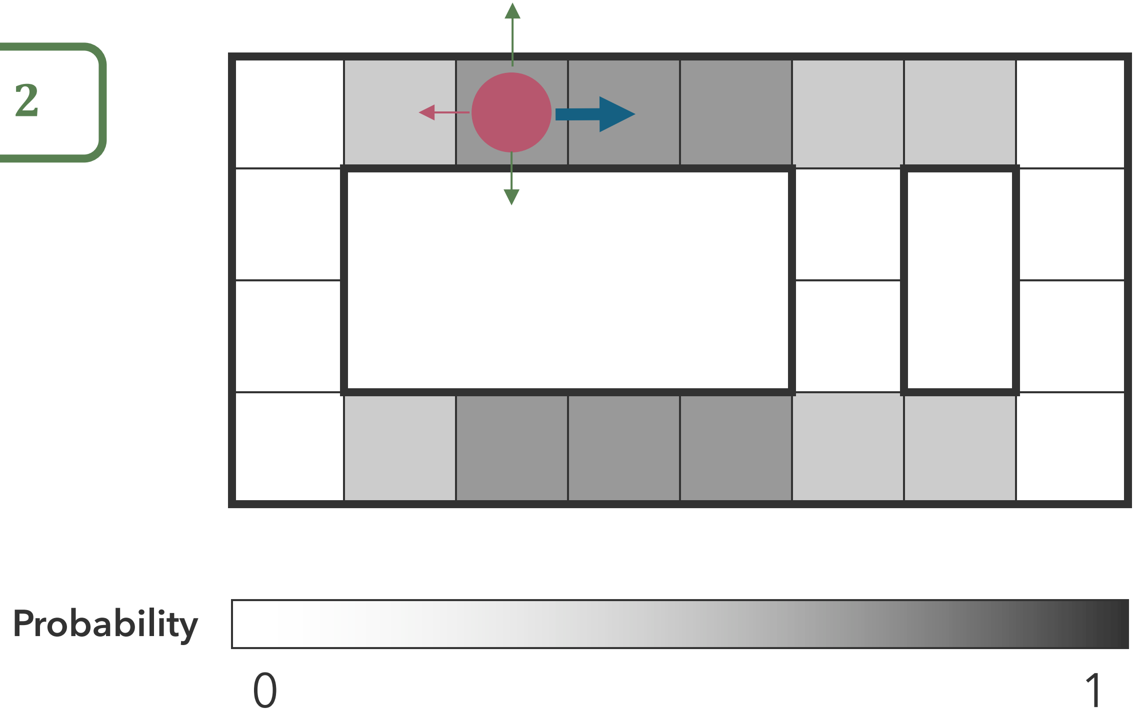
Lighter grey: *possible* to get reading, but less likely (requires 1 mistake)

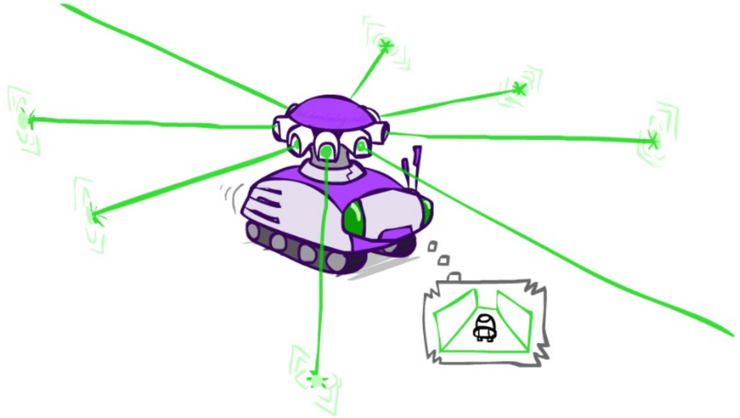
Example from Michael Pfeiffer

Robot Localization (2)

EXAMPLE

$t = 2$



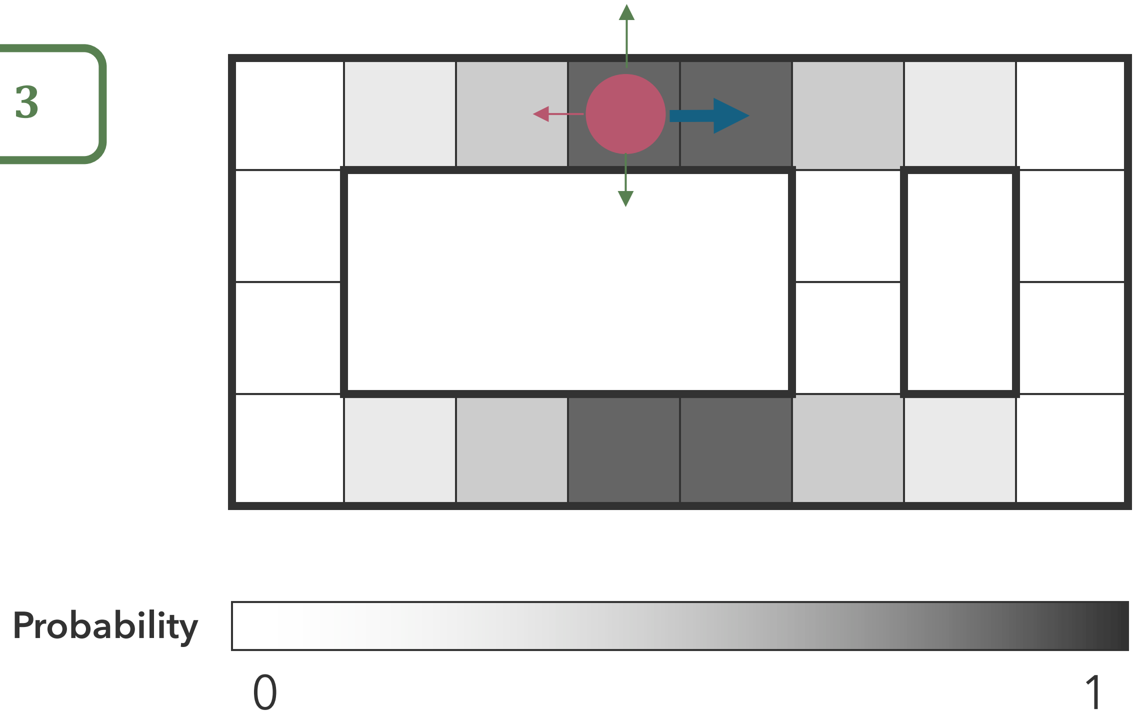


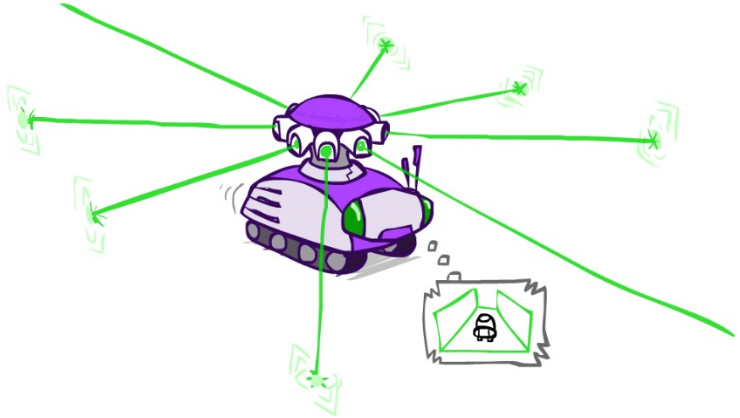
Example from Michael Pfeiffer

Robot Localization (3)

EXAMPLE

$t = 3$



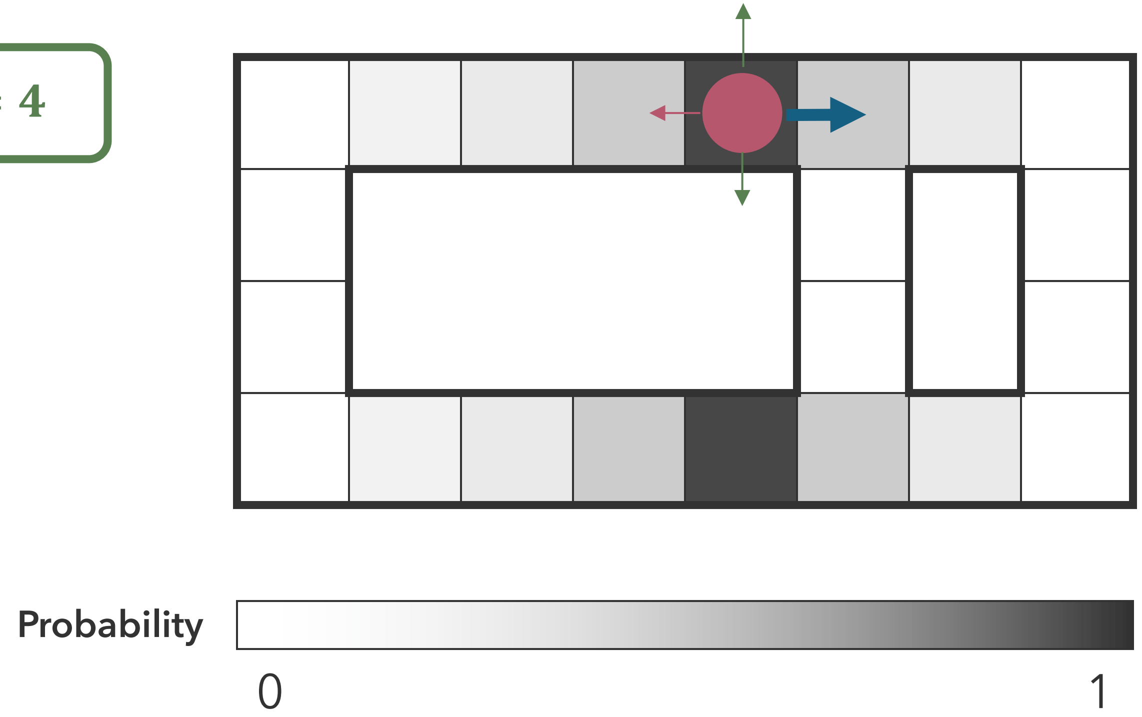


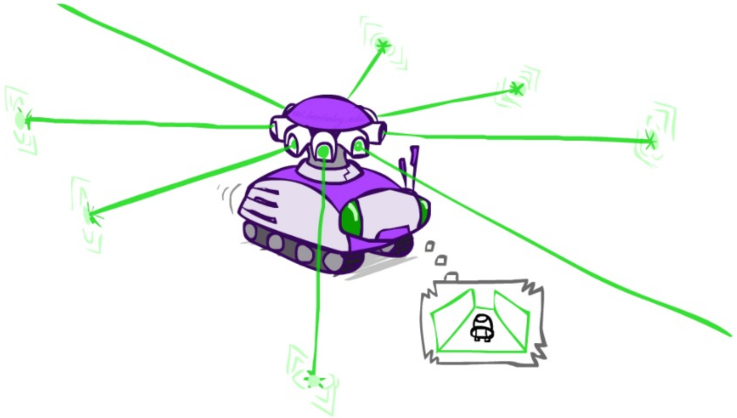
Example from Michael Pfeiffer

Robot Localization (4)

EXAMPLE

$t = 4$



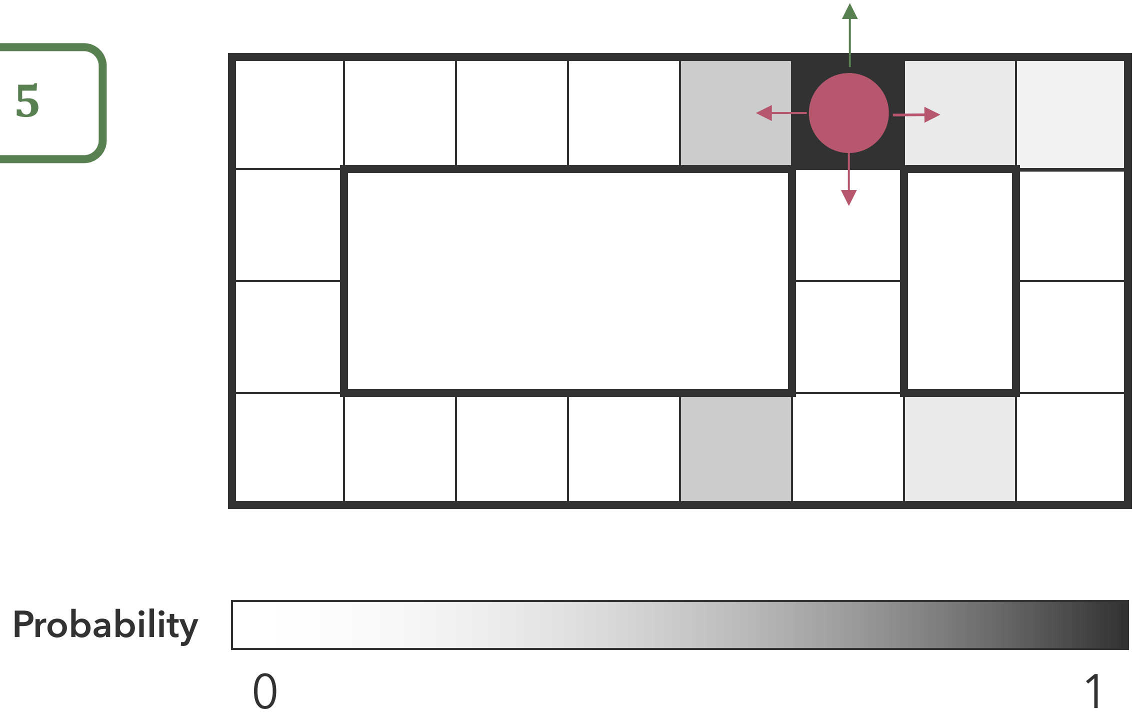


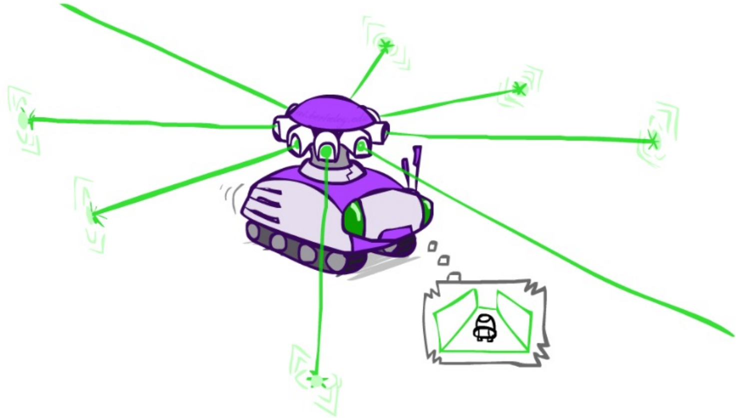
Example from Michael Pfeiffer

Robot Localization (5)

EXAMPLE

$t = 5$





Example from Michael Pfeiffer

Filtering Algorithm

ALGORITHM

Goal: Device a **recursive filtering** algorithm of the form

$$P(X_{t+1}|e_{1:t+1}) = g(e_{t+1}, P(X_t|e_{1:t}))$$

Approach:

$$\begin{aligned} P(X_{t+1}|e_{1:t+1}) &= P(X_{t+1}|e_{1:t}, e_{t+1}) && \text{apply Bayes rule} \\ &= \alpha \cdot P(e_{t+1}|X_{t+1}, e_{1:t}) \cdot P(X_{t+1}|e_{1:t}) && \text{apply conditional independence} \\ &= \alpha \cdot \underbrace{P(e_{t+1}|X_{t+1})}_{\text{Likelihood}} \cdot \underbrace{P(X_{t+1}|e_{1:t})}_{\text{Prior Prediction}} \end{aligned}$$

Filtering Algorithm (2)

ALGORITHM

Goal: Device a **recursive filtering** algorithm of the form

$$P(X_{t+1}|e_{1:t+1}) = g(e_{t+1}, P(X_t|e_{1:t}))$$

Approach:

$$\begin{aligned}
P(X_{t+1}|e_{1:t+1}) &= P(X_{t+1}|e_{1:t}, e_{t+1}) && \text{apply Bayes rule} \\
&= \alpha \cdot P(e_{t+1}|X_{t+1}, e_{1:t}) \cdot P(X_{t+1}|e_{1:t}) && \text{apply conditional independence} \\
&= \alpha \cdot P(e_{t+1}|X_{t+1}) \cdot \boxed{P(X_{t+1}|e_{1:t})} && \text{condition on } X_t \\
&= \alpha \cdot P(e_{t+1}|X_{t+1}) \cdot \textstyle\sum_{x_t} P(x_t|e_{1:t}) \cdot P(X_{t+1}\boxed{|x_t, e_{1:t})} && \text{apply conditional independence} \\
&= \alpha \cdot \underbrace{P(e_{t+1}|X_{t+1})}_{\text{Likelihood}} \cdot \textstyle\sum_{x_t} \underbrace{P(x_t|e_{1:t})}_{\text{Recursive}} \underbrace{P(X_{t+1}|x_t)}_{\text{Transition Prob.}}
\end{aligned}$$

Filtering Properties

DEFINITION

$$f_{1:t+1} = P(X_{t+1}|e_{1:t+1}) = \alpha \underbrace{P(e_{t+1}|X_{t+1})}_{\text{Update}} \underbrace{\sum_{x_t} P(x_t|e_{1:t}) \cdot P(X_{t+1}|x_t)}_{\text{Predict}}$$

- $f_{1:t+1} = \text{FORWARD}(f_{1:t}, e_{t+1})$
- Cost per time step: $O(|X|^2)$ for $|X|$ number of states
  - Time and space costs are independent of t (i.e. constant)
  - But $O(|X|^2)$ is infeasible for models with many state variables
  - This motivates *approximate* filtering algorithms

Questions?

live and on sli.do #cse473

Matrix Formulation

DEFINITION

Equivalent formulation to recursive equation given

- Transition matrix T
- Observation matrix O_t, with state likelihoods for E_t along diagonal
- For example, $T = \begin{bmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{bmatrix}$, $U_1 = \text{true}$, $O_1 = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.9 \end{bmatrix}$

- **Filtering**
  - $f_{1:t+1} = \alpha \cdot O_{t+1} T^{\top} f_{1:t}$

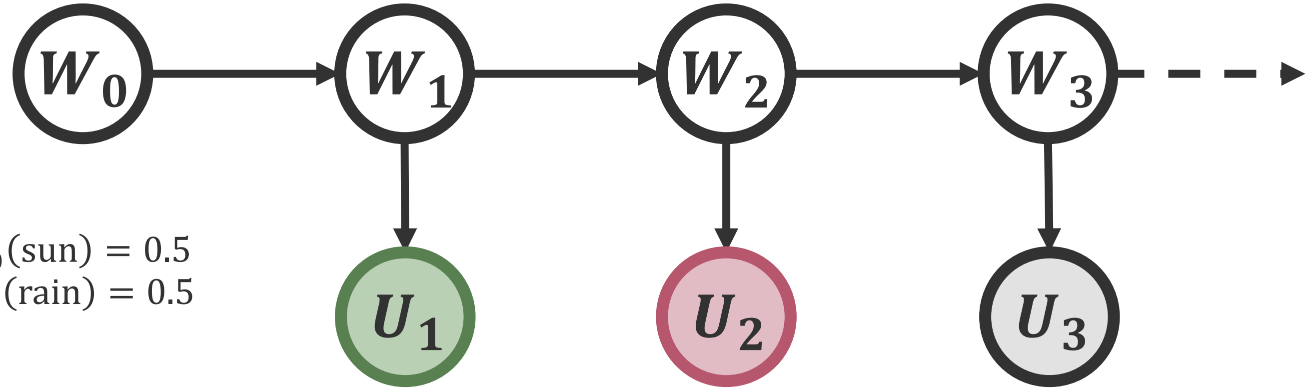
$\boldsymbol{W_{t-1}}$	$\boldsymbol{W_t}$	$\boldsymbol{P(W_t \mid W_{t-1})}$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	sun	0.7

$\boldsymbol{W_t}$	$\boldsymbol{U_t}$	$\boldsymbol{P(U_t \mid W_t)}$
sun	true	0.2
sun	false	0.8
rain	true	0.9
rain	false	0.1

Weather Filtering

$$P(X_{t+1}|e_{1:t+1}) = \alpha \cdot P(e_{t+1}|X_{t+1}) \sum_{x_t} P(x_t|e_{1:t}) \cdot P(X_{t+1}|x_t)$$

EXAMPLE



$f_0(\text{sun}) = 0.5$
$f_0(\text{rain}) = 0.5$

$U_1 = \text{true}$

$U_2 = \text{false}$

W_{t-1}	W_t	$P(W_t\|W_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	sun	0.7

Predict

$P(s_1|W_0) = 0.5 \cdot 0.9 + 0.5 \cdot 0.3 = 0.6$
$P(r_1|W_0) = 0.5 \cdot 0.1 + 0.5 \cdot 0.7 = 0.4$

Update

$f_1(\text{sun}) = \alpha \cdot 0.2 \cdot 0.6 = 0.25$
$f_1(\text{rain}) = \alpha \cdot 0.9 \cdot 0.4 = 0.75$
$\alpha = 0.2 \cdot 0.6 + 0.9 \cdot 0.4$

Predict

$P(s_2|W_1) = 0.25 \cdot 0.9 + 0.75 \cdot 0.3 = 0.45$
$P(r_2|W_1) = 0.25 \cdot 0.1 + 0.75 \cdot 0.7 = 0.55$

Update

$f_2(\text{sun}) = \alpha \cdot 0.8 \cdot 0.45 = 0.867$
$f_2(\text{rain}) = \alpha \cdot 0.1 \cdot 0.55 = 0.133$
$\alpha = 0.8 \cdot 0.45 + 0.1 \cdot 0.55$

W_t	U_t	$P(U_t\|W_t)$
sun	true	0.2
sun	false	0.8
rain	true	0.9
rain	false	0.1

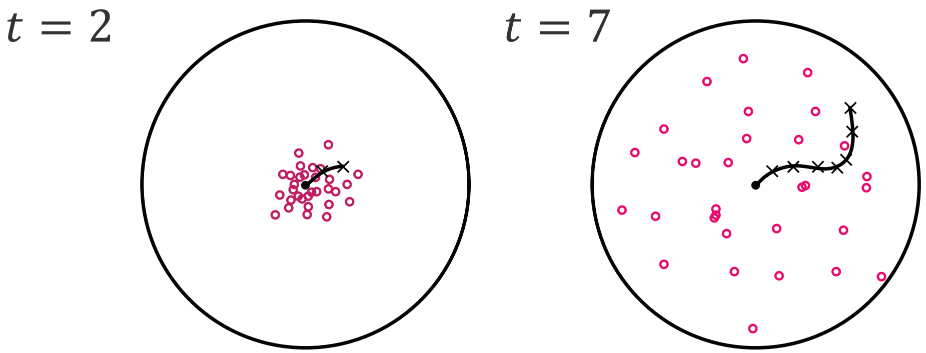
Approximate Inference

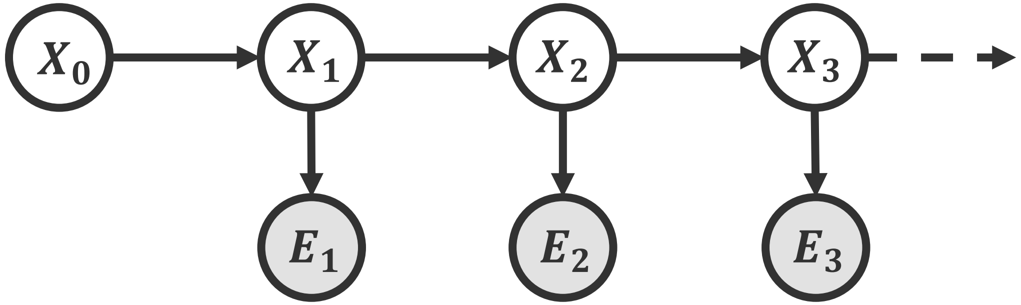
CONTEXT

When $|X|$ is more than $\sim 10^6$, exact inference becomes infeasible

We need a new approach!

- Likelihood Weighting?
  - Fails completely (samples grow exponentially with T)





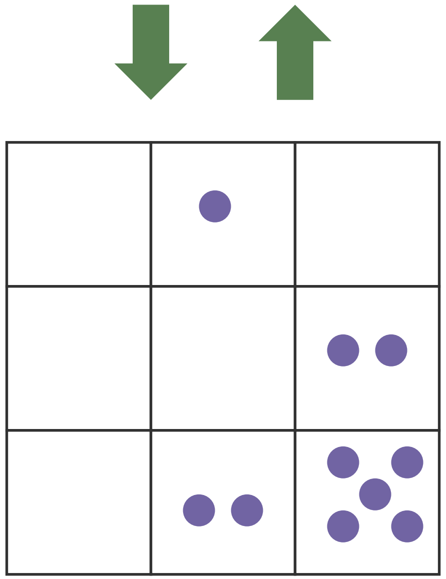
Solution: Eliminate 'bad' samples, make more 'good' samples

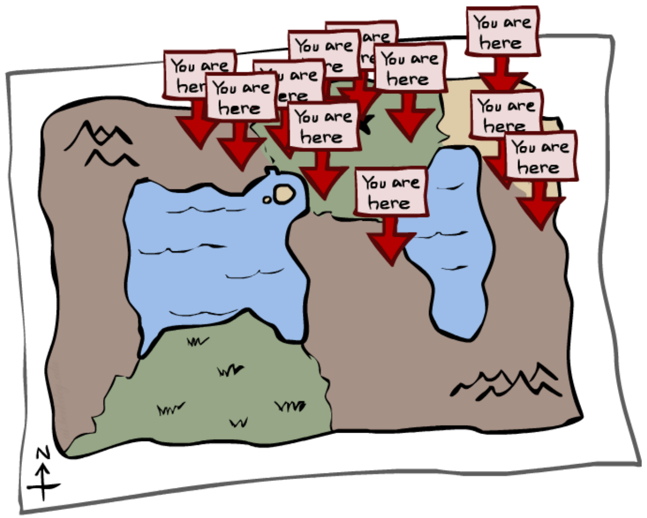
Particle Filtering

FOUNDATIONS

- Represent belief state by a set of samples
  - Samples are also called **particles**
  - Time per step is linear in the number of samples…
    - …but number needed may be large
  - Store list of particles, not states!

0	0.1	0
0	0	0.2
0	0.2	0.5



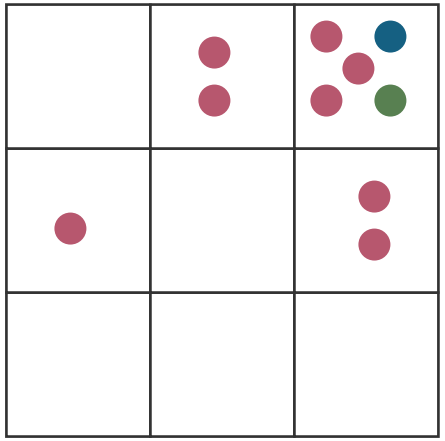


Particles

DEFINITION

- Representation of $P(X)$ is now a list of $N \ll |X|$
- $P(X)$ is approximated by the number of particles with the value x
  - Many x may have $P(x) = 0$
  - More particles, more accuracy! (converges to true distribution)

Track samples of states rather than an explicit distribution!



Particles
(3,3)
(2,3)
(3,3)
(3,2)
(3,3)
(3,2)
(1,2)
(3,3)
(3,3)
(2,3)

QUESTIONS

Questions?

live and on sli.do #cse473

Particle Filtering Prediction

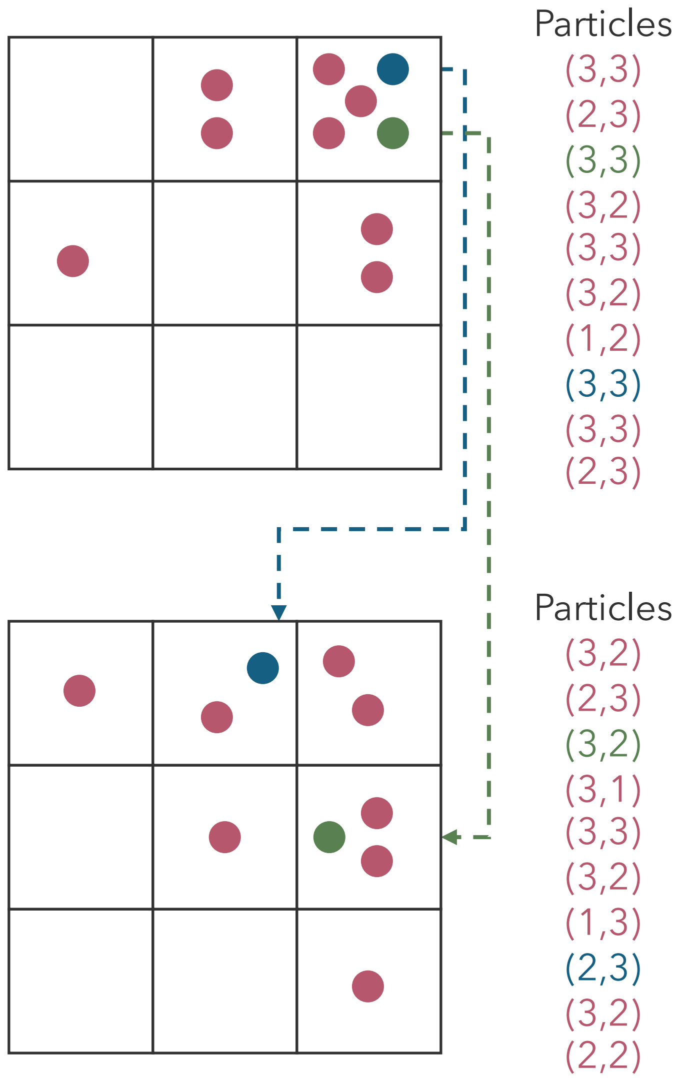
DEFINITION

- Particle j in state $x_t^{(j)}$ samples a new state directly from the transition model:

$$x_{t+1}^{(j)} \sim P(X_{t+1} | x_t^{(j)})$$

  - Most samples move clockwise, but some move in another direction or stay in place
  - For example:

$$x_{t+1}^{(j)} \sim P\left(X_{t+1} \middle| x_t^{(\text{green})}\right) = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix}$$



Particle Filtering Update

DEFINITION

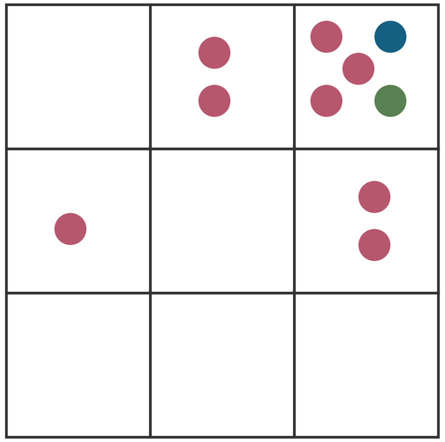
- After observing e_{t+1}:
  - As in likelihood weighting, weight each sample based on the evidence

$$w^{(j)} = P(e_{t+1} | x_{t+1}^{(j)})$$

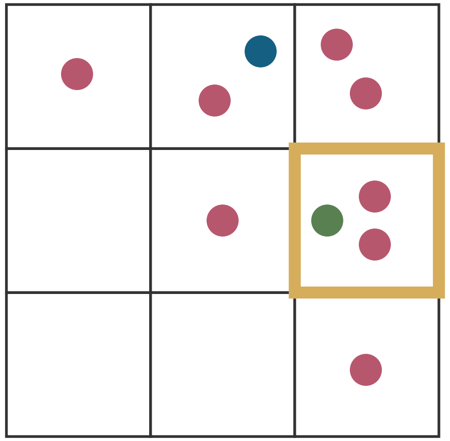
  - Normalize by the weights: particles that fit the data better get higher weights, others get lower weights
  - For example, with $e_{t+1} = (3,2)$

$$w^{(\text{green})} = P\big((3,2)|(3,2)\big) = 0.9$$

$$w^{(\text{blue})} = P\big((3,2)|(2,3)\big) = 0.2$$



Particles
(3,3)
(2,3)
(3,3)
(3,2)
(3,3)
(3,2)
(1,2)
(3,3)
(3,3)
(2,3)



Particles
(3,2) $w = 0.9$
(2,3) $w = 0.2$
(3,2) $w = 0.9$
(3,1) $w = 0.4$
(3,3) $w = 0.4$
(3,2) $w = 0.9$
(1,3) $w = 0.1$
(2,3) $w = 0.2$
(3,2) $w = 0.9$
(2,2) $w = 0.4$

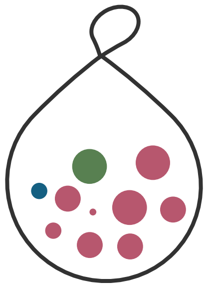
Particle Filtering Resampling

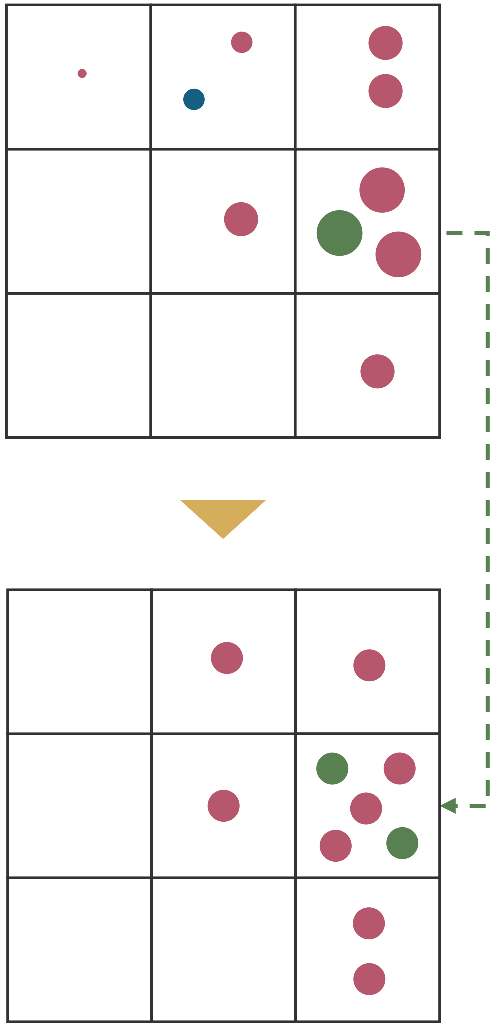
DEFINITION

- Rather than tracking weighted samples, ***resample***
  - Choose from weighted sample distribution N times (draw with replacement)

$$x_{t+1}^{(j)} \sim \alpha \cdot \mathrm{Num}(X_{t+1} | e_{1:t}) \cdot W(X_{t+1} | e_{1:t})$$

$1 \cdot 0.1 \cdot \alpha$	$2 \cdot 0.2 \cdot \alpha$	$2 \cdot 0.4 \cdot \alpha$
0	$1 \cdot 0.4 \cdot \alpha$	$3 \cdot 0.9 \cdot \alpha$
0	0	$1 \cdot 0.4 \cdot \alpha$





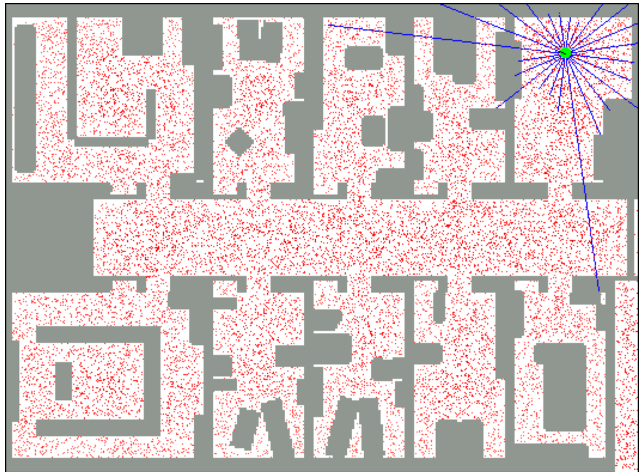
Particles
(3,2) $w = 0.9$
(2,3) $w = 0.2$
(3,2) $w = 0.9$
(3,1) $w = 0.4$
(3,3) $w = 0.4$
(3,2) $w = 0.9$
(1,3) $w = 0.1$
(2,3) $w = 0.2$
(3,2) $w = 0.9$
(2,2) $w = 0.4$

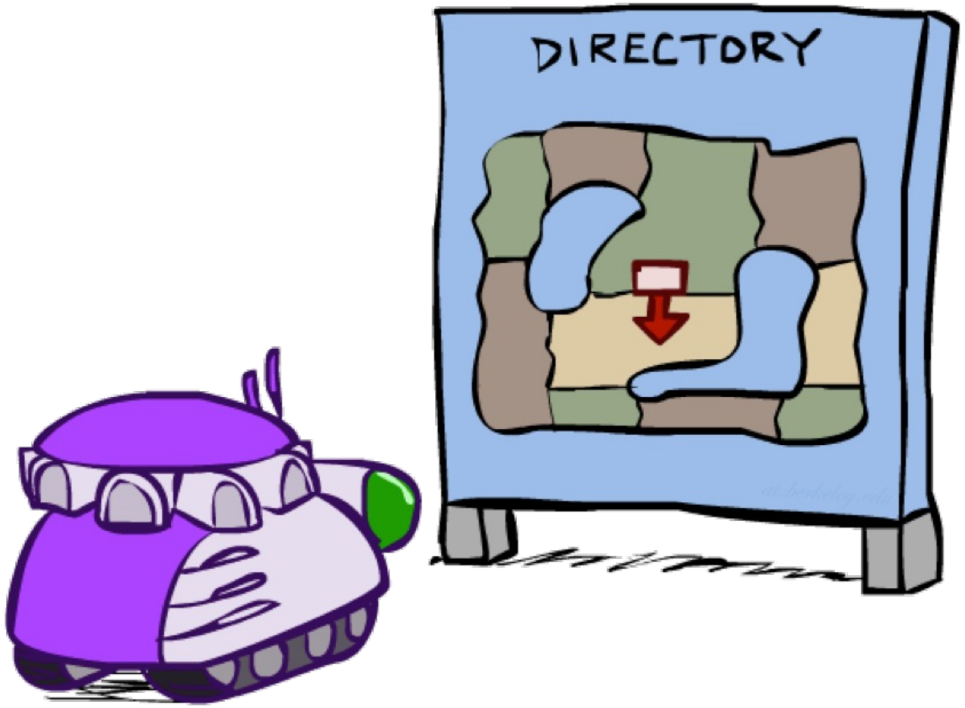
Particles
(3,1)
(2,3)
(3,2)
(3,2)
(3,3)
(3,2)
(3,1)
(2,2)
(3,2)
(3,2)

Localization

EXAMPLE

- In a robot localization problem:
  - We know the map, but not the robot's position
  - Observations may be vectors of range finder readings
  - State space and readings are typically continuous
    - Large problem size motivates particle filtering





Particle Filter Localization

EXAMPLE

QUESTIONS

Questions?

live and on [sli.do](sli.do) #cse473

that's it for today!

SUMMARY

- Markov chains model sequential observations
- Hidden Markov Models use evidence to infer unobservable state
- Filtering involves updating belief given evidence

UPCOMING

- Machine Learning

REMINDERS

- Complete **Practice Problem 14**
- Do **Project 3** (due 7.30)
- Do **Homework 3** (due 8.6)