Probability and Time: Markov Models

Computer Science cpsc322, Lecture 31

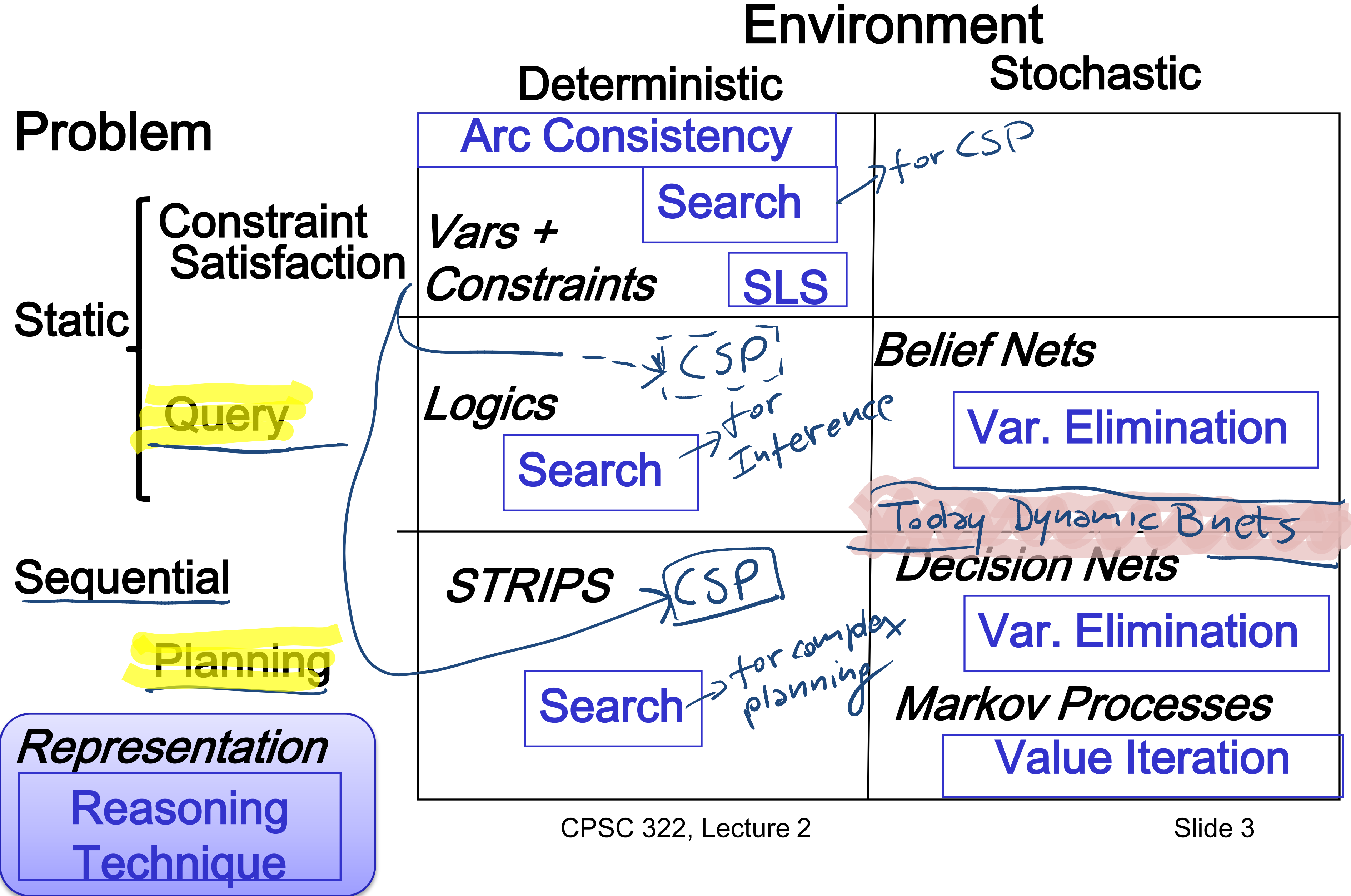
(Textbook Chpt 6.5)

March, 25, 2009

Lecture Overview

- **Recap**
- Temporal Probabilistic Models
- Start Markov Models
  - Markov Chain
  - Markov Chains in Natural Language Processing

Big Picture: R&R systems



Environment

Deterministic

Stochastic

Problem

Static

Constraint Satisfaction

Vars + Constraints

Arc Consistency

Search

SLS

→ for CSP

Query

Logics

Search

→ for Inference

CSP

Belief Nets

Var. Elimination

Today Dynamic Bnets

Sequential

Planning

STRIPS

CSP

Search

→ for complex planning

Decision Nets

Var. Elimination

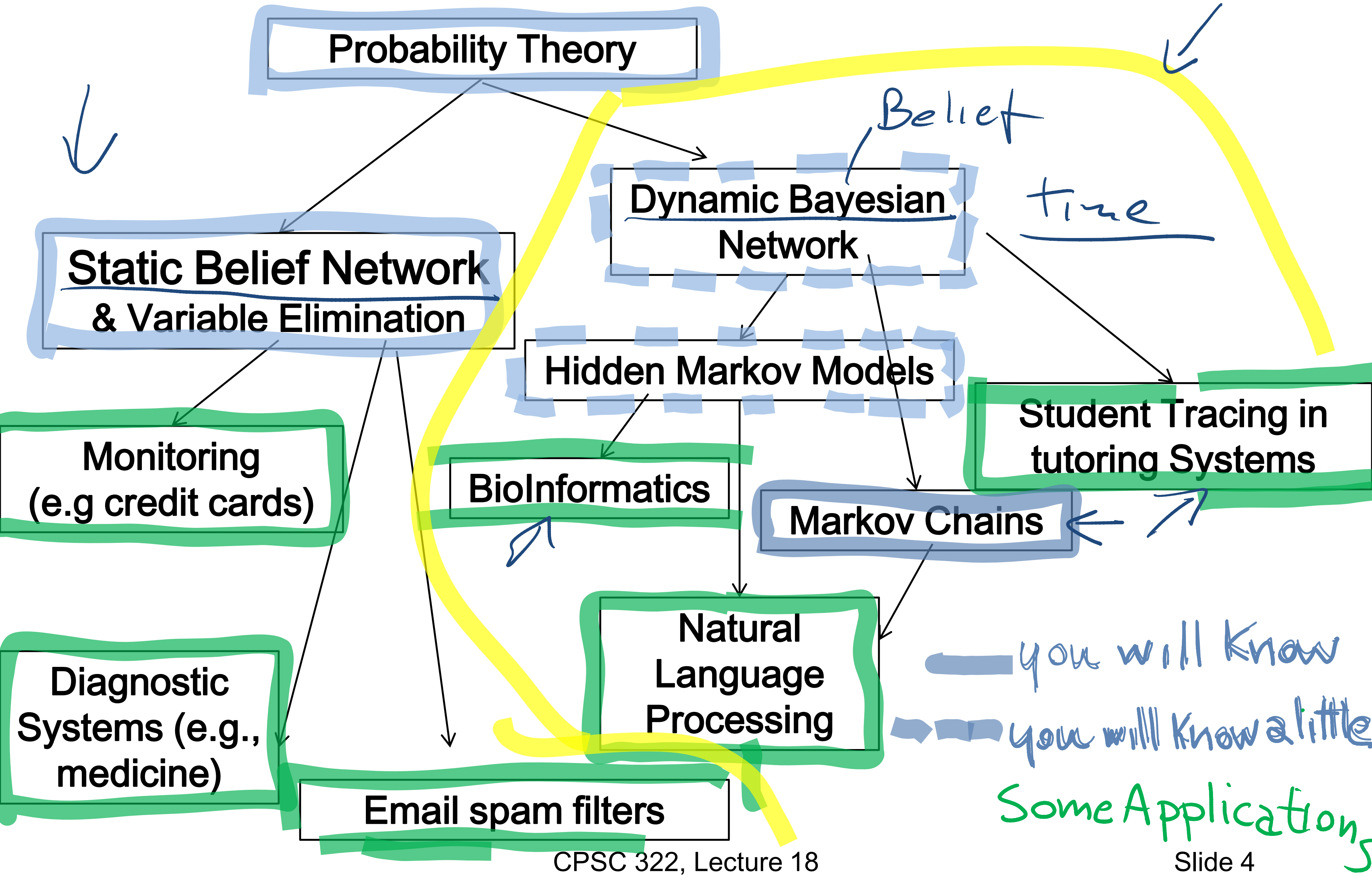
Markov Processes

Value Iteration

Representation

Reasoning Technique

Answering Query under Uncertainty



Probability Theory

Static Belief Network & Variable Elimination

Dynamic Bayesian Network

Belief

time

Hidden Markov Models

Monitoring (e.g credit cards)

BioInformatics

Student Tracing in tutoring Systems

Markov Chains

Diagnostic Systems (e.g., medicine)

Natural Language Processing

Email spam filters

you will Know

you will Know a little

Some Applications

Lecture Overview

- Recap
- **Temporal Probabilistic Models**
- Start Markov Models
  - Markov Chain
  - Markov Chains in Natural Language Processing

Modelling static Environments

So far we have used Bnets to perform inference in static environments

- For instance, the reasoning system keeps collecting evidence to diagnose the cause of a fault in a system (e.g., a car).

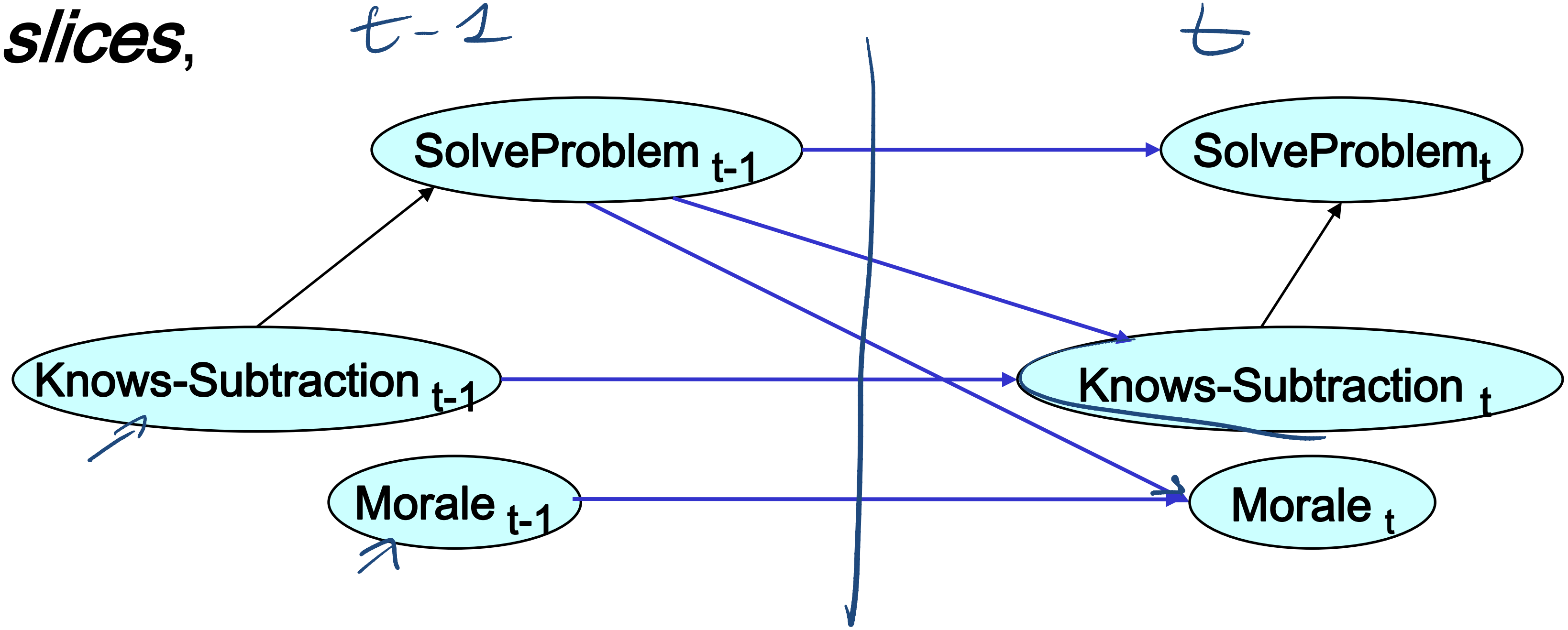
- The environment (values of the evidence, the true cause) does not change as I gather new evidence

- What does change?

The reasoning system's beliefs over possible causes

Modeling Evolving Environments

- Often we need to make inferences about evolving environments.
- Represent the state of the world at each specific point in time via a series of snapshots, or ***time slices***,



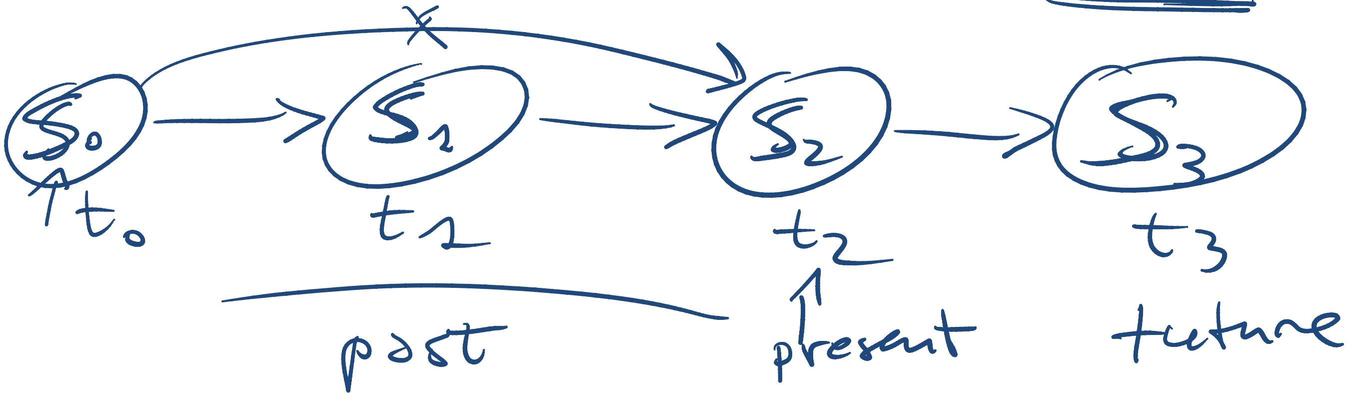
Tutoring system tracing student *knowledge* and *morale*

Lecture Overview

- Recap
- Temporal Probabilistic Models
- Start Markov Models
  - **Markov Chain**
  - Markov Chains in Natural Language Processing

Simplest Possible DBN

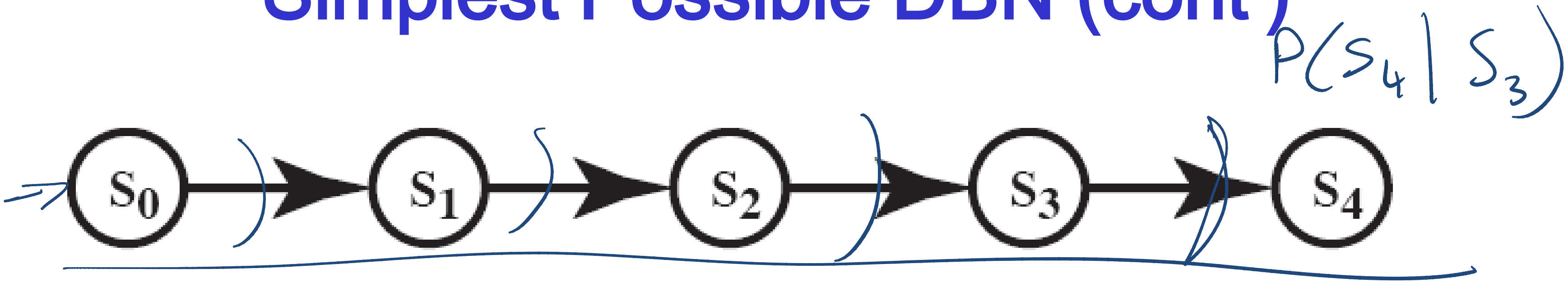
- **One random variable** for each time slice: let's assume S_t represents the state at time *t*. with domain $\{s_1 \ldots s_n\}$



- **Each random variable depends only on the previous one**
- Thus $P(S_{t+1} | S_0 \cdots S_t) = P(\underline{S_{t+1}} | \underline{S_t})$

- Intuitively S_t conveys all of the information about the history that can affect the future states.
- "The future is independent of the past given the present."

Simplest Possible DBN (cont')



- How many CPTs do we need to specify?

4 P(S1 | S0) P(S2 | S1) etc.

- *Stationary process assumption:* the mechanism that regulates how state variables change overtime is **stationary**, that is it can be described by a single transition model

- $P(S_t|S_{t-1})$ is the same for all t

Stationary Markov Chain (SMC)

A stationary Markov Chain : for all t >0

- $P(S_{t+1}| S_0,\ldots,S_t) = P(S_{t+1}|S_t)$ and Markov assumption
- $P(S_{t+1}|S_t)$ is the same stationary

We only need to specify $P(S_0)$ and $P(S_{t+1} | S_t)$

- Simple Model, easy to specify ←
- Often the natural model ←
- The network can extend indefinitely ←
- **Variations of SMC are at the core of most Natural Language Processing (NLP) applications!**

also used in the PageRank algo (used by Google to rank web pages)

Stationary Markov-Chain: Example

Domain of variable S_i is {t , i, p, a, h, e}

six possible values

We only need to specify…

$P(S_0)$

Probability of initial state

t	.6
q	.4
p	0
a	0
h	0
e	0

Stochastic Transition Matrix

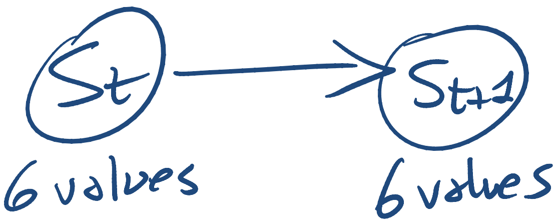
$P(S_{t+1}|S_t)$

S_{t+1}

S_t	t	q	p	a	h	e
t	0	.3	0	.3	.4	0
q	.4	0	.6	0	0	0
p	0	0	1	0	0	0
a	0	0	.4	.6	0	0
h	0	0	0	0	0	1
e	1	0	0	0	0	0

← $P(S_{t+1}|S_t = q)$

← $P(S_{t+1}|S_t = p)$



Markov-Chain: Inference

Probability of a sequence of states $S_0 \dots S_T$

$$\underline{P(S_0,\dots,S_T)} = P(S_0)\, P(S_1|S_0)\, P(S_2|S_1) \dots\dots$$

$P(u,e,e) \rightarrow$

$P(S_0)$

t	.6
q	.4
p	0
a	0
h	0
e	0

$P(S_{t+1}|S_t)$

	t	q	p	a	h	e
t	0	.3	0	.3	.4	0
q	.4	0	.6	0	0	0
p	0	0	1	0	0	0
a	0	0	.4	.6	0	0
h	0	0	0	0	0	1
e	1	0	0	0	0	0

Example:

$$\underline{P(t, q, p)} =$$

$$P(t) * P(q|t) * P(p|q)$$

$$.6 \times .3 \times .6 = .108$$

Lecture Overview

- Recap
- Temporal Probabilistic Models
- **Markov Models**
  - Markov Chain
  - Markov Chains in Natural Language Processing

Key problems in NLP

Noun Verb

"Book me a room near UBC"

w_1 w_2 w_3 w_4 w_5 w_6

$$\boxed{P(w_1,..,w_n)?}$$

Assign a probability to a sentence (a sequence of words)

- Part-of-speech tagging
- Word-sense disambiguation,
- Probabilistic Parsing

Summarization, Machine Translation……..

Predict the next word

$$P(w_n | w_1 \dots \dots w_{n-1}) = $$
$$= P(w_1 \dots w_n) / P(w_1 .. w_{n-1})$$

- Speech recognition
- Hand-writing recognition
- Augmentative communication for the disabled

$$\boxed{P(w_1,..,w_n)?}$$

Impossible to estimate ☹

$\boxed{P(w_1,..,w_n)?}$

Impossible to estimate!

Assuming 10^5 words and average sentence contains 10 words ....

$(10^5)^{10} = 10^{50}$

would contain ↗ probabilities

Google language repository (22 Sept. 2006) → collected from the whole web

contained "only": 95,119,665,584 sentences

$\sim 10^{11}$

Most sentences will not appear or appear only once ☹

What can we do?

Make a strong simplifying assumption!

Sentences are generated by a Markov Chain

$$\underline{P(w_1,..,w_n)} = \overbrace{P(\underline{w_1}|<S>)}^{w_1 \text{ at the beginning of a sentence}}\underline{\prod_{k=2}^{n} P(w_k \mid w_{k-1})}$$

$$= P(w_1|<S>)\,P(w_2|w_1)\,P(w_3|w_2)\,..\,P(w_K|w_{K-1})$$

P(The big red dog barks)=

**P(The|<S>) *** P(big|the) * P(red|big) * ...

* P(dog|red) * P(barks|dog)

These probs can be assessed in practice!

Estimates for Bigrams

$P(w_i \mid w_{i-1})$

Silly language repositories with **only two sentences**:

"<S> *The big red dog barks against the big pink dog*"

"<S> *The big pink dog is much smaller*"

Count How many times in your documents you have "big red" and "big"

$$P(red \mid big) = \frac{P(big, red)}{P(big)} = \frac{\overset{\text{ount}}{\dfrac{C(big, red)}{\cancel{N_{pairs}}}}}{\dfrac{C(big)}{\cancel{N_{words}}}} = \frac{C(big, red)}{C(big)} = \frac{1}{3}$$

10^5

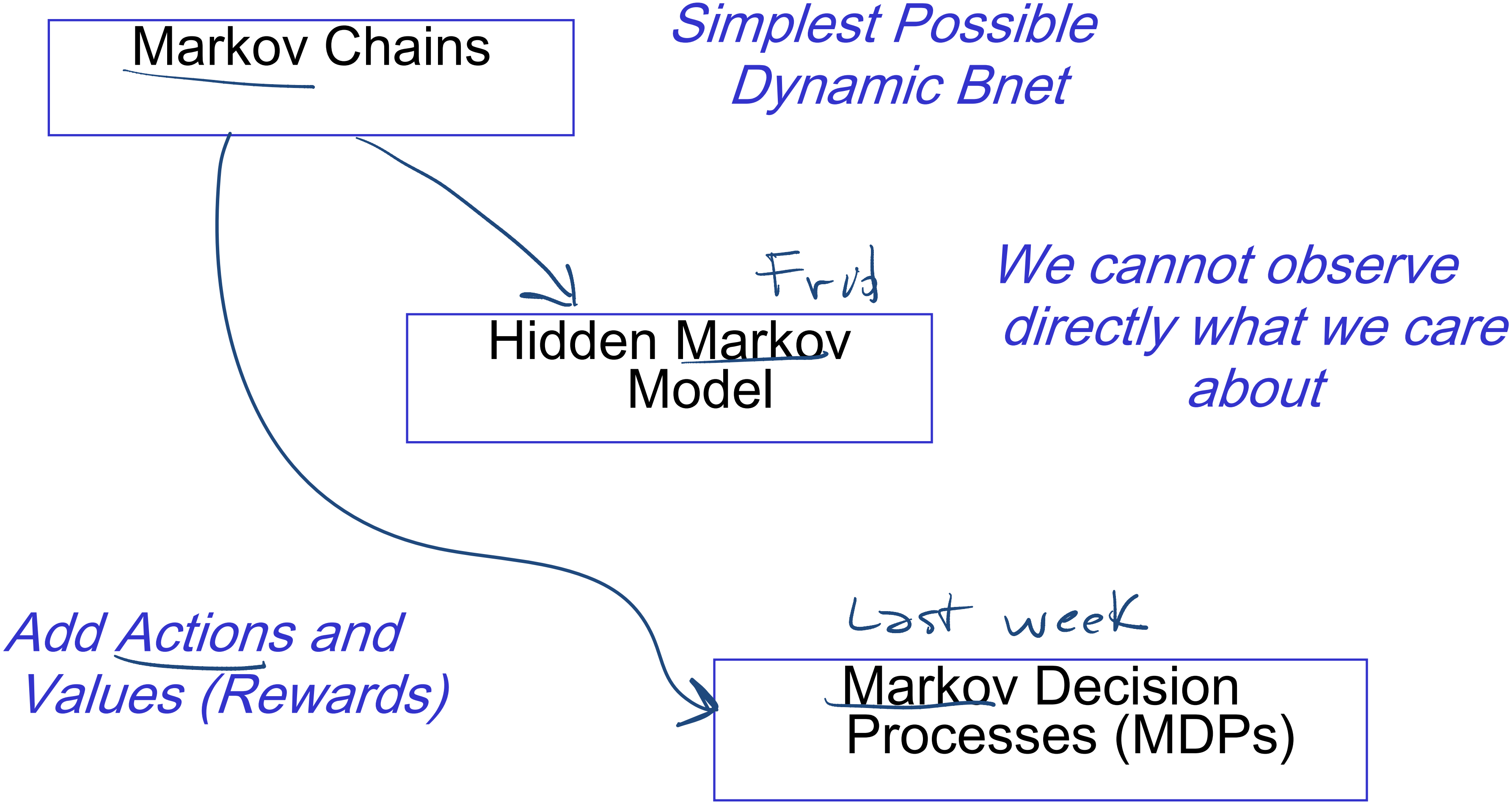
$$\boxed{P(w_i \mid w_{i-1}, w_{i-2})}$$

Learning Goals for today's class

You can:

- Specify a Markov Chain and compute the probability of a sequence of states
- Justify and apply Markov Chains to compute the probability of a Natural Language sentence (NOT to compute the conditional probabilities - slide 18)

Markov Models



Markov Chains

Simplest Possible Dynamic Bnet

Frid

Hidden Markov Model

We cannot observe directly what we care about

Add Actions and Values (Rewards)

Last week

Markov Decision Processes (MDPs)

Next Class

- **Finish Probability and Time:** Hidden Markov Models (HMM) *(TextBook 6.5)*
- **Start Decision networks** *(TextBook chpt 9)*

Course Elements

- **Assignment 4 is available on webCT . It is due on Apr the 8th (last class).**
  - You can now work on the first 3 questions. For the 4th one you have to wait until next week.