Today

- Project 1 Solution
- Assignment 2 Solutions
- Probability Distributions
- FSMs

Announcements

- Project 2 due next Tuesday 8/14 at 11:45pm
- Assignment 3 due Thursday 8/16 at 4:30pm

Project 1

- A python implementation is now available on the course website.
- Balancing Parentheses is the biggest issue.
- Best solution: use a recursive function to track each level of nesting

Project 1

```
#assumption: index is at '('
def parse_level(index, search_str):
  next_space = search_str.find(' ', index+1)
  this_node = search_str[index+1:next_space]
  if this_node == 'S':
    Counts.s_count += 1
  elif this_node == 'NP':
    Counts.np_count += 1
  elif this_node == 'VP':
    Counts.vp_count += 1
  next_lpar = search_str.find('(', index+1)
  next_rpar = search_str.find(')', index+1)
  daughters = []
  while (next_lpar != -1 and next_lpar < next_rpar):
    ret_val = parse_level(next_lpar, search_str)
    daughters = daughters + [ret_val[0]]
    next_lpar = search_str.find('(', ret_val[1])
    next_rpar = search_str.find(')', ret_val[1])
  if this_node == 'VP':
    if daughters == []:
      Counts.itv_count += 1
    elif daughters == ['NP', 'NP']:
      Counts.dtv_count += 1
  if next_rpar == -1:
    return (this_node, -1)
  else:
    return (this_node, next_rpar+1)
```

Assignment 2

1. Using the following sets, we run a trial which selects exactly one word from each set. Within each set, all words are equally likely.

A = { monkey, donkey, yak, kangaroo, aardvark, antelope, puma, cheetah }

B = { whale, shark, dolphin, eel }

$$|A| = 8$$
$$|B| = 4$$

There are 32 tuples

E = { either of the words contain a ‘y’ }

(monkey,whale) (monkey,shark) (monkey,dolphin) (monkey,eel) (donkey,whale) (donkey,shark) (donkey,dolphin) (donkey,eel) (yak,whale) (yak,shark) (yak,dolphin) (yak,eel)

F = { both words contain an ‘e’ }

(monkey,whale) (monkey,eel) (donkey,whale) (donkey,eel) (antelope,whale) (antelope,eel) (cheetah,whale) (cheetah,eel)

G = { both words contain the same number of letters }

(yak,eel) (cheetah,dolphin)

H = { either (or both) of the words contains *more than two* vowels { a e i o u }. This count includes repeated uses of the same vowel. }

(kangaroo,whale) (kangaroo,shark) (kangaroo,dolphin) (kangaroo,eel) (aardvark,whale) (aardvark,shark) (aardvark,dolphin) (aardvark,eel) (antelope,whale) (antelope,shark) (antelope,dolphin) (antelope,eel) (cheetah,whale) (cheetah,shark) (cheetah,dolphin) (cheetah,eel)

At this point the tuples are fixed for this problem so you could assign each tuple a unique number

E = { either of the words contain a ‘y’ }

0 1 2 3 4 5 6 7 8 9 10 11 $\frac{|E|}{|\Omega|} = \frac{12}{32}$=.375

F = { both words contain an ‘e’ }

0 3 4 7 20 23 28 31 $\frac{|F|}{|\Omega|} = \frac{8}{32} = .25$

G = { both words contain the same number of letters }

11 30 $\frac{|G|}{|\Omega|} = \frac{2}{32} = .0625$

H = { either (or both) of the words contains *more than two* vowels { a e i o u }. This count includes repeated uses of the same vowel. }

12 13 14 15 16 17 18 19 20 21 22 23 28 29 30 31 $\frac{|H|}{|\Omega|} = \frac{16}{32} = .5$

E = { either of the words contain a ‘y’ }

0 1 2 3 4 5 6 7 8 9 10 11

F = { both words contain an ‘e’ }

0 3 4 7 20 23 28 31

G = { both words contain the same number of letters }

11 30

H = { either (or both) of the words contains *more than two* vowels { a e i o u }. This count includes repeated uses of the same vowel. }

12 13 14 15 16 17 18 19 20 21 22 23 28 29 30 31

$$P(E \cup H) = \frac{28}{32} = .875$$

E = { either of the words contain a ‘y’ }

0 1 2 3 4 5 6 7 8 9 10 11

F = { both words contain an ‘e’ }

0 3 4 7 20 23 28 31

G = { both words contain the same number of letters }

11 30

H = { either (or both) of the words contains *more than two* vowels { a e i o u }. This count includes repeated uses of the same vowel. }

12 13 14 15 16 17 18 19 20 21 22 23 28 29 30 31

$$P(F \cap H) = \frac{4}{32} = .125$$

E = { either of the words contain a ‘y’ }

0 1 2 3 4 5 6 7 8 9 10 11

F = { both words contain an ‘e’ }

0 3 4 7 20 23 28 31

G = { both words contain the same number of letters }

11 30

H = { either (or both) of the words contains *more than two* vowels { a e i o u }. This count includes repeated uses of the same vowel. }

12 13 14 15 16 17 18 19 20 21 22 23 28 29 30 31

$$E \cap F \cap G = \emptyset$$
$$P(E \cap F \cap G) = 0$$

E = { either of the words contain a ‘y’ }

0 1 2 3 4 5 6 7 8 9 10 11

F = { both words contain an ‘e’ }

0 3 4 7 20 23 28 31

G = { both words contain the same number of letters }

11 30

H = { either (or both) of the words contains *more than two* vowels { a e i o u }. This count includes repeated uses of the same vowel. }

12 13 14 15 16 17 18 19 20 21 22 23 28 29 30 31

$$P(H \cup G) = \frac{17}{32} = .53125$$

F = { both words contain an ‘e’ }

0 3 4 7 20 23 28 31

$\bar{F}$: 1 2 5 6 8 9 10 11 12 13 14 15 16 17 18 19 21 22 24 25 26 27 29 30

H = { either (or both) of the words contains *more than two* vowels { a e i o u }. This count includes repeated uses of the same vowel. }

12 13 14 15 16 17 18 19 20 21 22 23 28 29 30 31

$$P(H \cap \bar{F}) = \frac{12}{32} = .375$$

	E	F	G
H	X	I	I
G		X	
F			

(E and G) are dependent
(E and F) are dependent

2. Working in Yunnan, a field linguist has discovered an extinct version of the Dongba pictographic script. So far, his team has found 32 distinct glyphs in this script, and the linguist has deciphered 22 of them. He just received news that another researcher has discovered a new inscription that consists of 8 glyphs. These 8 have all previously been encountered, but he doesn't yet know if the new inscription has repeated glyphs, or not.

a. What is the probability that the linguist will fully understand the newly discovered inscription?

"Inscription" implies that a glyph can appear more than once in the new discovery, so we assume trials that select with replacement:

$$\left(\frac{22}{32}\right)^8 = \frac{214{,}358{,}881}{4{,}294{,}967{,}296} \approx 0.05$$

b. What is the probability that the linguist will understand at least half of the glyphs in the newly discovered inscription?

$$\sum_{k=4}^{8}\left(\frac{22}{32}\right)^{k}\left(\frac{10}{32}\right)^{8-k}\binom{8}{k} \approx .9318$$

Explanation:

Canvas → Files → A2Q2

extra credit: If each of the 8 glyphs in the newly discovered inscription **are distinct from each other** (but still in the set of 32 known glyphs), what is the probability that the linguist will understand at least half of them?

number of sets containing k fully-understood glyphs

number of sets of $8 - k$ not-understood glyphs

$$\sum_{k=4}^{8} \frac{\binom{22}{k}\binom{10}{8-k}}{\binom{32}{8}} = \frac{134387}{140244} \approx 0.9582$$

number of possible subsets

Properties of probability distributions

- Let’s look at some of the important probability distributions
- First, these are the parameters we will use to describe those distributions:
  - Expected Value (Mean)
  - Variance
  - Standard Deviation

Expected Value

- Notation: $E[X]$
- Discrete: $E[X] = \sum x P_X(x)$
  - This should not be confused with "most probable value."
  - The expected value may be a value that is not in the domain
  - The expected value is only meaningful if the random variable's values are chosen meaningfully
- Continuous: $E[X] = \int x\, f(x) dx$
  - A weighted sum of all the possible values

Expected value as average

$$E[X] = \sum xP_X(x)$$

If the distribution is uniform (e.g., each outcome is equally likely):

$$E[X] = \sum_{i}^{n} x_i \frac{1}{n}$$

$$E[X] = \frac{1}{n}\sum_{i}^{n} x_i$$

$$E[X] = \frac{\sum x}{n} = \mu = \bar{x}$$

Measuring “spread”

$$E[X] = 50$$

$$X = \{50, 50, 50, 50, 50, 50\}$$
$$X = \{\,47, 48, 49, 51, 52, 53\,\}$$
$$X = \{\,0, 0, 0, 100, 100, 100\,\}$$

$$E[X - \mu] = ?$$

Variance

- Discrete

$$Var(X) = \sum_{i}^{n} P(x_i)(x_i - \mu)^2$$

- Continuous

$$Var(X) = \int (x_i - \mu)^2 f(x_i) dx$$

Variance

Q: why do we get to cancel this term?

$$\sigma_X{}^2 = \mathrm{Var}(X) = E[(X-\mu)^2]$$
$$= \sum (x-\mu)^2 P(x)$$
$$= \sum (x^2 - 2x\mu + \mu^2) P(x)$$
$$= \sum x^2 P(x) - 2P(x)x\mu + \mu^2 P(x)$$
$$= \sum x^2 P(x) - \sum 2P(x)x\mu + \sum \mu^2 P(x)$$

$$= E[X^2] - 2\mu \sum P(x)x + \mu^2 \cancel{\sum P(x)}$$
$$= E[X^2] - 2\mu^2 + \mu^2$$
$$= E[X^2] - \mu^2$$
$$= E[X^2] - E[X]^2$$

Standard deviation

Defined as the square root of the variance

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

Covariance

- How much does X vary with regard to Y

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$
$$= E[XY] - E\big[XE[Y]\big] - E\big[YE[X]\big] + E\big[E[X]E[Y]\big]$$
$$= E[XY] - E[X]E[Y] - E[X] - E[Y] + E[X]E[Y]$$
$$= E[XY] - E[X]E[Y]$$

Probability distributions

- Discrete distributions
  - Uniform
  - Bernoulli
  - Binomial
  - Geometric
  - Poisson
- Continuous distributions
  - Uniform
  - Normal

Discrete probability distributions

Uniform distribution (discrete)

- Every discrete value is equally likely to occur

$$a, b \in \mathbb{Z}, \mathrm{a} \leq \mathrm{b}$$

$$n = b - a + 1$$

$$\mu = \frac{a + b}{2}$$

$$\sigma^2 = \frac{n^2 - 1}{12}$$

Only two outcomes?

- Remember random variables
- Events alone were not convenient for correlating probabilities with stochastic trials because they
  - only partition sample spaces into two subsets
  - each imply independent, well-formed probability spaces without regard to other outcomes that we might be interested in.

Having said this, what if there *are* only two outcomes in our experiment?

Bernoulli Trial

- A Bernoulli trial is an experiment with only two outcomes

$$\Omega = \{\, yes, no \,\}$$

- If the outcome is modeled by a random variable

$$X = \begin{cases} 1, if\ the\ result\ is\ yes, \\ 0, if\ the\ result\ is\ no. \end{cases}$$

then random variable X has a Bernoulli distribution

- This discrete probability distribution can be described with a single parameter

$$p = P(X = 1)$$

Bernoulli distribution

- Two outcomes: { success, failure }
- Parameter: $0 \leq p \leq 1, p \in \mathbb{R}$

$$P(X = x) = \begin{cases} p, & \text{if x=success} \\ 1 - p, & \text{if } x = \text{failure} \\ 0, & \text{otherwise} \end{cases}$$

$$\mu = p$$

$$\sigma^2 = p(1 - p)$$

Binomial distribution

- Model the number of successes in n Bernoulli trials
- Parameters: p, n

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$\mu = np$$
$$\sigma^2 = np(1-p)$$

- Binomial(n, p) = Bernoulli distribution

Binomial distribution

Q: A corpus contains 4,000 newswire articles, covering every day of the week. An article is selected at random. Let E be the event that the article is from a Sunday. What is the probability distribution for E?

A: Binomial distribution with:

$$p = \frac{1}{7}$$

$$\mu = \frac{1}{7}$$

$$\sigma^2 = \frac{1}{7}\left(1 - \frac{1}{7}\right)$$

Geometric distribution

$$X = \{\text{ number of Bernoulli trials until obtaining success }\}$$

Parameter: p from Bernoulli trial

$$(1-p)(1-p)(1-p)\ldots(1-p)p$$
$$P(X=x) = (1-p)^{x-1}p$$
$$P(X>x) = (1\ -p)^{x}$$
$$\mu = \frac{1}{p}$$
$$\sigma^2 = \frac{1-p}{p^2}$$

Geometric distribution

Q: A fair coin is flipped T times until it comes up heads. Characterize $P(T)$.

A: Geometric distribution with

$$p = .5$$
$$\mu = 2$$
$$\sigma^2 = 2$$

Poisson distribution

- The number of independent events that will probably occur during a period of time, given the rate of events
- Parameter: $\lambda = \text{expected \# of events per interval}$

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$\mu = \lambda$$
$$\sigma^2 = \lambda$$

Poisson Distribution

- The phone rings 5 times per hour on average
- What is the probability of an hour going by without the phone ringing?

$$p(0) = \frac{5^0 e^{-5}}{0!}$$
$$= .0067$$

Continuous probability distributions

Uniform distribution (continuous)

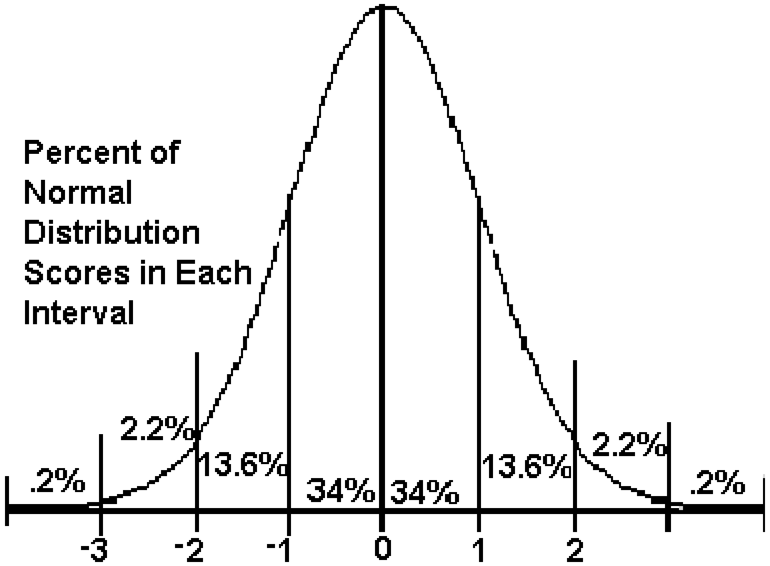
$$f(x) = \begin{cases} \dfrac{1}{b-a}, if\ a \leq x \leq b \\ 0, if\ x < a\ or\ x > b \end{cases}$$

$$\mu = \frac{a+b}{2}$$

$$\sigma^2 = \frac{(b-a)^2}{12}$$

Normal Distribution

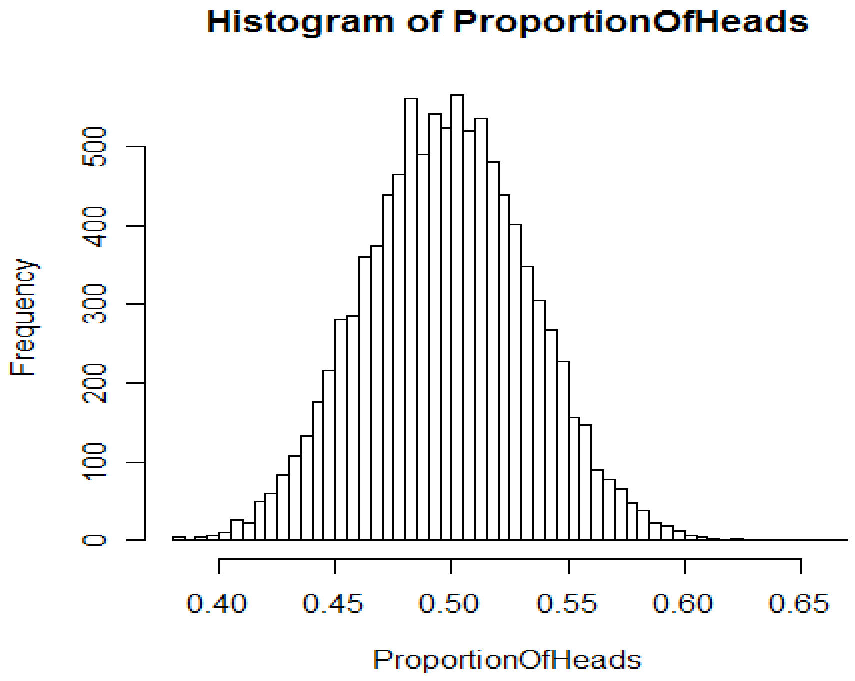
- aka Gaussian distribution
- Parameters:
  - μ
  - σ^2
- $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$



Central Limit Theorem

- When a large number of independent random variables is added together, its sum approaches a normal distribution
- Consider a fair coin toss
- $X = \{\, it\ shows\ heads \,\}$
- $P(X = heads) = .5$
- Many trials of this r.v.

will be normally distributed



Finite state machines

or, finite state automata

- Deterministic
- Non-deterministic

{ set of states, transitions, start state, input alphabet, final states }

- Finite state transducers
- Acceptor

Deterministic FSM

$q \in S$	States
$S_0 \in S$	Start state
$x \in \Sigma$	Input alphabet
$F \in S$	Final states (or Ø)
$\delta : S \times \Sigma \rightarrow S$	Transitions

Each state/input pair has no more than one transition

Non-deterministic FSM

$q \in S$ States

$S_0 \in S$ Start state

$x \in \Sigma$ Input alphabet

$F \in S$ Final states (or Ø)

P_S Transition probabilities

$\delta: S \times \Sigma \times P_S \rightarrow S$ Transitions

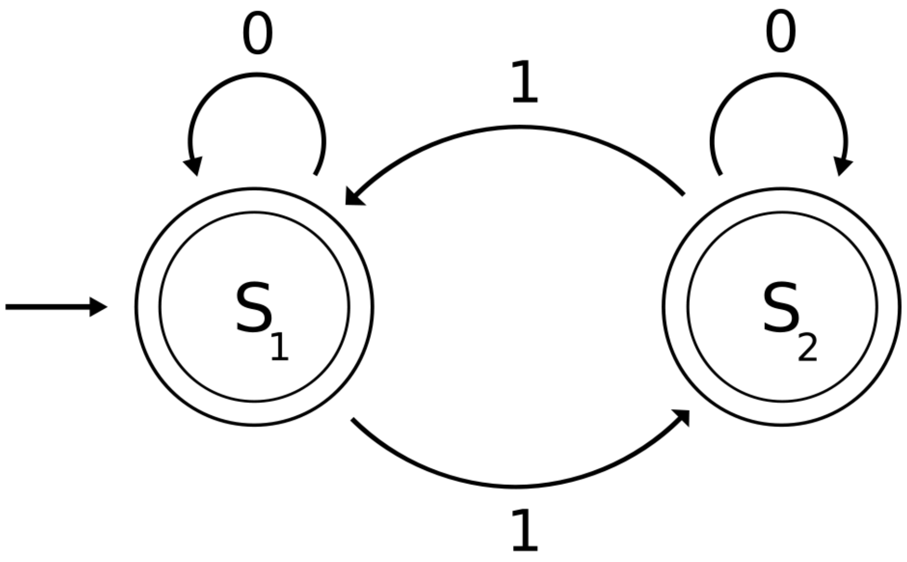
For a given state/input, there may be more than one possible transition

At runtime

- This is sufficient description of the machine. At runtime, an input stream composed of symbols from alphabet Σ is provided
- If $\delta(q, x)$ is not present, the FSM is said to reject the input

Example

parity: the number of bits in a binary value that are 'set' to one (1)

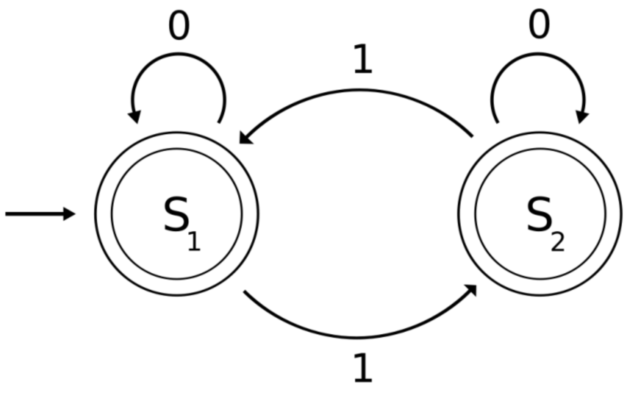


Double-circles are used to indicate accepting states

parity of the binary input:
S1 : even
S2 : odd

1 0 1 1 0 0 1 → S1
0 0 0 1 0 0 0 → S2

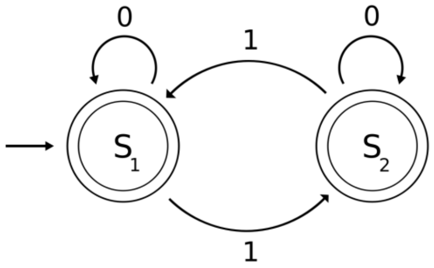
Example



State	Transition
S1	0 → S1, 1 → S2
S2	0 → S2, 1 → S1

Programming FSTs

```
int Parity(String s)    // i.e. "00101010"
{
    int state = 1;
    foreach (Char ch in s)
        switch (state)
        {
            case 1:
                if (state == '1')
                    state = 2;
                break;
            case 2:
                if (state == '1')
                    state = 1;
                break;
        }
    return state;
}
```



Example

- Write an FSA for the RegEx:

  a[ab]*b[cd]

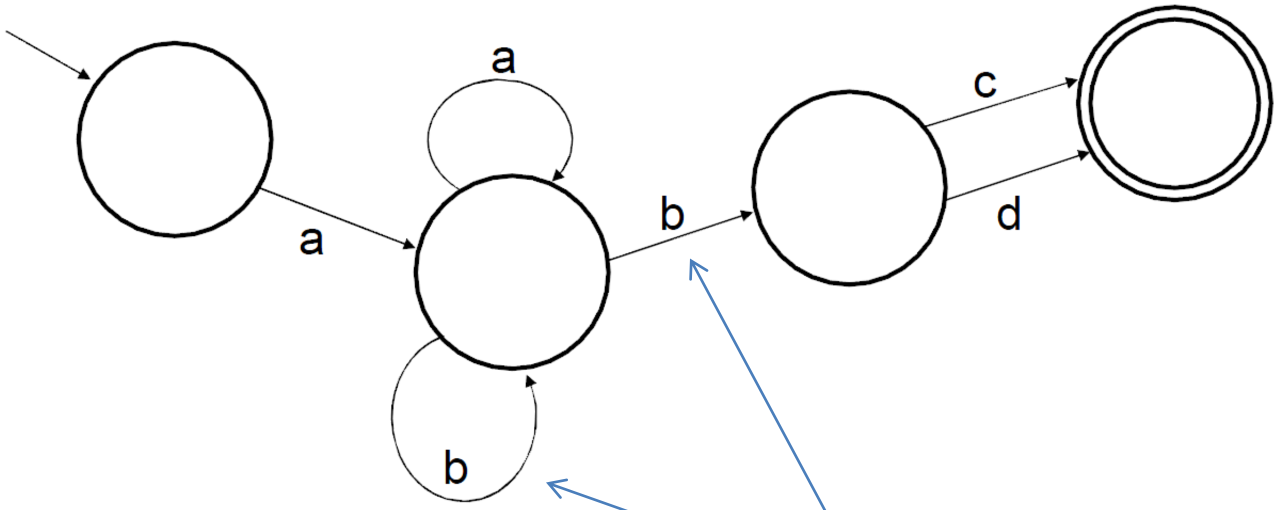
FSM example

Is your FSM deterministic or non-deterministic?

Example

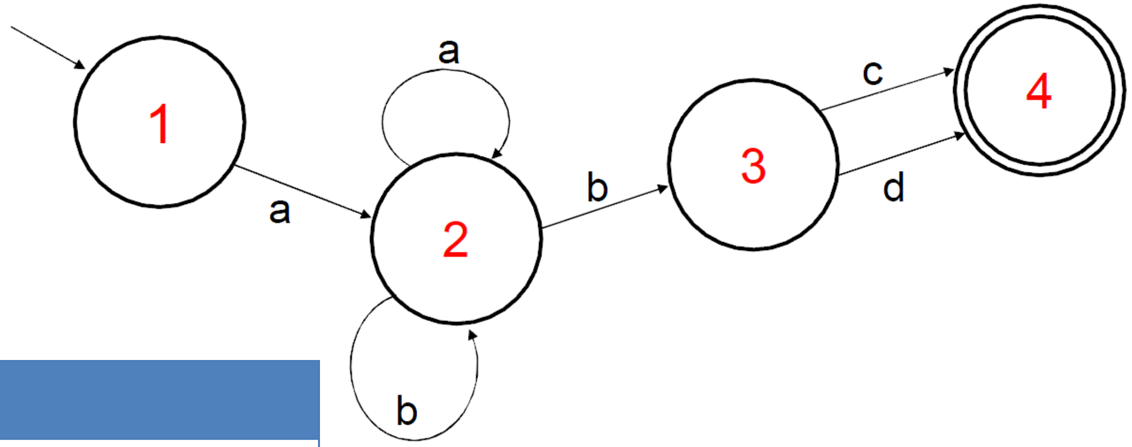
- Non-deterministic

  a[ab]*b[cd]



Two ways to go

a[ab]*b[cd]



State	Transition
1	a → 2
2	a → 2, b → 2, b → 3
3	c → 4, d → 4

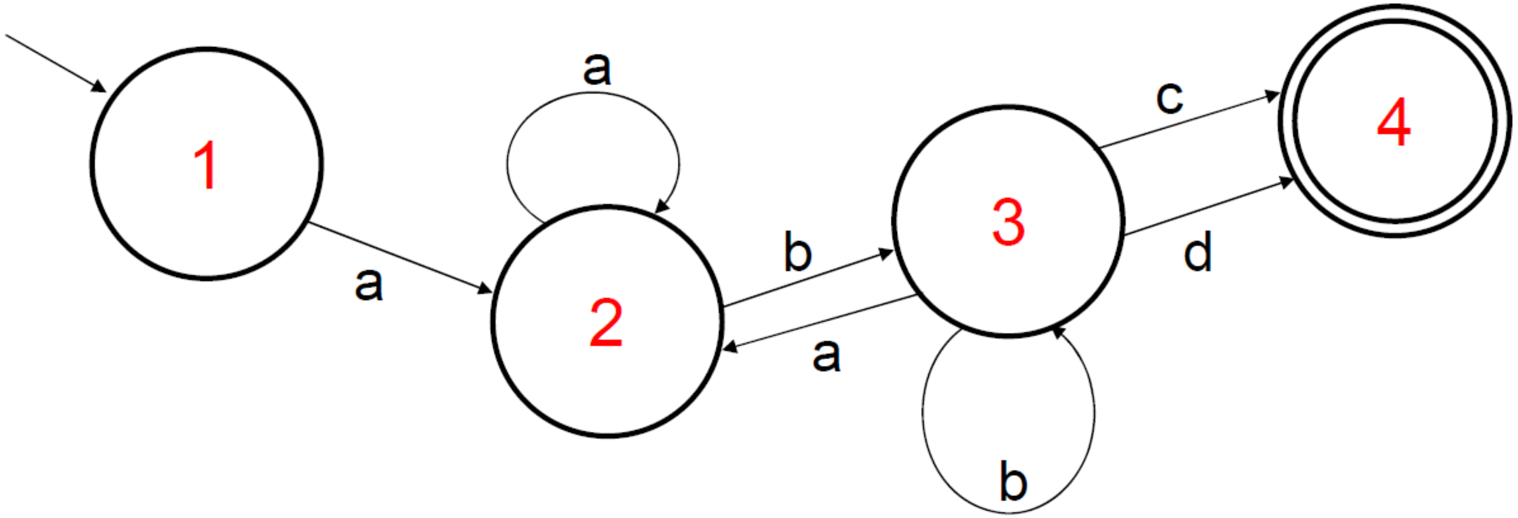
How would we implement this state machine?
- backtracking
- running multiple paths

Example: abbcd
if we choose state 3 here, we will fail to accept this pattern when we should have

Example

- deterministic

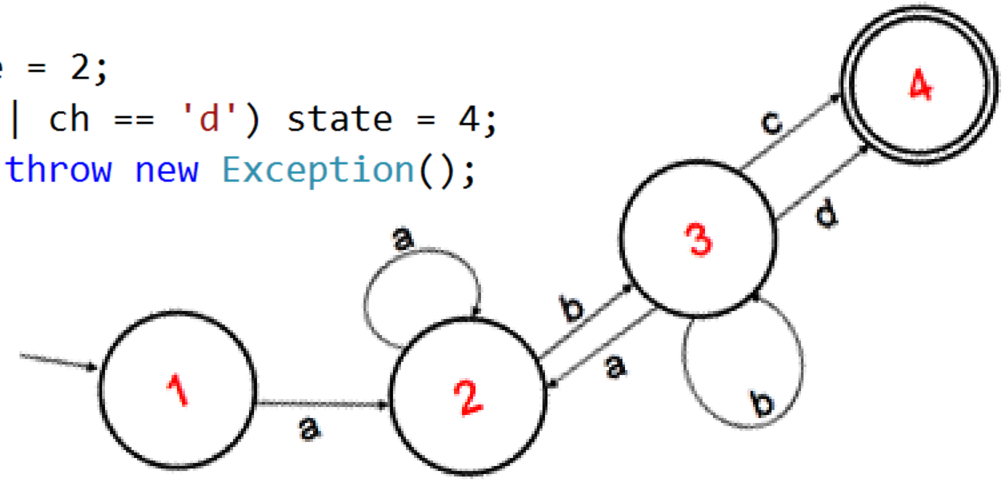
  a[ab]*b[cd]



FSA

```
IEnumerable<int> FST(String input) {
    int state = 1;
    foreach (Char ch in input) {
        switch (state) {
            case 1:
                if (ch == 'a') state = 2;
                else throw new Exception();
                break;
            case 2:
                if (ch == 'b') state = 3;
                else if (ch != 'a') throw new Exception();
                break;
            case 3:
                if (ch == 'a') state = 2;
                else if (ch == 'c' || ch == 'd') state = 4;
                else if (ch != 'b') throw new Exception();
                break;
            case 4:
                yield break;
        }

    }
}
```



Finite state transducer (FST)

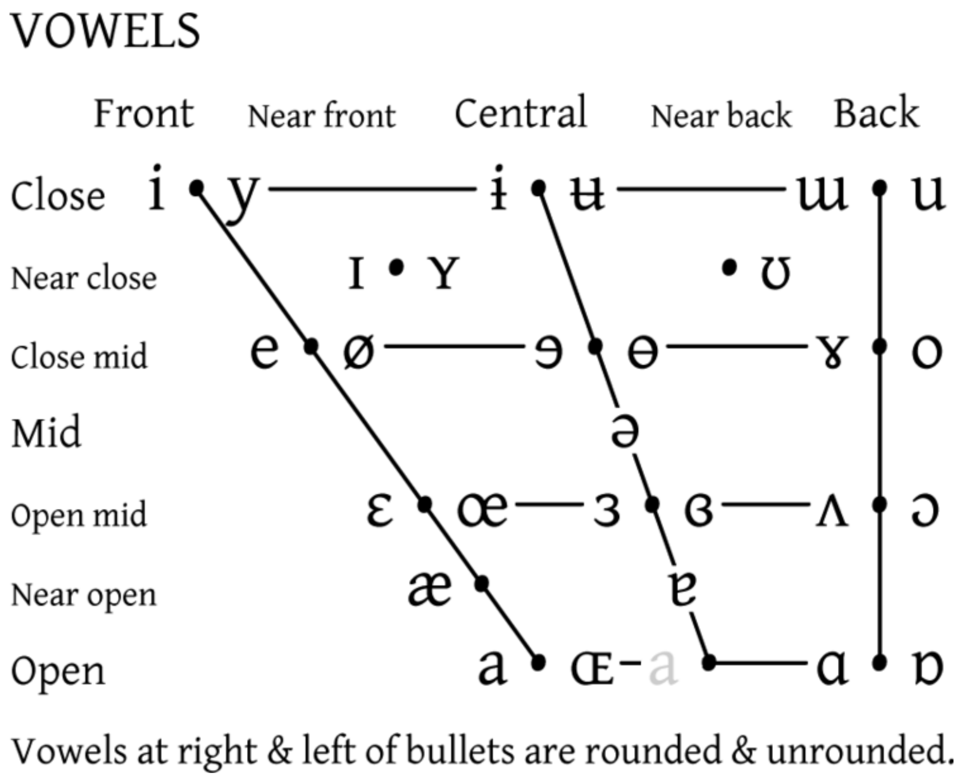
- Add an output function (per state or per trannsition) to an FSM
- The function fires upon arriving at a state or on transition
- Output models:
  - Mealy model: the output depends on both the current input and the state (“on transition”)
  - Moore model: the output depends only on the state

Assignment 3

- Due next Thursday, August 16th at 4:30pm PDT
- 4 probability/statistics problems that will draw together what we’ve been studying:
  - Conditional probability
  - Random variables
  - Bayes theorem

Open/Closed Vowels

Quick intro to vowel phonetics, since this is mentioned in Assignment 3



POS tag probabilities

The Red Badge of Courage, by Stephen Crane

DT	NN	VBD	RB	IN	DT	NN	,	CC	DT	VBG	NNS	VBD	DT	NN	VBD	IN	IN	DT
the	cold	passed	reluctantly	from	the	earth	,	and	the	retiring	fogs	revealed	an	army	stretched	out	on	the

NNS	,	VBG	.	IN	DT	NN	VBN	IN	JJ	TO	VB	,	DT	NN	VBN	,	CC	VBD	TO	VB
hills	,	resting	.	as	the	landscape	changed	from	brown	to	green	,	the	army	awakened	,	and	began	to	tremble

IN	NN	IN	DT	NN	IN	NNS	.	PRP	NN	PRP$	NNS	IN	DT	NNS	,	WDT	VBD	VBG	IN	JJ
with	eagerness	at	the	noise	of	rumors	.	it	cast	its	eyes	upon	the	roads	,	which	were	growing	from	long

NNS	IN	NN	NN	TO	JJ	NNS	.	DT	NN	,	JJ	IN	DT	NN	IN	PRP$	NNS	,
troughs	of	liquid	mud	to	proper	thoroughfares	.	a	river	,	amber-tinted	in	the	shadow	of	its	banks	,

VBD	IN	DT	NN	POS	NNS	:	CC	IN	NN	,	WRB	DT	NN	VBD	VBN	IN	DT	JJ	NN	,	NN
purled	at	the	army	's	feet	;	and	at	night	,	when	the	stream	had	become	of	a	sorrowful	blackness	,	one

MD	VB	IN	PRP	DT	JJ	,	JJ	NN	IN	JJ	NNS	VBN	IN	DT	JJ	NNS	IN	JJ	NNS	.
could	see	across	it	the	red	,	eyelike	gleam	of	hostile	camp-fires	set	in	the	low	brows	of	distant	hills	.

unigram count: 123
bigram count: 122

$$P(NN) = \frac{18}{123} \quad P(IN) = \frac{20}{123} \quad P(NN|IN) = \frac{3}{20} \quad P(DT\ NN) = \frac{10}{122}$$

Next Time

- FSMs
- Complexity