

Probabilistic Languages

Some Models of Translation

- IBM Models 1-5
- Hidden Markov Model
- Phrase-Based Models

Q: What do all of these things have in common?

“Language”

Definition 1: *Natural* language

“Language”

Definition 1: *Natural* language



Guid tae see ye at the Scots Wikipædia, the first encyclopædia in the Scots leid!

Wikipædia is a project tae big a free encyclopædia in mony leids.
This Scots edeetion wis shapit on 23rd Juin 2005. We hae 8,469 airticles the nou.
There's 12,465 veesitors/uisers here the nou.

 **Gettin Stairtit**
Walcome page — Writin Scots — Editin Lessons

 **Scotland**
Aw the airts — Cultur — Economy — Fowk — Law — Leids — Politics

 **Applee'd sciences**
Agriculturn — Airchitectur — Communication — Electronics — Engineerin — Heal — Industry — Medicine — Transport — Wather

 **Naitural Sciences an Maths**
Astronomy — Biology — Chemistry — Computers — Yird science — Mathematics — Pheesics

 **Fowk an Social Studies**
Anthropology — Airchaeology — Geography — Eddication — History — Leids — Philosophy — Psychology — Releegion — Sociology

 **Govrenment an Law**
Economics — Govrenment — Law — Military — Politics

 **Airt an Cultur**
Airt - Fuid & Farin — Cultur — Dance — Habbies — Media — Muisic — Gemmes & Sports — Theatre

Tends to not be well-defined.

“Language”

Definition 1: *Natural* language



Tends to not be well-defined.

“Language”

Definition 1: *Formal* language

Well-defined, so that a computer can process it:
a (possibly infinite) set of strings.

- All of the English words in a dictionary.
- All sequences of any length over those words.
- All English sentences with non-zero $p(e | f)$ for some French sentence f , according to your model.

Desiderata

- We need efficient algorithms and data structures to:
 - Encode all of the strings in the language.
 - Assign probabilities to all of those strings.
 - Via products such as $p(e)p(f|e)$.
 - Find the string with the highest probability.
 - Compute expectations over substrings.
 - Compute mappings between strings.

Regular Languages

Regular Languages

$$\mathcal{L}_1 = \left\{ \begin{array}{l} a a a \\ a b a \\ a a b \\ a b b \end{array} \right\}$$

Regular Languages

$$\mathcal{L}_1 = \left\{ \begin{array}{c} a a a \\ a b a \\ a a b \\ a b b \end{array} \right\}$$

$$\mathcal{L}_2 = a^* = \{a, aa, aaa, \dots\}$$

Regular Languages

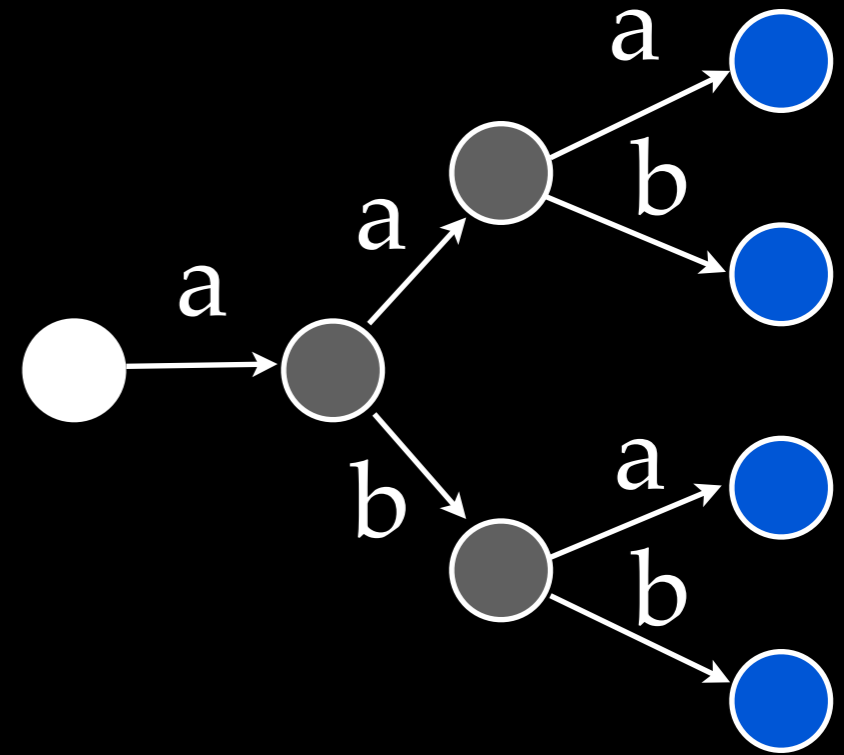
$$\mathcal{L}_1 = \left\{ \begin{array}{c} a a a \\ a b a \\ a a b \\ a b b \end{array} \right\}$$

$$\mathcal{L}_2 = a^* = \{a, aa, aaa, \dots\}$$

$$\mathcal{L}_3 = \{ \text{“the north wind howls”} \}$$

Regular Languages

$$\mathcal{L}_1 = \left\{ \begin{array}{l} a a a \\ a b a \\ a a b \\ a b b \end{array} \right\}$$

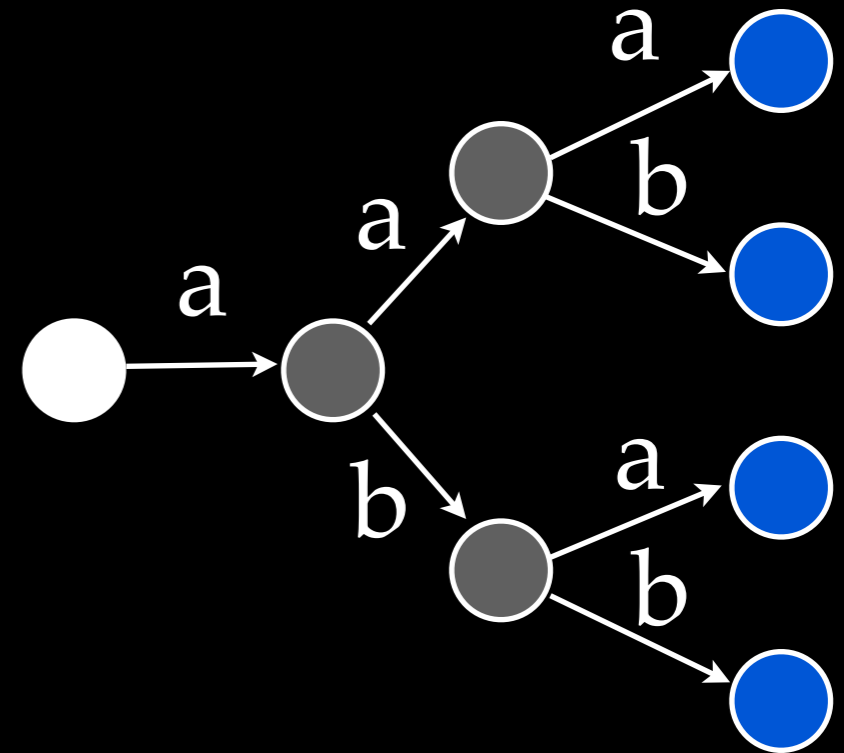


$$\mathcal{L}_2 = a^* = \{a, aa, aaa, \dots\}$$

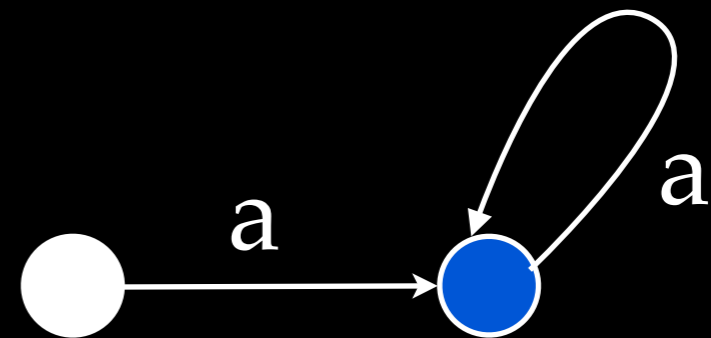
$$\mathcal{L}_3 = \{ \text{"the north wind howls"} \}$$

Regular Languages

$$\mathcal{L}_1 = \left\{ \begin{array}{l} a a a \\ a b a \\ a a b \\ a b b \end{array} \right\}$$



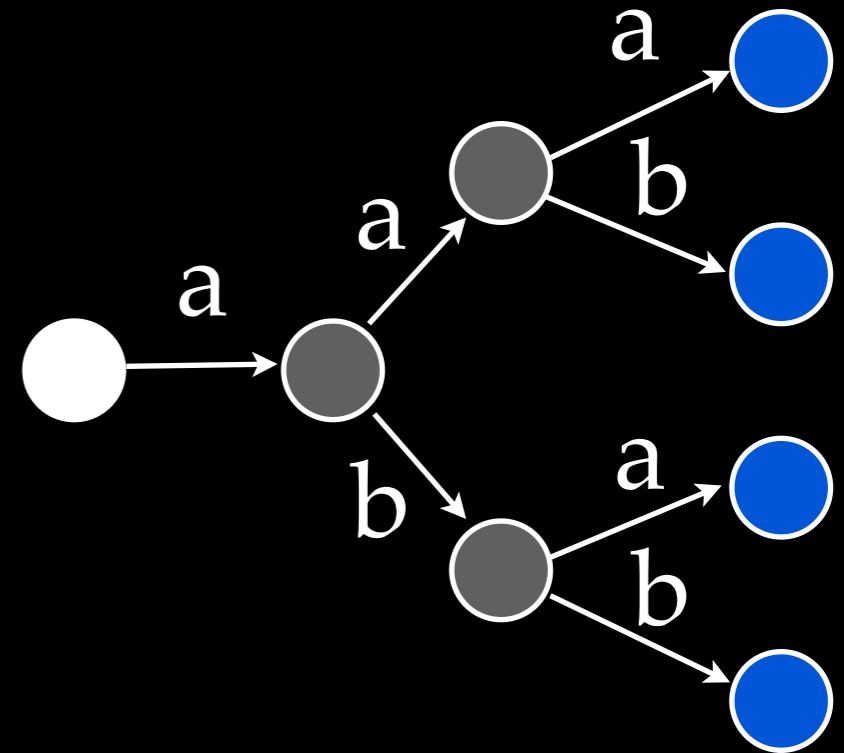
$$\mathcal{L}_2 = a^* = \{a, aa, aaa, \dots\}$$



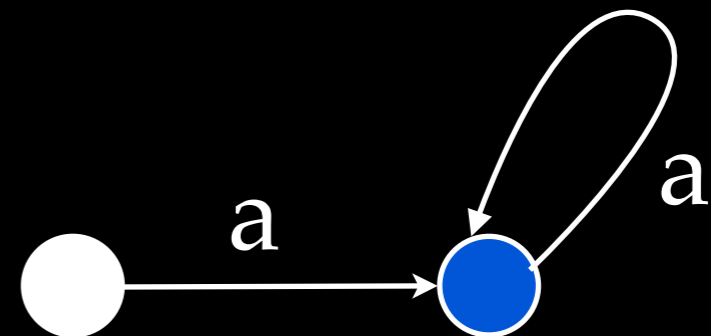
$$\mathcal{L}_3 = \{ \text{"the north wind howls"} \}$$

Regular Languages

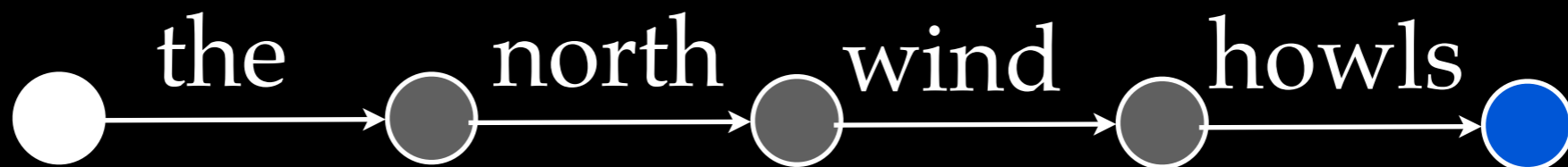
$$\mathcal{L}_1 = \left\{ \begin{array}{l} a a a \\ a b a \\ a a b \\ a b b \end{array} \right\}$$



$$\mathcal{L}_2 = a^* = \{a, aa, aaa, \dots\}$$

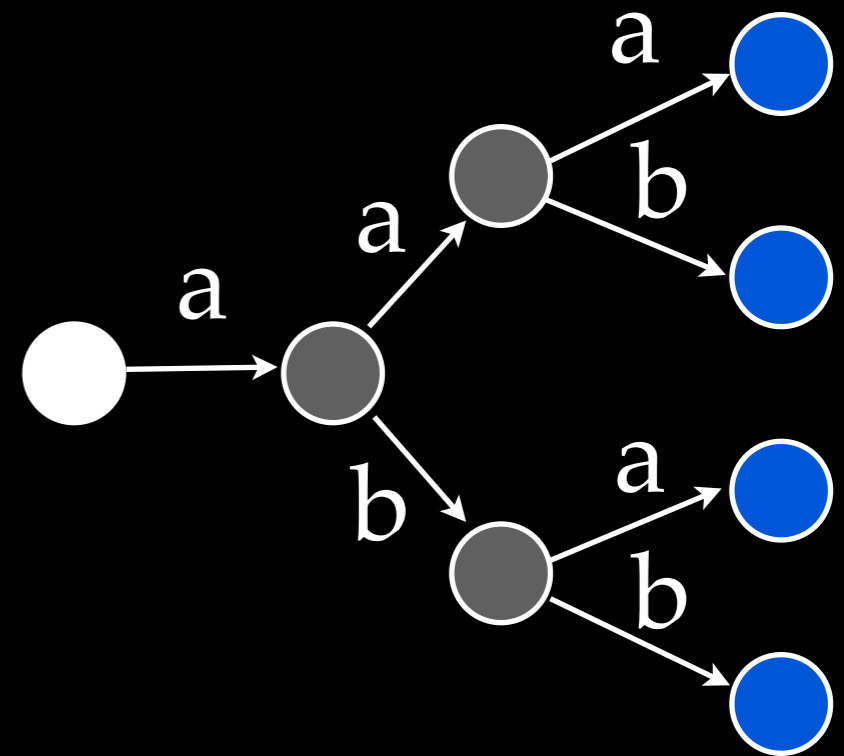


$$\mathcal{L}_3 = \{ \text{"the north wind howls"} \}$$

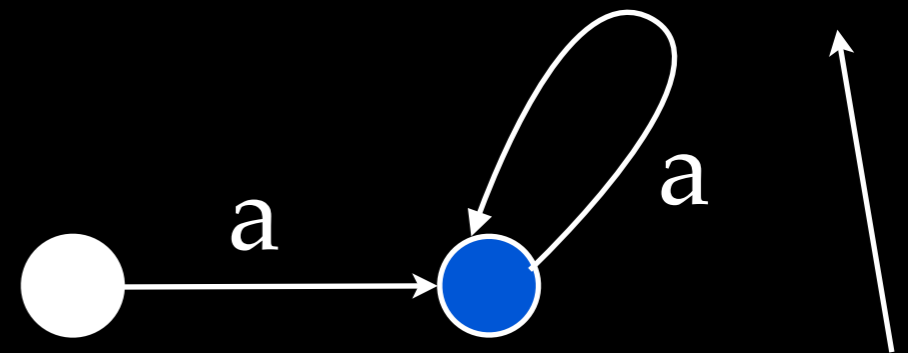


Regular Languages

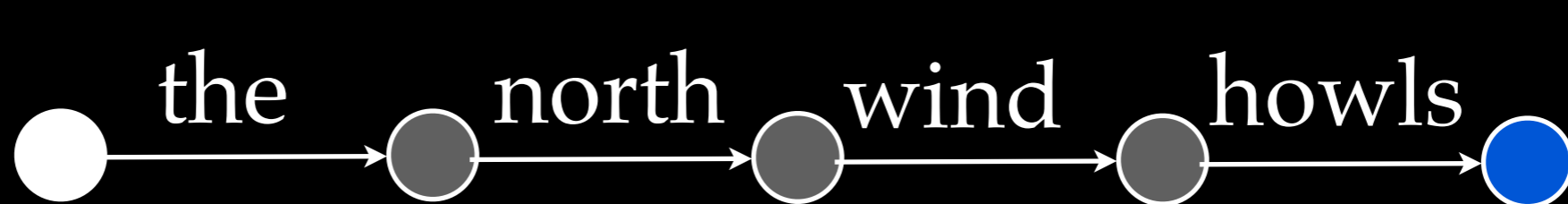
$$\mathcal{L}_1 = \left\{ \begin{array}{l} a a a \\ a b a \\ a a b \\ a b b \end{array} \right\}$$



$$\mathcal{L}_2 = a^* = \{a, aa, aaa, \dots\}$$



$$\mathcal{L}_3 = \{ \text{"the north wind howls"} \}$$



finite-state automata

Regular Languages

Regular Languages

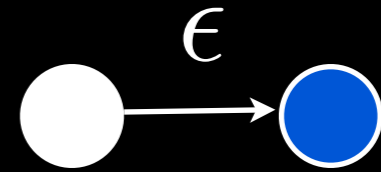
$\{\epsilon\}$ is regular

Regular Languages

$\{\epsilon\}$ is regular 

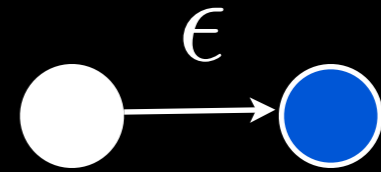
Regular Languages

$\{\epsilon\}$ is regular



Regular Languages

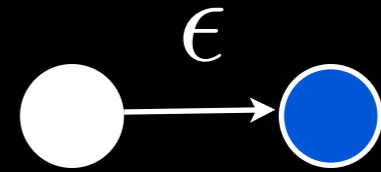
$\{\epsilon\}$ is regular



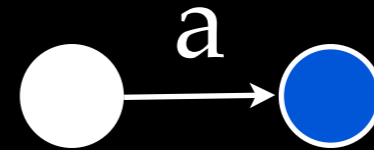
$\{a\}$ is regular

Regular Languages

$\{\epsilon\}$ is regular



$\{a\}$ is regular



Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

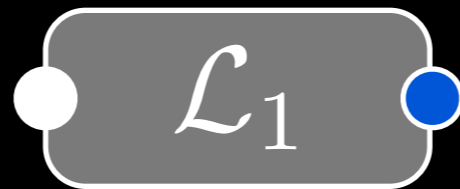
$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular

Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular

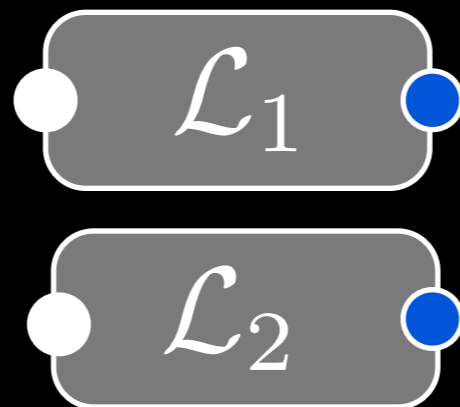


Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular

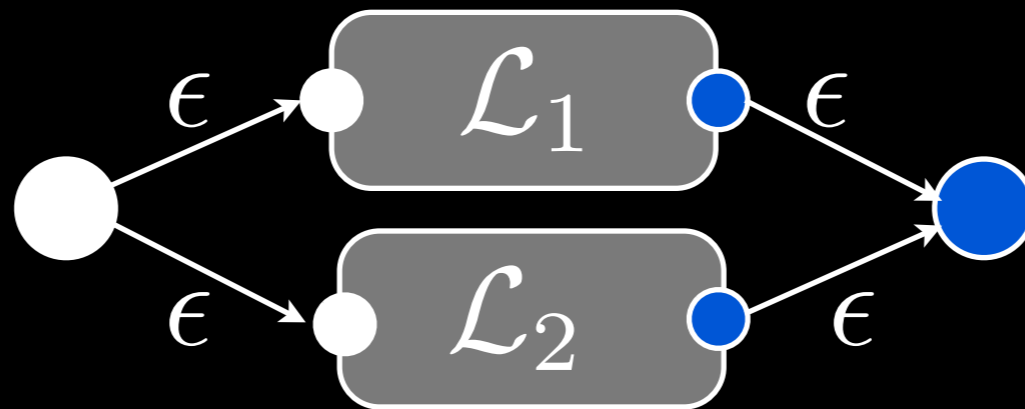


Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular

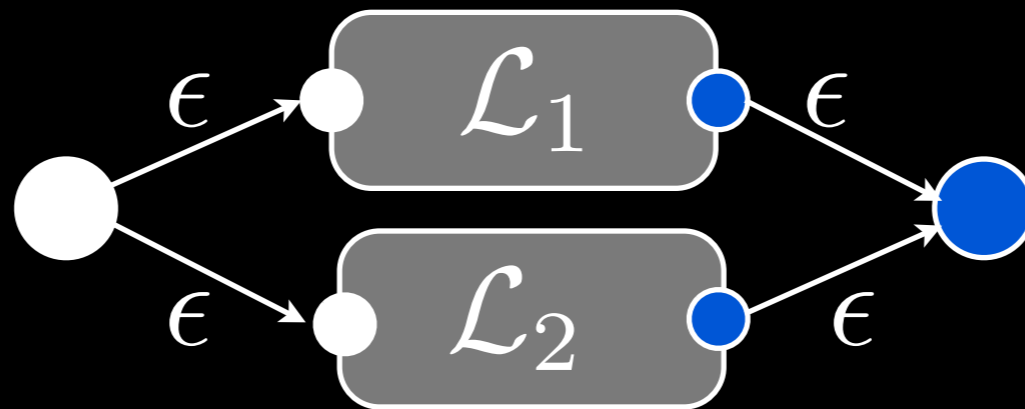


Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular



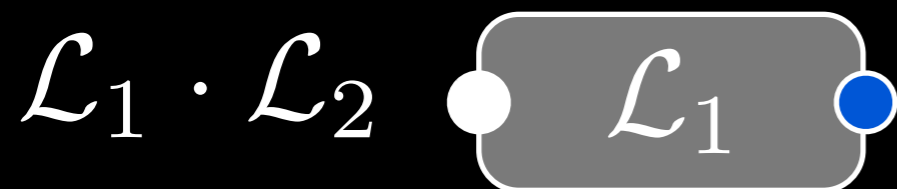
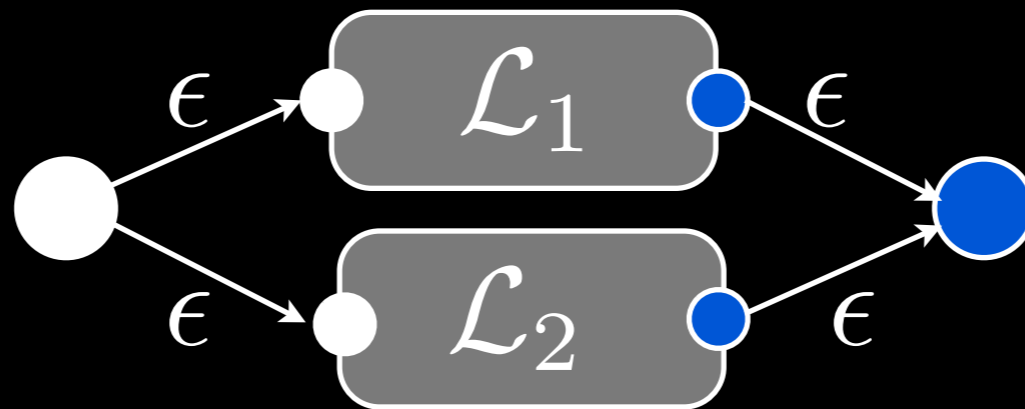
$\mathcal{L}_1 \cdot \mathcal{L}_2$

Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular

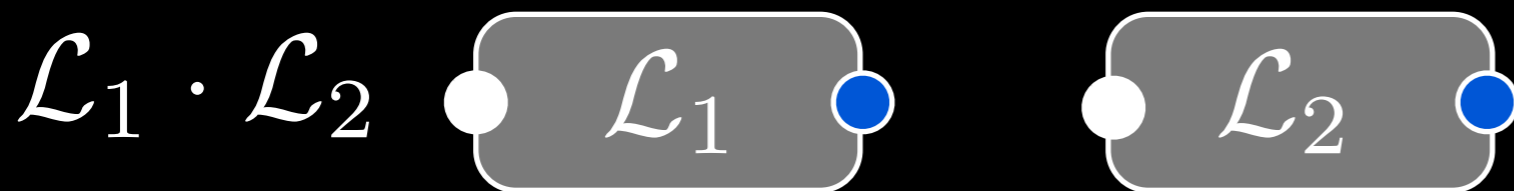
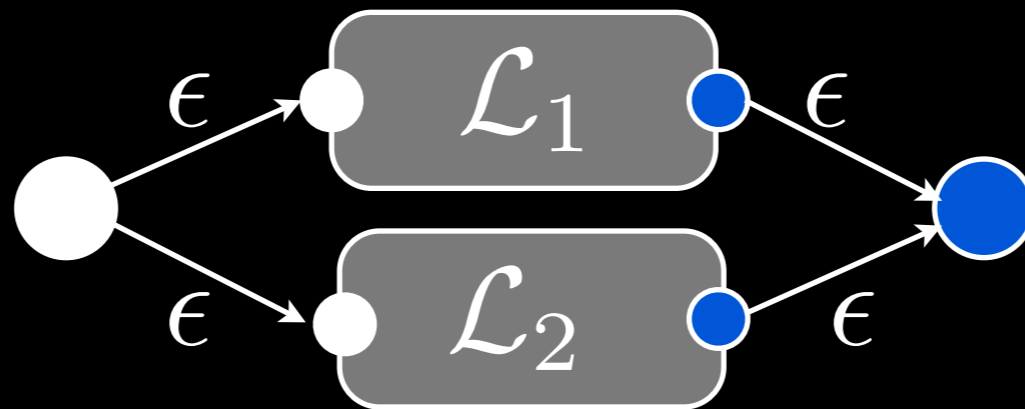


Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular

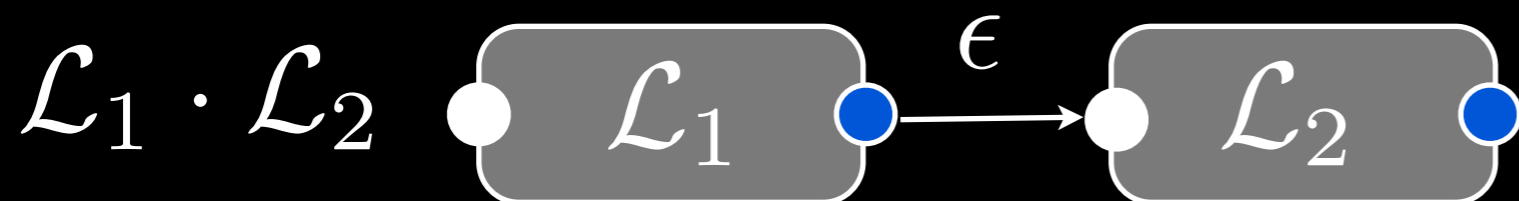
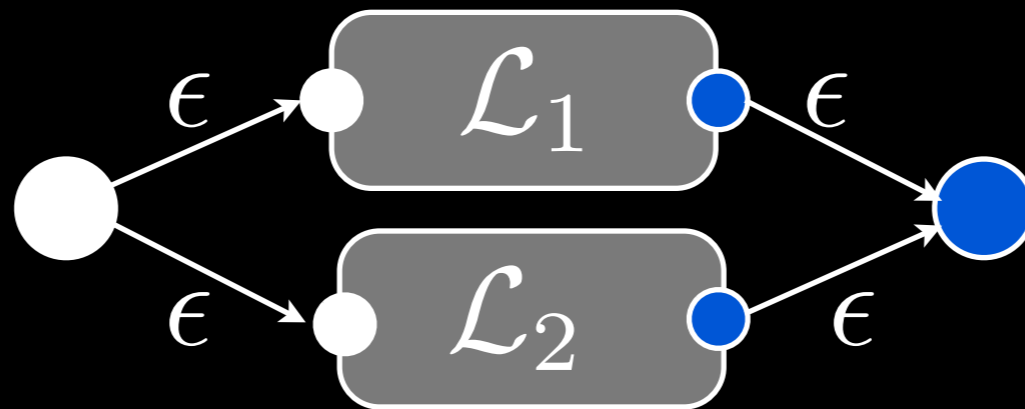


Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular

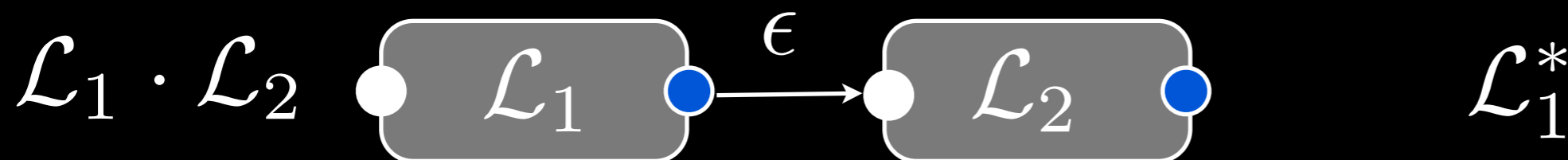
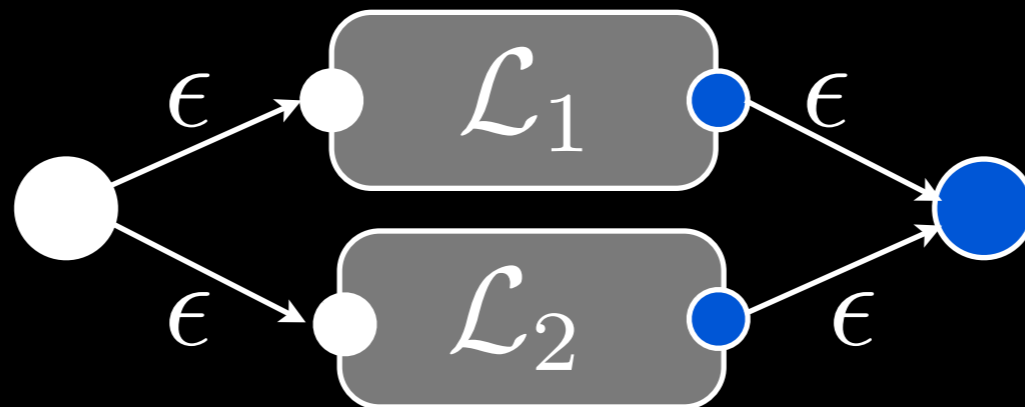


Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular

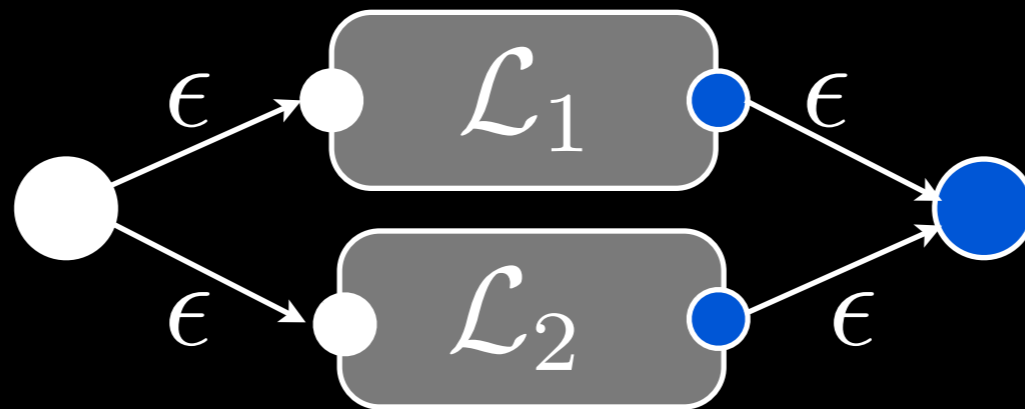


Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular

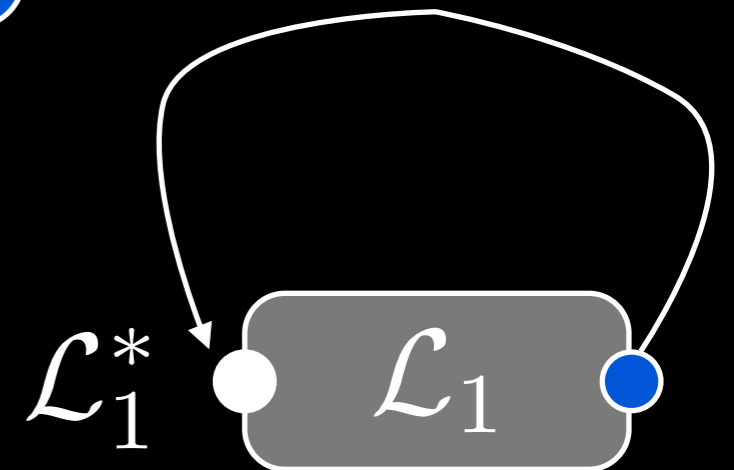
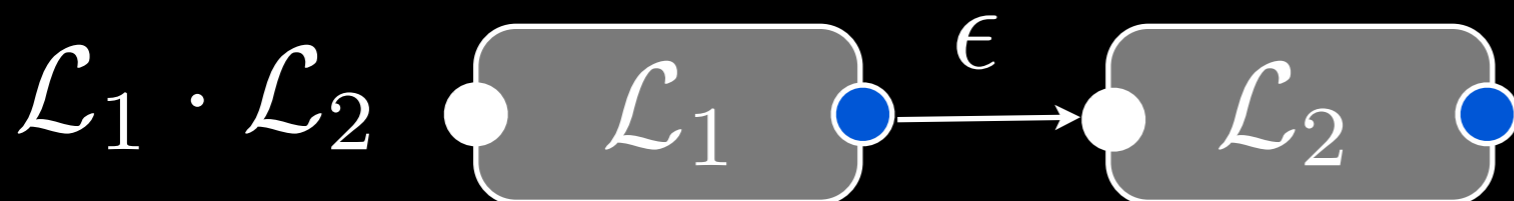
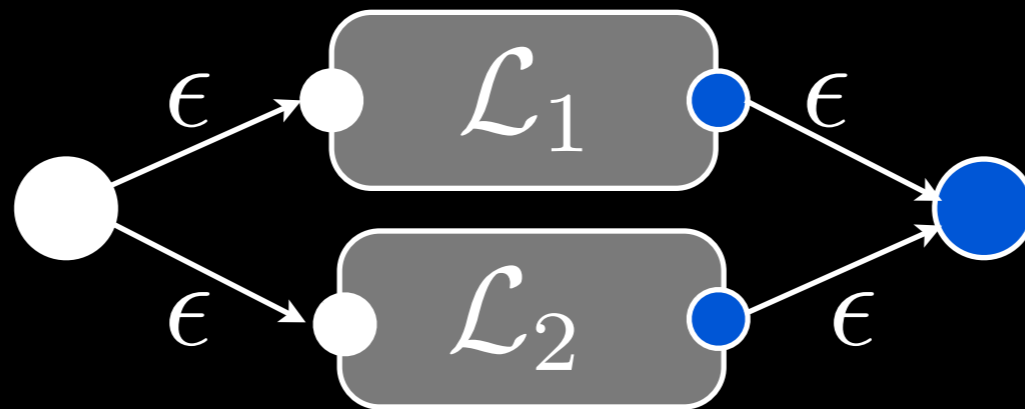


Regular Languages

$\{\epsilon\}$ is regular  

$\{a\}$ is regular 

$\mathcal{L}_1 \cup \mathcal{L}_2$ is regular if \mathcal{L}_1 and \mathcal{L}_2 are regular



Regular Languages

Not all languages are regular!

Regular Languages

Not all languages are regular!

$$\mathcal{L}_4 = \{ab, aabb, aaabbb, \dots\} = \forall_{n \in [1, \text{inf})} a^n b^n$$

Regular Languages

Not all languages are regular!

$$\mathcal{L}_4 = \{ab, aabb, aaabbb, \dots\} = \forall_{n \in [1, \text{inf})} a^n b^n$$

We'll talk about such *context-free* languages next week

Regular Languages

Not all languages are regular!

$$\mathcal{L}_4 = \{ab, aabb, aaabbb, \dots\} = \forall_{n \in [1, \text{inf})} a^n b^n$$

We'll talk about such *context-free* languages next week

But not all languages are context-free, either!

Probabilistic Regular Languages

We want a function:

$$f : \mathcal{L} \rightarrow \mathbb{R}^+$$

Probabilistic Regular Languages

We want a function:

$$f : \mathcal{L} \rightarrow \mathbb{R}^+$$

such that:

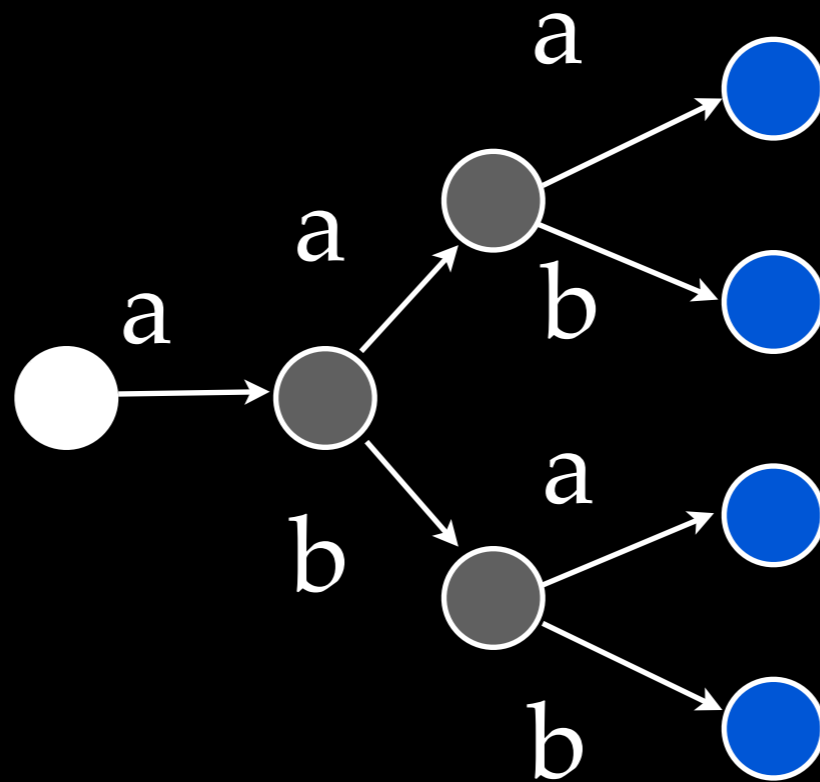
$$f(w) \in [0, 1]$$

$$\sum_w f(w) \in [0, 1]$$

Probabilistic Regular Languages

We want a function:

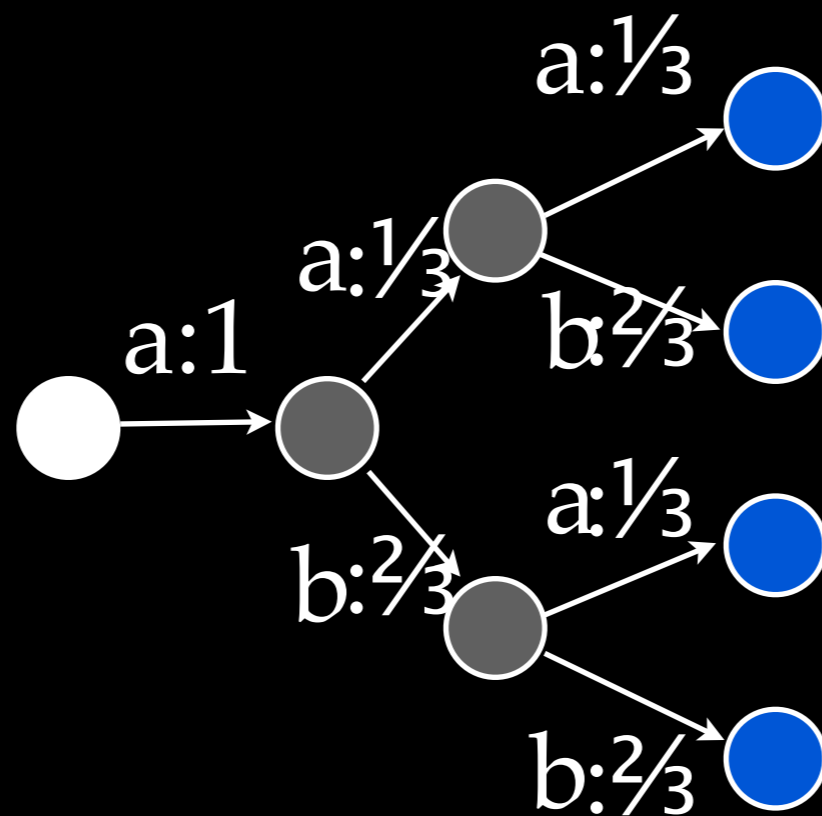
$$f : \mathcal{L} \rightarrow \mathbb{R}^+$$



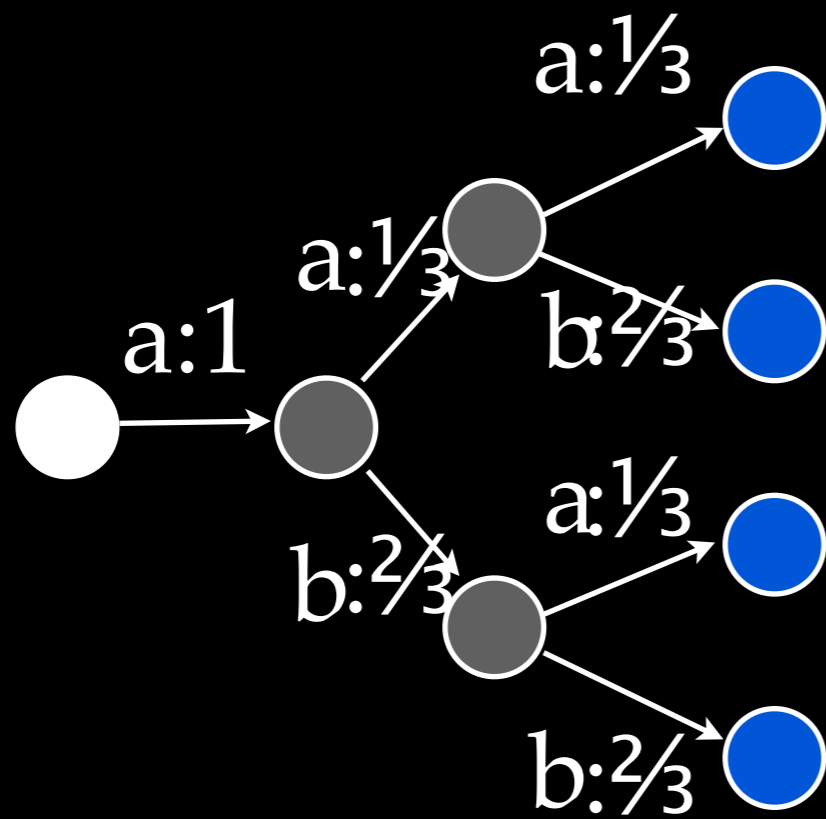
Probabilistic Regular Languages

We want a function:

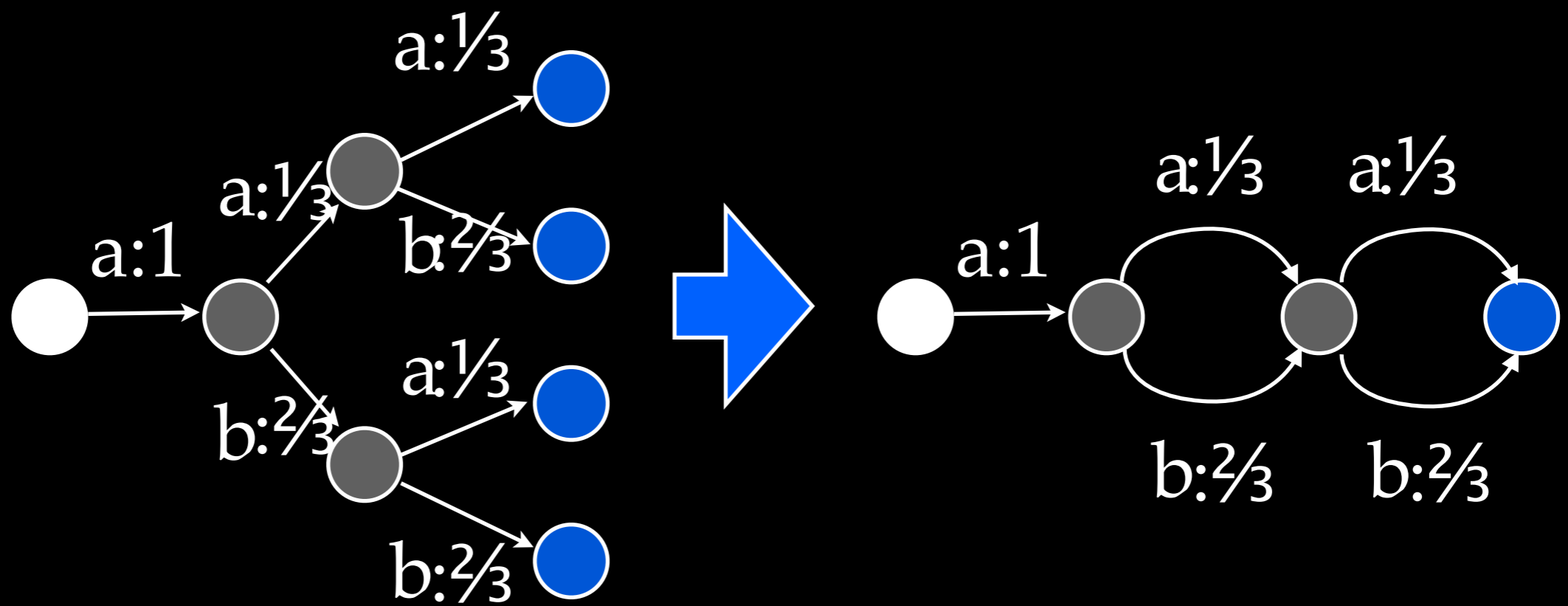
$$f : \mathcal{L} \rightarrow \mathbb{R}^+$$



Minimization



Minimization

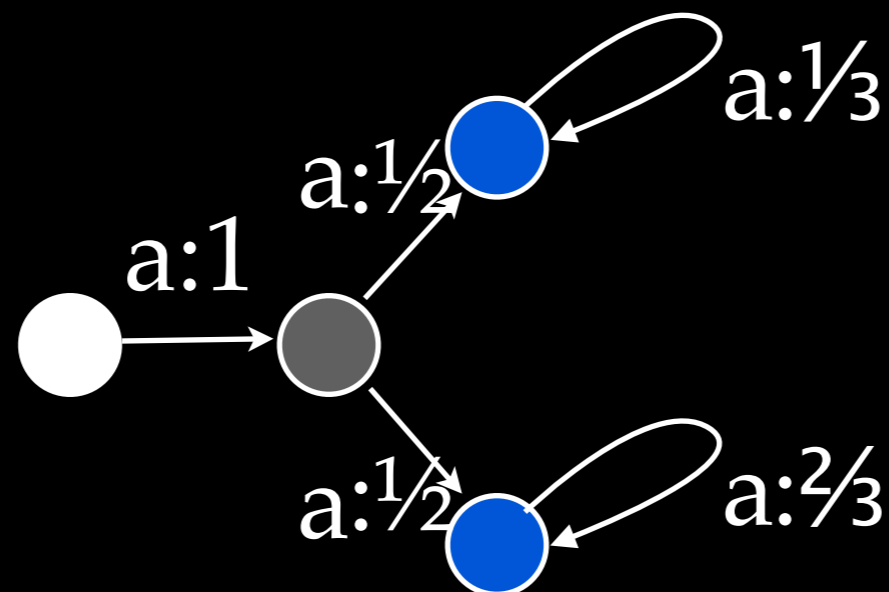


Other Algorithms

- Shortest path (e.g. Dijkstra, A^*): most probable
- Determinization (not all can be determinized)
- Epsilon-removal
- Lazy composition (e.g. intersection): $p(e)p(f|e)$

Other Algorithms

- Shortest path (e.g. Dijkstra, A^*): most probable
- Determinization (not all can be determinized)
- Epsilon-removal
- Lazy composition (e.g. intersection): $p(e)p(f|e)$



Practical Issues

- OpenFST (openfst.org)
 - Efficient C++ implementation.
 - Used in speech recognition (Google, Kaldi @ JHU)
 - Machine translation (JHU → Cambridge, Google)

Some Models of Translation

- IBM Models 1-5
- Hidden Markov Model
- Phrase-Based Models

Q: What do all of these things have in common?

A: They all define *weighted regular languages* over a set of output sentences. Details Thursday.

Questions

Questions

Are natural languages regular?

Questions

Are natural languages regular?

Does it matter for MT if they aren't?