Probability Distributions on Structured Objects

Bhiksha Raj

7 Feb 2018

Probability Outline

- Why probability?
- Probability review
- Multinomials vs. exponential parameterization
- Locally vs. globally normalized models & partition functions
- Examples

Defining Probability

- What is “probability”?

Defining Probability

- You roll a six-sided dice. What is the probability that you will get a 6?
  - Why?
- What is the probability that you will get a A in this course?
  - Why?
- Winter temperatures in Pittsburgh have fallen below 0oF in 143 of the past 1000 years. What

Defining Probability

- Probability of 6 in the roll of a six-sided dice
  - *Classical Definition:* Ratio of number of "favorable" *outcomes* to total outcomes
- Probability that you will get a A in this course
  - Belief
- Winter temps in Pitt have hit < 0oF in 143 of the past 1000 years. Probability that it will hit < 0oF this year

Defining Probability

- A *numerical* way of specifying a belief that a particular *experiment* will have one of a set of *outcomes*
  - The set of outcomes is called an *event*

- The belief may be based on a variety of criteria
  - Total number of outcomes
  - Pure belief
  - Past experience

Defining Probability

- What is "probability"?
- No real meaning
- Best understood as a *measure* computed over a set
- But what is in this set?
  - "Outcomes"…

Definitions

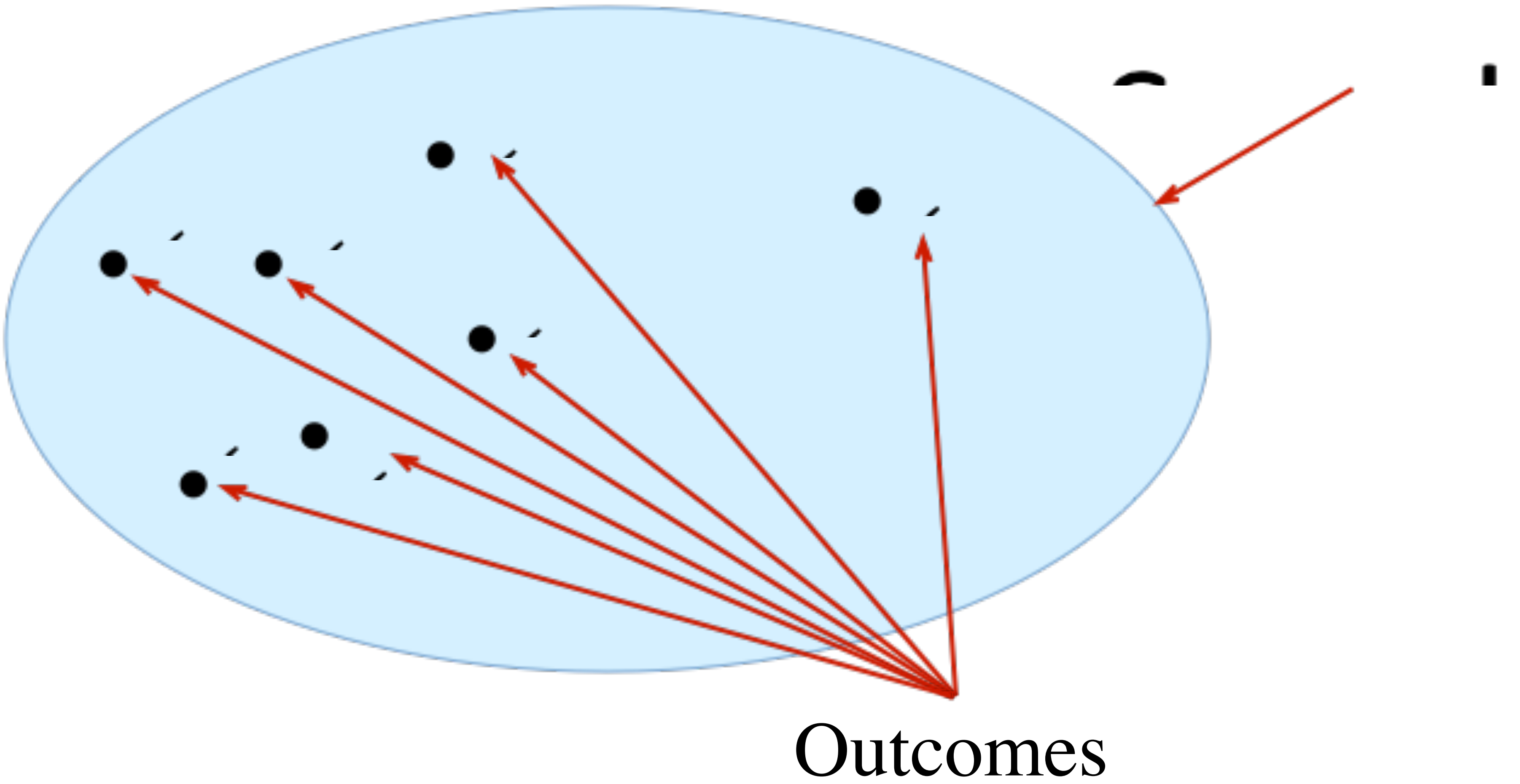
- **Experiment**: A single run of the process we are trying to characterize
  - E.g. Toss of a coin
  - E.g. Roll of a dice
  - E.g. Producing a sequence of words
  - E.g. Car driving down Forbes Ave

- **Outcome**: A result from this process
  - Heads vs. tails
  - Outcome 1 through 6

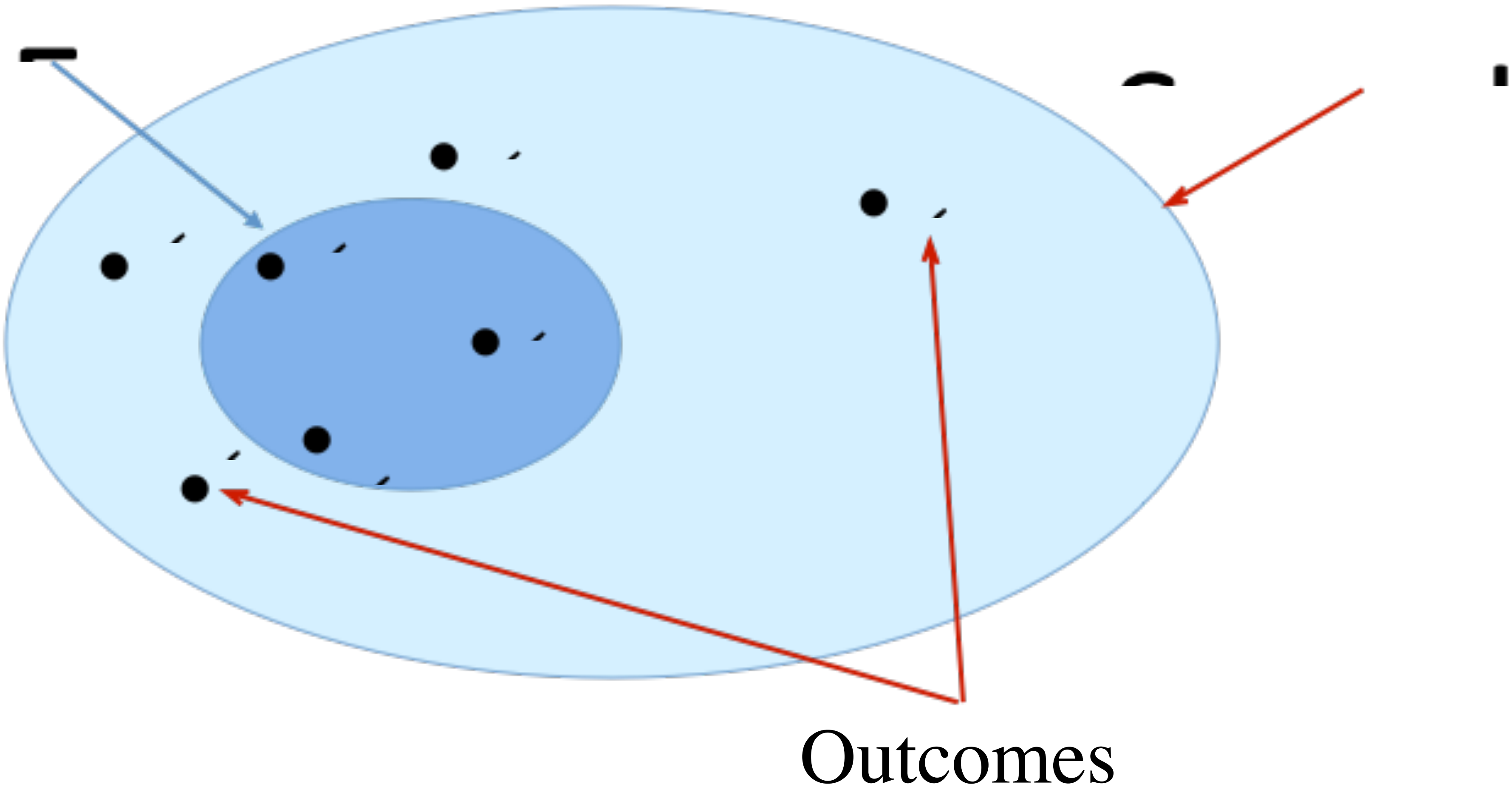
Outcomes and Events

- **Outcome**: A single result of
  - Typically represented by an
  - Outcomes must be
    - Mutually exclusive (any part
      not happen)
    - Collectively exhaustive: The
      specified as one of the outc

Outcomes, Events, and Sample Space



Outcomes, Events, and Sample Space



Axiomatic definition of probability

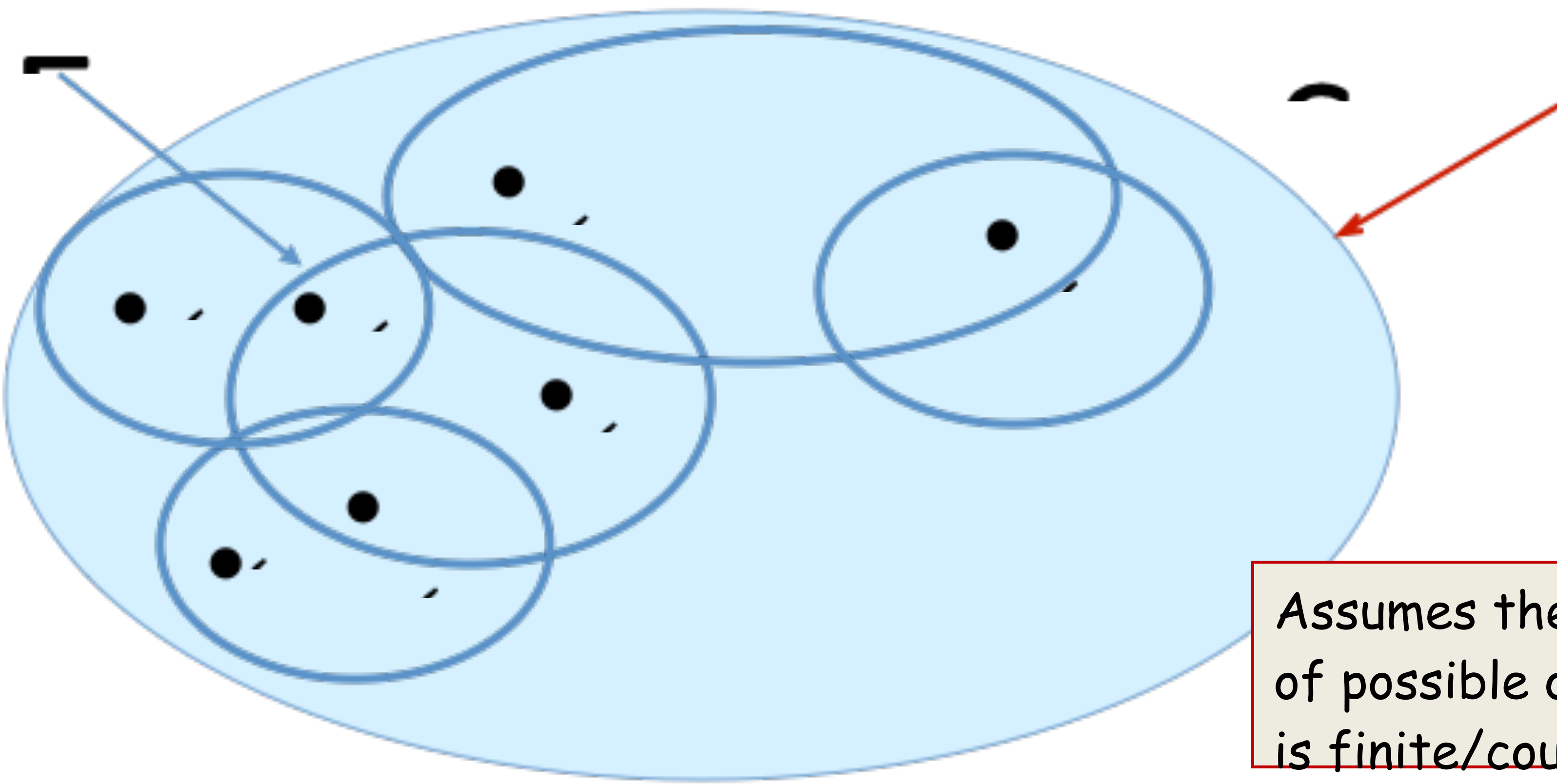
- From Kolmogorov..
- Probability is a *measure* over tl
  following properties
  1. The probability of an event i

     $$\forall E \;\; P(E) \in$$

  2. The probability of the entire

     $$P(\Omega)$$

Outcomes, Events, and Sample Space



Assumes the number of possible outcomes is finite/countable

- “Discrete” sample spaces: Number of ways in which we can define events is *finite* or *countable*.

Definition: Probability *distribution*

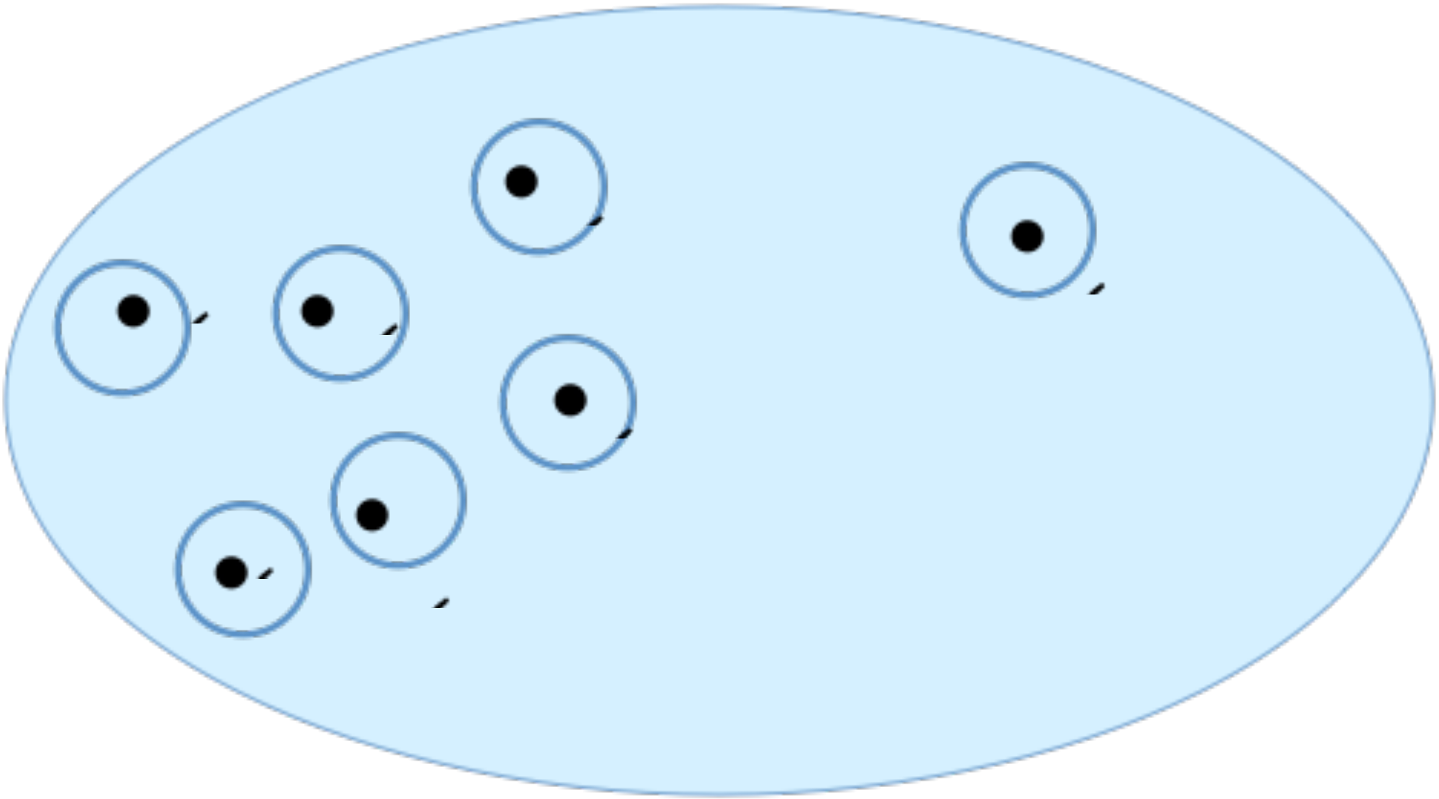
-
- Let $E_1, E_2, \ldots$ be a set of events
  - The events are disjoint

$$E_i \cap E_j = \phi(nu$$

  - The events cover the sample s

$$\bigcup_i E_i$$

Outcomes, Events, and Sample Space



- • Defining individual outcomes as e
  - $E_i = \{\omega_i\}$

Notation (don’t blame me)

-
  - Introducing some (ba
  - For $E_i = \{\omega_i\}$, notat
    elementary events a

$$f(\omega_i)$$

Probabilities over outcomes

$$\forall \omega \in \Omega, \quad f(\omega) \in [0, 1]$$

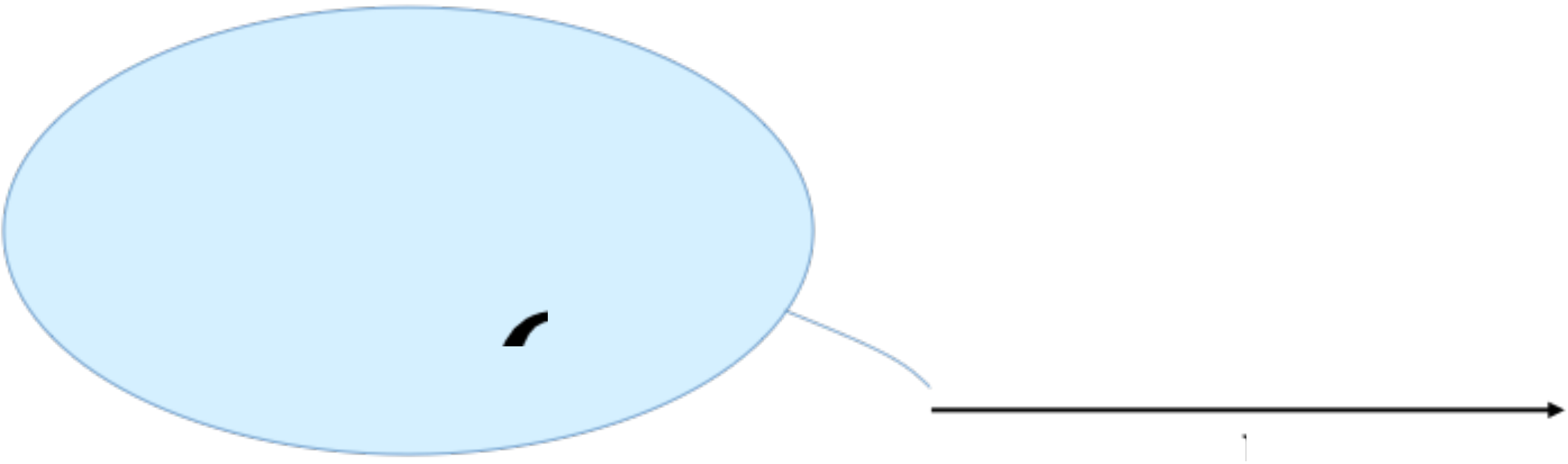
$$\sum f(\omega) = 1$$

Probability mass function

An **event** is a subset (maybe one element) of the sample space, $E \subseteq \Omega$

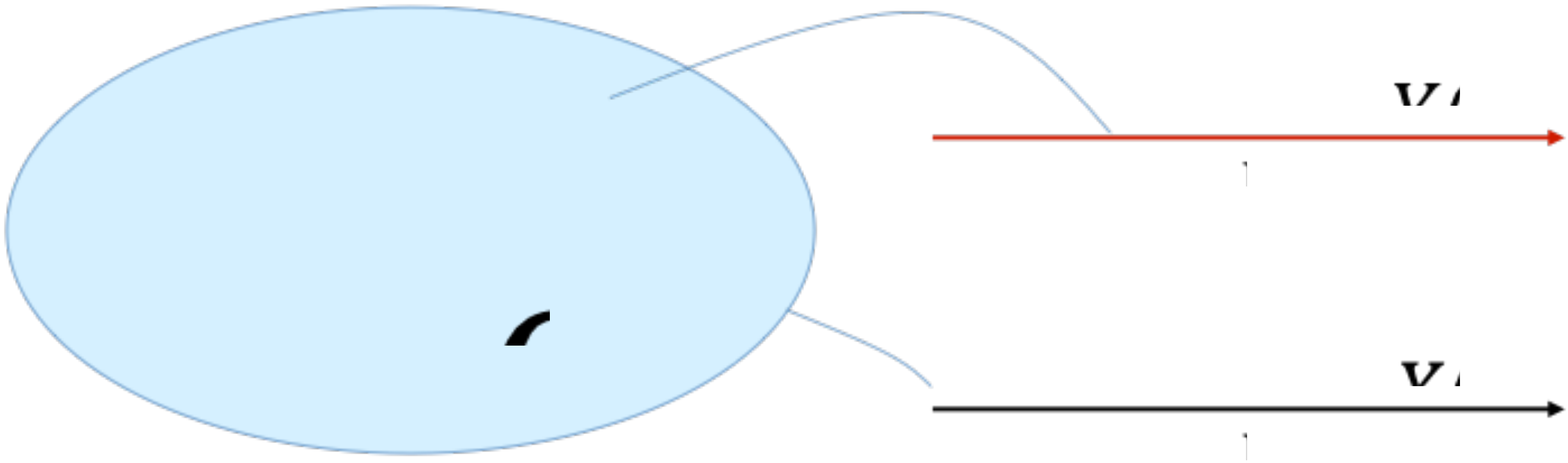
$$P(E) = \sum f(\omega)$$

Random Variable



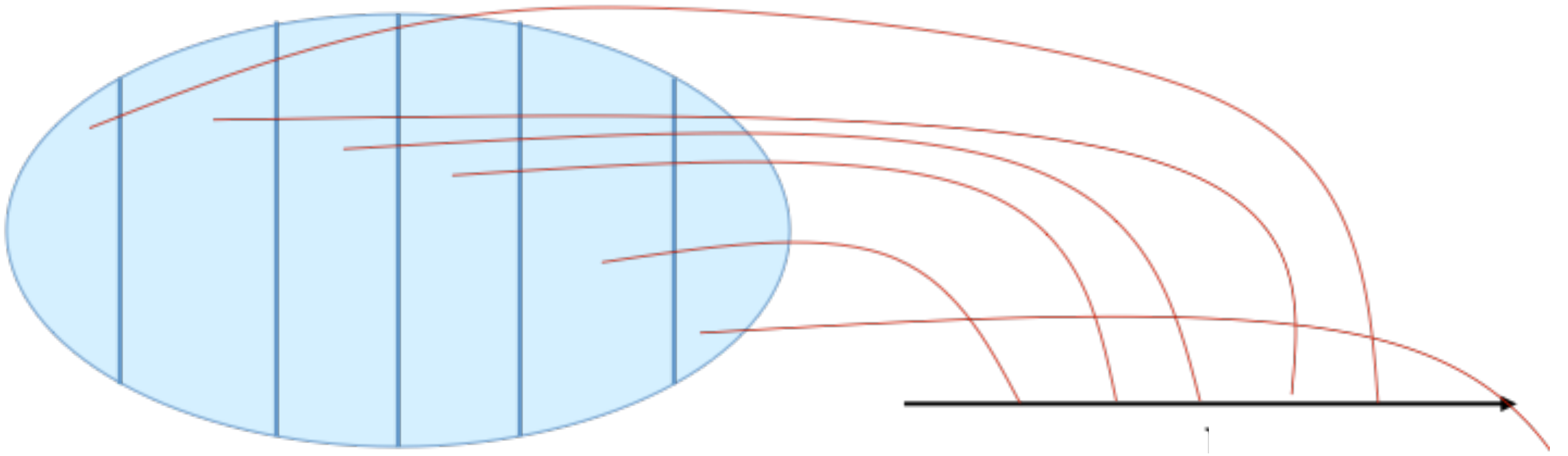
- A random variable is a function that maps the sample space onto the real line
  - Can only use some portion of the real line

Random Variable



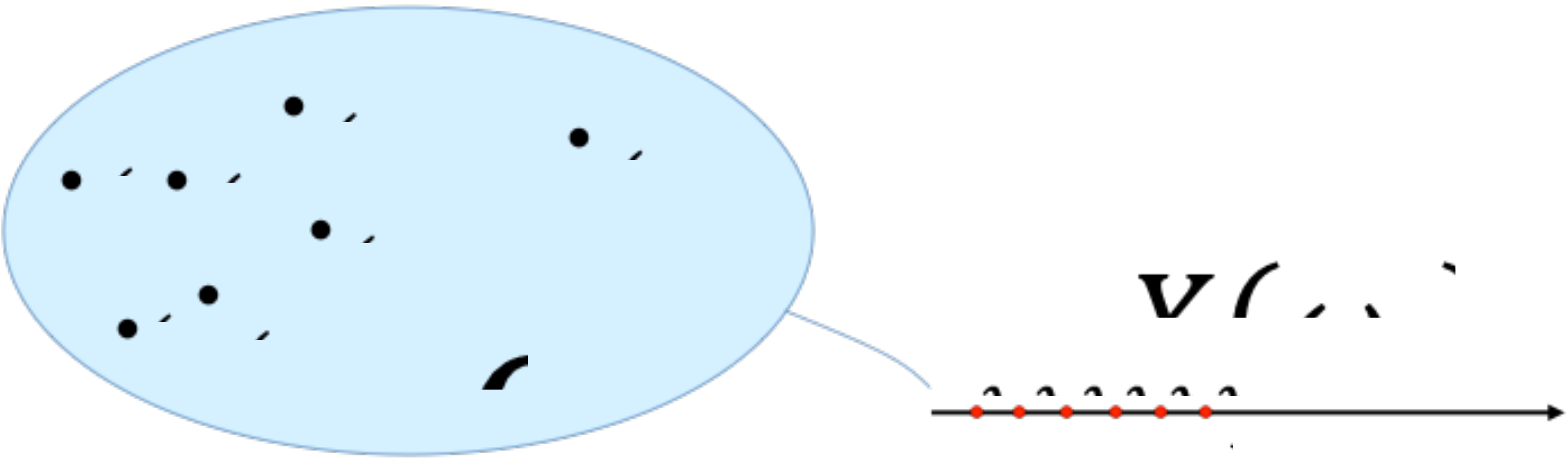
- A random variable is a function
  the real line
  - Can only use some portion of

Random Variable



- A random variable that maps the sample space onto a discrete set of points on the real line is called a *discrete Random Variable*
  - You can compose the discrete RVs even if the sample space is not discrete!!

Random Variable



- For a discrete sample space, the RV must necessarily be discrete
  - But a discrete RV does not imply a discrete sample space

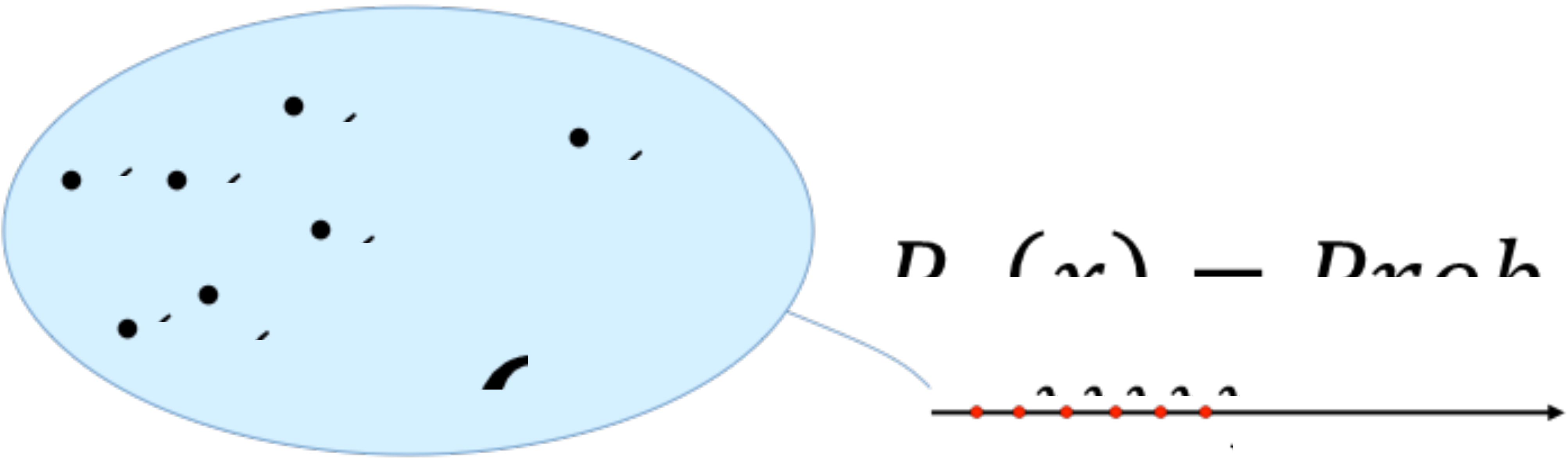
Discrete Random Variable

{0,1}

{1,2,3,4,5,6}

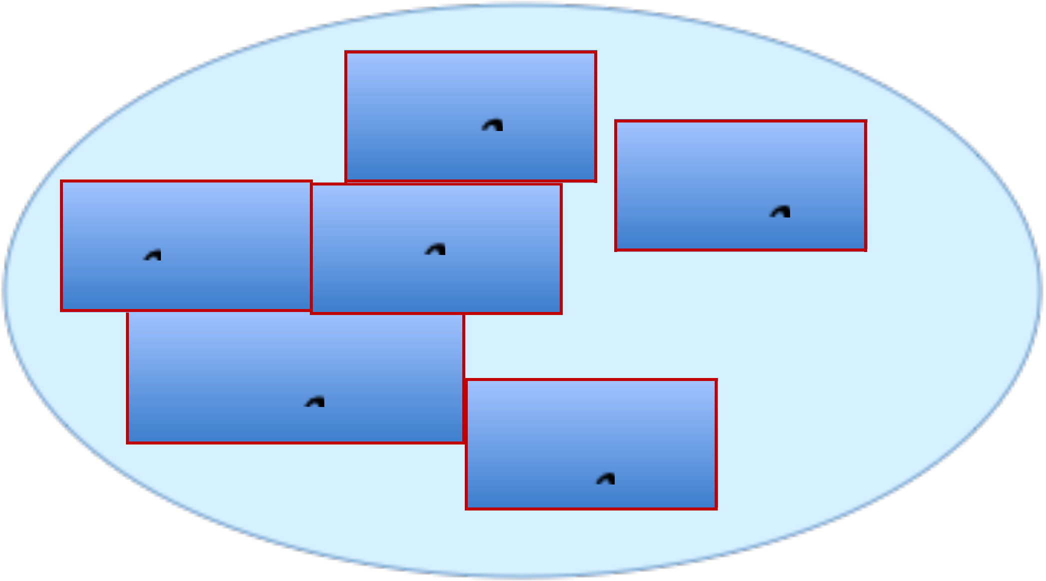
- For a discrete sample space, the RV must necessarily be discrete
  - But a discrete RV does not imply a discrete sample space

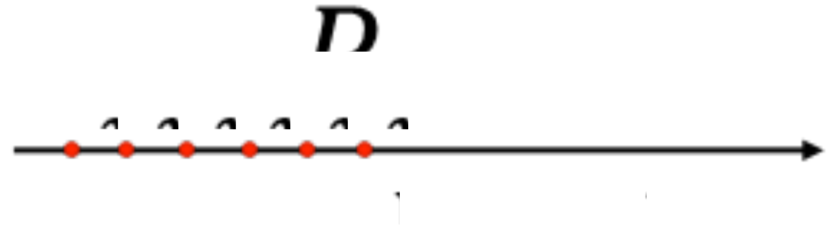
Notation



- $P_X(x)$ is the probabili
  variable X takes the v

Notation





$P_Y(y) = Prob$

-
- $P_Y(y)$ is the probabili
  variable Y takes the v

Discrete Random Variable

{0,1}

For a "fair" coin

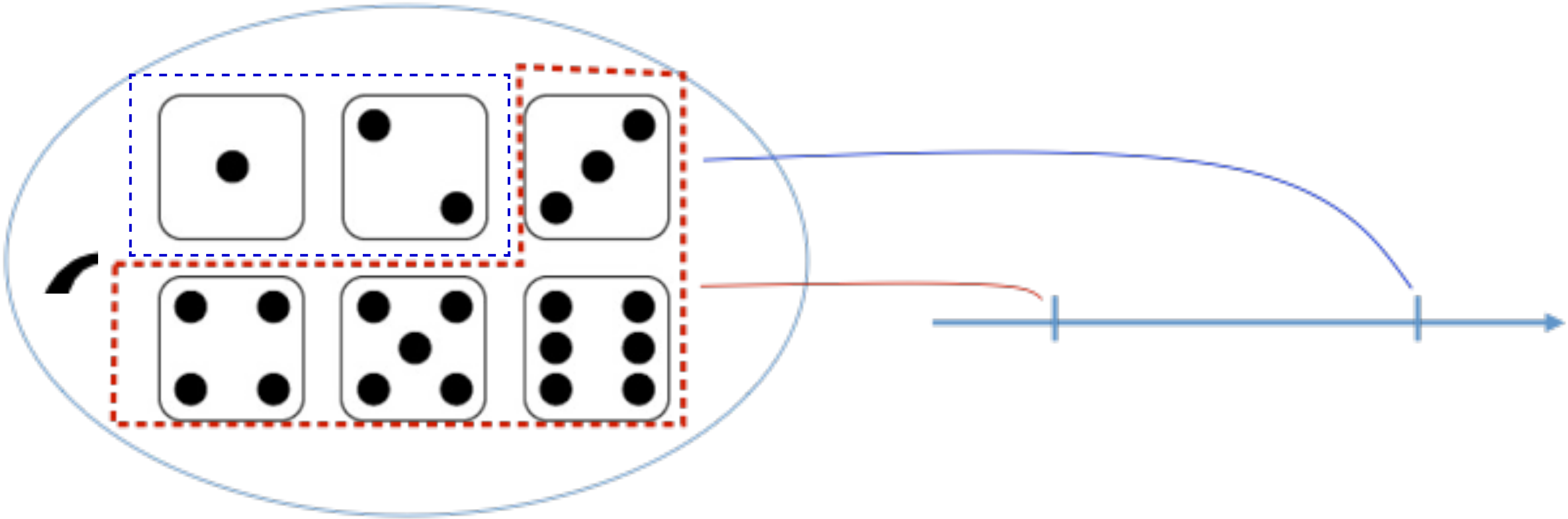
1

{1,2,3,4,5,6}

For a "fair" dice

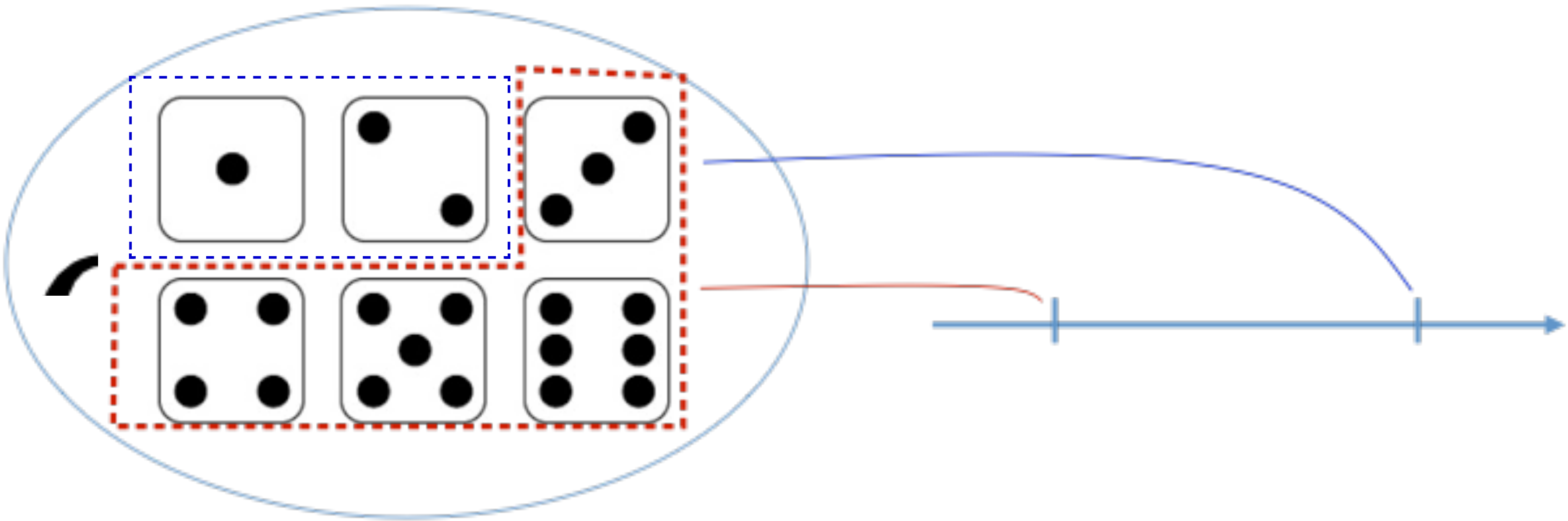
1

A different RV from the dice



For a “fair” dice

A different RV from the dice

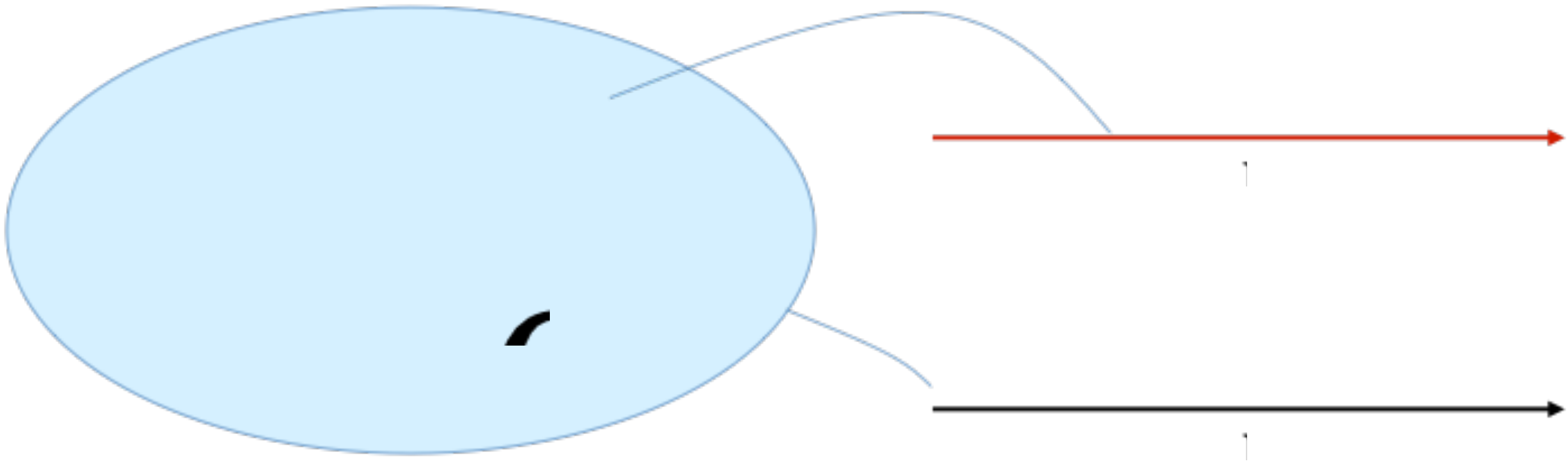


For a "fair" dice

P (0

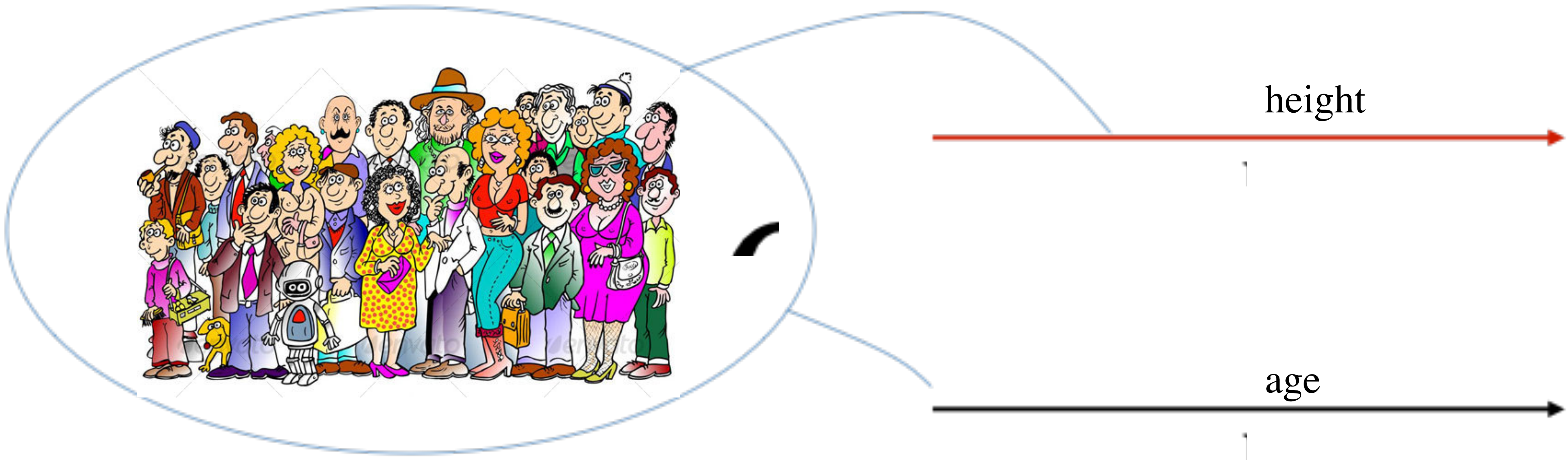
P (1

Joint RVs



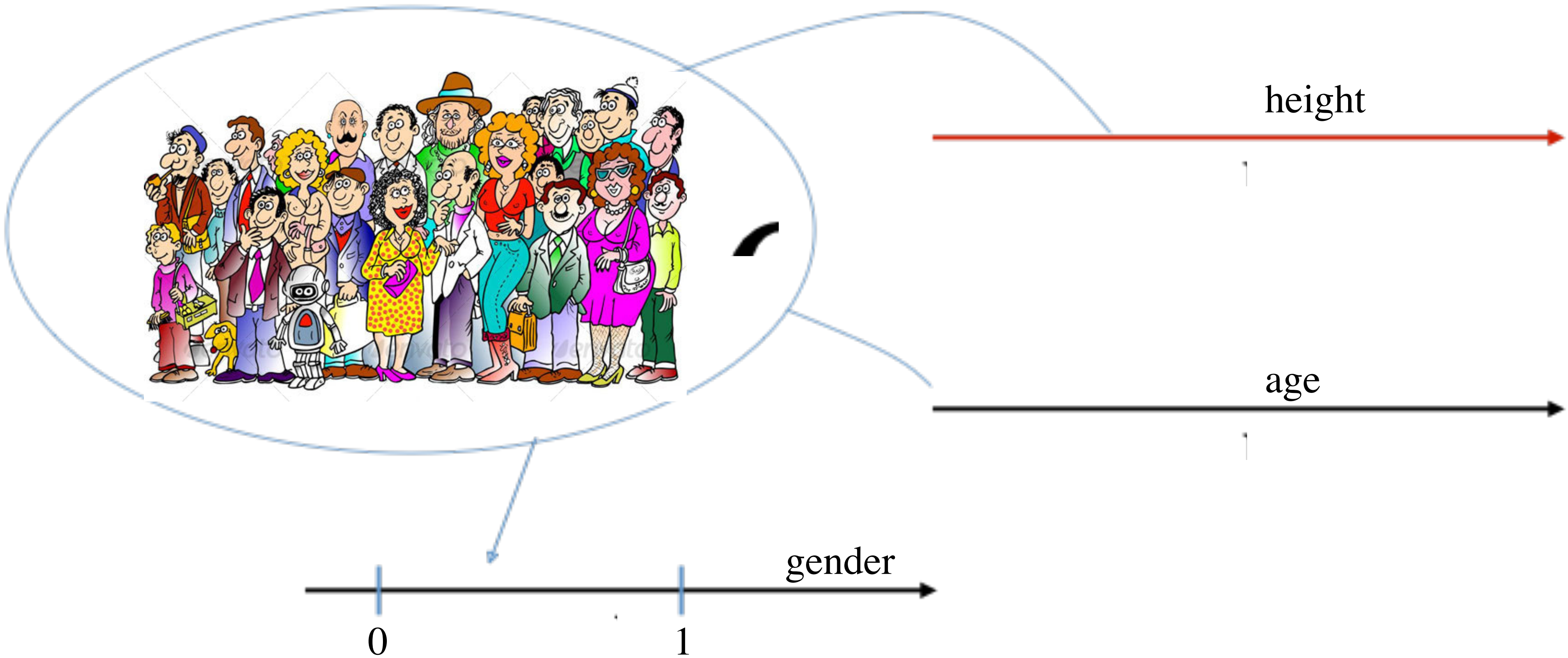
- When we produce *multiple* RVs from the same sample space..

Joint RVs



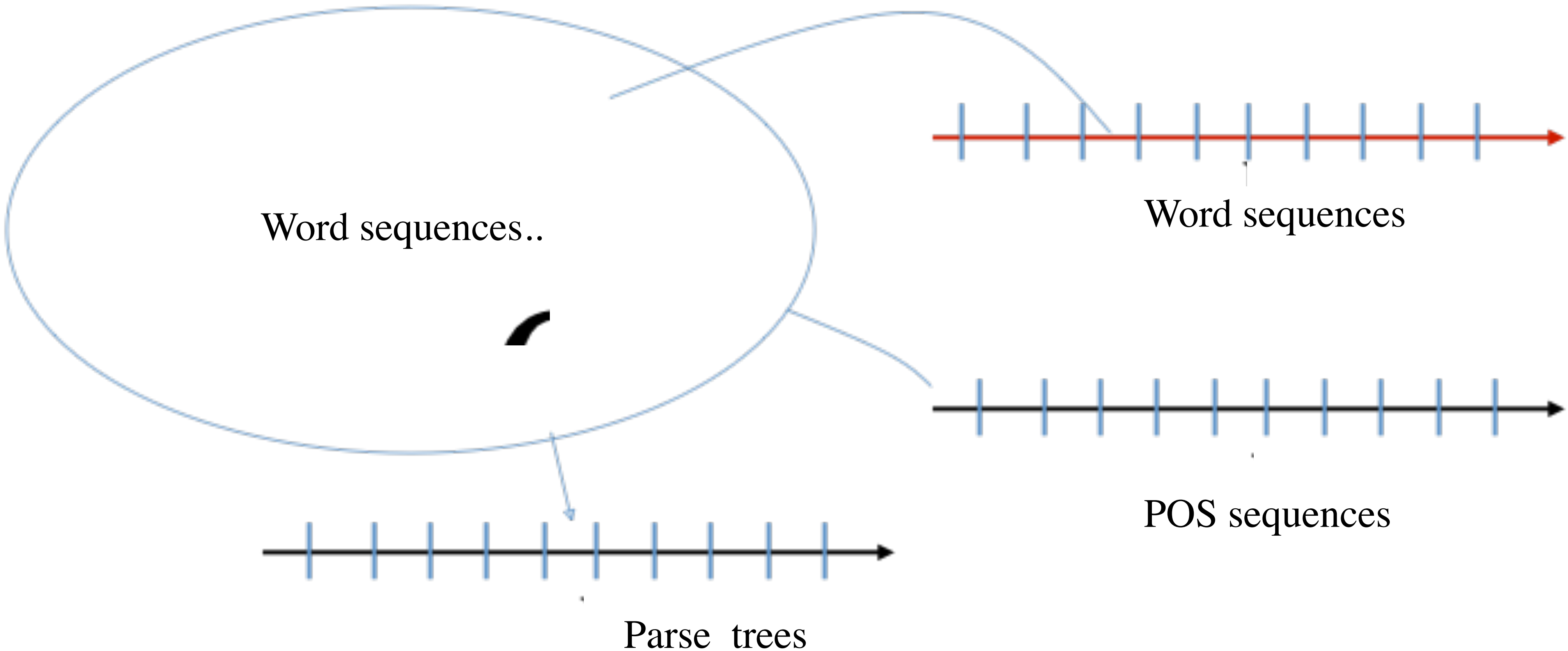
- When we produce *multiple* RVs from the same sample space..

Joint RVs



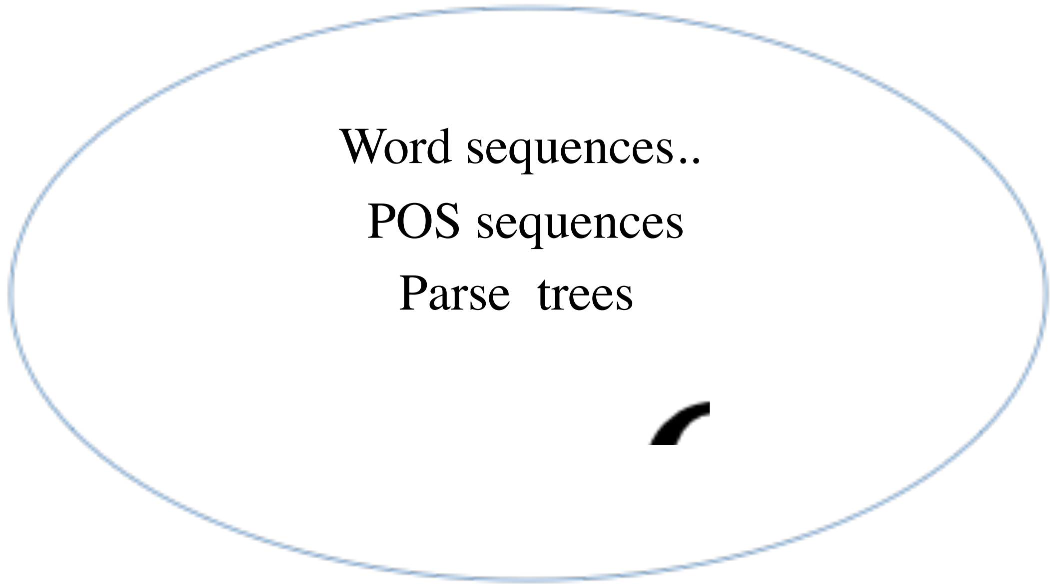
- We can even mix discrete and continuous RVs

Joint RVs



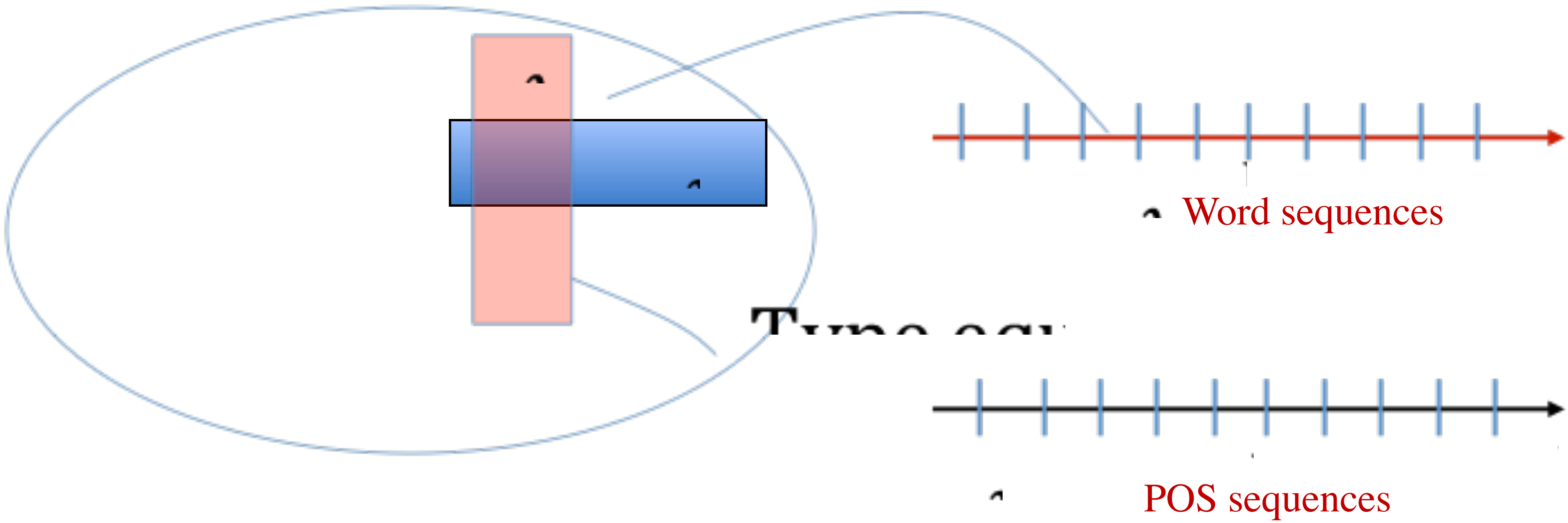
- When we produce *multiple* RVs from the same sample space..

No need for RVs..



- For discrete sample spaces we will often dispense with the entire business of RVs and deal directly with events in the sample space
  - Each RV is just a different way of creating a cover of events over the sample space

Joint RVs



- Each value of each (discrete) RV re
  - Each RV represents a different "d
- The *joint* RV is combination of eve

Joint Probability

- Probability over multiple event types
- Tool for reasoning about dependent (correlated) events

A **joint probability distribution** is a probability distribution over joint r.v.'s with the following form:

$$Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix}$$

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = 1 \qquad \rho_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) \geq 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

Joint Probability

- Probability over multiple event types
- Tool for reasoning about dependent (correlated) events

A **joint probability distribution** is a probability distribution over joint r.v.'s with the following form:

$$Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix}$$

Words

Tag
s

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = 1 \qquad \rho_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) \geq 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

Joint Probability

- Probability over multiple event types
- Tool for reasoning about dependent (correlated) events

A **joint probability distribution** is a probability distribution over r.v.'s with the following form:

$$Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} \begin{array}{l} \leftarrow \text{Words} \\ \leftarrow \text{Trees} \end{array}$$

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = 1 \qquad \rho_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) \geq 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

Joint Probability

- Probability over multiple event types
- Tool for reasoning about dependent (correlated) events

A **joint probability distribution** is a probability distribution over r.v.'s with the following form:

$$Z = \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix}$$

DNA sequence

Proteins

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \rho_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = 1 \qquad \rho_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) \geq 0 \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

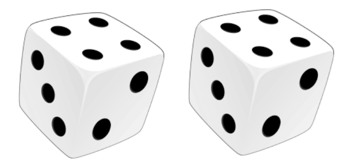
$\Omega = \{1,$

$X(\omega)$

$\Omega = \{1,$

$X(\omega)$

$$\Omega = \begin{matrix} \{(1,1) & (1,2) & (1, \\ (2,1) & (2,2) & (2, \\ (3,1) & (3,2) & (3, \end{matrix}$$

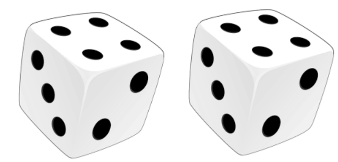


$X(\omega)\ Y(\omega)$

Probability of joint RV for
"fair" dice:

$P \quad (x,$

$$\{(1,1) \quad (1,2) \quad (1,$$
$$(2,1) \quad (2,2) \quad (2,$$
$$(3,1) \quad (3,2) \quad (3,$$



Probability of joint RV for
"fair" dice:

Marginal probability

-
  - Given a joint RV $(X, \text{I}$
    probabilities of the c
    marginal probability

Marginal Probability

$$p(X = x, Y = y) = \rho_{X,Y}(x, y)$$

$$p(X = x) = \sum_{y' \in \mathcal{Y}} p(X = x, Y = y')$$

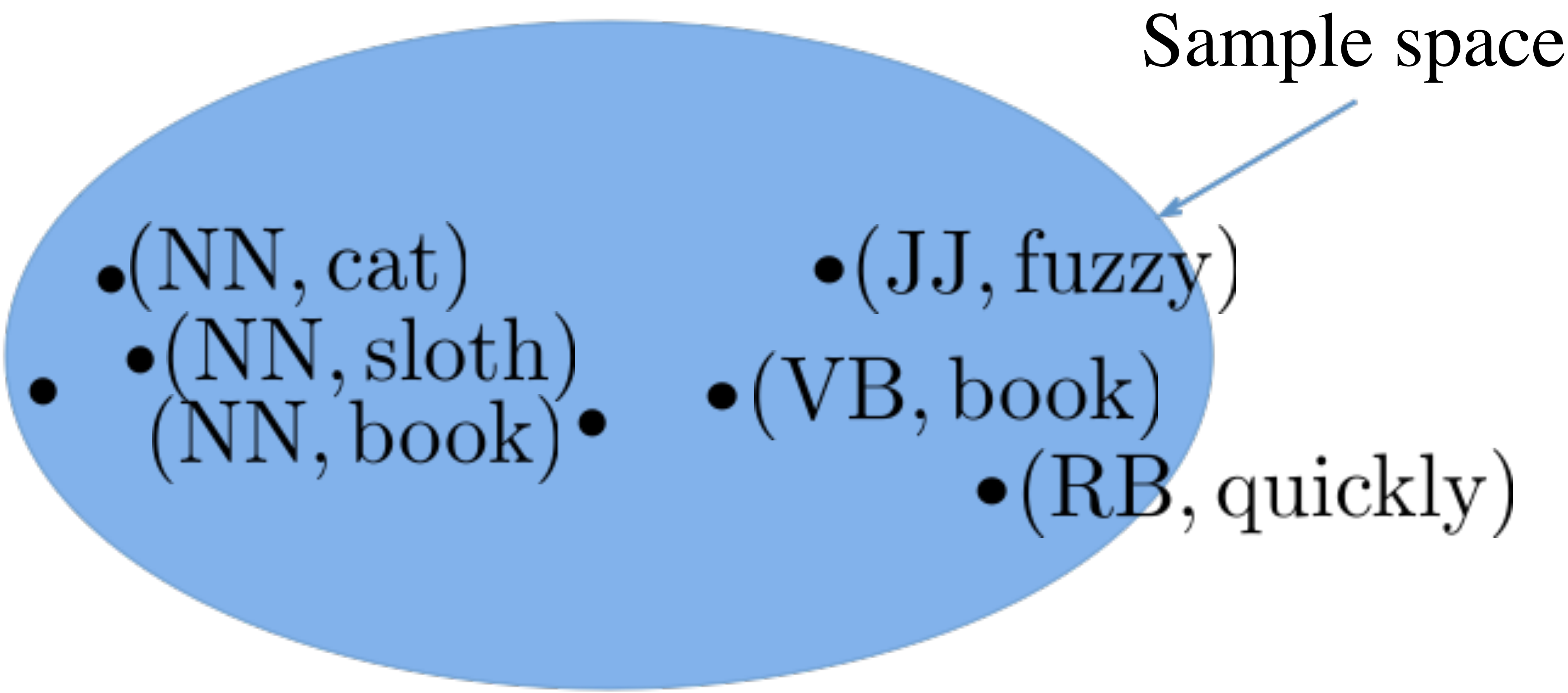
$$p(Y = y) = \sum_{x' \in \mathcal{X}} p(X = x', Y = y)$$

$$\begin{aligned}\Omega = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),\}\end{aligned}$$

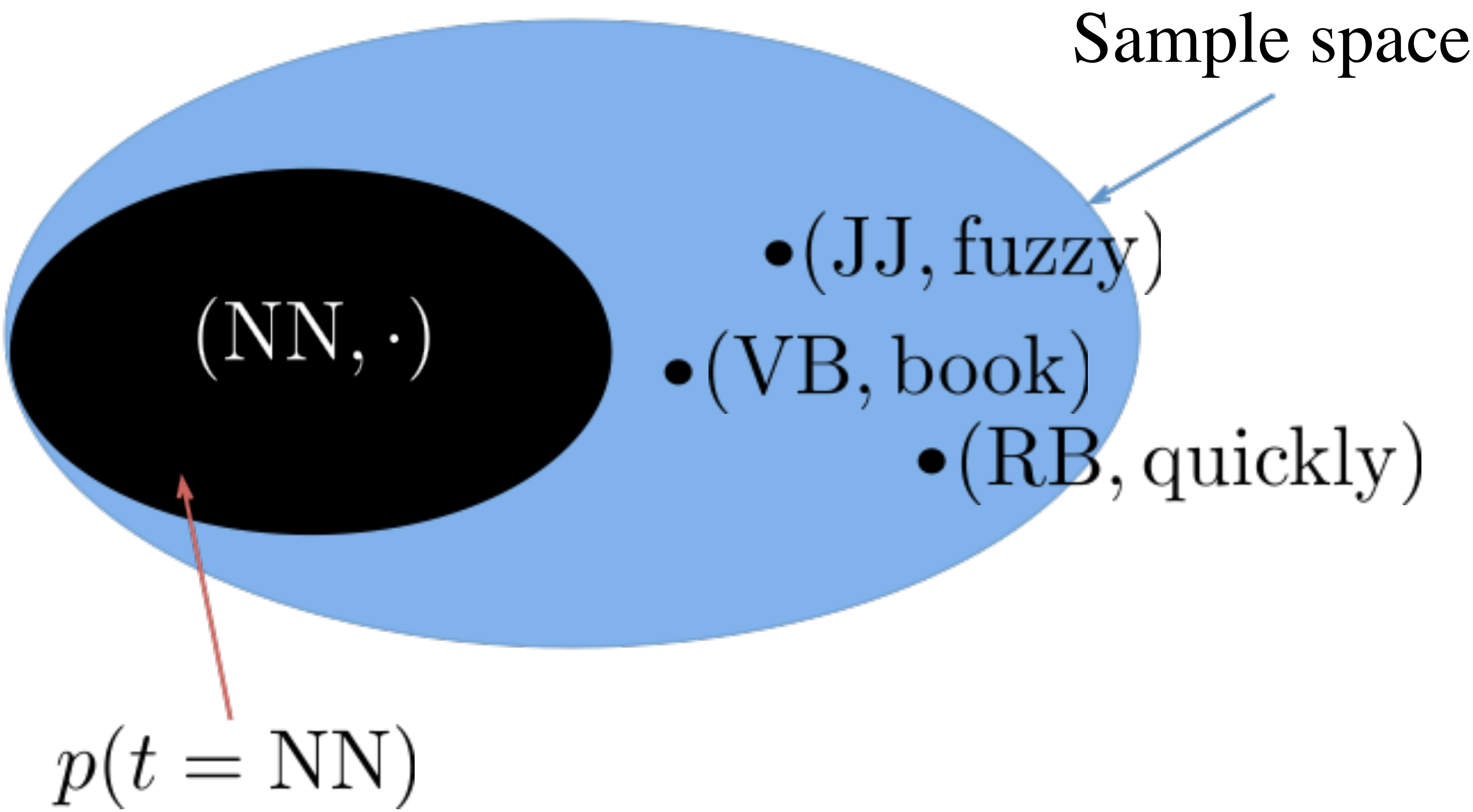
$$p(X = 4) = \sum_{y' \in [1,6]} p(X = 4, Y = y')$$

$$p(Y = 3) = \sum_{x' \in [1,6]} p(X = x', Y = 3)$$

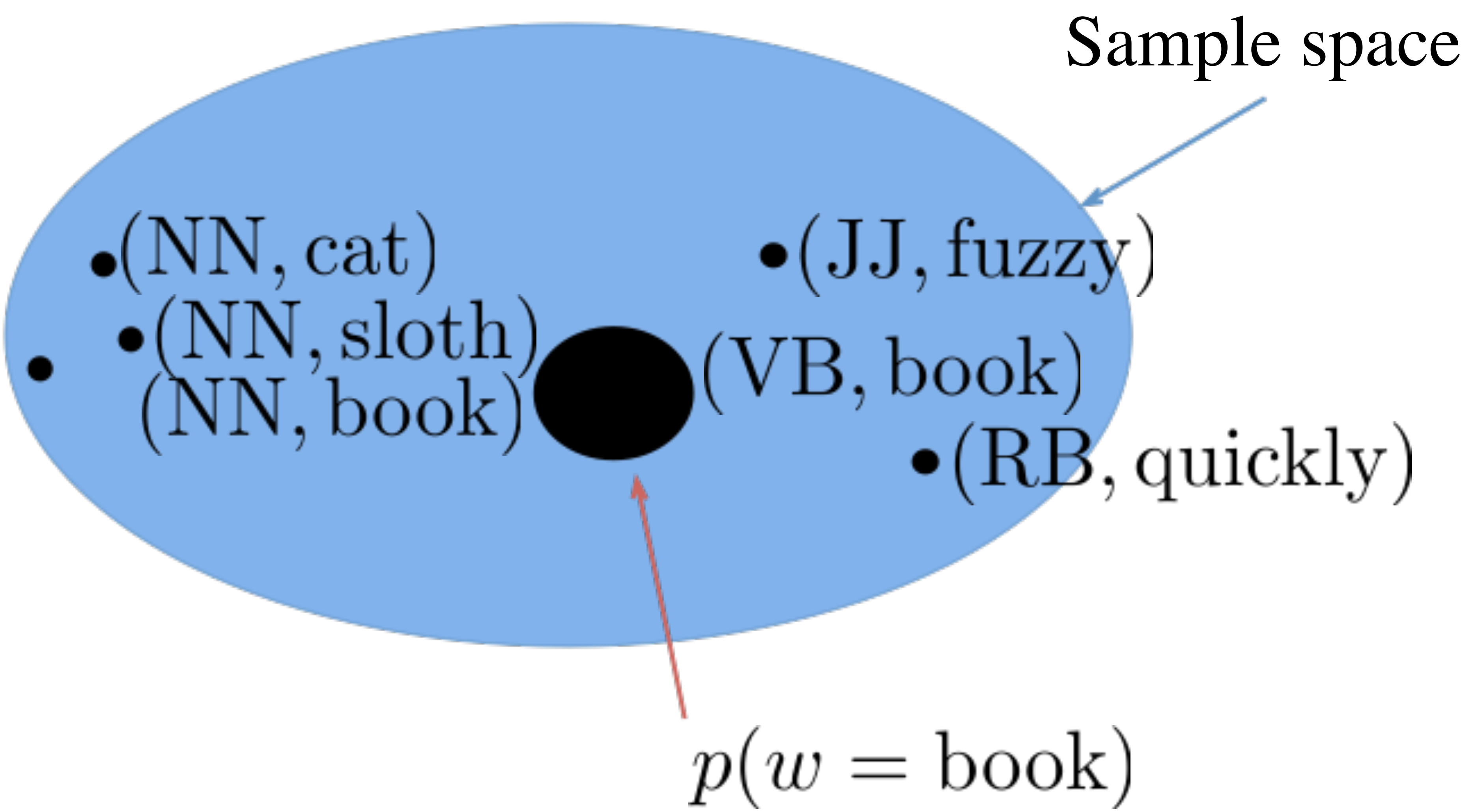
Marginal Probability



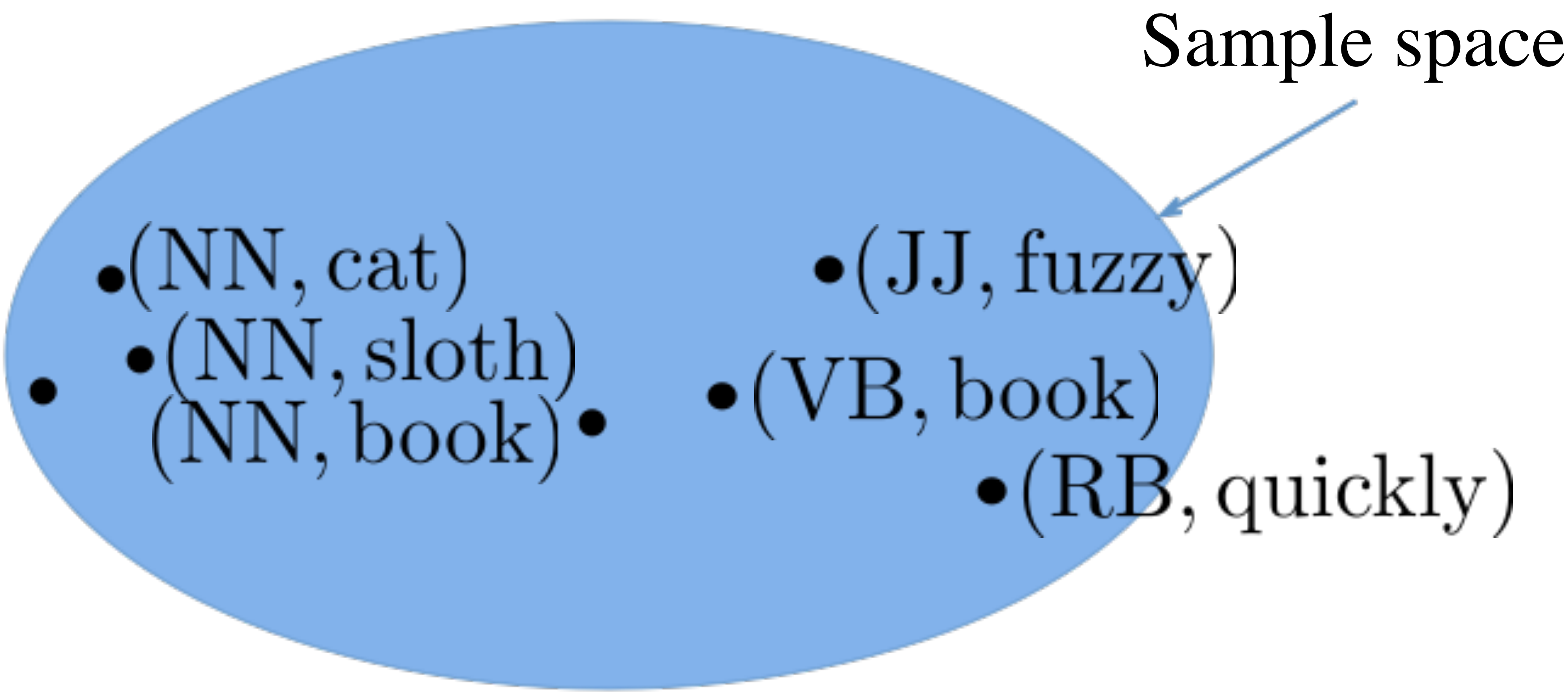
Marginal Probability



Marginal Probability



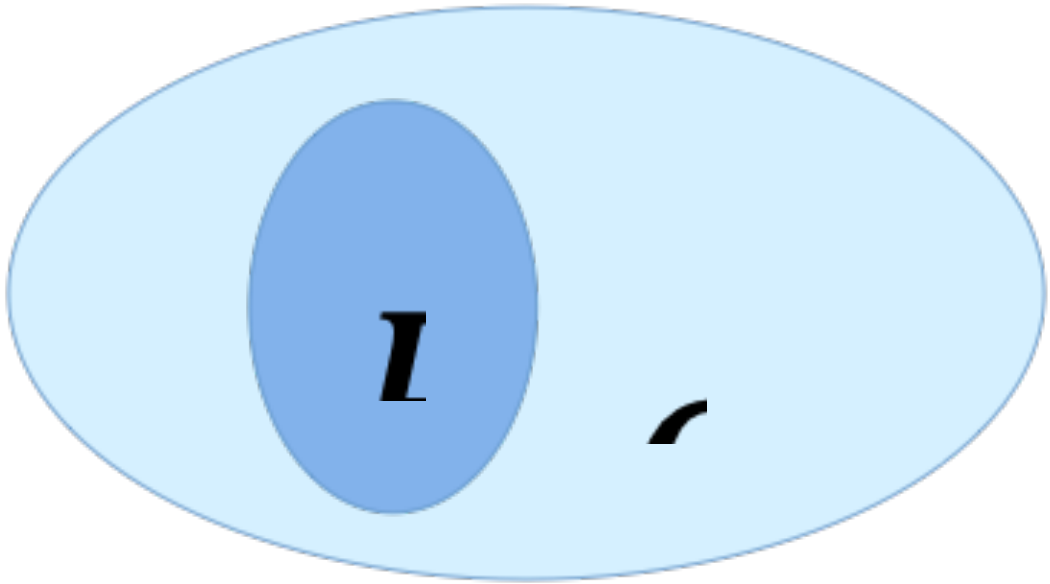
Marginal Probability



Marginal Probabilities

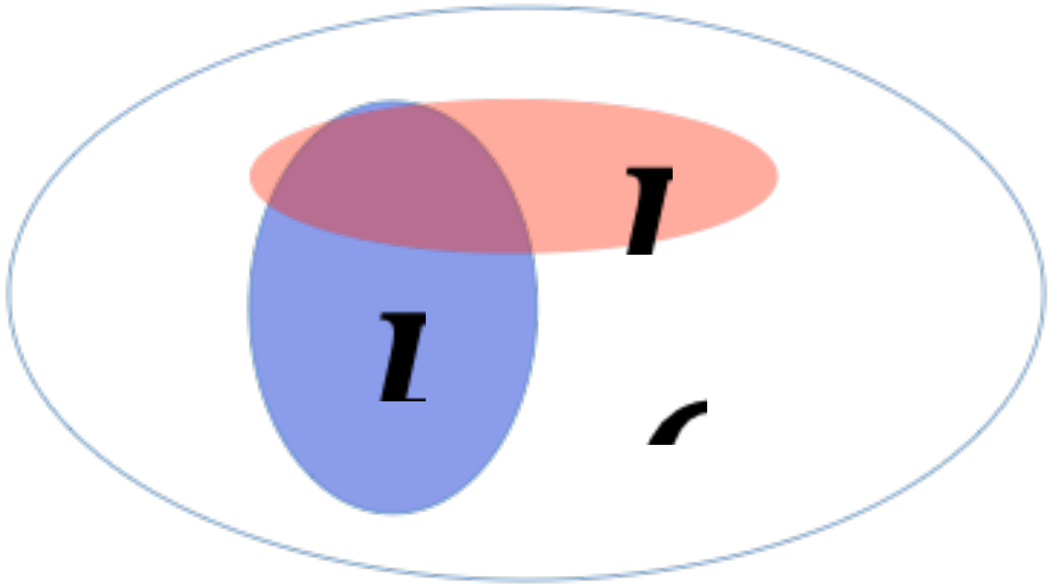
- In a joint model of word and tag sequences p(**w**,**t**)
  - The probability of a word sequence p(**w**)
  - The probability of a tag sequence p(**t**)
  - The probability of a word sequence with the word "cat" somewhere in it
  - The probability of a tag sequence containing three verbs in a row

Conditional probability



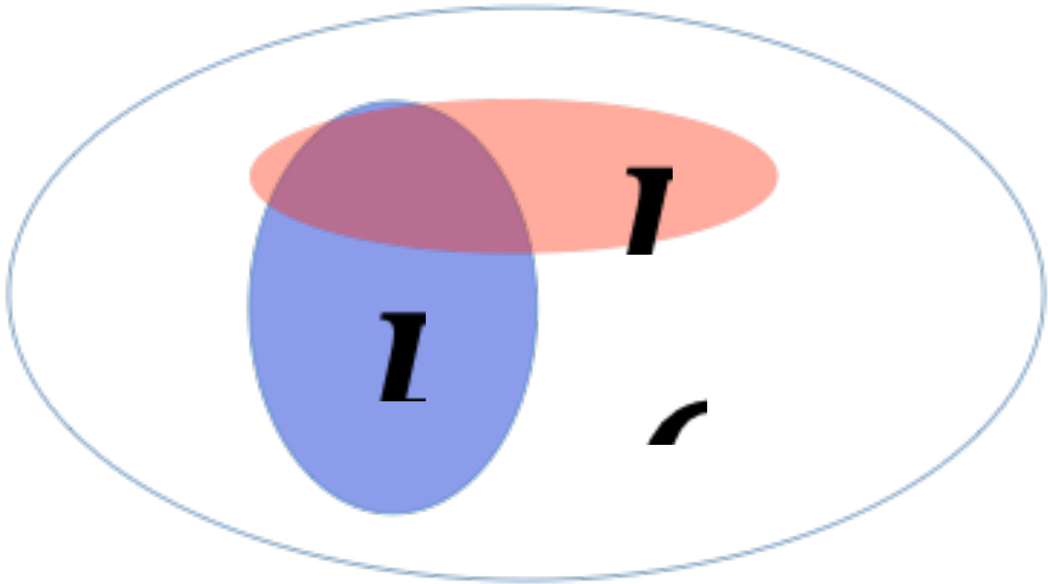
- *Conditioning* events are events th
  - They represent isolated worlds th sample space
  - The new sample space is the con

Conditional probability of events



-
- $P(R|E)$ = the probabi
  scaled by the inverse

Conditional probability of events



•

$$P(R|E)P(E$$

$$P_{X,Y}(x,y) =$$

Conditional Probability

The **conditional probability** is defined as follows:

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{\text{joint probability}}{\text{marginal}}$$

This assumes $p(Y = y) \neq 0$

We can construct joint probability distributions out of conditional distributions:

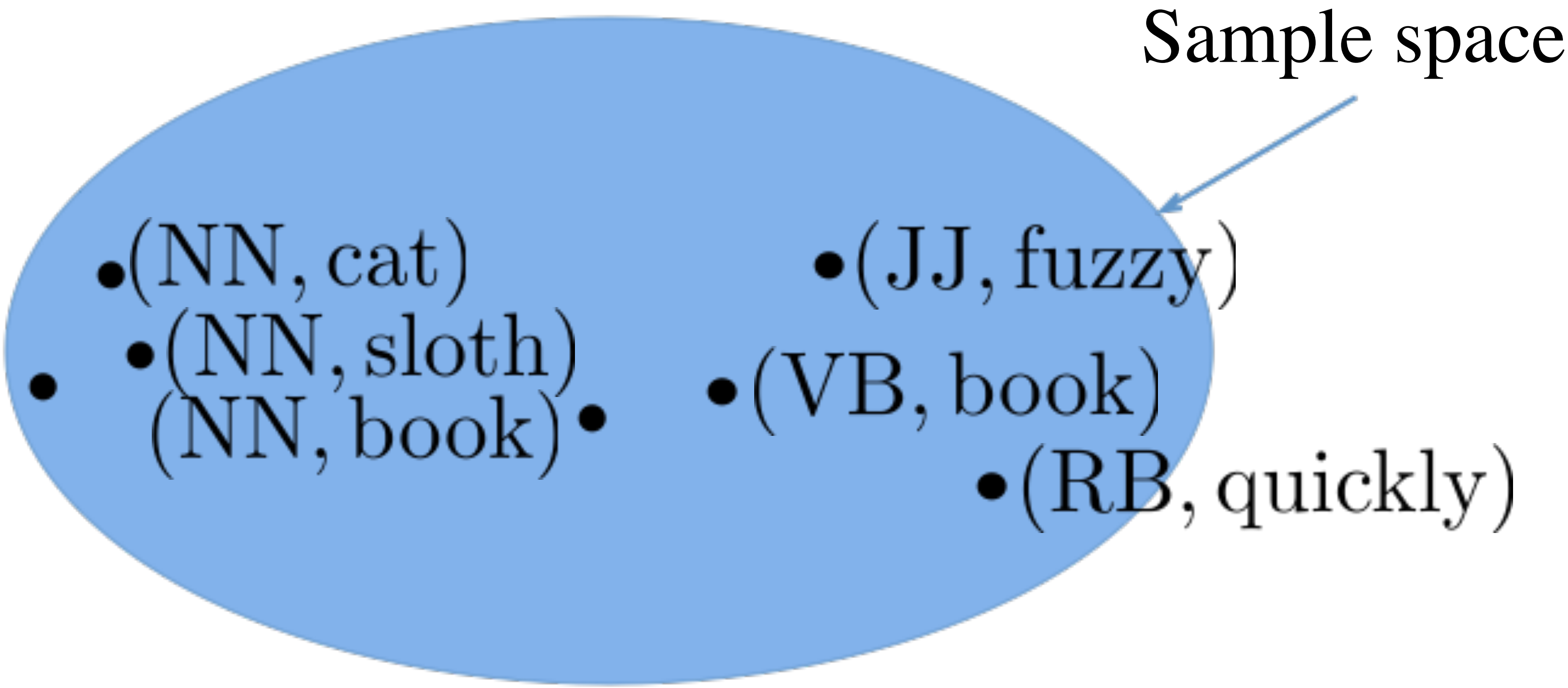
$$p(x \mid y)p(y) = p(x, y) = p(y \mid x)p(x)$$

Conditional Probability Distributions

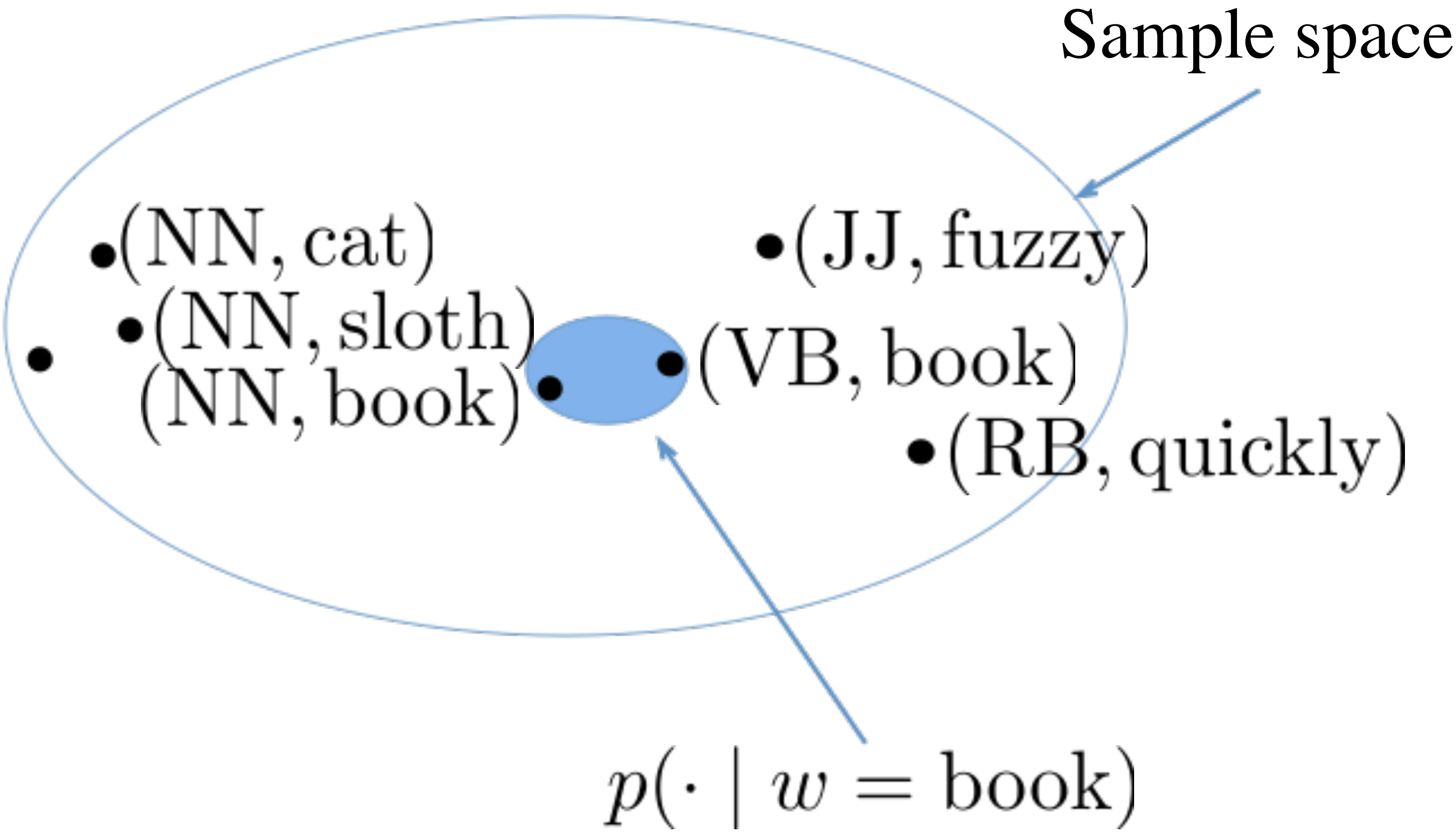
The **conditional probability distribution** of a variable X given a variable Y has the following properties:

$$\forall\ y \in Y,\ \sum_{x \in X} p(X = x \mid Y = y) = 1$$

Conditional Probability



Conditional Probability

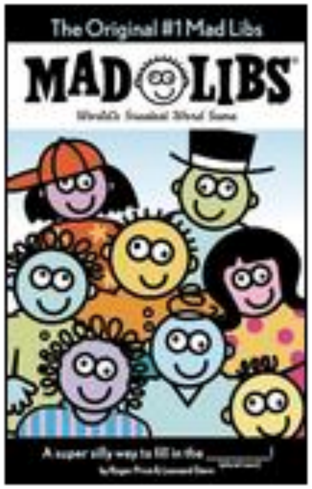


Conditional Probabilities

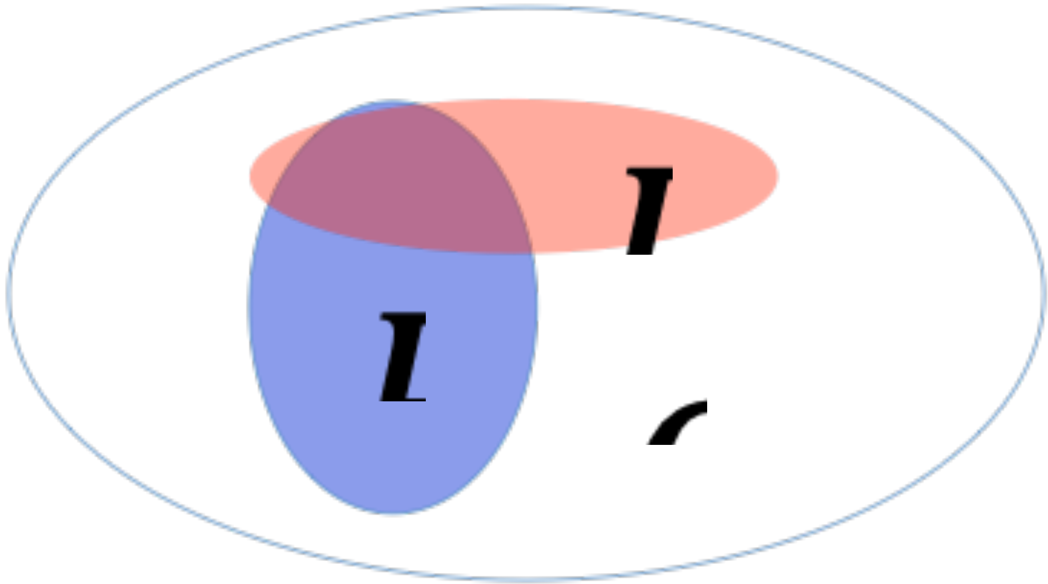
- In a joint model of word and tag sequences p(**w**,**t**)
  - The probability of a tag sequence given a word sequence p(**t** | **w**)
  - The probability of a word sequence given a tag sequence p(**w** | **t**)

Joint and Marginal Probabilities

- In a joint model of word and tag sequences p(**w**,**t**)
  - The probability that the 3rd tag is VERB, given **w** = "Time flies ***like*** an arrow"
    p(t3 = VERB| **w** = Time flies like an arrow)
  - The probability that the 3rd word is ***like***, given **w** = "Time flies ______ an arrow", t3 = VERB
    p(t3 = ***like*** | **w** = Time flies _____ an arrow, t3 = VERB)



Conditional probability of events



R itself may be
a set intersection

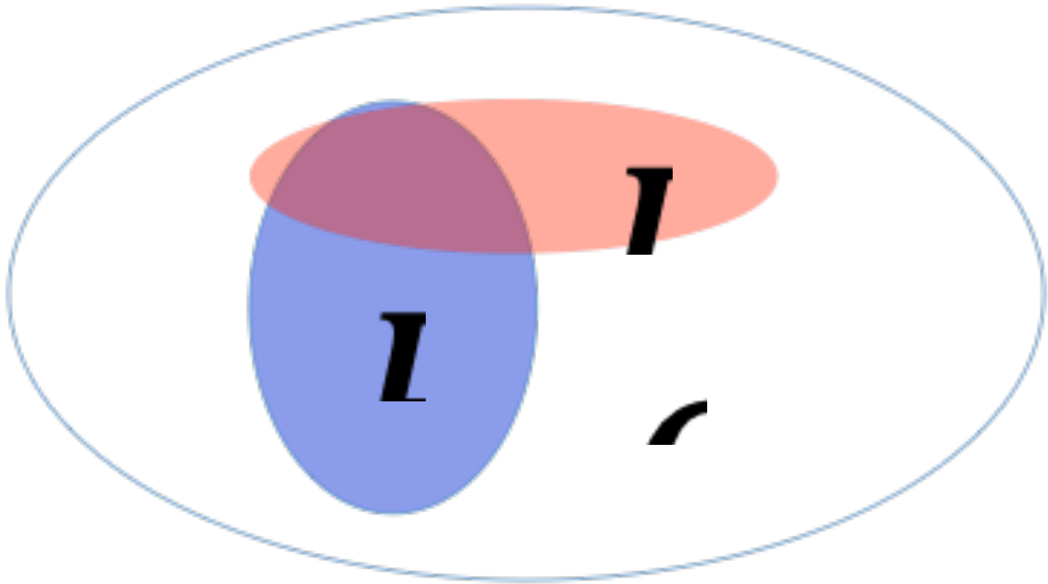
•

$$P(R|E)P(E$$

Chain Rule

$$
\begin{aligned}
p(a, b, c, d, \ldots) = & \; p(a) \times \\
& \; p(b \mid a) \times \\
& \; p(c \mid a, b) \times \\
& \; p(d \mid a, b, c) \times \\
& \; \vdots
\end{aligned}
$$

Conditional probability of events



•

$$P(R|E)P(E) = P(R$$

Bayes Rule

-

$$P(x|y) =$$

$$P(x|y) =$$

Bayes Rule

Posterior Likelihood Prior

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} \qquad \left(= \frac{p(y \mid x)p(x)}{\sum_{x'} p(y \mid x')p(x')} \right)$$

Evidence

Independence

Two r.v.'s are **independent** iff

$$p(X = x, Y = y) = p(X = x) \times p(Y = y)$$

Equivalently (prove with def. of cond. prob.)

$$p(X = x \mid Y = y) = p(X = x)$$

Alternatively,

$$p(Y = y \mid X = x) = p(Y = y)$$

Conditional Independence

Two equivalent statements of conditional independence:

$$p(a, c \mid b) = p(a \mid b)p(c \mid b)$$

and:

$$p(a \mid b, c) = p(a \mid b)$$

"If I know B, then C doesn't tell me about A"

$$p(a \mid b, c) = p(a \mid b)$$

$$\begin{aligned} p(a, b, c) &= p(a \mid b, c)p(b, c) \\ &= p(a \mid b\cancel{, c})p(b \mid c)p(c) \end{aligned}$$

Conditional Independence

Two equivalent statements of conditional independence:

$$p(a, c \mid b) = p(a \mid b)p(c \mid b)$$

and:

$$p(a \mid b, c) = p(a \mid b)$$

"If I know B, then C doesn't tell me about A"

$$p(a \mid b, c) = p(a \mid b)$$

$$\begin{aligned} p(a, b, c) &= p(a \mid b, c)p(b, c) \\ &= p(a \mid b \cancel{, c})p(b \mid c)p(c) \\ &= p(a \mid b)p(b \mid c)p(c) \end{aligned}$$

Conditional Independence

- Useful thing to assume when designing models
  - Limit the variables that influence distributions
  - Classical example: Markov assumption
- Questions
  - Does conditional independence imply marginal independence?
  - Does marginal independence imply conditional independence?

Expected Values

$$\mathbb{E}_{p(X=x)}\left[f(x)\right] \doteq \sum_{x \in \mathcal{X}} p(X=x) \times f(x)$$

Some special expectations:

$$p(X=y) = \mathbb{E}_{p(X=x)}\left[\mathbb{I}_{x=y}\right]$$

$$H(X) = \mathbb{E}_{p(X=x)}[-\log_2 x]$$

Why Probability?

-
  - Regardless of the purely probability provides us a in the state of the world.

  - It also helps us make pre

The **true** *probability distribution of something*

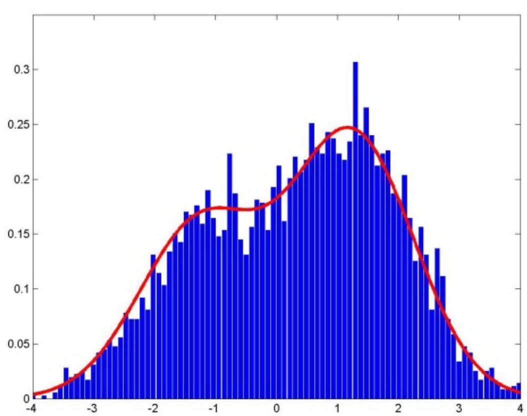
- What on earth is a *true* probability distribution of anything?

- What do we mean by *sampling?*
- *What is generation?*

The notion of a model

- A model for a probability distribution is a distribution that approximates the "true" distribution of an RV according to some metric

- Models are typically parameterized



$$P(X) = \sum_k P(k) N(X; \mu_k, \Theta_k) = \sum_k \frac{P(k)}{\sqrt{(2\pi)^d |\Theta_k|}} \exp\left(-0.5 (X - \mu_k)^T \Theta_k^{-1} (X - \mu_k)\right)$$

The notion of parmetrization

-
  - A parameterization c
    distribution is a set c
    sufficient to comput
    probabilities over th

$$P(X) = \sum_k P(k) N(X; \mu_k, \Theta_k) = \sum_k \frac{P(k)}{\sqrt{(2\pi)^d |\Theta_k|}} \exp\left(-0.5(X-\mu_k)^T \Theta_k^{-1} (X-\mu_k)\right)$$

Sampling Notation

$$y \sim \mathrm{Distribution}(\boldsymbol{\theta})$$

Distribution

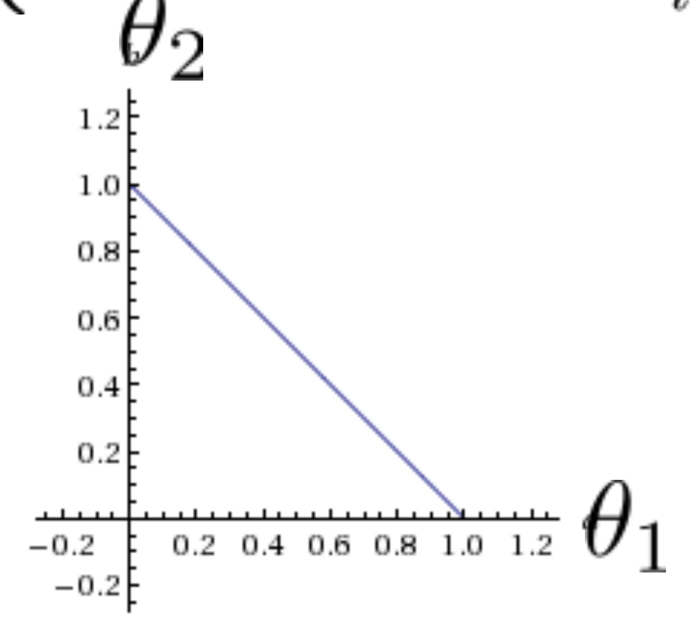
Random variable

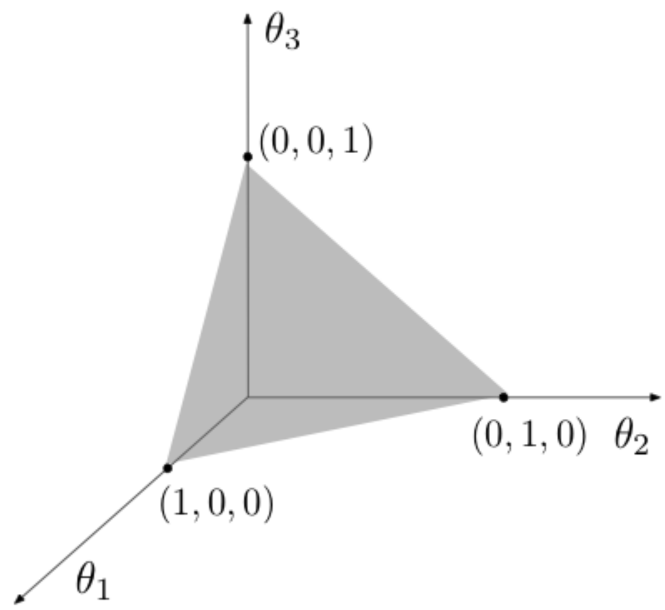
Parameter

Categorical (Multinomial) Distributions

- Generalized model of a die to *k* dimensions
- Option 1: Parameters lie on the ***k*-simplex**

$$\Delta^k = \left\{ (\theta_1, \theta_2, \ldots, \theta_k) \;\middle|\; \sum_{i=1}^{k} \theta_i = 1 \wedge \theta_i \geq 0 \;\forall\; i \in [0, k] \right\}$$





Log-linear Parameterization

Weight vector

Feature vector function

$$p(x) = \frac{\exp \boldsymbol{w}^\top \boldsymbol{f}(x)}{Z}$$

$$\text{where } Z = \sum_{x' \in \mathcal{X}} \exp \boldsymbol{w}^\top \boldsymbol{f}(x)$$

Assumption: Z converges

Categorical (Multinomial) Distributions

- "Naïve" parameterization
  - k outcomes, k(-1) independent parameters
  - Model as tables of (conditional) probabilities
  -

- Log-linear parameterization
  - k outcomes, n, possibly overlapping parameters
    -
    -

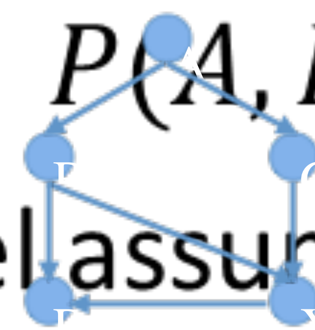
Modelling, inference and conditional independence

- Probabilistic inference usually req probability

$$P(X|A,$$

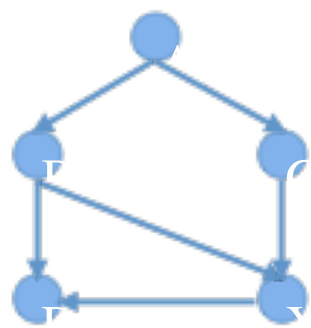
- Or equivalently a joint probability

$$P(A,B,$$



- We will often make model assum

Locally Normalized Models



- Each conditional term is a probability distribution by itself
  - It is *locally normalized*
  - Although the actual model for the distribution may vary

Parameterization

- For each node in the graph
  - We have a multinomial distribution
  - We can use independent parameters (on simplex)
  - We can use log-linear models
    - “Locally normalized model” (cf. Appendix D.2)
    - Z is “local” to the decision being made

Globally Normalized Models

- Extension of the exponential parameterization to structured output spaces

$$p(\mathbf{x}) = \frac{\exp \mathbf{w}^\top \mathbf{F}(\mathbf{x})}{Z}$$

$$\text{where } Z = \sum_{\mathbf{x}' \in \mathcal{X}} \exp \mathbf{w}^\top \mathbf{F}(\mathbf{x}')$$

Conditional Random Fields

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{\exp \mathbf{w}^\top \mathbf{F}(\mathbf{x})}{Z(\mathbf{x})}$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y}' \in \mathcal{Y}_{\mathbf{x}}} \exp \mathbf{w}^\top \mathbf{F}(\mathbf{x})$$

Conditional Random Fields

$$P(y|x) = \quad \mathrm{ex}$$

$$Z(\mathbf{x}) = \sum$$

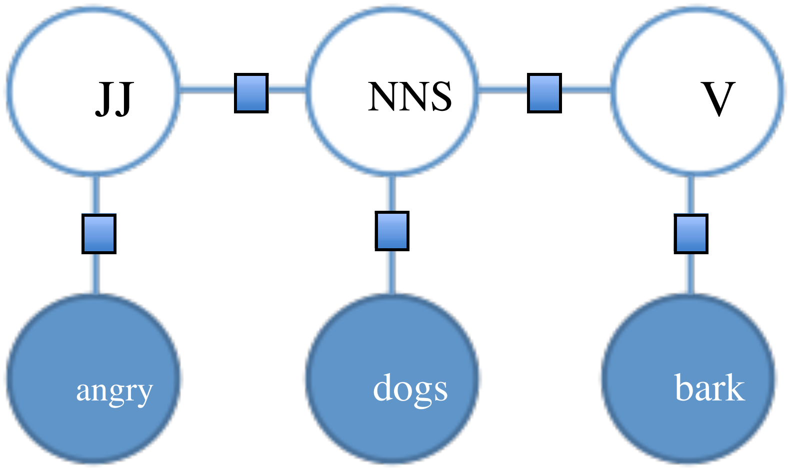
Decoding is nice:

$$y^* = \operatorname{argmax}$$

$$= \operatorname{argmax} \mathrm{e}$$

$$= \operatorname{argmax}$$

Conditional Random Fields



$$\mathbf{F}(\mathbf{x}, \mathbf{y}) = \sum_{C \in G} \mathbf{f}(C)$$

Comparison of Feature-Based Models

- Locally Normalized Models
  - Good joint models
  - Easy to train
  - Downside: decoding can be expensive
- Globally Normalized Models
  - Very popular conditional models (CRFs)
  - Challenge: computing Z / training
  - Advantage: decoding can be cheap