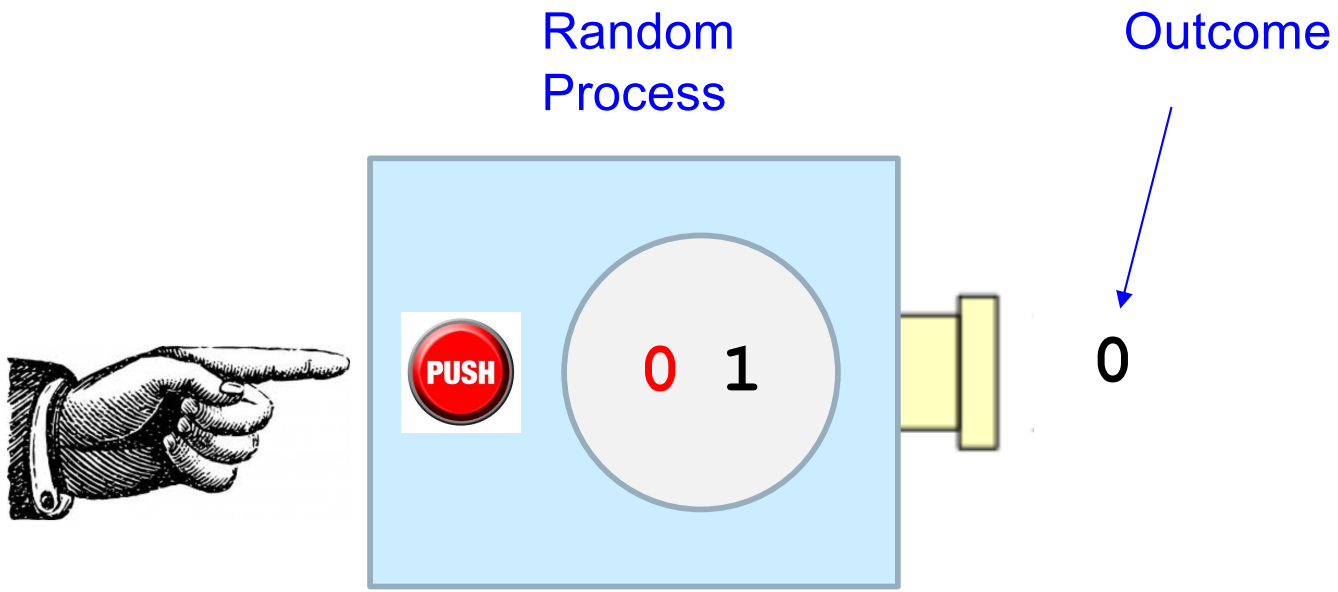
CS 4100: Introduction to AI

Wayne Snyder
Northeastern University

Lecture 11: Review of Probability and Likelihood for Machine Learning

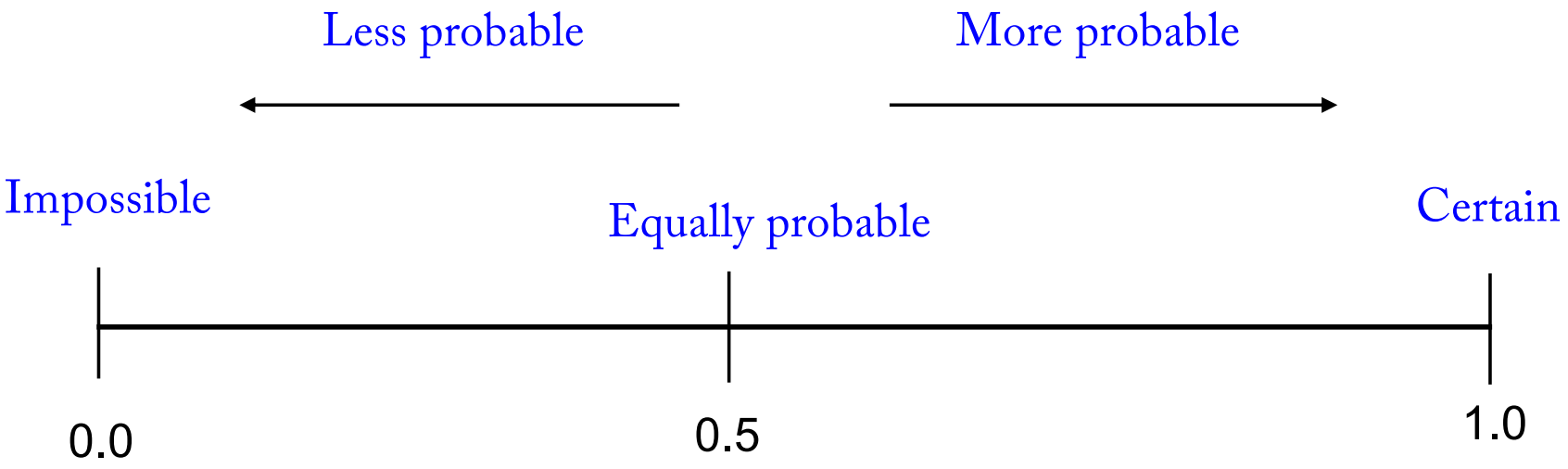
Random Experiments

A Random Experiment is a process that produces uncertain outcomes from a well-defined set of possible outcomes. Usually there is some kind of physical experiment which is being modeled theoretically.



Probability

We measure the probability of events on a real-number scale from 0 to 1:



Finite Equiprobable Probability Spaces

For finite and equiprobable probability spaces,

it is easy to calculate the probability:

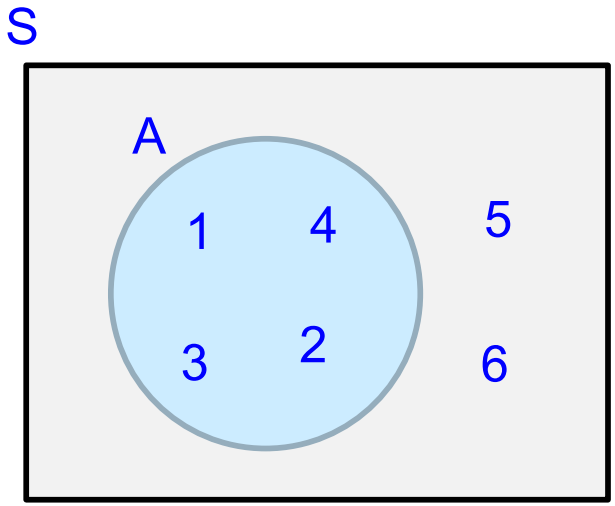
$$P(A) = \frac{|A|}{|S|}$$

Example: Roll a die, how many dots showing on the top face? Let A = "less than 5 dots are showing."

S = { 1, 2, ... , 6 }

P = { 1/6, 1/6, ... , 1/6 }

P(A) = 4/6 = 2/3 = 0.6666...



Conditional Probability: P(A | B) = P(A) if we know B has happened

Example: Roll two dice. A = "the total # dots showing is > 8" and

B = "the first roll was 3"

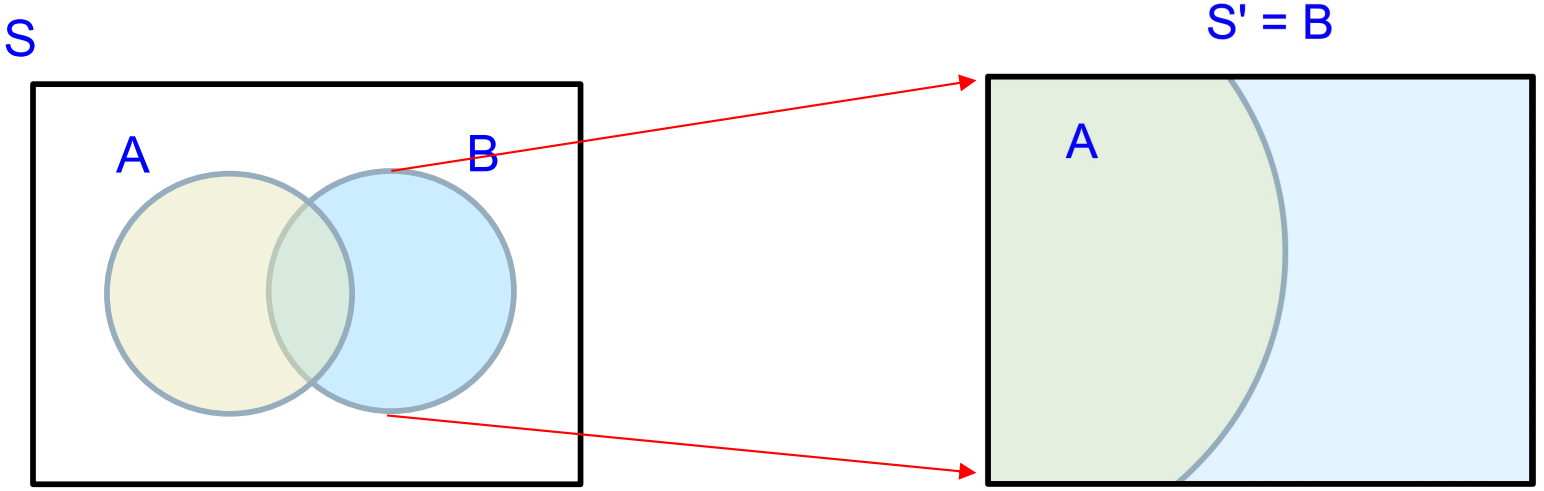
What is P(A | B)?

The key to solving such problems is to realize that there are two probability spaces:

- **the one before you know whether B has happened, and**
- **the one that has been "conditioned" by knowing that B has definitely happened, so the sample space has shrunk and the proportion representing event A may have changed:**

Original

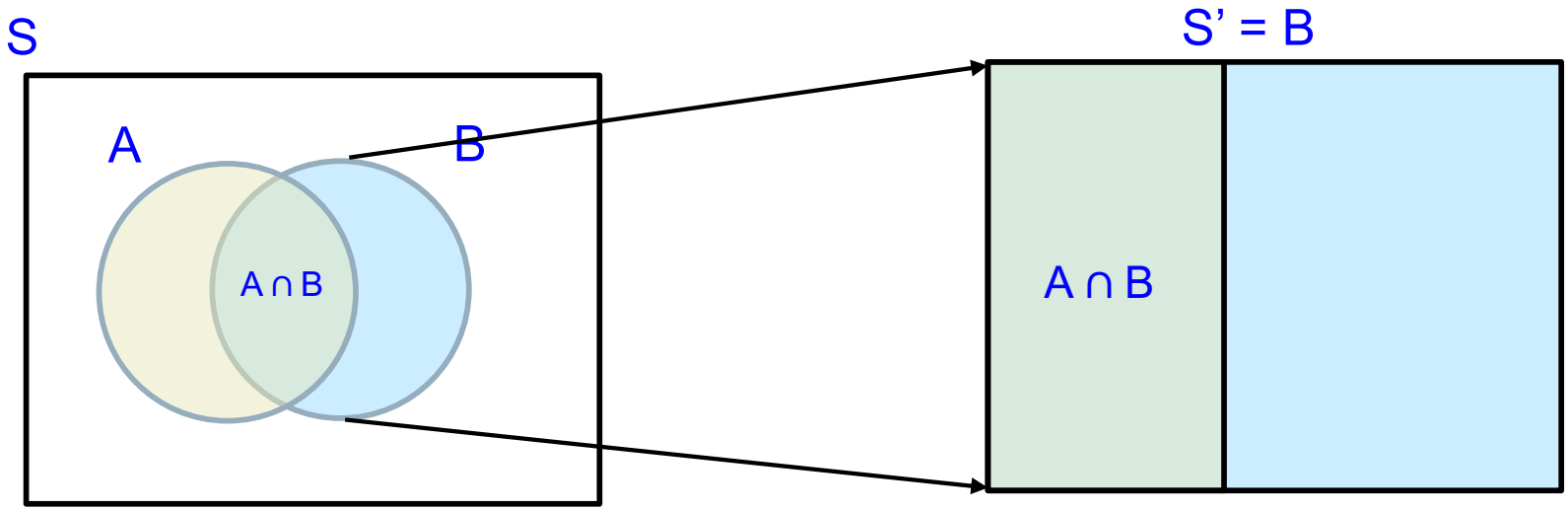
Conditioned by knowing B happened:



Conditional Probability

Conditioning the original sample space means changing the perspective: instead of finding the area of A inside S, we are finding the area of A ∩ B inside B:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



Independence and Dependence

We say that two events A and B are independent if

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

or:

$$P(A \mid B) = P(A)$$

$$P(A \cap B) = P(A) * P(B)$$

Example:

Suppose in a particular city, 40% of the population is male, and 60% female, and 20% of the population smokes. If male smokers are 8% of the population, then are smoking and gender independent? That is, are the following two events independent?

YES. Check:

A = Smoker

B = Male

$P(A \cap B) = 0.08 = 0.4 * 0.2 = P(A) * P(B)$

Bayes' Rule

We can rearrange the conditional probability rule in a way that makes the sequence of the events irrelevant -- which happened first, A or B? Or did they happen at the same time? Does it matter?

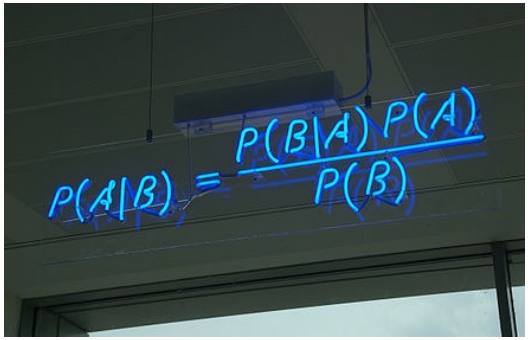
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

We can do a little algebra to define conditional probabilities in terms of each other:

$$P(B \mid A) * P(A) = P(B \cap A) = P(A \mid B) * P(B)$$

so:

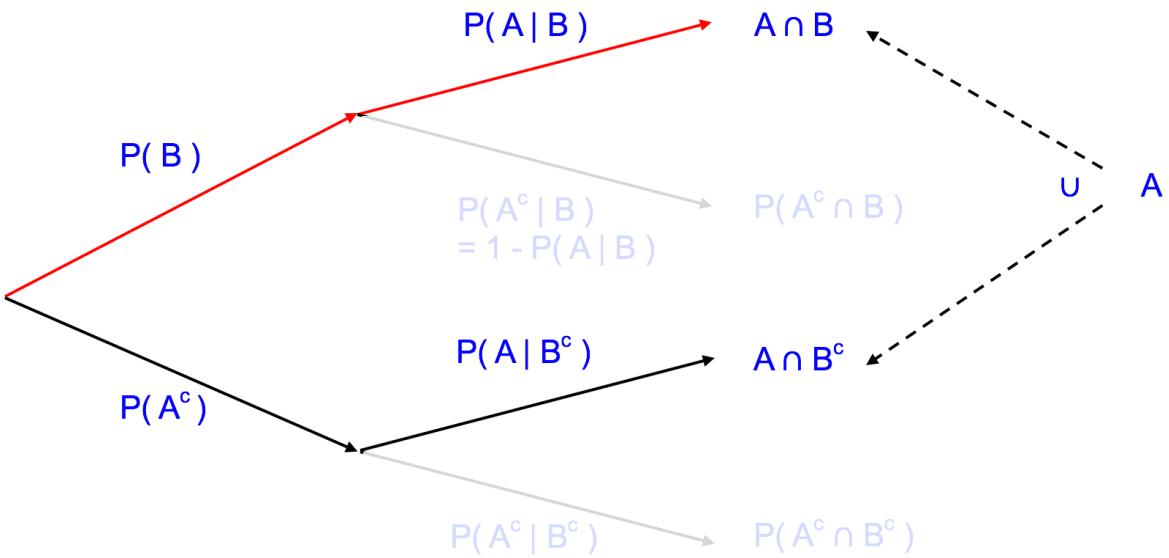
$$P(B \mid A) = \frac{P(A \mid B) * P(B)}{P(A)}$$



Bayes' Rule

The best way to understand this is to view it with a tree diagram!

P(B | A) = the probability that when A happens, it was "preceeded" by B:



If A has happened, what is the probability that it did so on the path where B also occurred?

Note:

$$A = P(A \cap B) \cup P(A \cap B^c)$$

So what percentage of A is due to $A \cap B$?

Same calculation as:

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

Bayes' Rule

This has an interesting flavor, because we can ask about causes of outcomes:

A Priori Reasoning -- "I randomly choose a person and observe that he is male; what the probability that it is a smoker?"

"The first toss of a pair of dice is a 5; what is the probability that the total is greater than 8?"

A Posteriori Reasoning -- "I find a cigarette butt on the ground, what is the probability that it was left by a man?"

"The total of a pair of thrown dice is greater than 8; what is the probability that the first toss was a 5?"

This seems odd, because instead of reasoning forward from "causes to effects" we are reasoning backwards from "effects to causes" but really it is just different ways of phrasing the mathematical formulae. Time is not really relevant!

Discrete RandomVariables: Probability Mass Function

The probability function of a **discrete** random variable X is a function

f_X = Probability Mass Function (PMF)

which assigns a probability to each real number in the range of X and follows the normal rules for a probability space:

$$f_X : R_x \rightarrow \mathcal{R}$$

$$\forall a \in R_x \quad f_X(a) \geq 0$$

$$\sum_{a \in R_x} f_X(a) = 1.0$$

Discrete RandomVariables: Probability Distributions

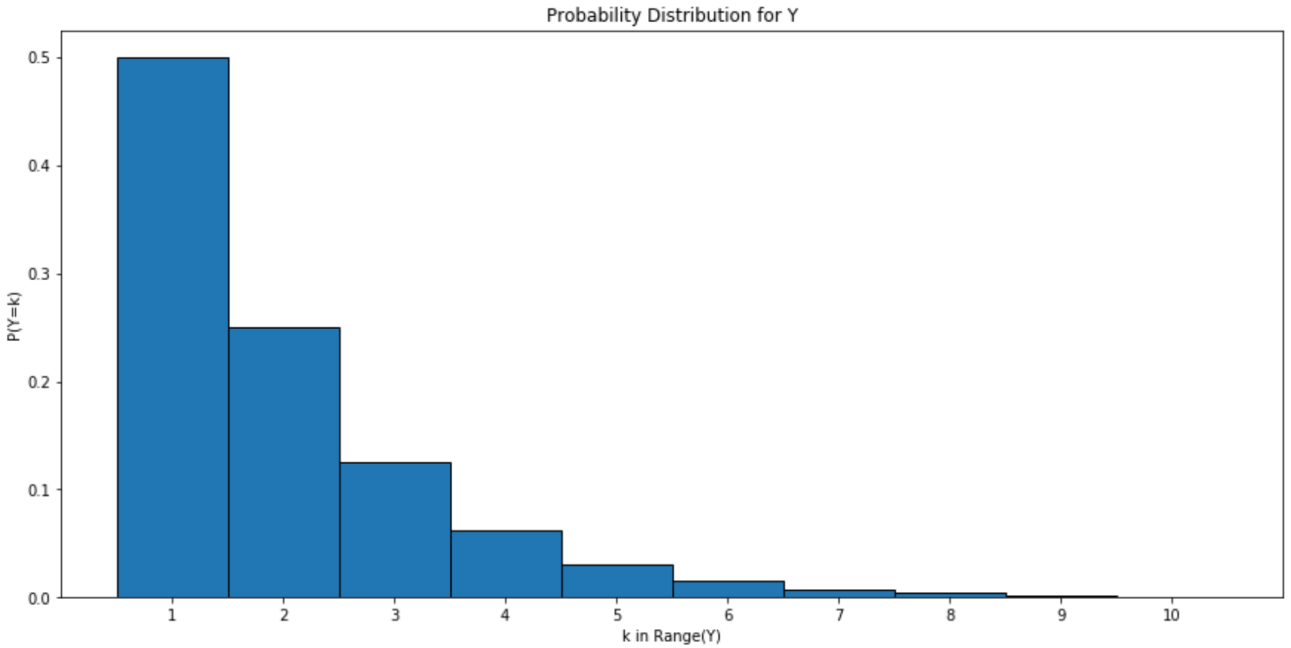
We will emphasize the distributions of random variables, using graphical representations to help our intuitions.

Example:

Y = "The number of tosses of a fair coin until a head appears"

$$R_Y = \{ 1, 2, 3, \ldots \}$$

$$f_Y = \{ \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \ldots \}$$



Discrete vs Continuous Distributions

Discrete Random Variables

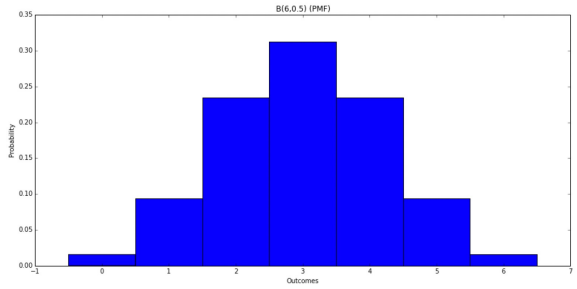
The Probability Mass Function (PMF) of a discrete random variable X is a function from the range of X into $\mathcal{R}$:

$$P_x : R_X \mapsto \mathcal{R}$$

such that

(i) $\forall y \in R_X \; P_X(y) \geq 0.0$

(ii) $\sum_{y \in R_X} P_X(y) = 1.0$



Continuous Random Variables

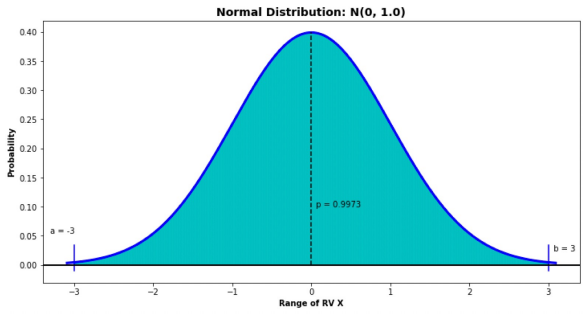
The Probability Density Function (PDF) of a continuous random variable X is a function from $\mathcal{R}$ to $\mathcal{R}$:

$$f_x : \mathcal{R} \mapsto \mathcal{R}$$

such that

(i) $\forall y \; f_X(y) \geq 0.0$

(ii) $\int_{-\infty}^{\infty} f_X(y)\, dy = 1.0$



Examples of Distributions from NLP

I am going to show examples from the Brown Corpus, an early (1961) and still-used corpus of various English texts:

Category	Genre (Code)	# of texts	Total Tokens	%
INFORMATIVE	Learned (J)	80	160,000	16.0%
INFORMATIVE	Belles Lettres, Biography, Memoirs, etc (G)	75	150,000	15.0%
INFORMATIVE	Popular Lore (F)	48	96,000	9.6%
INFORMATIVE	Press: Reportage (A)	44	88,000	8.8%
INFORMATIVE	Skills and Hobbies (E)	36	72,000	7.2%
INFORMATIVE	Miscellaneous (H)	30	60,000	6.0%
IMAGINATIVE	General Fiction (K)	29	58,000	5.8%
IMAGINATIVE	Adventure and Western Fiction (N)	29	58,000	5.8%
IMAGINATIVE	Romance and Love Story (P)	29	58,000	5.8%
INFORMATIVE	Press: Editorial (B)	27	54,000	5.4%
IMAGINATIVE	Mystery and Detective Fiction (L)	24	48,000	4.8%
INFORMATIVE	Press: Reviews (theatre, books, music, dance) (C)	17	34,000	3.4%
INFORMATIVE	Religion (D)	17	34,000	3.4%
IMAGINATIVE	Humor (R)	9	18,000	1.8%
IMAGINATIVE	Science Fiction (M)	6	12,000	1.2%
	TOTAL	500	1,000,000	100.0%

The composition of the Brown Corpus

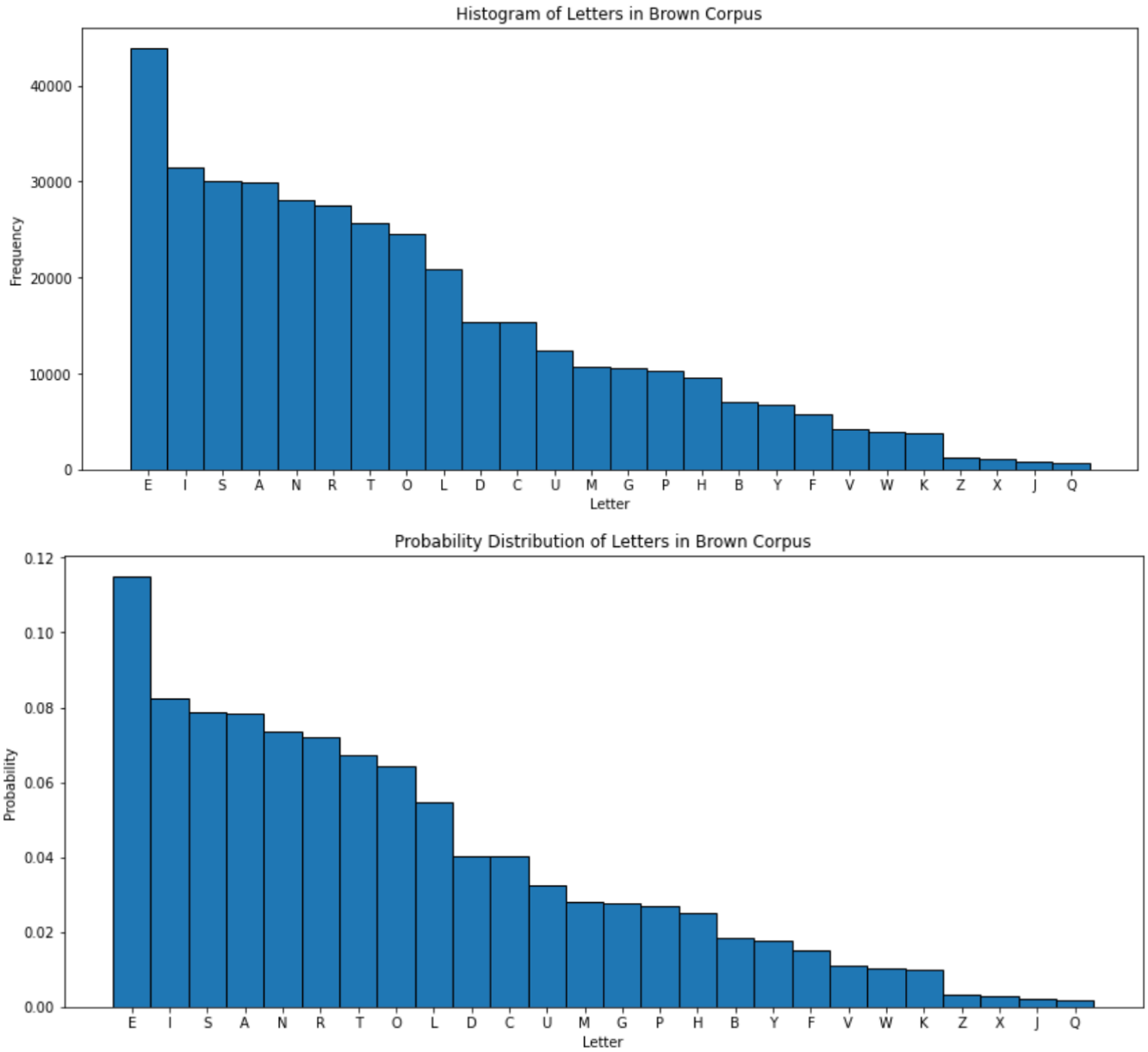
Brown Corpus Sample (untagged)

A01 0010 The Fulton County Grand Jury said Friday an investigation
A01 0020 of Atlanta's recent primary election produced "no evidence" that
A01 0030 any irregularities took place. The jury further said in term-end
A01 0040 presentments that the City Executive Committee, which had over-all
A01 0050 charge of the election, "deserves the praise and thanks of the
A01 0060 City of Atlanta" for the manner in which the election was conducted.

Brown Corpus Sample (tagged)

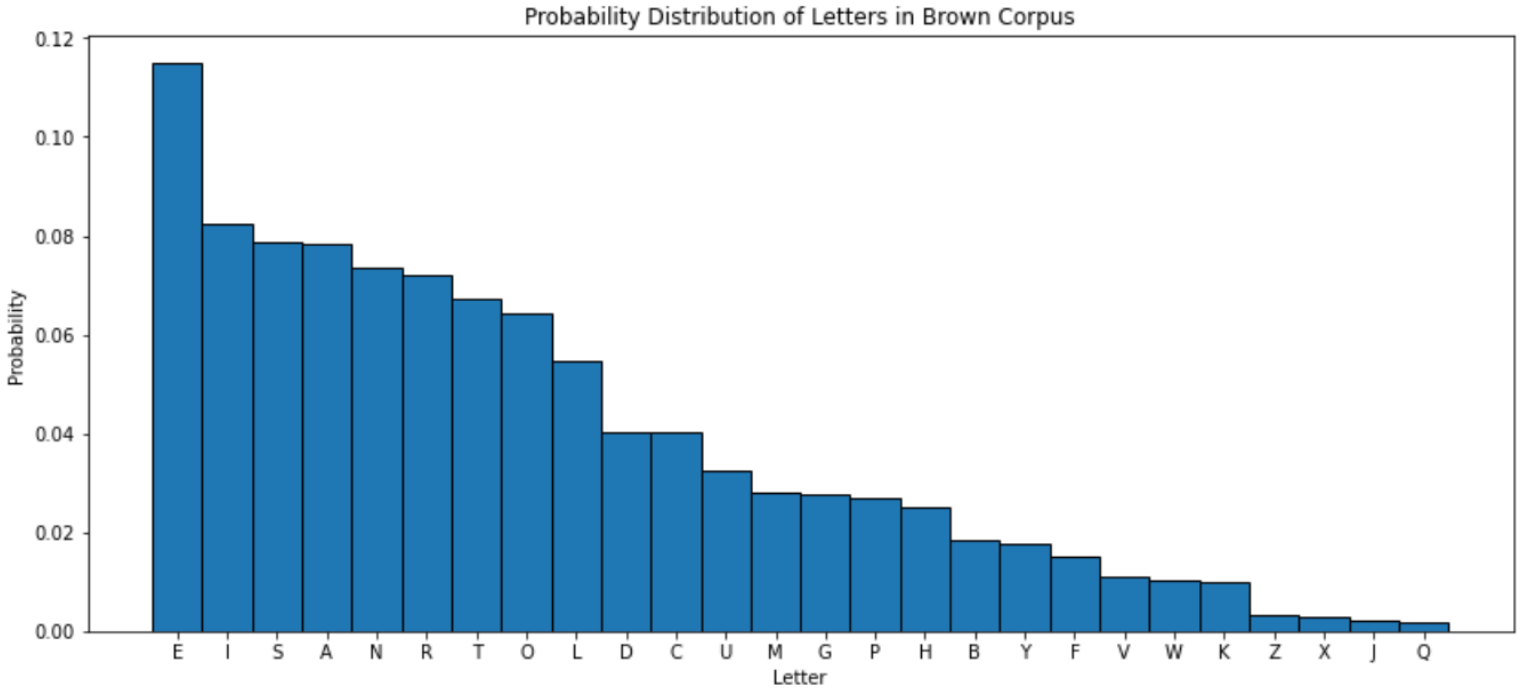
A01_FO 0010_MC The_AT Fulton_NP1 County_NN1 Grand_JJ Jury_NN1 said_VVD Friday_NPD1 an_AT1 investigation_NN1
A01_FO 0020_MC of_IO Atlanta_93 's_03 recent_JJ primary_JJ election_NN1 produced_VVD "_" no_AT evidence_NN1 "_" that_CST
A01_FO 0030_MC any_DD irregularities_NN2 took_VVD place_NN1 ,_.
The_AT jury_NN1 further_RRR said_VVD in_II term-end_NN1
A01_FO 0040_MC presentments_NN2 that_CST the_AT City_NN1 Executive_NN1 Committee_NN1 ,_ which_DDQ had_VHD over-all_RR
A01_FO 0050_MC charge_NN1 of_IO the_AT election_NN1 ,_ "_" deserves_VVZ the_AT praise_NN1 and_CC thanks_NN2 of_IO the_AT
A01_FO 0060_MC City_NN1 of_IO Atlanta_NP1 "_" for_IF the_AT manner_NN1 in_II which_DDQ the_AT election_NN1 was_VBDZ conducted_VVN ._.

Examples of Distributions from NLP

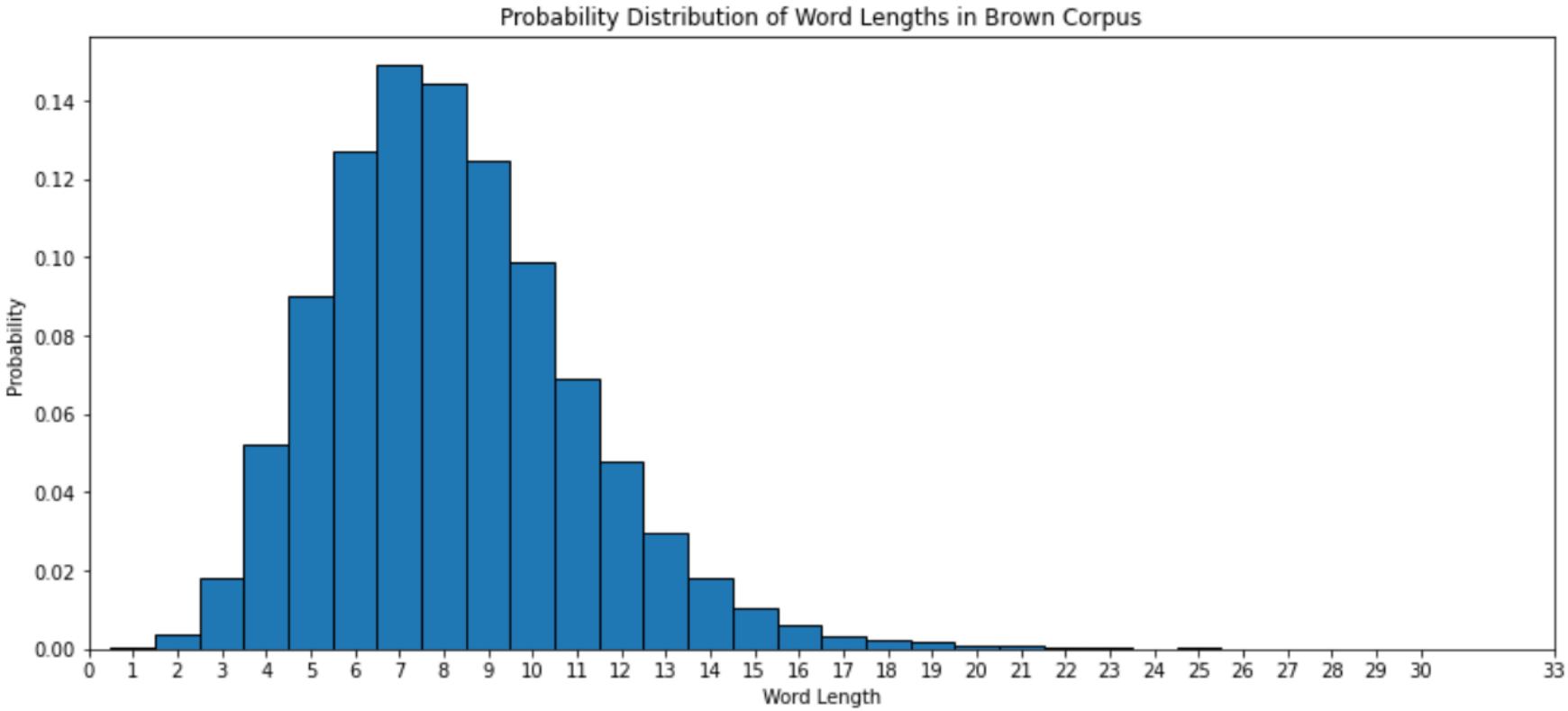


First of all, be sure you understand the difference between a histogram and a probability distribution!

Examples of Distributions from NLP



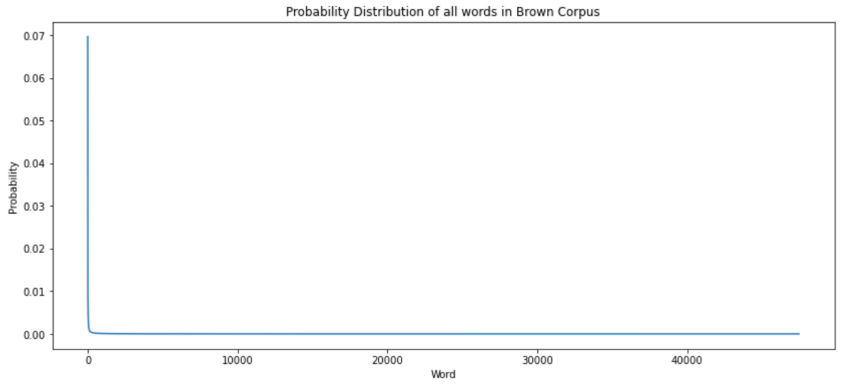
Examples of Distributions from NLP

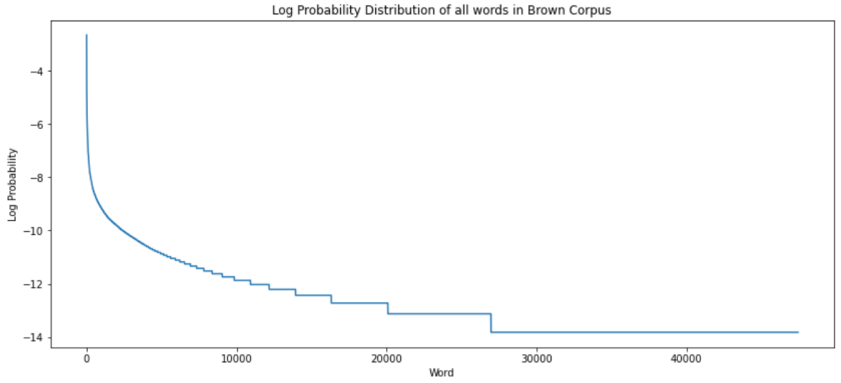


Examples of Distributions from NLP

There are 1,161,192 occurrences of words in the Brown Corpus.

There are 47,451 unique words; 20,478 of these words are *hapax legomena*, i.e., they occur only once in the corpus.

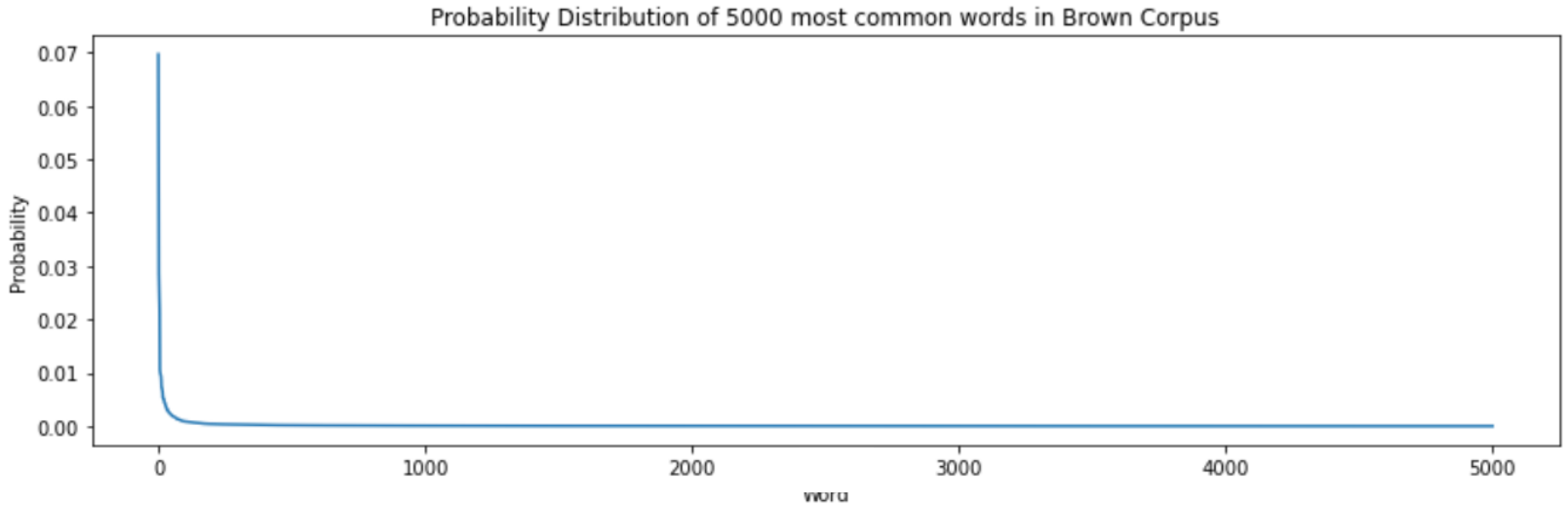


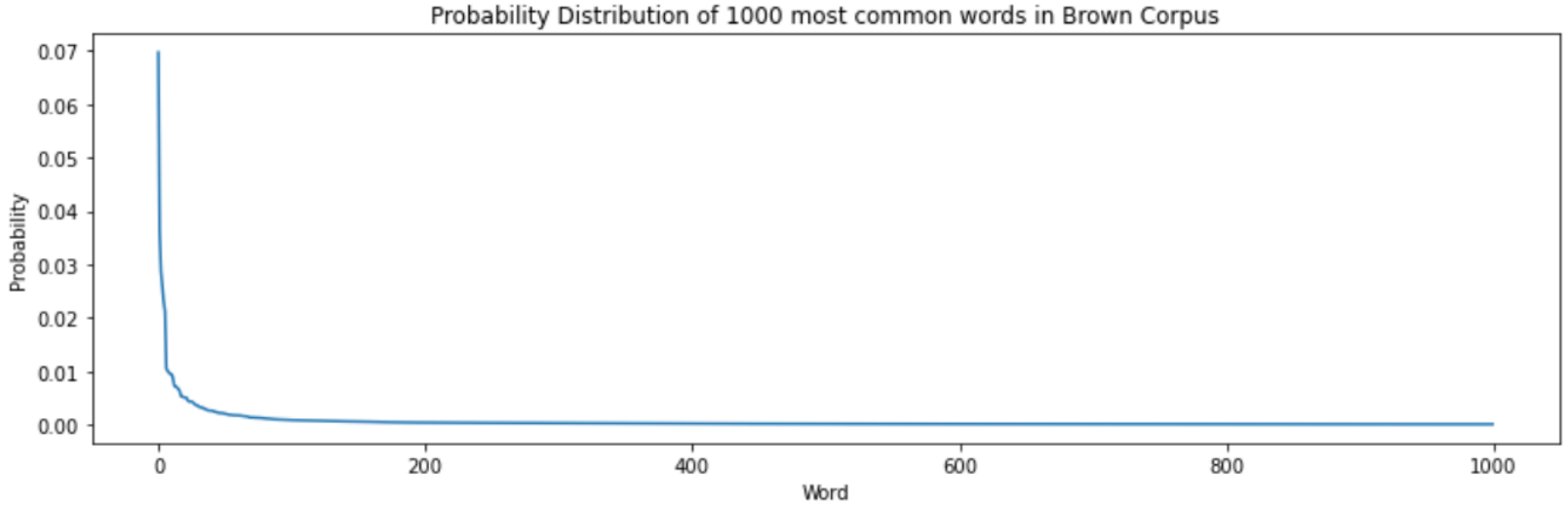


The first 100 words are:

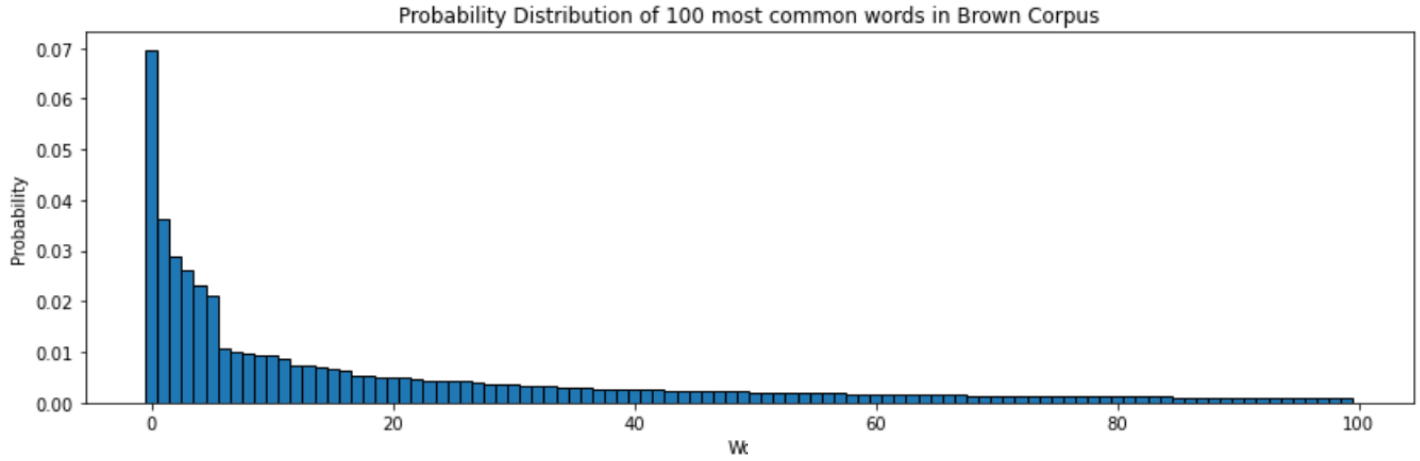
```
['the', 'of', 'and', 'to', 'a', 'in', 'that', 'is', 'was', 'he', 'for', 'it', 'with', 'as', 'his', 'on', 'be', 'at',
'by', 'i', 'this', 'had', 'not', 'are', 'but', 'from', 'or', 'have', 'an', 'they', 'which', 'one', 'you', 'were', 'he
r', 'all', 'she', 'there', 'would', 'their', 'we', 'him', 'been', 'has', 'when', 'who', 'will', 'more', 'if', 'no']
```

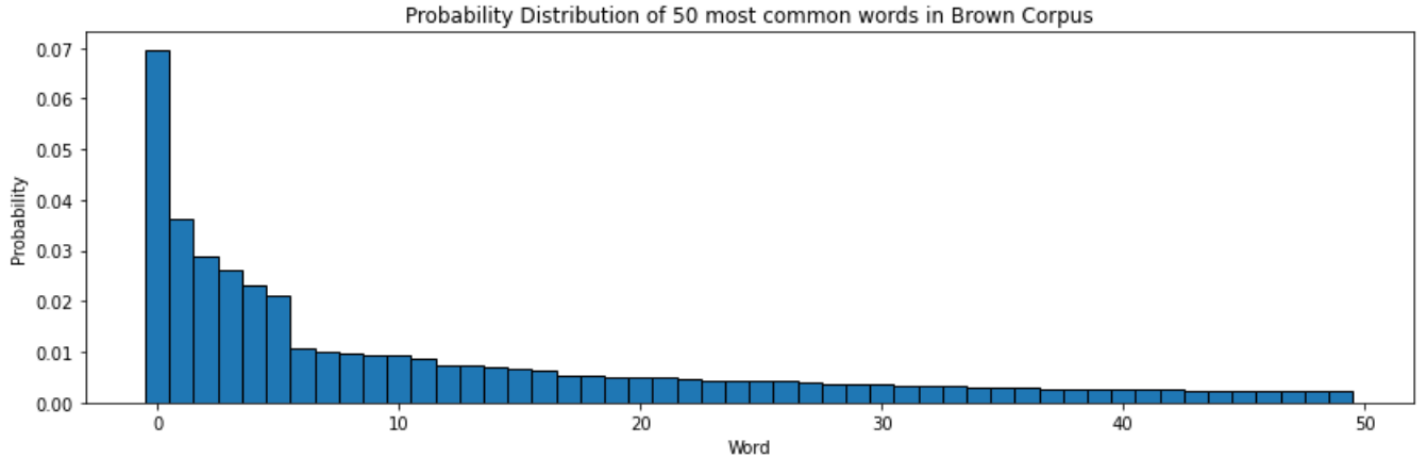
Examples of Distributions from NLP





Examples of Distributions from NLP

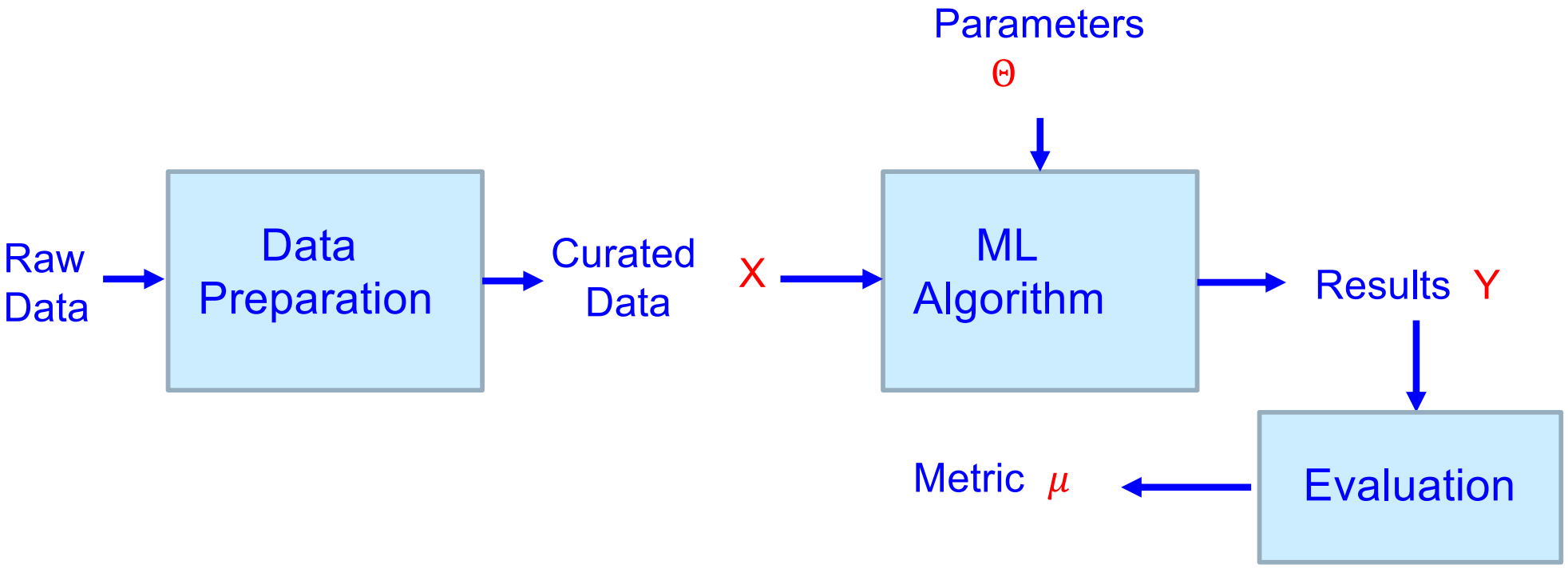





```
['the', 'of', 'and', 'to', 'a', 'in', 'that', 'is', 'was', 'he', 'for', 'it', 'with', 'as', 'his', 'on', 'be', 'at',
'by', 'i', 'this', 'had', 'not', 'are', 'but', 'from', 'or', 'have', 'an', 'they', 'which', 'one', 'you', 'were', 'he
r', 'all', 'she', 'there', 'would', 'their', 'we', 'him', 'been', 'has', 'when', 'who', 'will', 'more', 'if', 'no']
```

Bayesian Reasoning and Maximum Likelihood Estimates (MLE)

A typical machine learning workflow is the following:



Bayesian Reasoning and MLE

Typically:

Pick initial values for the parameters Θ, based on some initial knowledge (e.g., from a textbook, a similar experiment, a "hunch," etc.),

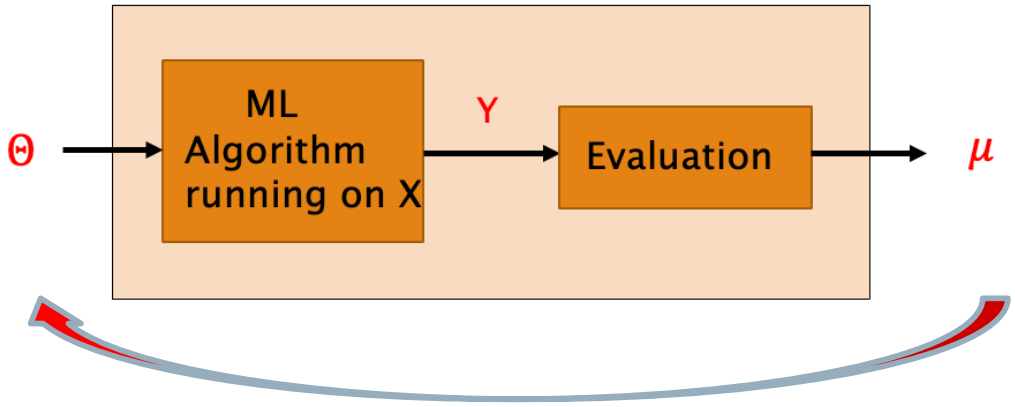
Repeat

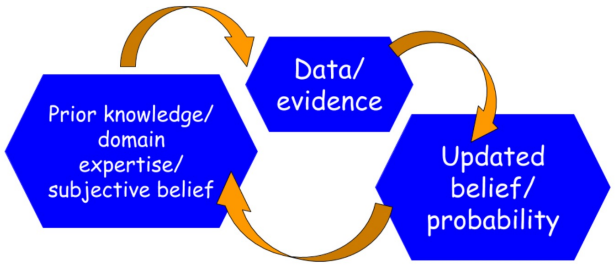
Perform experiment and find new μ

Based on knowledge gained from last run revise Θ

Until .....

But notice this is a bit... weird! We are used to thinking of programs as starting with inputs and producing outputs, but in ML the emphasis is on using the outputs to determine the inputs!

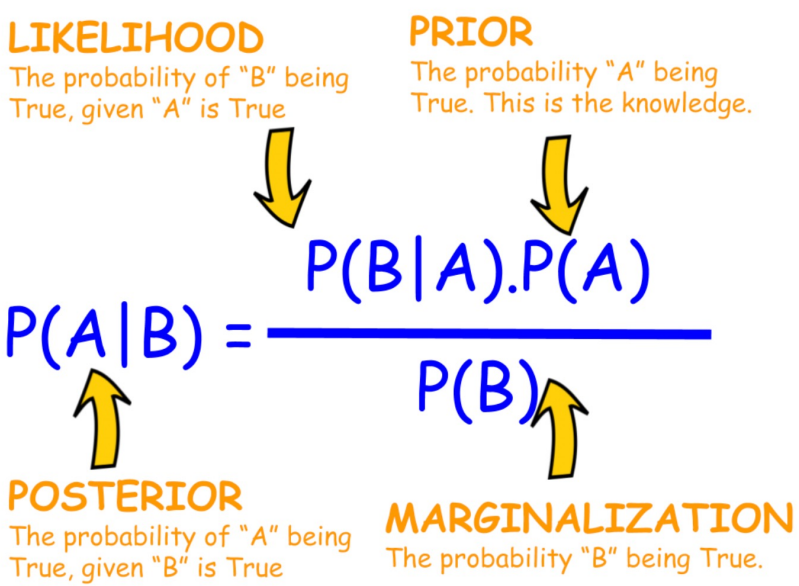


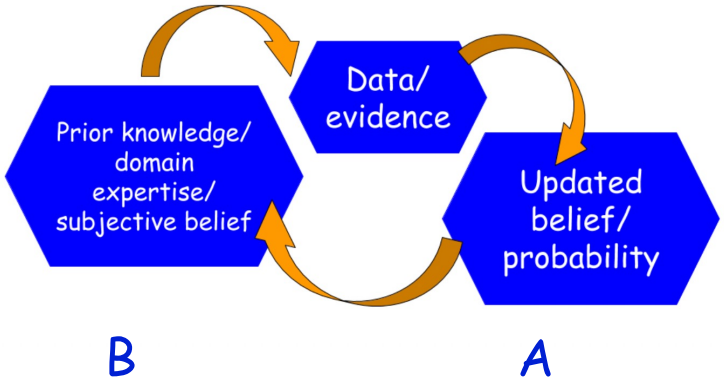


Bayesian Reasoning and MLE

Bayesian Reasoning is a way of thinking about probability which emphasizes this point of view, and has become standard in NLP and ML.

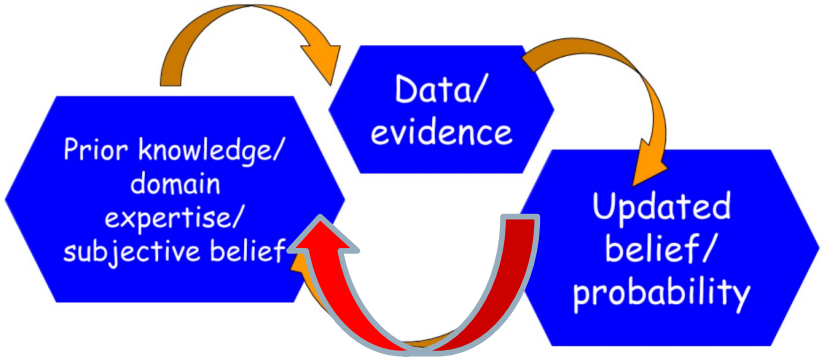
Formally, we have Bayes Rule and some terminology:





Bayesian Reasoning and MLE

But notice this is a bit... weird! We are used to thinking of programs as starting with inputs and producing outputs, but in ML the emphasis is on using the outputs to determine the inputs!



This is generally called the Bayesian interpretation of the experiment: Given some experimental result, how likely is the parameter θ to be some particular value?

Note: Probability refers to outcomes, and likelihood refers to parameters.

Bayesian Reasoning and MLE

Let's consider a simple random experiment:

We flip a (possibly) unfair coin n times and record the outcomes.

Frequentist: The data will converge in the limit to a specific probability θ which existed before we started to flip:

```
[ ]  from numpy import mean

     theta = 0.65

     for n in range(7):
       N = 10**n
       print("Over",N," iterations, the mean is", mean([X(theta) for k in range(N)]))
```

```
Over 1  iterations, the mean is 1.0
Over 10  iterations, the mean is 0.7
Over 100  iterations, the mean is 0.67
Over 1000  iterations, the mean is 0.648
Over 10000  iterations, the mean is 0.657
Over 100000  iterations, the mean is 0.65094
Over 1000000  iterations, the mean is 0.650158
```

Bayesian: Given some experimental result, what is the MOST LIKELY value for the parameter θ?

Bayesian Reasoning and MLE

In our example we would naturally assume that θ is simply the mean of the outcomes observed so far, for example, if we get outcomes

```
[ ]  theta = 0.25
     n = 100
     outcomes = [X(theta) for k in range(n) ]
     print(outcomes)
```

```
[0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0
```

we would assume that θ is the mean of the n outcomes, i.e.,

$$\hat{\theta} = \frac{count(\text{outcomes} = 1)}{count(\text{all outcomes})}$$

Bayesian Reasoning and MLE

```
[ ]  theta_hat = mean(outcomes)

     print("theta_hat =", theta_hat)

     theta_hat = 0.24
```

This is the most likely estimate for θ, since we have no other information.

The more outcomes we see, the more accurate an estimate of θ we would get. However, notice that in general, we will not get θ exactly right, but only an *estimate* based on our experiment. This is appropriate, because we did not know θ to start with, and we want the *most likely* value.

Bayesian Reasoning and MLE

Now we can introduce the general notion of the Maximum Likelihood Estimate (MLE), which applies to our simple experiment and to many more complex situations which occur often in NLP and ML.

If we express the situation mathematically, keeping the parameter θ in the foreground, let us consider the n outcomes of the previous experiment as a sequence of n independent and identically-distributed random variables, X_1, X_2, ..., X_n, where

$$X_i(\theta) \approx Bernoulli(\theta).$$

and the PDF (which assigns a probability to each outcome x) is

$$f(x, \theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

In the frequentist interpretation, we would ask: Given the parameter θ, how likely are the outcomes $X_1,\ldots,X_n$, and get the answer

$$f(X_1, \theta) * f(X_2, \theta) * \cdots * f(X_n, \theta) = \prod_{i=1}^{n} f(X_i, \theta).$$

Bayesian Reasoning and MLE

```
[ ]  import numpy as np

     def f(x,theta):
       if(x==1):
         return theta
       else:
         return 1-theta


     outcome_probs = [f(x,theta) for x in outcomes]
     print("Outcomes", outcomes)
     print("Outcome probabilities:",outcome_probs)
     print("With theta =",theta,"the probability of this particular list of outcomes is ",
           np.prod( outcome_probs))
```

```
Outcomes [0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
Outcome probabilities: [0.75, 0.75, 0.75, 0.75, 0.25, 0.75, 0.75, 0.75, 0.75, 0.25, 0.75, 0.75, 0.75, 0.75,
With theta = 0.25 the probability of this particular list of outcomes is  1.1355760538867237e-24
```

Bayesian Reasoning and MLE

In the Bayesian interpretation, where we don't assume we know θ, we would ask a more useful question: Given the set of outcomes, **what is the *most likely* value for θ?** Alternately, **what estimate of θ is most likely given our outcomes?**
This is formalized as the *likelihood* function for n outcomes with parameter θ:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i, \theta).$$

The **maximum likelihood estimator** (MLE) is then

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}_n(\theta) = \text{the value of } \theta \text{ that maximizes } \mathcal{L}_n(\theta)$$

If we try many values of the parameter, we would see that it reaches a maximum at the value found by simple counting....

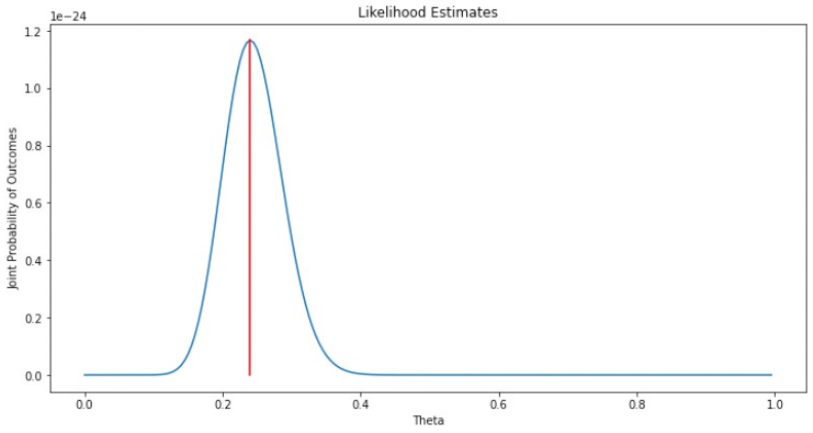
Bayesian Reasoning and MLE

```
from matplotlib import pyplot as plt

T = np.arange(0,1,0.005)
P = [np.prod( [f(x,theta) for x in outcomes] ) for theta in T]

mx = np.max(P)

plt.figure(figsize=(12,6))
plt.title("Likelihood Estimates")
plt.xlabel("Theta")
plt.ylabel("Joint Probability of Outcomes")
plt.plot(T,P)
plt.plot([theta_hat,theta_hat],[0,mx],color='r')
plt.show()
```



The MLE formalizes the idea that we are searching for the best possible (= most likely) set of parameters; we shall see (e.g., in Logistic Regression) that when there is an explicit formula for $\mathcal{L}_n$, we can calculate the MLE directly using the derivative of $\mathcal{L}_n$.