Illinois Institute of Technology
Department of Computer Science

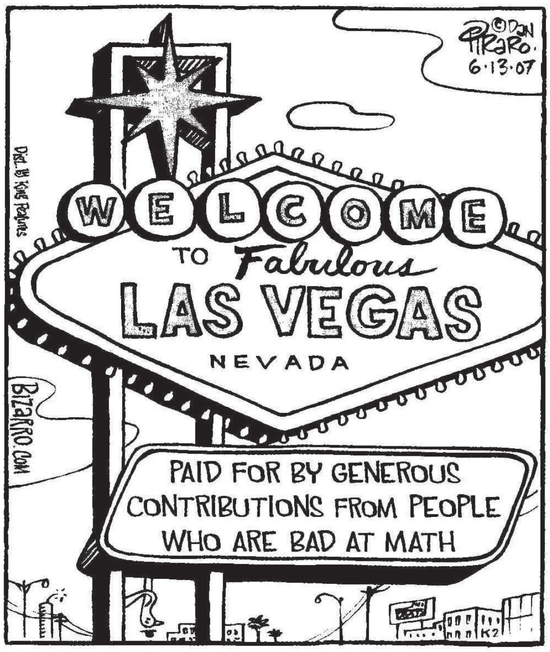
Lecture 10: February 20, 2019

CS 330 Discrete Structures
Spring Semester, 2019

> *"I think you're begging the question," said Haydock, "and I can see looming ahead one of those terrible exercises in probability where six men have white hats and six men have black hats and you have to work it out by mathematics how likely it is that the hats will get mixed up and in what proportion. If you start thinking about things like that, you would go round the bend. Let me assure you of that!"*
>
> —Agatha Christie: *The Mirror Crack'd* (1962)

> *Misunderstanding of probability may be the greatest of all impediments to scientific literacy.*
>
> —Stephen Jay Gould : *Dinosaur in a Haystack: Reflections on Natural History* (1996)

> *I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be.*
>
> —Lord Kelvin (William Thomson) (1883)



1 Probability

When the weather-woman says “there is a 30% chance of rain”, what does she mean? Does she mean that:

- rain will fall on 30% of the viewing area?
- in the last 100 years, it rained 30 times on this date?
- under present conditions, recorded history for comparable conditions shows it rained 30% of the time?

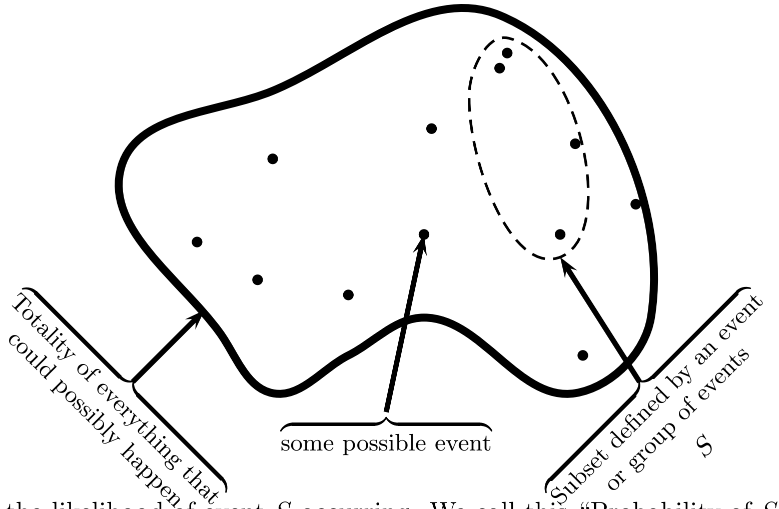
Similarly, what does it mean for an algorithm to be correct 99% of the time? Can we trust it to guard our nuclear warheads?

In order to answer these questions, we first need to know a little about **probability**.

There are two distinct forms of probability: *Aleatory* (from the Latin meaning “dice player”) looks at theoretical events such as rolling dice, drawing balls out of an urn, or random input to an algorithm. *Epistemic* (knowledge-based) probability looks not at randomness, but at lack of knowledge—such as the question of how likely it is rain tomorrow; in this case we could reduce uncertainty by better knowledge. In this course we will be concerned only with aleatory probabilty.

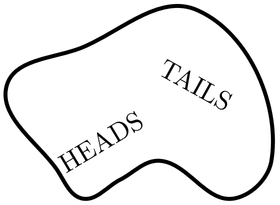
2 Basics of Probability

Consider the following amoeba-like graph:


Totality of everything that could possibly happen
some possible event
Subset defined by an event or group of events
S

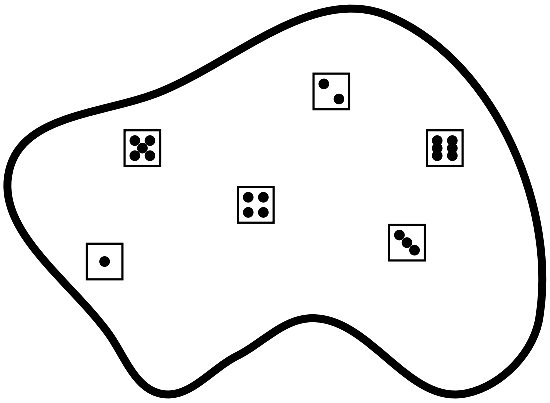
We wish to measure the likelihood of event S occurring. We call this “Probability of S” and we write it as $\mathbf{Pr}(S)$.

Consider the following experiment: We will flip a “fair” coin. We have the following “universe” of outcomes:



Our intuition tells us that $\mathbf{Pr}(\text{TAILS}) = 1/2$ because TAILS happens about half the time.

Consider this experiment: We will roll a "fair" die. We have the following "universe" of outcomes:



Our intuition tells us that $\mathbf{Pr}(\text{Rolling a one}) = 1/6$, because rolling a "one" occurs about one sixth of the time.

From these we can see a simple definition of Probability as:

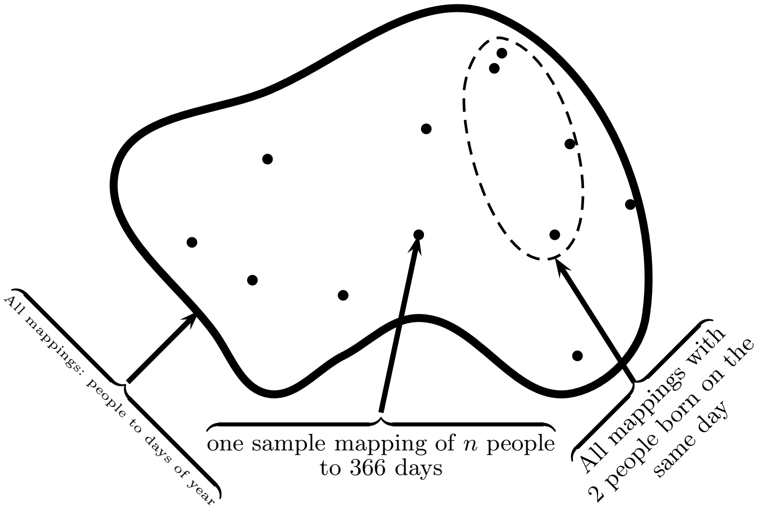
$$\mathbf{Pr}(S) = \frac{\text{number of events in } S}{\text{number of events altogether}}$$

This definition will be ample for our purposes.

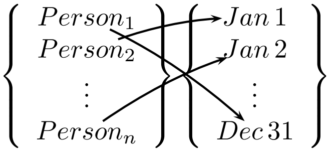
3 The Birthday Problem

An illustration of the power of probability is the Birthday problem: If we have n people in a room, what is the probability that two people celebrate their birthday on exactly the same day?

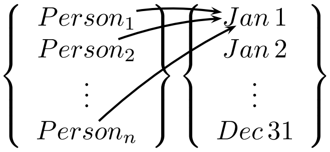
After a careful examination, you will notice that the set we have defined is:


All mappings: people to days of year
one sample mapping of n people to 366 days
All mappings with 2 people born on the same day

A mapping simply relates people to a day of the year corresponding to their birthday. For instance



and



are mappings.

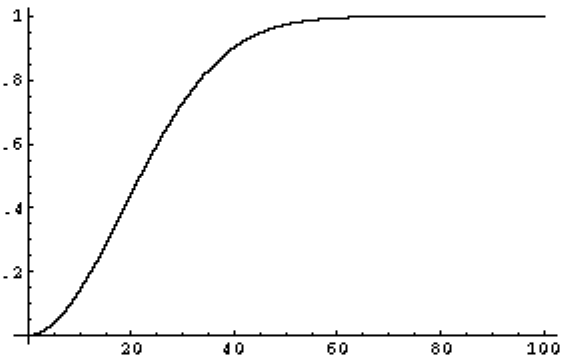
There are 366^n ways to map n people to birthdays. There are $\frac{366!}{(366-n)!}$ ways to pick the birthdays so that no two people share the same birthday. Using our definition of probability:

$$\mathbf{Pr}(\text{No two birthdays on the same day}) = \frac{\frac{366!}{(366-n)!}}{366^n} = 1(1-\frac{1}{366})(1-\frac{2}{366})\cdots(1-\frac{n-1}{366})$$

Here we can see how this probability decreases when n grows:

n	1	2	3	4	5	6	7	8	9	10
$\frac{\frac{366!}{(366-n)!}}{366^n}$	1	0.9973	0.9918	0.9837	0.9729	0.9596	0.9439	0.9259	0.9056	0.8834
n	11	12	13	14	15	16	17	18	19	20
$\frac{\frac{366!}{(366-n)!}}{366^n}$	0.8592	0.8334	0.8061	0.7774	0.7477	0.7171	0.6857	0.6539	0.6217	0.5894
n	21	22	23	24	25	26	27	28	29	30
$\frac{\frac{366!}{(366-n)!}}{366^n}$	0.5572	0.5252	0.4937	0.4627	0.4323	0.4028	0.3742	0.3466	0.3201	0.2947

Here is a graph of the probabilty of having a shared birthday for a given number of people:



It turns out that when there are 23 people in the room the **Pr**(No two birthdays on the same day) ≈ 0.5. This result has direct relation to computer science and hash tables, because it says that a hash table can be nearly empty (23 people compared to 366 days of the year, in our analogy), but it is still very likely that there will be a hashing conflict (that you will have two people with the same birthday, in our analogy).

For much more information on the birthday problem, including some fascinating graphs, see

`http://en.wikipedia.org/wiki/Birthday_problem`

The birthday problem makes at least one appearance in the law! In the court of appeals of the State of California, in the case of The People of the State of California versus John Puckett, (Case No. A121368, San Francisco County Superior Court). See

`http:www.personal.psu.edu/dhk3/dhblog/ROB%28Puckett-CA%29.pdf`

Exercise How many people are needed to have the probability of the same birth *month* be at least 0.5? The same week of the year?

Exercise How many people are needed to have probability 50% that one of them has a birthday on December 25?

4 An Application—GUIDs

According to Wikipedia,

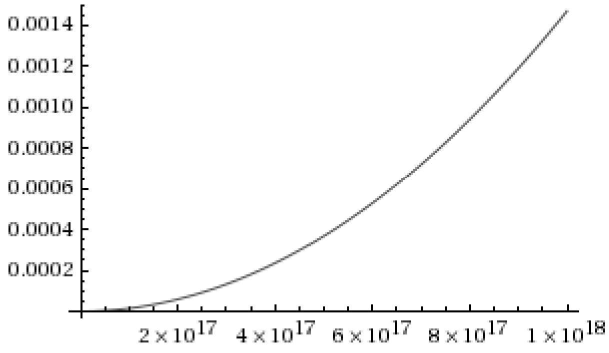
> A globally unique identifier or GUID is a 128-bit identifier used in software applications to provide a reference number which is unique in any context (hence, “globally”), for example, in defining the internal reference for a type of access point in a software application, or for creating unique keys in a database. While each generated GUID is not guaranteed to be unique, the total number of unique keys ($2^{128} \approx 3.4 \times 10^{38}$) is so large that the probability of the same number being generated twice is extremely small.

But how small? Specifically, suppose that a million GUIDs are generated every hour of every day for 100 years; what is the probability of a duplicate GUID?

The total number of GUIDs generated will be less than $10^6 \times 24 \times 366 \times 100 < 10^{12}$. Reasoning as in the birthday problem, the probability of a duplication is

$$1 - \frac{2^{128}!}{(2^{128})^{10^{12}}(2^{128} - 10^{12})!} < 10^{-38}$$

by Stirling's approximation. Here is a graph of the probabilty of getting a duplicate GUID for a given number of GUIDs generated



For details of the calculation, see

`http://alumnus.caltech.edu/~psaipetc/birthdayparadox/birthdayparadox.htm`

5 Another Application—Elevators

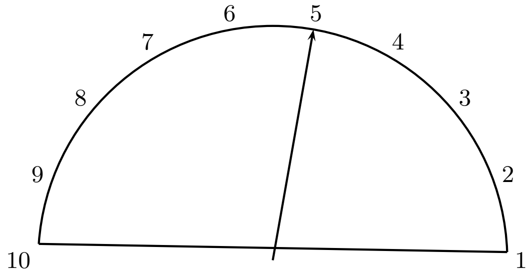
Ever wonder why the wait for an elevator seems interminable? Why the next elevator to come along is usually going in the wrong direction? Let's figure out why. Assume we have elevators that go up and down continuously from the bottom floor to the top floor and then back again in cyclic fashion; assume all elevators are independent of one another (this is *not* usually true: elevator software uses sophisticated techniques to balance the traffic). Finally, we assume that at the moment we begin to wait for an elevator, the elevators are all at random floors. The key value is

$$p = \frac{\text{distance from our floor to the bottom floor}}{\text{distance from top floor to bottom floor}}.$$

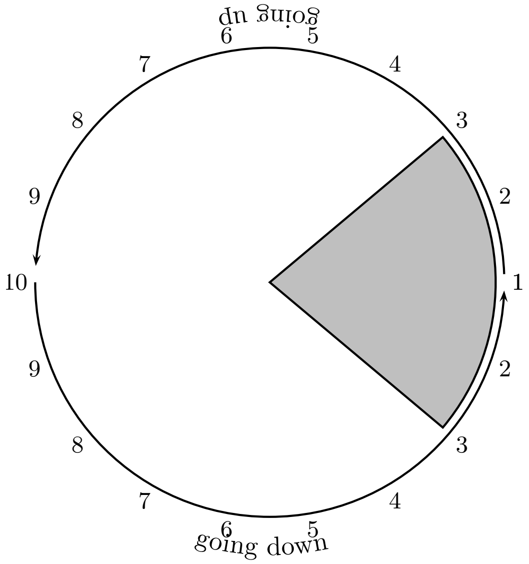
5.1 One elevator

Approaching the one elevator at a random time, there is probability p that it is below us (and hence will be going up when it reaches our floor) and probability $1 - p$ that it is above us (and hence will be going down when it reaches our floor). So, if we are waiting on the third floor to go up in an ten story building, $p = 2/9$, so seven times out of nine the next elevator to stop where we are waiting will be going the wrong way!

Recall the very old style elevator indicator:



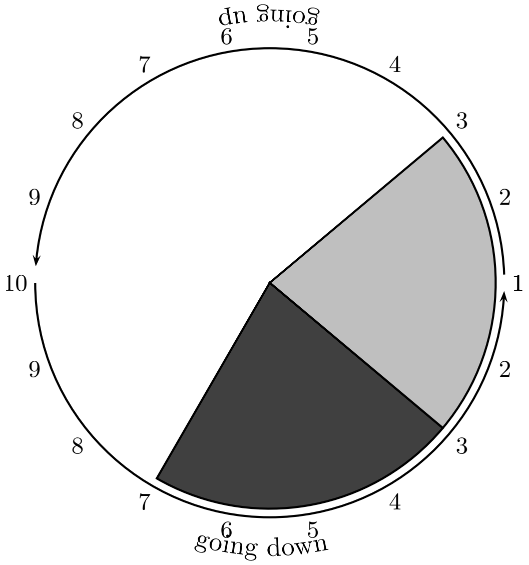
The arrow sweeps counterclockwise when the elevator is going up and it sweeps clockwise when the elevator is going down. To make the movement always counterclockwise, the elevator's position can be thought of as being at points on a circle:



If the elevator is in the white region, it will be going down when it comes to the third floor; if it is in the gray region, it will be going up when it comes to the third floor. The gray region is 2/9 of the circle and we can thus understand the 2/9 probability as being the likelihood that the elevator is in that region. So if we want to go up from the third floor of a 10-story building, 7/9 ≈ 78% of the time the elevator will be going the wrong way.

5.2 Two elevators

With multiple elevators the problem is more complex because we are concerned with the direction of the *first* elevator that stops on our floor. Suppose there are two elevators. Now the diagram looks like this:



If both elevators are in the unshaded region, the first elevator to come to our floor will be going down.

If there is an elevator in the dark gray region and no elevator in the light gray region, the first elevator to come to our floor will be going down. If there is an elevator in the light gray region and no elevator in the dark gray region, the first elevator to come to our floor will be going up. These cases are mirror images.

If there is an elevator in each of the gray regions, the closer elevator reaches us first; but for every such case in which an elevator going down, the mirror image case has it going up. Thus in half of the cases of an elevator in each region, it will be going up and in half of the cases it will be going down.

So, what is the probability that the first elevator is going down? It is the probability that both elevators are in the unshaded region, plus half of the probability that they are not:

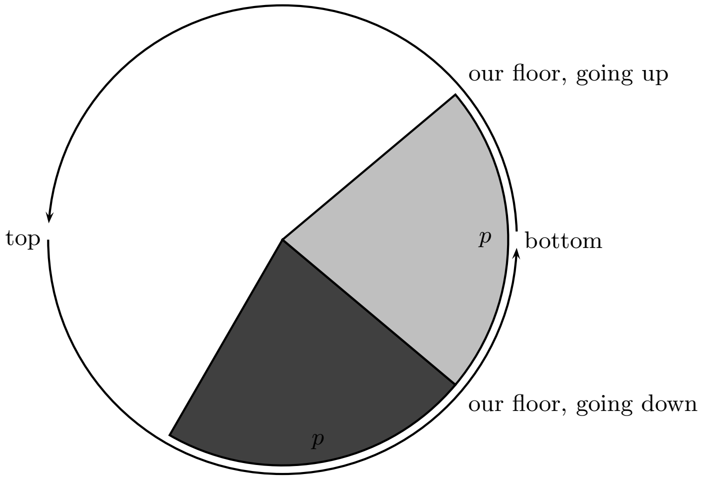
$$(1-4/9)^2 + \frac{1}{2}\left(1-(1-4/9)^2\right) = \frac{53}{81},$$

so the probability that the first elevator is going up is 28/81 (versus $2/9 = 18/81$ for one elevator).

Exercise Do this same analysis (as we just did for 2 elevators) for 1 elevator.

5.3 Many elevators

Now suppose there are n elevators; scale the picture so the circumference of the circle is 1. We have



Exercise Explain what it means if $p > 1/2$.

with each gray region having arc length p. We can assume that $p \leq 1/2$ because if $p > 1/2$ we could consider the analogous problem with $1-p$ and the directions reversed. The calculation is now

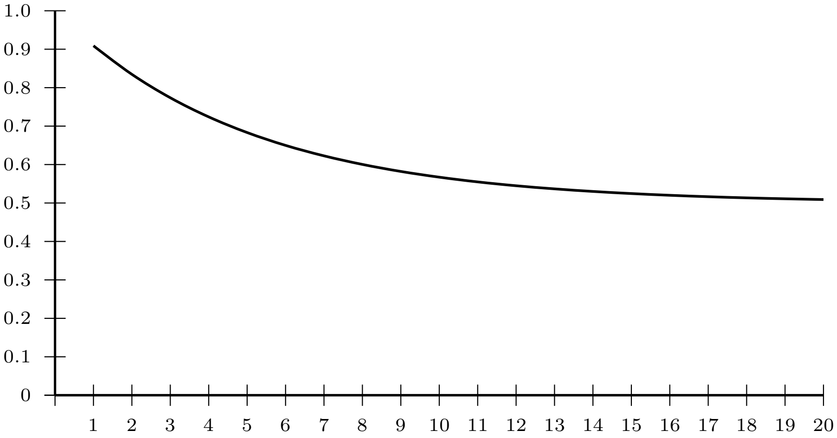
$$\Pr(\text{next elevator going down}) = (1-2p)^n + \frac{1}{2}\left(1-(1-2p)^n\right) = \frac{1}{2} + \frac{1}{2}(1-2p)^n.$$

Exercise What happens when $p = 1/2$?

For example, in a 23-story building with 6 elevators, if we are waiting on the third floor, we have $p = 2/22 = 1/11$, $n = 6$, so the probability that the next elevator is going down is

$$\frac{1}{2} + \frac{1}{2}(1 - 2/11)^6 \approx 0.65,$$

so almost 2/3 of the time the first elevator to come by will be going down. Here is a graph of the number of elevators (horizontal axis) versus the probability that the next elevator to stop at the third floor of a 23-story building will be going up:



As is clear from the formula and the graph, as the number n of elevators increases, the probability approaches 1/2.