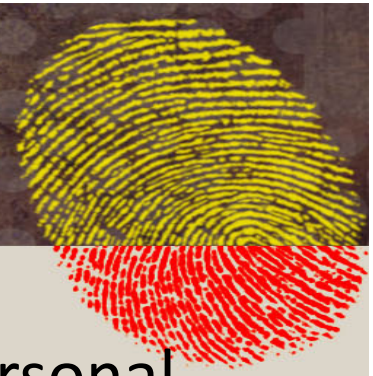
Probabilistic reasoning for digital forensics

Pavel Gladyshev

Disclaimer



- The material presented in this workshop reflects my personal opinion on the subject.
- I don’t have every detail worked out. It is a work in progress and an invitation to discussion.
- There is much more material than can comfortably fit into this workshop, so do speak with me afterwards if you have questions.

Plan of the workshop

14:00-15:30

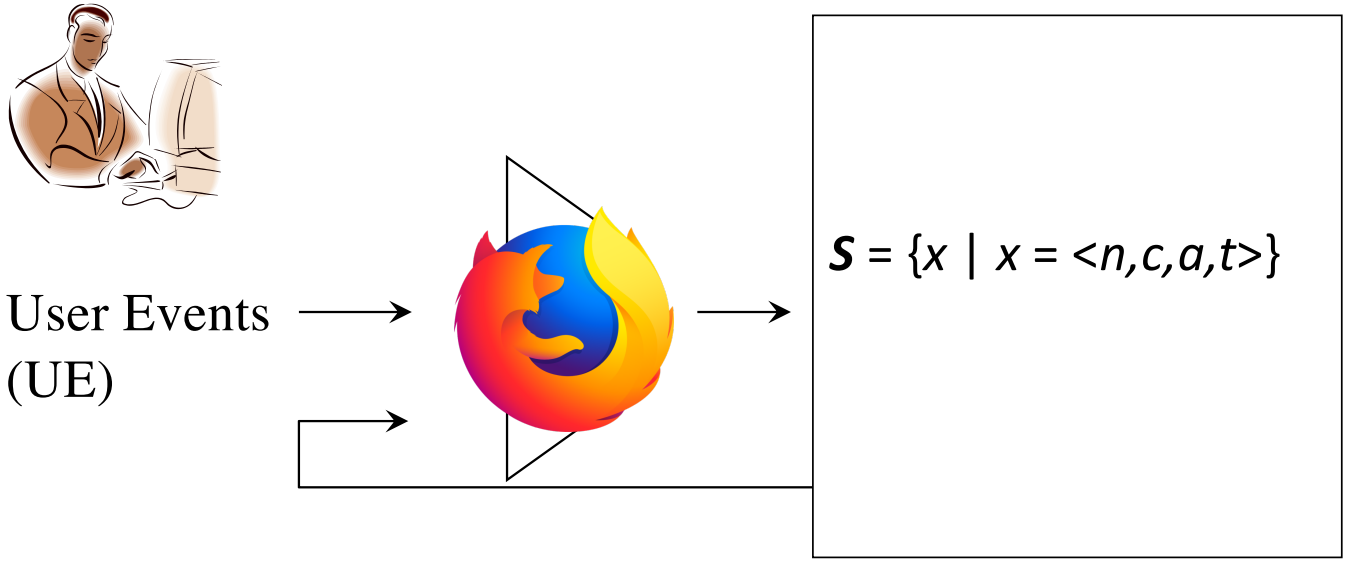
- FSM theory of digital event reconstruction
- Essential concepts from probability theory

15:45-17:00

- Probabilistic / evidential reasoning about algorithms
- Probabilistic reasoning for forensic automation (DeCa)

1. FSM theory of digital event reconstruciton

Making Inferences about User's behavior

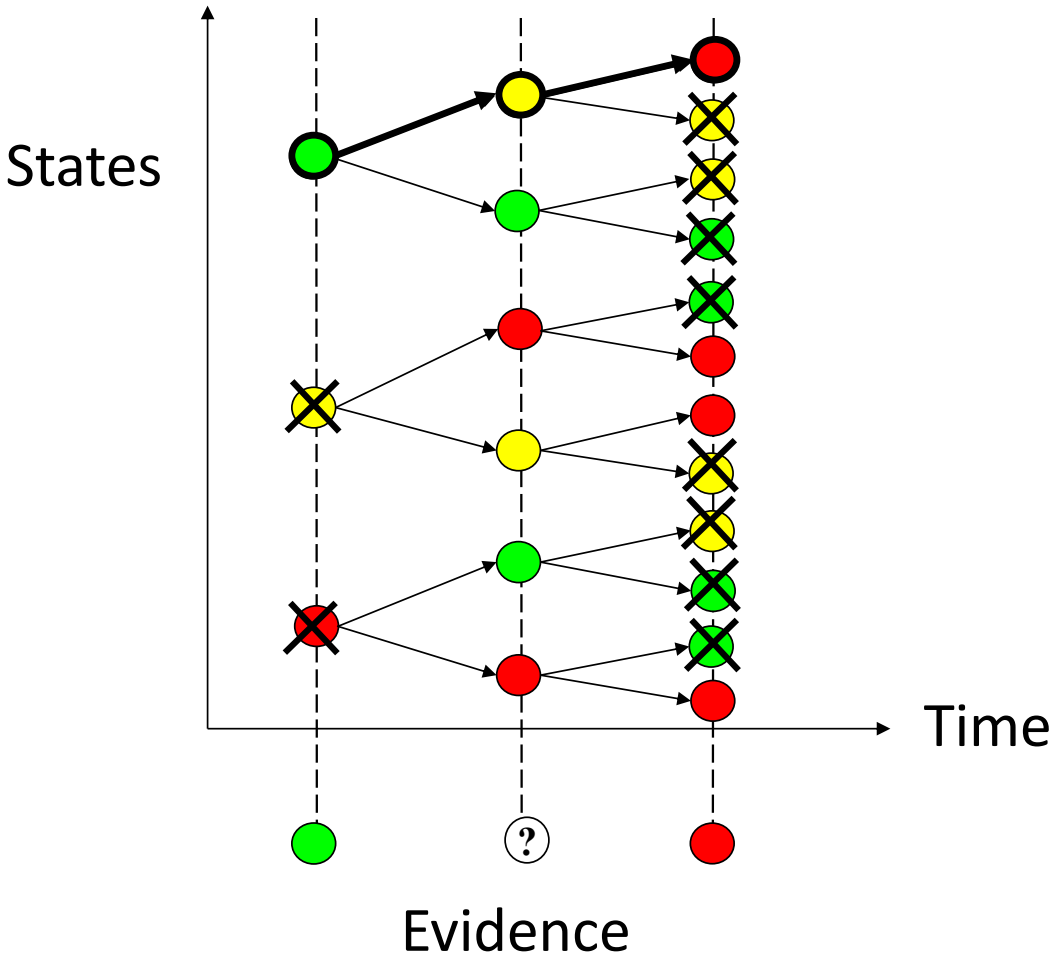


http://www.formalforenscis.org/

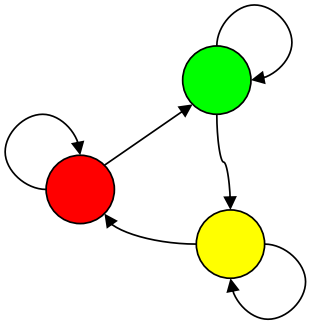
Formalizing evidence

- The incident is modeled as a “computation” – i.e. the sequence of state updates
- Evidence is modeled as restrictions on possible computations.

The naïve approach to finding explanation of evidence



State machine



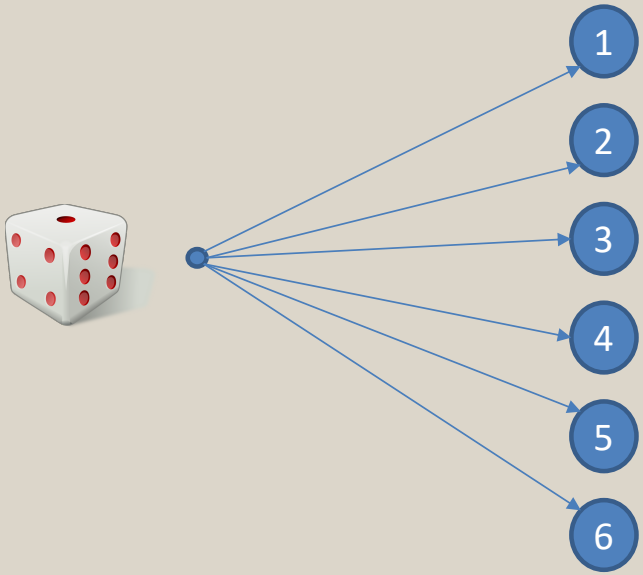
Automated hypothesis testing

1. Finding “explanation” of evidence - the set of all computations of the model that agree with the given evidence.
2. If the resulting set of computations is empty then the evidence contradicts the model
3. Matching the resulting explanation against the hypothesis
4. If the explanation contains at least one computation consistent with the hypothesis, the hypothesis **is plausible.**

2. Essential concepts of probability theory

Origins of the word “event”

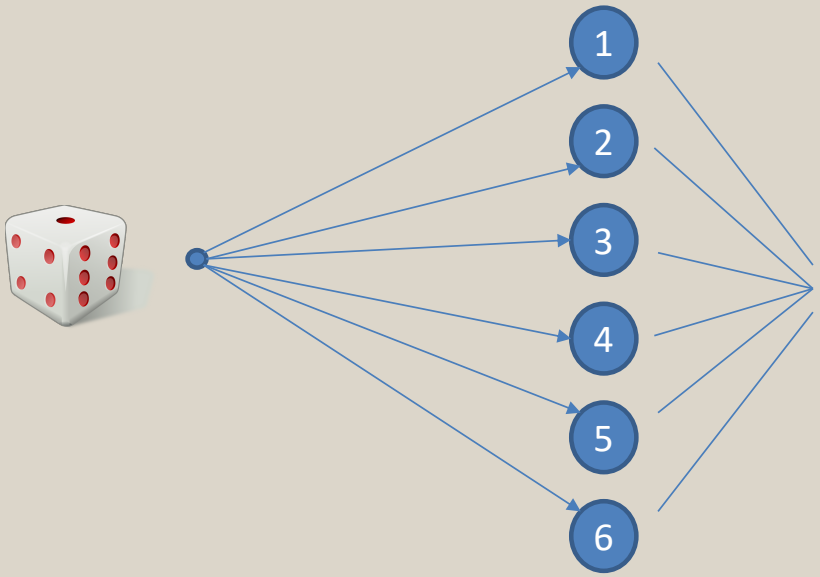
- The word *event* originated in the late 16th century from Latin *eventus* from *evenire* ‘result, happen’, from *e*-(variant of *ex*-) ‘out of’ and *venire* ‘come’.



“event” = “outcome”

Sample space and primitive events

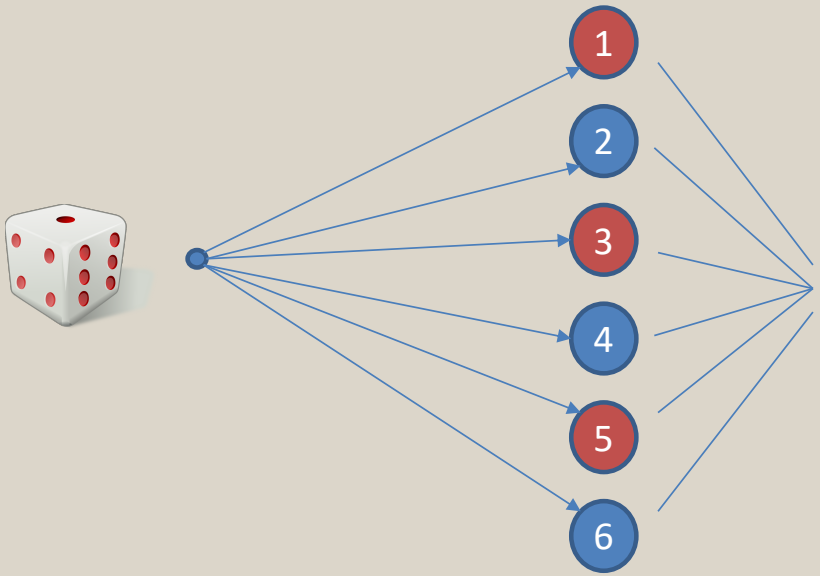
The **sample space** Ω is the set of all mutually exclusive and exhaustive **primitive events** that may result from a game of chance



$\Omega = \{1,2,3,4,5,6\}$

Events

An **event** $E \subseteq \Omega$ is a subset of primitive events that possess certain property (e.g. even, odd, red, blue, etc.). The set of all possible events is denoted $\mathcal{F}$.



Ω = {1,2,3,4,5,6}

Even = {2,4,6}

Odd = {1,3,5}

Primitive events / outcomes

Mutually exclusive events

- Mutually exclusive events do not have primitive events in common

$\Omega = \{1,2,3,4,5,6\}$

Even $= \{2,4,6\}$

Odd $= \{1,3,5\}$

$$\text{Even} \cap \text{Odd} = \emptyset$$

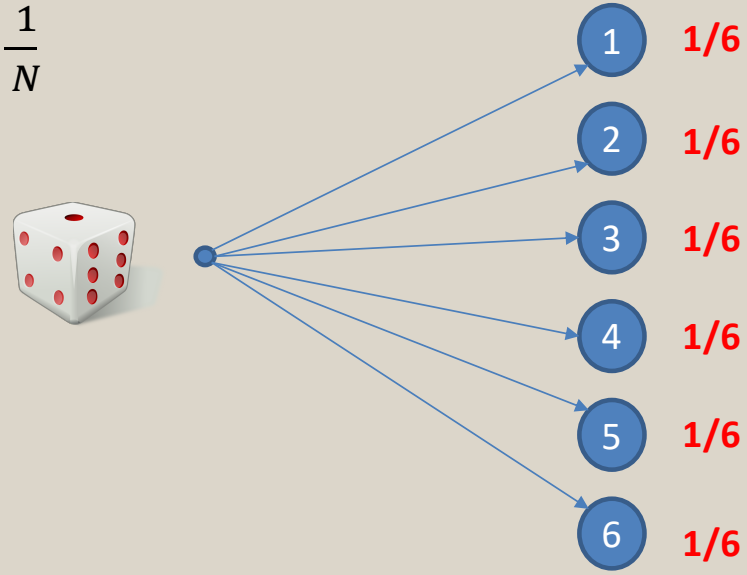
Probabilty

Probability is a degree of certainty and differs from absolute certainty as a part differs from the whole. If, for example, the whole and absolute certainty — which we designate by the letter a or by the unity symbol 1 — is supposed to consist of five probabilities or parts, three of which stand for the existence or future existence of some event, the remaining two standing against its existence or future existence, this event is said to have 3/5 a or 3/5 certainty.

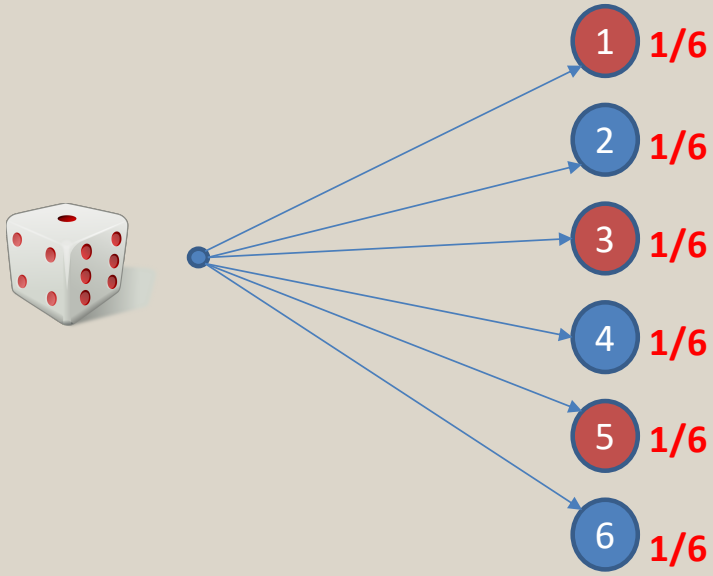
J. Bernoulli (*Ars Conjectandi*, translated by Bin Sung)

Classical definition of probability

- For a ***fair*** of game of chance with N possible, mutually exclusive, and equally likely outcomes $o_1, o_2, \dots, o_N$.
- Sample space $\Omega = \{o_1, o_2, \dots, o_N\}$,
- The probability of a primitive event $P(o_i) = \frac{1}{N}$



Probability of an event



$\Omega = \{1,2,3,4,5,6\}$

Even = {2,4,6}

Odd = {1,3,5}

$$P(\text{Odd}) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

$$P(\text{Even}) = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Probability of an event

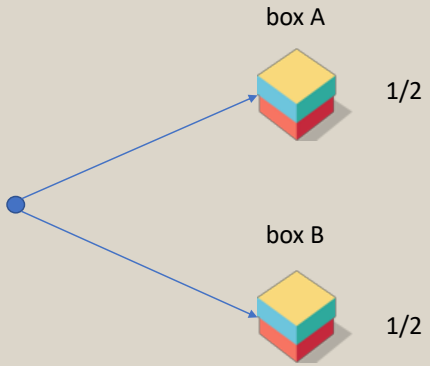
- **Probability $\boldsymbol{P}$** of an event $E \in \mathcal{F}$ is a real-valued function $0 \leq P(E) \leq 1$. Probability function has the following properties:
  - $P(\Omega) = 1$
  - $P(\emptyset) = 0$
  - Given a set of mutually exclusive events $S = \{E_i\}$, the probability of a *union* of these events is the sum of probabilities of individual events $P\left(\bigcup_{E_i \in S} E_i\right) = \sum_{E_i \in S} P(E_i)$.

Problem 1

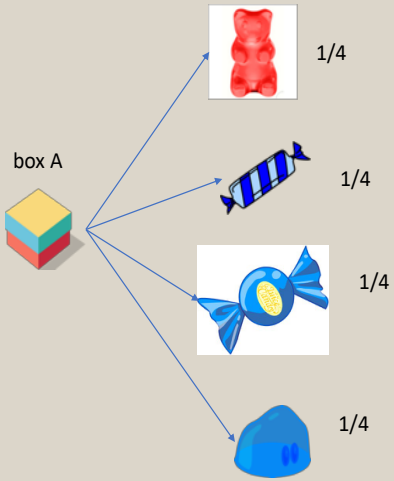
There are two identically looking boxes: A and B. Box A contains 1 red candy and 3 blue candies. Box B contains 2 red candies and 3 blue candies. One box is chosen at random and a candy is drawn from it. Assuming that the candy drawn is red, determine the probability that box A was chosen in the first step.

Modeling probabilistic problems as sequences of random events

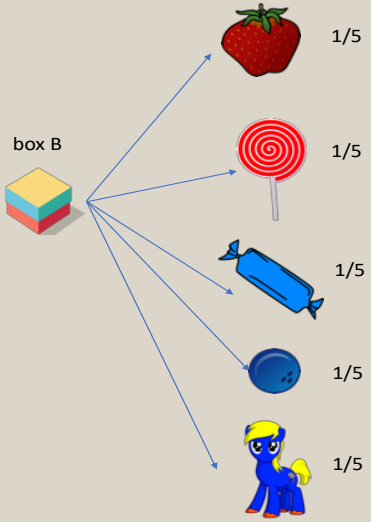
Step 1: Choose a box



Step 2: Pull candy from the chosen box

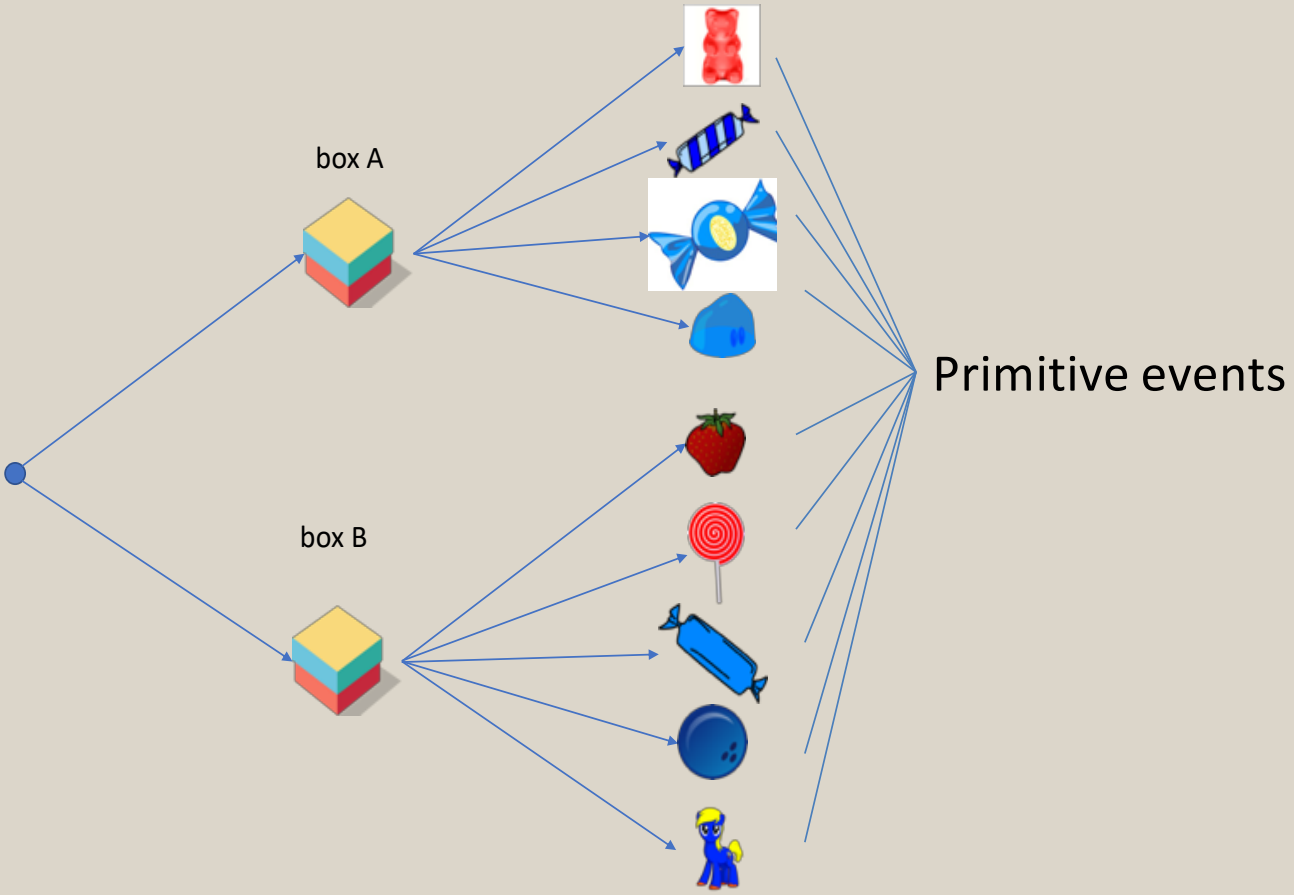


OR



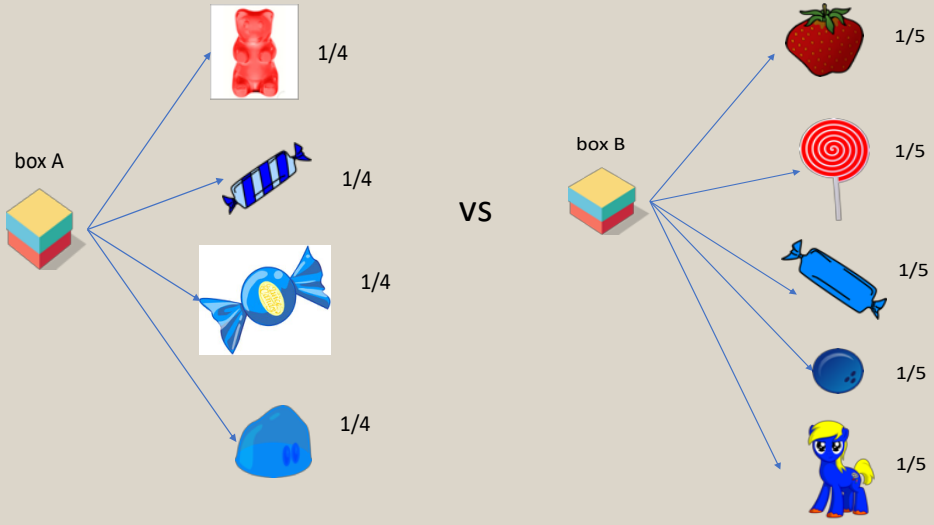
Note: the game “played” in the step 2 depends on the outcome of the step 1 !

Combined game



Conditional probability

- **Conditional probability** $\boldsymbol{P(A|B)}$ is the probability of some event A if it is known that another event B has occurred.

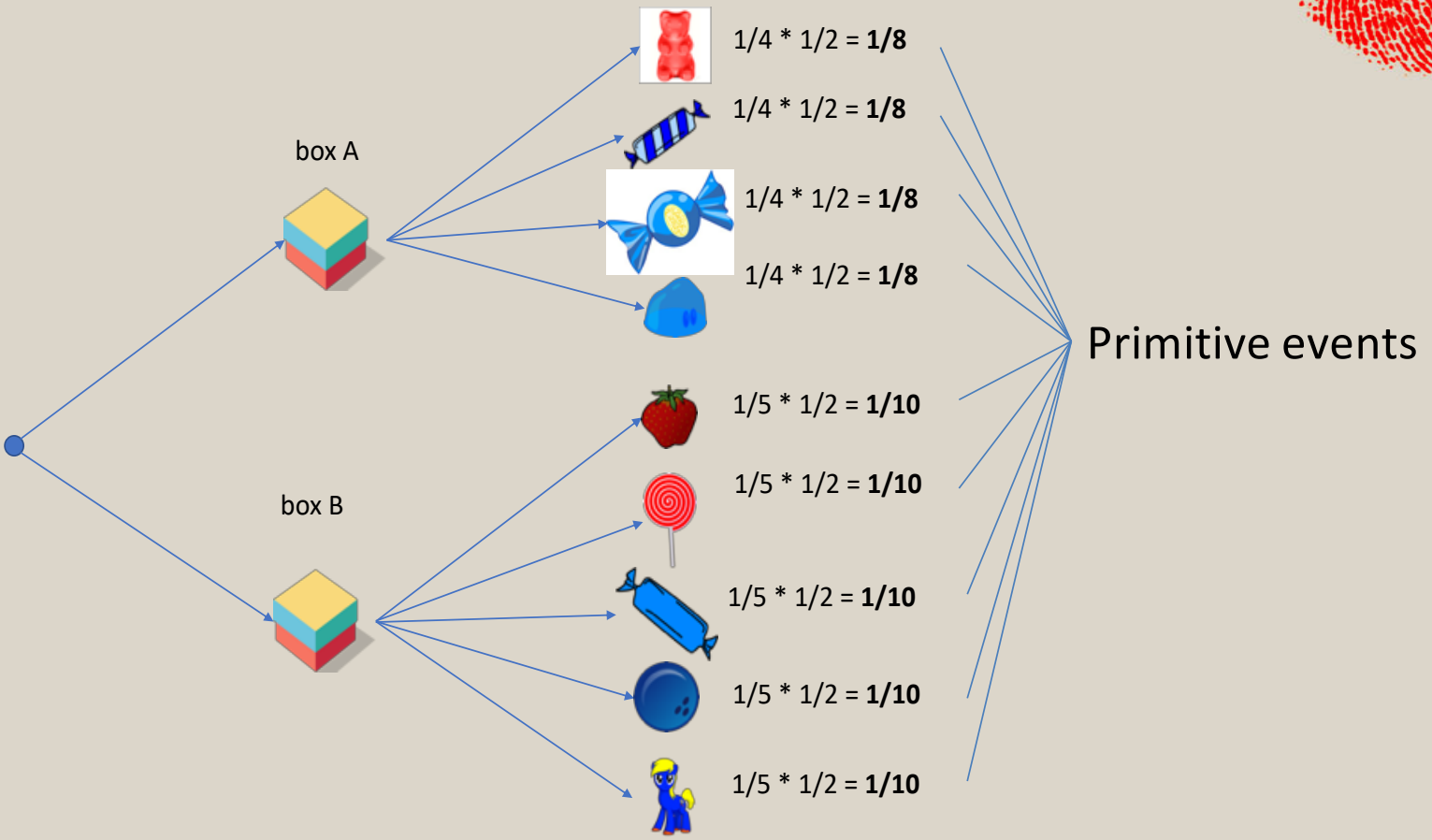


Chain rule

For two random events A and B:

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Combined game



Joint Probability Table

Outcome of step 1	Outcome of step 2	Elementary event probability	Pulled red candy?
box A 		$P(boxA) \times P(candyA1 \mid boxA) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	**True**
box A 		$P(boxA) \times P(candyA2 \mid boxA) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	False
box A 		$P(boxA) \times P(candyA3 \mid boxA) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	False
box A 		$P(boxA) \times P(candyA4 \mid boxA) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	False
box B 		$P(boxB) \times P(candyB1 \mid boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	**True**
box B 		$P(boxB) \times P(candyB2 \mid boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	**True**
box B 		$P(boxB) \times P(candyB3 \mid boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	False
box B 		$P(boxB) \times P(candyB4 \mid boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	False
box B 		$P(boxB) \times P(candyB5 \mid boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	False

$$P(pulledRedCandy) = P(boxA, candyA1) + P(boxB, candyB1) + P(boxB, candyB2) = \frac{1}{8} + \frac{1}{10} + \frac{1}{10} = 0.325$$

$$P(boxA) = P(boxA, candyA1) + P(boxA, candyA2) + P(boxA, candyA3) + P(boxA, candyA4) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = 0.5$$

Evidential reasoning

- The function of evidence in reasoning is to restrict possibilities.
- When some evidence about the outcome of a game of chance is introduced, it *means* that elementary events *inconsistent* with the evidence *could not have happened*, and their probability must now be changed to 0.
- The total probability of the entire sample space, however, must still be equal to 1, so the remaining probabilities must be scaled up.

Calculating posterior probabilities

- If we simply change probabilities of all inconsistent elementary events in the Joint Probability Table to 0, the sum total of the remaining probabilities in the table will be exactly $P(evidence) \leq 1$.
- .We want the sum total of probabilities in the new table to be equal to 1, so we need to multiply all probabilities in the new table by some scaling factor k such that

$$P(evidence) \times k = 1$$

therefore,

$$k = \frac{1}{P(evidence)}$$

Procedure for updating joint probability table

1. Calculate scaling factor $k = \frac{1}{P(Evidence)}$
2. Zero out probabilities of all primitive event in the table, which are inconsistent with the evidence
3. Multiply probabilities of all remaining primitive events by k

Updated Joint Probability Table

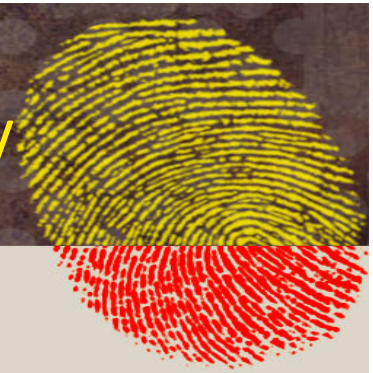
Outcome of step 1	Outcome of step 2	Old elementary event probability (prior)	Pulled red candy?	New elementary event probability (posterior)
box A 		$P(boxA) \times P(candyA1\|boxA) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	**True**	$\frac{1}{8} \times \frac{1}{0.325} \approx 0.3846$
box A 		$P(boxA) \times P(candyA2\|boxA) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	False	0
box A 		$P(boxA) \times P(candyA3\|boxA) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	False	0
box A 		$P(boxA) \times P(candyA4\|boxA) = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$	False	0
box B 		$P(boxB) \times P(candyB1\|boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	**True**	$\frac{1}{10} \times \frac{1}{0.325} \approx 0.3077$
box B 		$P(boxB) \times P(candyB2\|boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	**True**	$\frac{1}{10} \times \frac{1}{0.325} \approx 0.3077$
box B 		$P(boxB) \times P(candyB3\|boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	False	0
box B 		$P(boxB) \times P(candyB4\|boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	False	0
box B 		$P(boxB) \times P(candyB5\|boxB) = \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$	False	0

$$P(pulledRedCandy) = P(boxA, candyA1) + P(boxB, candyB1) + P(boxB, candyB2) = \frac{1}{8} + \frac{1}{10} + \frac{1}{10} = 0.325$$

$$k = \frac{1}{0.325}$$

$$P(boxA) = P(boxA, candyA1) + P(boxA, candyA2) + P(boxA, candyA3) + P(boxA, candyA4) = 0.3846 + 0 + 0 + 0 = 0.3846$$

Challenges with practical application of probability theory



- Need to determine mutually exclusive and exhaustive set of outcomes

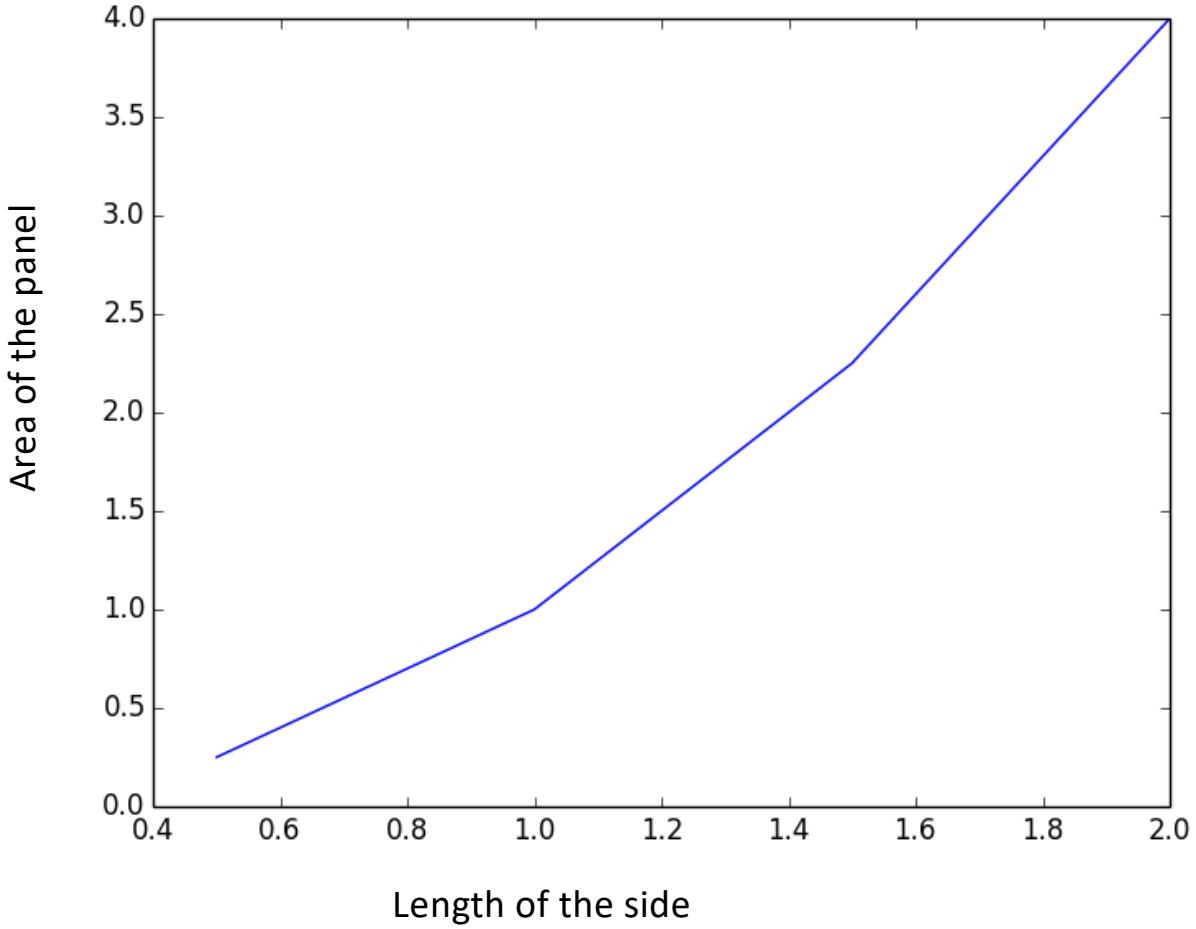
- Need to assign probabilities to outcomes

Problems with classical definition of probability

- Factory making square panels with the length of the side equal either 0.5m, or 1m, or 1.5m, or 2m.

- What is the expected area of a panel on average?

Problems with classical definition of probability



Determining probabilities: frequentist view

- Probabilities can be discussed **only** when dealing with well-defined random experiments (or random samples from a large population)
- Relative frequency of occurrence of an event in a series of trials is a measure of probability of that event. If n_x is the number of trials in which outcome x occurred, and n_t is the total number of trials,

$$P(x) \approx \frac{n_x}{n_t}$$

- As the number of trials approaches infinity the relative frequency converges to *true probability:*

$$P(x) = \lim_{n_t \to \infty} \frac{n_x}{n_t}$$

UCD DUBLIN

Determining probabilities: subjectivist viewpoint

- Probability corresponds to the **personal** belief of the decision maker that that a particular event will occur (or that a particular proposition is true). The stronger the belief, the higher the probability.

- The degree of probability that an individual attaches to a particular outcome can be measured by finding what odds the individual would accept when betting on that outcome.

(source: https://en.wikipedia.org/wiki/Frank_P._Ramsey).

UCD DUBLIN

Exercise 1

- The floor at the crime scene is uniformly covered with fragments of clear window glass and green bottle glass. There are a very large number of fragments, of which 10% are green. A suspect's shoe is found to have two glass fragments embedded in the sole. If the suspect acquired the fragments randomly by walking over the glass, picking up each fragment in turn and assuming each is equally likely to adhere, determine
  - the probability of there being two clear fragments
  - the probability of there being two green fragments
  - the probability of there being one fragment of each color

Exercise 2

- A pub is frequented by some individuals who engage in fights involving assault on each other with their beer glasses. On the other hand, the bar staff are careless in handling the glassware and often drop glasses whilst serving or washing up. Over a period of time, observation leads to the information that 4% of the customers engage in fights and 65% of fights involve glass breakages. In addition, accidental glass breakage is found to occur when serving 15% of customers.
  - What is the probability that some broken glass will result from a customer going to that pub?
  - Given that some glass evidence is found, what is the probability that it was a result of a fight?