Lecture 1

Basics of randomness

Plan of the Lecture

The various nuances of “unpredictable”

1 Defining words

2-3 History of randomness

4 The fair coin

5-6 Definition of randomness

THREE WORDS

An exercise in analogy

Randomness

Lack of predictability

- Example 1: coin tossing
  - Ignorance in each toss
  - lack of pattern in the sequence (normally)
- Example 2: polling
  - We ignore which item will be picked
- Single events vs. statistics

Chance

Usually it refers to “coincidence” that is **concurrent uncorrelated causes**

- – T. Wilder, *The Bridge of San Luis Rey*.

- Desire to see a deeper cause beneath the coincidence, especially when tragic
- Even if both causes are not “random”, their conjunction may be unpredictable.

Free will

Usually it refers to **our choices**, especially in a moral context.

- Example: “nobody is looking at me, should I steal this object?”
  - The decision is not “random” for me…
  - … but is unpredictable for someone who does not know me.

Common element

Lack of information (⇒ of control)

- I don’t know exactly how the coin is going to be tossed
- I can’t imagine all that happens around me and that may affect me
- I can’t predict the behavior of the others

This course

We deal with

- Mathematical definition of randomness
- Randomness as a resource (a.k.a. when unpredictable is good)
- Sources of randomness in the physical world
- Science on free will

We don’t deal with

- Chance in evolution
- Statistics, decision making
- Prediction in finance, politics, sports
- “Destiny”, “fate” etc.
- Randomness and free will in religions

Suggested Readings

On probabilities and statistics:

- Deborah J. Bennett, Randomness (Harvard University Press, Cambridge MA, 1998)
- Leonard Mlodinow, The drunkard's walk (Pantheon, New York, 2008)

On chance: almost any book or newspaper article deals with it.

- Thornton Wilder, The bridge of San Luis Rey, 1927: http://en.wikipedia.org/wiki/The_Bridge_of_San_Luis_Rey

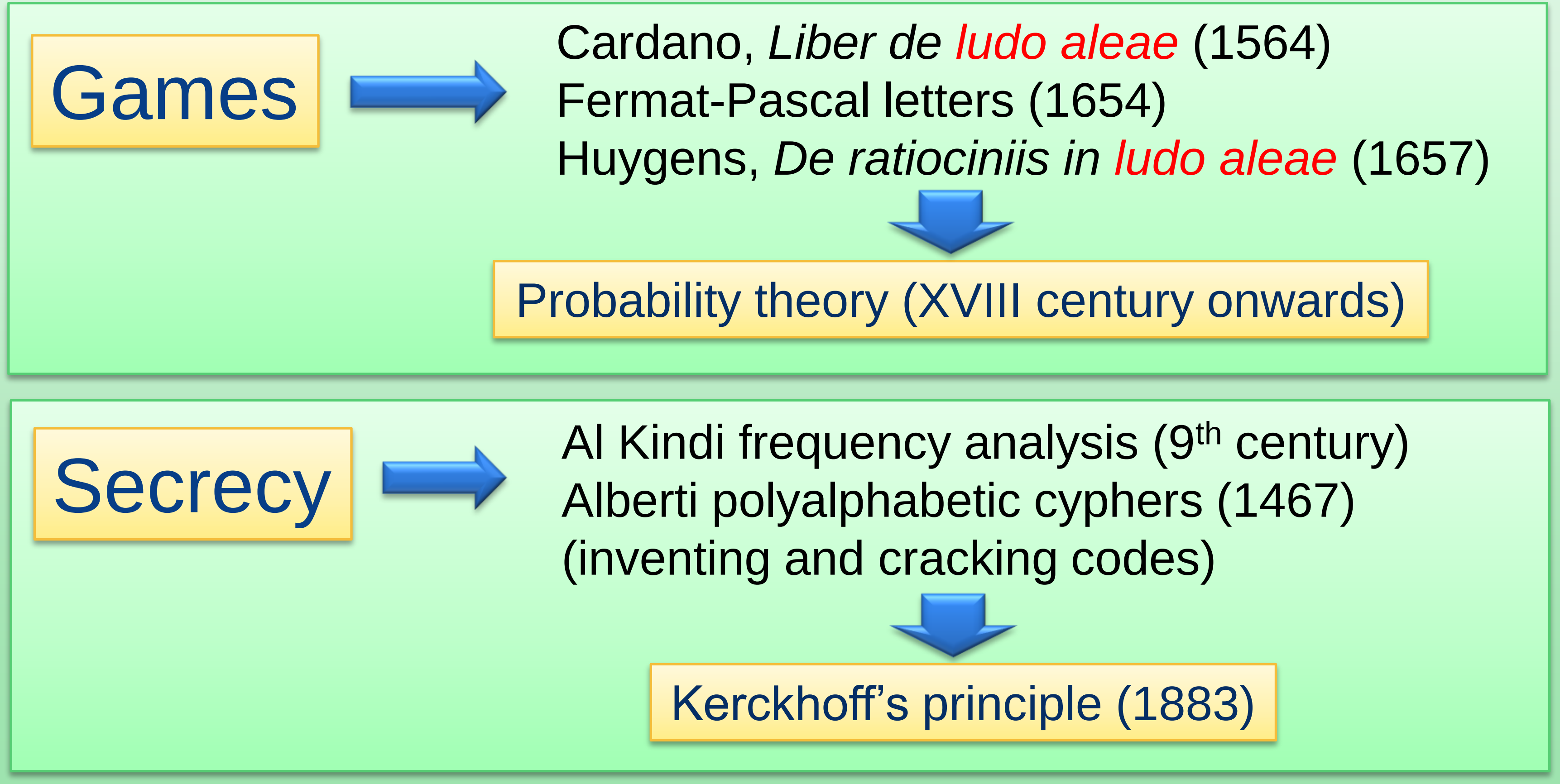
HISTORY OF RANDOMNESS (1)

Fast forward to the 19th century

Since Antiquity

Games	Ignorance ⇒ fairness
Secrecy	Ignorance by the adversary
Divination	… but someone must know (and thus be in control)

Mathematics begin



Games

Cardano, *Liber de ludo aleae* (1564)
Fermat-Pascal letters (1654)
Huygens, *De ratiociniis in ludo aleae* (1657)

Probability theory (XVIII century onwards)

Secrecy

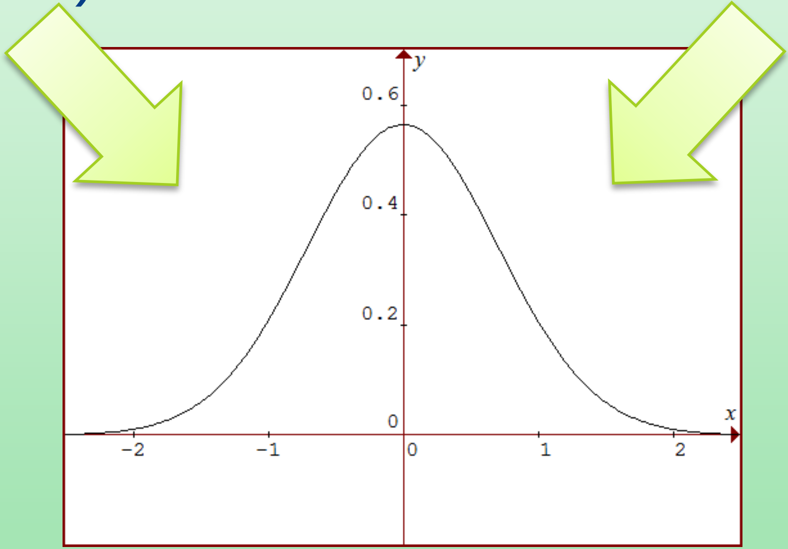
Al Kindi frequency analysis (9th century)
Alberti polyalphabetic cyphers (1467)
(inventing and cracking codes)

Kerckhoff's principle (1883)

The bell curve

Tossing many coins (De Moivre, 1733)

Distribution of errors in measurements (Laplace, 1812)



“Certainty in uncertainty”: We can’t know the exact result, but we can quantify rigorously how far we may be from it

Very frequent… but not universal (assumptions go into it)

Laplace, *Théorie analytique des probabilités*, 1812

« We may regard the present state of the universe as the effect of its past and the cause of its future.

An intellect which at a certain moment would **know all forces** that set nature in motion, **and all positions** of all items of which nature is composed,

if this intellect were also **vast enough to submit these data to analysis**, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom;

for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes. »

End of 19th century

- Physics: all is predictable
- Statistics: basics solidly established

« Of all objects, the planets are those which appear to us under the least varied aspect. We see how we may determine their forms, their distances, their bulk, and their motions, but *we can never known anything of their chemical or mineralogical structure*. »

August Comte, *The Positive Philosophy*, Book II (1842)

« *The more important fundamental laws and facts of physical science have all been discovered*, and these are so firmly established that the possibility of their ever being supplanted in consequence of new discoveries is exceedingly remote. »

Albert Michelson, *Light waves and their uses* (1903)

Suggested Readings

Artur Ekert, Complex and unpredictable Cardano,
http://arxiv.org/abs/0806.0485

Simon Singh, The Code Book, 1999:
http://en.wikipedia.org/wiki/The_Code_Book

Wikipedia pages:

- http://en.wikipedia.org/wiki/History_of_randomness
- http://en.wikipedia.org/wiki/Probability
- http://en.wikipedia.org/wiki/Normal_distribution
- http://en.wikipedia.org/wiki/Jia_Xian

HISTORY OF RANDOMNESS (2)

6 messages from the 20th century

1

Quantum physics

- Precursors ~1900; fully born 1926
- Intrinsic randomness in elementary physics
  - one cannot specify both position and momentum with arbitrary precision
  - NOT a default of measurement
- Lectures 6-7

2 Deterministic chaos

- Laplace: “vast enough to submit these data to analysis” $\Rightarrow$ use computers
- 1970s: “butterfly effect”
  - Discovered studying models of the weather
  - Even simple deterministic dynamics can become unpredictable in the long run
- Lectures 4 & 5

3

Chance in evolution

- As far as science can say, an observer in the past would have had hard time predicting the “rise of humans”:
  - Macroscopic chance events: asteroids, climate changes…
  - Microscopic chance events: mutations
- Not dealt with in this course.

4

Statistics >> predictions

- Statistics are everywhere
  - Polls, insurance, sports…
- Still, we are not in full control
  - Natural: earthquakes
  - Human: finance, terrorism
- Not dealt with in this course

Predict… us?

- Desire to predict human behavior
  - 19th century: recognize criminals from physical traits, calligraphy… ⇒ failure!
  - Maybe undesirable (privacy, control…)
- Libet’s experiments (1980s)
- Lecture 8

6

Information theory

- Turing 1940s, Shannon 1948
- Unpredictability is not always bad ⇒ randomness as a resource
  - Computing: efficient “randomized algorithms”.
  - Privacy, secrecy: e.g. online transactions
- Lectures 2 & 3

Summary

1. Quantum: intrinsic ignorance in physics
2. Even with computers and without quantum, ignorance in the long term
3. Chance in evolution: we ignore how we got here
4. The limits of statistics
5. While unpredictability creeps in nature, our own decisions seem to become more predictable!
6. Randomness as a resource (computing, privacy)

Suggested Readings

Here are some classic popular books on the topics we left out, which can help you to start in thinking

On chance in biological evolution:

- Stephen J. Gould, Wonderful life, 1989: http://en.wikipedia.org/wiki/Wonderful_Life_(book)

On use and misuse of statistics in finance and in life:

- Nate Silver, The signal and the noise, 2012: http://en.wikipedia.org/wiki/The_Signal_and_the_Noise
- Nassim Nicholas Taleb, The Black Swan, 2007: http://en.wikipedia.org/wiki/The_Black_Swan_(2007_book)

THE FAIR COIN

Definition

“Fair coin” = $\left\{ \begin{array}{l} \text{• Alphabet: two values \{0,1\} = bit} \\ \text{• Single run (toss): fully unpredictable} \\ \text{• Several runs: uncorrelated} \end{array} \right.$

Remark: “Ideal” as a resource, not “exemplary” or “universal”

Single run:

$$P(0) = P(1) = \frac{1}{2}$$

Sequence of *n* runs:

$$P(d_1 d_2 \dots d_n) \equiv P[n] = \frac{1}{2^n}$$

$$\boxed{n = -\mathbf{log_2}\, P[n]}$$

Change alphabet (1 die = 2.6 coins)

“Fair die” = $\left\{ \begin{array}{l} \text{• Alphabet: six values \{1,2,3,4,5,6\}} \\ \text{• Single run (cast): fully unpredictable} \\ \text{• Several runs: uncorrelated} \end{array} \right.$

Single run:

$$P(1) = \cdots = P(6) = \frac{1}{6}$$

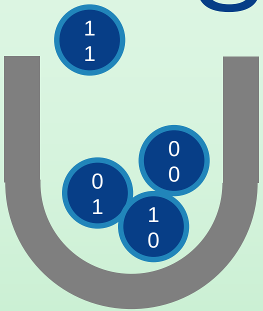
Sequence of *m* runs:

$$P(b_1 b_2 \dots b_m) \equiv P_6[m] = \frac{1}{6^m}$$

Bits: $n = -\log_2 P_6[m] = m \log_2 6 \approx 2.585\, m$

Alphabet of size A: multiply by $\log_2 A$ to get bits

Statistics of sequences of bits



Each sequence of length n is drawn with probability $P[n] = \frac{1}{2^n}$

- one sequence consisting of n 0s $\Rightarrow P(n\,\mathbf{0}, 0\,\mathbf{1}) = \frac{1}{2^n}$
- n sequences with only one 1 $\Rightarrow P(n-1\,\mathbf{0}, 1\,\mathbf{1}) = n\frac{1}{2^n}$
- n-choose-k sequences with k 1s $\Rightarrow P(n-k\,\mathbf{0}, k\,\mathbf{1}) = \binom{n}{k}\frac{1}{2^n}$

$\Rightarrow$ The sequences with n/2 0 and 1 are the most numerous.

Their number is $\binom{n}{n/2} \approx 2^n\sqrt{\frac{2}{\pi n}} \approx 2^n$ for $n \to \infty$.

Long sequences have almost certainly n/2 0 and n/2 1

For those who are interested in deriving the approximation

$$\binom{n}{n/2} \approx 2^n \sqrt{\frac{2}{\pi n}}$$

, you can check out Stirling approximation , which is used to expand

$$\binom{n}{n/2} = \frac{n!}{((n/2)!)^2}$$

(Note that one has to go to the third term of the series for a non-trivial result)

Statistics within a sequence

…001010011111111100101010001110101001111001110100010111010100111…

How are **d-bits strings** distributed within a sequence of n bits?

Rule of thumb: all the strings with length d satisfying $d2^d<n$ should appear in a sequence of length n.

Example: n = 4096

- Number of 0 and 1: $m_1 = 2084 \pm 32$
- Number of 00, 01, 10 and 11: $m_2 = 512 \pm 20$
- Number of 000, 001,… and 111: $m_3 = 170 \pm 12$
- Number of 0000, 0001,… and 1111: $m_4 = 64 \pm 8$
- Etc. until $m_8 = 2 \pm 2$ (indeed $8 * 2^8 = 2048, 9 * 2^9 = 4608$)

Proof

Lemma

Let $\delta_{j=v} = \begin{cases} 1, v\ in\ run\ j \\ 0, otherwise \end{cases}$ be the indicator of value v of the alphabet. Then $\overline{\delta_v}$ is the probability of drawing the value v in each run; since $\delta_{j=v}^2 = \delta_{j=v}$, it holds $\overline{\delta_v^2} = \overline{\delta_v}$. It follows that an estimator of the average on N runs will yield $\langle \delta_v \rangle_n = \overline{\delta_v} \pm \frac{\sigma}{\sqrt{N}}$ with $\sigma = \sqrt{\overline{\delta_v}(1 - \overline{\delta_v})}$.

Proof

Drawing a sequence of n bits is equivalent to drawing N=n/d strings of d bits; for a fair coin, each such string has probability $\overline{\delta_d} = \frac{1}{2^d}$.

By plugging these observations in the Lemma, and with a bit of algebra, one finds that the number of appearances of any string of length d obeys

$$\langle m_d \rangle_{n\ bits} = \frac{n}{d}\frac{1}{2^d}\left(1 \pm \sqrt{\frac{2^d - 1}{n/d}}\right)$$

The randomness dilemma

0000000000000000
does not “look random”

1001001111101100
“looks random”

but for a fair coin, they are equally probable to be drawn!

How to define randomness

- From the probabilities (mathematical description)
- Capturing what we believe “random” to mean

Check out this cartoon of Dilbert:
http://dilbert.com/strips/comic/2001-10-25/

DEFINE RANDOMNESS (1)

Randomness of a sequence

Kolmogorov complexity (KC) a.k.a. algorithmic complexity

Complexity of a sequence = the length of the shortest algorithm that can generate it.

0000000000000000	Print a list of sixteen 0
1001001111101100	Print *twice a 1 followed by two 0*, then five 1, one 0, two 1, two 0

Captures: randomness = lack of pattern

Martin-Löf definition (1966)

“Algorithmic randomness”

A *sequence* is random if [equivalent statements]:

- It passes all possible **statistical tests** (with fair coin as ideal case)
- It cannot be produced by a program shorter than itself (“**incompressible**”)
  - Inspired by Kolmogorov complexity
  - 0000000, or the digits of π, are not random in this sense

Problem: KC is uncomputable

Not just "difficult to compute": there is no consistent way of defining "the shortest algorithm".

Similar to *Berry's paradox*:

- Hypothesis: positive integers can be ordered by the smallest number of English words needed to describe each of them.
- Then "the smallest positive integer not definable in less than N English words" can then be described in 13 words $\Rightarrow$ paradox for N>13

Sketch of formal proof

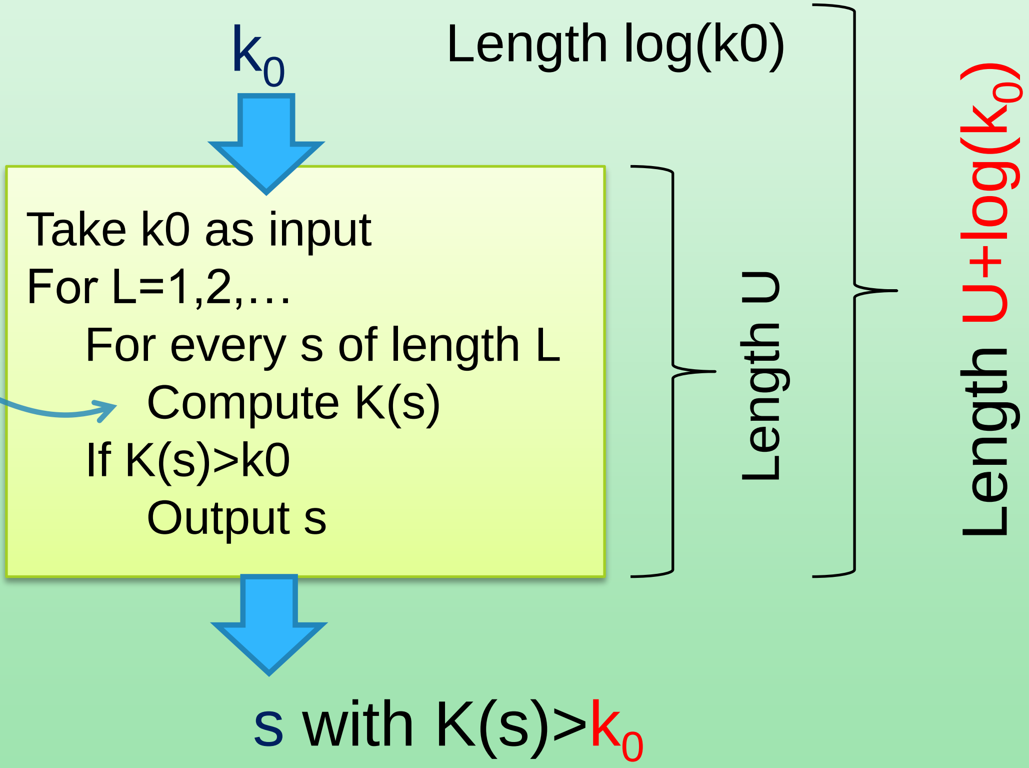
Hypothesis

$\exists\, K(s)$ that computes KC for every string s

$$\Downarrow$$

$$\log k_0 + U > k_0$$

Wrong for k_0 large enough



k_0

Length log(k0)

```
Take k0 as input
For L=1,2,…
    For every s of length L
        Compute K(s)
    If K(s)>k0
        Output s
```

Length U

Length U+log(k_0)

s with K(s)>k_0

Suggested Readings

The NIST 800-22 battery of statistical tests for random number generators:
http://csrc.nist.gov/publications/nistpubs/800-22-rev1a/SP800-22rev1a.pdf

Wikipedia pages:

- http://en.wikipedia.org/wiki/Kolmogorov_complexity
- http://en.wikipedia.org/wiki/Martin-Lof
- http://en.wikipedia.org/wiki/Statistical_randomness
- http://en.wikipedia.org/wiki/Algorithmic_randomness

DEFINE RANDOMNESS (2)

Randomness of the process

Unpredictable process

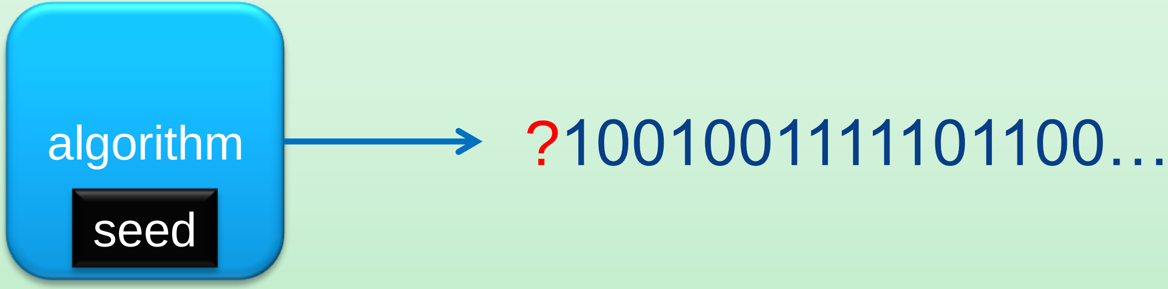
1001001111101100…

Can anyone predict the next one?

Ideal random process if $P(b_n|s_{n-1}) = \frac{1}{2}$

- The next million bits may all be 0…
- … but of course, if I have never seen a 1, I won’t trust the process to be random

Not-too-black boxes



algorithm

seed

?1001001111101100…

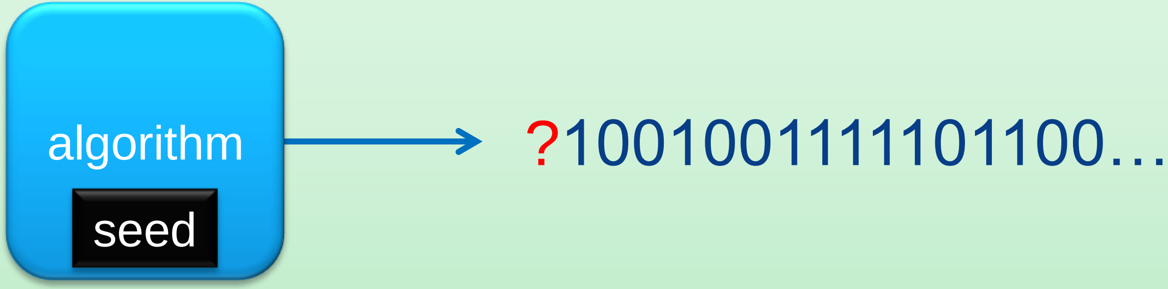
The full black-box model may be too strong:

⇒ Kerckhoff-Shannon: no “security by obscurity”, let the algorithm be known, but not the seed.

Example (not very practical)

Algorithm: “the sequence is drawn from the second letter of each line of a book” ⇒ Pretty unpredictable if one can’t guess *which book*.

Pseudorandomness


algorithm

seed

?1001001111101100…

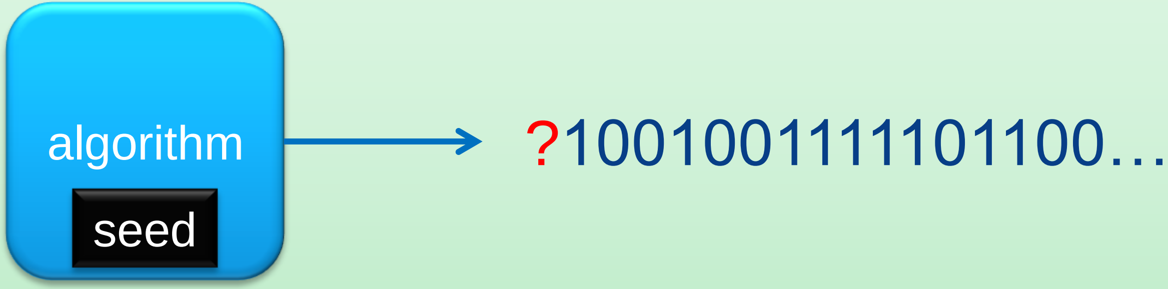
More practical: mechanical algorithms:

- Mathematical function (seed: e.g. some divider)
- Scramble the last digits of the clock of the computer (seed: the precise instant the key is touched)

This is pseudo-random: nothing is *really* random, but it is unpredictable for someone who does not know the seed.

Is it really?

Pseudorandomness (2)



algorithm

seed

?1001001111101100…

Can the seed be inferred by observing the process?
YES it can – but if things are well done, only with high computational power ☺

Very frequent definition: the process is random if it is **unpredictable with limited computational power** [more precise Lectures 2-3]

At a glance

Random sequence

- No pattern ☺
- May not be “secret”: it could have been recorded once and copied many times ☹
- Practice: uncomputable; trust a finite battery of tests

Random process

- If randomness “is being generated”, it is secret ☺
- May (occasionally) generate a “simple” sequence ☹
- Practice: need to trust the characterization of the process

Summary of Lecture 1

The various nuances of “unpredictable”

- Randomness = lack of predictability
  - Lack of information ⇒ lack of control
- Mathematical tool: statistics
  - Fair coin as “ideal” case
- Definition of randomness
  - Of a sequence? Of a process?
  - Verification requires some element of trust