Lecture 5a
Hash Functions II

Stefan Dziembowski
www.crypto.edu.pl/Dziembowski

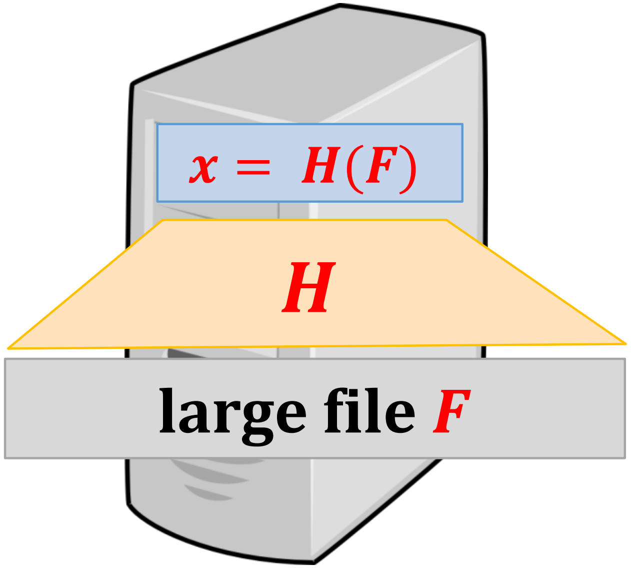
University of Warsaw

24.10.18 **version 1.0**

Plan

1. Other uses of hash functions
   1. Merkle trees
   2. Practical randomness extraction and the random oracle model
   3. Password storage and Proofs of Work
2. Real-life constructions

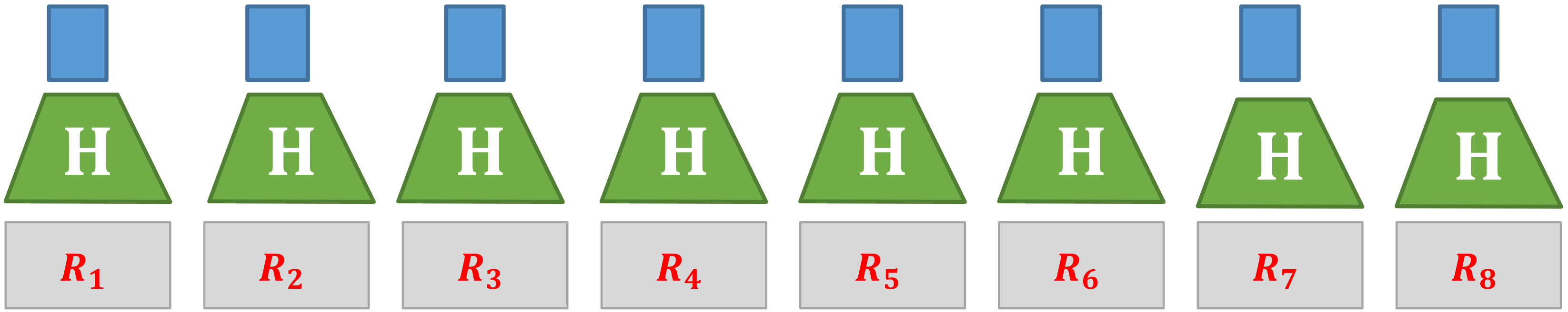
Consider again file fingerprinting



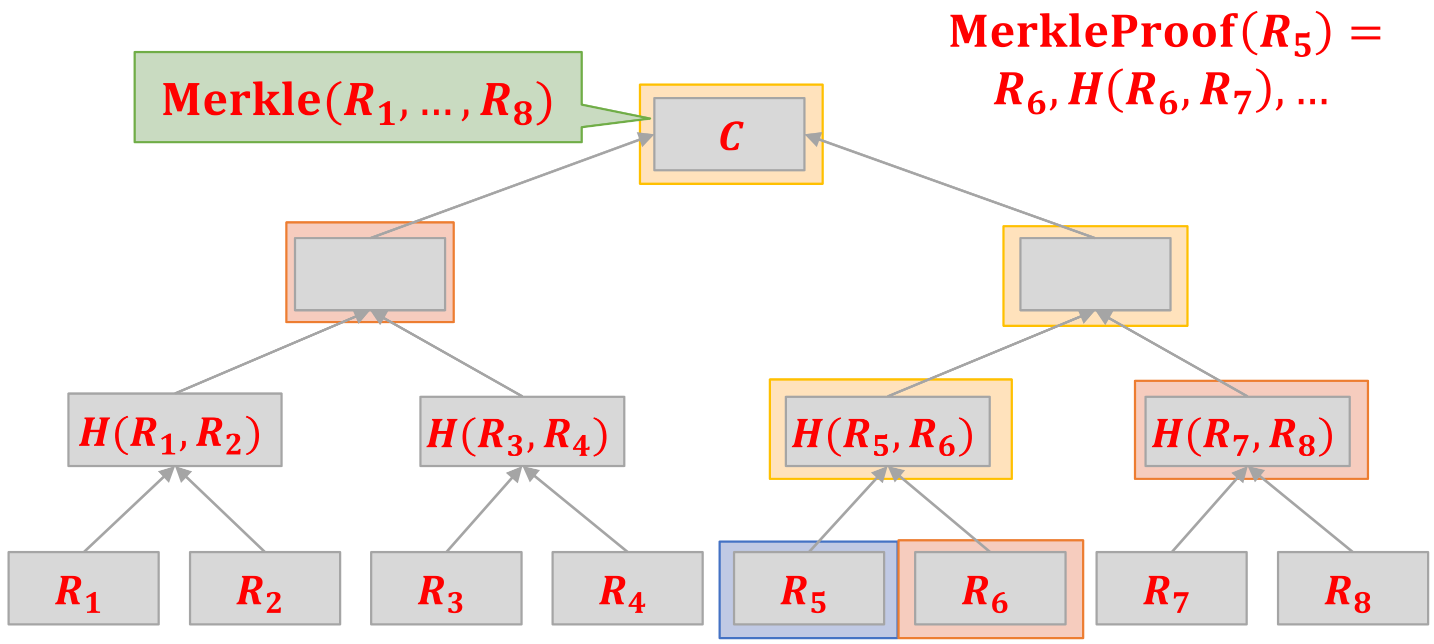
A question

Suppose a file F consists of many smaller blocks $R_1, \ldots, R_n$, and the user may want to access only one of them. How to "fingerprint F"?

Naive solution: fingerprint each of them independently.



Better solution: construct a **Merkle tree:**



Recall: Merkle trees allow to efficiently prove that each block R_i was included into the hash C.
This is done by sending $\mathbf{MerkleProof}(R_i)$.
Easy to see: if H is collision resistant then so is **Merkle**.

File sharing

BitTorrent, **Gnutella**, **Gnutella2**, and **Direct Connect P2P**... - a variant of the idea from the previous slides:

- files in the peer-to-peer networks are **identified by their hashes**
- each file consists of "**pieces**"
- the users download the pieces from each other.
- some of them use **Merkle trees**.

Plan

1. Other uses of hash functions
   1. Merkle trees
   2. Practical randomness extraction and the random oracle model
   3. Password storage and Proofs of Work
2. Real-life constructions

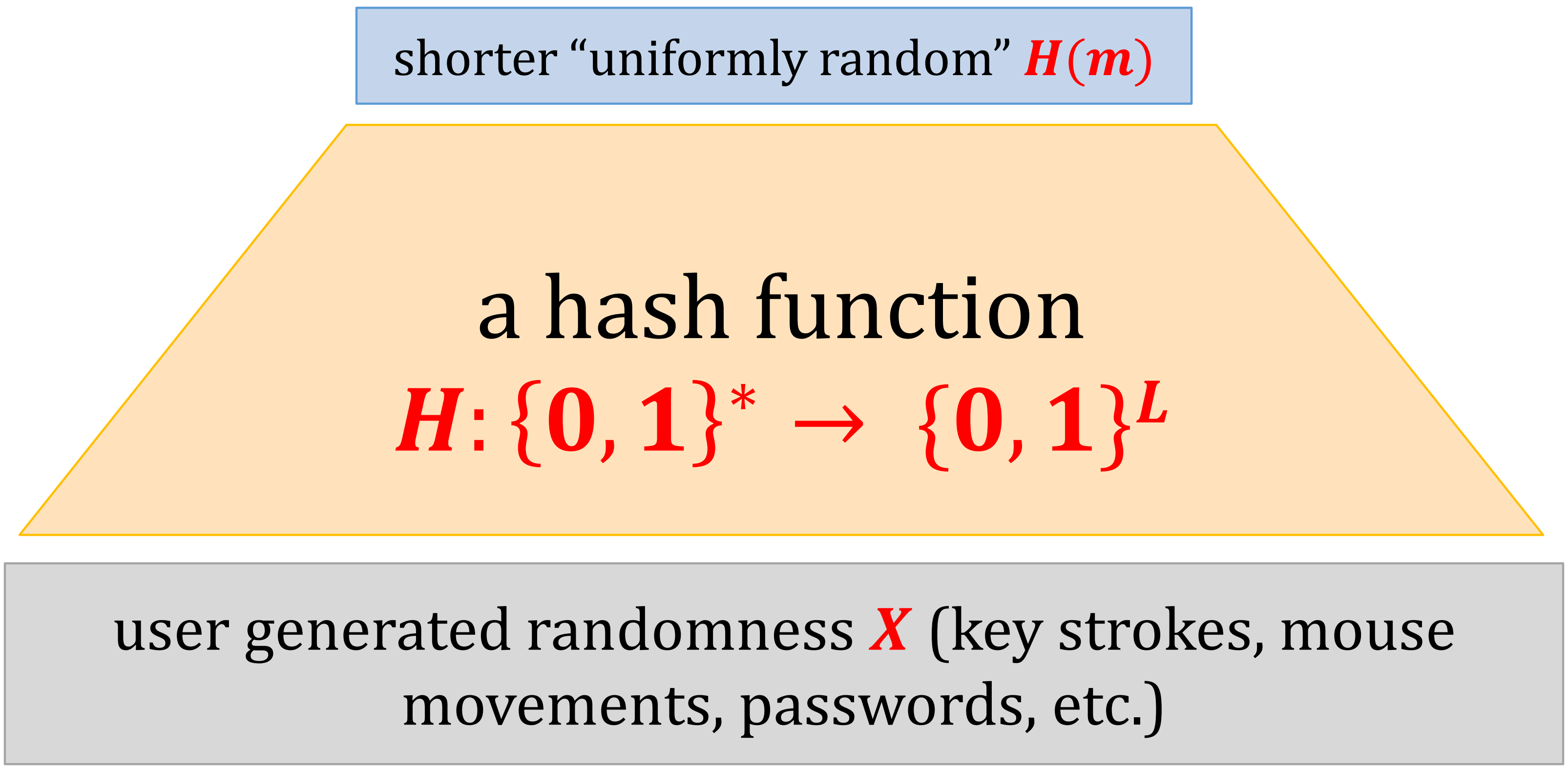
How the outputs of hash functions look in real life?

```
C:\> echo -n `Wydział Matematyki, Informatyki i Mechaniki` | openssl sha1
30428440c00bd45d2e2fd93ed980fbd8aa063428

C:\> echo -n `Wydział Matematyki, Informtyki i Mechaniki` | openssl sha1
9937fe966d988e8163fe07f6a1dbd9caf624e1c8

C:\> echo -n `Wydział Matematyki Informatyki i Mechaniki` | openssl sha1
456b370c5afe5f45c0af4a6290d02f6d2f557381
```

Observation: the outputs on different inputs are "unrelated" and "completely random".

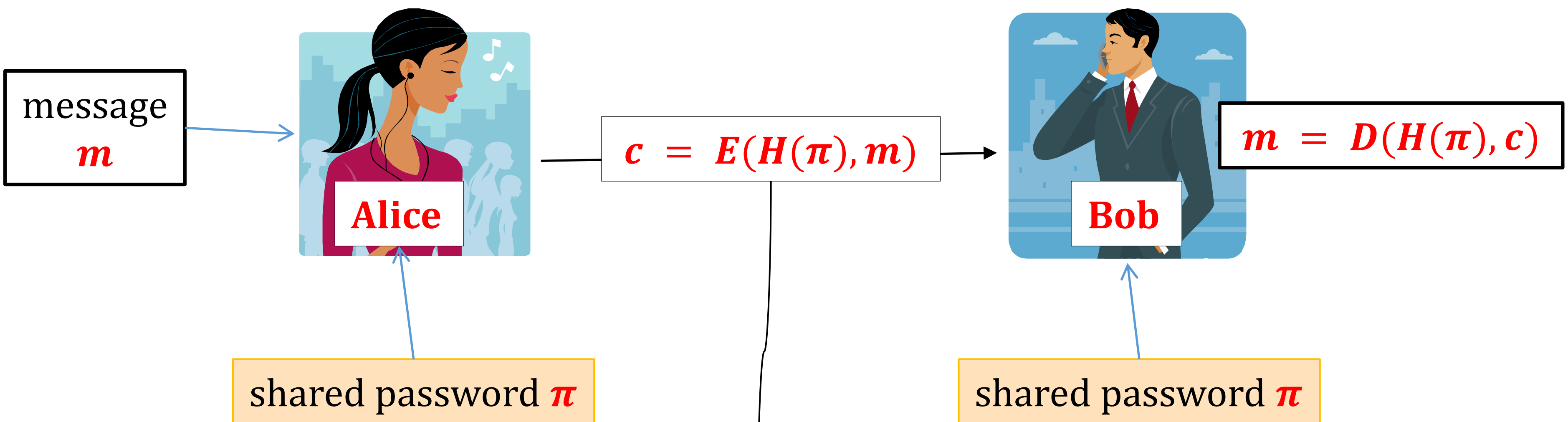
we will formalize this property in a moment

Example of how this property is used: deriving “uniformly random keys” from “non-uniform randomness"



Example: password-based encryption

H – hash function
(E, D) – encryption scheme



Warning:
there exist much better solutions for this problem

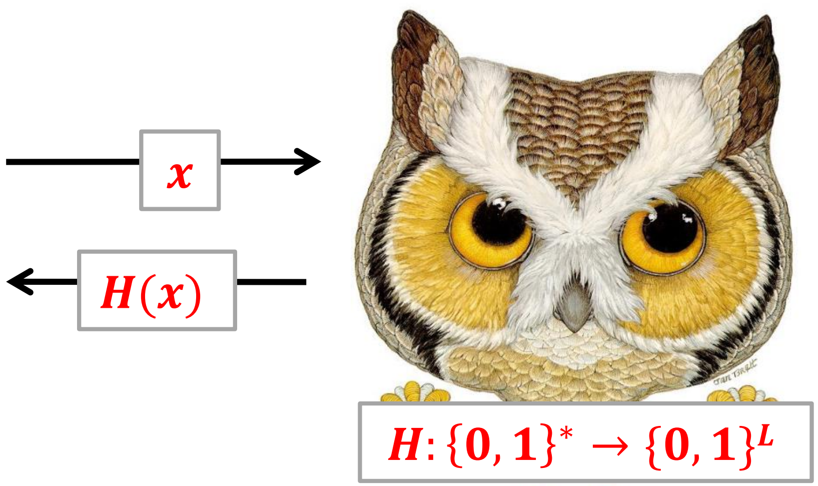
Informally:
The only thing that Eve can do is to examine all possible passwords .

Random oracle model

[Fiat, Shamir: How to Prove Yourself: Practical Solutions to Identification and Signature Problems. 1986]

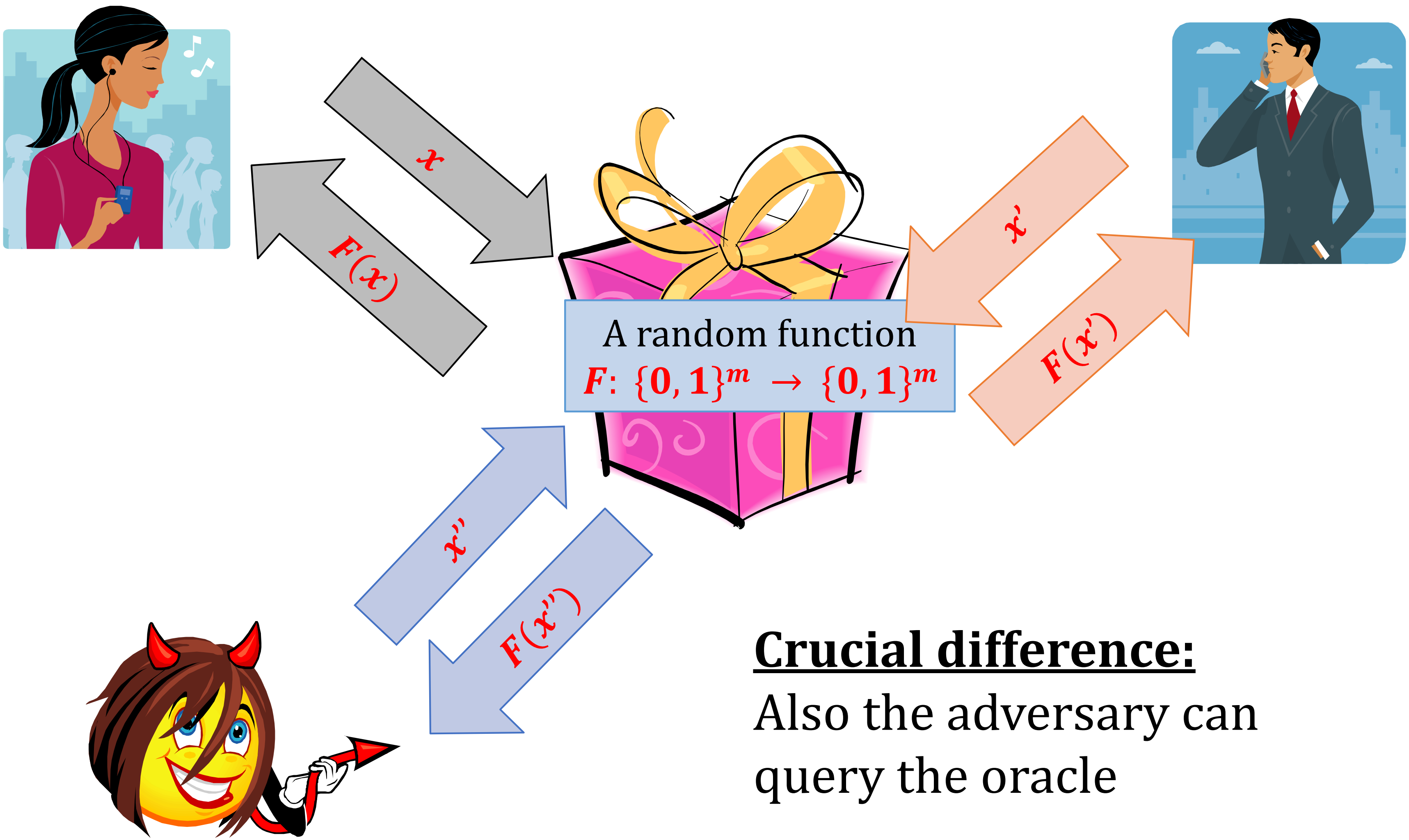
[Bellare, Rogaway: Random Oracles are Practical: A Paradigm for Designing Efficient Protocols, 1993]

Idea: model the hash function as a **random oracle**.



a completely random function

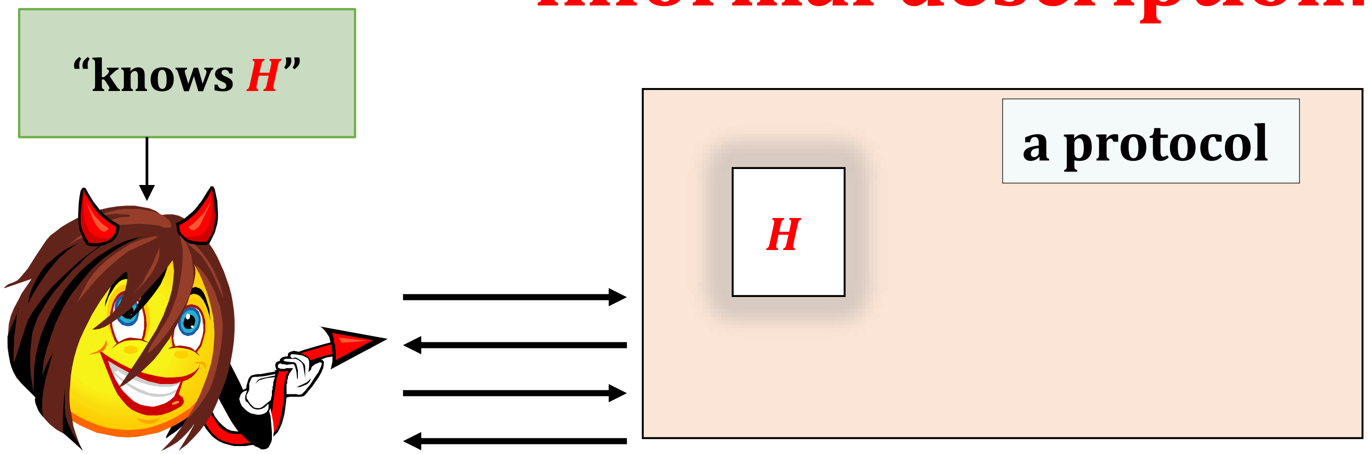
Remember the pseudorandom functions?



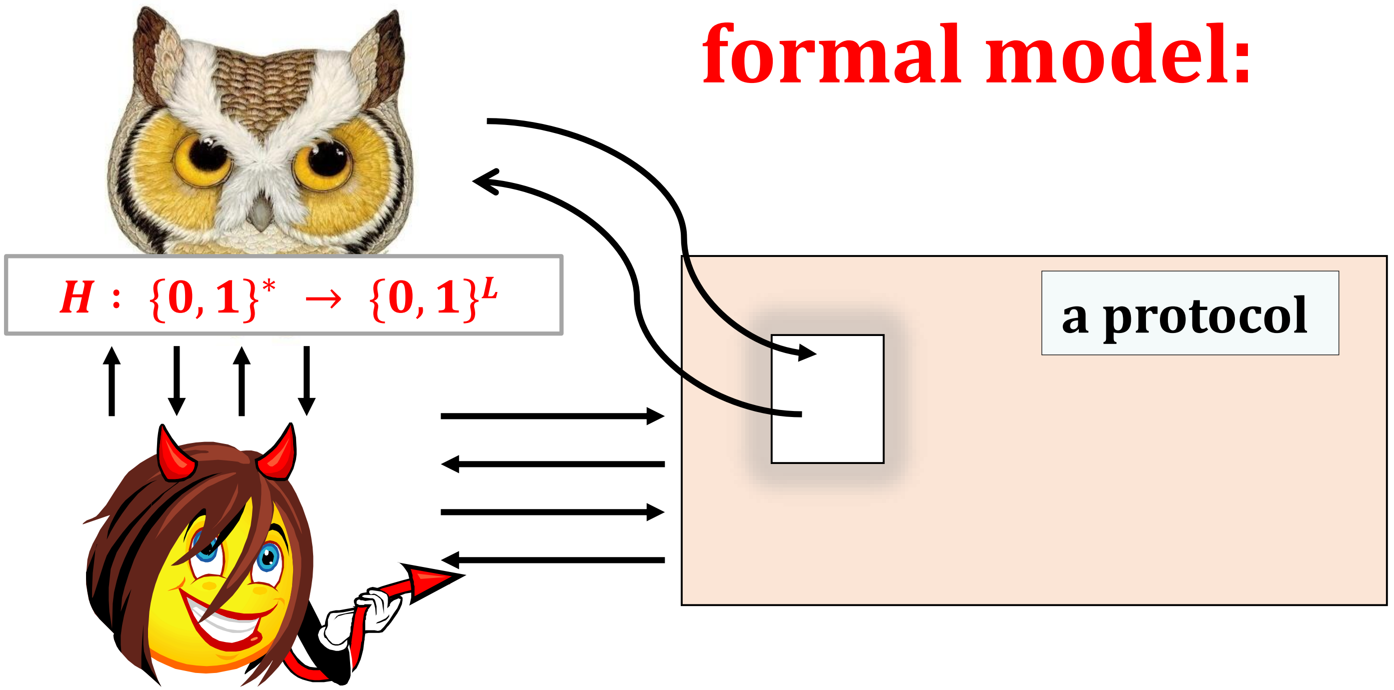
Crucial difference:

Also the adversary can query the oracle

informal description:



formal model:



Every call to ***H*** is replaced with a query to the oracle.

also the adversary is allowed to query the oracle.

How would we use it in the proof?

shorter “uniformly random” $H(X)$

a hash function

$$H: \{0,1\}^* \rightarrow \{0,1\}^L$$

user generated randomness X

As long as the adversary never queried the oracle on X the value $H(X)$ “looks completely random to him”.

Criticism of the Random Oracle Model

[Canetti, Goldreich, Halevi: **The random oracle methodology, revisited**. 1998]

There exists a signature scheme that is

- **secure** in **ROM**

  but

- is **not secure** if the random oracle is replaced with **any** real hash function.

This example is **very artificial**. No "realistic" example of this type is know.

Terminology

Model without the random oracles:
- "**plain model**"
- "**cryptographic model**"

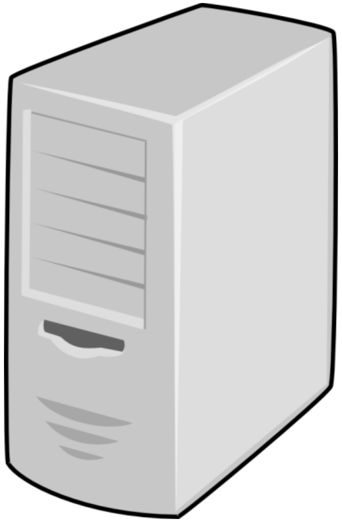
Random Oracle Model is also called:
the "**Random Oracle Heuristic**".

Common view: a proof in **ROM** is better than nothing.

Plan

1. Other uses of hash functions
   1. Merkle trees
   2. Practical randomness extraction and the random oracle model
   3. Password storage and Proofs of Work
2. Real-life constructions

Password storage



Simple idea: instead of storing user's passwords π in plaintext store their hashes.

Better: "salted hashes" $(s, H(s, \pi))$

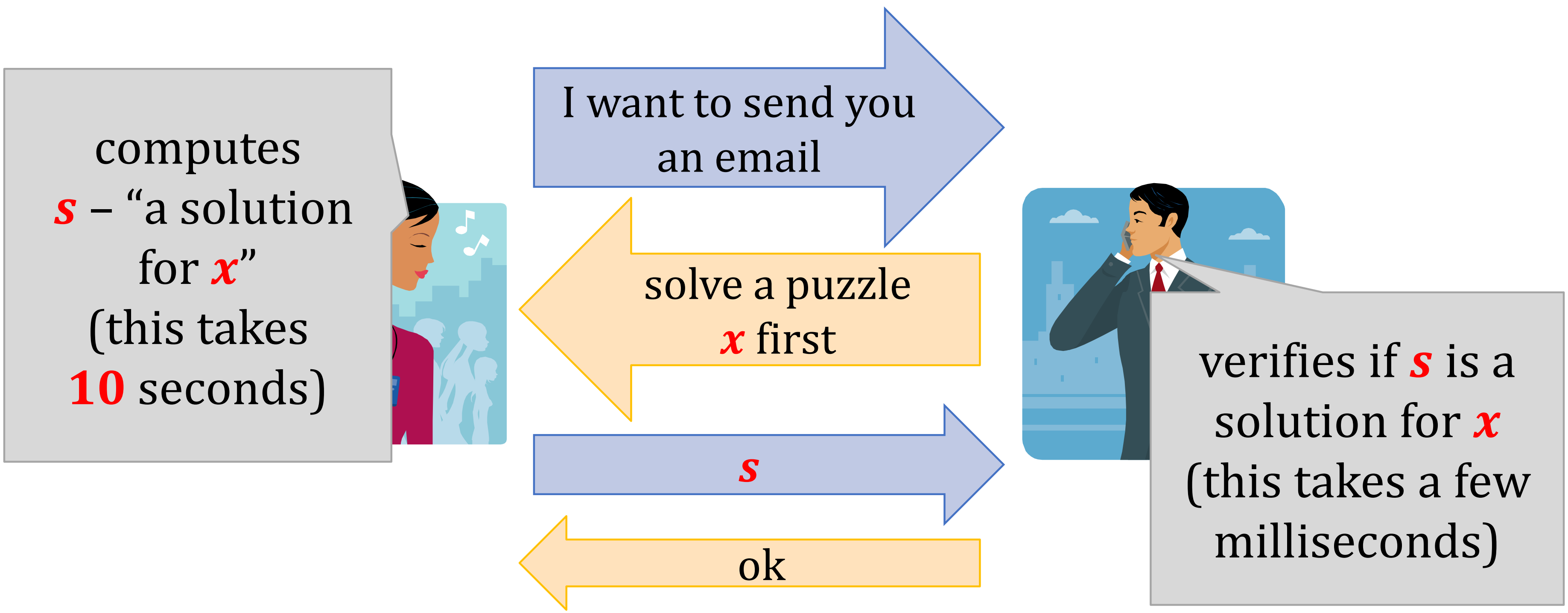
advantages:

1. makes "precomputation attacks" harder (we discussed these attacks on the last exercises)
2. if two users have the same password then the stored values are different.

Proofs of work

Introduced by **Dwork and Naor** [Crypto 1992] as a countermeasure against spam.

Basic idea: Force users to do some computational work: solve a **moderately difficult** "puzzle" (checking correctness of the solution has to be fast).



A simple hash-based PoW

H – a hash function whose computation takes time **TIME(*H*)**

random **x**

s

Prover
finds ***s*** such that
$H(s, x)$ starts with n zeros (in binary)

salt	"hardness parameter

(in **ROM**) takes expected time $2^n \cdot$ **TIME(*H*)**

Verifier
checks if
$H(s, x)$ starts
with **n** zeros

takes time **TIME(*H*)**

This **PoW** is used in Bitcoin.

Problem

Computing typical hash functions is much faster when done in **parallel** and in **hardware** (this can give advantage to a powerful adversary).

For example “**Bitcoin mining**” is done almost entirely on ASICs

AntMiner S7

Advertised Capacity:
4.73 Th/s

Power Efficiency:
0.25 W/Gh

Weight:
8.8 pounds

Guide:
Yes

Price:
$479.95

Avalon6

Advertised Capacity:
3.5 Th/s

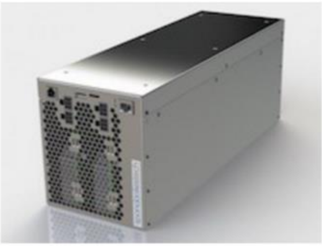
Power Efficiency:
0.29 W/Gh

Weight:
9.5 pounds

Guide:
No

Price:
$499.95

SP20 Jackson



Advertised Capacity:
1.3-1.7 Th/s

Power Efficiency:
0.65 W/Gh

Weight:
20 pounds

Guide:
Yes

Price:
$248.99

Idea for a solution

Design hash functions whose computation **needs to lot of memory**, so it's hard to implement it efficiently in hardware

Example: **scrypt** hash function introduced in:

> **Colin Percival, *Stronger Key Derivation via Sequential Memory-Hard Functions*, 2009**.

Used in **Litecoin**

Has one practical drawback: it's access pattern is **data-dependent** (hence: it reveals the input).

bad for the side-channel resilience

How **scrypt** works?

computing **scrypt(X)**

init phase: fill-in at table of length N with pseudorandom expansion of X.

$$V_0 = X \rightarrow V_1 = H(X) \rightarrow V_2 = H(V_2) \rightarrow \cdots \rightarrow V_{N-1} = H(V_{N-2})$$

result (for $N = 10$):

V_0	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9

second phase: compute the output by accessing the table ”pseudorandomly”

for $i = 0$ **to** $N - 1$ **do**

$j := X \bmod N$

$X := H(X \oplus V_j)$

output X

What is known about **scrypt**?

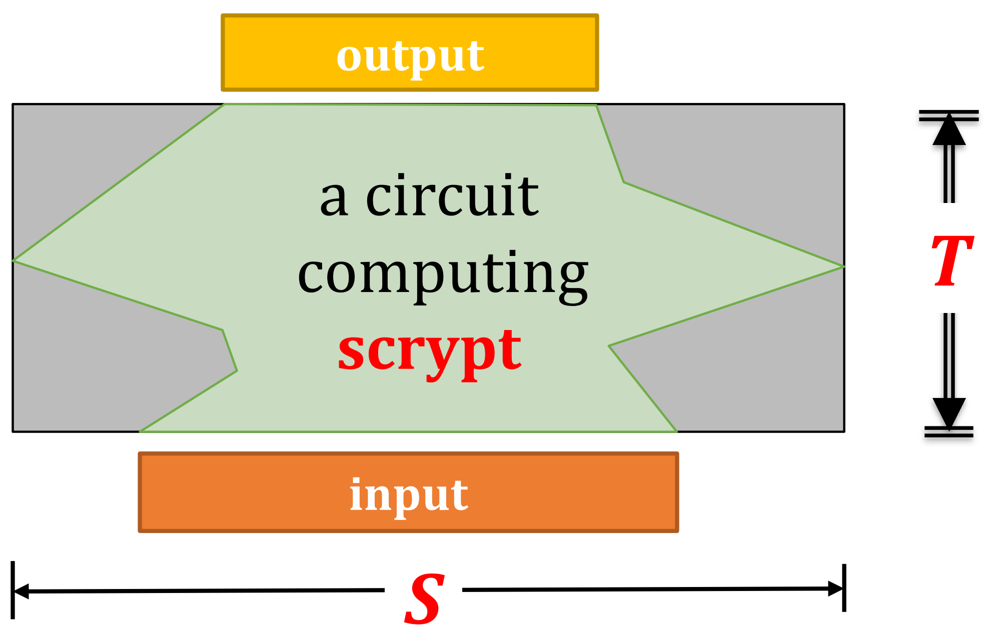
[Percival, 2009]:

- it can be computed in time $O(\mathrm{N})$,
- to compute it one needs time T and space S such that

$$S \times T = \Omega(N^2)$$

this holds even on a parallel machine.

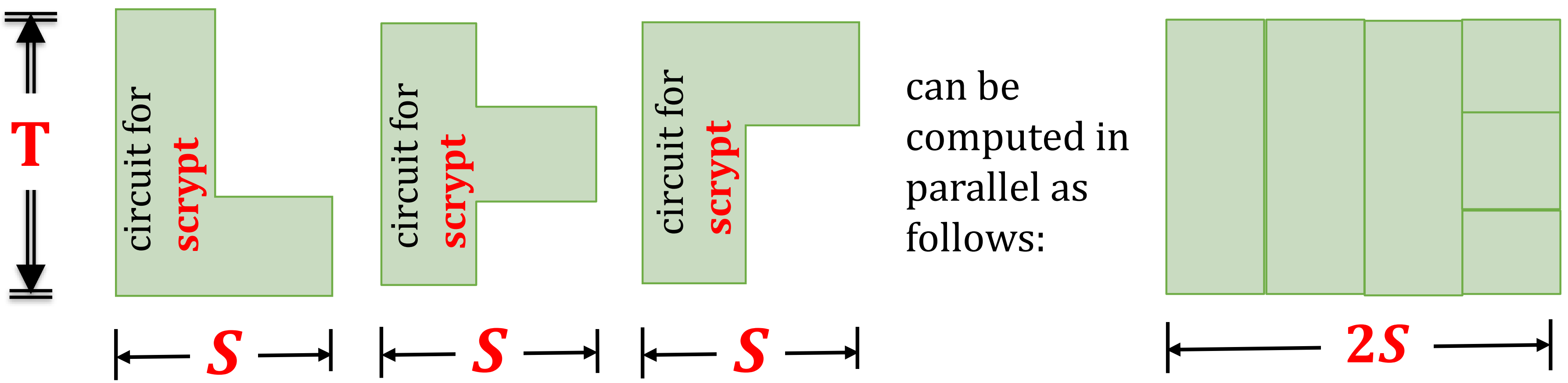
Pictorially:



An observation

[Alwen, Serbinenko, STOC'15]: this definition is **not strong enough**.

The adversary that wants to compute scrypt in parallel can "amortize space". **Example**:



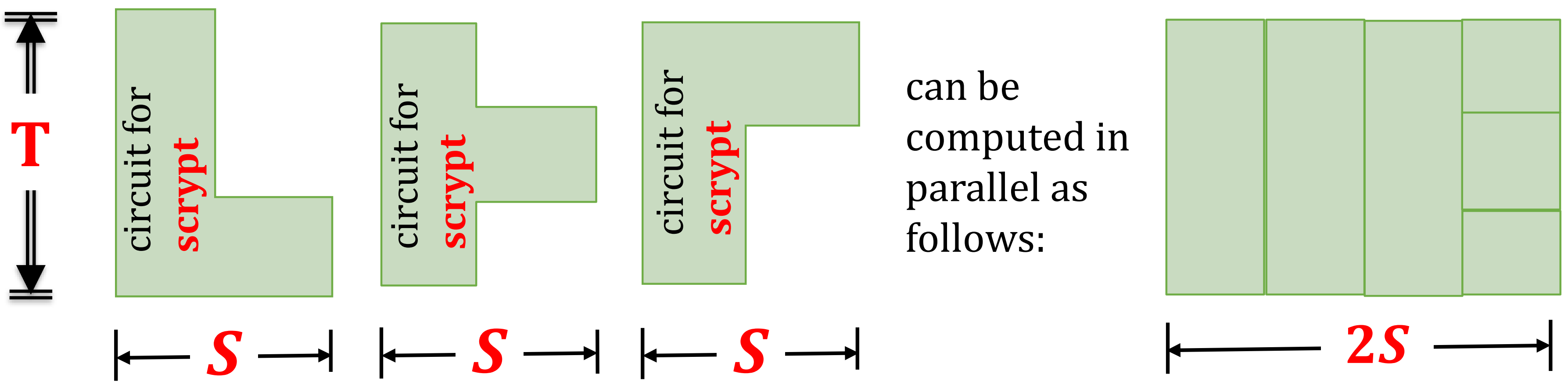
Note: $2S \ll 3S$.

So: the bound provided by Percival is meaningless.

An observation

[Alwen, Serbinenko, STOC'15]: this definition is **not strong enough**.

The adversary that wants to compute scrypt in parallel can "amortize space". **Example**:

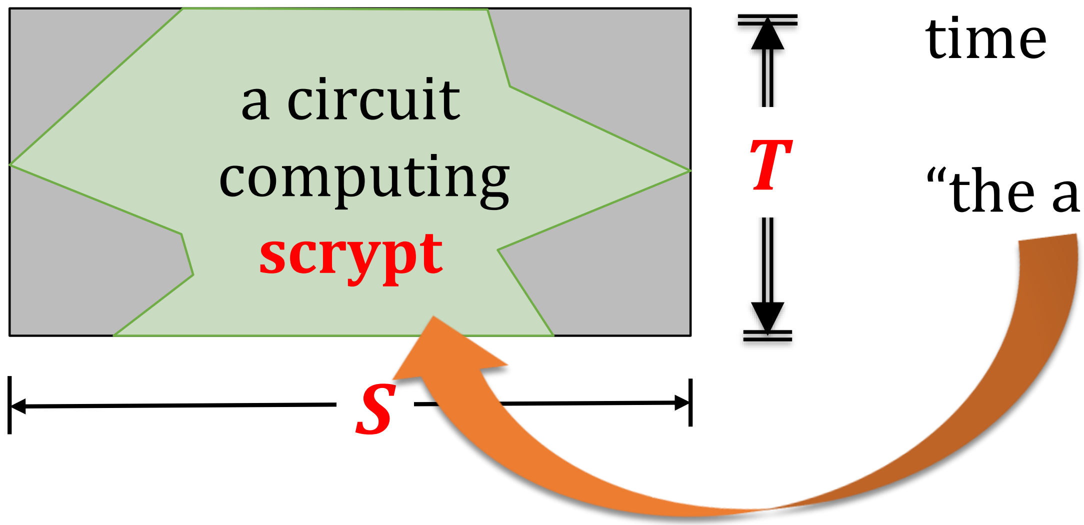


Note: $2S \ll 3S$.

So: the bound provided by Percival is meaningless.

The “right” definition **[Alwen and Serbinenko]**

instead of looking at $S \times T$…



look at the sum of memory cells used over time

“the area on the picture”

A recent result

Alwen, Chen, Pietrzak, Reyzin, and Tessaro: **Scrypt is Maximally Memory-Hard, Cryptology ePrint Archive**, Oct 2016

Password Hashing Competition

- announced in **2013**
- run by an **independent panel** of experts
- **website**: password-hashing.net
- winner (2015): **Argon2** (by Biryukov, Dinu, and Khovratovic)

broken (together with several other competition finalists) by Alwen, Gaži, Kamath, Klein, Osang, Pietrzak, Reyzin, Rolínek, Rybár: **On the Memory-Hardness of Data-Independent Password-Hashing Functions**, Aug 2016

Several improvements to **Argon2** (e.g.: "**Argon2i 1.3**" were also broken)

Plan

1. Other uses of hash functions
    1. Merkle trees
    2. Practical randomness extraction and the random oracle model
    3. Password storage and Proofs of Work
2. Real-life constructions

Fact

There exists a generic attack on any hash function

$$H: \{0,1\}^* \rightarrow \{0,1\}^n$$

that finds a **collision with probability** $\frac{1}{2}$ and works in **time and space** $O\left(2^{\frac{n}{2}}\right)$.

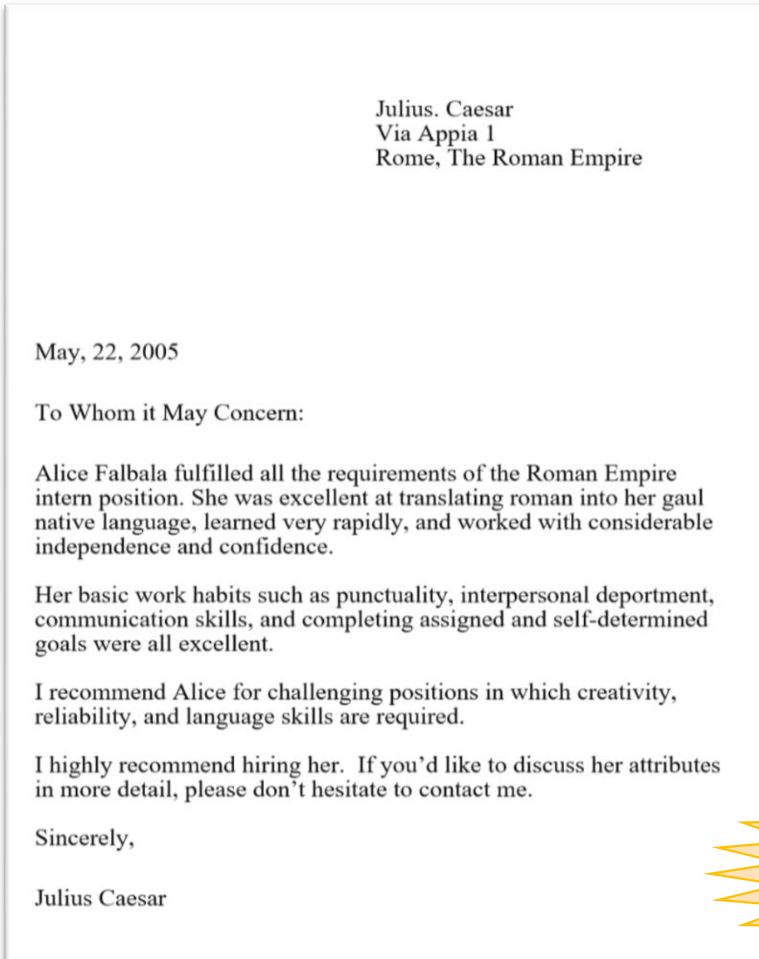
It's called a **birthday attack** and we will discuss it during the exercises.

Consequence: to achieve "m bits of security" one needs to set $n = 2 \cdot m$.

MD5 (Message-Digest Algorithm 5)

- based on the **Merkle-Damgard paradigm**
- **output length**: **128 bits**,
- **designed** by **Rivest** in **1991**,
- in **1996**, **Dobbertin** found collisions in the compressing function of **MD5**,
- in **2004** a group of **Chinese mathematicians** designed a method for finding collisions in **MD5**,

June 2005: researchers at the Bochum University produce 2 postscript documents with the same **MD5** hash



Julius. Caesar
Via Appia 1
Rome, The Roman Empire

May, 22, 2005

To Whom it May Concern:

Alice Falbala fulfilled all the requirements of the Roman Empire intern position. She was excellent at translating roman into her gaul native language, learned very rapidly, and worked with considerable independence and confidence.

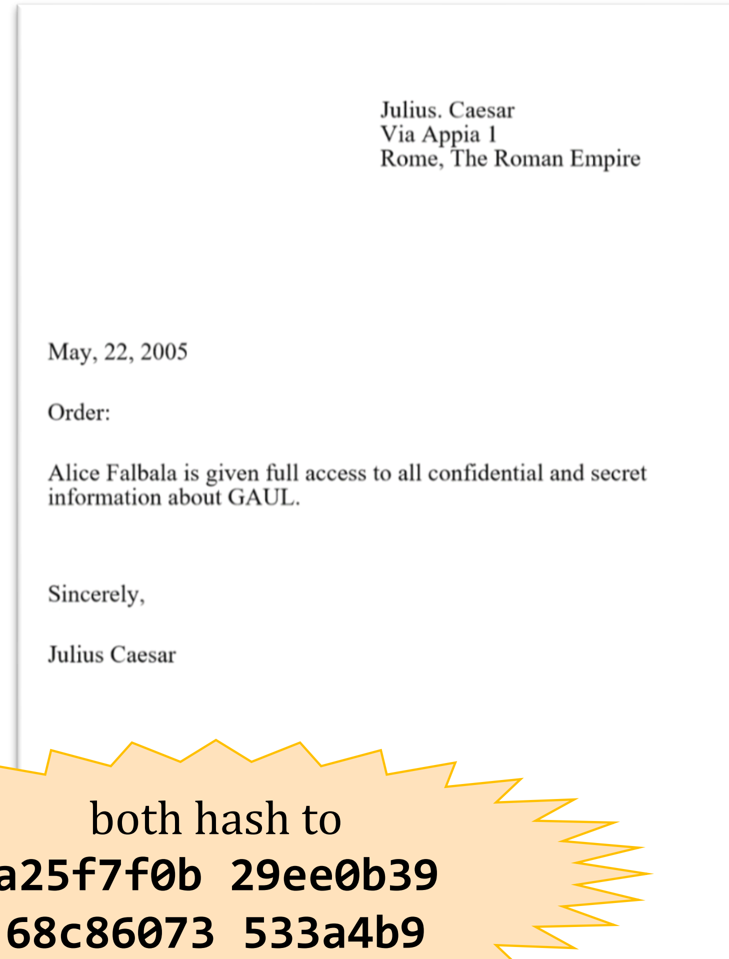
Her basic work habits such as punctuality, interpersonal deportment, communication skills, and completing assigned and self-determined goals were all excellent.

I recommend Alice for challenging positions in which creativity, reliability, and language skills are required.

I highly recommend hiring her. If you'd like to discuss her attributes in more detail, please don't hesitate to contact me.

Sincerely,

Julius Caesar



Julius. Caesar
Via Appia 1
Rome, The Roman Empire

May, 22, 2005

Order:

Alice Falbala is given full access to all confidential and secret information about GAUL.

Sincerely,

Julius Caesar

both hash to
`a25f7f0b 29ee0b39 68c86073 533a4b9`

This is done by exploiting the redundancy in postscript.

Colliding certificates

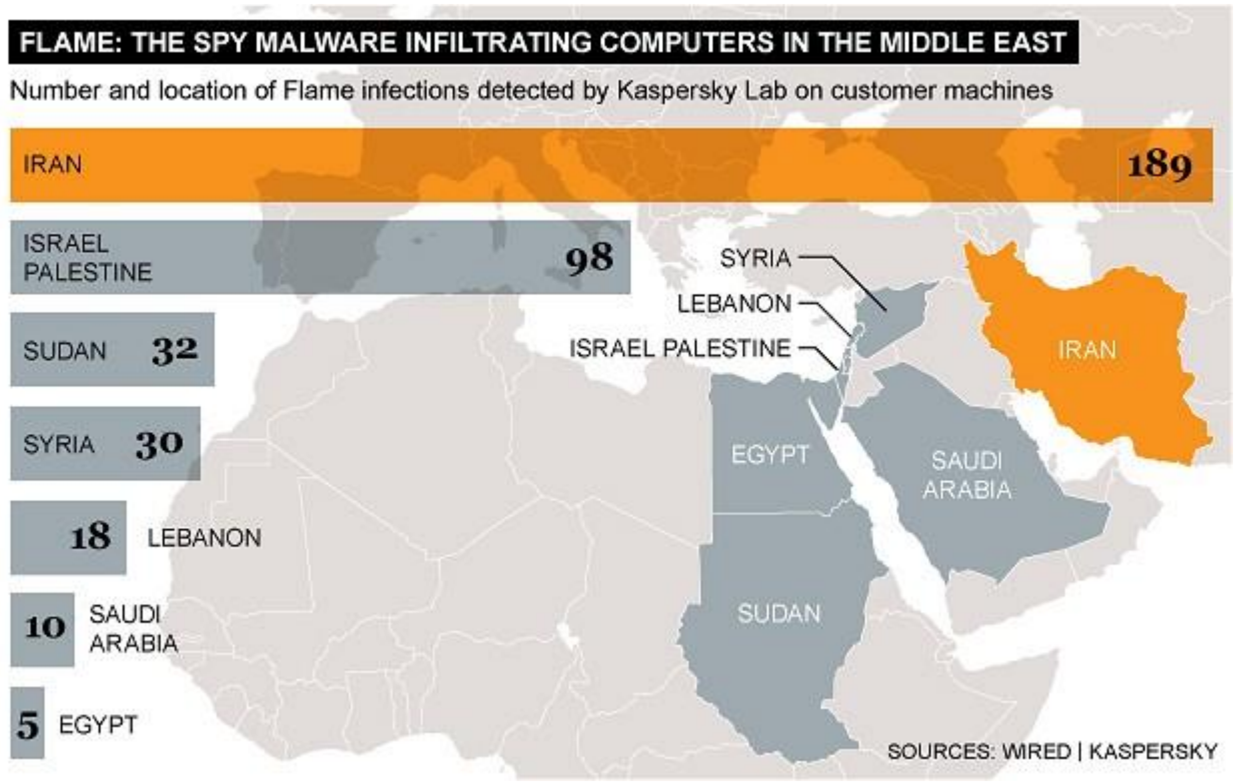
2005 and 2006: A. Lenstra, X. Wang, and B. de Weger found **X.509** certificates with different public keys and the same **MD5** hash.

(we will discuss the **X.509** certificates later)

Two certificates for **different names** ("Arjen K. Lenstra" and "Marc Stevens") **and different public keys**.

Flame malware

attack on the Microsoft Windows Update Mechanism
exploiting **MD5** collision



SHA-1 (Secure Hash Algorithm)

- based on the **Merkle-Damgard paradigm**
- **output length**: **160 bits**,
- designed in **1993** by the **NSA**,
- in **2005 Xiaoyun Wang, Andrew Yao and Frances Yao** presented an attack that runs in time 2^{63}.
- Oct 2015: **[Stevens, Karpman, and Peyrin: Freestart collision on full SHA-1]** a collision in the compression function in 2^{57} **SHA-1** evaluations.
- Feb 2017: **[Stevens, Bursztein, Karpman, Albertini , Markov]** a collision in full **SHA-1**

The collision that was found in 2015

	Input 1
IV1	50 6b 01 78 ff 6d 18 90 20 22 91 fd 3a de 38 71 b2 c6 65 ea
M1	9d 44 38 28 a5 ea 3d f0 86 ea a0 fa 77 83 a7 36 33 24 48 4d af 70 2a aa a3 da b6 79 d8 a6 9e 2d 54 38 20 ed a7 ff fb 52 d3 ff 49 3f c3 ff 55 1e fb ff d9 7f 55 fe ee f2 08 5a f3 12 08 86 88 a9
SHA1_compression_function (**IV1**,**M1**)	f0 20 48 6f 07 1b f1 10 53 54 7a 86 f4 a7 15 3b 3c 95 0f 4b

	Input 2
IV2	50 6b 01 78 ff 6d 18 91 a0 22 91 fd 3a de 38 71 b2 c6 65 ea
M2	3f 44 38 38 81 ea 3d ec a0 ea a0 ee 51 83 a7 2c 33 24 48 5d ab 70 2a b6 6f da b6 6d d4 a6 9e 2f 94 38 20 fd 13 ff fb 4e ef ff 49 3b 7f ff 55 04 db ff d9 6f 71 fe ee ee e4 5a f3 06 04 86 88 ab
SHA1_compression_function (**IV2**,**M2**)	f0 20 48 6f 07 1b f1 10 53 54 7a 86 f4 a7 15 3b 3c 95 0f 4b

Hardware used in **[Stevens, Karpman, and Peyrin: Freestart collision on full SHA-1]**:

*"We have computed the SHA-1 freestart collision on **Kraken**, our 64-GPU cluster. More precisely Kraken is composed of 16 nodes, each node being made of simple, cheap and widely available hardware: 4 GTX-970 GPUs, 1 Haswell i5-4460 processor and 16GB of RAM."*

An estimation

[Stevens, Karpman, and Peyrin: Freestart collision on full SHA-1]

"Concretely, we estimate the SHA-1 collision cost today (i.e., Fall 2015) between 75K$ and 120K$ renting Amazon EC2 cloud computing over a few months."

Reaction of the industry

Microsoft may block SHA1 certificates sooner than expected

Encrypted sites running old certificates will be inaccessible from modern browsers.

By Zack Whittaker for Zero Day | November 9, 2015 -- 13:16 GMT (13:16 GMT) | Topic: Security

Mozilla Security Blog

OCT 20 2015

Continuing to Phase Out SHA-1 Certificates

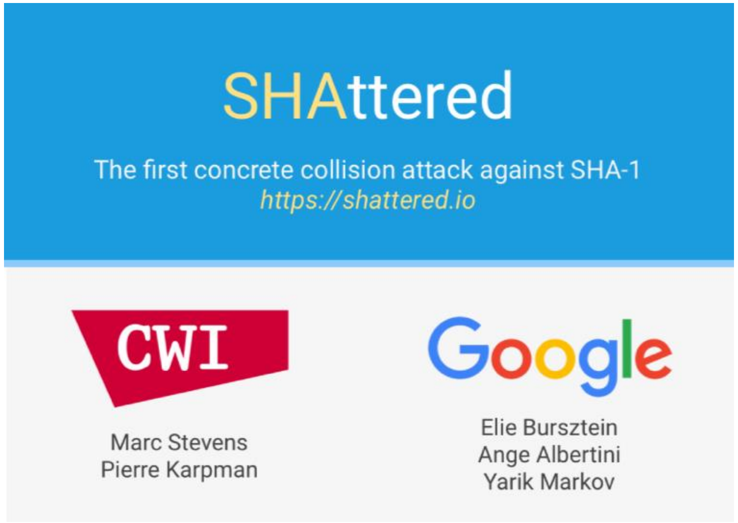
Richard Barnes

In our previous blog post about phasing out certificates with SHA-1 based signature algorithms, we said that we planned to take a few actions with regard to SHA-1 certificates:

1. Add a security warning to the Web Console to remind developers that they should not be using a SHA-1 based certificates
2. Show the "Untrusted Connection" error whenever a SHA-1 certificate issued after January 1, 2016, is encountered in Firefox
3. Show the "Untrusted Connection" error whenever a SHA-1 certificate is encountered in Firefox after January 1, 2017

The attack from 2017

Two colliding pdf files:



An unexpected victim of this attack

PCWorld FROM IDG

GET OUR DIGITAL MAGAZINE

NEWS | REVIEWS | HOW-TO | VIDEO | BUSINESS | LAPTOPS | TABLETS | PHONES | HARDWARE | SECURITY | SOFTWARE | GADGETS

Privacy Encryption Antivirus

Home / Security

NEWS

SHA-1 collision can break SVN code repositories

The WebKit repository was corrupted after someone committed two colliding PDF files to it

By Lucian Constantin

Romania Correspondent, IDG News Service |

FEB 27, 2017 10:23 AM PT

A new hash algorithm: SHA-3

Selected by the **National Institute of Standards and Technology (NIST)** in an open competition.

5 finalists: **BLAKE, Grøstl, JH, Keccak, Skein**.

Winner (2012): **Keccak**.

SHA-3: Keccak

authors: Guido Bertoni, Joan Daemen, Michaël Peeters, and Gilles Van Assche

output lengths: **224**, **256**, **384**, **512**, or unbounded

speed: **12.5** cycles per byte on Core 2

Not based on the **Merkle-Damgard paradigm**.

Instead: it uses the **sponge construction**.

Standardized Keccak's parameters for fixed output length

state width *b*	**rate *r***	**capacity *c***	**output length**
1600	1344	256	224
1600	1344	256	256
1600	1088	512	384
1600	1088	512	512