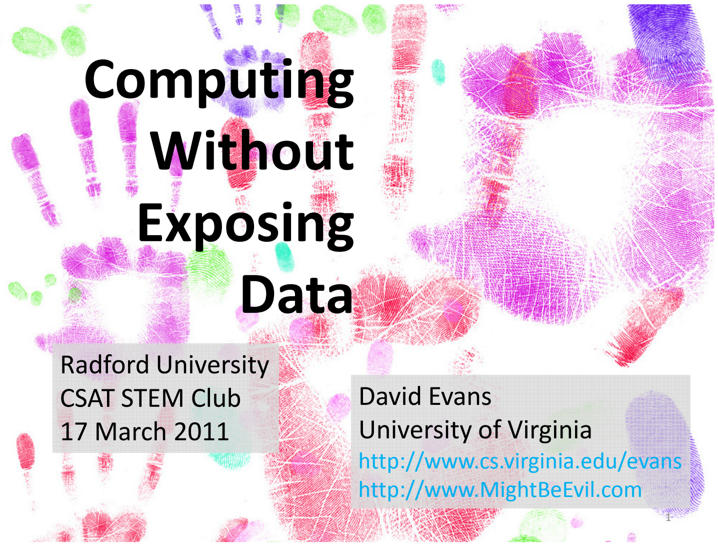
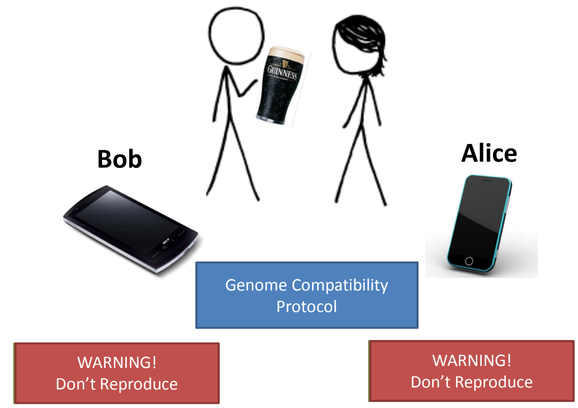


Computing Without Exposing Data

Radford University
CSAT STEM Club
17 March 2011

David Evans
University of Virginia
http://www.cs.virginia.edu/evans
http://www.MightBeEvil.com

"Genetic Dating"



GenePartner™ Love is no coincidence

23andMe

TheScientist News Current Issue Archive Sur

2 comments
Comment on this news story
By Kerry Grens

Forget mistletoe - what about DNA?

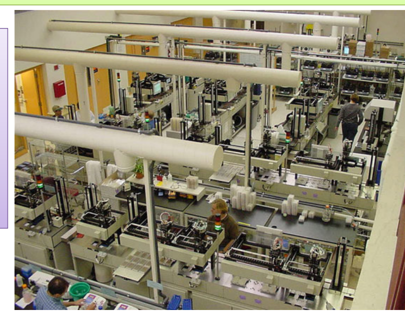
A new dating service matches singles using major histocompatibility complex genes

ScientificMatch.com "The Science of Love"

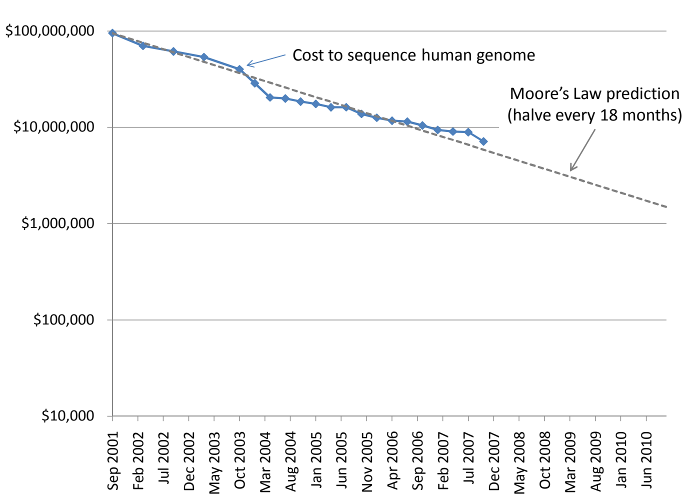
Genome Sequencing

1990: Human Genome Project starts, estimate $3B to sequence one genome ($0.50/base)

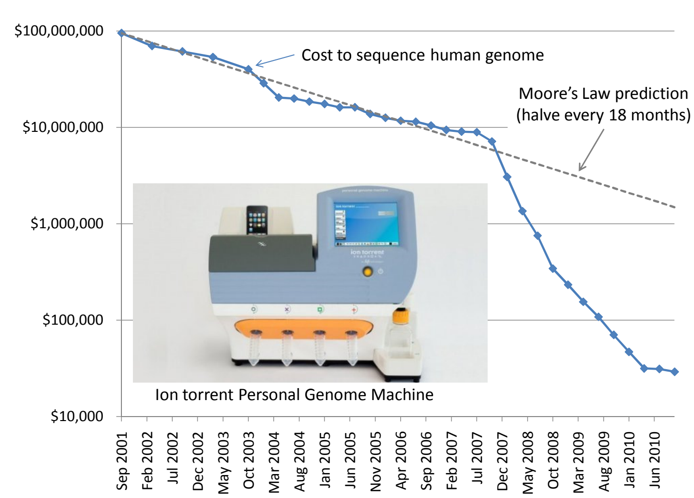
2000: Human Genome Project declared complete, cost ~$300M



Whitehead Institute, MIT



Data from National Human Genome Research Institute: http://www.genome.gov/sequencingcosts



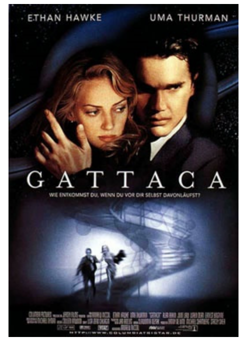
Data from National Human Genome Research Institute: http://www.genome.gov/sequencingcosts

Year	r		Reported sequencing consumables cost	Estimated cost per 40-fold coverage
[redacted]			$10,000,000	$57,000,000
			$1,000,000	$5,700,000
			$250,000	$330,000
			$48,000	$69,000
2009	t		$8,005	$3,700
2009	t		$3,451	$2,200
2009	t		$1,726	$1,500

George Church (Personal Genome Project)
Richmond Forum, Saturday

Human Geno ... ***ling DNA Nanoarrays*****.** Radoje Drmanac, An... L. Burns, Bahram G. Kermani, Paolo Carnevali, Igo... Andres Fernandez, Bryan Staker, Krishna P. Par... Ryan Cedeno, Linsu Chen, Dan Chernikoff, Alex... t, Coleen R. Hacker, Robert Hartlage, Brian Hauser, St... lvin Kong, Tom Landers, Catherine Le, Jia Liu, Celeste ... Helena Perazich, Kimberly Perry, Brock A. Peters, Joe Peterson, Charit L. Pethiyagoda, Kaliprasad Pothuraju, Claudia Richter, Abraham M. Rosenbaum, Shaunak Roy, Jay Shafto, Uladzislau Sharanhovich, Karen W. Shannon, Conrad G. Sheppy, Michel Sun, Joseph V. Thakuria, Anne Tran, Dylan Vu, Alexander Wait Zaranek, Xiaodi Wu, Snezana Drmanac, Arnold R. Oliphant, William C. Banyai, Bruce Martin, Dennis G. Ballinger, George M. Church, Clifford A. Reid. ***Science*****, January 2010.**

Dystopia



GenePartner™
Love is no coincidence

Personalized Medicine

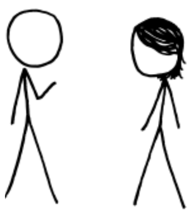
Secure Two-Party Computation

Bob's Genome: ACTG…
Markers (~1000): [0,1, …, 0]

Alice's Genome: ACTG…
Markers (~1000): [0, 0, …, 1]

Bob

Alice



$$x = f(g_A, g_B)$$

Can Alice and Bob compute a function of their private data, without exposing anything about their data besides the result?

Secure Function Evaluation

Alice (circuit generator)

Picks $a \in \{0,1\}^s$

Agree on
$f(a,b) \to x$

Bob (circuit evaluator)

Picks $b \in \{0,1\}^t$

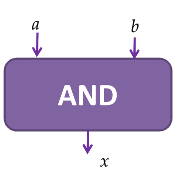
Garbled Circuit Protocol

Outputs $x = f(a,b)$ without revealing a to Bob or b to Alice.

Andrew Yao, 1982/1986

Computing with Lookup Tables

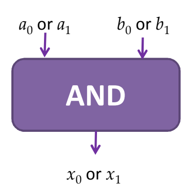
Inputs		Output
a	b	x
0	0	0
0	1	0
1	0	0
1	1	1



Computing with Meaningless Values?

Inputs		Output
a	b	x
a_0	b_0	x_0
a_0	b_1	x_0
a_1	b_0	x_0
a_1	b_1	x_1

a_i, b_i, x_i are **random** values, chosen by the **circuit generator** but **meaningless** to the **circuit evaluator**.

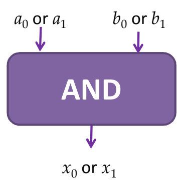


Computing with Garbled Tables

Inputs		Output
a	b	x
a_0	b_0	$Enc_{a_0,b_0}(x_0)$
a_0	b_1	$Enc_{a_0,b_1}(x_0)$
a_1	b_0	$Enc_{a_1,b_0}(x_0)$
a_1	b_1	$Enc_{a_1,b_1}(x_1)$

Bob can only decrypt **one** of these!

a_i, b_i, x_i are **random** values, chosen by the **circuit generator** but **meaningless** to the **circuit evaluator**.



Garbled And Gate
$Enc_{a_0,b_1}(x_0)$
$Enc_{a_1,b_1}(x_1)$
$Enc_{a_1,b_0}(x_0)$
$Enc_{a_0,b_0}(x_0)$

Garbled Circuit Protocol

Alice (circuit generator)

Creates random keys: $a_0, a_1, b_0, b_1, x_0, x_1$

Bob (circuit evaluator)

And Gate
$Enc_{a_0,b_1}(x_0)$
$Enc_{a_1,b_1}(x_1)$
$Enc_{a_1,b_0}(x_0)$
$Enc_{a_0,b_0}(x_0)$

Sends a_i to Bob based on her input value a_0

How does the Bob learn his own input wires?

Primitive: **Oblivious Transfer**

Alice Knows b_0, b_1

Bob Picks $i \in \{0, 1\}$

Oblivious Transfer Protocol

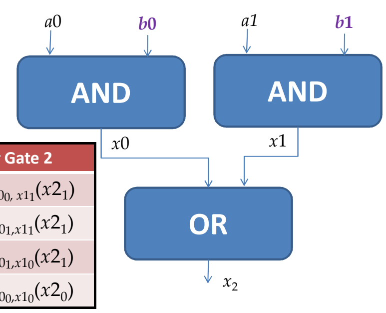
Alice: Learns nothing

Bob: Learns b_i (only)

Oblivious: Alice doesn't learn which secret Bob obtains
Transfer: Bob learns one of Alice's secrets

Rabin, 1981; Even, Goldreich, and Lempel, 1985; many subsequent papers

Chaining Garbled Circuits

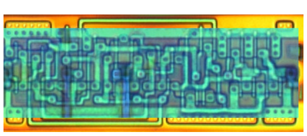


And Gate 1
$Enc_{a10,b11}(x1_0)$
$Enc_{a11,b11}(x1_1)$
$Enc_{a11,b10}(x1_0)$
$Enc_{a10,b10}(x1_0)$

Or Gate 2
$Enc_{x00,x11}(x2_1)$
$Enc_{x01,x11}(x2_1)$
$Enc_{x01,x10}(x2_1)$
$Enc_{x00,x10}(x2_0)$

We can do ***any*** computation privately this way!

Building Computing Systems



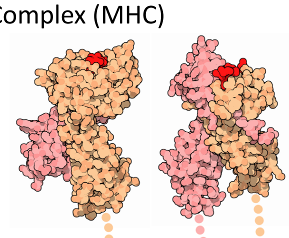
$Enc_{x00,x11}(x2_1)$
$Enc_{x01,x11}(x2_1)$
$Enc_{x01,x10}(x2_1)$
$Enc_{x00,x10}(x2_0)$

Digital Electronic Circuits	Garbled Circuits
Operate on **known data**	Operate on **encrypted wire labels**
One-bit logical operation requires moving a few electrons a few nanometers (many Billions per second)	One-bit logical operation requires performing (up to) 4 encryption operations (~100,000 gates per second)
Reuse is great!	Reuse is not allowed!
All basic operations have similar cost	Some logical operations nearly free (XOR)

Offspring Immune System Test

Major Histocompatibility Complex (MHC)

Alice's MHC genes: [0 1 1 0 … 0 1 1]
Bob's MHC genes: [1 1 1 0 … 1 0 1]



Diversity is key to good immune systems!

Goal: count number of indices where A[i] ≠ B[i]

XOR
Every Cryptographer's Favorite Function

⊕

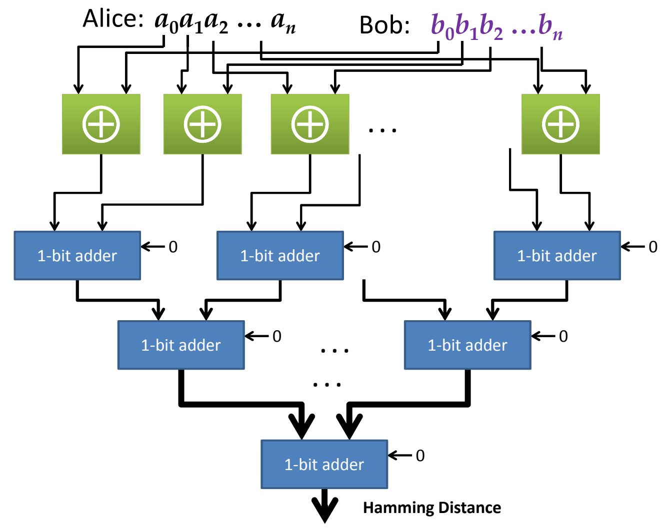
$a \oplus a = 0$

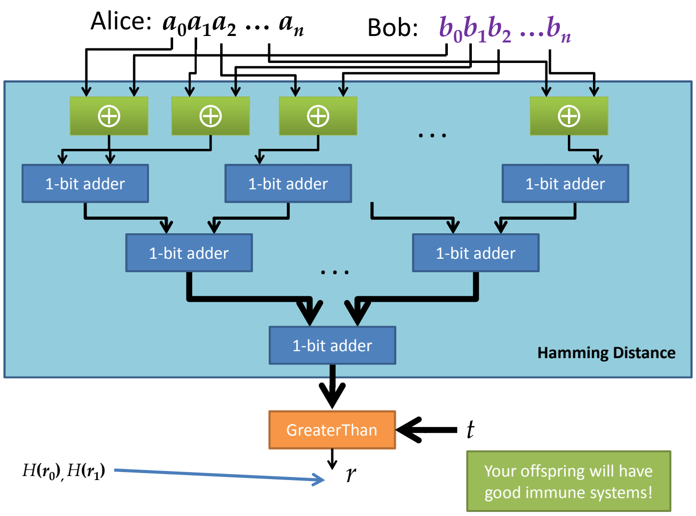
$a \oplus r$ is uniformly random

$a \oplus r \oplus r = a$

XOR		
a	b	x
0	0	0
0	1	1
1	0	1
1	1	0

Can compute $a \oplus b$ on garbled inputs without and encryptions ("free").





Heterozygous Recessive Risk

Alice

Bob	A	a
A	AA	Aa
a	aA	aa

carrier (Aa)

cystic fibrosis (aa)

Alice's Heterozygous Recessive genes: { 5283423, 1425236, 839523, … }

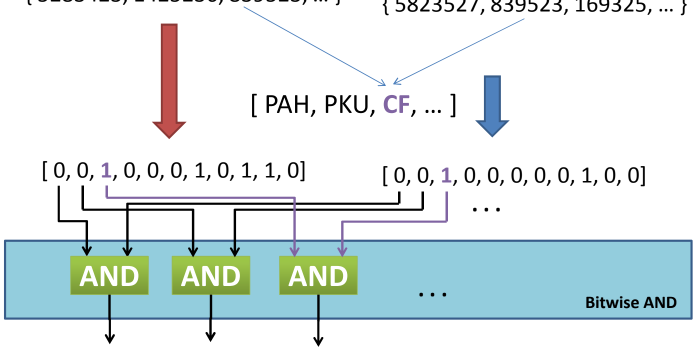
Bob's Heterozygous Recessive genes: { 5823527, 839523, 169325, … }

Goal: find the intersection of A and B

Bit Vector Intersection

Alice's Recessive genes:
{ 5283423, 1425236, 839523, … }

Bob's Recessive genes:
{ 5823527, 839523, 169325, … }



Scaling

What if there are millions of possible diseases?

Length of bit vector:

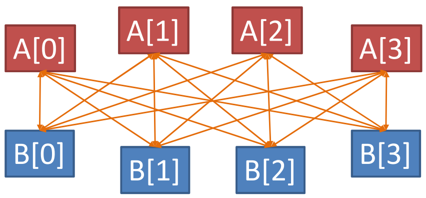
number of possible values

(2^L where L is number of bits for each value)

Other private set intersection problems:
- Do Alice and Bob have any friends in common?
- Data mining problems: combine medical records across hospitals
- Two companies want to do joint marketing to common customers

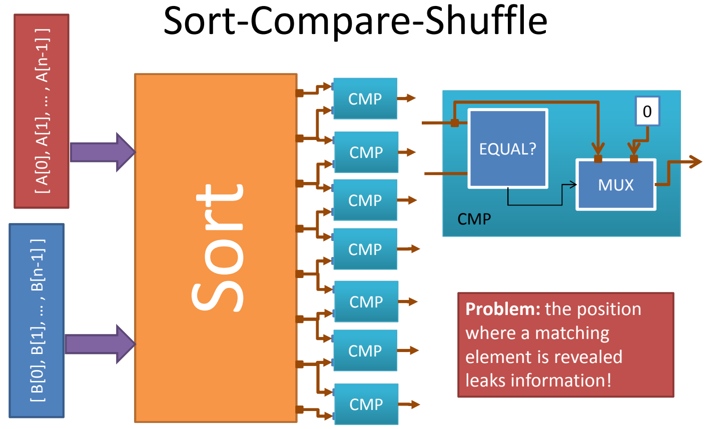
Pairwise Comparison

```
for i in range(0, n-1):
    for j in range(0, n-1):
        if A[i] = B[j] output A[i]
```

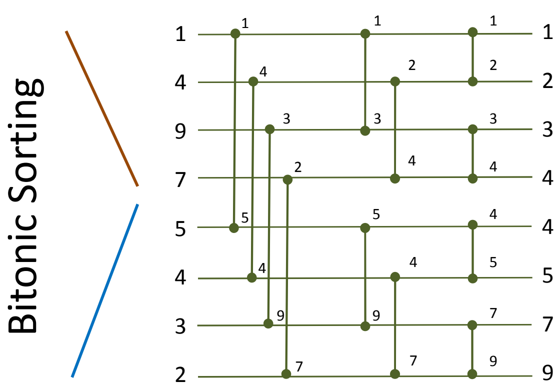


n^2 comparisons

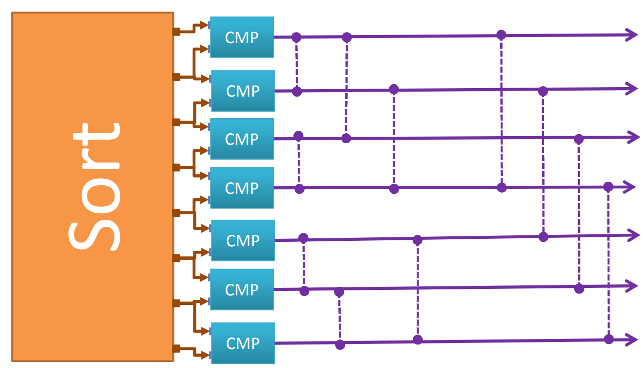
Sort-Compare-Shuffle



Problem: the position where a matching element is revealed leaks information!



(Imperfect) Shuffling



Some Results

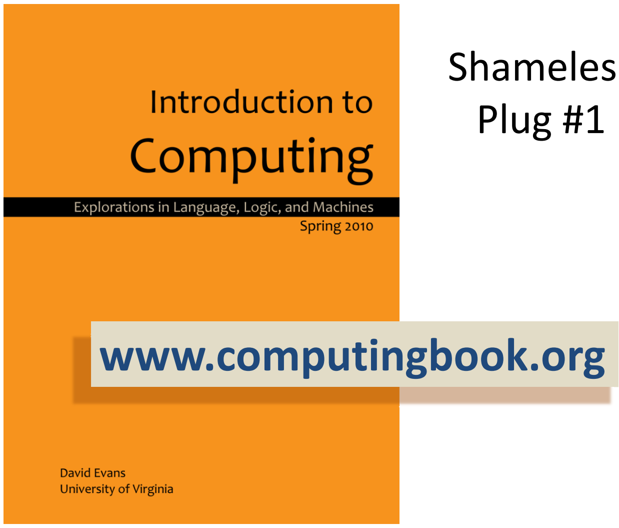
Problem	Best Previous Result	Our Result	Speedup
Hamming Distance (Face Recognition, Genetic Dating) – two 900-bit vectors	213s [SCiFI, 2010]	0.051s	4176
Levenshtein Distance (genome, text comparison) – two 200-character inputs	534s [Jha+, 2008]	18.4s	29
Smith-Waterman (genome alignment) – two 60-nucleotide sequences	[Not Implementable]	447s	-
AES Encryption	3.3s [Henecka, 2010]	0.2s	16.5
Fingerprint Matching (1024-entry database, 640x8bit vectors)	~83s [Barni, 2010]	18s	4.6

Scalable: 1 Billion gates evaluated at ~100,000 gates/second on laptop

Yan Huang
(UVa Computer Science PhD Student)

Jonathan Katz
(University of Maryland)

Aaron Mackey
(UVa Center for Public Health Genomics)



Shameless
Plug #1