Advanced Internet Systems

CSE 454
Daniel Weld

To do

- Add picture of original MT
- Add greg little or casting words flowchart
- Discussion included qualifications, contracts,

CSE 454 Overview

HTTP, HTML, Scaling & Crawling
Cryptography & Security

Transfer of Confidential Data

You (client) — Merchant (server)

I want to make a purchase.

Here is my RSA public key E *and* a cert *and* a "nonce".

My Credit Card and your nonce are E("6543 2345 6789 8765", nonce).

Cyptography

- Symmetric + asymmetric ciphers
- Stream + block ciphers; 1-way hash
- $Z=Y^X \bmod N$

DNS, HTTP, HTML

- Get, put, post
- Cookies, log file analysis

DNS

- Hierarchical namespace
- Susceptible to DOS attacks
- Recent news: Google Public DNS

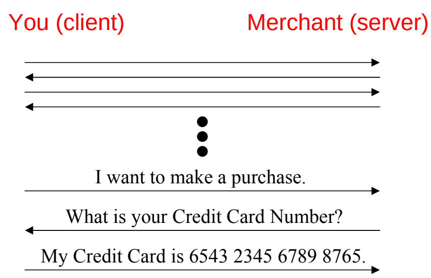
HTTP

- Get, put, post
- Cookies, log file analysis

HTML

- Link extraction

Transfer of Confidential Data



Transfer of Confidential Data

- But the Internet provides no privacy.
- Is there any way to protect my data from prying eyes at intermediate nodes?

Slides by Josh Benaloh

Symmetric Encryption

- If the user has a pre-existing relationship with the merchant, they may have a shared secret key K – known only to the two parties.
- User encrypts private data with key K.
- Merchant decrypts data with key K.

Slides by Josh Benaloh

Asymmetric Encryption

- What if the user and merchant have no prior relationship?
- Asymmetric encryption allows me to encrypt a message for a recipient without knowledge of the recipient's decryption key.

Slides by Josh Benaloh

The Fundamental Equation

$$E = mc^2$$

Slides by Josh Benaloh

The Fundamental Equation

$$Z = Y^X \bmod N$$

Slides by Josh Benaloh

The Fundamental Equation

$$Z=Y^X \bmod N$$

When Z is unknown, it can be efficiently computed.

The Fundamental Equation

$$Z=Y^X \bmod N$$

When X is unknown, the problem is known as the ***discrete logarithm*** and is generally believed to be hard to solve.

The Fundamental Equation

$$Z=Y^X \bmod N$$

When Y is unknown, the problem is known as ***discrete root finding*** and is generally believed to be hard to solve …

without the factorization of N.

The Fundamental Equation

$$Z=Y^X \bmod N$$

The problem is not well-studied for the case when N is unknown.

How to compute Y^X mod N

Compute Y^X and then reduce mod N.

- If X, Y, and N each are 1,000-bit integers, Y^X consists of $\sim 2^{1010}$ bits.
- Since there are roughly 2^{250} particles in the universe, storage is a problem.

How to compute Y^X mod N

- Repeatedly multiplying by Y (followed each time by a reduction modulo N) X times solves the storage problem.
- However, we would need to perform $\sim 2^{900}$ 32-bit multiplications per second to complete the computation before the sun burns out.

How to compute Y^X mod N

Multiplication by Repeated Doubling

To compute $X \cdot Y$,

compute $Y, 2Y, 4Y, 8Y, 16Y, \ldots$

and sum up those values dictated by the binary representation of X.

Example: $26Y = 2Y + 8Y + 16Y$.

How to compute Y^X mod N

Exponentiation by Repeated Squaring

To compute Y^X,

compute $Y, Y^2, Y^4, Y^8, Y^{16}, \ldots$

and multiply those values dictated by the binary representation of X.

Example: $Y^{26} = Y^2 \bullet Y^8 \bullet Y^{16}$.

How to compute Y^X mod N

We can now perform a 1,000-bit modular exponentiation using ~1,500 1,000-bit modular multiplications.

- **1,000 squarings: $y, y^2, y^4, \ldots, y^{2^{1000}}$**
- **~500 "ordinary" multiplications**

The Fundamental Equation

$$Z = Y^X \bmod N$$

When Y is unknown, the problem is known as *discrete root finding* and is generally believed to be hard to solve … without the factorization of N.

RSA Encryption/Decryption

- **Select two large primes p and q.**
- **Publish the product $N = pq$.**
- **The exponent X is typically fixed at 65537.**
- **Encrypt message Y as $E(Y) = Y^X \bmod N$.**
- **Decrypt ciphertext Z as $D(Z) = Z^{1/X} \bmod N$.**
- **Note $D(E(Y)) = (Y^X)^{1/X} \bmod N = Y$.**

RSA Signatures and Verification

- **Not only is $D(E(Y)) = (Y^X)^{1/X} \bmod N = Y$, but also $E(D(Y)) = (Y^{1/X})^X \bmod N = Y$.**
- **To form a signature of message Y, create $S = D(Y) = Y^{1/X} \bmod N$.**
- **To verify the signature, check that $E(S) = S^X \bmod N$ matches Y.**

Transfer of Confidential Data

You (client) | Merchant (server)

⋮

I want to make a purchase. →

← What is your Credit Card Number?

My Credit Card is 6543 2345 6789 8765. →

Slides by Josh Benaloh

Transfer of Confidential Data

You (client) | Merchant (server)

⋮

I want to make a purchase. →

← Here is my RSA public key E.

My Credit Card is E(6543 2345 6789 8765). →

Slides by Josh Benaloh

Intermediary Attack

You (client) | Intermediary | Merchant (server)

⋮

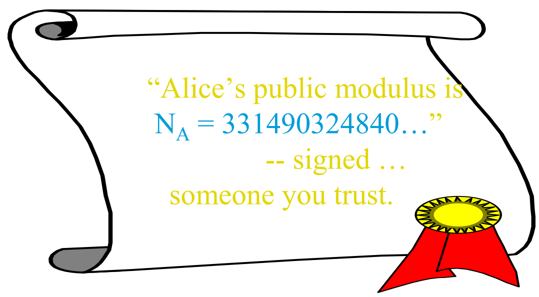
I want to make a purchase. → | I want to make a purchase. →

← My public key is Ẽ. | ← My public key is E.

Ẽ(CC#) → | E(CC#) →

Slides by Josh Benaloh

Digital Certificates



Slides by Josh Benaloh

Transfer of Confidential Data

You (client) | Merchant (server)

⋮

I want to make a purchase. →

← Here is my RSA public key E *and* a cert.

My Credit Card is E(6543 2345 6789 8765). →

Slides by Josh Benaloh

Replay Attack

You (client) | Merchant (server)

I want to make a purchase. →

← Here is my RSA public key E *and* a cert.

My Credit Card is E(6543 2345 6789 8765). →

Later …

Eavesdropper | Merchant (server)

I want to make a different purchase. →

← Here is my RSA public key E *and* a cert.

My Credit Card is E(6543 2345 6789 8765). →

Slides by Josh Benaloh

Transfer of Confidential Data

You (client) **Merchant (server)**

I want to make a purchase.

Here is my RSA public key E *and* a cert *and* a "nonce".

My Credit Card and your nonce are E("6543 2345 6789 8765", nonce).

SSL/PCT/TLS History

- 1994: Secure Sockets Layer (SSL) V2.0
- 1995: Private Communication Technology (PCT) V1.0
- 1996: Secure Sockets Layer (SSL) V3.0
- 1997: Private Communication Technology (PCT) V4.0
- 1999: Transport Layer Security (TLS) V1.0

SSL/PCT/TLS

You (client) **Merchant (server)**

Let's talk securely.
Here are the protocols and ciphers I understand.

I choose this protocol and these ciphers.
Here is my public key, a cert, a nonce, etc.

Using your public key, I've encrypted a
random symmetric key (and your nonce).

SSL/TLS

All subsequent secure messages are sent using the symmetric key and a keyed hash for message authentication.

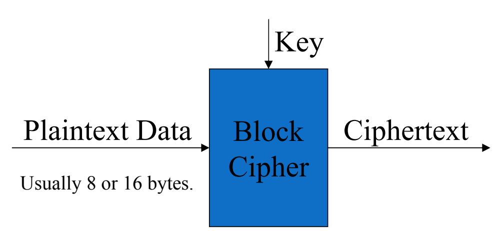
Symmetric Ciphers

Private-key (symmetric) ciphers are usually divided into two classes.

- **Block ciphers**
- **Stream ciphers**

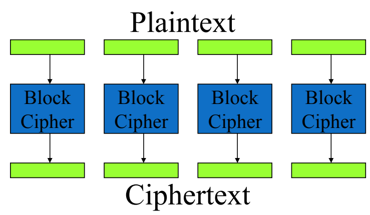
Block Ciphers

DES, AES, RC2, RC5, etc.



Block Cipher Modes

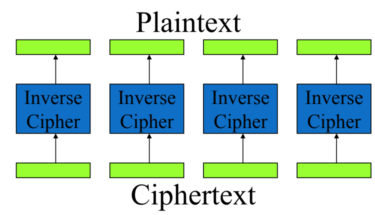
Electronic Code Book (ECB) Encryption:



Slides by Josh Benaloh

Block Cipher Modes

Electronic Code Book (ECB) Decryption:



Slides by Josh Benaloh

Block Cipher Integrity

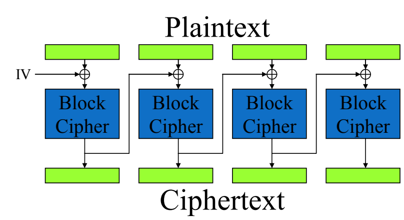
With ECB mode, identical blocks will have identical encryptions.

This can enable replay attacks as well as re-orderings of data. Even a passive observer may obtain statistical data.

Slides by Josh Benaloh

Block Cipher Modes

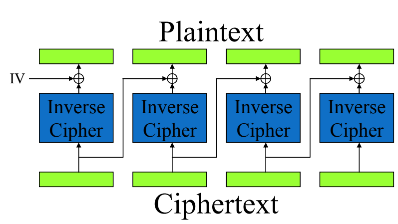
Cipher Block Chaining (CBC) Encryption:



Slides by Josh Benaloh

Block Cipher Modes

Cipher Block Chaining (CBC) Decryption:



Slides by Josh Benaloh

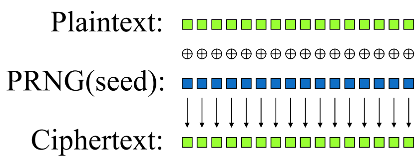
Stream Ciphers

RC4, SEAL, etc.

- Use the key as a seed to a pseudo-random number-generator (PRNG).
- Take the stream of output bits from the PRNG and XOR it with the plaintext to form the ciphertext.

Slides by Josh Benaloh

Stream Cipher Encryption



Stream Cipher Integrity

- It is easy for an adversary (even one who can't decrypt the ciphertext) to alter the plaintext in a known way.

Bob to Bob's Bank:
Please transfer $0,000,002.00 to the account of my good friend Alice.

Stream Cipher Integrity

- It is easy for an adversary (even one who can't decrypt the ciphertext) to alter the plaintext in a known way.

Bob to Bob's Bank:
Please transfer $1,000,002.00 to the account of my good friend Alice.

One-Way Hash Functions

The idea of a *check sum* is great, but it is designed to prevent accidental changes in a message.

For cryptographic integrity, we need an integrity check that is resilient against a smart and determined adversary.

One-Way Hash Functions

MD4, MD5, SHA-1, SHA-256, etc.

A *one-way hash function* is a function

$H : \{0,1\}^* \rightarrow \{0,1\}^k$ (typically k is 128 or 160)

such that, given an input value x, one can't find $x' \neq x$ such that $H(x) = H(x')$.

One-Way Hash Functions

There are many measures for one-way hashes.

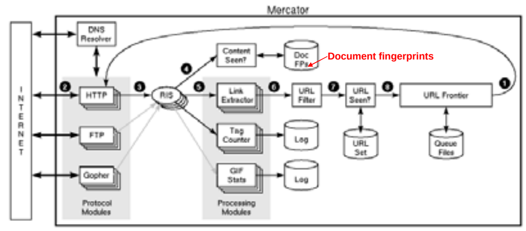
- Non-invertability: given y, it's difficult to find any x such that $H(x) = y$.
- Collision-intractability: one cannot find a pair of values $x' \neq x$ such that $H(x) = H(x')$.

One-Way Hash Functions

- When using a stream cipher, a hash of the message can be appended to ensure integrity. [Message Authentication Code]
- When forming a digital signature, the signature need only be applied to a hash of the message. [Message Digest]

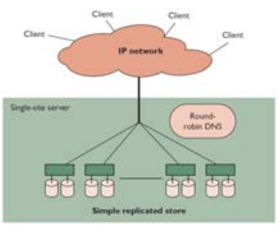
Slides by Josh Benaloh

Structure of Mercator Spider

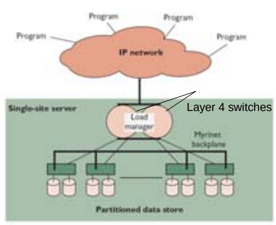


1. Remove URL from queue
2. Simulate network protocols & REP
3. Read w/ RewindInputStream (RIS)
4. Has document been seen before? (checksums and fingerprints)
5. Extract links
6. Download new URL?
7. Has URL been seen before?
8. Add URL to frontier

Common Types of Clusters



Simple Web Farm



Search Engine Cluster

Inktomi (2001) Supports programs (not users) Persistent data is partitioned across servers:
⇑ capacity, but ⇓ data loss if server fails

From: Brewer *Lessons from Giant-Scale Services*

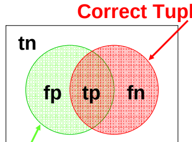
CSE 454 Overview

	Search Engines
HTTP, HTML, Scaling & Crawling	
Cryptography & Security	

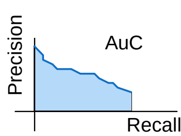
The Precision / Recall Tradeoff

- **Precision** $\frac{tp}{tp + fp}$
  - Proportion of selected items that are correct
- **Recall** $\frac{tp}{tp + fn}$
  - Proportion of target items that were selected
- **Precision-Recall curve**
  - Shows tradeoff

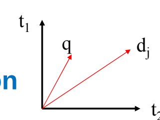
Correct Tuples



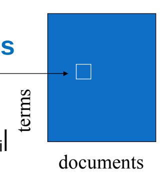
Tuples returned by System



- **Vector Space Representation**
  - Dot Product as Similarity Metric



- **TF-IDF for Computing Weights**
  - $w_{ij} = f(i,j) * \log(N/n_i)$
  - Where q = … $word_i$…
  - N = |docs| n_i = |docs with $word_i$|



- **How Process Efficiently?**

Thinking about Efficiency

- **Clock cycle: 2 GHz**
  - Typically *completes* 2 instructions / cycle
    - ~10 cycles / instruction, but pipelining & parallel execution
  - Thus: 4 billion instructions / sec
- **Disk access: 1-10ms**
  - Depends on seek distance, published average is 5ms
  - Thus perform 200 seeks / sec
  - (And we are ignoring rotation and transfer times)
- **Disk is *20 Million* times slower !!!**

Inverted Files for Multiple Documents

LEXICON

WORD	NDOCS	PTR
jezebel	20	
jezer	3	
jezerit	1	
jeziah	1	
jeziel	1	
jezliah	1	
jezoar	1	
jezrahliah	1	
jezreel	39	

DOCID OCCUR POS 1 POS 2 . . .

"jezebel" occurs
6 times in document 34,
3 times in document 44,
4 times in document 56 . . .

34	6	1	118	2087	3922	3981	5002
44	3	215	2291	3010			
56	4	5	22	134	992		

...

566	3	203	245	287

67	1	132

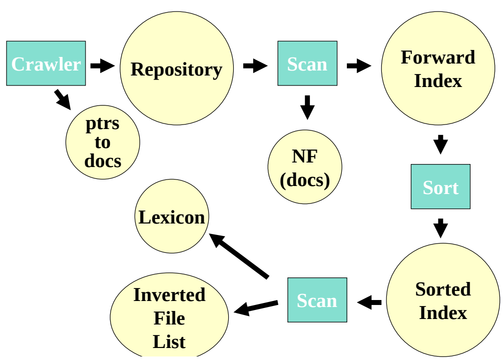
. . .

107	4	322	354	381	405		
232	6	15	195	248	1897	1951	2192
677	1	481					
713	3	42	312	802			

OCCURENCE INDEX

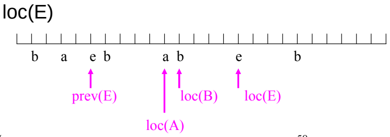
- **One method. Alta Vista uses alternative**

How Inverted Files are Created



AltaVista

- **Basic Framework**
  - Flat 64-bit address space
  - Index Stream Readers: Loc, Next, Seek, Prev
  - Constraints
- **Let E be ISR for word** enddoc
- **Constraints for conjunction a AND b**
  - prev(E) ≤ loc(A)
  - loc(A) ≤ loc(E)
  - prev(E) ≤ loc(B)
  - loc(B) ≤ loc(E)



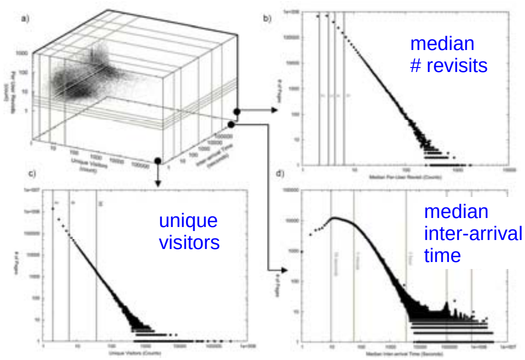
Beyond Size - Search User Interfaces

- **Specialized Search**
- **Suggestions**
- **Spelling Correction**

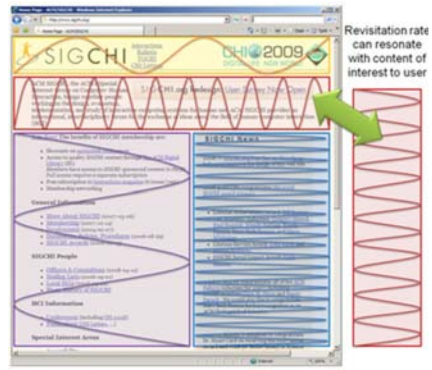
50-80% Page Views = Revisits

Cluster Group	Name	Shape	Description
Fast Revisits (< hour) 23611 pages	F1		Pornography & Spam, Hub & Spoke, Shopping & Reference Web sites, Auto refresh, Fast monitoring
	F2		
	F3		
	F4		
	F5		
Medium (hour to day) 9421 pages	M1		Popular homepages, Communication, .edu domain, Browser homepages
	M2		
Slow Revisits (> day) 18422 pages	S1		Entry pages, Weekend activity, Search engines used for revisitation, Child-oriented content, Software updates
	S2		
	S3		
	S4		
Hybrid 3334 pages	H1		Popular but infrequently used, Entertainment & Hobbies, Combined Fast & Slow

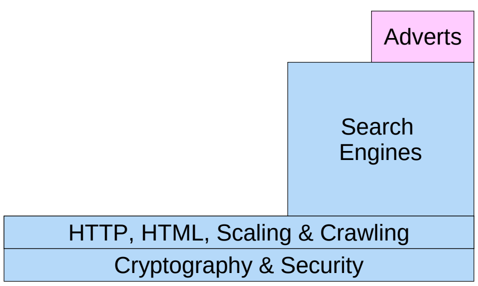
Revisitation



Resonance



CSE 454 Overview



A-B testing

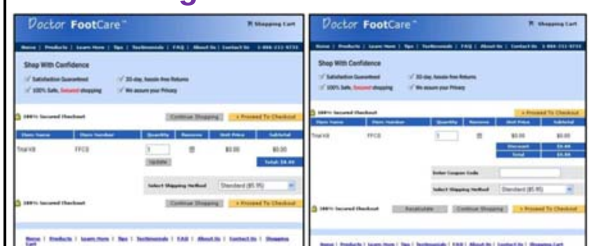
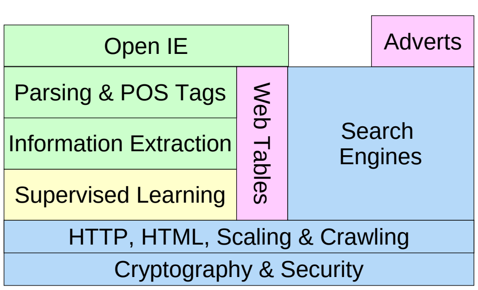


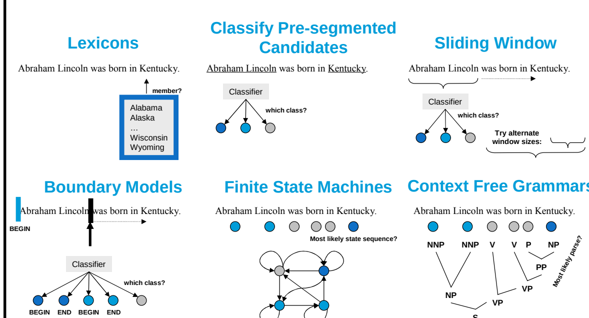
Figure 1: Variant A on left, Variant B on right.
Can you guess which one has a higher conversion rate and whether the difference is significant?

Nine Changes in Site Above

CSE 454 Overview



Landscape of IE Techniques: Models



Any of these models can be used to capture words, formatting or both.

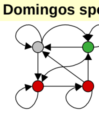
Slides from Cohen & McCallum

IE with Hidden Markov Models

Given a sequence of observations:

Yesterday Pedro Domingos spoke this example sentence.

and a trained HMM:



- person name
- location name
- background

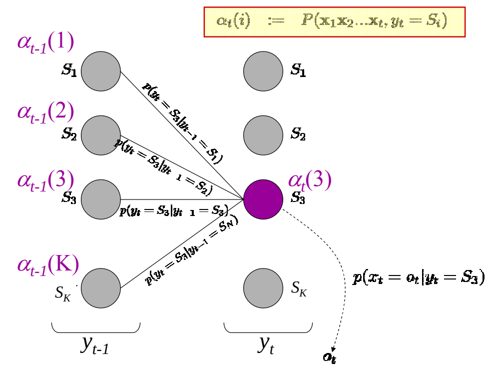
Find the most likely state sequence: (Viterbi) $\arg\max_{\vec{s}} P(\vec{s}, \vec{o})$

Any words said to be generated by the designated "person name" state extract as a person name:

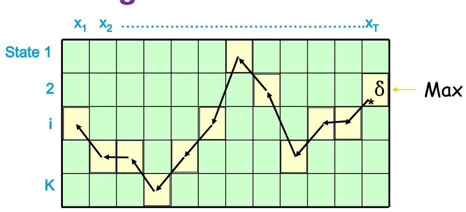
Person name: Pedro Domingos

Slide by Cohen & McCallum

The Forward Algorithm



Terminating Viterbi



How did we compute δ*? $\text{Max}_i\ \delta_{T-1}(i) * P_{trans} * P_{obs}$

Now Backchain to Find Final Sequence

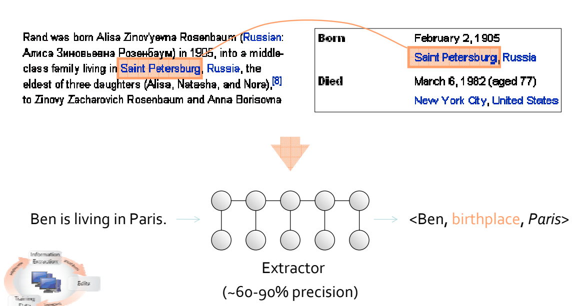
Time: $O(K^2T)$
Space: $O(KT)$ ← Linear in length of sequence

What is Open Information Extraction?

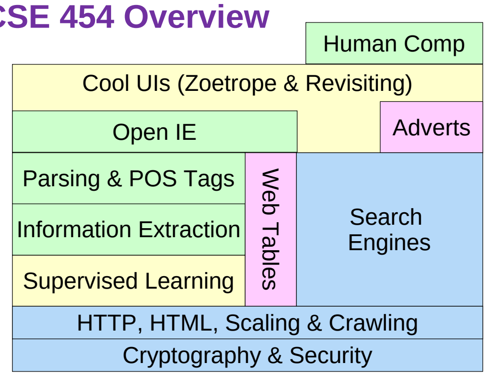
	Traditional IE	Open IE
Input	Corpus + Labeled Data	Corpus + Domain-Independent Methods
Relations	Specified In Advance	Discovered Automatically
Complexity	$O(D * R)$ D documents, R relations	$O(D)$ D documents

Self-Supervised Learning from WIkipedia

[Wu et.al. CIKM'07]

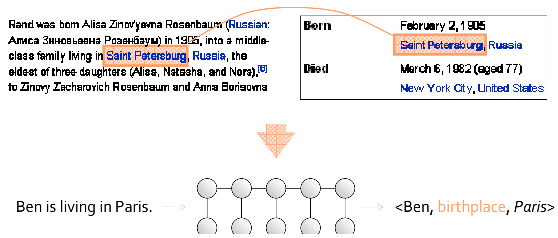


CSE 454 Overview



Self-Supervised Learning from WIkipedia

[Wu et.al. CIKM'07]



(~60-90% precision)

How Motivate People to Help?

- Money
- Fun
- Altruism
- Esteem
- Self-Interest

Altruism

WIKIPEDIA
The Free Encyclopedia

Self-Esteem

Self-Interest

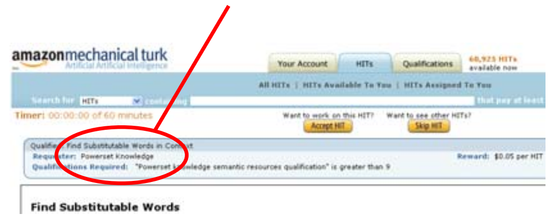
How Motivate People to Help?

- Pay them…

amazonmechanical turk
beta
Artificial Artificial Intelligence

Built in 1770 by Wolfgang von Kempelen

Powerset



Find Substitutable Words

In the sentence below, what words or phrases could replace the **bolded** word without changing the meaning? F

Example:
In most countries **children** are required by law to attend school.

You might enter:
kid
youngster
pupil
young person

Try to enter single words or short phrases like "water bottle" or "post office." You are encouraged to use the ta station".
Avoid descriptive phrases, e.g. "a container you drink out of," or "a place you mail things from" unless you abs
Further, tell us how easy or difficult it is to assign one of several possible meanings for the **bolded** word in the

Your sentence is: The term silver dollar is often used for any large white metal coin issued by the United States with a **face** value of one dollar ; although purists insist that a dollar is not silver unless it contains some of that metal .

Enter *one term* per box.

$0.05

Fast & Cheap, but is it Good?

[Snow *et al.* EMNLP-08

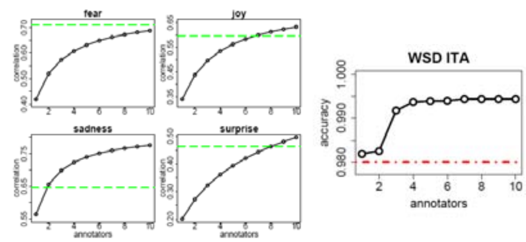


Figure 1: Non-expert correlation for affect recognition

How Cheap + Fast?

[Snow *et al.* EMNLP-08

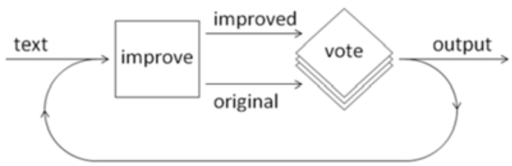
In our experiment we ask for 10 annotations each of the full 30 word pairs, at an offered price of $0.02 for each set of 30 annotations (or, equivalently, at the rate of 1500 annotations per USD). The most surprising aspect of this study was the speed with which it was completed; the task of 300 annotations was completed by 10 annotators in less than 11 minutes …

1724 annotations / hour.

Who are those Turkers?

Complex Jobs

- Casting Words
- TurkIt



Iterative Improvement

A parial view of a pocket calculator together with some coins and a pen.

Iterative Improvement

A parial view of a pocket calculator together with some coins and a pen.

Iterative Improvement

"A close-up photograph of the following items:

A CASIO multi-function, solar powered scientific calculator.

A blue ball point pen with a blue rubber grip and the tip extended.

Six British coins; two of £1 value, three of 20p value and one of 1p value.

Seems to be a theme illustration for a brochure or document cover treating finance - probably personal finance."

Motivating People

- Money
- Fun

IMAGE SEARCH ON THE WEB

USES FILENAMES AND HTML TEXT

Slides by Luis von Ahn

ACCESSIBILITY

LESS THAN 10% OF THE WEB IS ACCESSIBLE TO THE VISUALLY IMPAIRED

REASON: MOST IMAGES DON'T HAVE A CAPTION

LABELING IMAGES WITH WORDS

FACE
MAN
SUPER SEXY

STILL A COMPLETELY OPEN PROBLEM

DESIDERATA

A METHOD THAT CAN LABEL ALL IMAGES ON THE WEB FAST AND CHEAP

THE ESP GAME

TWO-PLAYER ONLINE GAME

PARTNERS DON'T KNOW EACH OTHER AND CAN'T COMMUNICATE

OBJECT OF THE GAME: TYPE THE SAME WORD

THE ONLY THING IN COMMON IS AN IMAGE

THE ESP GAME

PLAYER 1	PLAYER 2
	
GUESSING: CAR	GUESSING: BOY
GUESSING: HAT	GUESSING: CAR
GUESSING: KID	SUCCESS! YOU AGREE ON CAR
SUCCESS! YOU AGREE ON CAR	

THE ESP GAME IS FUN

3.2 MILLION LABELS WITH 22,000 PLAYERS

MANY PEOPLE PLAY OVER 20 HOURS A WEEK

Slides by Luis von Ahn

LABELING THE ENTIRE WEB

5000 PEOPLE PLAYING SIMULTANEOUSLY CAN LABEL ALL IMAGES ON GOOGLE IN 30 DAYS!

INDIVIDUAL GAMES IN YAHOO! AND MSN AVERAGE OVER 10,000 PLAYERS AT A TIME

Slides by Luis von Ahn

9 BILLION MAN-HOURS OF SOLITAIRE WERE PLAYED IN 2003

EMPIRE STATE BUILDING
7 MILLION MAN-HOURS
(6.8 HOURS OF SOLITAIRE)

PANAMA CANAL
20 MILLION MAN-HOURS
(LESS THAN A DAY OF SOLITAIRE)

Slides by Luis von Ahn

GWAP

- **Problem?**

Motivating Vision

Next-Generation Search = Information Extraction
+ Ontology
+ Inference

Which German Scientists Taught at US Universities?

Next-Generation Search

- Information Extraction
  - <Einstein, Born-In, Germany>
  - <Einstein, ISA, Physicist>
  - <Einstein, Lectured-At, IAS>
  - <IAS, In, New-Jersey>
  - <New-Jersey, In, United-States> ...
- Ontology
  - Physicist (x) ➔ Scientist(x) ...
- Inference
  - Einstein = Einstein ...