

ManTech SMA

Computer Forensics and Intrusion Analysis

Fuzzy Hashing



Jesse Kornblum

- **Too Many Pictures**
- **Cryptographic Hashing**
- **Fuzzy Hashing**
- **Demonstration**
- **Issues**
- **Future Research**
- **Questions**

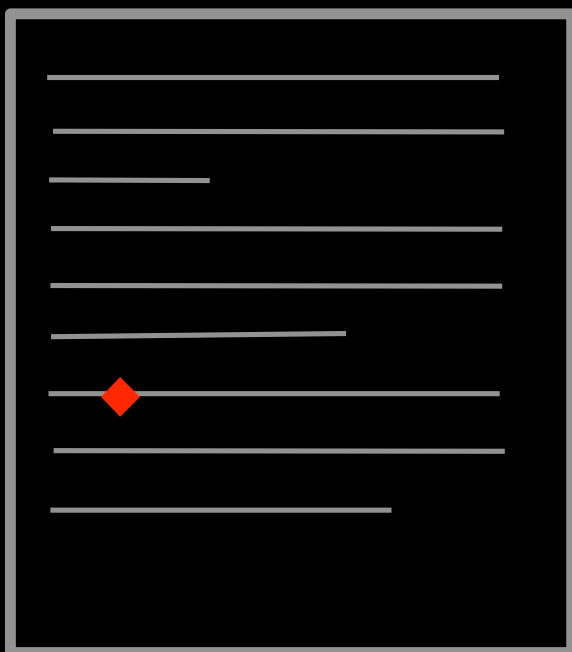
Too Many Pictures

- **Child Pornography cases**
 - **Hundreds of thousands of images**
 - **MD5 not effective for carved files**



Cryptographic Hashing

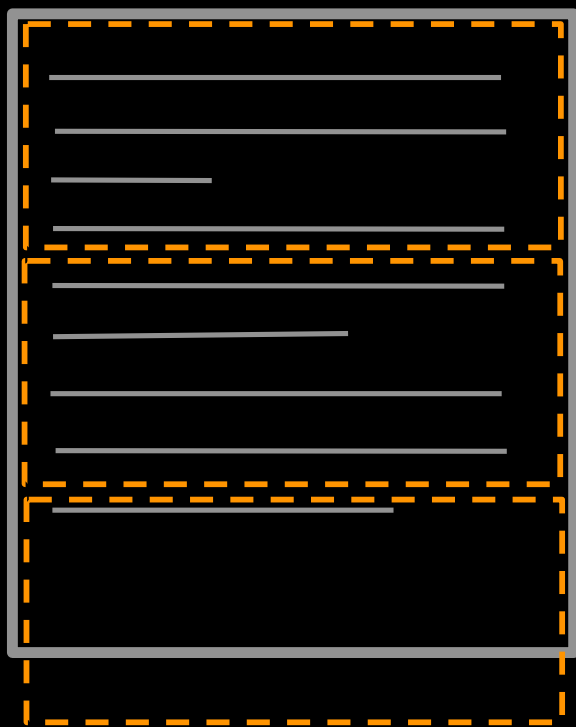
- Algorithms like MD5, SHA-1
- Generate single hash for entire input
- Any change greatly alters hash



~~e41b1427a018fbb264c8adf0a~~
7f48e4b990a2d637fc363efc8

Piecewise Hashing

- Developed for integrity during imaging
- Divide input into equal sized sections and hash
- Insert or delete changes all subsequent hashes



3b152e0baa367a8038373f6df

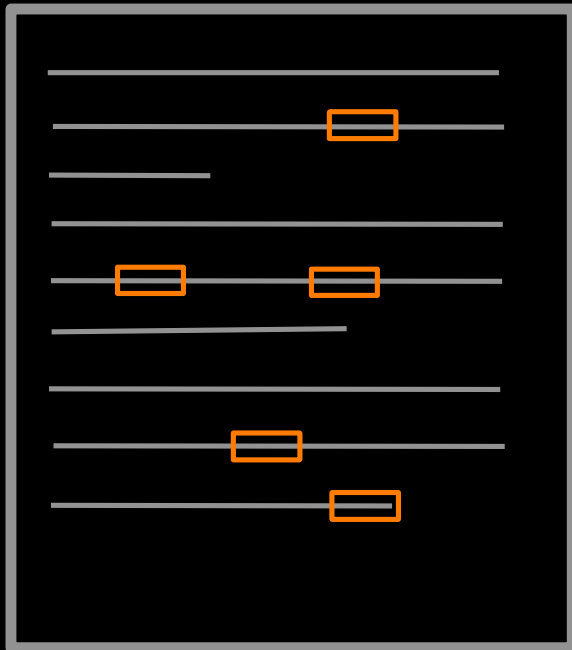


40c39f174a8756a2c266849b



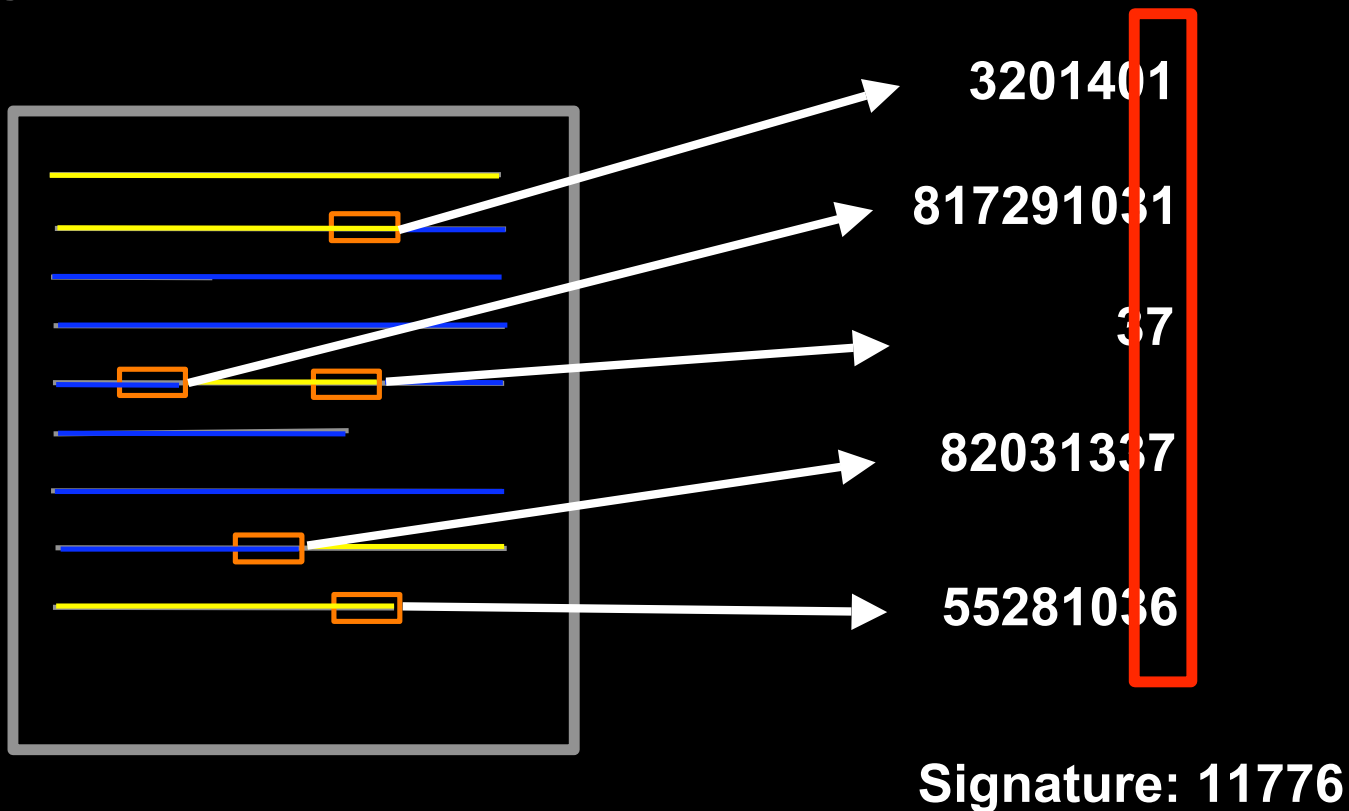
fdb05977978a8bc69ecc46ec

- Function triggered by current context of input



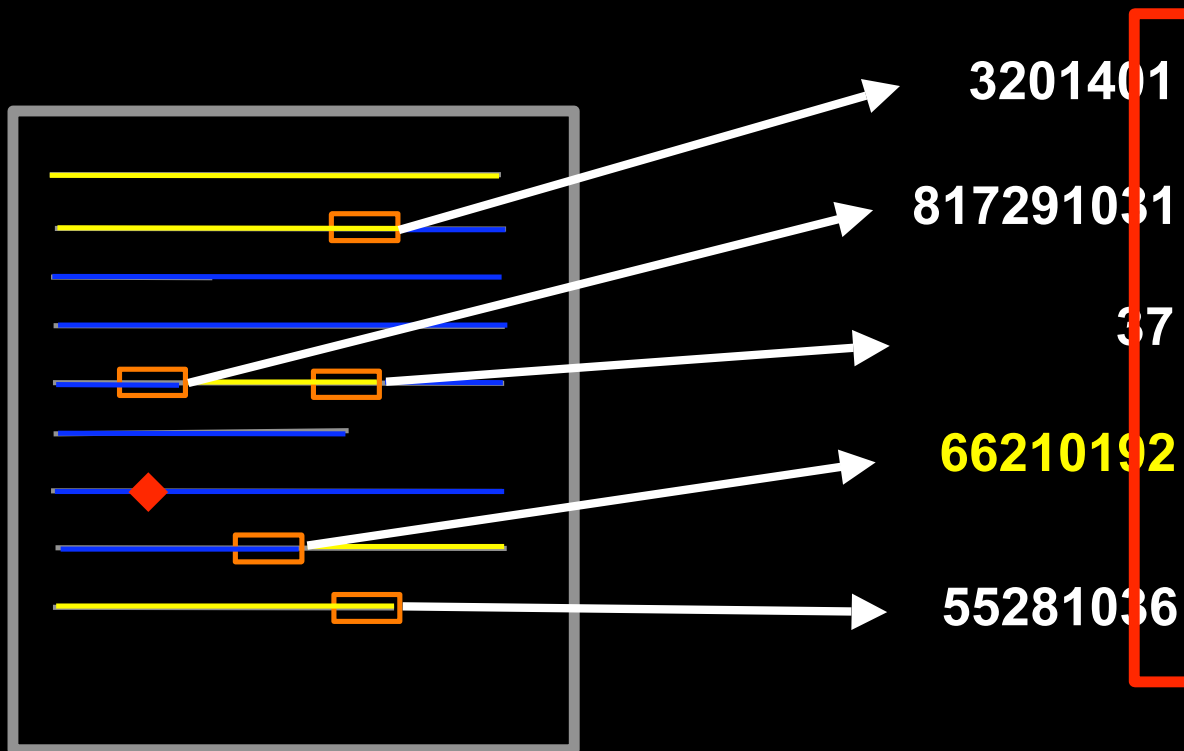
Fuzzy Hashing

- Piecewise hashing with boundaries defined by when rolling hash triggers



Fuzzy Hashing

- Changes only affect one small part of signature



New Signature: 11726

Original: 11776

To update the hash for a byte d:

$y = y - x$

$y = y + \text{size} * d$

$x = x + d$

$x = x - \text{window}[c \bmod \text{size}]$

$\text{window}[c \bmod \text{size}] = d$

$c = c + 1$

$z = z \ll 5$

$z = z \text{ XOR } d$

return (x + y + z)

- Choose triggers such that
 - $\text{rolling_hash}(d) \bmod \text{block_size} = \text{block_size} - 1$
 - Depends only on previous seven bytes
- Example
 - Excerpt from "The Raven" by Edgar Allan Poe
 - Based on file size, triggers on ood and ore

Rolling Hash

Deep into the darkness peering, long I stood there, wondering,
fearing

Doubting, dreaming dreams no mortals ever dared to dream before;

But the silence was unbroken, and the stillness gave no token,

And the only word there spoken was the whispered word,

Lenore?, This I whispered, and an echo murmured back the word,

"Lenore!" Merely this, and nothing more.

Rolling Hash

Deep into the darkness peering, long I **stood** there, wondering,
fearing

Doubting, dreaming dreams no mortals ever dared to dream **before**;

But the silence was unbroken, and the stillness gave no token,

And the only word there spoken was the whispered word,

Lenore?, This I whispered, and an echo murmured back the word,

"**Lenore!**" Merely this, and nothing **more**.

Rolling Hash

Deep into the darkness peering, long I stood **243732**

there, wondering, fearing Doubting, dreaming dreams no mortals
ever dared to dream **before 610**

; But the silence was unbroken, and the stillness gave no token,
And the only word there spoken was the whispered word, **Lenore 3270168**

?, This I whispered, and an echo murmured back the word, "**Lenore 53280**

!" Merely this, and nothing **more. 8381002**



Demonstration

- Needle in a haystack



Known kitty porn



MATCH

- No false positives



Known kitty porn

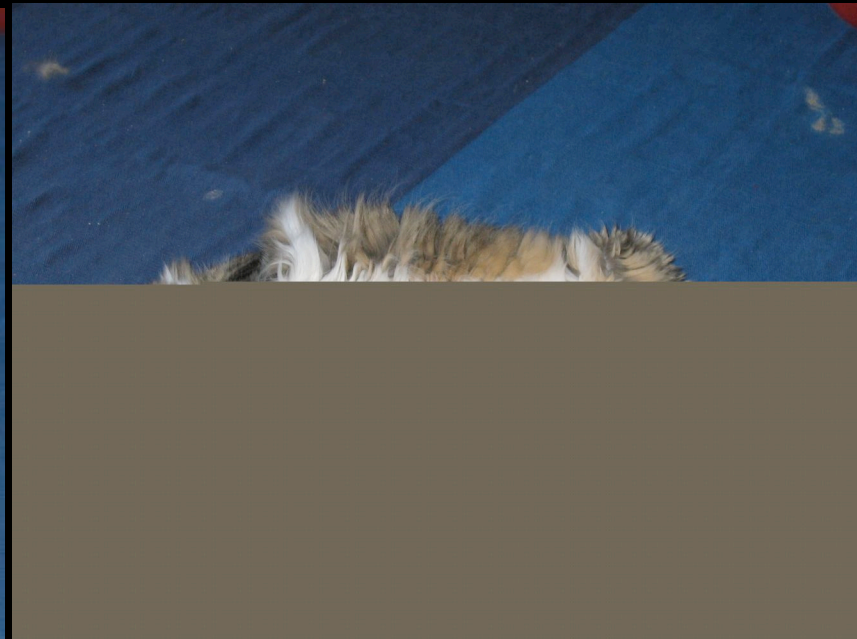


**no match
(00000380.JPG)**

Demonstration



Known kitty porn



MATCH

- File footers



Known kitty porn

MATCH

- **Not perfect**
 - Confused by many small changes throughout input
 - Unable to handle cropping, resizing, and other edits
- **Computationally intensive**
 - 7-10 times slower than MD5
- **No way to sort signatures**
 - Must compare each input to all known signatures



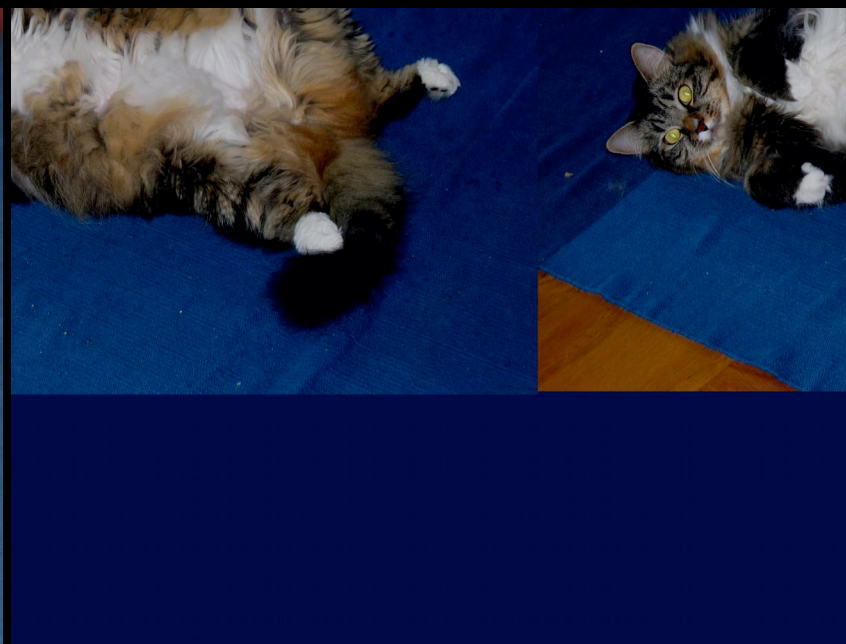
- **Need Error Rate Computation**
 - I am a practitioner, not math geek
- **For court, need error rate**
 - How similar is similar?

- **File Footer Reconstruction**
 - Record headers when making signatures
 - Append recovered footers
- **Need to parse known files**
 - How much information to record?
 - Best storage method?

- File footer Reconstruction



Known kitty porn



File header with footer appended

- **Finding footers and middles**
 - **Current carvers require true footer**
 - **Encase, iLook, Foremost, Scalpel, etc.**
- **The formatted drive scenario**
- **Find blocks that are "JPEgY" or "GIFy"**
 - **Lots of academic research**
 - **No practical tools**

Coming Soon!

- **ssdeep to be published August 14th**
 - **Free software!**
 - **<http://ssdeep.sf.net/>**

Questions



Jesse Kornblum - ManTech CFIA

jesse.kornblum@mantech.com - 410-312-5548