Beyond Fuzzy Hashing

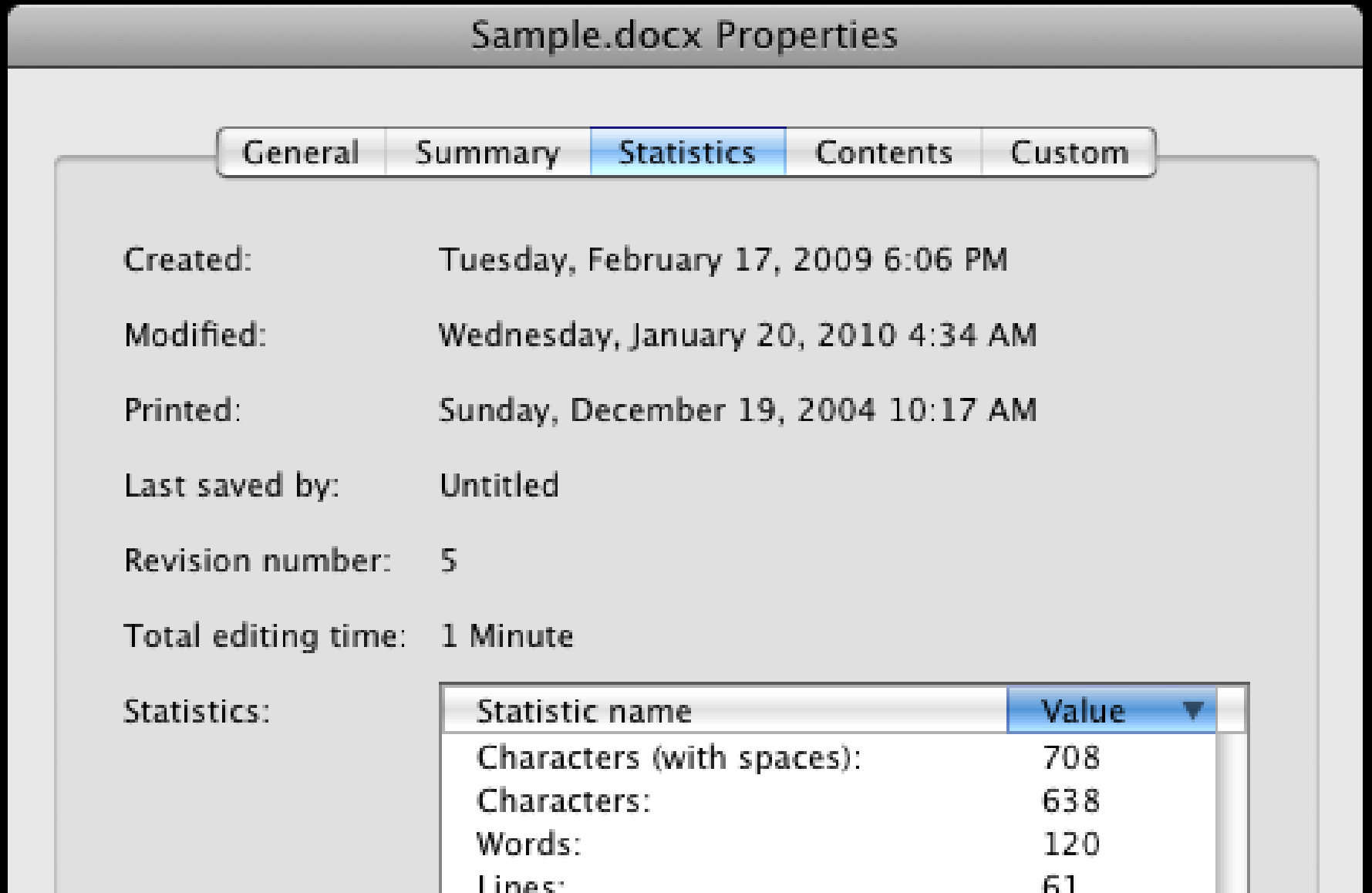
Jesse Kornblum

Outline

- Introduction
- Identical Data
- Similar
- Generic Data
- Fuzzy Hashing
- Image Comparisons
- General Approach
- Documents
- Applications
- Questions

KYRUS TECHNOLOGY

Motivation


Sample.docx Properties
General
Summary
Statistics
Contents
Custom
Created: Tuesday, February 17, 2009 6:06 PM
Modified: Wednesday, January 20, 2010 4:34 AM
Printed: Sunday, December 19, 2004 10:17 AM
Last saved by: Untitled
Revision number: 5
Total editing time: 1 Minute
Statistics:
Statistic name
Value
Characters (with spaces): 708
Characters: 638
Words: 120
Lines: 61

Identical

- A == B
- Difficult for humans (for large documents)
- Easy for computers

- Requires storing the original A and B
  - Big files
  - Could be illegal or private content

KYRUS TECHNOLOGY

Identical

- Cryptographic Hashing shortcut
- MD5 and friends
- If MD5(A) == MD5(B) then A == B*
  - * to within a high degree of certainty
  - Chance of random collision is 2^{-128}, or about 10^{-38}

- Hashes signatures are small
- Impossible to recover input from signature

KYRUS TECHNOLOGY

Identical Data

- Cryptographic hashes are spoiled by even a single byte difference in the input
- Very similar things have wildly different cryptographic hashes

Image courtesy of Flickr user krystalchu and used under Create Commons license.

KYRUS TECHNOLOGY

Similar Data

- What does it mean for two things to be similar?

KYRUS TECHNOLOGY

Similar Data

- Depends on:
  - The kind of things be compared
  - How they're being compared

- Pictures
  - Looks the same
  - Same subject
  - Same location
  - Taken by the same camera
  - Taken by the same person

KYRUS TECHNOLOGY

Generic Data

- Don’t care about the structure
- Assume any differences are byte aligned
  - No insertions or deletions

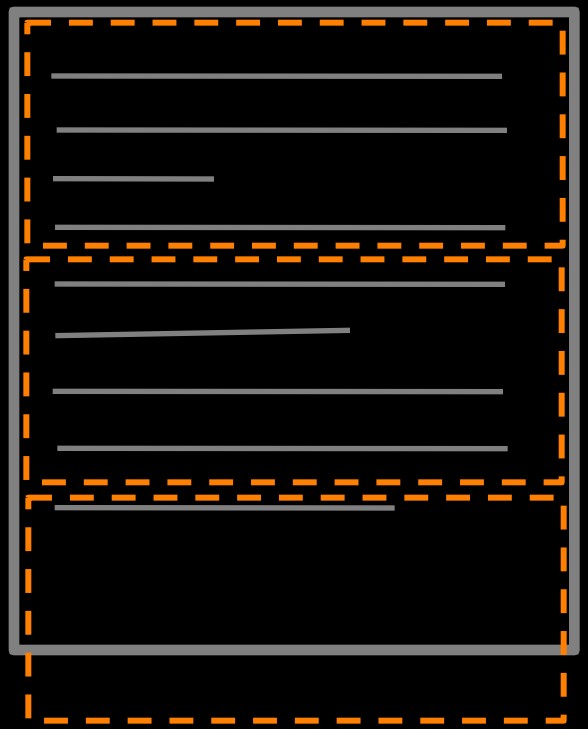
The quick brown fox jumped over the lazy dog. How much wood could

The quick brown fax jumped over the lazy dog. How much good could

KYRUS TECHNOLOGY

Piecewise Hashing

- Developed by Nick Harbour
  - Designed for errors in drive imaging
  - Found in dcfldd, dc3dd, md5deep, etc
- Divide input into fixed size sections and hash separately



3b152e0baa367a8038373f6df

40c39f174a8756a2c266849b

fdb05977978a8bc69ecc46ec

KYRUS TECHNOLOGY

Bytewise Comparison

The quick brown fox jumped over the lazy dog. How much wood could

The quick brown fax jumped over the lazy dog. How much good could

97% of the data is identical

KYRUS TECHNOLOGY

Bytewise Comparison

- Scenario:
  - Image computer
  - Lose control of computer
  - Regain control, image again

- 97% of the the data on the drive was identical
- What changed?

KYRUS TECHNOLOGY

Visual Representation

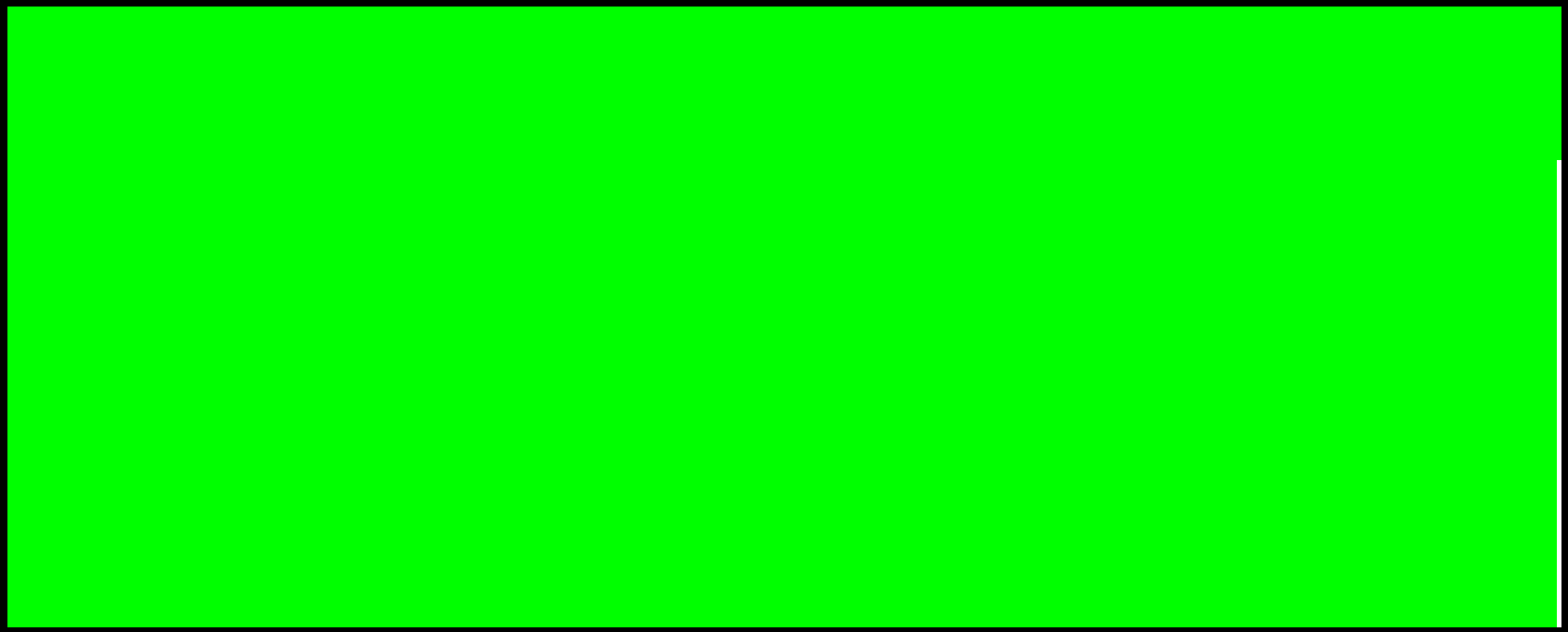
- Compare the data in each block
  - Can specify block size later
- If identical, add a green pixel
- If different, add a red pixel

The quick brown fox jumped over the lazy dog. How much wood could

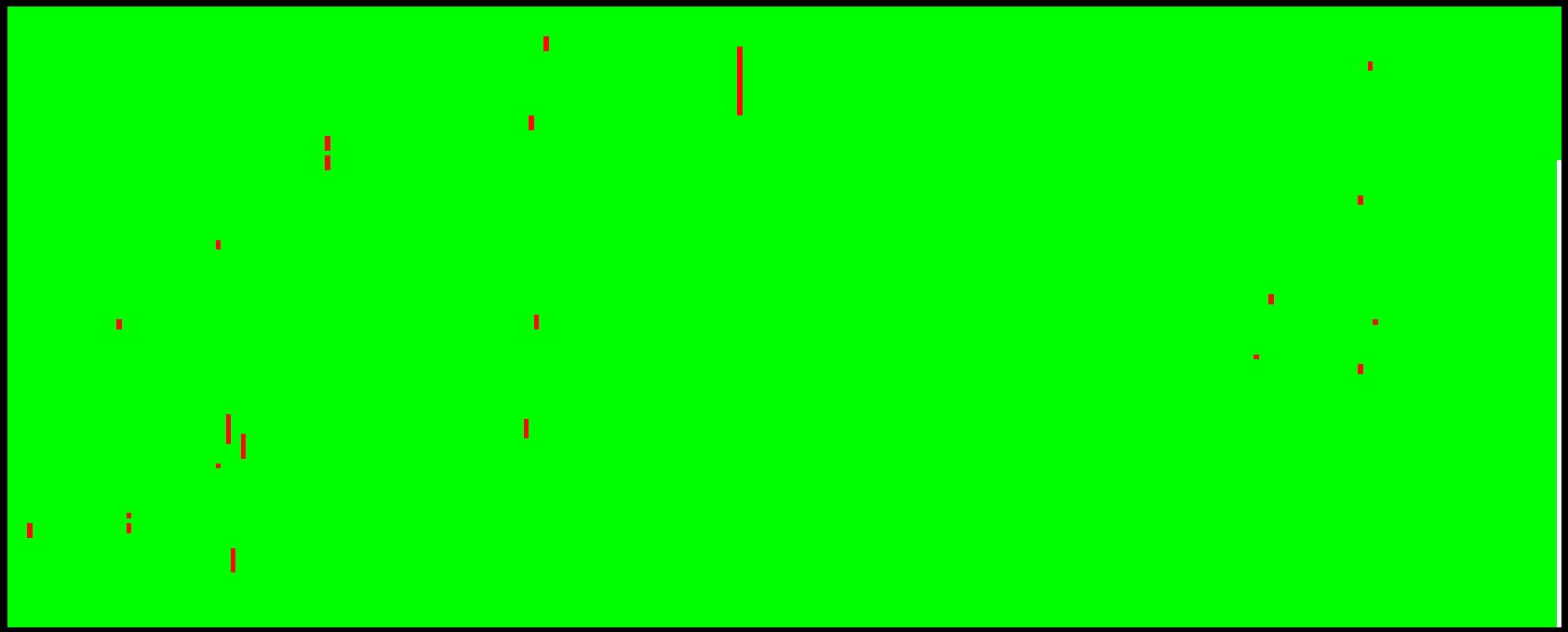
The quick brown fax jumped over the lazy dog. How much good could

KYRUS TECHNOLOGY

No changes made



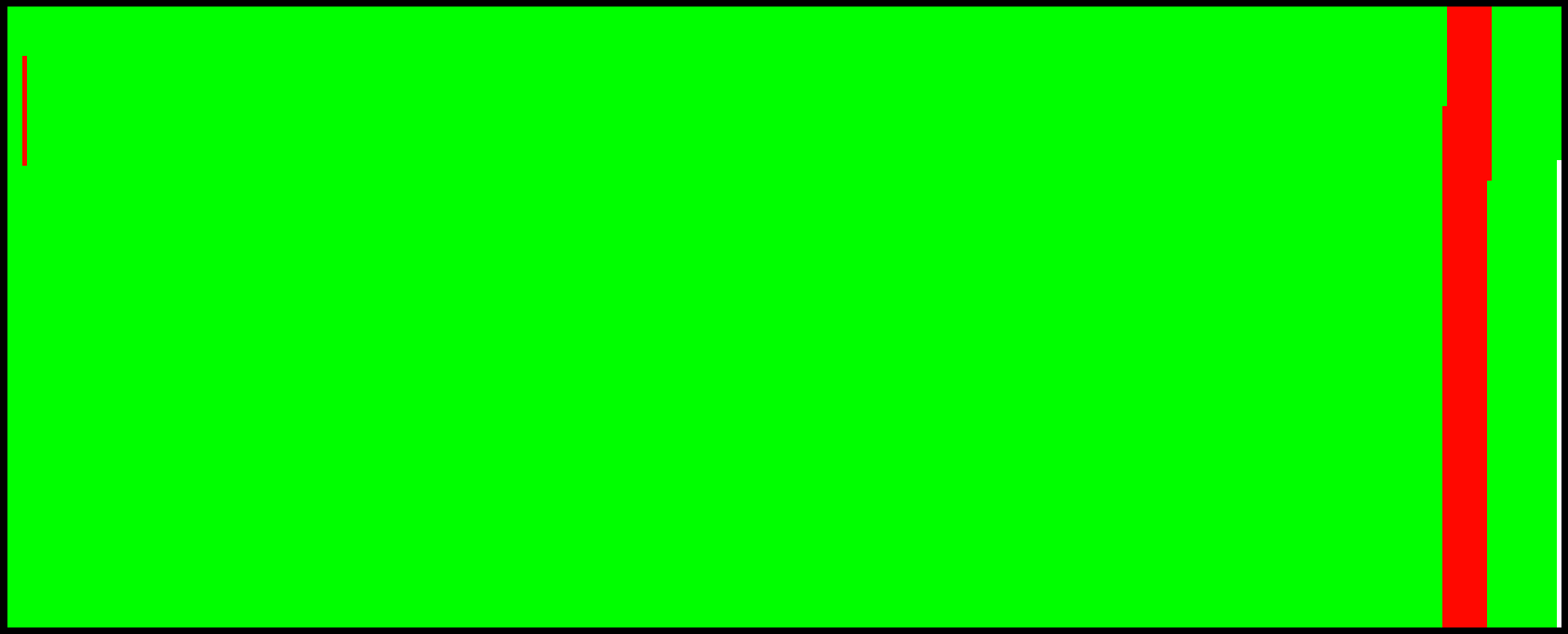
Powered on and off



97% of the data is identical

KYRUS TECHNOLOGY

Actual Result



97% of the data is identical

KYRUS TECHNOLOGY

Generic Data

- What if the data is not byte-aligned?

The quick brown fox jumped over the lazy dog. How much wood could

The quick brown fox jumped up and over the lazy dog. How much wood

KYRUS TECHNOLOGY

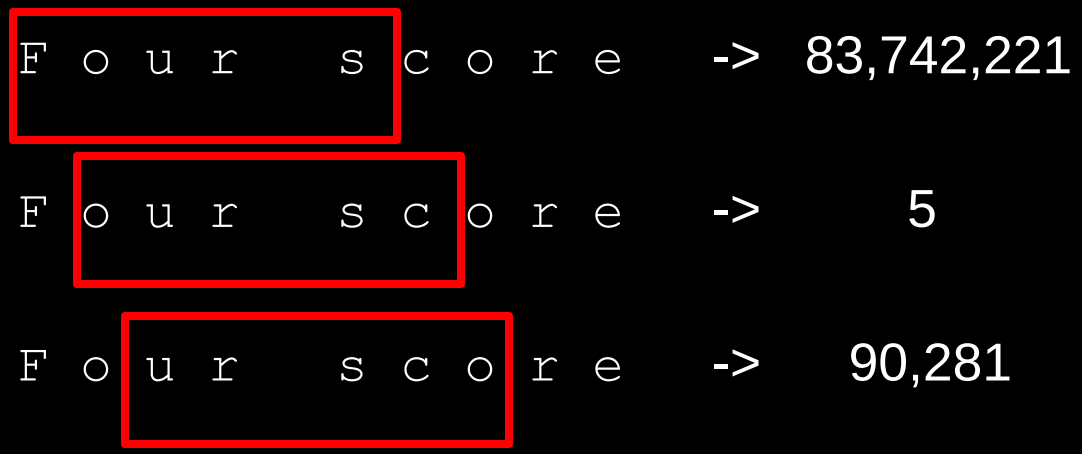
Disclaimer

- I didn’t invent this math
- Originally Dr. Andrew Tridgell
  - Samba
  - rsync was part of his thesis
  - Modified slightly for spamsum
    - Spam detector in his “junk code” folder
- There is LOTS of academic research into ‘similarity’

KYRUS TECHNOLOGY

Fuzzy Hashing

- Combination of a rolling hash and traditional hash
- Rolling hash looks only at last few bytes



- If rolling hash mod block size = 1, it's a trigger point
  - Block size guessed based on file size

KYRUS TECHNOLOGY

Fuzzy Hashing

- Compute traditional hash while processing file
- On each trigger point, record value
- Reset traditional hash and continue

- Example
  - Excerpt from "The Raven" by Edgar Allan Poe
  - Triggers on ood and ore

KYRUS TECHNOLOGY

Fuzzy Hashing

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

KYRUS TECHNOLOGY

Fuzzy Hashing

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

KYRUS TECHNOLOGY

Fuzzy Hashing

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

KYRUS TECHNOLOGY

Fuzzy Hashing

Deep into the darkness peering, long I stood **28163**

there, wondering, fearing Doubting, dreaming dreams no mortals ever dared to dream before **491522**

; But the silence was unbroken, and the stillness gave no token,
And the only word there spoken was the whispered word, Lenore **57**

?, This I whispered, and an echo murmured back the word, "Lenore **145410213**

!" Merely this, and nothing more **738210**

KYRUS TECHNOLOGY

Fuzzy Hashing

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

Signature = 32730

KYRUS TECHNOLOGY

Fuzzy Hashing

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

KYRUS TECHNOLOGY

Fuzzy Hashing

Deep into the darkness peering, long I stood

there, wondering, I AM THE LIZARD KING!!!1! 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

KYRUS TECHNOLOGY

Fuzzy Hashing

Deep into the darkness peering, long I stood **28163**

there, wondering, I AM THE LIZARD KING!!!1! fearing Doubting,
dreaming dreams no mortals ever dared to dream before **381739**

; But the silence was unbroken, and the stillness gave no token,
And the only word there spoken was the whispered word, Lenore **57**

?, This I whispered, and an echo murmured back the word, "Lenore
145410213

!" Merely this, and nothing more
738210

KYRUS TECHNOLOGY

Fuzzy Hashing

Deep into the darkness peering, long I stood

there, wondering, I AM THE LIZARD KING!!!1! 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

Original Signature = 32730
New Signatures = 39730

KYRUS TECHNOLOGY

Demonstration

WARNING:

EXPLICIT IMAGERY

KYRUS TECHNOLOGY

Demonstration

KYRUS TECHNOLOGY

Corrupted File

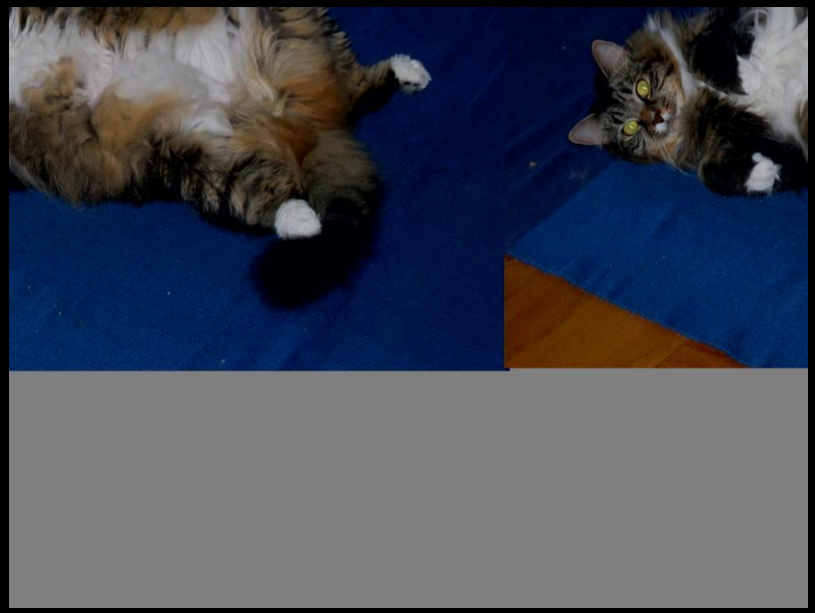
MATCH!

KYRUS TECHNOLOGY

File Footer

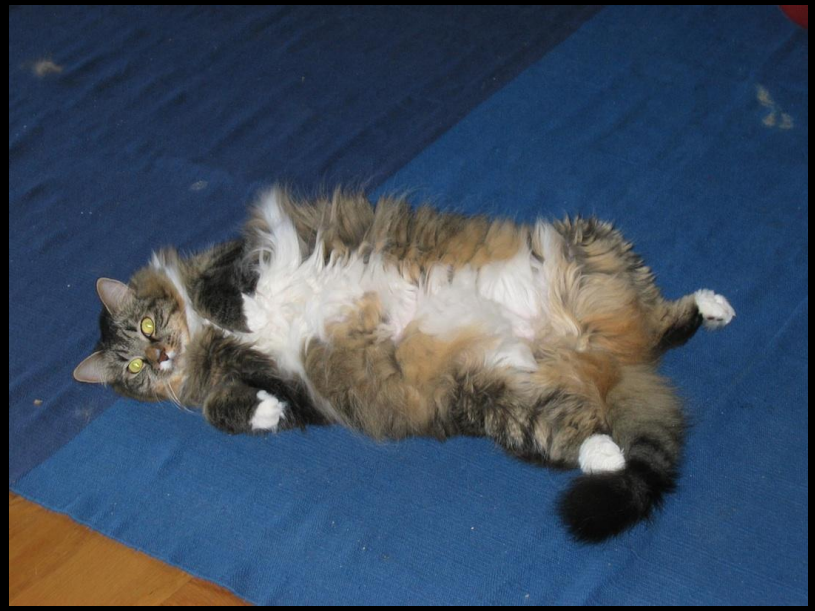
MATCH!

File Footer



MATCH!

Where Fuzzy Hashing Fails



Do not match

Comparing Pictures

- Visual Comparisons
- Easy for humans
- Somewhat difficult for computers
- Content Based Image Retrieval (CBIR)
  - There are companies tripping over themselves to do this
  - Nobody has it quite nailed yet
- A free product is ImgSeek
  - http://imgseek.net/
- Search Styles
  - Search by drawing
  - Search by example

KYRUS TECHNOLOGY

Search by Example

Query

Result

Image courtesy Flickr user andrewbain and licensed under the Creative Commons

KYRUS TECHNOLOGY

Comparing Pictures

- Non-visual comparisons
  - EXIF information
  - Same camera
    - Looks at imperfections in CCDs
    - Requires thousands of pictures and some mathy stuff

KYRUS TECHNOLOGY

General Approach

1. Feature Selection
2. Feature Extraction
3. Comparison
4. Clustering (optional)
5. Profit!

KYRUS TECHNOLOGY

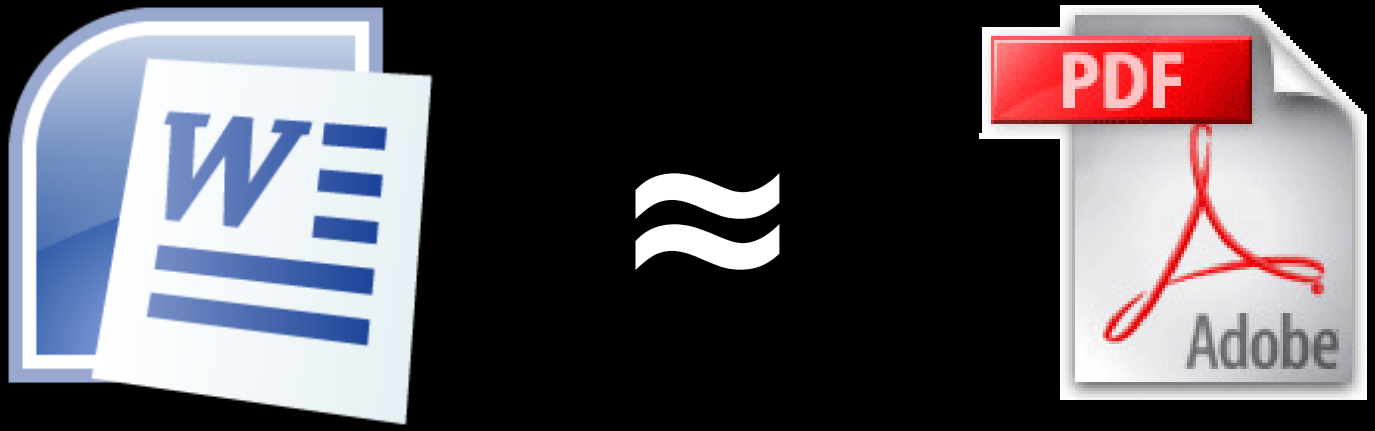
Documents

- File Format Similarity
- Need to parse documents
  - Non-trivial task for complex formats
  - Development time for Word is measured in man-centuries
- Simplified parser can work in some cases
  - Ignore timestamps, for example
- OLE is a filesystem
- DOCX is a zip file of XML documents

KYRUS TECHNOLOGY

Documents

- Content similarity
  - Same document in Word and PDF formats
- Features are text



KYRUS TECHNOLOGY

Programs

- What makes programs similar?

KYRUS TECHNOLOGY

Programs

- Do the same thing
- "Look and feel"
- Same author
  - Shared libraries
  - Code reuse
- Compilation method

KYRUS TECHNOLOGY

Video and Music

- Generally solved
  - But costs money
- Audible Magic
  - Used by YouTube
  - Semi-scientific test at http://www.csh.rit.edu/~parallax/

KYRUS TECHNOLOGY

General Approach

1. Feature Selection
2. Feature Extraction
3. Comparison
4. Clustering (optional)
5. Profit!

KYRUS TECHNOLOGY

Outline

- Introduction
- Identical Data
- Similar
- Generic Data
- Fuzzy Hashing
- Image Comparisons
- General Approach
- Documents
- Applications
- Questions

KYRUS TECHNOLOGY

Questions?

Jesse Kornblum
jesse.kornblum@kyrus-tech.com

KYRUS TECHNOLOGY