

Far More Than You Ever Wanted To Tell

# Hidden Data in Internet Published Documents



2004-12-27

21. Chaos Communication Congress 2004

Steven J. Murdoch & Maximillian Dornseif

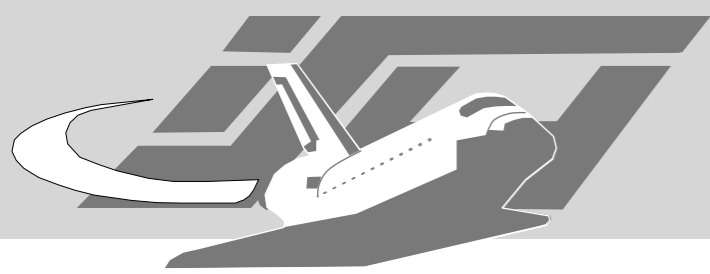
See <http://md.hudora.de/presentations/#hiddendata-21c3>

This Research was supported by the Carnegie Trust for the Universities of Scotland



UNIVERSITY OF  
CAMBRIDGE





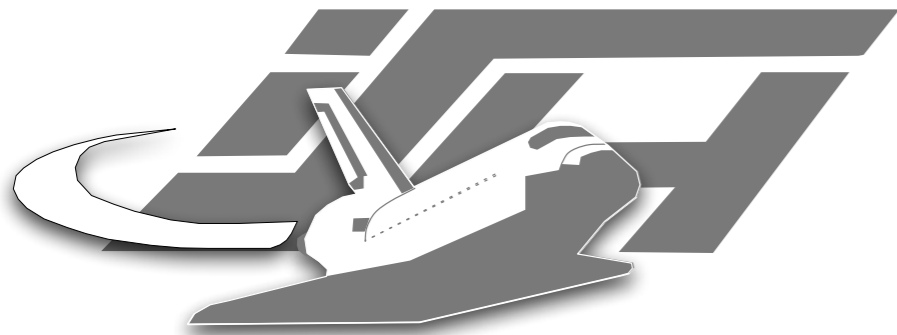
# The Problem

- Software we do not understand and trust
- Complex data formats
  - We are not supposed to understand
  - or we are not willing to understand
- Massive exchange of documents in this complex formats.
- Covert channels everywhere!

# Who we are

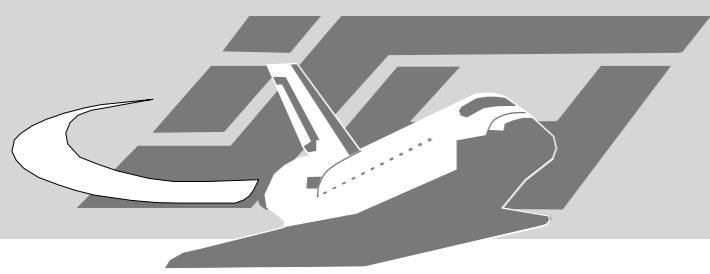


UNIVERSITY OF  
CAMBRIDGE



Laboratory for Dependable Distributed Systems

- Cambridge Security Group - if you don't know them you must have been living under a rock.
- Laboratory for Dependable Distributed Systems at RWTH-Aachen University
  - Founded in late 2003 for theoretical & practical security research, topics include:
    - Security Education
    - Honeypot technology
    - Sensor Networks
  - Notable classes include “Hacker Seminar”, “Hacker Praktikum”, “Pen-Test Praktikum”, “Aachen Summerschool applied IT-Security”, “Computer Forensics”
  - <http://mail-i4.informatik.rwth-aachen.de/mailman/listinfo/lufgtalk/>



# Agenda

- The MS Office Document problem
- Problems with PDFs
- So go for simple formats?
- p0rn!
  - Never trust a girl named .jpeg

# The MS Office Document Problem

Monstrous!

next steps for **North Sea taxation**, the Government's approach would be guided not by short-term factors but by the need for a regime that raises a fair share of revenue and promotes long-term investment in the North Sea. In line with this commitment, the Government has now decided on the reforms it wishes to bring forward.

- It is widely recognised that the present North Sea fiscal regime does not strike the right balance between promoting investment and taking an adequate share of revenue derived from a national resource, ~~adequately reflect the large profits derived from the exploitation of a national resource and so has become unsustainable. To ensure the nation obtains a fair share of the profits from the exploitation of the North Sea,~~ the Government has therefore decided to introduce from today a supplementary charge on profits from the production of oil and gas in the UK and on the UK Continental Shelf (UKCS). The charge will apply at 10 per cent, in addition to the standard corporation tax rate of 30 per cent, and will be deducted from the UKCS calculation. The Government wants to encourage long-term investment in the North Sea. From today, therefore, most capital investment in the North Sea will qualify for an immediate 100 per cent allowance against general corporation tax and the supplementary charge, rather than the 25 per cent allowance currently available.

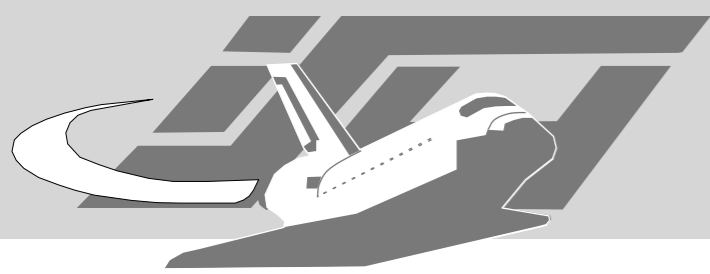
**Accept or Reject Changes**

Changes: No changes selected. Use Find buttons or select a change.

View:
 

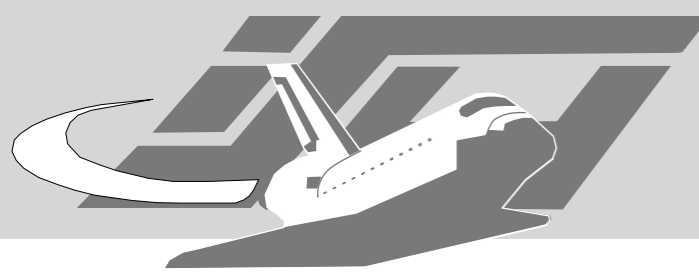
- Changes with highlighting
- Changes without highlighting
- Original

Buttons: Find (left), Find (right), Accept, Reject, Accept All, Reject All, Undo, Close



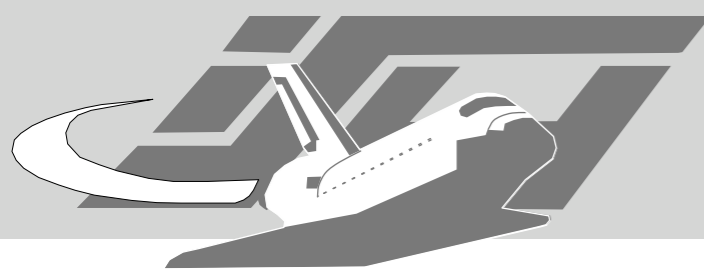
# Tools to investigate

- Antiword
  - Word 2, 6, 7, 97, 2000 and 2002
  - <http://www.winfield.demon.nl/>
- catdoc & xls2csv
  - no support for OLE streams
  - <http://www.45.free.net/~vitus/ice/catdoc/>
- word2x
  - <http://word2x.sourceforge.net/>



- Laola “is a collection of documentations and perl programs dealing with binary file formats of Windows program documents.”
- Contains
  - **Iclean** - Laola Clean: “Saves the trash sections of e.g. Word 6, Word 7 or Excel documents to own files.”
  - **Idat** - Laola Display Authress Title: “Lists author, title, creation date and some other information sticked in a laola file. Gets printer information from Excel and Word files.”
  - **Ils** - Laola List: “Lists the structure of a Laola document.”
  - **Elser** - “password resolving, macro decoding”.
- Development ceased for 5 years.
- <http://www.cs.tu-berlin.de/~schwartz/pmh/index.html>

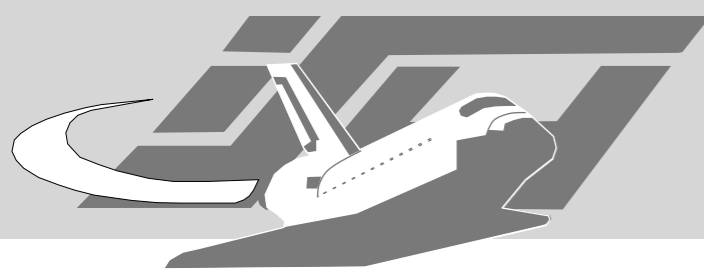




- used by abiword
- tested by kword
- actively developed, but development lines are hard to understand: WordView, wv, wv2, wvWare ...
- Tools
  - wvText, wvHtml
  - wvSummary, wvVersion



<http://wware.sourceforge.net/>

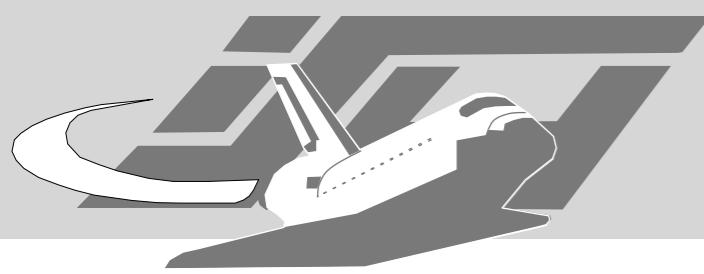


# WordDumper

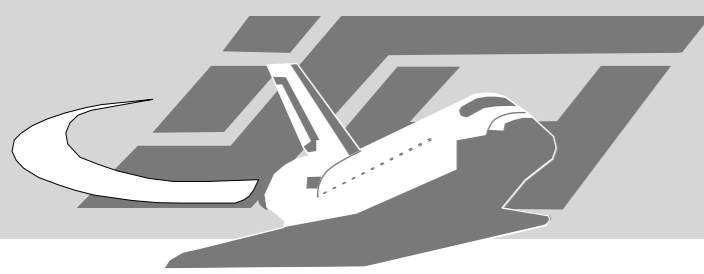


# Problems with PDFs

A document exchange format is becoming a document editing format.

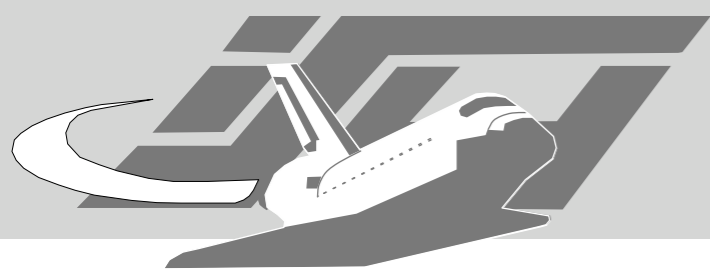


- Looks like an “open standard” ...
- ... but very hard to decode in depth
- Designed for document publishing distribution.
- Very wide deployment
- Adobe is pushing PDF as the default file format of their applications
- The Problem of [REDACTED] / redaction



# Redacted Documents

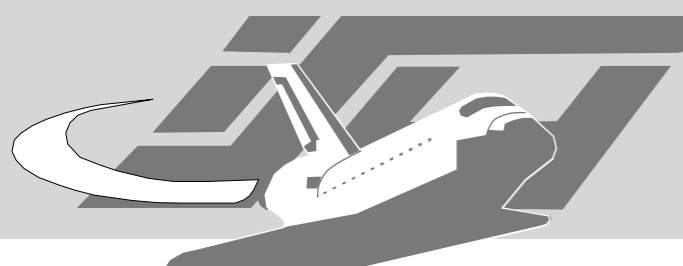
- Documents where the public has “a right to know” ...
  - ... but contain confidential or private information
- Or documents a party is forced to hand over to another party
- Typical classes of documents:
  - court documents
  - public files



# Who is using redaction?

- The “legal community”
- Historians
- Journalists





exported on its behalf. Once the energy was delivered to COB, TransAlta would either use its available transmission rights to schedule the energy to Mid

Ce

# Columbia or

un  
tha  
the



These e

8. WAPA/MWD – Hoover.....22

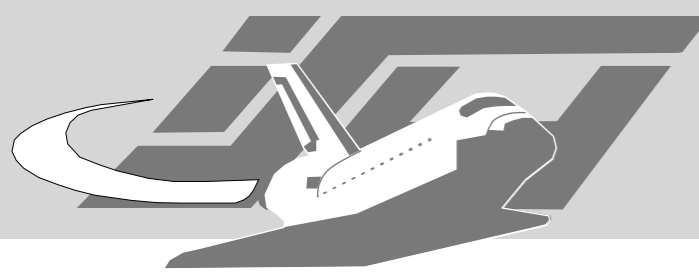
9. 1

# =9.

# DV







# Legal Redaction

## VIA AIRBORNE EXPRESS AND E-MAIL

Singingfish

Attn: **xxx** (DMCA Notification)

2401 Fourth Avenue, Suite 400

Seattle, WA 98121

Copyright\_issues@singingfish.com

Dear **xxx**:

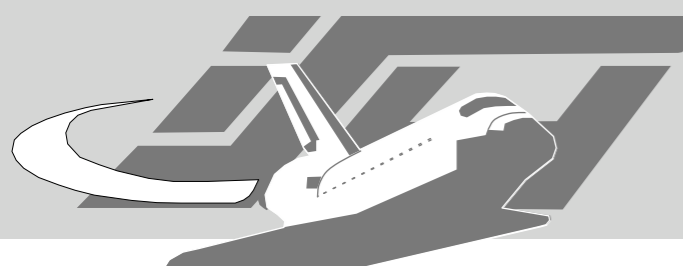
Q10. What is the index and adjustment of your first adjustment?

A. My first adjustment is a reduction of • , based on updating the United States vs. Canadian foreign exchange rate.

This is the Final Report pertaining to the above-referenced special education compliance complaint (the "Complaint") compiled and submitted pursuant to Admin. R. Mont. 10.16.3662. \*\*\*\*\* (the "Complainant") alleges that the \*\*\*\*\* Public Schools (the "District") did not implement the Complainant's child's, \*\*\*\* (the "Student"), Individualized Education Program ("IEP") "properly and in a timely manner." In particular the Complainant alleges that the


3. 30(b)(6) deposition of Defendant regarding MUR 5181, [redacted], the enforcement process, alternate dispute resolution, the Enforcement Priority System, interrogatory responses provided, and documents produced;

In July 2003, NSPI commissioned M to conduct a global coal supply basin survey. M recommended that NSPI should find ways to reduce its dependence on belt, self-unloading vessels, and should develop the capability to unload gearless Panamax and Capesize vessels. M recommended that NSPI would realize a significant reduction in freight cost given access to this much larger fleet of standard gearless bulk carriers, which



# PDF Scrubbing

Sample.pdf



**The Growing Threat to Information Systems Security**

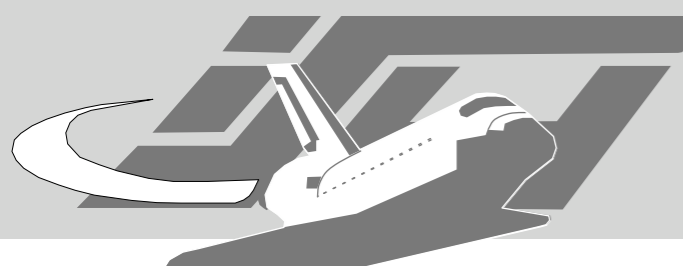
Information technology costs for the Federal Government exceeded \$25 billion in 1995. Within its civilian agencies, the Government employed 120,000 information technology workers, and operated 25,000 medium and large mainframe computers and more than two million individual work stations.<sup>1</sup> The Department of Defense has over two million computers, 10,000 local area networks, and 100 long-distance networks. The civilian sector has a critical responsibility to maintain privacy and services for the public using automated data processing and relying on the National Information Infrastructure. Just as critical to the Department of Defense is its ability to carry out any mission that is dependent on information carried on and supported by the NII. If key responsibilities of both the civilian and military sectors of government are heavily dependent upon an unsecured, potentially unavailable Internet, the Government must address whether this reliance on the NII (and GII) is acceptable and, if so, how to manage the risks involved.

Notwithstanding considerable expenditures on information technology, there exists a widening chasm between the security requirements of and the protection provided for unclassified systems government-wide and those applied to the classified

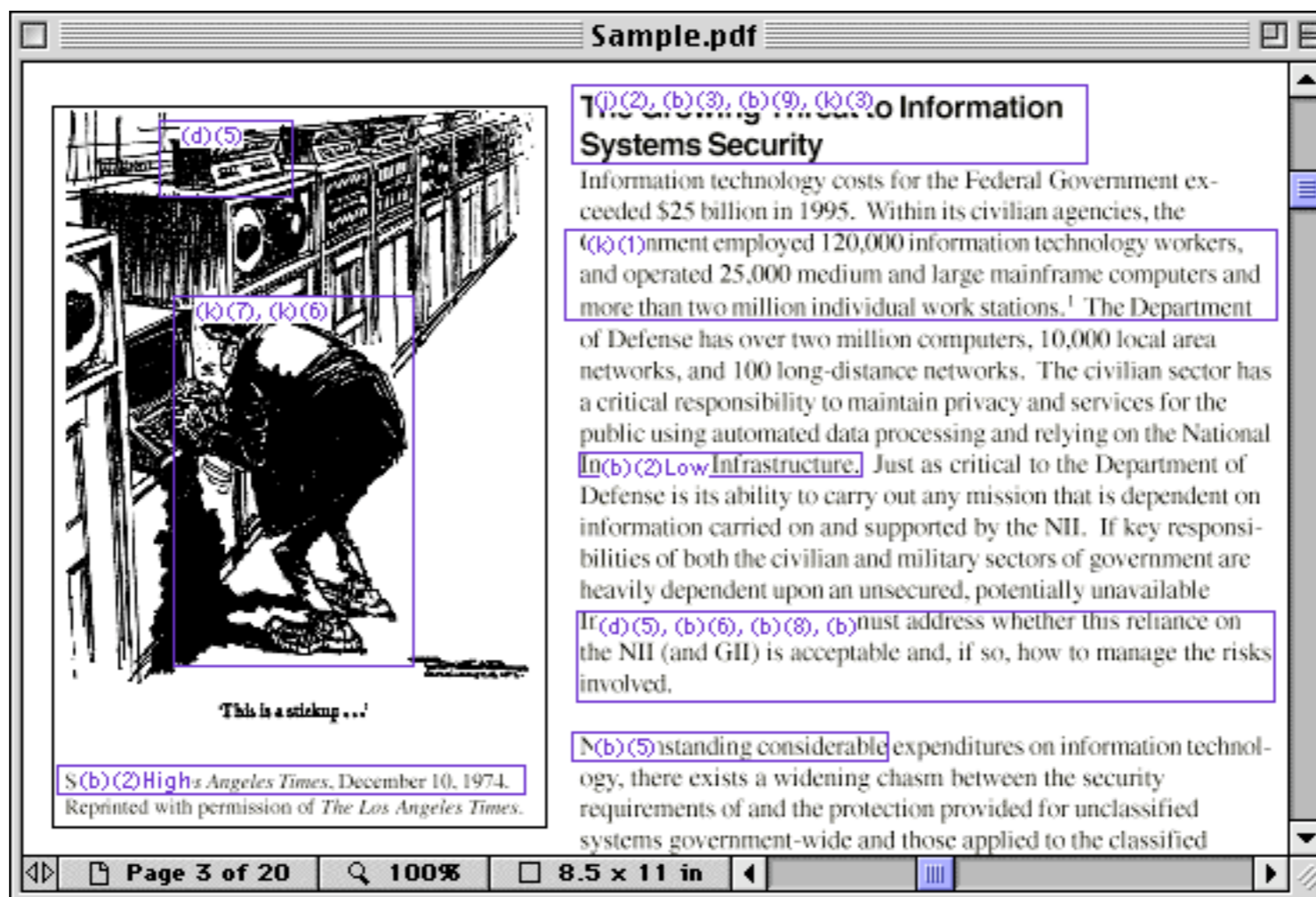
*This is a sticking ...!*

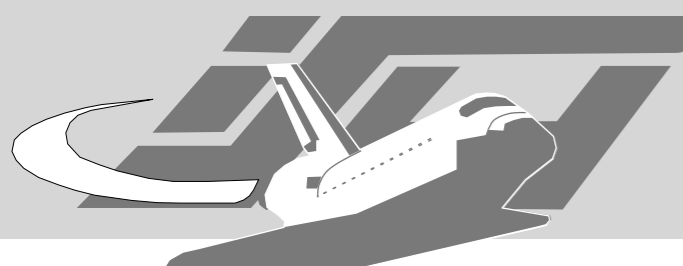
Source: *The Los Angeles Times*, December 10, 1974.  
Reprinted with permission of *The Los Angeles Times*.

Page 3 of 20 100% 8.5 x 11 in

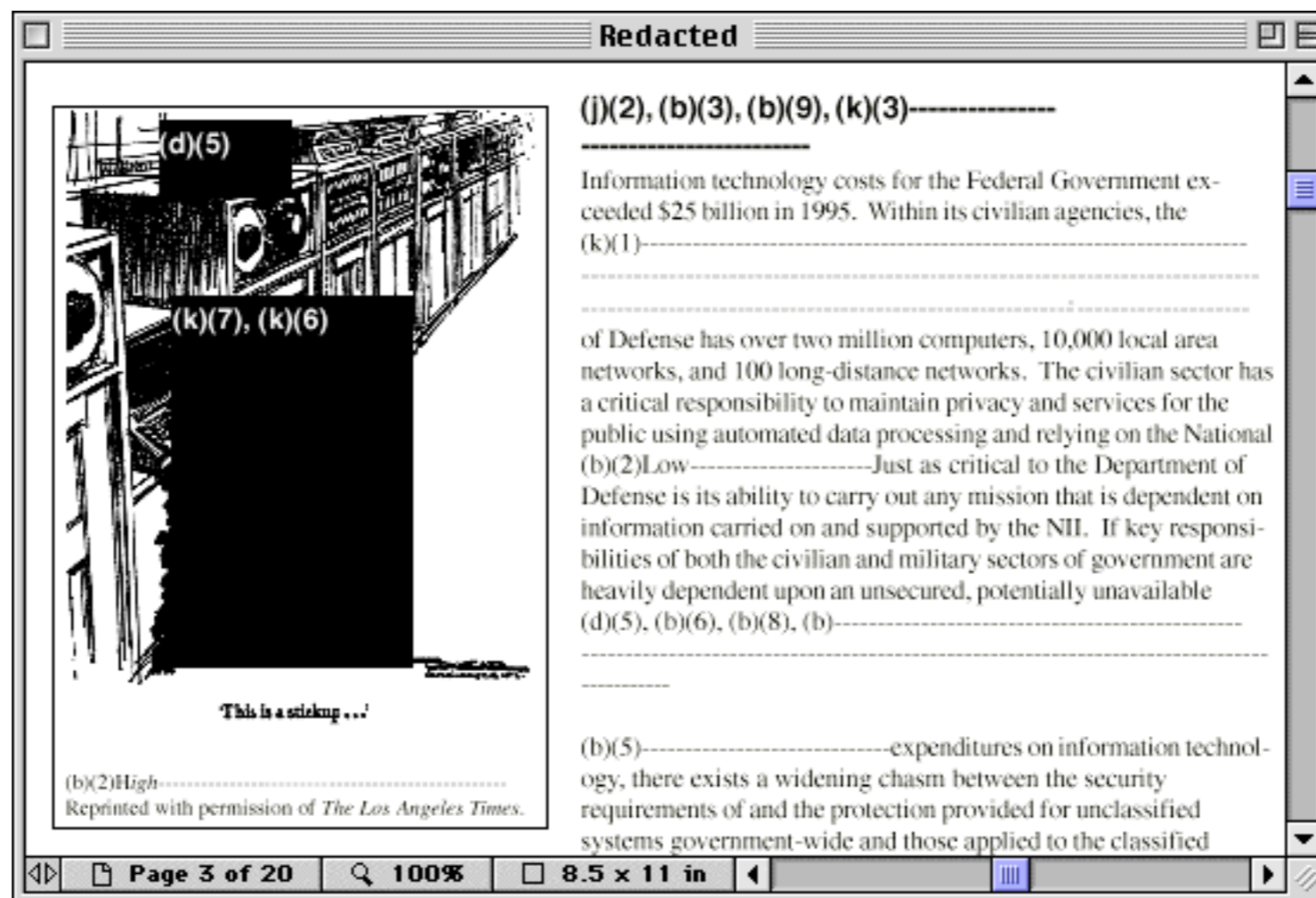


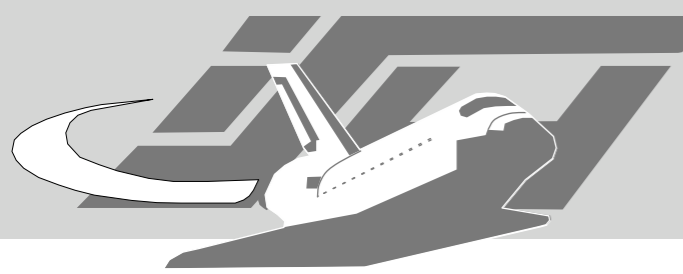
# PDF Scrubbing





# PDF Scrubbing





# Removing Redactions

- Methods
  - Very dependant on the amount of Adobe software you have at hand.
    - Copy black/white text on same ground
    - Copy text under black bars
    - Copy graphics under black bars
    - Remove overlaying graphics
  - Write your own tool

copy underlying text

Schlissel Testimony - Redacted.pdf (49 Pages)

Drawer Back/Forward Page (of 49) Page Up Page Down Zoom In Zoom Out Tool Mode

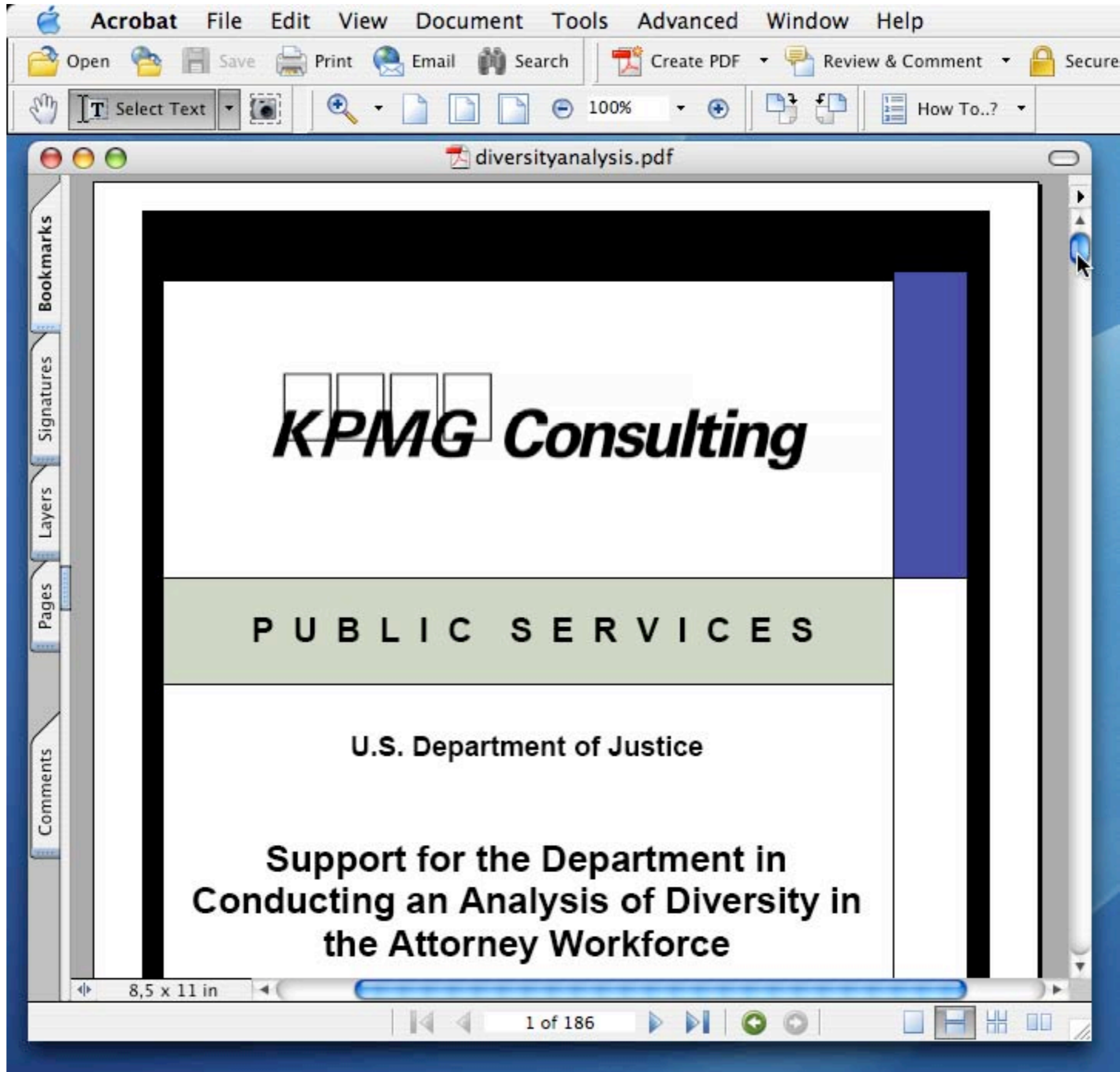
Docket No. E-01345A-03-0437 Direct Testimony of David A. Schlissel

1 12. Numerous APS and PWEC planning studies from the years 1998-2002  
2 indicated that the PWEC units were being built to facilitate power sales to  
3 areas outside Arizona, not primarily to serve APS load.

4 13. [REDACTED]  
5 [REDACTED]  
6 [REDACTED]  
7 [REDACTED]  
8 [REDACTED]  
9 [REDACTED]  
10 [REDACTED]  
11 [REDACTED]  
12 [REDACTED]

13 14. The PWEC units were built in locations where they could serve APS loads  
14 and supply power to markets outside Arizona.

black text on black ground

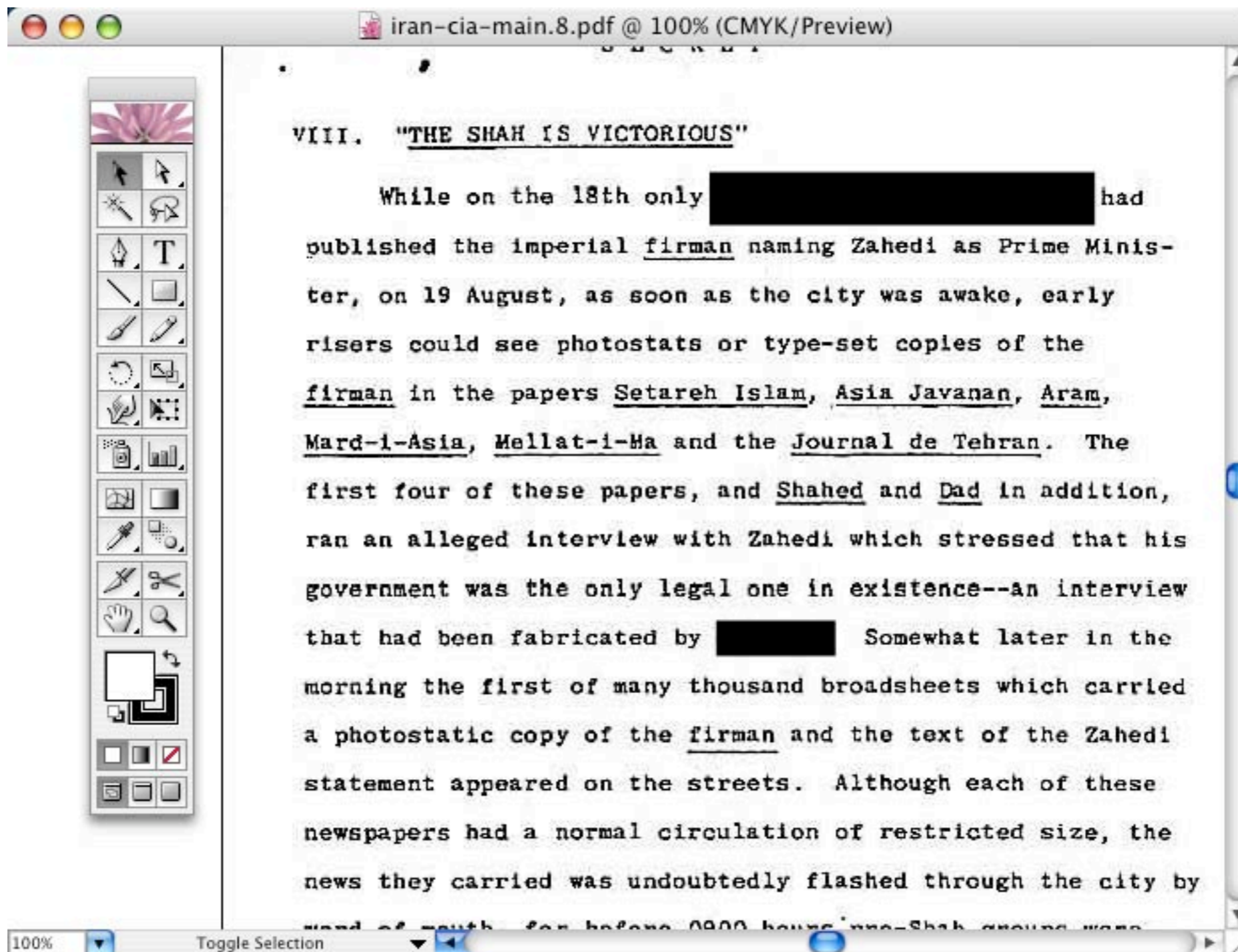


copy underlying graphics

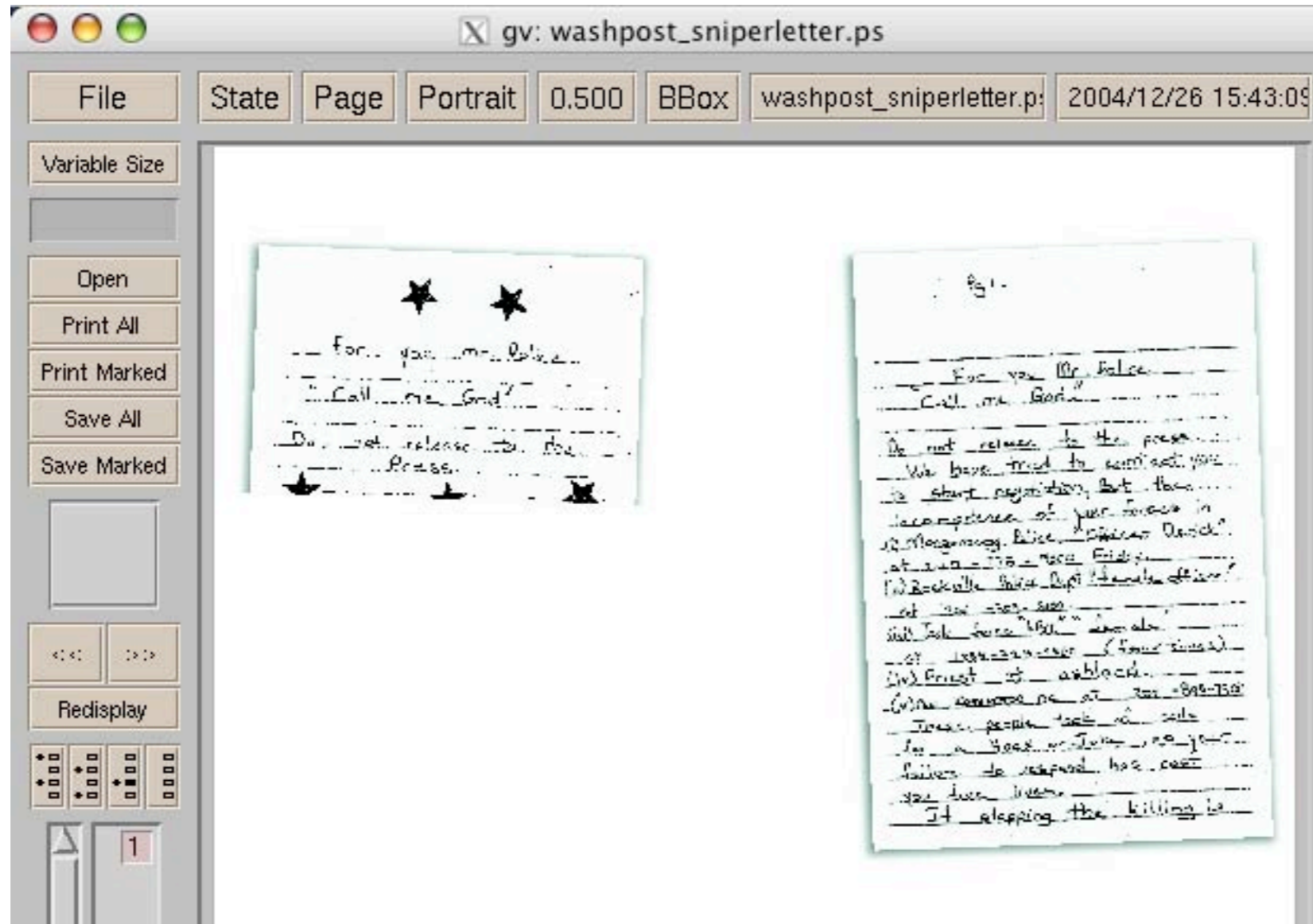
The image shows a screenshot of the Adobe Acrobat application interface. The top menu bar includes 'Acrobat', 'File', 'Edit', 'View', 'Document', 'Tools', 'Advanced', 'Window', and 'Help'. Below the menu bar is a toolbar with icons for 'Open', 'Save', 'Print', 'Email', 'Search', 'Create PDF', 'Review & Comment', and 'Secure'. The main window displays a PDF document titled 'washpost\_sniperletter.pdf'. The document content is handwritten text on lined paper, divided into two pages. The left page is labeled 'Pg 2.' and contains the following text: 'more important than catching us... now, then you will accept our demand which are non-negotiable. (i) You will place ten million dollar in Bank of America account no. [redacted] Pin no. [redacted] Activation date [redacted] Exp. date [redacted] Name: [redacted] member since [redacted] Platinum Visa Account. We will have unlimited withdrawal at any atm worldwide. You will activate the bank [redacted] and Pin [redacted]'. A text box on the left side of the page contains the following text: 'Sniper instructs authorities to transfer \$10 million into a Visa credit card account. The account belongs to a woman who reported the card stolen in California. The card was later used in Tacoma, Wash.'. The right page is labeled 'Pg. 3.' and contains the following text: '(ORLANDO, VA) Anderson's Buffet will be an Sunday. You have until Monday morning transaction. Try to catch at least you will body bags. (But) (ii) If trying to catch more important you body bags. If we give that is what + "Word is Bond!" P.S. your children...'. The bottom of the window shows a status bar with '11 x 17 in', '1 of 1', and navigation icons.

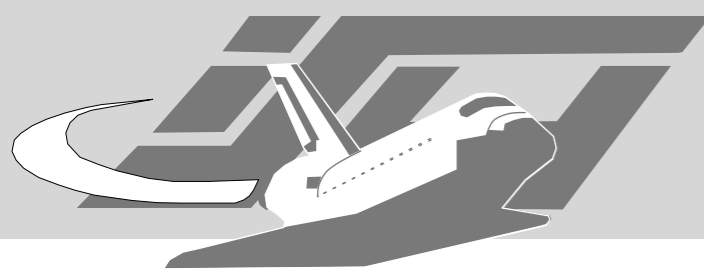


remove black bars



just wait



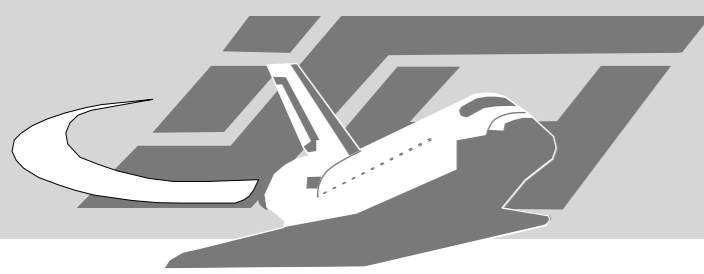


# Coding your own

- Strategy:
  - convert to Postscript
  - replace 'box' operators by NOOPs
  - (actually by popping the parameters to box into the bitbucket)
- Problem: Real world postscript uses no boxes

```
2204.84 5683.09 2.21 -63.26 1198.27 41.84 -2.21 63.26 -1198.27 -41.84 f*
1299.72 5515.11 2.21 -63.26 340.15 11.88 ^ ^ f*
1805 5374.75 2.21 -63.26 340.15 11.88 ^ ^ f*
2375.79 5245.32 2.21 -63.26 489.41 17.09 ^ ^ f*
2116.53 5081.14 2.21 -63.26 351.07 12.26 -2.21 63.26 -351.07 -12.26 f*
1833.88 4950.36 3.29 -94.24 1179.92 41.2 ^ ^ f*
2620.39 4798.75 2.21 -63.26 277.01 9.67 ^ ^ f*
5772.52 6352.31 2.21 -63.26 527.48 -12.31 ^ ^ f*
6151.04 8283.32 2.21 -63.26 705.89 19.75 ^ ^ f*
```

```
/^ {3 index neg 3 index neg}!
/f* {P eofill}!
/! {bind def} bind def
/P {N 0 gt {N -2 roll moveto p} if}!
/p {N 2 idiv {N -2 roll rlineto} repeat}!
...
```



# Works!

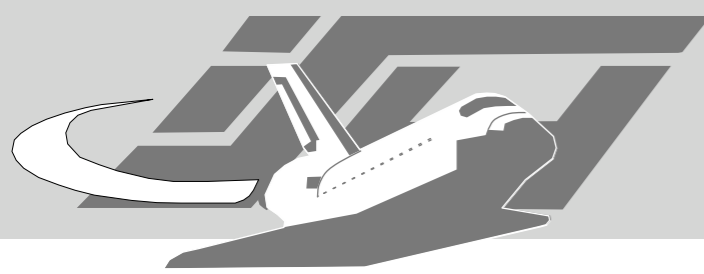
account no. [REDACTED]  
[REDACTED]  
pin no. [REDACTED]  
activation date [REDACTED]

account no. [REDACTED]  
[REDACTED]  
pin no. [REDACTED]  
activation date [REDACTED]

account no. 402  
-9173  
pin no. 4545  
activation date 01

```
% pdf2ps washpost_sniperletter.pdf \  
washpost_sniperletter.ps
```

```
% perl -npe 's/ f*$//;' \  
< washpost_sniperletter.ps \  
> washpost_sniperletter-\  
unredacted.ps
```



# Miserable Failure

n was received from the Tehran  
neral [REDACTED] had  
ary attache and had requested

n was received from the Tehran  
neral [REDACTED] had  
ary attache and had requested

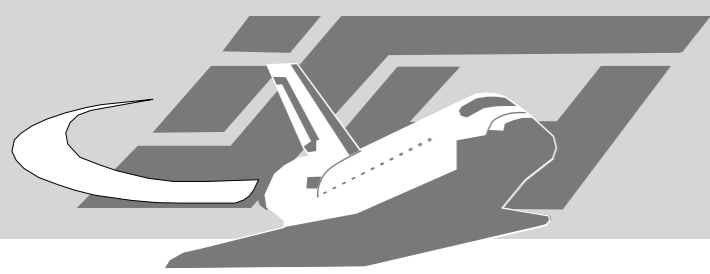
n was received from the Tehran  
neral [REDACTED] had  
ary attache and had requested

```
% pdf2ps 01.pdf 01.ps
```

```
% perl -npe \  
's/^\d+ \d+ \d{3,10} \d+ rf$//' \  
< 01.ps > 01-unredacted.ps
```

So go for simple  
formats?

Simple things are easy to understand, aren't they?



# Plain Text Formates bite

- Mail/News headers
- Signatures
- Configuration files
- HTML
- META, Comments

```

```



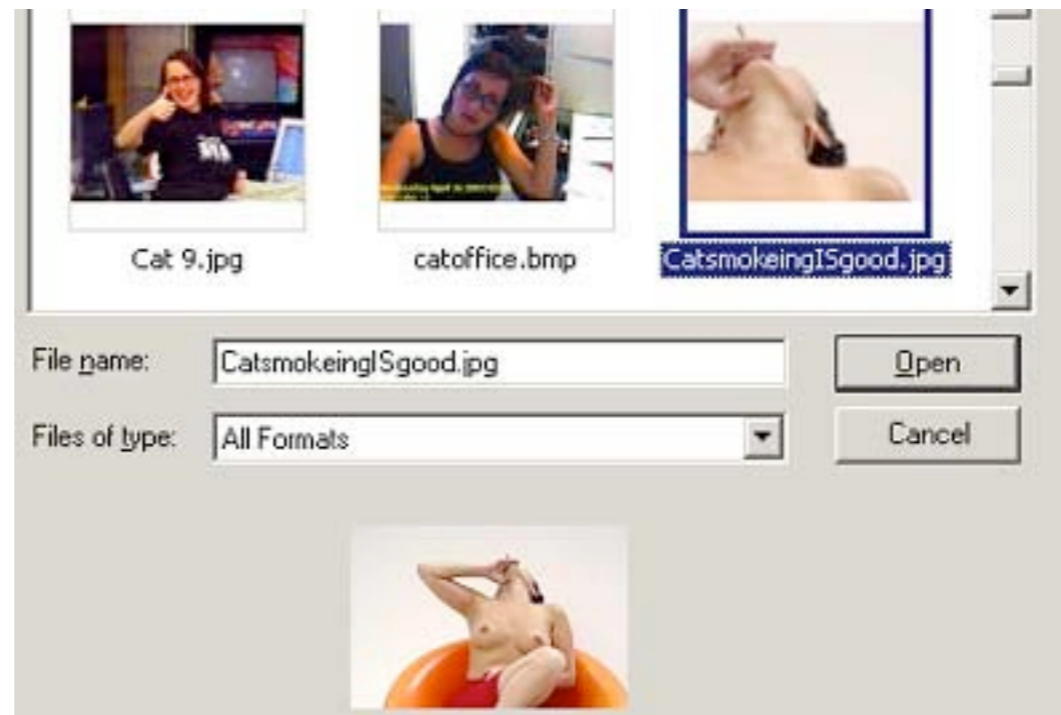
Frohes Weihnachtsfest

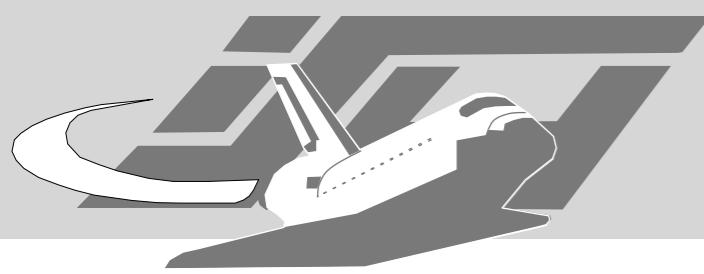
**From:** Prof. Dr. Thomas Hoeren <Hoeren@uni-muenster.de>  
**Subject:** Frohes Weihnachtsfest  
**Date:** 23. Dezember 2004 16:37:07 MEZ  
**To:** quicklinks-owner@egroups.com , David Rosenthal <rosenthal@insider.ch> ,  
peters-herrdum@phsanwaelte.de , www.klages@klages-berlin.de ,  
Prof. Dr. Thomas Hoeren <hoeren@uni-muenster.de> ,  
Snke Schrder <j8scso@nds.rz.uni-jena.de> , Andreas.Franke@ernst-young.de ,  
L-Soft list server at GMD (1.8d) <LISTSERV@LISTSERV.GMD.DE> ,  
Mail Delivery System <MAILER-DAEMON@uni-muenster.de> , wipr-l@bna.com ,  
fidele ndeshyo <fidele.ndeshyo@fundp.ac.be> , ruse@uni-muenster.de ,  
Fischer Dieter <Dieter.Fischer@icn.siemens.de> , dekan03@uni-muenster.de ,  
Arthur Waldenberger <a.wald@vdz.de> , johannes.paul <fortunamedien@t-online.de> ,  
Mller, Ulf <UM@Piepenbrock-Schuster.de> , owner-urhge-2000@urheberrecht.org ,  
Beate Lambrecht <Beate.Lambrecht@stud.uni-goettingen.de> ,  
Christian Boecker <c.boecker@kmk.org> ,  
Prof. Dr. Klaus Peter Berger, LL.M. <kpberger@netcologne.de> ,  
NET-LAWYERS@PEACH.EASE.LSOFT.COM ,  
Cornelia Holsten <conni.holsten@gmx.de> , zhoulin@ht.rol.cn.net ,  
Cundiff, Fred <fcundiff@netsol.com> , KaiKruger@t-online.de ,  
manfred.wittkowski@gerling.de , Claudia Thomas <cthomas@uni-muenster.de> ,  
Fachschaft Jura <fs-jura@uni-bonn.de> ,  
Regina Dalluege <dalluege@berlin.wvk-lawyers.de> ,  
Alexander, Ines <Ines.Alexander@euroforum.com> , jhuebner@uni-muenster.de ,  
Feldmann, Rolf-Dirk (VK) <Rolf-Dirk.Feldmann@victoria.de> ,  
dagmar.woester@uni-koeln.de , Martin Schermaier <schermm@uni-muenster.de> ,  
<Birgit.Kaufmann@ercgroup.com> <Birgit.Kaufmann@ercgroup.com> ,  
Welters, Andre (INTL) <Andre.Welters@ace-ina.com> ,  
Sonja Hebenstreit <shebenstreit@hotmail.com> ,  
Jochen Stauder <jochenstauder@gmx.de> , Stefanie.Pusch@arcor.net ,  
Domain.Disputes <Domain.Disputes@wipo.int> , Nina Walter <ninewalter@yahoo.de> ,  
Yves Pouillet <yves.pouillet@fundp.ac.be> ,  
Claudia Priemer <claudia.priemer@beck.de> , aid <aid@bn.kowi.de> ,  
vdv127@uni-muenster.de , Severine Dusollier <severine.dusollier@fundp.ac.be> ,  
Ulrich <ubadde@uni-muenster.de> , jurist-l@lawlibdns.wuacc.edu ,  
Simone Westpfahl <westpfahl@german-law.com> , ludickk@uni-muenster.de ,  
<networks@tm.net.my> <networks@tm.net.my> , Frithjof.Maennel@cec.eu.int ,  
iumkf@mail.yahoo.co.jp , M. Markmann <m.markmann@gmx.de> ,  
norbert.meder@uni-bielefeld.de , VRothkirch@aol.com ,  
Dr. K.-E. Maass <maass@dfn.de> , info@eifonline.org ,



```
% curl -q http://www.affordablehairtransplants.com/robots.txt
<?php
header("Content-type: text/plain");
if (strstr($_SERVER["HTTP_USER_AGENT"],"lurp")) print "User-
Agent: Slurp\nDisallow: /";
?>
```

# Girls named .jpeg

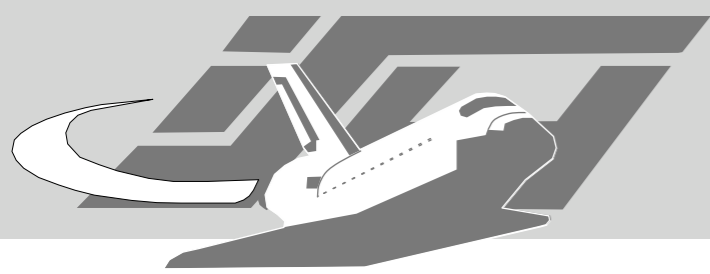




# The techtv moderator incident

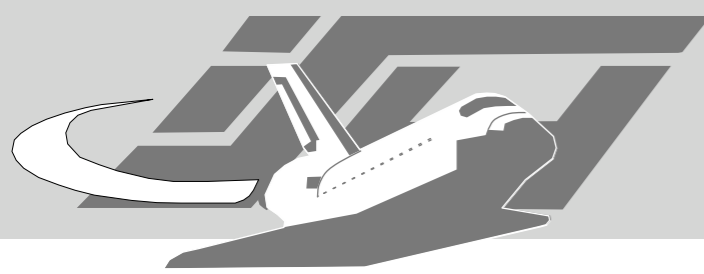
- Moderator adds picture to her weblog
- People download it, archive it, view it with image browser
- Picture was cropped, thumbnail remains uncropped
- Male teenage geeks get totally mad





# How did it happen?

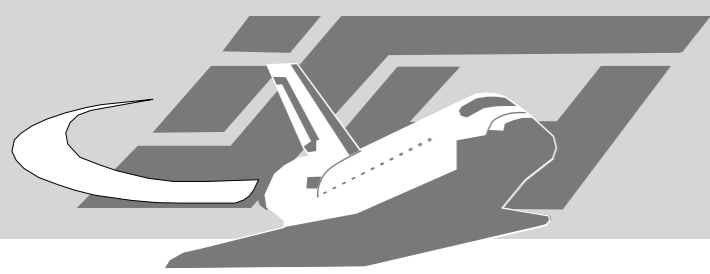
- Software glitch?
- Widespread?
- Desired behavior?
  - ... actually it is.



# EXIF

- JPEG works surprisingly well considering that there is such a wide variety of JPEG standards and implementations.
- EXIF is the standard way to store headers
  - Applications usually are leaving unknown EXIF headers (thumbnails?) untouched.
- So we expect the problem to be quite widespread.

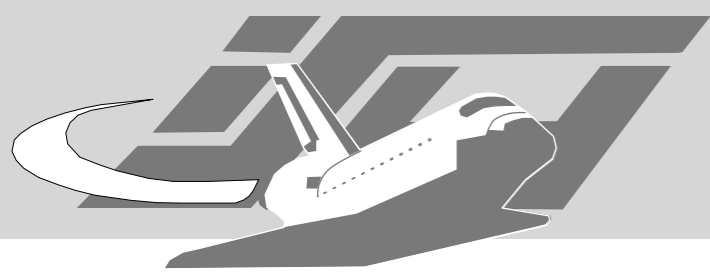
```
JPEG image data, EXIF standard 0.73, 10752 x 2048
JPEG image data, EXIF standard 0.77, "AppleMark", 42 x 0
JPEG image data, EXIF standard 0.77, 42 x 0
JPEG image data, JFIF standard 1.01, aspect ratio, 1 x 1
JPEG image data, JFIF standard 1.01, resolution (DPI), 180 x 180
JPEG image data, JFIF standard 1.02, resolution (DPI), 150 x 150
```



# Experimental Setup

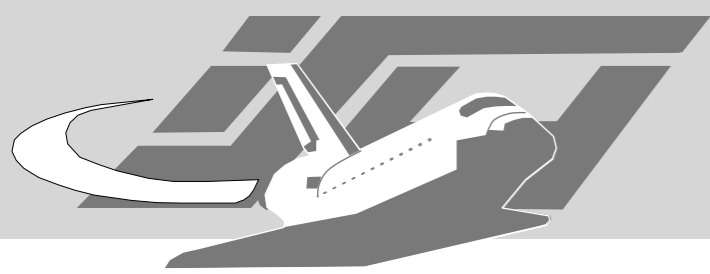
- Get as many images as possible from the Internet
- Compare thumbnails to images





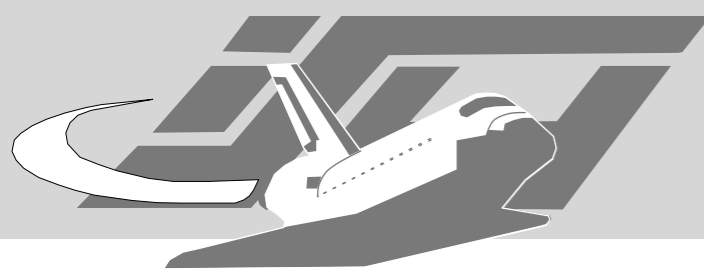
# Spidering the Web

- We use a patched Version of Niels' Provos' crawl-0.4. Modifications:
- Do not overload filesystem with 100.000 entries in a directory
- Keep HTTP headers for fingerprinting
- See [http://core.23.nu/code/misc/crawl-\\*.patch](http://core.23.nu/code/misc/crawl-*.patch)



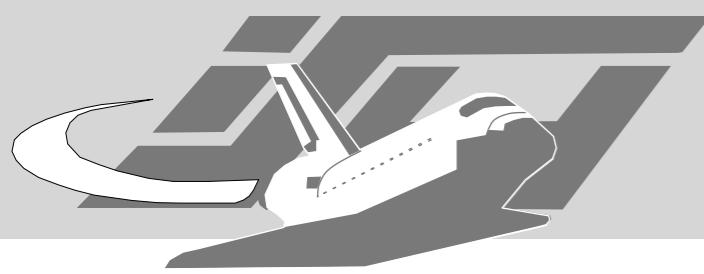
# Comparing Images

- We need a way to find among a million pictures the ones with a substantial difference between thumbnail and image.
- Steven J. Murdoch found a Way for doing so
  - compare image proportion
  - compare image contents
  - analysis



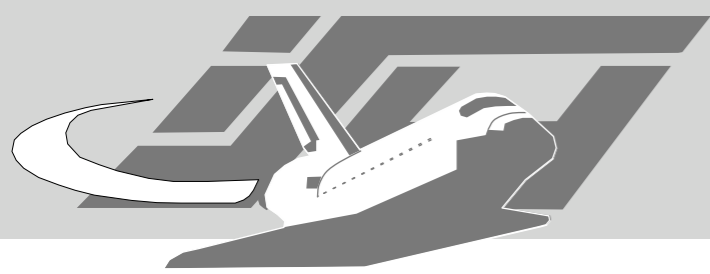
# Image Proportion

- Scale both dimensions of the full size image equally, so that the larger dimension of the full size image is equal to the larger dimension of the thumbnail
- Compare the smaller dimension of the scaled full size image to the smaller dimension of the thumbnail
- The difference should be 0 but, if the generator used a different rounding technique, it could be +/- 1
- Repeat for the full size image rotated 90 degrees, and choose the minimum



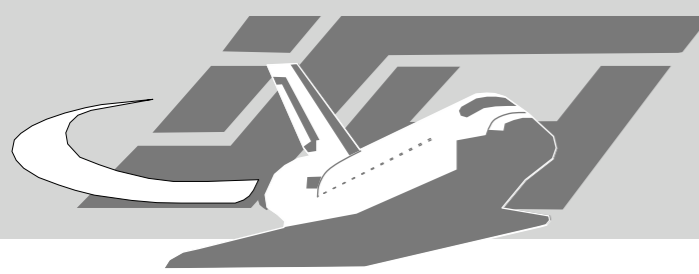
# Image Content

- Scale the full size image to the size of the thumbnail
- Use "nearest" interpolation method for speed
- Subtract one image from the other, and calculate to root-mean-squared
- If the ratio was closer with the swapped dimensions then do this for 90 degree rotation (clockwise and anti-clockwise) and choose the minimum



# Analysis

- Use GNU R to find a suitable criteria on ratio and RMS difference
- Pick a random sample, check manually and compare histograms
- Output full size image and scaled thumbnail side-by-side, for comparison



# Analysis

- Filter out false positives manually, mainly due to:
  - Images with sharp edges cause phase difference in scaled image because of “nearest” interpolation, and so increases RMS difference
  - Images where thumbnail has been padded to a fixed ratio, different from that of the full size image

```
% sh process.sh
```

```
372105 files in 7073s processed (0.019s per image), 69603 thumbnails found (18.7%)
```

```
processing in './results.data', writing output to './flagged.data'
```

```
372105 files processed, 6441 found interesting(1.7%) out of 69603 with thumbnails (9.3%)
```

**ca. 19% of the images have thumbnails**

**ca. 9% of the thumbnails are “interesting”**

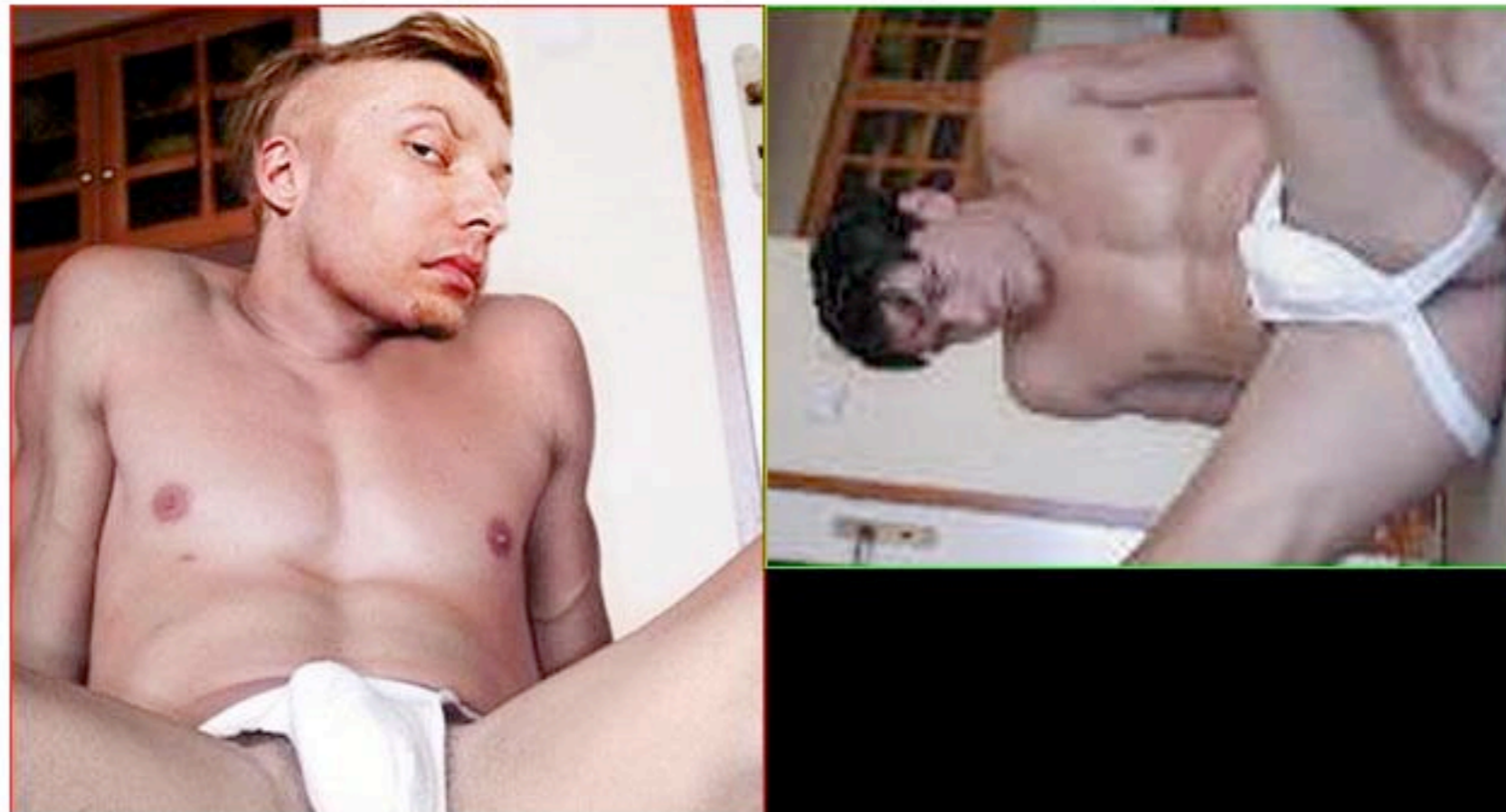
**how screen ca. thousands of images?**

# Demo: Differences between JPEG Images and their EXIF Thumbnails

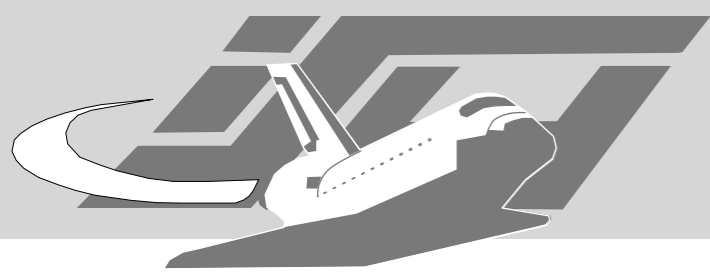
## The Image

Comparing the thumbnail to the image ...

No visible Differences - Boring - Worth a look - Interesting - A must see! - Completely different Images







# What did we find?

- Completely unrelated images
- Cropping
- People removing their friends
- Stolen Images
- Privacy violations

# Removing Friends



8.00, 92.37

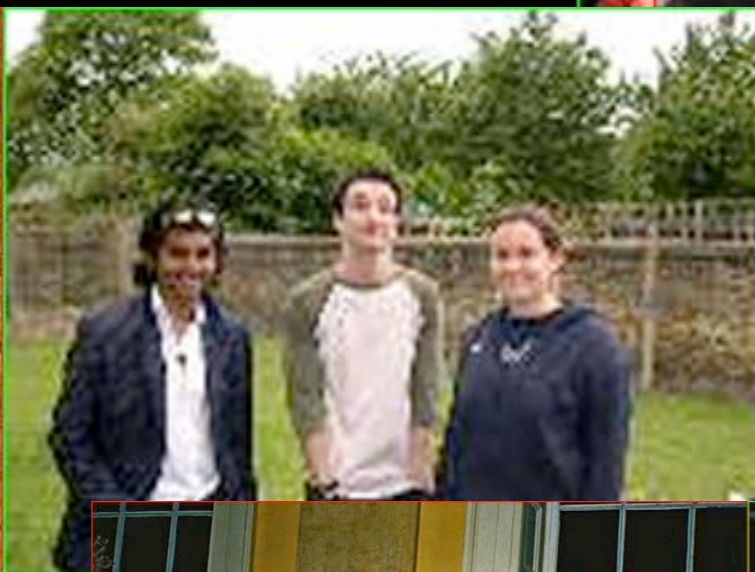
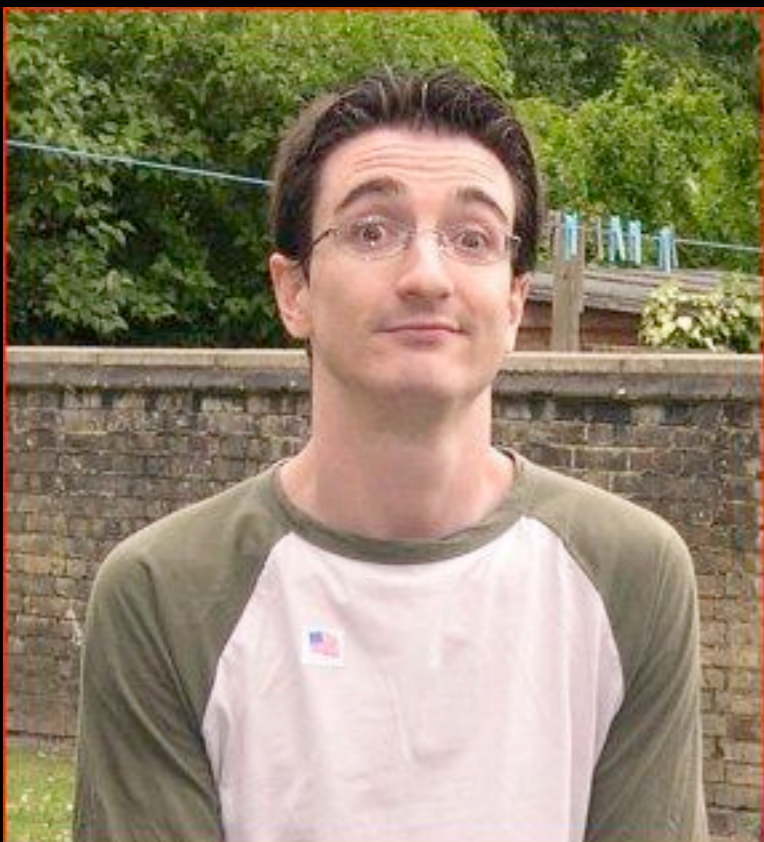
0.64, 74.54

46.61, 73.58



4.81, 91.24

78.62



25.27, 102.21



24.00, 64.13

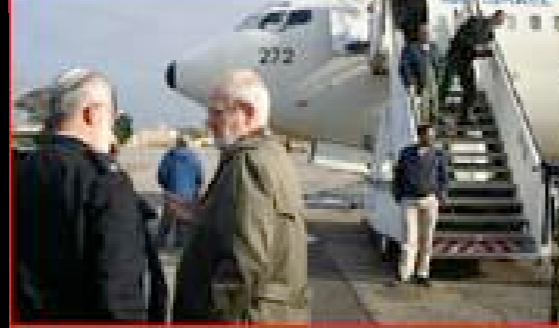


12.80, 73.68

# Stolen Images



00, 100.09



41.09, 108.74



Small text block, possibly a caption or description, located below the image of the two men in military uniforms.



D112-348 MATT DAMON as Jason Bourne and FRANKA POTENTE as Marie in the espionage thriller The Bourne Supremacy  
Credit: Jason Boland  
©2004 Universal Studios. All Rights Reserved.



10.80, 128.92



0.97, 96.05



# YOUR SOLUTIONS



15.45, 122.72



4.95, 41.61



Small text at the bottom of the grayscale image, possibly a caption or credit.

475.86, 58.76



CHASE MARYSON

# CREATURE

# UNKNOWN

475.86, 58.76



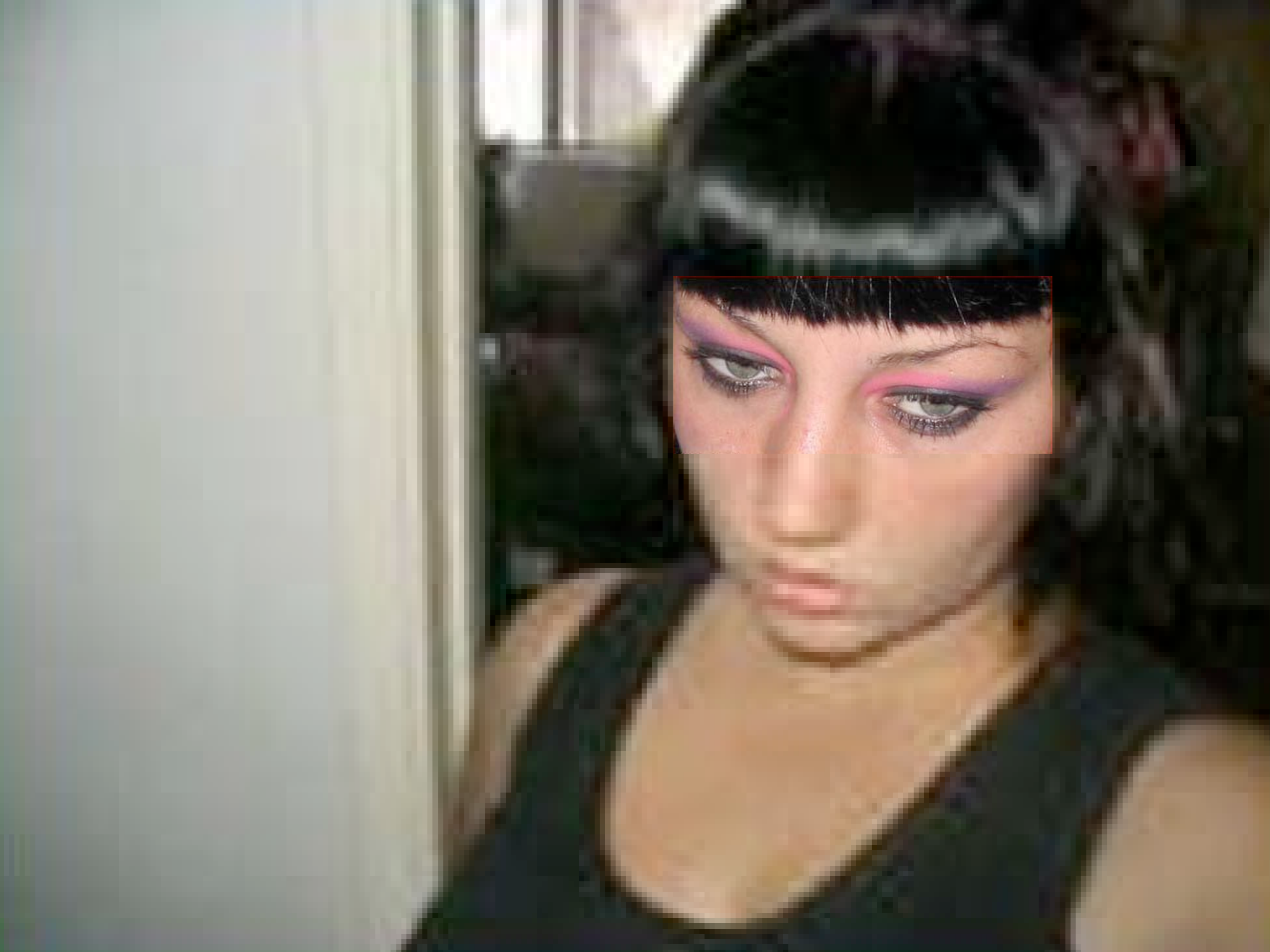
THE  
HIGGS AS  
DUTY FOR RAIN

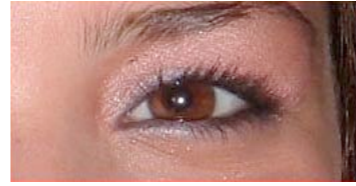


# Identity Hiding













33.60, 126.71



5.36, 99.52



10, 95.67



38.50, 97.99



28.57, 104.63



0.00, 125.86



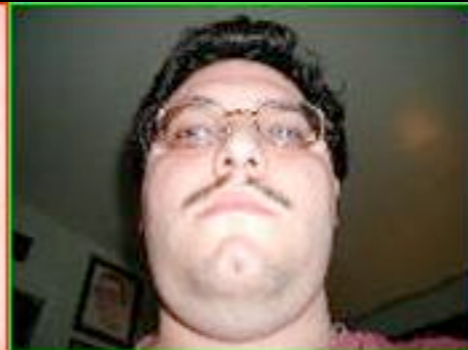
60.24, 106.99



0.23, 108.08



14.40, 144.38



40



23.45, 63.47



20.00, 193.66





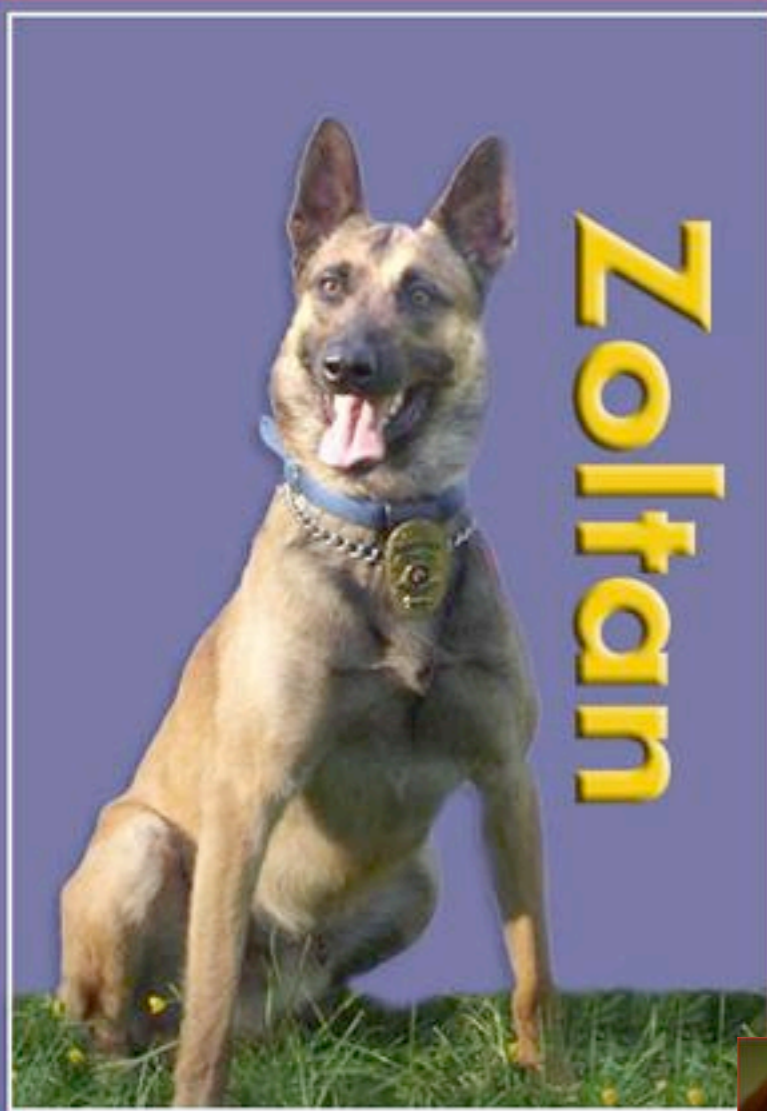






# Photoshopping

# Unrelated Images



Zoltan



40.00, 67.16



11.02, 97.06



11.02, 119.36



0.11, 103.31



5.55, 108.87



Desert Ace  
Flying low-very low-over Baja



Saving Maine  
Sonar technology could rescue oceans from collapse-is anyone listening?



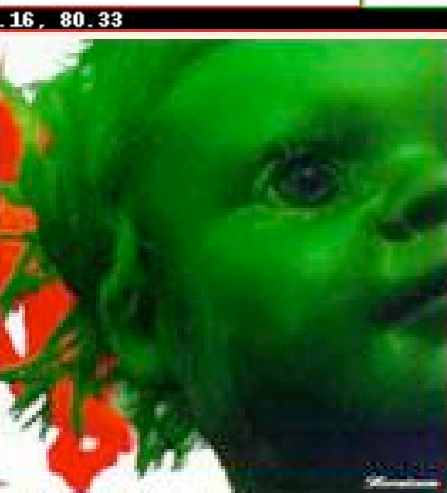
07.24, 188.97



Ich hasse Mama, ich hasse Dadi  
Mama hasst Dadi, Dadi hasst Mama  
Ach, was macht mir das (Lisa Kurper)



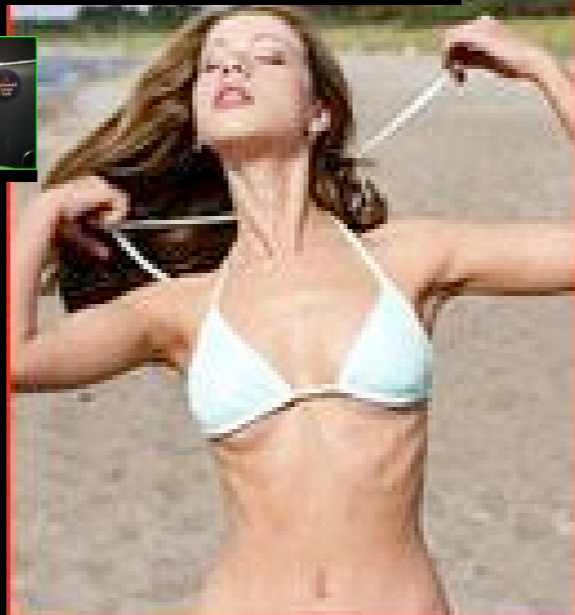
0.00, 98.68



0.00, 89.94



04.16, 205.52



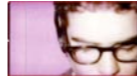
0.07, 76.92

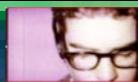


0.00, 183.89

# Cropping











34.44, 110.51

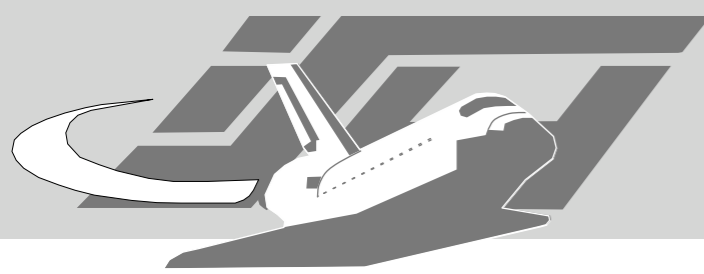








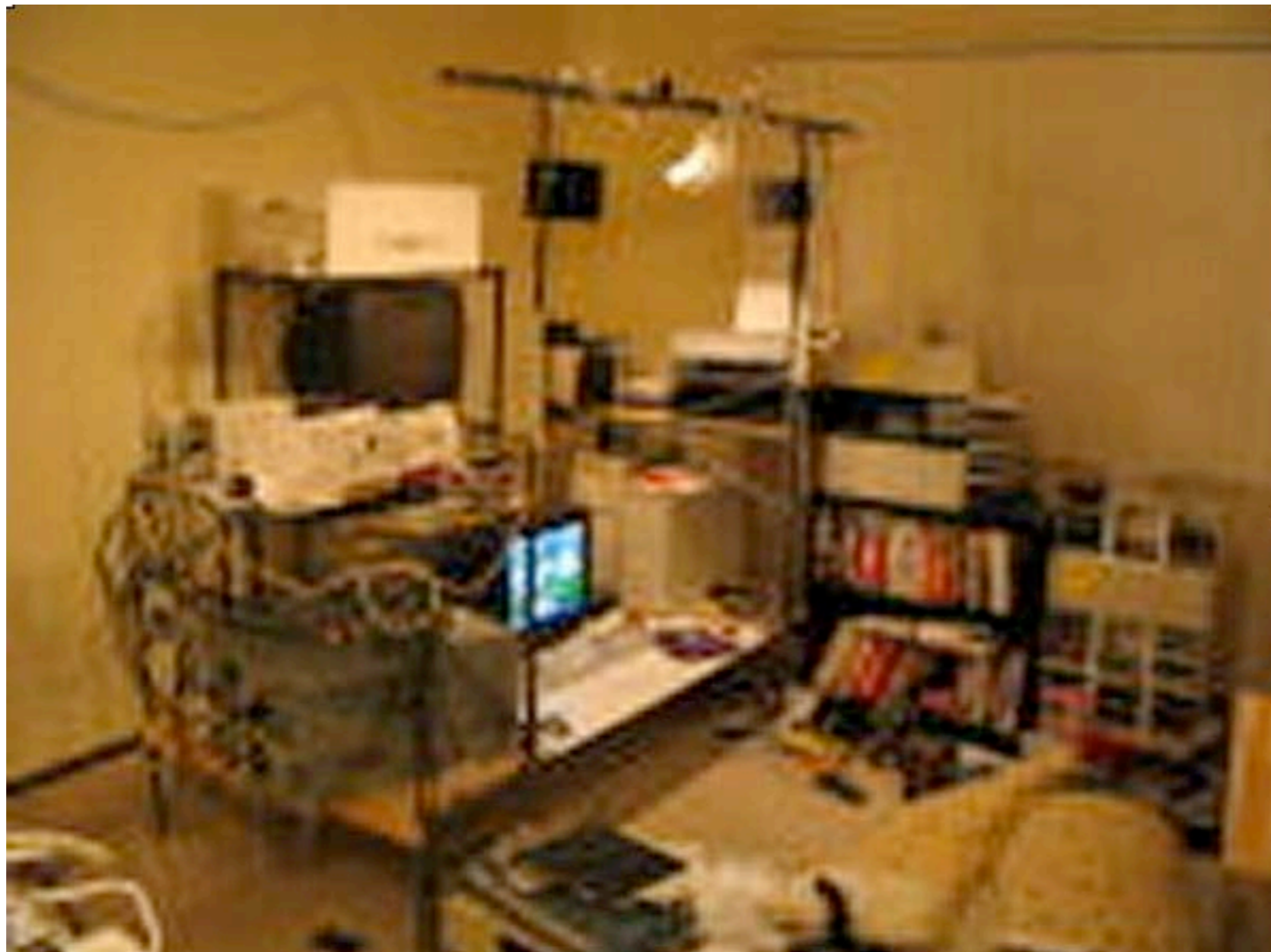
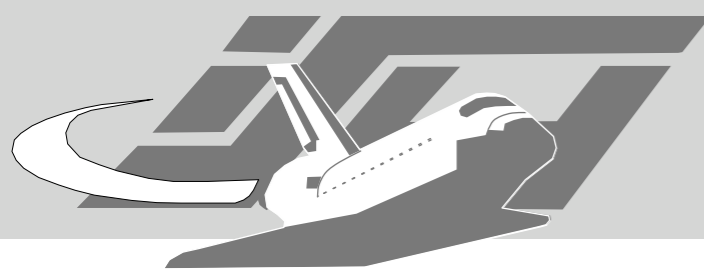
0.00, 74.97





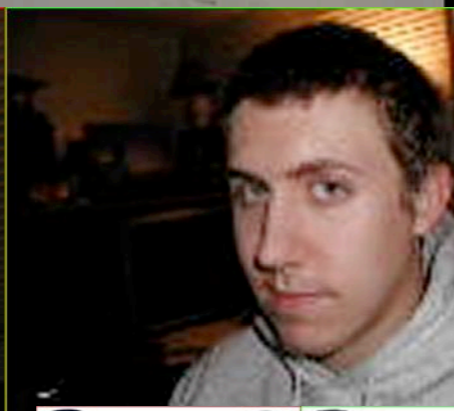
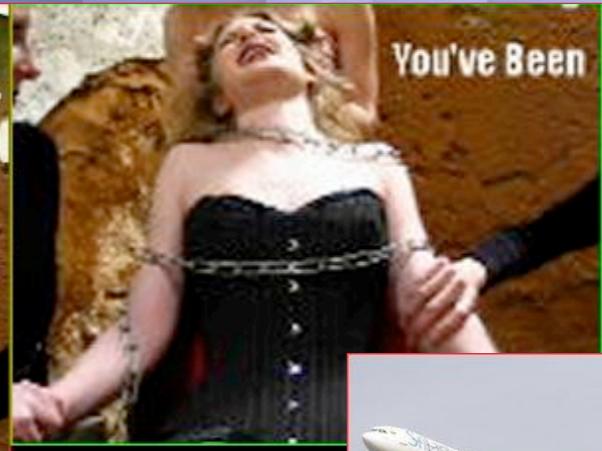




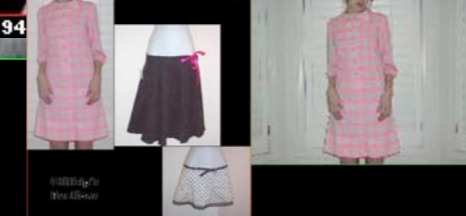


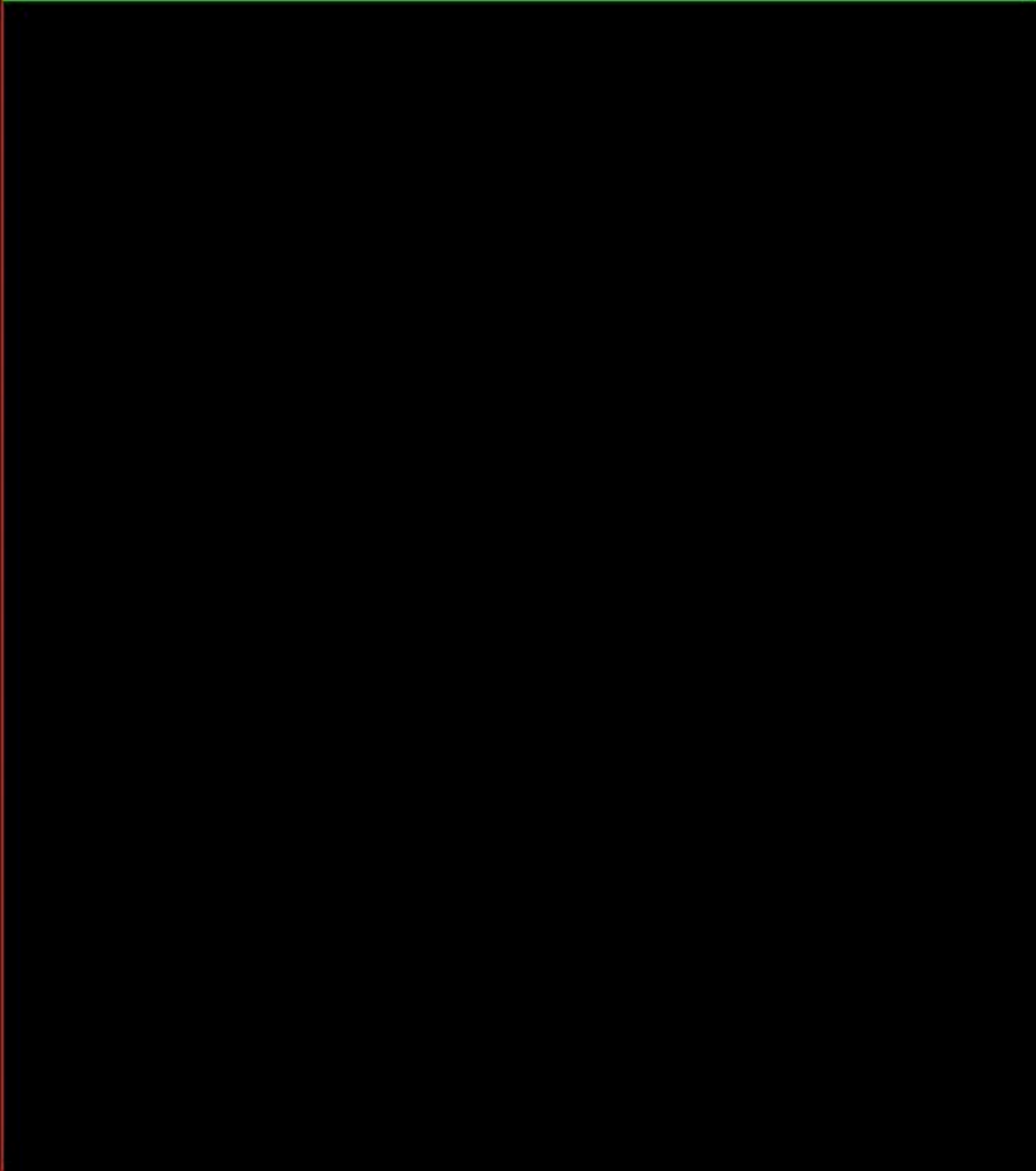
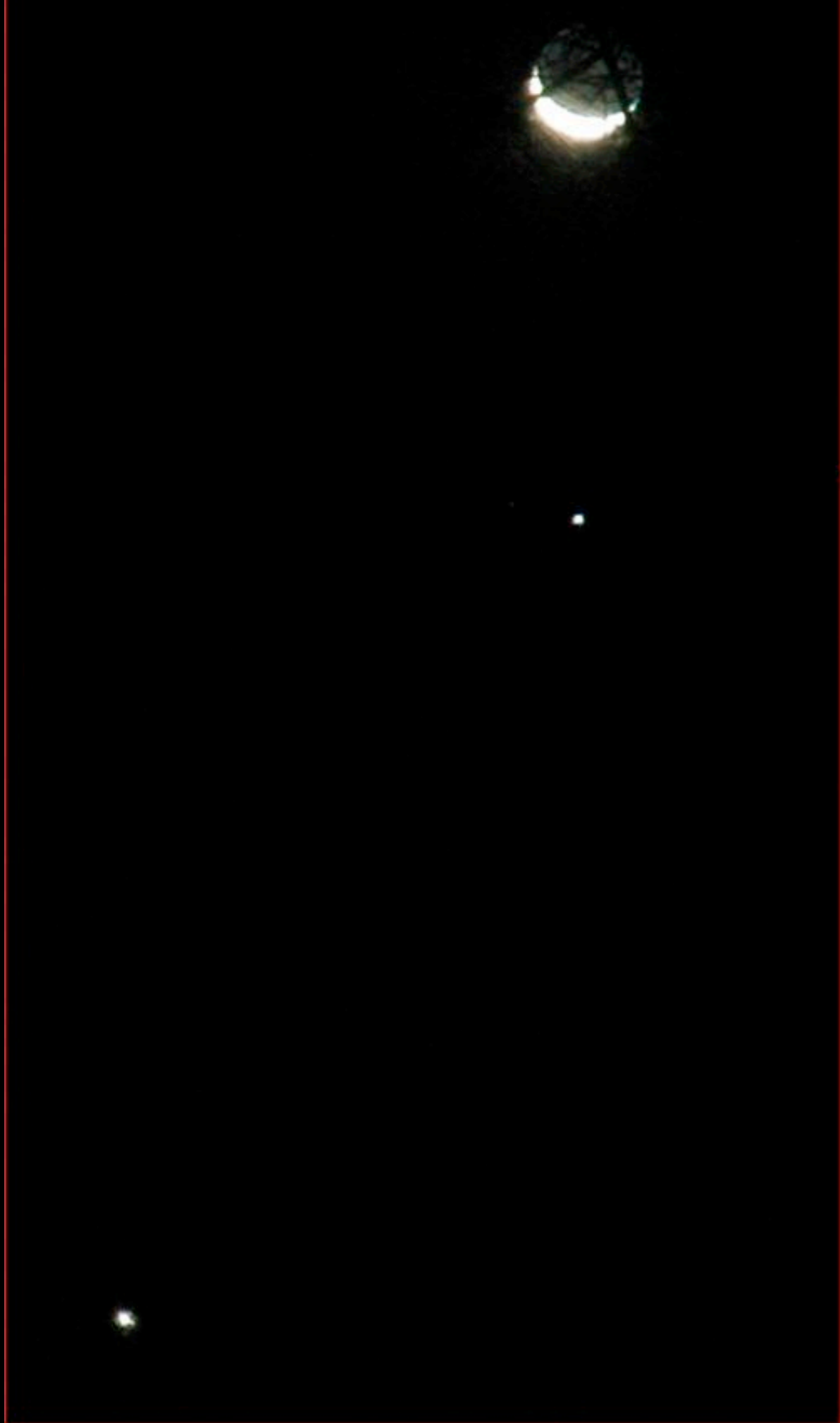
6.67, 56.53





ACCEPTED

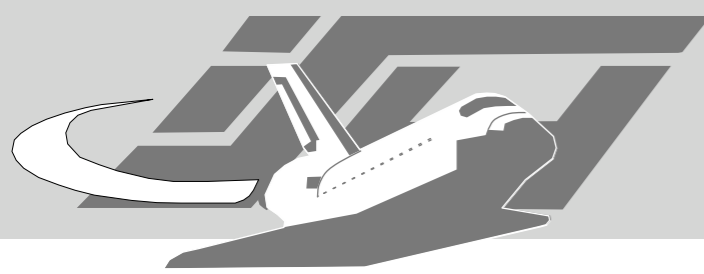






47.33, 67.70





# Reference

- Scalable Exploitation of, and Responses to Information Leakage Through Hidden Data in Published Documents Simon Byers  
byers@research.att.com 2003/04/03
- [http://www.user-agent.org/word\\_docs.pdf](http://www.user-agent.org/word_docs.pdf)
- <http://md.hudora.de/presentations/#hiddendata-21c3>
- presentations, crawl patches, exif\_thumb
- <http://sauna.5711.org/~md/thumbnails/>