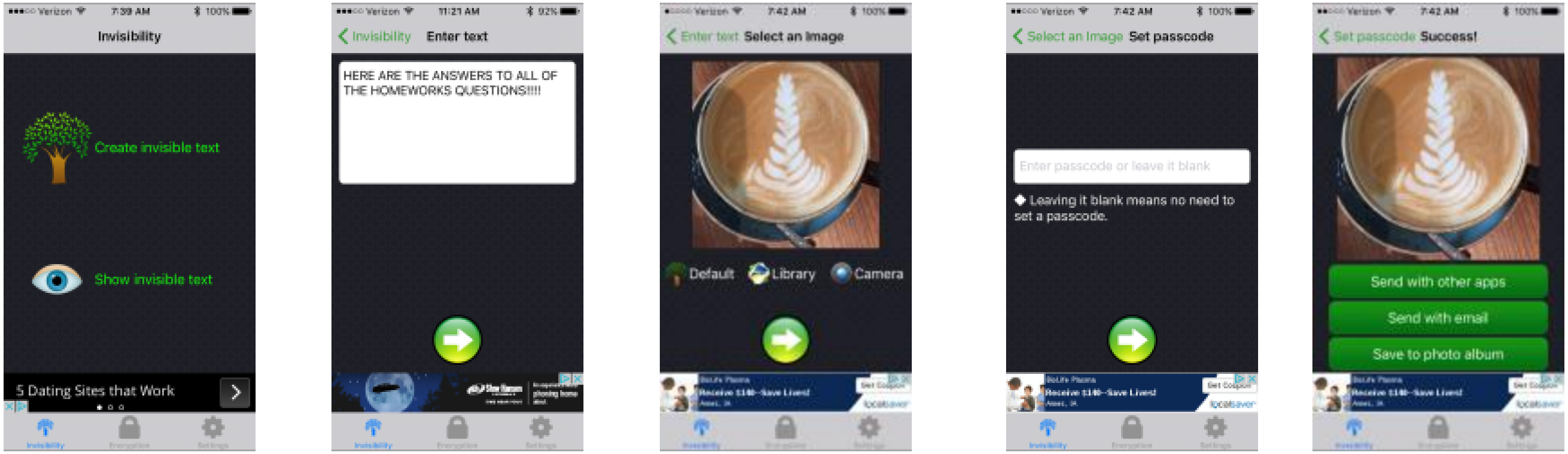
StegoDB: A Statistically-designed Image Dataset for Benchmarking Steganalysis Algorithms

Jennifer Newman*, PI
Yong Guan**, Co-PI
Min Wu#, Co-PI
Stephanie Reinders and Li Lin, PhD students*

*Mathematics Department, Iowa State University, Ames IA
**Electrical & Computer Engineering Department, Iowa State University, Ames IA
#Department of Electrical and Computer Engineering, and Institute for Advanced Computer Studies, University of Maryland, College Park

WeHide app



Invisibility

Create invisible text

Show invisible text

Invisibility | Enter text

HERE ARE THE ANSWERS TO ALL OF THE HOMEWORKS QUESTIONS!!!!

Enter text | Select an Image

Default | Library | Camera

Select an Image | Set passcode

Enter passcode or leave it blank

◆ Leaving it blank means no need to set a passcode.

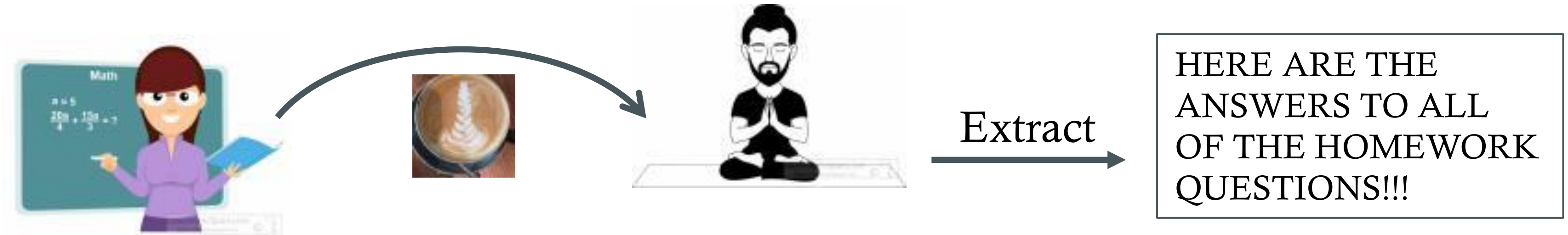
Set passcode | Success!

Send with other apps

Send with email

Save to photo album

Eve is spying

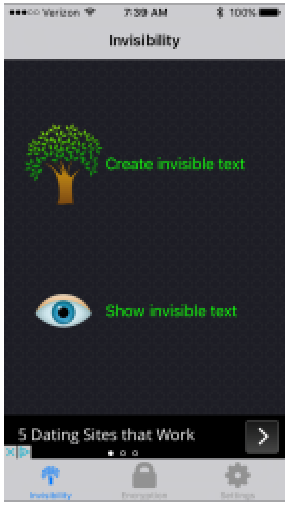


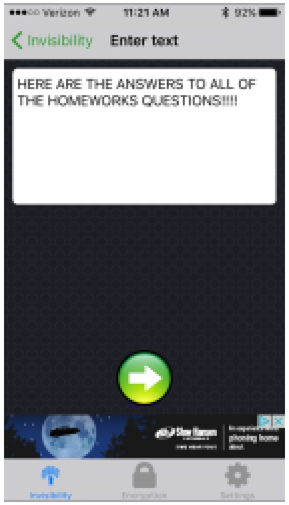
Extract

HERE ARE THE ANSWERS TO ALL OF THE HOMEWORK QUESTIONS!!!

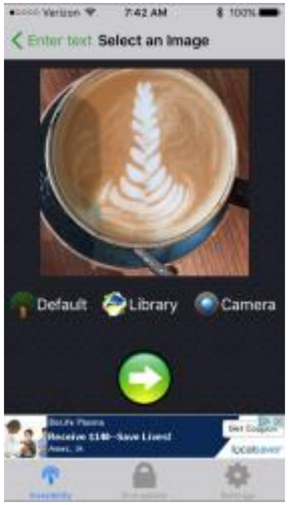
Alice sends a stego image to Bob

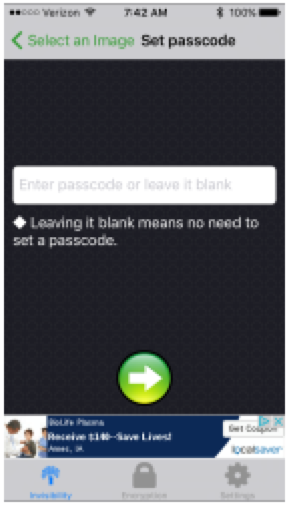
WeHide app



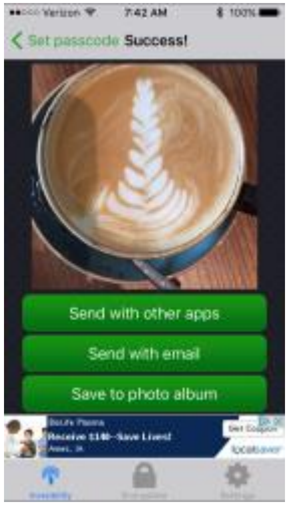


HERE ARE THE ANSWERS TO ALL OF THE HOMEWORKS QUESTIONS!!!!





Leaving it blank means no need to set a passcode.



Eve is spying

Alice sends a stego image to Bob

Question: If you are Eve, how do you determine if an image has a hidden message?

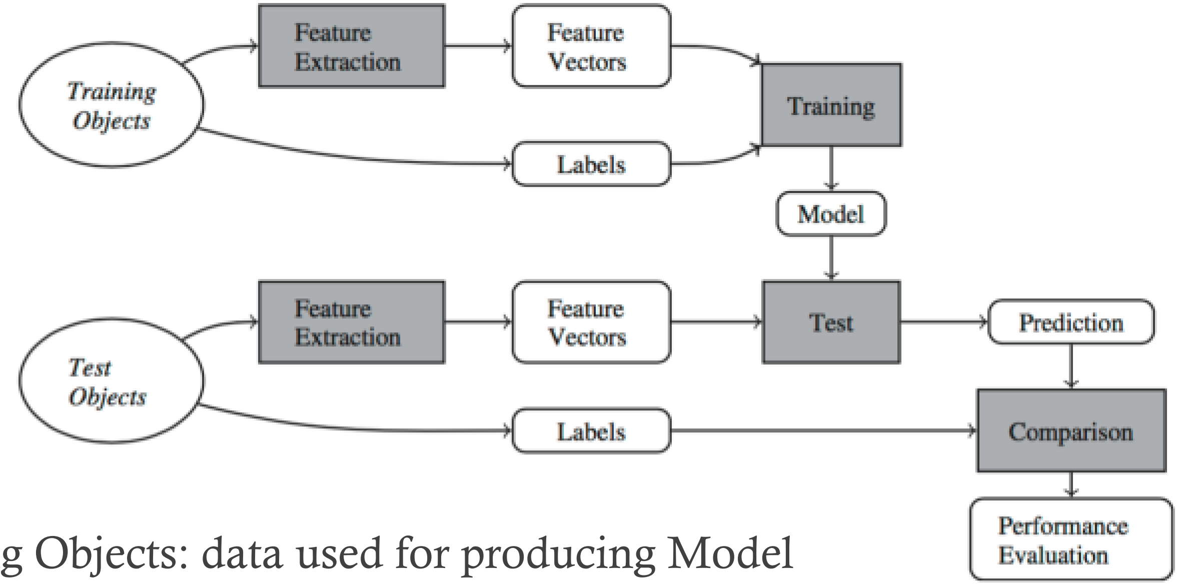
Overview

- What is steganography and steganalysis?
- Motivation for creating a database for steganalysis
- Why our work is novel
- What is in our dataset
- Project update

What is steganography and steganalysis?

- *Steganography*: hiding a message (payload) inside a digital photo by changing its pixel values in a *visually* imperceptible way
  - However, it may be statistically or otherwise detectable
  - *Cover image:* an image that has no payload
  - *Stego image:* an image that contains an embedded payload
- *Steganalysis*: classifying an image as stego or cover, typically using a statistical learning methods (SVM, linear classifiers, etc.) and a particular image dataset
  - We do not focus on extraction or decryption of any payload

Pattern Classifiers for Steganalysis

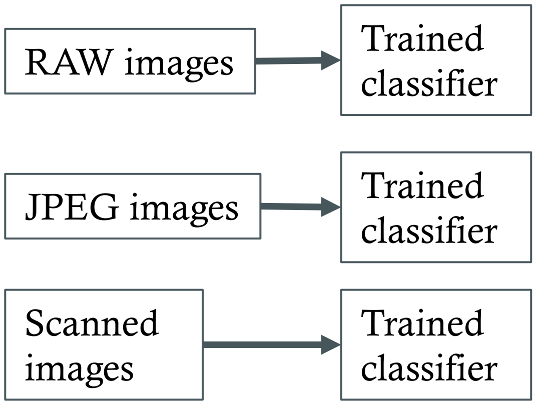


- Training Objects: data used for producing Model
- Features are a statistical representation of an image
- The set of images gives rise to a set of values that the classifier uses to distinguish between cover and stego
- The more information known about the training data, the better you can design your steganalysis experiments within a statistical framework
- **TRAINING DATASET IS VERY IMPORTANT!!!**

Figure from Schaathun, H.G.: Machine learning in image steganalysis. Wiley, pp. 26 (2012).

Why a standard dataset for steganalysis?

- Steganalysis researchers have noted that the type of images used for training and testing a classifier can have a important effect on the detection rate[2,3] . P_{MD} = probability of Missed Detection (of stego); P_{FA} = probability of False Alarm (cover identified as stego).

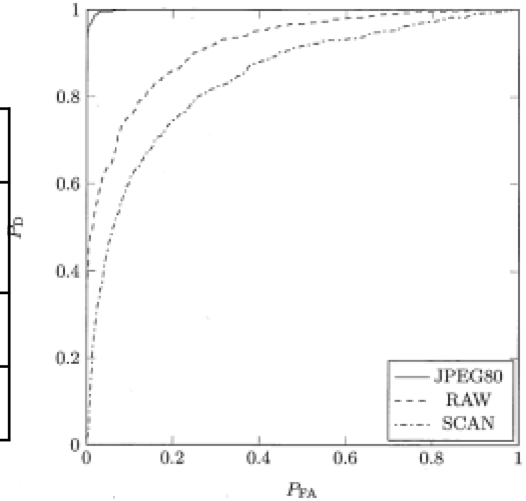

RAW images → Trained classifier

JPEG images → Trained classifier

Scanned images → Trained classifier

(ave. detection error)

Type	$(P_{MD}+P_{FA})/2$
RAW	16.6%
JPEG	1.7%
Scanned	22.8%



- There is no standard dataset of images used for steganalysis allowing different experiments to be conducted that takes into account the different sources of the image data used for training the classifiers

[2] Fridrich, J.: Stegnaography in Digital Media: Principles, Algorithms, and ApplicationsWiley, pp. 26 (2010).

[3] Sedighi, V., J. Fridrich, and R. Cogranne. "Toss that BOSSbase, Alice!." IS&T intl symposium on Electronic Imaging. 2016.

StegoDB: Guiding principles

- Our database follows these principles
  - Data are copyright-free and publicly available at no cost.
  - Authentication of the origin/pedigree of each image.
  - Collection of a minimum list of specific information on each image in the database.
  - Collect data in such a way so that experiments using StegoDB image data are repeatable.
- The novelty of our dataset is that the data is collected in this principled way, unlike other current stego datasets
- Initially we collect data from mobile phones and their apps

Project goals

- Design data collection process, what data items to be collected, and what additional parameters to add to database
- Collect data
- Use NoSQL database MongoDB to store data
  - Easy to use; convenient to maintain; extends to hold very large amounts of data; flexible to add information in the future;
  - Other databases are stored in organized folders; hard to query based on "image size" for example
- Add additional parameters to records, including parameters produced by processing image data
- Evaluate the data

What data we collect

- Previous researchers note these aspects of training data can affect detection results:
  - Image size; previous processing; image content; added noise; image format; saturation level; compression level
- These all can influence detection error rate
- We use this information to guide what image data we collect and what additional data we will store in our database, so that we will have a wider range of known parameters represented by the image data
- We will also collect as much other information as possible (meta data; ??) so that it may be possible to use these parameters in future experiments should there be other influences identified later

Information our records hold

- We collect phone information; image type; app(s) used; meta data; scene type;

- We will compute and/or identify in image data:
  - saturation levels;
  - double compression/multiple compression if JPEG;
  - any known pre-processing methods;
  - scene information (outdoor, indoor), if available

Image: cover1.jpg	Phone: phone2	Type: iPhone6sPlus	[exif]	sat=5%
Image: cover2.jpg	Phone: phone4	Type: iPhone6s	[exif]	sat=9%
Image: cover3.jpg	Phone: phone1	Type: iPhone6s	[exif]	sat=4%
Image: cover4.jpg	Phone: phone3	Type: iPhone7	[exif]	sat=5%
Image: stego1.jpg	Phone: phone2	Type: iPhone6sPlus	Cover: Cover1.jpg	App: WeHide
Image: stego2.jpg	Phone: phone2	Type: iPhone6sPlus	Cover: Cover1.jpg	App: WeHide
Image: stego3.jpg	Phone: phone4	Type: iPhone6s	Cover: Cover2.jpg	App: Cloak
Image: Stego4.jpg	Phone: phone4	Type: iPhone6s	Cover: Cover2.jpg	App: Cloak

Questions?

- Thank you!
- Question to audience:
  - Since knowledge of the target users and software is important, we are looking for a partner to collaborate with to help identify issues that forensic crime labs or forensic research labs have with using steganalysis