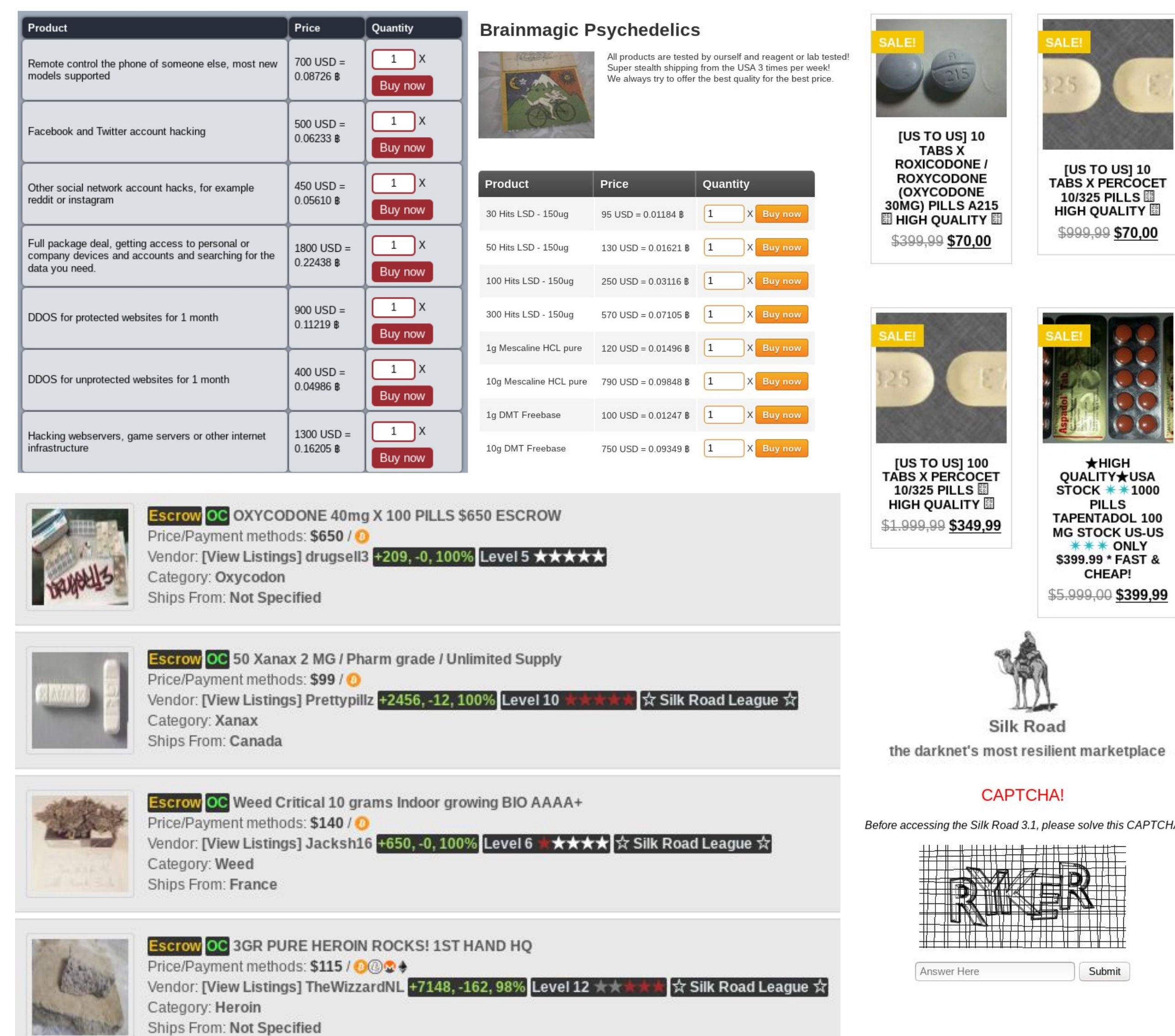


## Research Abstract

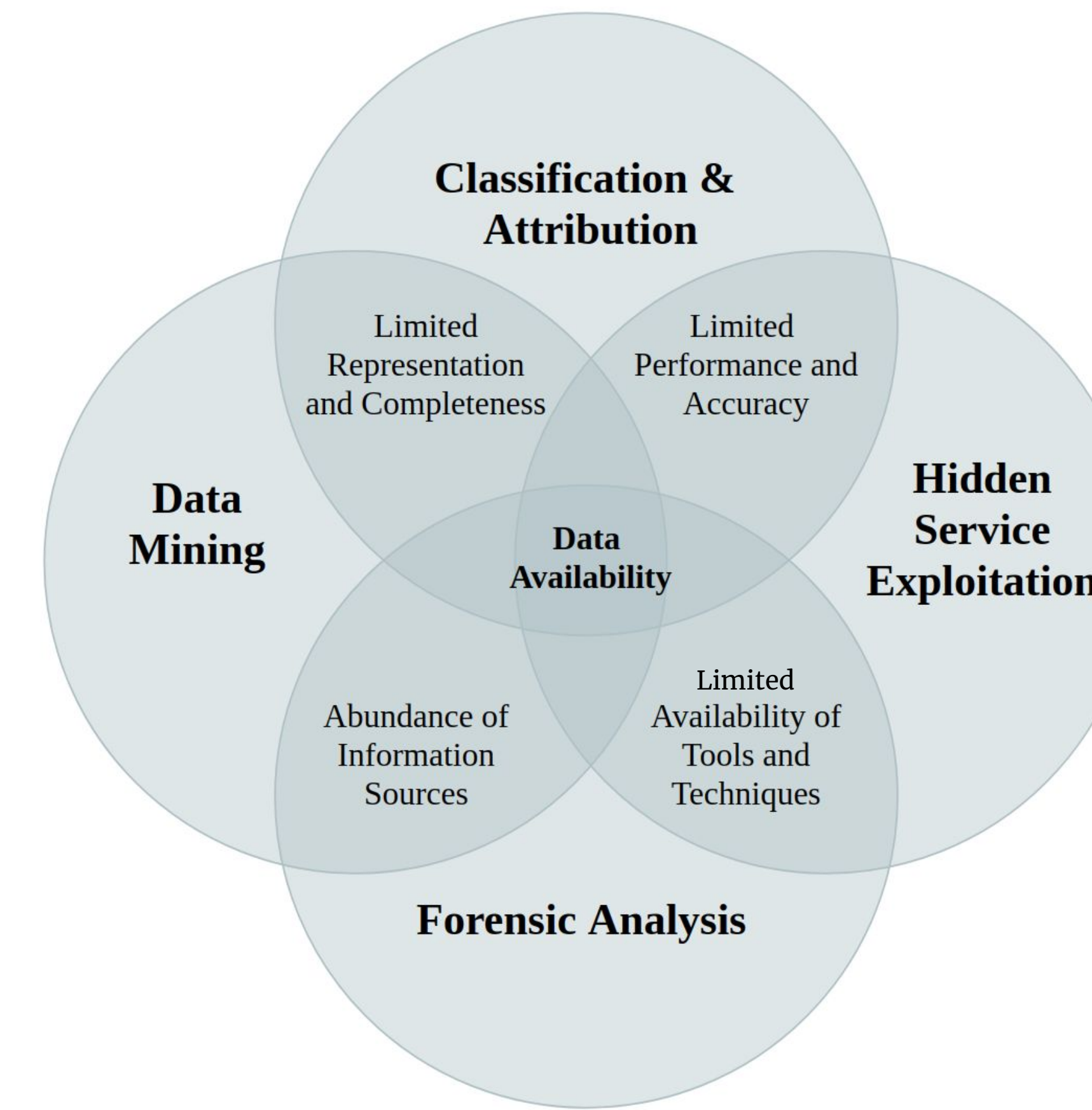
Tor hidden services and anonymity tools alike provide an avenue for cyber criminals to conduct illegal activities online without fear of consequences. In particular, dark marketplaces are hidden services that enable the trade of paraphernalia such as drugs, weapons, malware, counterfeit identities, and pornography among other items of criminal nature. Several effective Dark Web analysis techniques have been proposed for Dark Web Forums and primarily focus on authorship analysis where the goal is one of two tasks: (a) user attribution, where a user is profiled and identified given an artifact they own, and (b) alias attribution, where pairs of users are identified to belong to the same individual. While these techniques may support dark web investigations and help to identify and locate perpetrators, existing automated techniques are predominately forum-based and stylometry-based, leaving non-textual artifacts, such as images, out of consideration due to the illicit nature of dark marketplace listings. Thus, new methodologies for adequate evidence collection and image handling in dark marketplaces are essential. In this thesis, stylometric, image, and attribute-based artifacts are collected from 25 dark marketplaces and machine learning based Dark Vendor Profiling methodologies are proposed to achieve dark vendor attribution and alias attribution across dark marketplaces, thereby supporting investigative efforts in deanonymizing cyber criminals acting on the anonymous web.

## Example Marketplaces



## Problem Statement

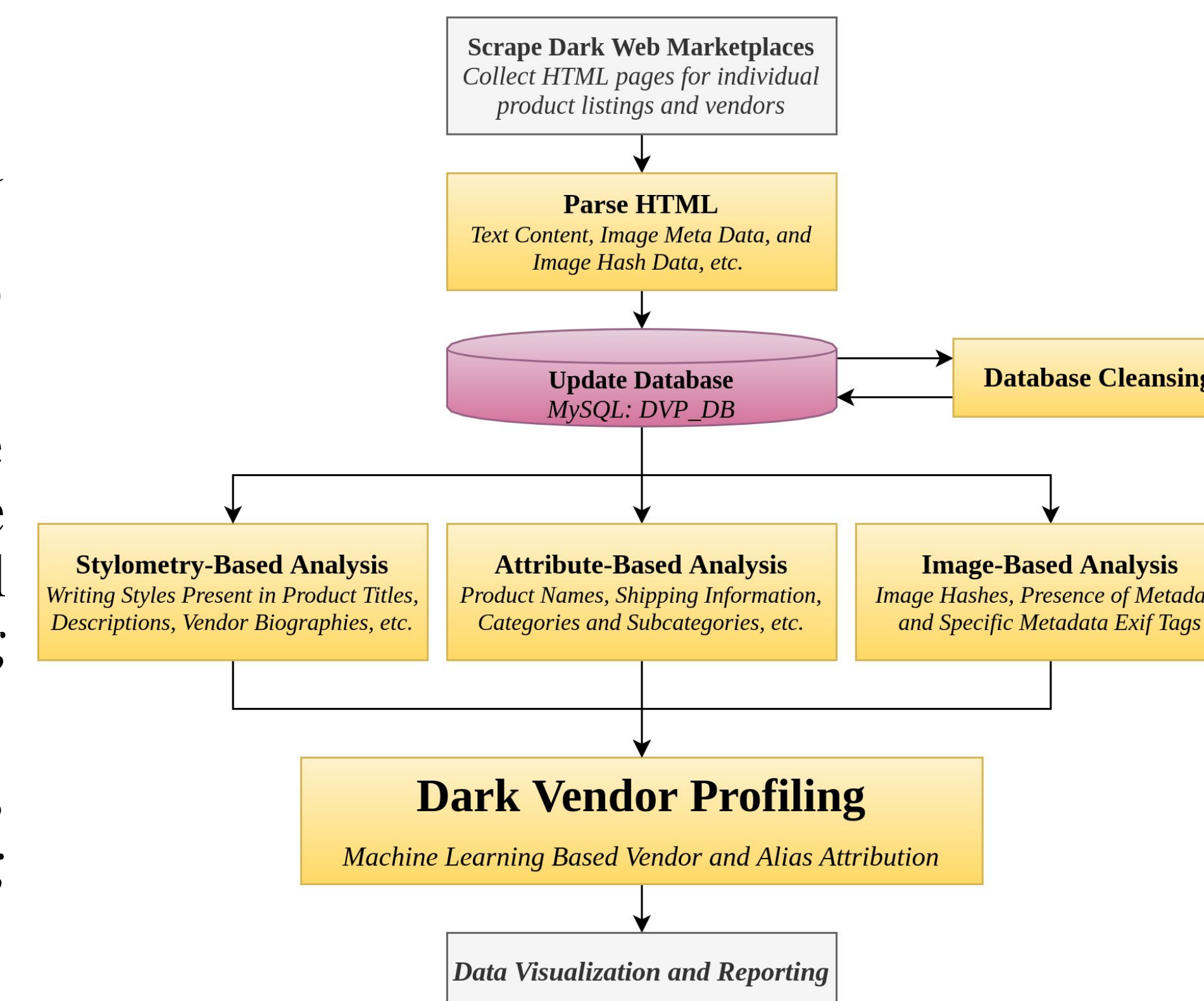
- Anonymous networks such as Tor, the Onion Router, provide cybercriminals with the means to conduct illegal business online.
- Traditional investigations regarding Dark Marketplaces are predominantly completed manually, thereby limiting the amount of markets any investigative agency may consider at one time.
- Current researches in the Dark Web domain possess several limitations as illustrated by the venn diagram on the right. These limitations are largely due to the hidden nature of anonymous networks.



## Research Contributions

- With this research, we propose *Dark Vendor Profiling*, a new method to automate the data collection and profiling of dark vendors which can support investigative efforts to deanonymize their identities.
- First, we propose the collection of image hashes in place of image content to reduce the storage demands of our proposed technique and reduce the risk of obtaining illicit digital material during data collection.
- Second, we design two unique feature sets for two authorship analysis tasks including *vendor* and *alias* attribution. These features are consequently extracted per listing and per vendor.
- Third, we propose a novel application of the Random Forest machine learning technique for the task of vendor attribution in dark marketplaces, achieving over 90% accuracy in distinguishing between over 2,500 unique dark vendors from various marketplaces.
- Fourth, we propose a novel application of the Record Linkage technique for the task of alias attribution and obtain imperative preliminary observations from Support Vector Machine and Logistic Regression based models that can assist in the design of future alias attribution models.
- Finally, we offer several future research directions for investigative analysis in dark web marketplaces.

## Proposed Workflow

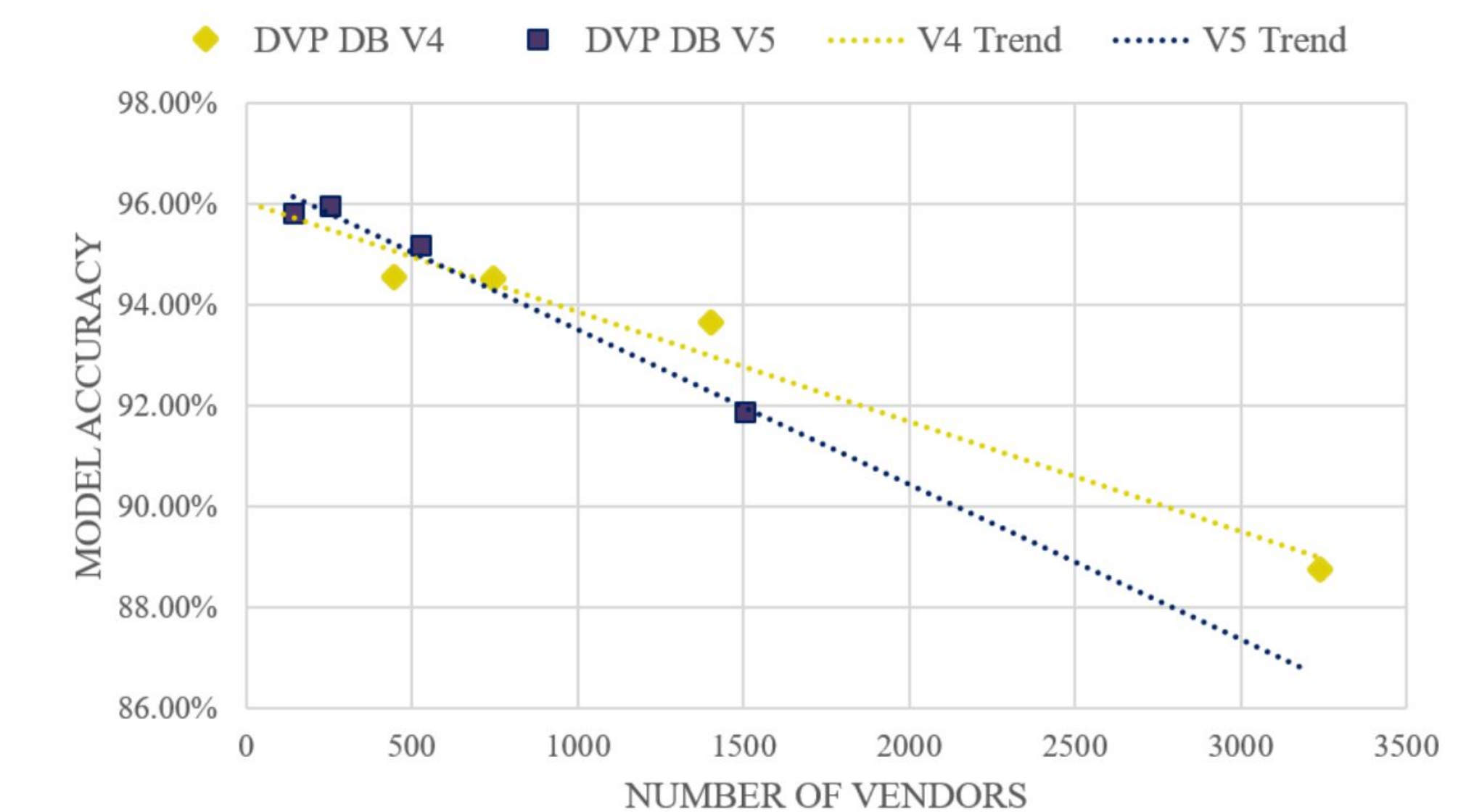


## Concluding Remarks

- Dark Vendor Profiling is a novel approach to authorship analysis in the Dark Web domain which has been implemented and tested with 25 dark marketplaces.
- Our promising results illustrate the practicality of this approach and motivate several future research directions to further support investigative efforts to deanonymize cybercriminals and reduce the impact of illegal online markets.
- We acknowledge the NSA Cybercorps (SFS) Scholarship for Service program for providing the means to complete these graduate studies.

## Results - Vendor Attribution

- A Random Forest Classifier is trained to classify vendors given their product listings. We evaluate the model in terms of performance, time complexity, and memory usage as the number of vendors, number of marketplaces, and number of listings per vendor changed in our training and testing datasets.
- To analyze the effect of 'missing data' on our model, we use a V4 dataset which contains missing data along with a V5 dataset which does not contain missing data.
- As depicted in the figure below, the random forest generally achieved over 90% accuracy which indicates that our feature set worked well to accomplish our task.



## Results - Alias Attribution

- For alias attribution, we designed two techniques for determining whether two vendors were potential aliases.
- In the first, we calculated an *unweighted Cosine similarity* metric across all features between pairs of vendors. This resulted in the vendor pairings illustrated in the table below. Note, many of the vendors had very similar names.

Marketplace	Vendor Name A	Vendor Name B	Similarity
Abraxas	fakeasanything	flawlessfakeids	0.9586
	fakeasanything	fake	0.9659
	flawlessfakeids	fake	0.9590
	Mountain	GreenMountainMan	0.9039
Agora	indiabenzos_ib	indiabenzos	0.9322
	Colorado	Colorado_Fantasy	0.9832
	UKPharma	UKPharmaceuticals	0.9620
	only	theonlysource	0.9296
Evolution	GotTheProduct	TheProduct	0.9628
	kingofcokeman	cokeman	0.9657
	TungstenGold	Goldenman	0.9730
	TheBitCoinGuru	Coin	0.9322
	monkey	howlingmonkey	0.9954
	crookscastle710	rook	0.9421
	Baron	Baron-JOY	0.9179

Model	TP	FP	FN	TN	Precision	Recall	F1 Score	Accuracy
SVM	1,278	0	0	444,778	1.0	1.0	1.0	1.0
LR	1,278	0	46	444,778	0.9653	1.0	0.9823	0.9999

Table 7.18: Results of SVM and LR training with Vendor Name as a feature in terms of True Positives (TP), False Positives (FP), False Negatives (FN), True Negatives (TN), Precision, Recall, F1 Score, and Accuracy.