“**Breaking Bad**: De-Anonymising Entity Types on the Bitcoin Blockchain Using Supervised Machine Learning ”

HICSS–51, January 2018
Internet and the Digital Economy
Distributed Ledger Technology: The Blockchain Minitrack

Awa Sun Yin
@awasunyin

About Me

Haohua (**Awa**) Sun Yin

Academic Background & Current Position

BSc. in Business & Statistics

MSc. in Information Systems

Data Scientist at Chainalysis, spec. in ML/DL & Blockchain

Research director at the Interchain Foundation*

Research Areas of Interest

Application of **ML/DL** to Blockchain data for clustering, de-anonymization, etc.

2nd and 3rd Generation Blockchains: Ethereum, Cøsmos*

Privacy Coins & Cryptography: Monero, ZCash

Agenda

1. Background & Motivations
2. Problem Formulation
3. Research Question
4. Basic Concepts
5. Methodology
6. Final Outcomes & Reflections
7. Future Research
8. Q&A

1

Background & Motivations

Adoption of Cryptocurrencies

2.9 to 5.8 million unique users (mostly Bitcoin), and increasing

Accepted as a payment method by over 100,000 merchants (~2014)

Affiliation with Illicit Activities

Used for: Money laundering, scamming, terror financing

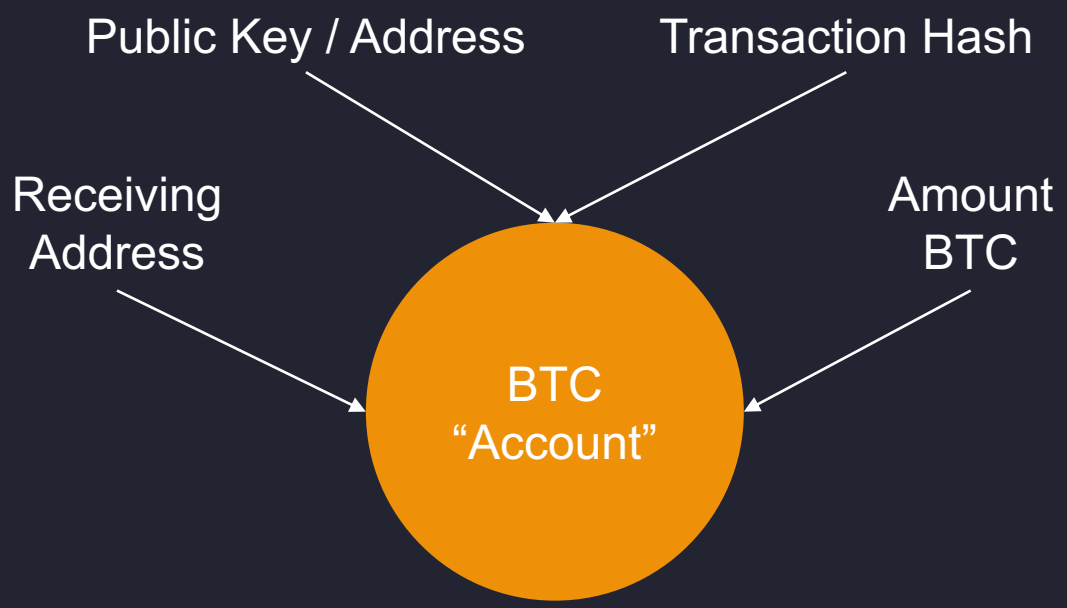
Used as payment method for: cyber-extortion (ransom payments), thievery, trading illegal goods in the Darknet

Need for Investigation & Compliance Tools

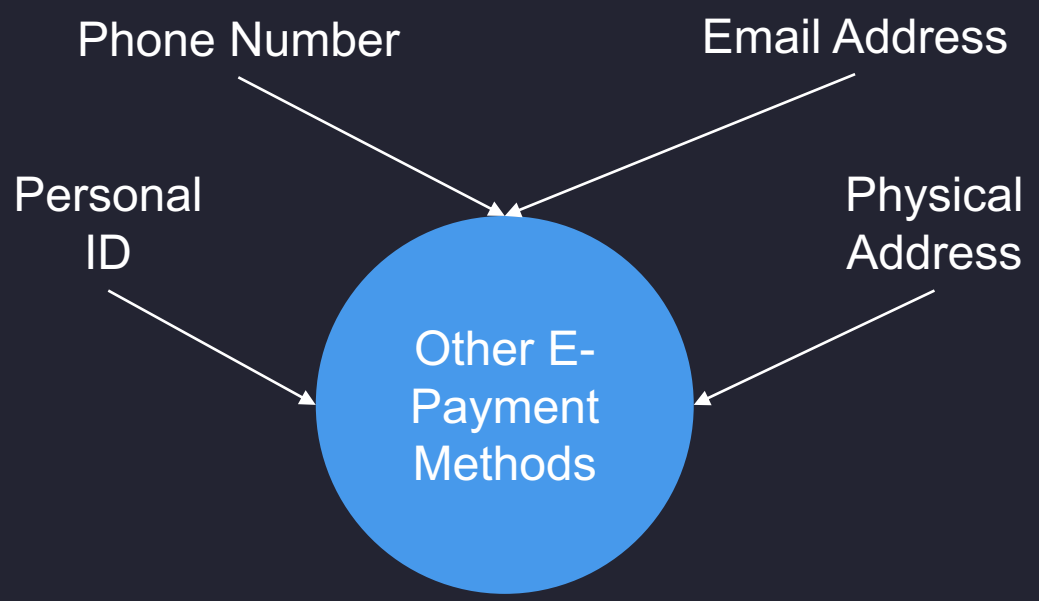
Businesses: Required by AML and KYC regulations, need tools to assess the risk of each of their customers

Law Enforcement: Need for domain specific analysis and investigation tools

2 Problem Formulation



Anyone can create anytime a BTC "account"



Issued by centralized organizations

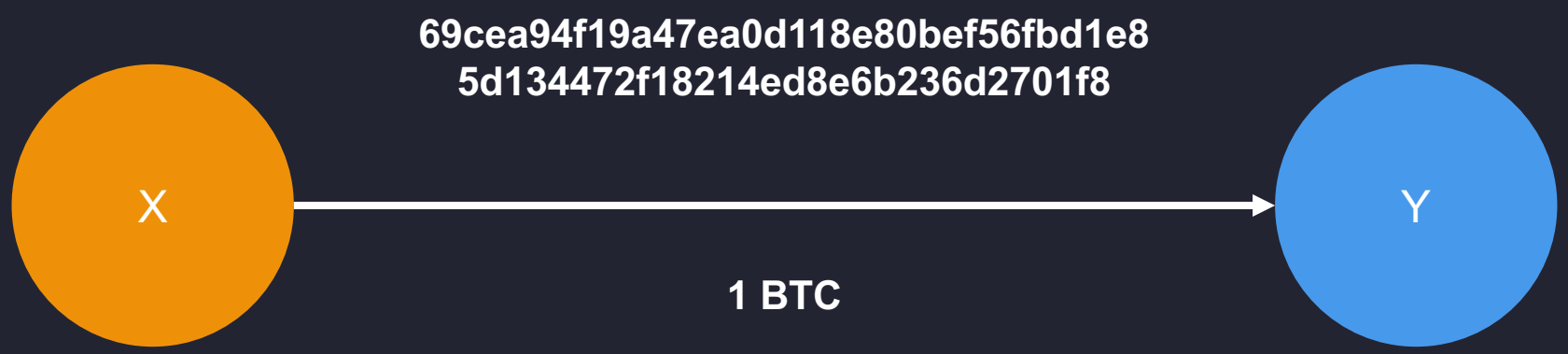
3

Research Question

To what extent can we predict the category of a yet-unidentified cluster on the Bitcoin Blockchain?

4a

Basic Concepts



12c6DSiU4Rq3P4ZxziKxzrL5LmMBrzjrJX

Category: Unknown

1DGoMT2uz6Dg59JbwtDSp8KyjiTPR7RnVQ

Category: Unknown

4b

Basic Concepts

B13HjAyJyAX6SYc6hrEqLf4naoKCWdeS
1FaEn6fGeDfDkHZvvSLsVRVxqgQt95JeyD
1JQFWqx9VFMagd8UryqiJriLmCvCuJLyf5

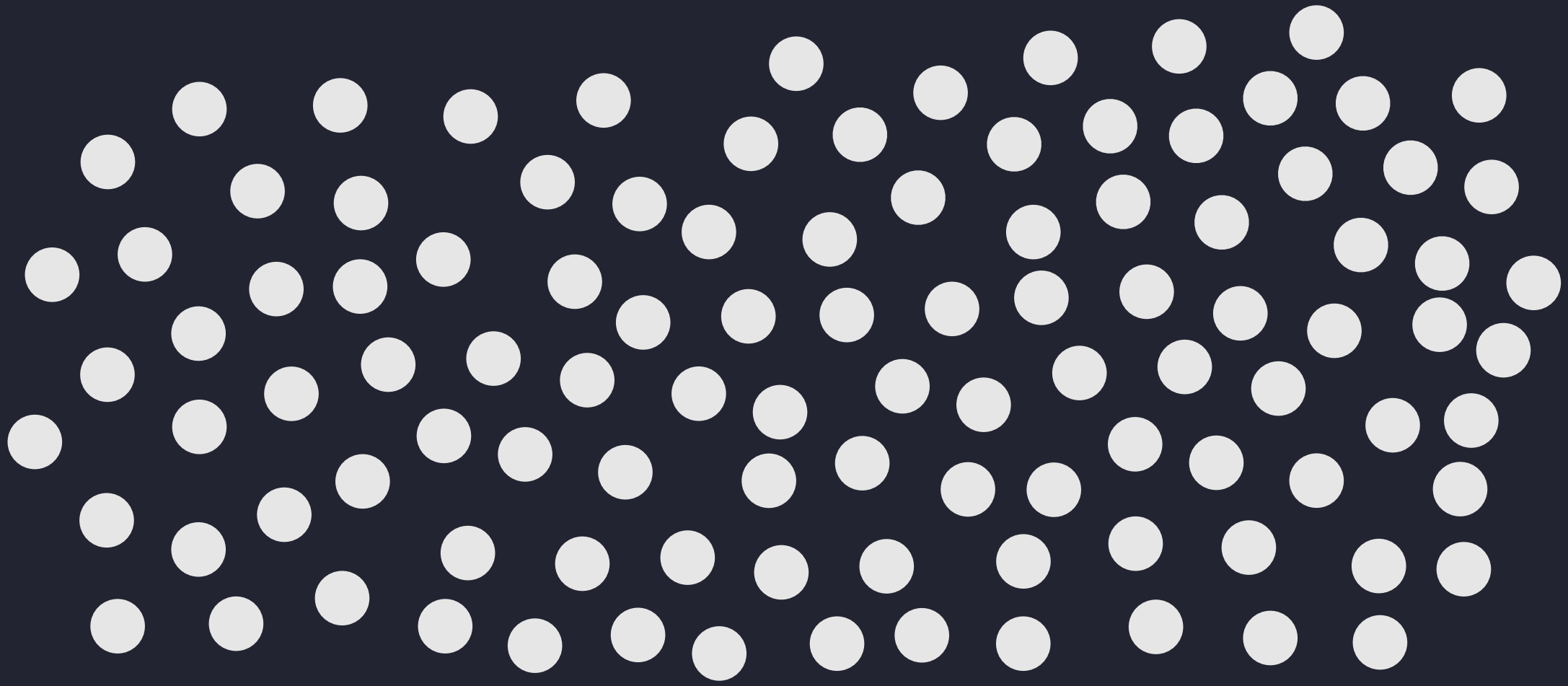
Category: Personal Wallet

1PxoZtvm3sb6LnYS5iNERuJBNR5oQZxuXJ
1K4EqXxSVd6NndkmVAMsSBd1n8VJNJLwr2

Category: Exchange

4c

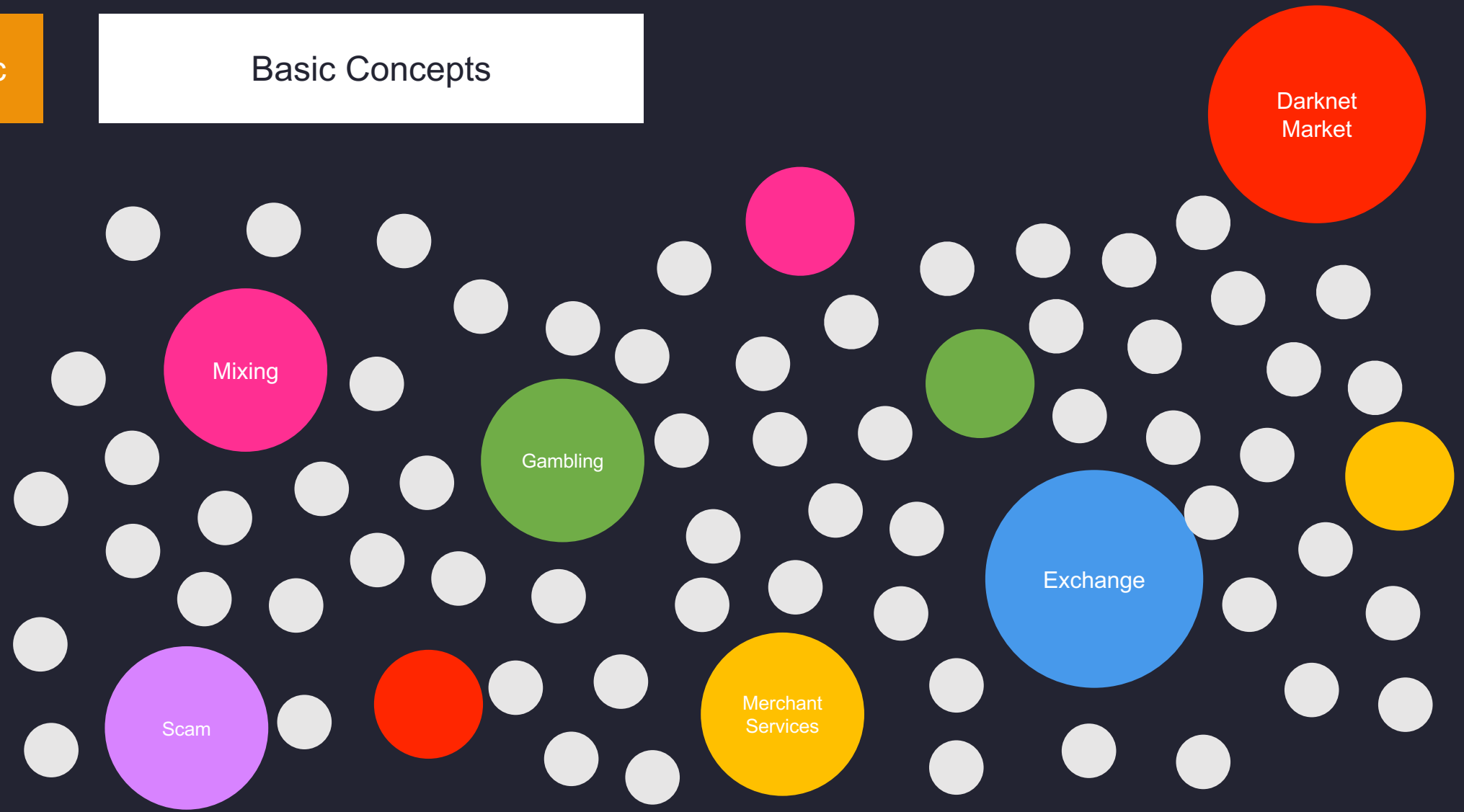
Basic Concepts



The Bitcoin Network I

4c

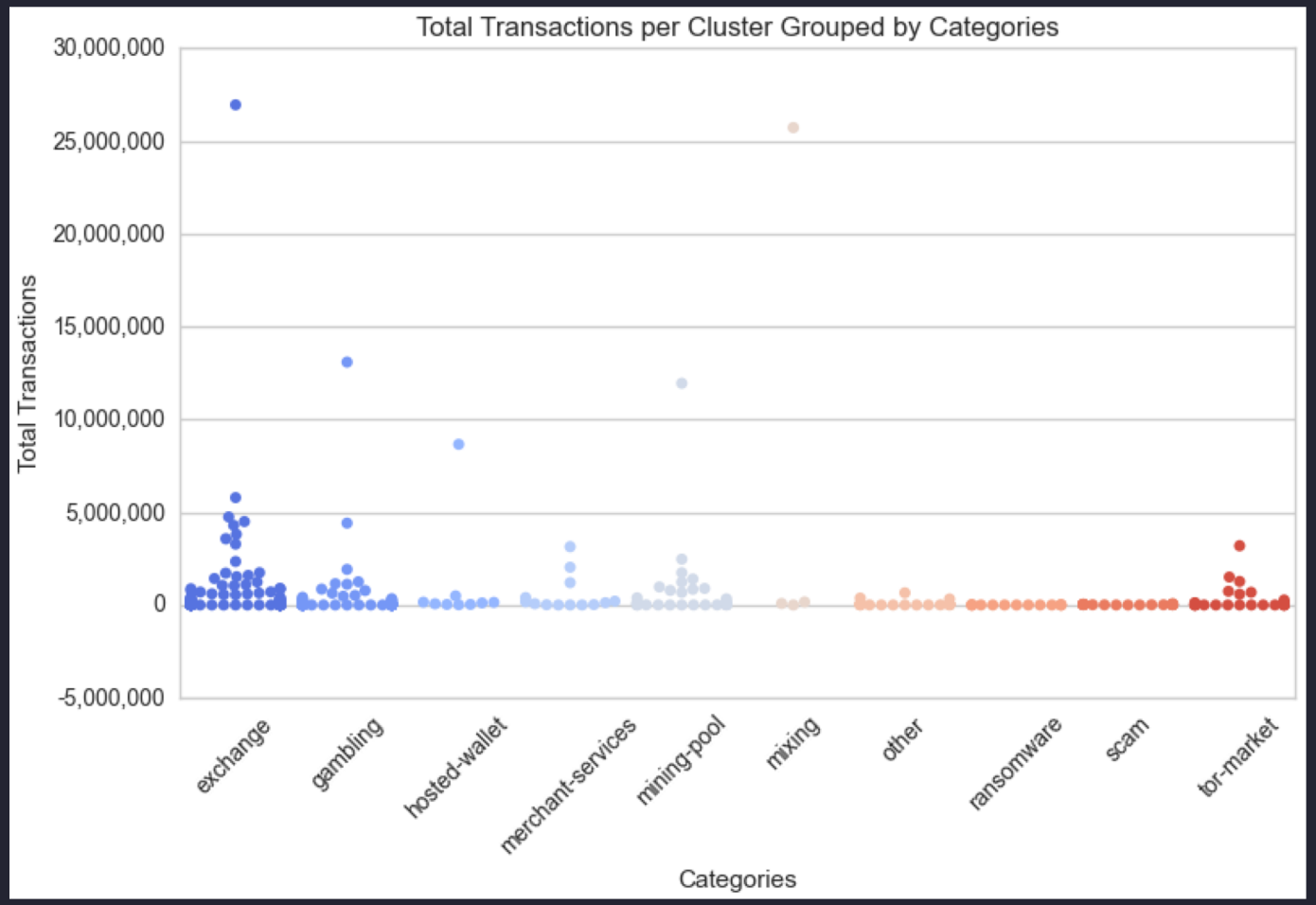
Basic Concepts



The Bitcoin Network II

5a

Methodology: Dataset

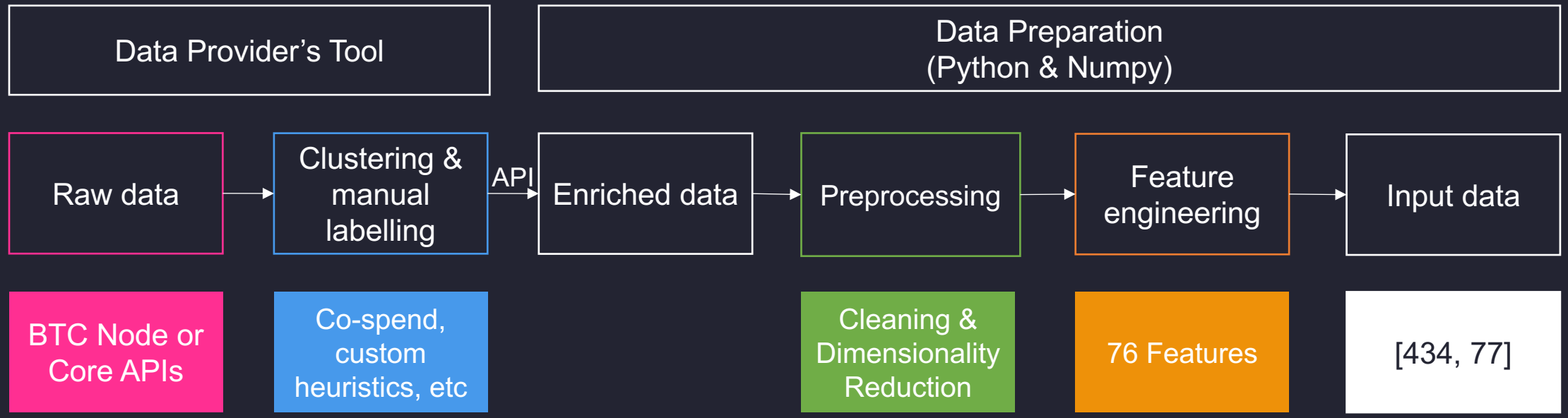


Dataset Description

- 434 Observations
- 10 categories
- + 200 M transactions
- Period: January 2009 – May 2017

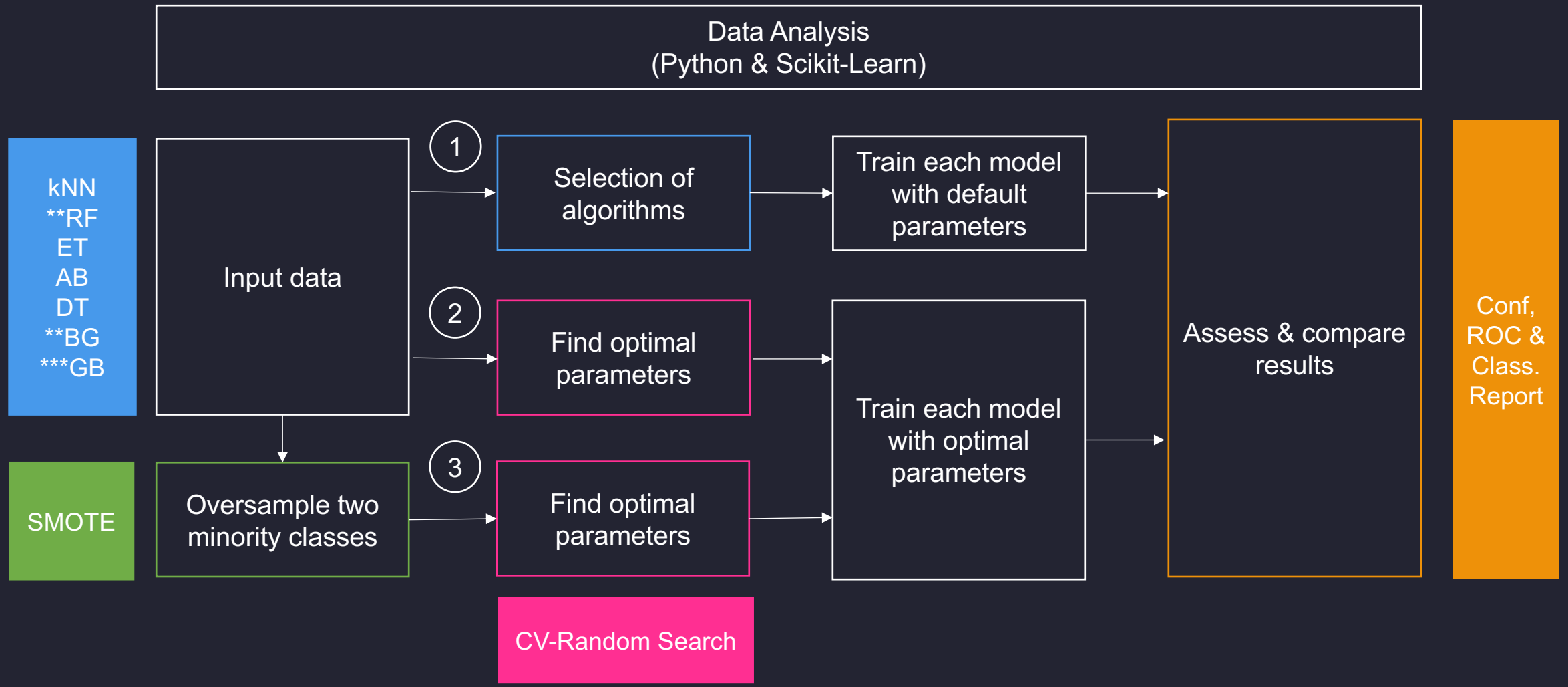
5b

Data Preparation



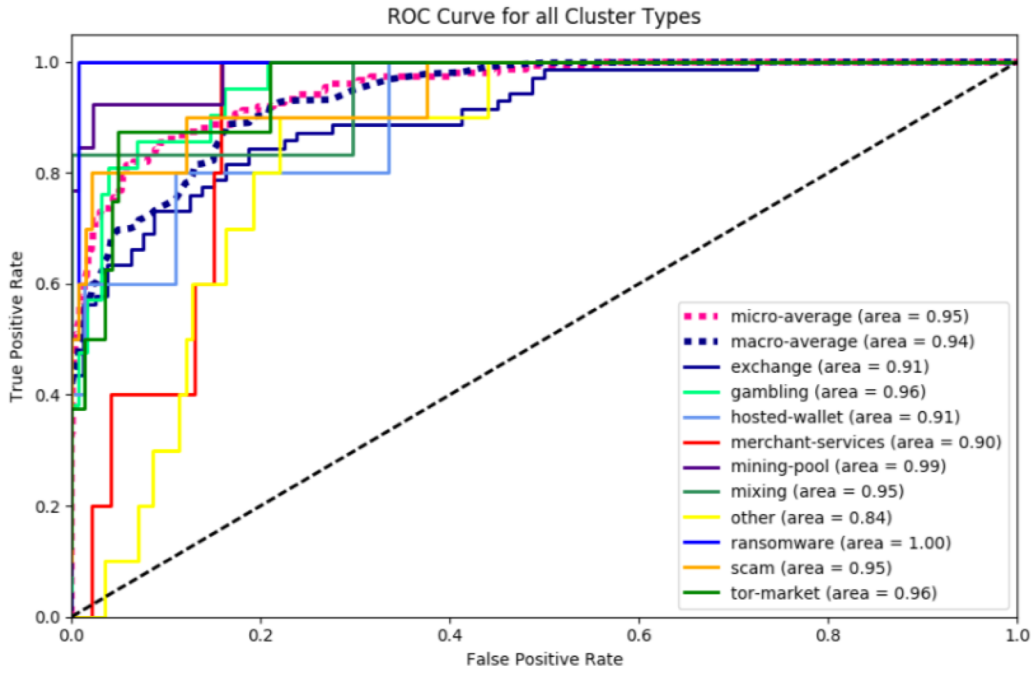
5c

Analysis: Supervised Learning



5d

Results



ROC Curve and Classification Report from Gradient Boosted Trees classifier, average confidence of 77%

Category	Precision	Recall	F1-score	Support
Exchange	0.79	0.94	0.86	201
Gambling	0.74	0.83	0.78	89
Hosted Wallet	0.25	0.11	0.15	9
Merchant Services	0.00	0.00	0.00	13
Mining Pool	0.96	0.84	0.90	31
Mixing	0.50	0.25	0.33	4
Other	0.43	0.15	0.22	20
Ransomware	0.91	0.77	0.83	13
Scam	0.68	0.59	0.63	22
Tor Market	0.79	0.59	0.68	32
Avg / Total	0.74	0.77	0.75	434

6b

Reflections

Implications

- It is possible to categorize unidentified clusters on Bitcoin using supervised learning
- Further challenging Bitcoin’s true level of anonymity
- Applicability to compliance, investigation tools

Limitations

- Dataset limited to 434 observations
- Low performance with under-sampled categories
- Features not reflecting all available data
- Lack of test set

6b

Conclusion

Multiclass Classification on the Bitcoin Blockchain

Goal: Predict the category of unidentified clusters

Methodology: Using a dataset of already identified clusters (a total of 434 observations across 10 categories)

Results: It is possible to classify with a confidence of 77% using Gradient Boosted Trees

Implications of the Research

A degree of de-anonymization can be achieved using this approach

Considering the limitations: Paving the way for future research

7a

Future Research

Expanding the Dataset of Identified Clusters

Number of observations per category

Number of categories

Refining & Testing Alternative Methodologies

Automatic feature engineering and extraction

Testing more classification algorithms

Binary Classification

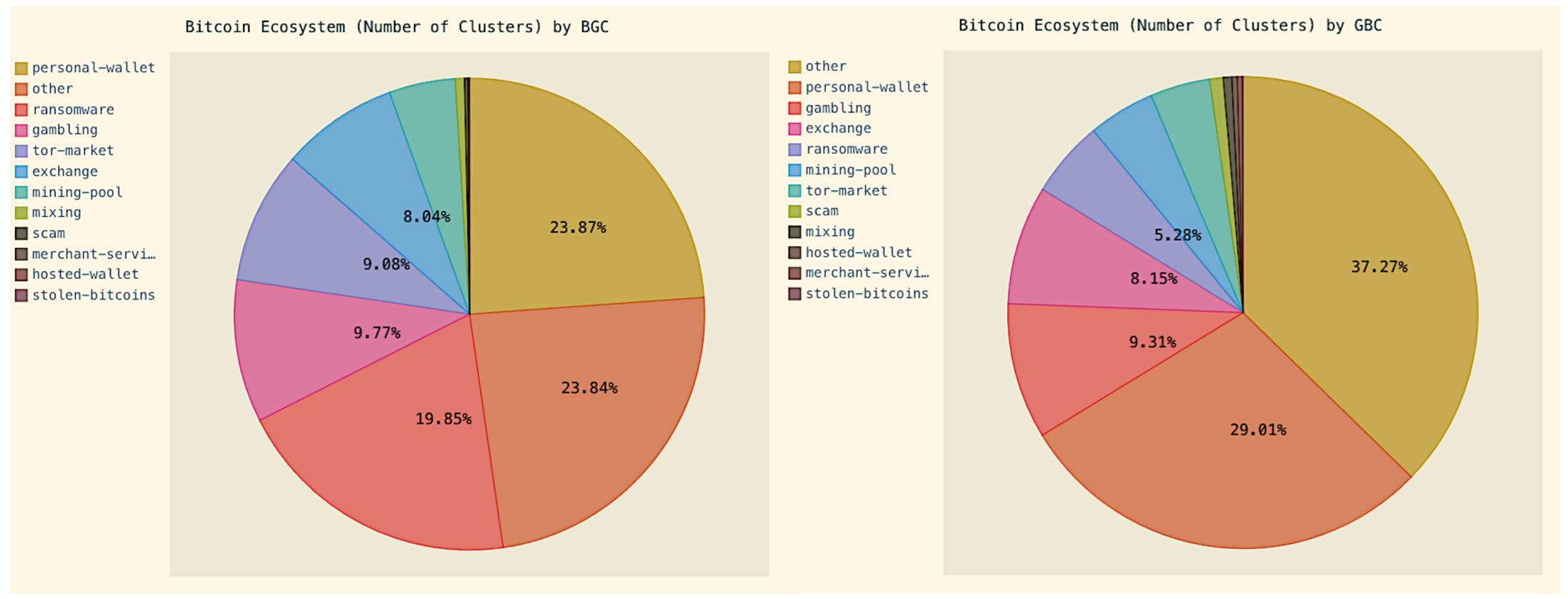
Applying the Model

Use the tested methodology to uncover the Bitcoin Blockchain for multiple purposes: cybercrime investigations, compliance tools, etc.

7b

Future Research

Haohua Sun Yin & Ravi Vatrapu, *A First Estimation of the Proportion of Cybercriminal Entities in the Bitcoin Ecosystem using Supervised Machine Learning*, IEEE Big Data for Cybercrime Prevention, 2017, Boston MA



8

Q&A

Proposed Questions

- How can we increase the dataset in both number of observations and categories?
- If Bitcoin is not truly anonymous, why has it been used for nefarious activities?
- Are there alternatives to Bitcoin that offer higher privacy?

Let’s stay in touch!

awasunyin@gmail.com

Get my pub PGP key here:
awasunyin.com