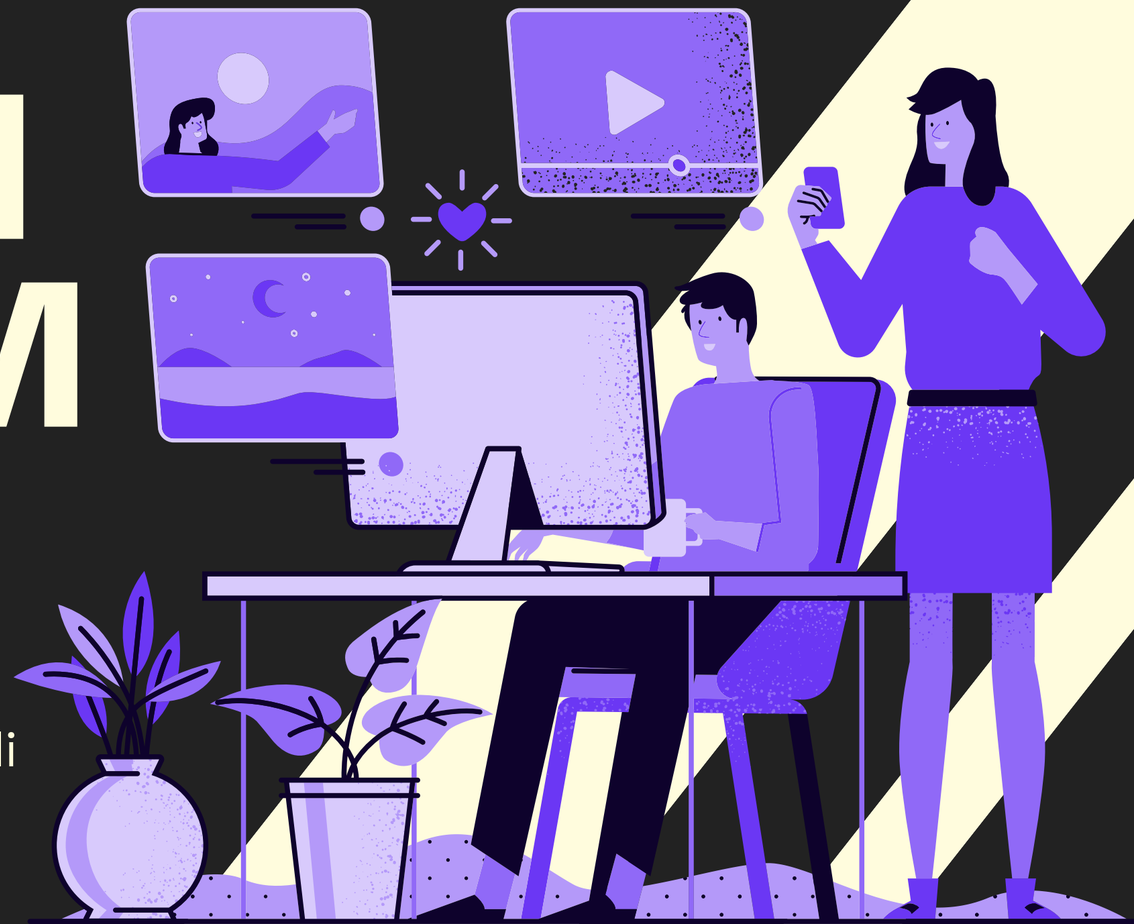
Penn Blockchain Conference

BOT DETECTION ALGORITHM

Leveraging Benford's Law and Semi-Supervised Machine Learning

Tommaso Campi, Pietro Del Bianco, Letizia Dimonopoli



Team members

Tommaso Campi
Final year Economics and Computer science student at Bocconi university, Milan, Italy

Pietro Del Bianco

Data Science - graduate 23

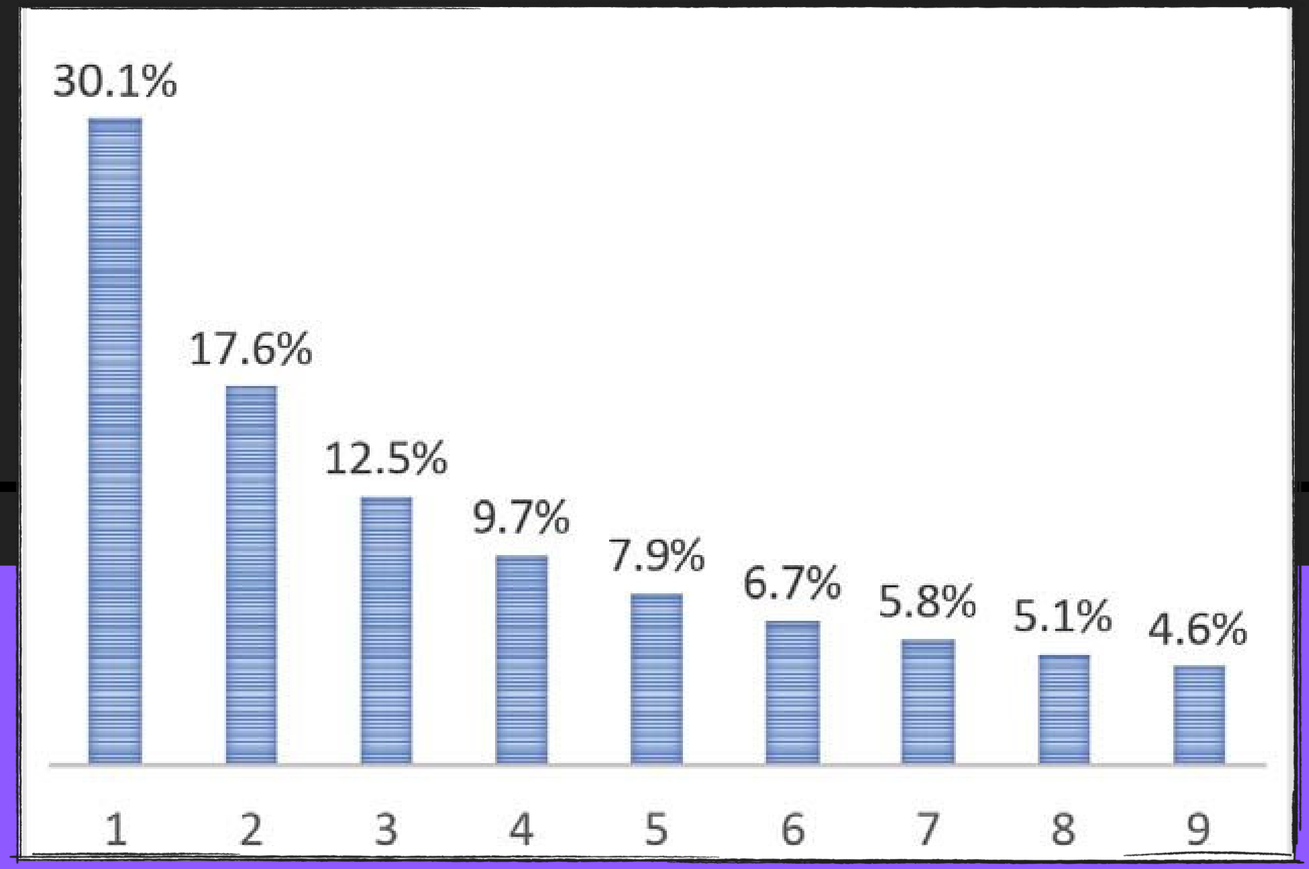
Letizia Dimonopoli
Final year Economics and Computer science student at Bocconi university, Milan, Italy

Bounty and solution approach

Bot Detection Algorithm Challenge:

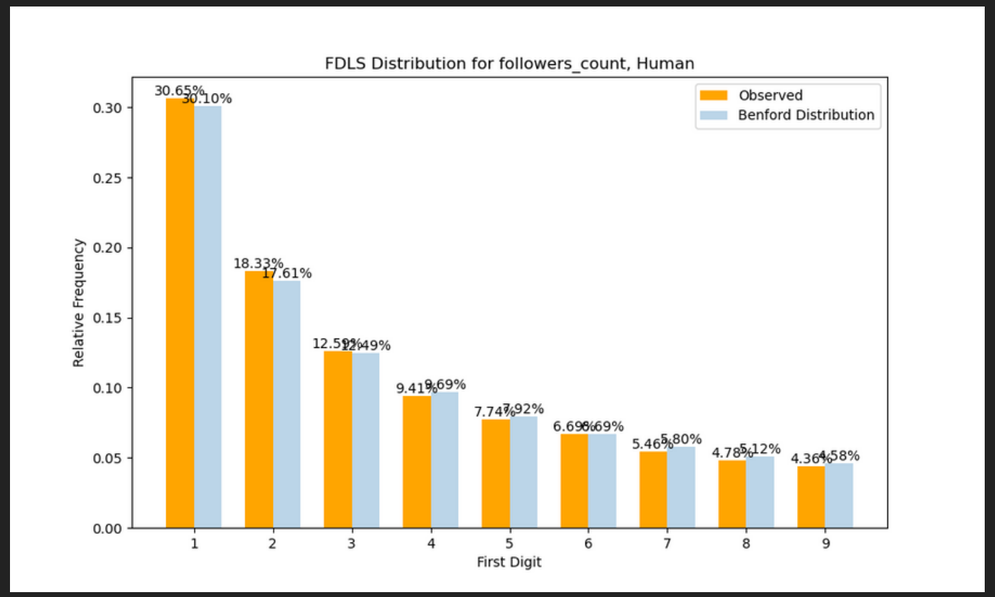
1. Analyzed Synoptic-like dataset, identifying common metrics;
2. Detected crucial variables using logistic regression, gradient boosting, random forest and SVM;
3. Applied Benford's Law;
4. Incorporated semi-supervised methods alongside Benford's Law for improved bot detection efficiency;
5. Conclusions

Benford's Law

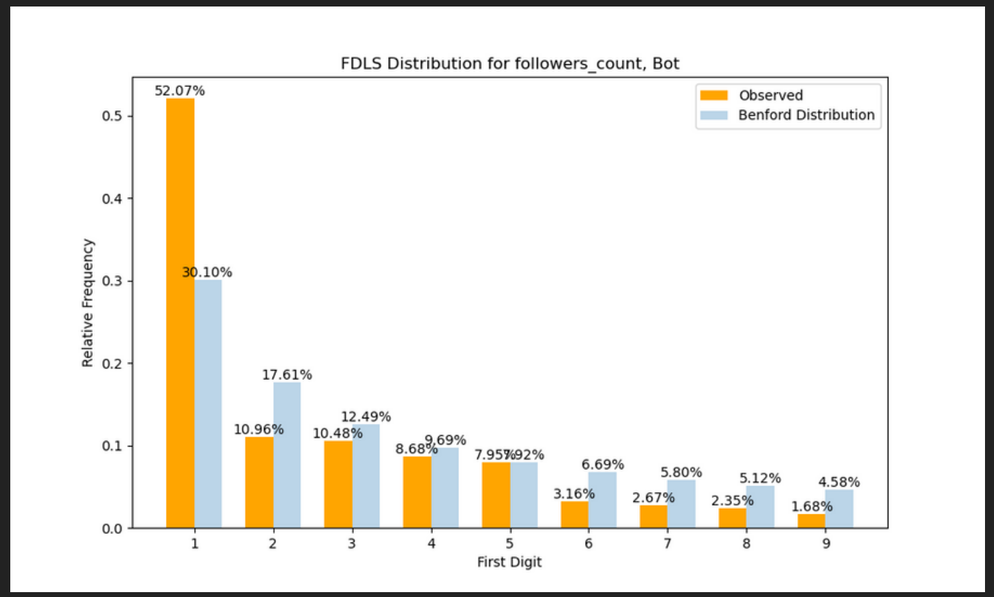


Benford's Law: our findings

Humans: cannot reject *H0*



Bots: reject *H0*



Semi-Supervised ML

According to our findings, we are now able to apply Semi-Supervised ML methods to achieve in every dataset of this kind:

1. **Efficiency**: less labeled data required compared to supervised methods, making them more efficient for tasks where labeled data is scarce.

2. **Scalability**: can handle large volumes of data efficiently, making them suitable for big data applications.

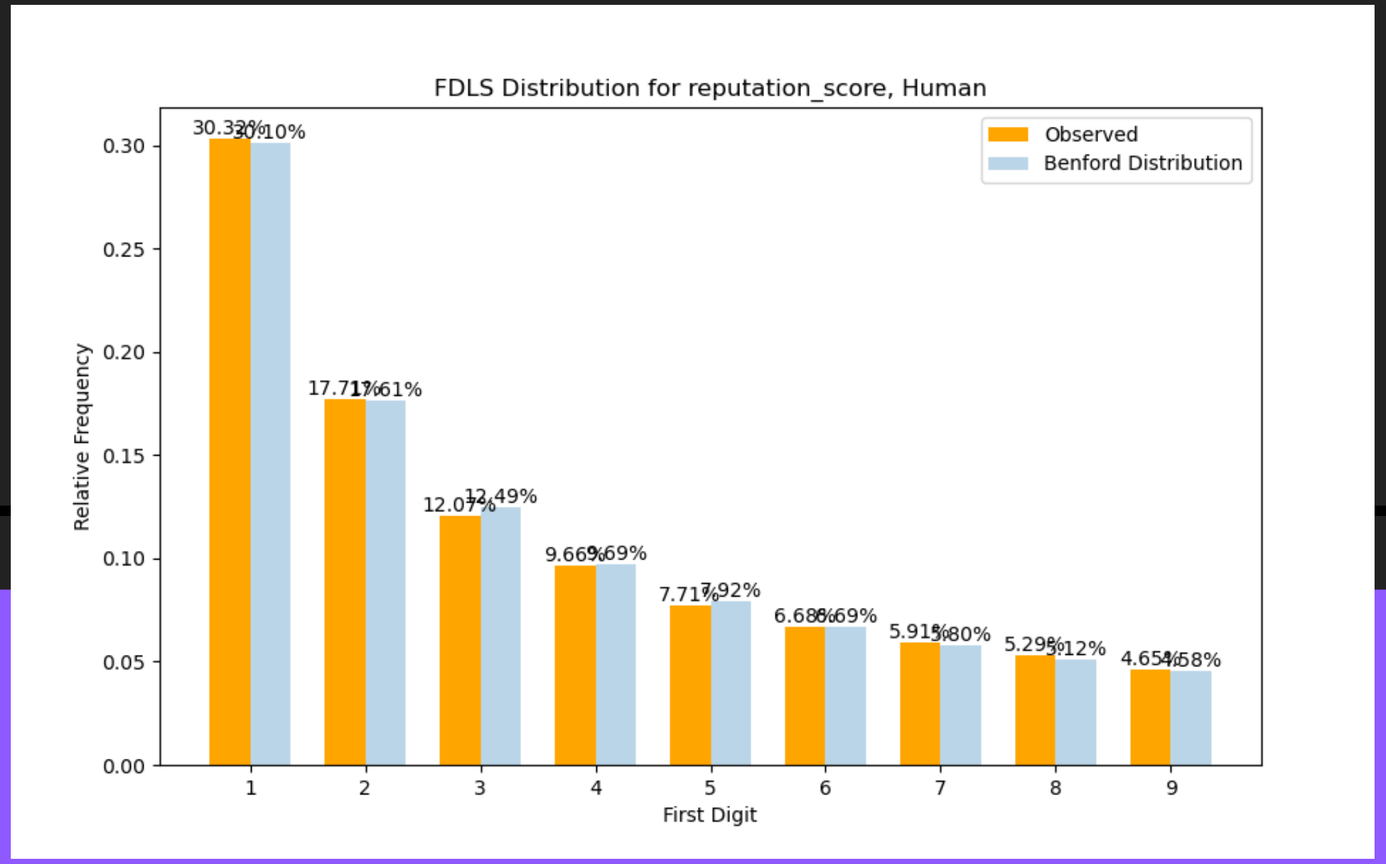
3. **Flexibility**: can discover hidden patterns and structures within data without relying on predefined labels, allowing for more flexible and exploratory analysis.

Technical overview of our algorithms

	Model	Metric				
		Accuracy	F1-Score	Precision	Recall	AUC
Supervised Learning	LR	0.96	0.57	0.69	0.49	0.74
	Lasso + SVM	0.97	0.58	0.70	0.49	0.74
	GB	0.98	0.86	0.90	0.82	0.90
	Random Forest	0.98	0.84	0.88	0.81	0.90

	Model	Metric				
		Accuracy	F1-Score	Precision	Recall	AUC
Semi Supervised Learning	GMM	0.78	0.26	0.15	0.92	0.85
	SVC	0.97	0.48	0.68	0.56	0.73
	Label Propagation	0.99	0.87	0.92	0.83	0.91
	Label Spreading	0.99	0.87	0.90	0.84	0.92

Practical example for new metrics

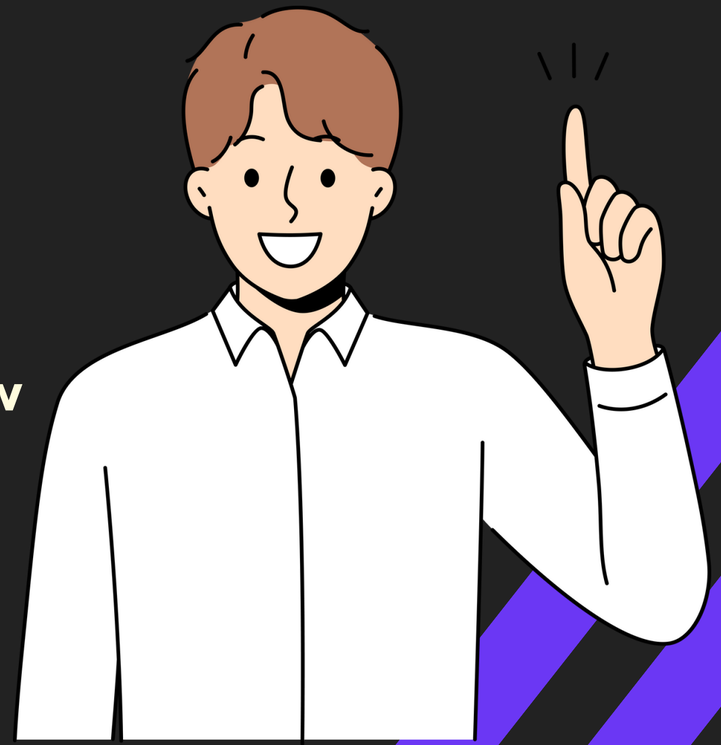


Next steps...

1. Explore semi-supervised ML algorithms, using your data and techniques available.

2. Redefine the existing metrics to make them consistent with Benford's Law.

3. Use other data that were not avilable for us to test this law



THANK YOU!