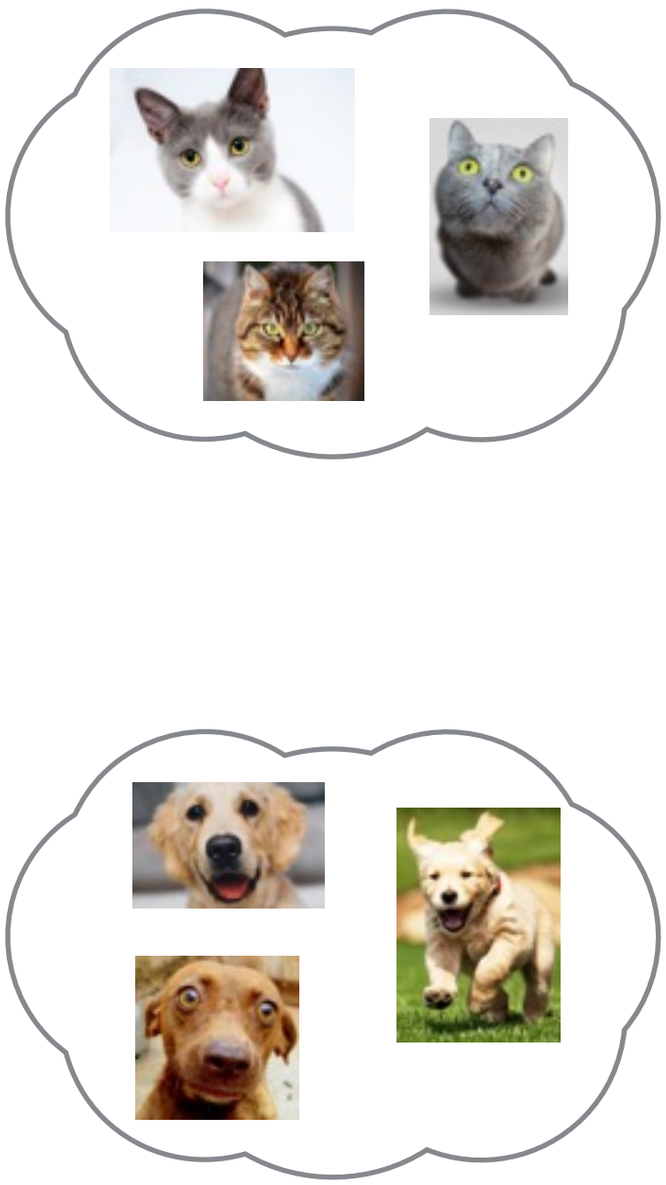
On the Security Against Machine Learning-based Attacks

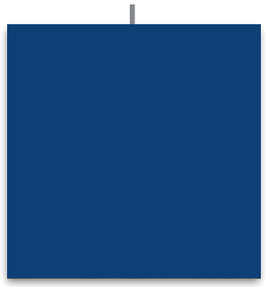
Giovanni Cherubin

@gchers

CDT Showcase Event
Evelyn Sharp Centre, Sunningdale Park
26 April, 2017

ML (Classification)

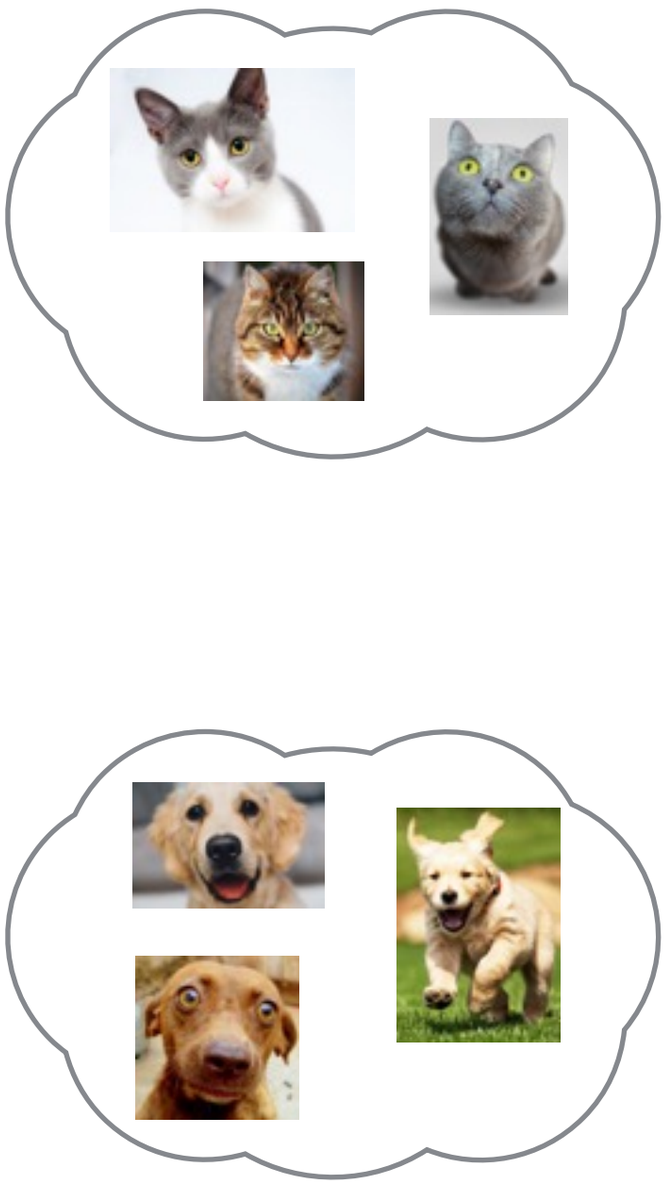


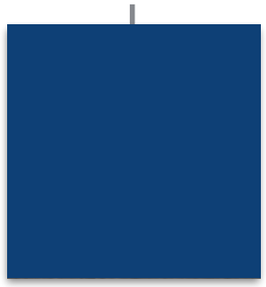


RGB
01001110001
01110000110
01110000101
01110010100
00110100111

ML System

ML (Classification)



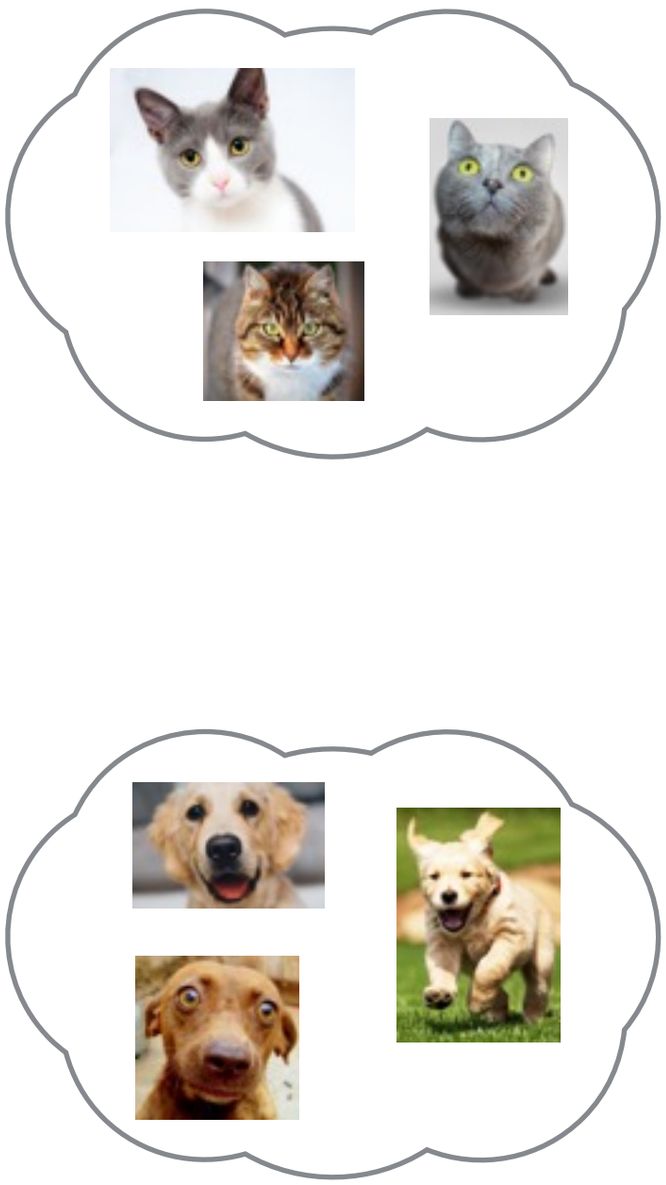


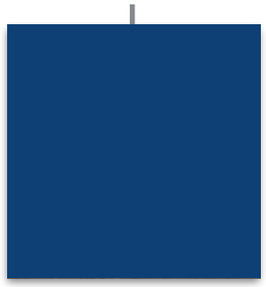
RGB
01001110001
01110000110
01110000101
01110010100
00110100111

ML System

- features (e.g., “ears shape”)

ML (Classification)





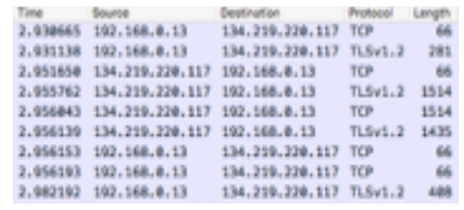
RGB
01001110001
01110000110
01110000101
01110010100
00110100111

ML System

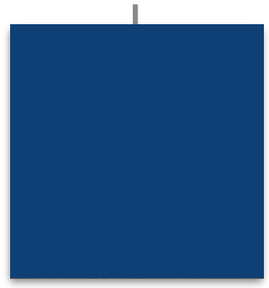
- features (e.g., “ears shape”)
- classifier (e.g., logistic regression, SVM)

Website Fingerprinting

(Back et al., 2001)



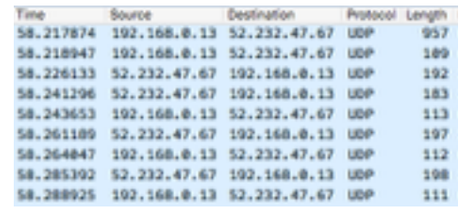
Time	Source	Destination	Protocol	Length
2.930665	192.168.0.13	134.219.220.117	TCP	66
2.931138	192.168.0.13	134.219.220.117	TLSv1.2	281
2.951650	134.219.220.117	192.168.0.13	TCP	66
2.955762	134.219.220.117	192.168.0.13	TLSv1.2	1514
2.956043	134.219.220.117	192.168.0.13	TCP	1514
2.956139	134.219.220.117	192.168.0.13	TLSv1.2	1435
2.956153	192.168.0.13	134.219.220.117	TCP	66
2.956193	192.168.0.13	134.219.220.117	TCP	66
2.982192	192.168.0.13	134.219.220.117	TLSv1.2	488



Adversary

Language of Voip Traffic

(Wright et al., 2007)

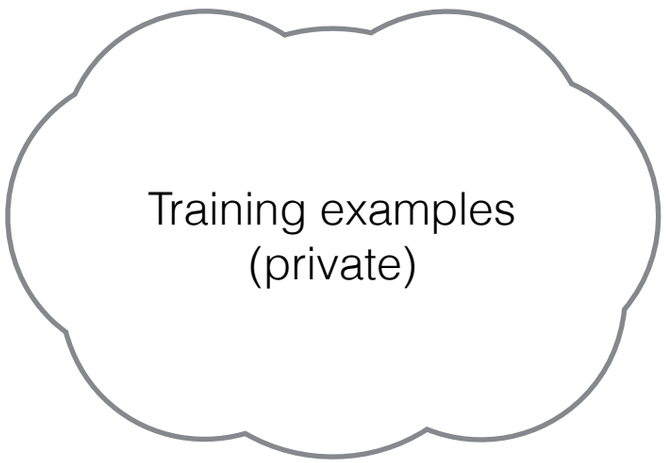


Time	Source	Destination	Protocol	Length
58.217874	192.168.0.13	52.232.47.67	UDP	957
58.218947	192.168.0.13	52.232.47.67	UDP	189
58.226133	52.232.47.67	192.168.0.13	UDP	192
58.241296	52.232.47.67	192.168.0.13	UDP	183
58.243653	192.168.0.13	52.232.47.67	UDP	113
58.261189	52.232.47.67	192.168.0.13	UDP	197
58.264847	192.168.0.13	52.232.47.67	UDP	112
58.285392	52.232.47.67	192.168.0.13	UDP	198
58.288925	192.168.0.13	52.232.47.67	UDP	111

Adversary

Membership Inference

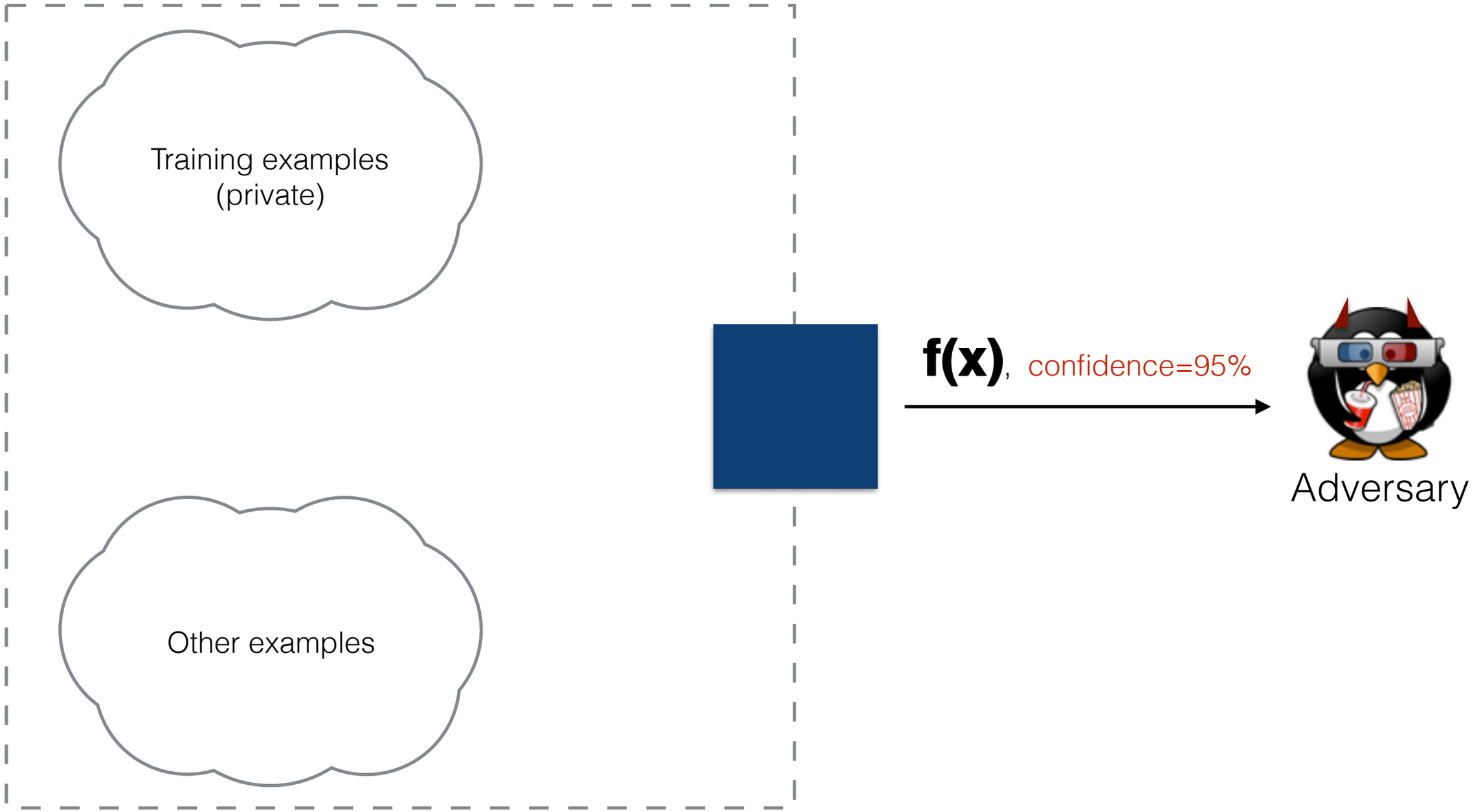
(Shokri et al., 2017)



f(x)

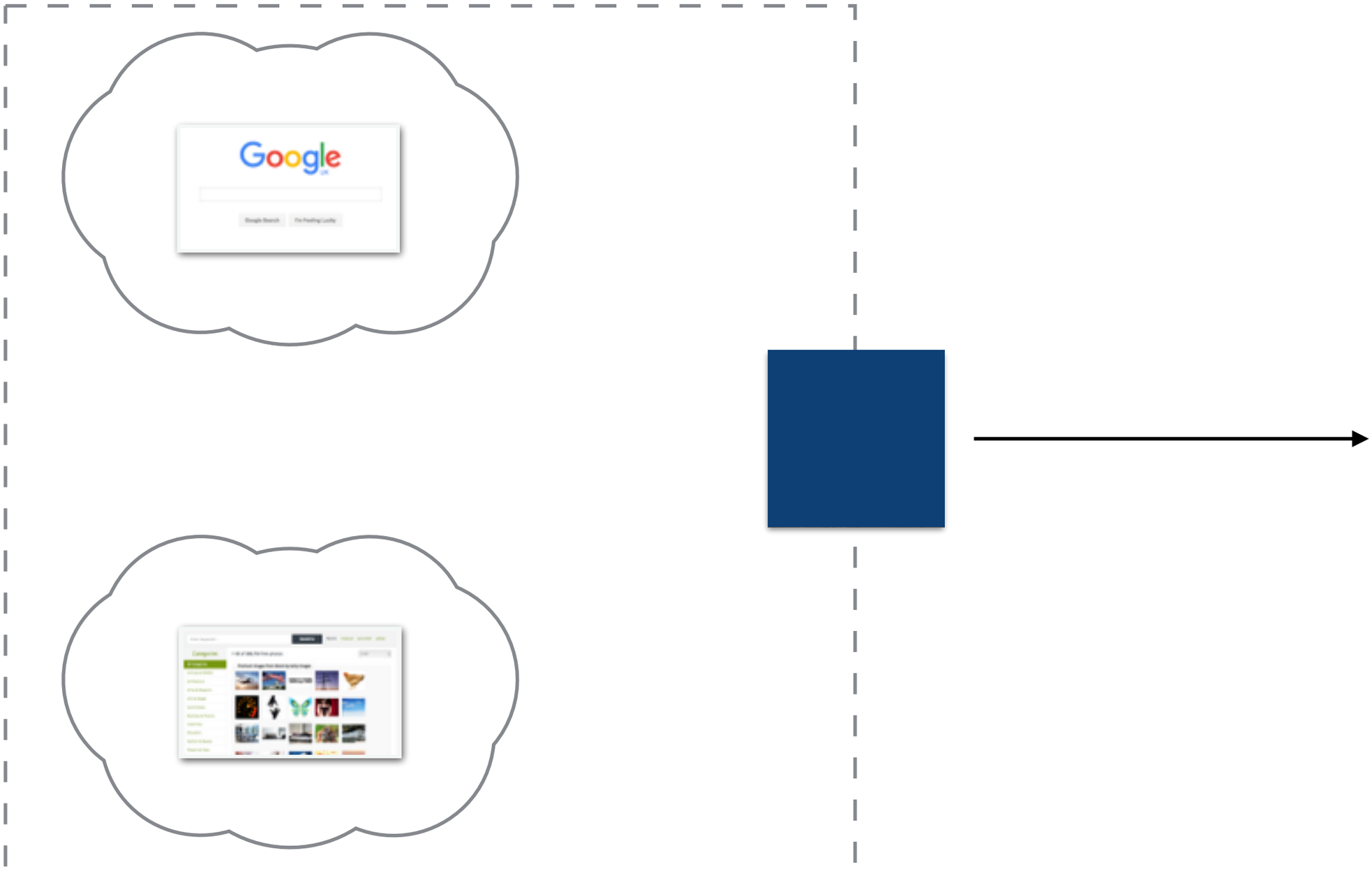
Membership Inference

(Shokri et al., 2017)



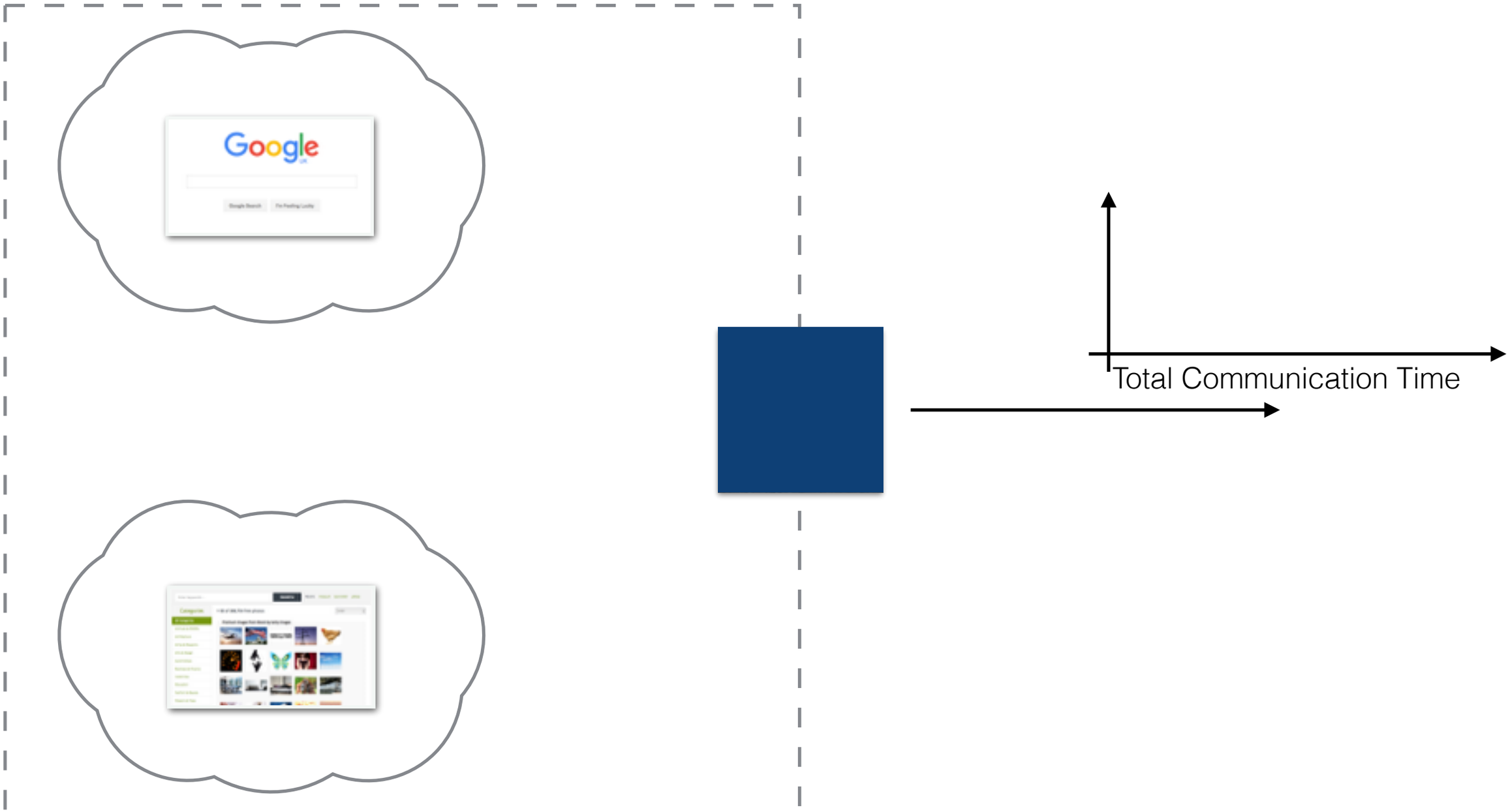
“Blackbox” approach

(Bayes error estimate)



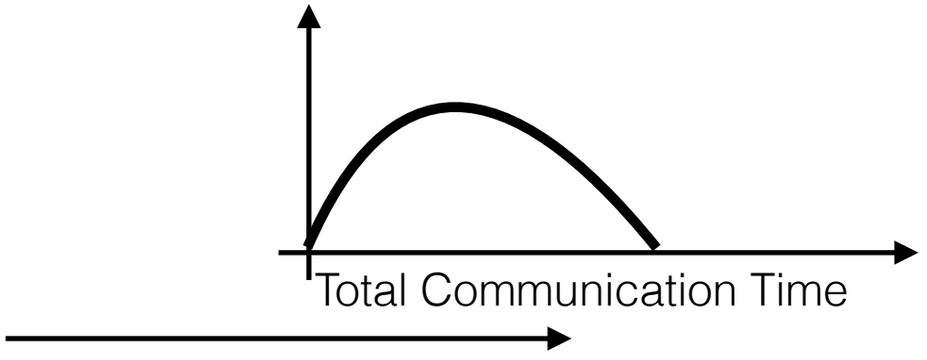
“Blackbox” approach

(Bayes error estimate)



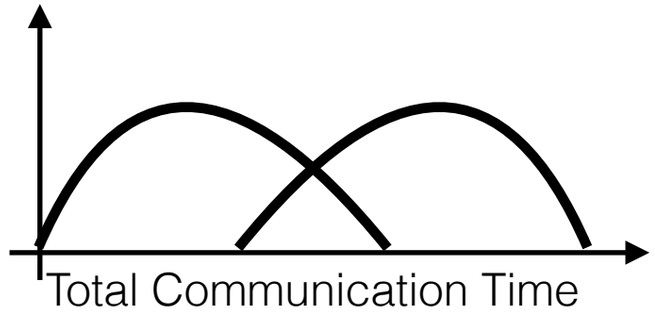
“Blackbox” approach

(Bayes error estimate)



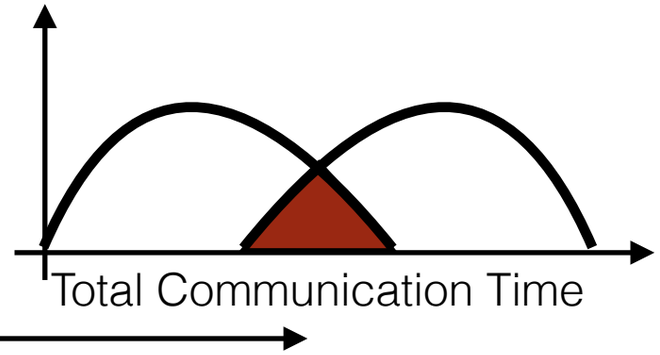
“Blackbox” approach

(Bayes error estimate)



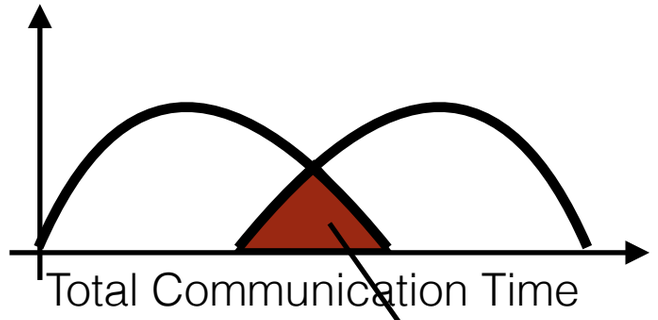
“Blackbox” approach

(Bayes error estimate)



“Blackbox” approach

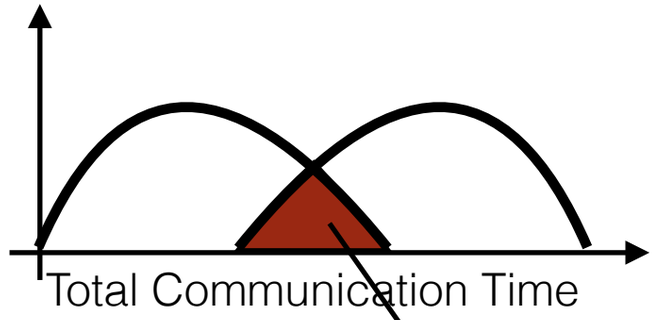
(Bayes error estimate)



R*

“Blackbox” approach

(Bayes error estimate)



$\Pr(\text{ is wrong}) \geq R^*$

ML System

Recap

Pros

- Blackbox, arbitrary systems/countermeasures
- Captures many attacks (potentially all ML-based ones)
- Probability of error (rather than “negligible indistinguishability”)

Cons & Future Research

- Asymptotic results (Bayes estimate)
- Needs “realistic” data
- Reliance on features (for some attacks)

References

- 2001, A. Back, U. Möller, A. Stiglic. "Traffic analysis attacks and trade-offs in anonymity providing systems." International Workshop on Information Hiding. Springer Berlin Heidelberg.

- 2007, C. V. Wright, L. Ballard, F. Monrose, G. M. Masson "Language identification of encrypted voip traffic: Alejandra y Roberto or Alice and Bob?." USENIX Security. Vol. 3. No. 3.6.

- 2017, R. Shokri, M. Stronati, V. Shmatikov. "Membership inference attacks against machine learning models." arXiv preprint arXiv: 1610.05820

- 2017, G. Cherubin. "Bayes, not Naïve: Security Bounds on Website Fingerprinting Defenses." arXiv preprint arXiv:1702.07707.

On the Security Against Machine Learning-based Attacks

Giovanni Cherubin

@gchers

CDT Showcase Event
Evelyn Sharp Centre, Sunningdale Park
26 April, 2017