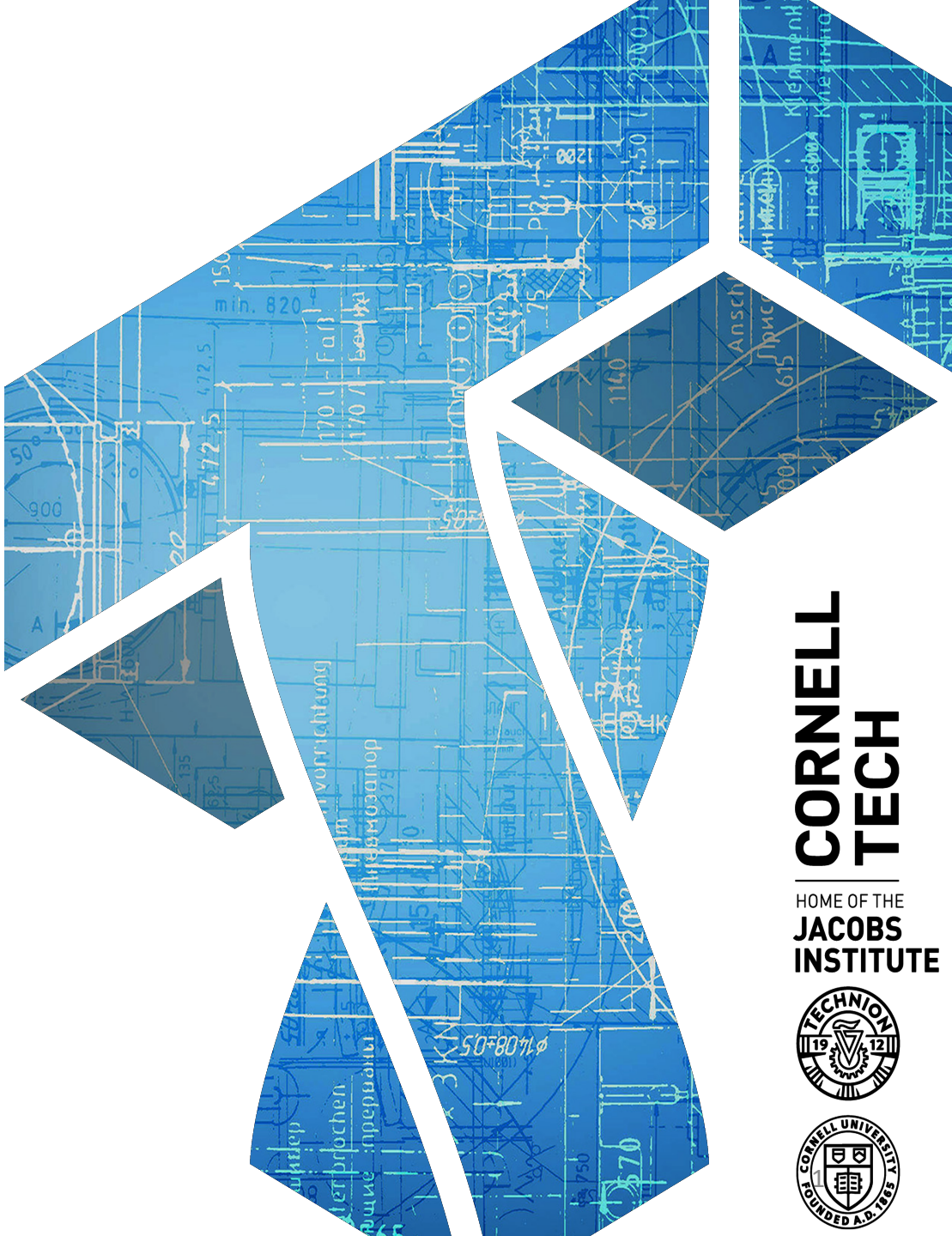
CS 6431:
ML Security



Instructor: Tom Ristenpart

https://github.com/tomrist/cs6431-fall2021

CORNELL TECH

HOME OF THE JACOBS INSTITUTE

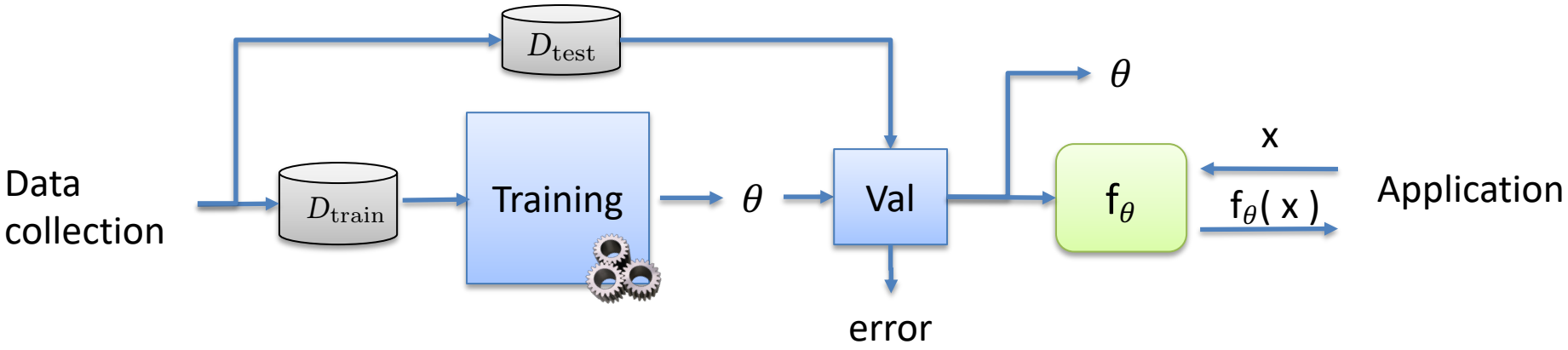
Reminder! Project proposal due soon

- Feel free to bounce ideas off me, and I'll try to provide feedback quickly. Slack DM is probably easiest

ML in security: detecting attacks

- **Unsupervised machine learning**
  - Anomaly detection (Denning 1987)
  - Attacks defined as outlier behavior
  - E.g.: how far away from mean is some monitored behavior?
- **Supervised machine learning**
  - Lots of malicious and benign examples
  - Train an ML model to classify behavior as malicious or not
- **Semi-supervised machine learning**
  - Lots of unlabeled data, some labeled data
- **Active learning**
  - Have oracle that labels data for you, minimize queries

Supervised machine learning (ML) pipeline



(1) Collect labeled data x[1] , y[1] x[2], y[2] …

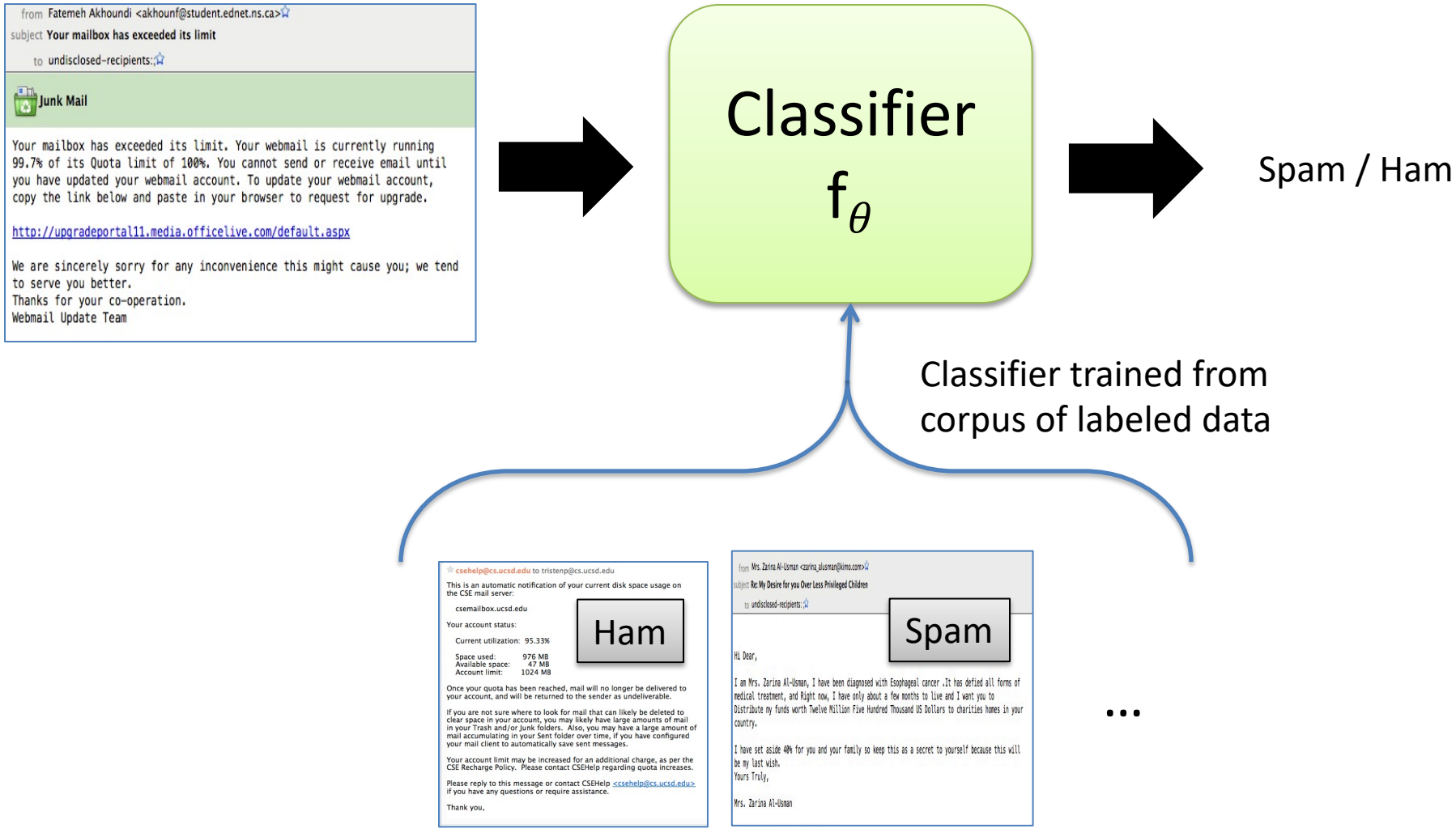
n-dimensional feature vector x

Dependent variable y

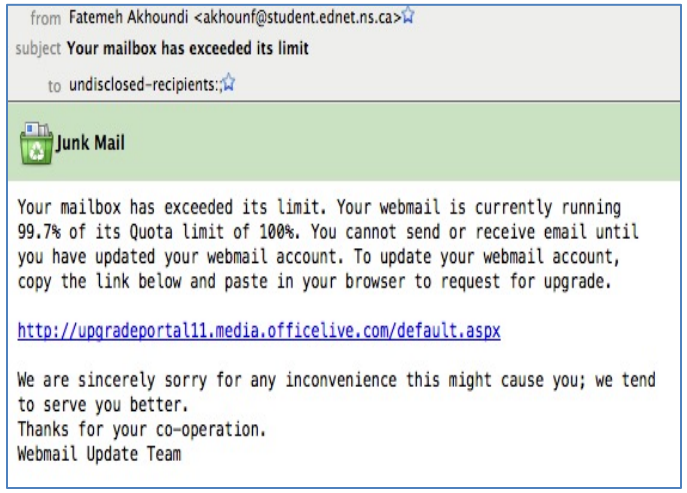
(2) Train & validate ML model θ to allow prediction: $f_\theta(x) = y$

(3) Use f_θ in some application or publish θ for others to use

Spam Classifiers



Multinomial Naïve Bayes Classifier


from Fatemeh Akhoundi <akhounf@student.ednet.ns.ca>
subject **Your mailbox has exceeded its limit**
to undisclosed-recipients:;

Junk Mail

Your mailbox has exceeded its limit. Your webmail is currently running 99.7% of its Quota limit of 100%. You cannot send or receive email until you have updated your webmail account. To update your webmail account, copy the link below and paste in your browser to request for upgrade.

http://upgradeportal11.media.officelive.com/default.aspx

We are sincerely sorry for any inconvenience this might cause you; we tend to serve you better.
Thanks for your co-operation.
Webmail Update Team

Represent email as "bag of words"

quota	1	x_1
webmail	4	x_2
cornell	0	x_3
fee	1	x_4
⋮	⋮	

Intuition: spam and ham have different distribution of keywords

$$\Pr[\text{spam} \mid x_1, x_2, \ldots, x_n] = \frac{\Pr[x_1, x_2, \ldots, x_n \mid \text{spam}]\ \Pr[\text{spam}]}{\Pr[x_1, x_2, \ldots, x_n]}$$

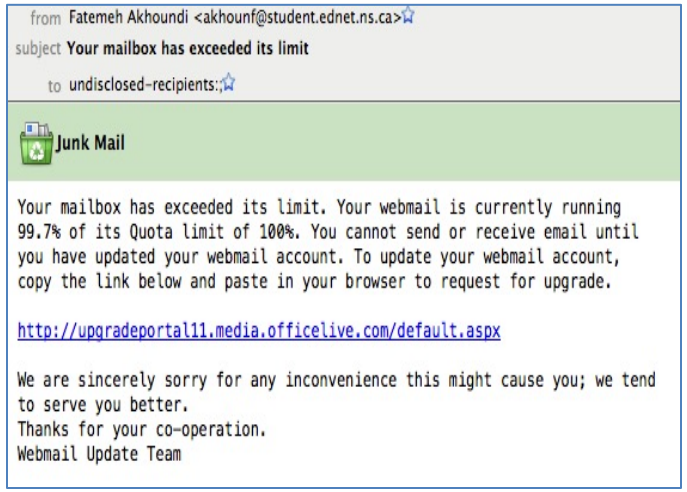
Bayes' theorem

$$= \frac{\Pr[\text{spam}] \prod \Pr[x_i \mid \text{spam}]}{\Pr[x_1, x_2, \ldots, x_n]}$$

"Naïve": assume words independent

$$\Pr[\text{ham} \mid x_1, x_2, \ldots, x_n] = \frac{\Pr[\text{ham}] \prod \Pr[x_i \mid \text{ham}]}{\Pr[x_1, x_2, \ldots, x_n]}$$

Multinomial Naïve Bayes Classifier


from Fatemeh Akhoundi <akhounf@student.ednet.ns.ca>
subject **Your mailbox has exceeded its limit**
to undisclosed-recipients:;

Junk Mail

Your mailbox has exceeded its limit. Your webmail is currently running 99.7% of its Quota limit of 100%. You cannot send or receive email until you have updated your webmail account. To update your webmail account, copy the link below and paste in your browser to request for upgrade.

http://upgradeportal11.media.officelive.com/default.aspx

We are sincerely sorry for any inconvenience this might cause you; we tend to serve you better.
Thanks for your co-operation.
Webmail Update Team

Represent email as "bag of words"

quota	1	x_1
webmail	4	x_2
cornell	0	x_3
fee	1	x_4
⋮	⋮	

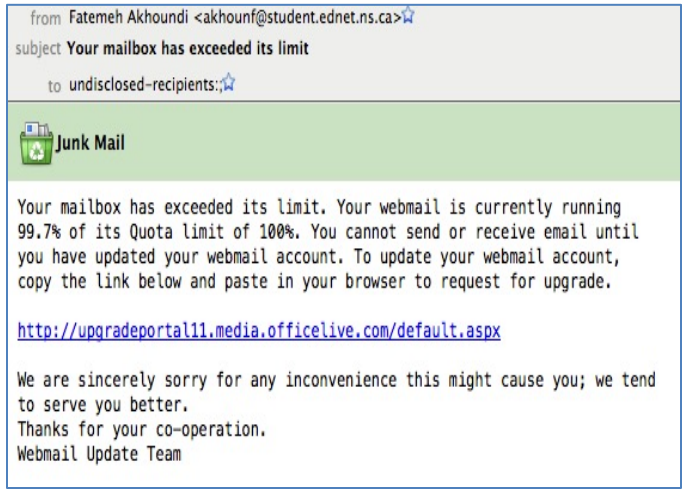
Intuition: spam and ham have different distribution of keywords

$$\Pr[\text{spam} \mid x_1, x_2, \ldots, x_n] = \frac{\Pr[\text{spam}] \prod \Pr[x_i \mid \text{spam}]}{\Pr[x_1, x_2, \ldots, x_n]}$$

$$\Pr[\text{ham} \mid x_1, x_2, \ldots, x_n] = \frac{\Pr[\text{ham}] \prod \Pr[x_i \mid \text{ham}]}{\Pr[x_1, x_2, \ldots, x_n]}$$

Classify as spam if: $\Pr[\text{spam} \mid x_1, x_2, \ldots, x_n] > \Pr[\text{ham} \mid x_1, x_2, \ldots, x_n]$

Multinomial Naïve Bayes Classifier


from Fatemeh Akhoundi <akhounf@student.ednet.ns.ca>
subject **Your mailbox has exceeded its limit**
to undisclosed-recipients:;

Junk Mail

Your mailbox has exceeded its limit. Your webmail is currently running 99.7% of its Quota limit of 100%. You cannot send or receive email until you have updated your webmail account. To update your webmail account, copy the link below and paste in your browser to request for upgrade.

http://upgradeportal11.media.officelive.com/default.aspx

We are sincerely sorry for any inconvenience this might cause you; we tend to serve you better.
Thanks for your co-operation.
Webmail Update Team

Represent email as "bag of words"

quota	1	x_1
webmail	4	x_2
cornell	0	x_3
fee	1	x_4
⋮	⋮	

Intuition: spam and ham have different distribution of keywords

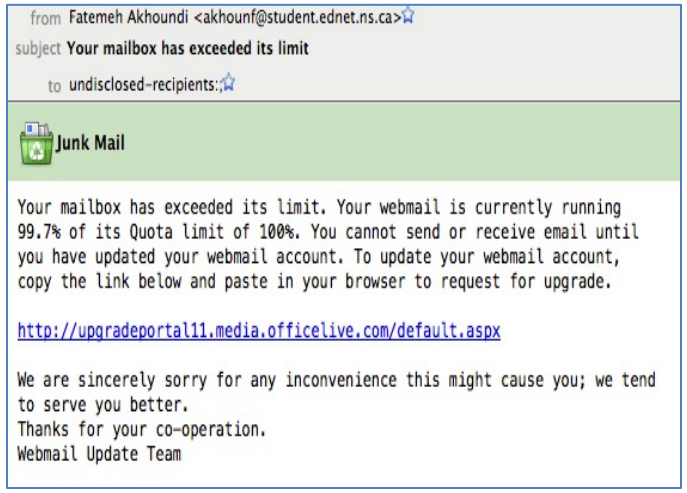
$$\Pr[\text{spam} \mid x_1, x_2, \ldots, x_n] = \frac{\Pr[\text{spam}] \prod \Pr[x_i \mid \text{spam}]}{\Pr[x_1, x_2, \ldots, x_n]}$$

$$\Pr[\text{ham} \mid x_1, x_2, \ldots, x_n] = \frac{\Pr[\text{ham}] \prod \Pr[x_i \mid \text{ham}]}{\Pr[x_1, x_2, \ldots, x_n]}$$

Estimate these from labeled training data.
Parameters θ are these probabilities, and
$p_{spam} = \Pr[spam]$
$p_{ham} = \Pr[ham]$

Classify as spam if: $\Pr[\text{spam}] \prod \Pr[x_i \mid \text{spam}] > \Pr[\text{ham}] \prod \Pr[x_i \mid \text{ham}]$

Multinomial Naïve Bayes Classifier


from Fatemeh Akhoundi <akhounf@student.ednet.ns.ca>
subject Your mailbox has exceeded its limit
to undisclosed-recipients:;

Junk Mail

Your mailbox has exceeded its limit. Your webmail is currently running 99.7% of its Quota limit of 100%. You cannot send or receive email until you have updated your webmail account. To update your webmail account, copy the link below and paste in your browser to request for upgrade.

http://upgradeportal11.media.officelive.com/default.aspx

We are sincerely sorry for any inconvenience this might cause you; we tend to serve you better.
Thanks for your co-operation.
Webmail Update Team

Represent email as "bag of words"

quota	1	x_1
webmail	4	x_2
cornell	0	x_3
fee	1	x_4
⋮	⋮	

Intuition: spam and ham have different distribution of keywords

$$\Pr[\text{spam} \mid x_1, x_2, \ldots, x_n] = \frac{\Pr[\text{spam}] \prod \Pr[x_i \mid \text{spam}]}{\Pr[x_1, x_2, \ldots, x_n]}$$

$$\Pr[\text{ham} \mid x_1, x_2, \ldots, x_n] = \frac{\Pr[\text{ham}] \prod \Pr[x_i \mid \text{ham}]}{\Pr[x_1, x_2, \ldots, x_n]}$$

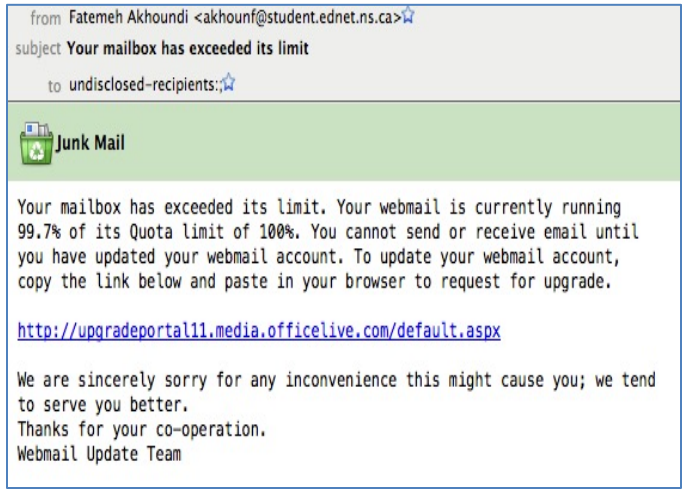
To estimate, fix event model, e.g. multinomial:

$\Pr[x_i \mid \text{spam}] = p_i^{x_i}$

where p_i is frequency of word i seen in spam messages

Classify as spam if: $\Pr[\text{spam}] \prod \Pr[x_i \mid \text{spam}] > \Pr[\text{ham}] \prod \Pr[x_i \mid \text{ham}]$

Multinomial Naïve Bayes Classifier


from Fatemeh Akhoundi <akhounf@student.ednet.ns.ca>
subject **Your mailbox has exceeded its limit**
to undisclosed-recipients:;

Junk Mail

Your mailbox has exceeded its limit. Your webmail is currently running 99.7% of its Quota limit of 100%. You cannot send or receive email until you have updated your webmail account. To update your webmail account, copy the link below and paste in your browser to request for upgrade.

http://upgradeportal11.media.officelive.com/default.aspx

We are sincerely sorry for any inconvenience this might cause you; we tend to serve you better.
Thanks for your co-operation.
Webmail Update Team

Represent email as "bag of words"

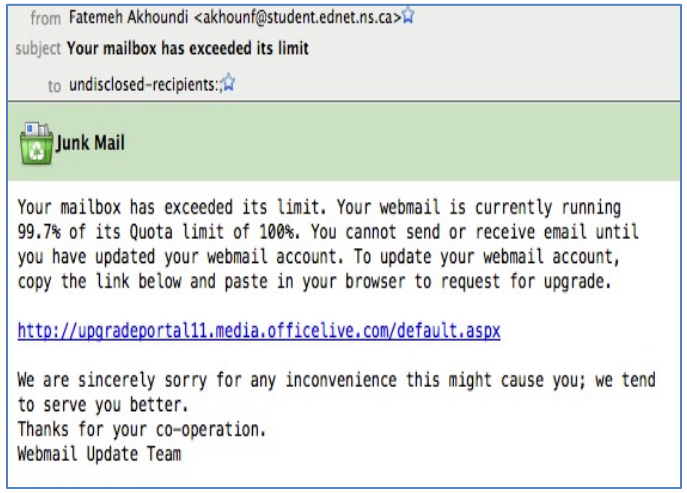
quota	1	x_1
webmail	4	x_2
cornell	0	x_3
fee	1	x_4
⋮	⋮	

Intuition: spam and ham have different distribution of keywords

Classify as spam if: $\Pr[\text{spam}] \prod p_i^{x_i} > \Pr[\text{ham}] \prod q_i^{x_i}$

$$\log\left(\Pr[\text{spam}] \prod p_i^{x_i}\right) = \log(\Pr[\text{spam}]) + \sum x_i \log p_i$$

Multinomial Naïve Bayes Classifier

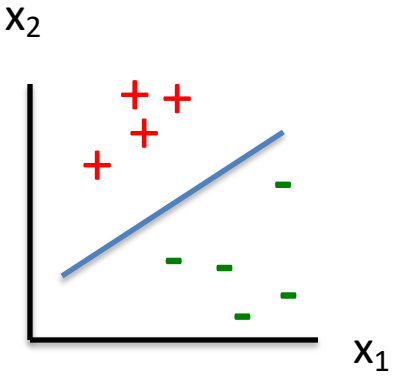


Represent email as "bag of words"

quota	1	x_1
webmail	4	x_2
cornell	0	x_3
fee	1	x_4
⋮	⋮	

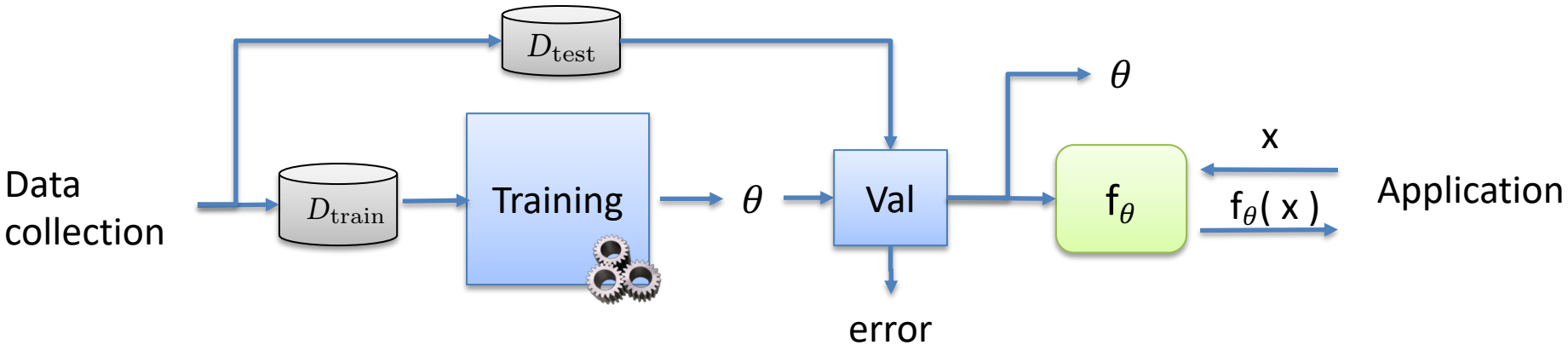
Intuition: spam and ham have different distribution of keywords

$$f_\theta(x) = \underset{c \in \{spam, ham\}}{\operatorname{argmax}} \left(\log p_c + \sum_{i=1}^{n} x_i \cdot \log p_i \right)$$



MNB is a linear classifier

Supervised machine learning (ML) pipeline



What are ML security risks that arise in such pipelines?

ML security issues

Poisoning attacks:

Insert bogus examples into training data

Trojan/backdoor attacks:

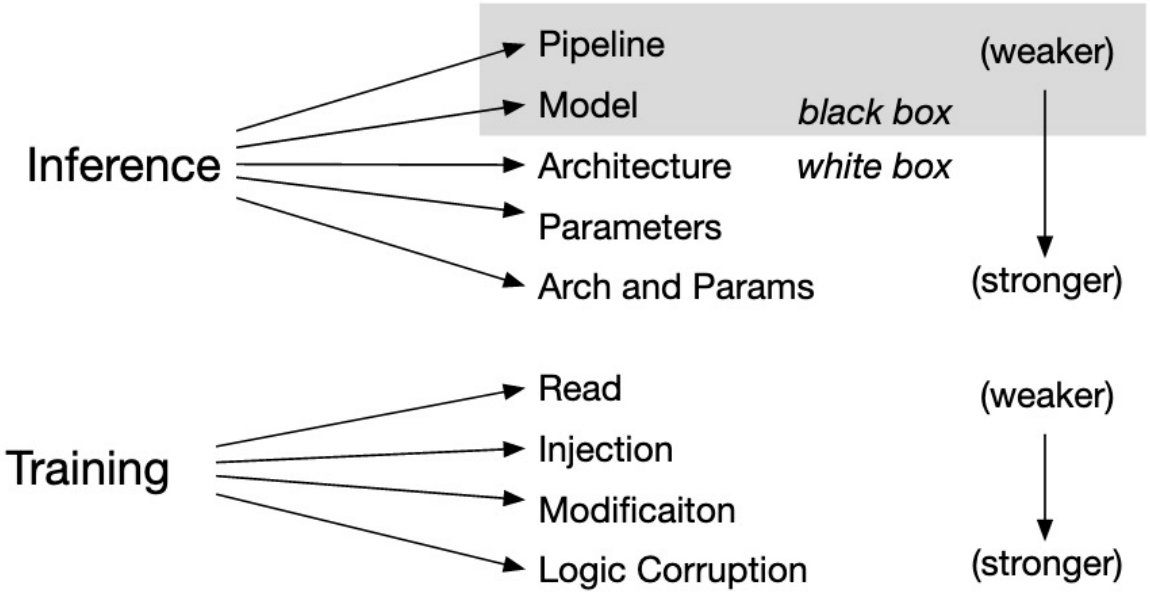
Insert backdoors into ML models

Evasion attacks:

a.k.a. “adversarial examples”

Mimicry attacks:

Used in intrusion detection system context to avoid detection



[Papernot et al. 2018] SoK paper

(We’ll discuss privacy issues in later lecture)

Science of security (??)

Paper never clarifies what they mean by science

"Science is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the world."

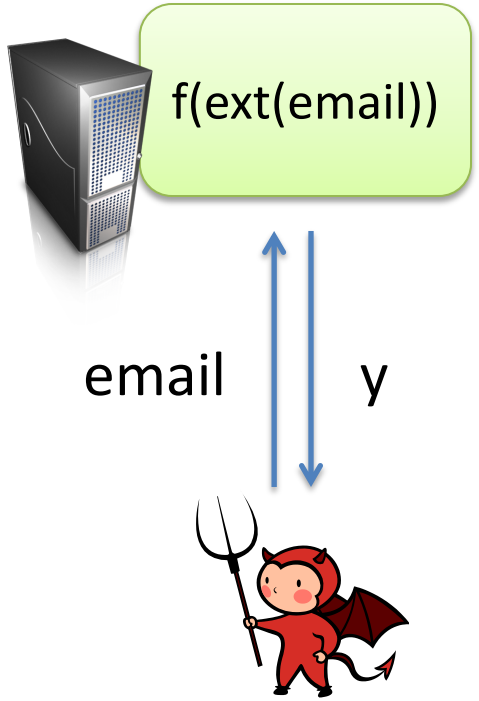
https://en.wikipedia.org/wiki/Science

Lowd, Meek 2005

- Investigate evasion attacks in spam setting
  - ACRE = adversarial classifier reverse engineering
  - This refers to model evasion in membership query setting
- First paper that looks at ***model extraction*** as a step towards evasion
  - Give algorithm for continuous features and linear classifiers
  - Show ACRE

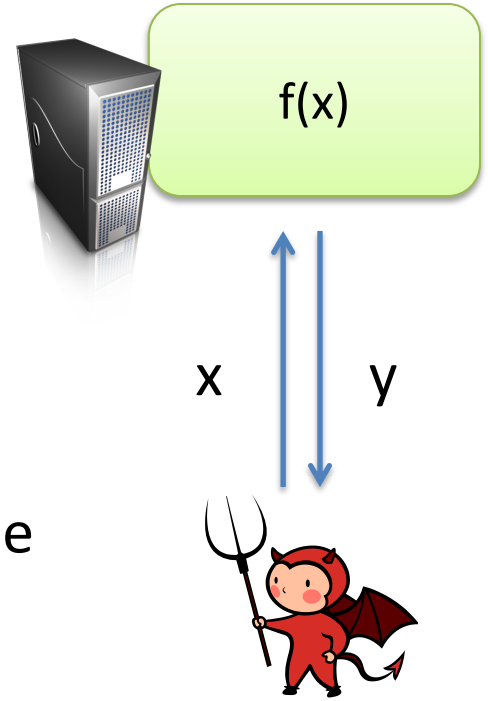
LM model extraction attack

Assume one knows features, or direct access to feature space ***membership queries.***
Membership queries just reveal class label

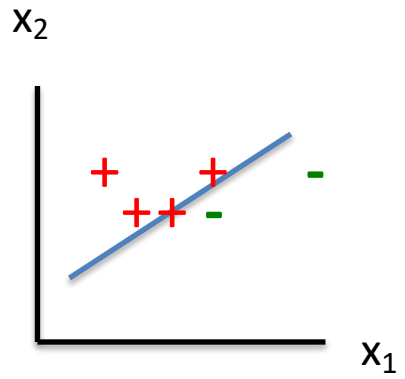


LM model extraction attack

Assume one knows features, or direct access to feature space membership queries



Assume model f is a linear function and it returns the sign of its output



Find points arbitrarily close to line, solve for line

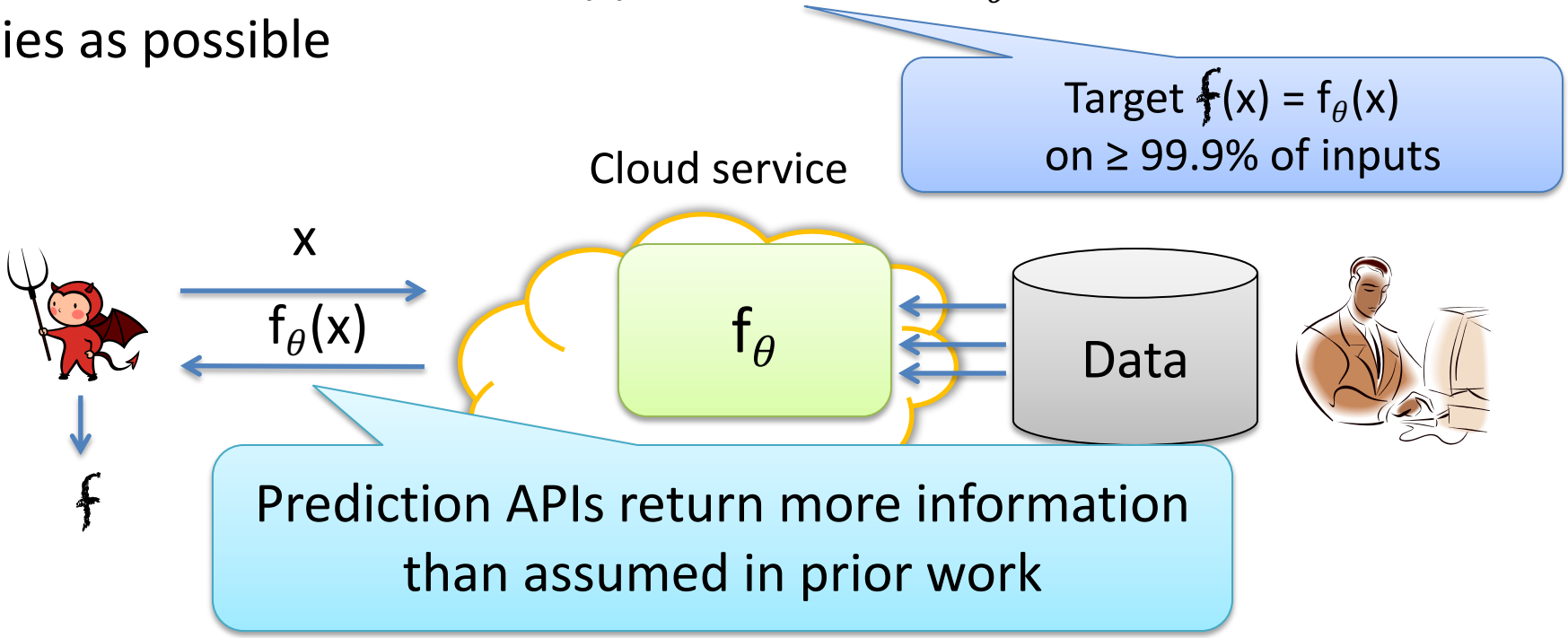
They describe it differently, but same idea

Point out that on boolean inputs, extraction is NP-hard. Show ACRE approximates well

Model extraction w/ confidences

[Tromer et al. 2016]

Adversarial client seeks to learn close approximation of f_θ in as few queries as possible



If $f_\theta(x)$ just class label: learning with membership queries setting

- [Kushilevitz, Mansour – 1993] for boolean decision trees
  Polytime but not practical
- [Lowd, Meek – 2005] linear models (e.g., binary logistic regression)
  See paper for generalizations and experimental results

Example: logistic regression

[Tromer et al. 2016]

Facial recognition of two people, Alice and Bob (the classes)

x[1] , Alice x[2] , Alice x[3] , Bob x[4] , Bob …

Feature vectors are pixel data
e.g.: n = 92 * 112 = 10,304

n+1 parameters θ = w,b chosen using training set to minimize expected error

$$f_\theta(x) = 1 / (1+e^{-(w*x + b)})$$

f_θ maps features to predicted probability of being "Alice"
≤ 0.5 classify as "Bob"
> 0.5 classify as "Alice"

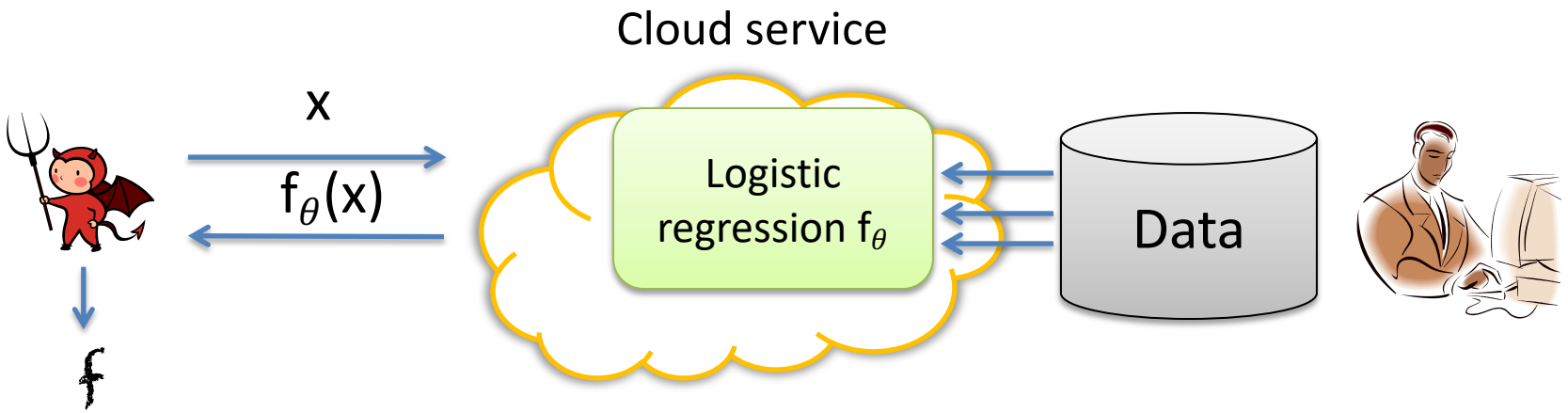
Generalize to c > 2 classes with *multinomial logistic regression*

$f_\theta(x) = [p_1, p_2, \ldots, p_c]$ predict label as $\text{argmax}_i\ p_i$

Model extraction attacks

[Tromer et al. 2016]

Adversarial client seeks to learn close approximation of f_θ in as few queries as possible



$$f_\theta(x) = 1 / (1+e^{-(w*x + b)})$$

$$\ln\left(\frac{f_\theta(x)}{1 - f_\theta(x)}\right) = w*x + b$$

Linear equation in n+1 unknowns w,b

Query n+1 random points → solve linear system of n+1 equations

~100x fewer queries than [Lowd, Meek 2005]

Model extraction attacks

[Tromer et al. 2016]

Adversarial client seeks to learn close approximation of f_θ in as few queries as possible

Model type	Attack approach
Binary logistic regression	Solve linear equations
Multinomial logistic regression	Solve non-linear equations
Neural network	Solve non-linear equations
Decision trees	Path-finding using pseudo-identifiers for leaves + partial feature vector queries

Tests with cloud services:

Amazon (multinomial LR)

BigML (decision trees)

100s to 1000s of queries

Seconds to minutes

100% accuracy

$\hat{f}(x) = f_\theta(x)$ on all x

Model extraction via active learning

[Tromer et al. 2016]

1. Query a set of random points
2. Train local model f on them
3. Choose new points to query close to f's decision boundary
4. Query new points, repeat steps 2-4 for set # of iterations

Model extraction summary [Papernot et al. 2020]

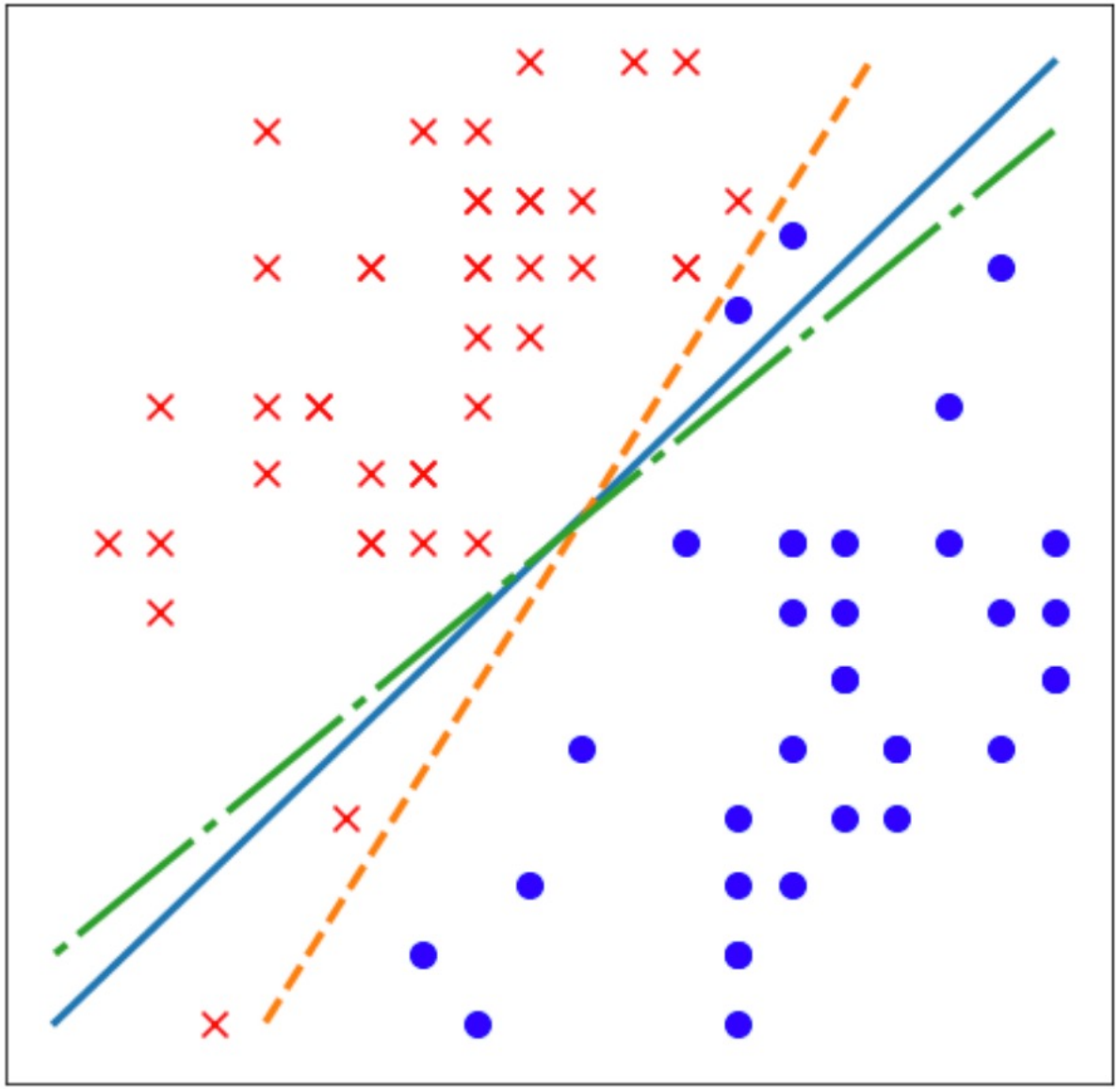
Attack	Type	Model type	Goal	Query Output
Lowd & Meek [8]	Direct Recovery	LM	Functionally Equivalent	Labels
Tramer *et al.* [11]	(Active) Learning	LM, NN	Task Accuracy, Fidelity	Probabilities, labels
Tramer *et al.* [11]	Path finding	DT	Functionally Equivalent	Probabilities, labels
Milli *et al.* [19] (theoretical)	Direct Recovery	NN (2 layer)	Functionally Equivalent	Gradients, logits
Milli *et al.* [19]	Learning	LM, NN	Task Accuracy	Gradients
Pal *et al.* [15]	Active learning	NN	Fidelity	Probabilities, labels
Chandrasekharan *et al.* [13]	Active learning	LM	Functionally Equivalent	Labels
Copycat CNN [16]	Learning	CNN	Task Accuracy, Fidelity	Labels
Papernot *et al.* [7]	Active learning	NN	Fidelity	Labels
CSI NN [25]	Direct Recovery	NN	Functionally Equivalent	Power Side Channel
Knockoff Nets [12]	Learning	NN	Task Accuracy	Probabilities
Functionally equivalent (this work)	Direct Recovery	NN (2 layer)	Functionally Equivalent	Probabilities, logits
Efficient learning (this work)	Learning	NN	Task Accuracy, Fidelity	Probabilities

DT = decision tree, LM = linear model, NN = neural network, CNN = convolutional NN

Papernot et al. 2016: extraction goals

Draw distinctions between:

- **Functionally equivalent extraction**
  - For all x, matches target model
- **Fidelity extraction**
  - For x from distribution, matches target model
- **Accuracy extraction**
  - For x from distribution, performs well for original task



Blue line = model decision boundary
Green line = high fidelity extraction
Orange dotted = high accuracy extraction

Papernot et al. approaches

- Learning based for ***fidelity extraction***
  - Have ***unlabeled*** training data. Same as data used in target model
  - Fully-supervised: query oracle to get labels and train
    - Need less data than training in first place
  - Semi-supervised: query oracle on small number of points, combine with unlabeled data
- ***Functionality extraction*** for 2-layer ReLU neural networks

$$O_L(x) = A^{(1)}\mathrm{ReLU}(A^{(0)}x + B^{(0)}) + B^{(1)}$$

Model extraction and other attacks

- Extraction often seen as stepping stone to other attacks:
  - Evasion
  - Privacy attacks (membership inference, training data leakage, etc.)