ADVERSARIAL AI

Yevgeniy Vorobeychik

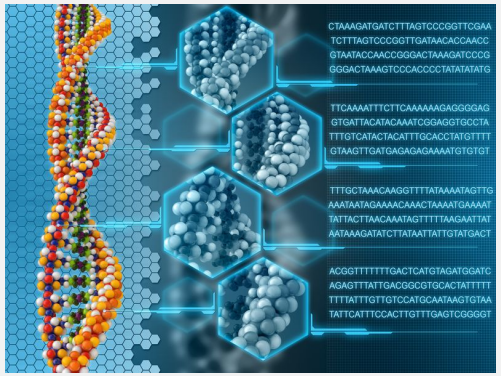
Computer Science

Vanderbilt University

AI IS TAKING OVER THE WORLD

Autonomous Driving

Cybersecurity

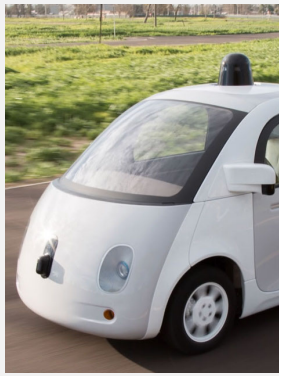


Bioinformatics

Physical security

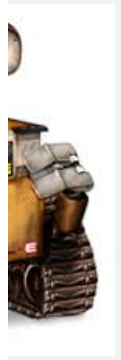
Healthcare

Robotics



Autonomous

Physical sec



s

SPEAKING OF SOCIAL BOTS

- Perform many useful functions, such as dissemination of information on social networks
- But can be exploited:
  - Useful social bots can be manipulated (e.g., to spread false rumors)
  - Malicious social bots (**sociobots**) can be used to spread misinformation, influence political campaigns, etc

AI IN ADVERSARIAL SETTINGS

AI IN ADVERSARIAL SETTINGS

AI applied in adversarial settings

AI IN ADVERSARIAL SETTINGS

AI applied in adversarial settings

AI used for malicious means

AI IN ADVERSARIAL SETTINGS

AI applied in adversarial settings

AI used for malicious means

EXAMPLE:
evasion attacks on machine learning

AI IN ADVERSARIAL SETTINGS

AI applied in adversarial settings

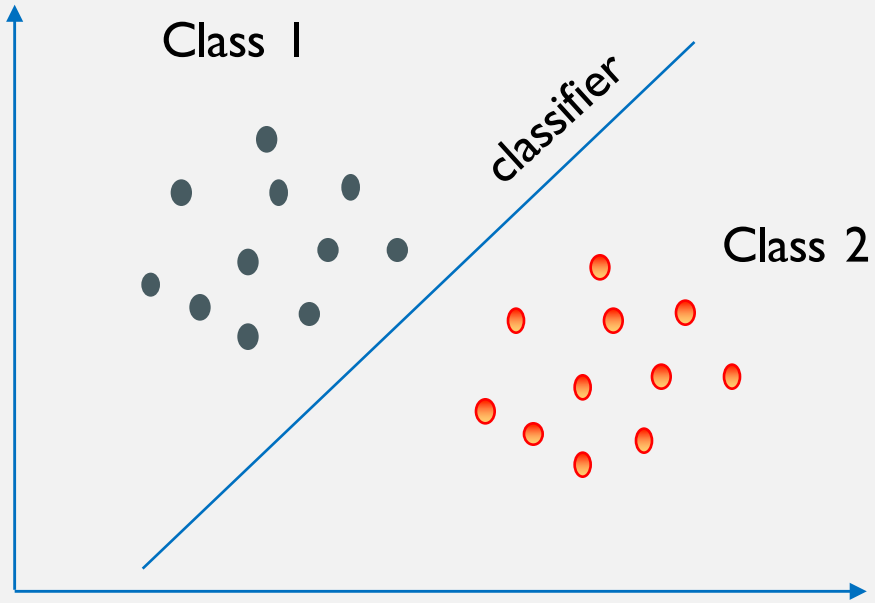
EXAMPLE:
evasion attacks on machine learning

AI used for malicious means

EXAMPLE:
automated re-identification of sanitized data

ADVERSARIAL CLASSIFIER EVASION

CLASSIFICATION LEARNING



Data set: $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, $(x,y) \sim \mathcal{D}$

x: feature vectors

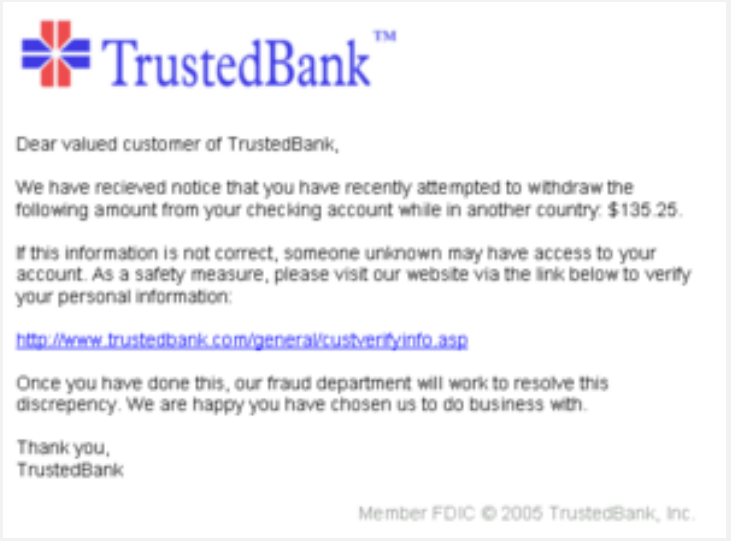
y: binary label in {-1, +1} representing which class x belongs to

Learning: train classifier f(x) on data

Prediction: use f(x) to predict label for arbitrary x

CLASSIFICATION IN ADVERSARIAL SETTINGS

- Often, classifiers are tasked with telling apart "good" from "bad"
  - Spam vs. non-spam (ham)
  - Benign vs. malicious software
  - Good bots vs. sociobots

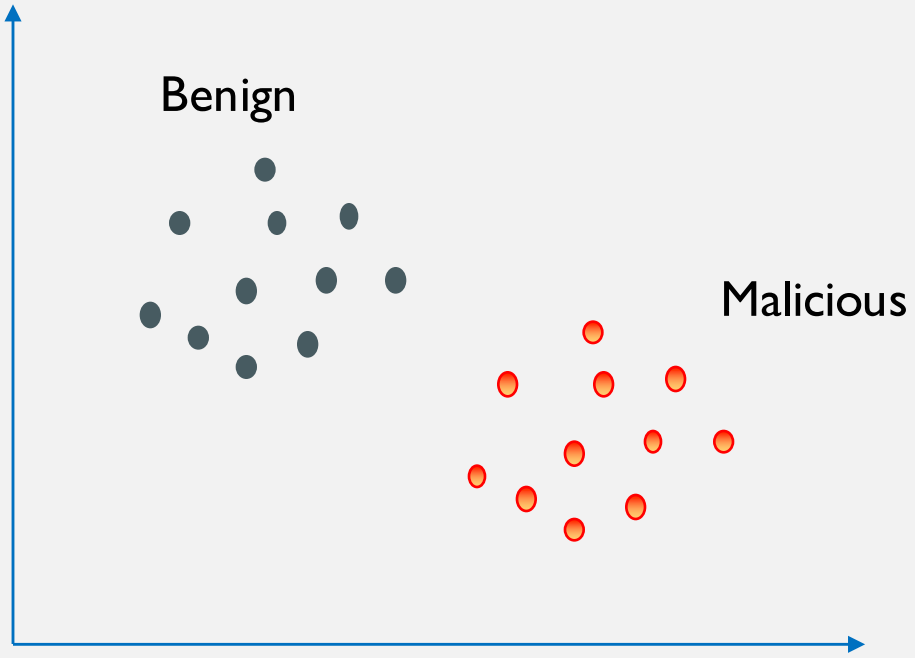


EVASION ATTACKS

- **Adversary who previously chose instance *x* (which is now classified as malicious) now chooses another instance *x'* which is classified as benign**

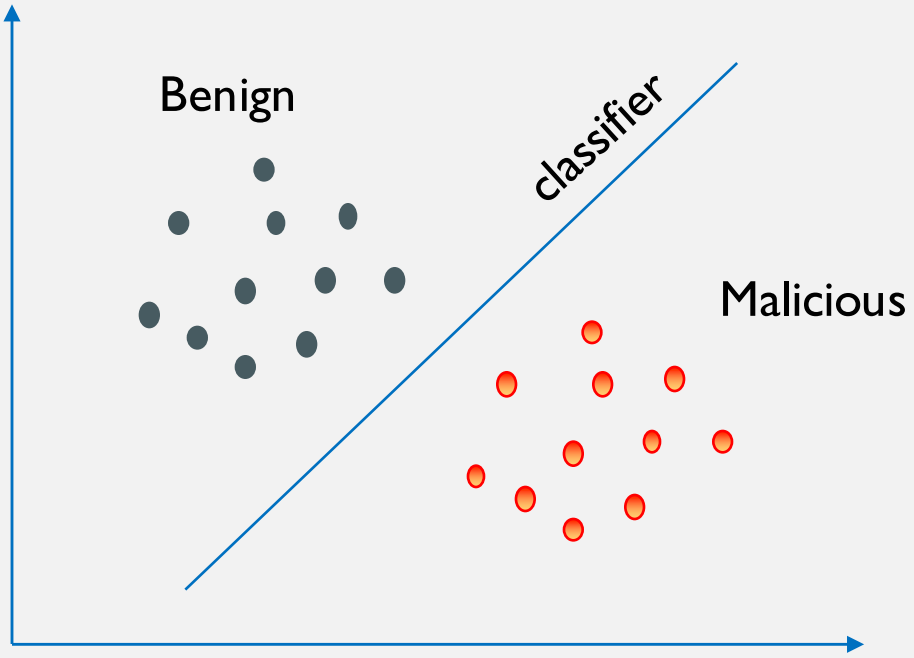
EVASION ATTACKS

- **Adversary who previously chose instance *x* (which is now classified as malicious) now chooses another instance *x'* which is classified as benign**



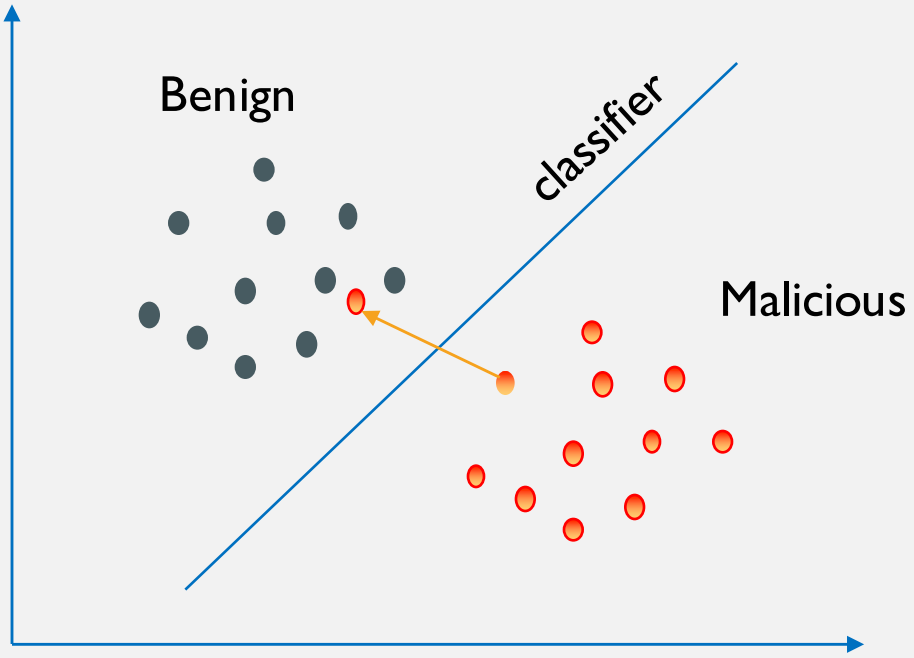
EVASION ATTACKS

- **Adversary who previously chose instance *x* (which is now classified as malicious) now chooses another instance *x'* which is classified as benign**



EVASION ATTACKS

- **Adversary who previously chose instance *x* (which is now classified as malicious) now chooses another instance *x'* which is classified as benign**



MODELING EVASION ATTACKS

- Attacker has an "ideal" feature vector x
  - These are the original malicious feature vectors in training data
- Modifying x into another feature vector x' incurs a cost $c(x,x')$
  - $c(x,x')$ can be measured using l_p norm, or consider feature substitution attacks
- There is a cost budget (maximum tolerance for cost incurred through evasion)
- Need to appear benign to the classifier
  - Problem: $\min_{x'} c(x,x')$ s.t.: $c(x,x') \leq$ cost budget, x' classified as benign

EVASION-ROBUST CLASSIFICATION

- Minimize total l_1 regularized (hinge) loss
- For each malicious instance, allow attacker to choose an alternative according to a "black-box" decision model, *A(x;w)*, parametrized by preferred attack *x*
- Minimize **adversarial loss**:

$$\min_w \sum_{j|y_j=-1} l(-w^T x_j) + \sum_{j|y_j=+1} l(w^T A(x_j; w)) + \lambda \|w\|_1$$

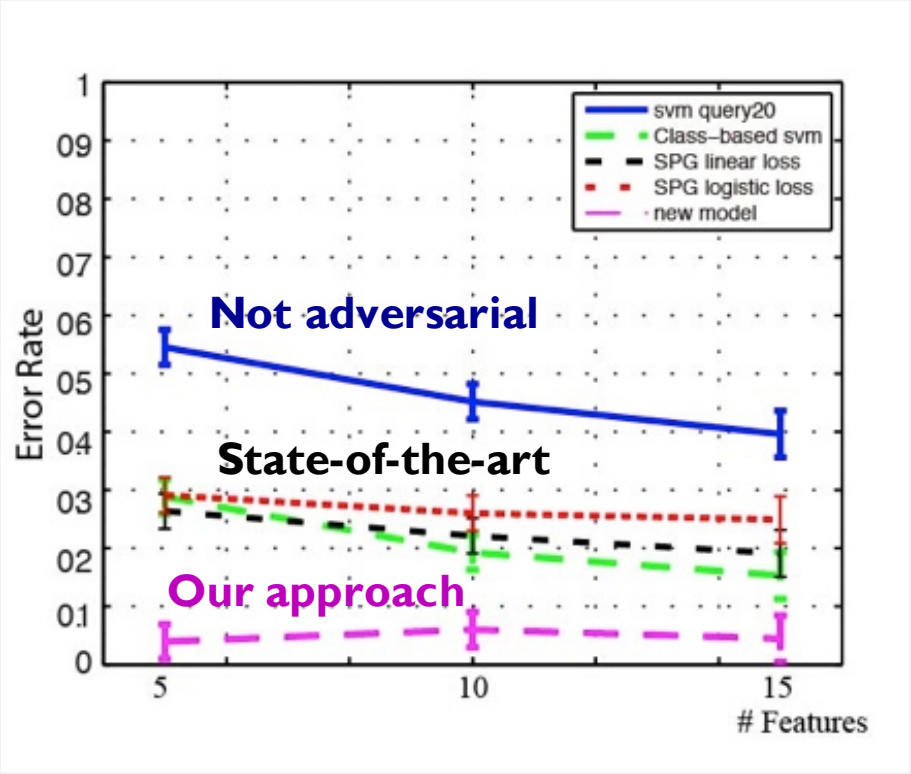
NIPS, 2014

EVASION-ROBUST CLASSIFICATION

- Can formulate the problem as a mixed integer linear program
- *In this formulation, we capture A(x;w) for a rational attacker using constraints*
  - Assume that the attacker aims to evade classifier and minimize distance from *x*, up to a cost budget (if too far from *x*, attacker is deterred)
- Scalability challenge: too many constraints
  - Each constraint corresponds to a malicious instance in the data and all possible alternative instances for the corresponding attacker
- Approach:
  - clustering attacks: cluster malicious feature vectors *in training data*
  - constraint generation: iteratively add "best response" attacks into master MILP

NIPS, 2014

EVASION-ROBUST CLASSIFICATION



NIPS, 2014

OTHER RESEARCH IN ADVERSARIAL CLASSIFIER EVASION

- Classifier reverse engineering is "easy in practice" (AAMAS 2014)
- Randomized decisions based on classifier prediction (AAMAS 2014, AISTATS 2015)
- Human subject experiment in adversarial evasion (AAAI 2016)

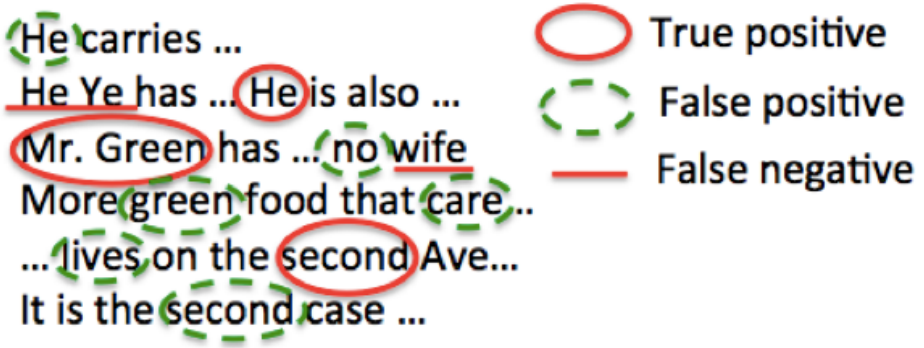
SANITIZING SENSITIVE DATA

SANITIZING DATA, AT SCALE

- Enormous amounts of data is available in unstructured form
  - Text data
  - Images
  - Video footage
- This data can be quite valuable for research and discovery
  - Medical records can reveal connections between symptoms and disease
- ***A lot of this data contains sensitive information***
  - Medical records (PII)
  - Images/Video (faces)

SANITIZING TEXT DATA USING MACHINE LEARNING

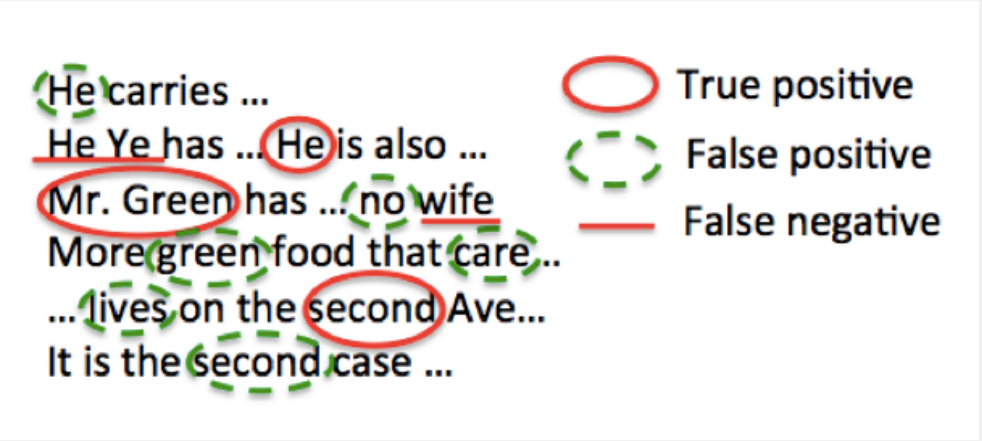
- Mark sensitive entities (names) as "+1"
- Mark non-sensitive entities (everything else) as "-1"
- Run classifier; remove predicted sensitive words
- *Szarvas et al. 2007; Uzuner et al., 2008; Gardner et al., 2009; Aberdeen et al., 2010; Ferrandez et al. 2013*



Essentially, a *named entity recognition* task

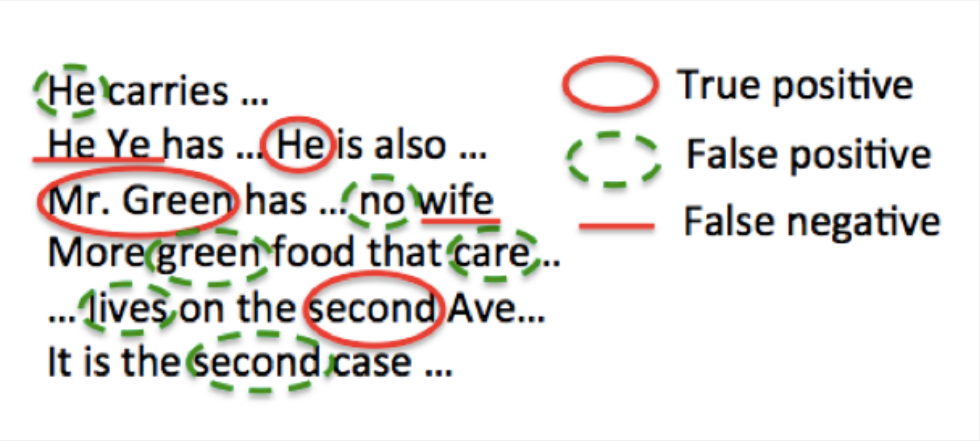
SANITIZING TEXT DATA USING MACHINE LEARNING

- But learning makes mistakes:
  - *False negatives (predicted negatives which are actual identifiers): under-redaction*
  - *False positives (predicted positives which are not sensitive): over-redaction*



SANITIZING TEXT DATA USING MACHINE LEARNING

- But learning makes mistakes:
  - ***False negatives (predicted negatives which are actual identifiers): under-redaction***
    - ***This is a major privacy concern***
  - *False positives (predicted positives which are not sensitive): over-redaction*



THREAT MODELING

- *Prior work: implicitly, threat model has been a human adversary who manually inspects instances*
- *Adversary can use machine learning* ***to find sensitive entities (e.g., PII, UCI, etc) at scale****!*

ICDM, 2015

THREAT MODELING

- Adversary has a manual inspection budget B
- Uses the following approach:
  - Learn a classifier h to predict sensitive entities in published data
  - Rank predicted positives before predicted negatives
  - Inspect top B instances
- Adversary has two super-powers:
  - **Perfect verification**: Any instance inspected is perfectly verified as sensitive or not
  - **Optimality**: Adversary will compute an ***optimal*** classifier h in terms of accuracy
    - In practice, adversary would label a subset of published data and learn a classifier on this "training" data

ICDM, 2015

DATA SANITIZING & PUBLISHING

- "Traditional" approach:
  - PUBLISHER:
    - Data publisher learns a classifier h
    - Removes all predicted positives
    - Releases data
  - ADVERSARY:
    - Acts according to threat model just described
    - (Learns a new classifier, h' on published data; sorts data; …)

ICDM, 2015

DATA SANITIZING & PUBLISHING

- "Traditional" approach:
  - PUBLISHER:
    - Data publisher learns a classifier h
    - Removes all predicted positives
    - Releases data
  - ADVERSARY:
    - Acts according to threat model just described
    - (Learns a new classifier, h' on published data; sorts data; …)
  - IDEA: **publisher can simulate the adversary**
    - If adversary can get a lot of value out of h', publisher should use h' and redact the corresponding predicted positives

ICDM, 2015

GREEDY ITERATIVE ALGORITHM

- GreedySanitize(*D* : input training data):

  Repeat:

    learn *h* on data *D*

    remove predicted positives (according to *h*) from *D*

    add *h* to *H*

  until (**stopping criterion**)

  Apply *H* to published data to remove the **union** of all predicted positives

ICDM, 2015

STOPPING CRITERION

- **Publisher utility**: when published data is released, the publisher loses L for each released identifier, and gains C for each non-sensitive entity (word)
  - Trading off value of releasing data and risk of identification
- Stop when applying a new classifier *h* to the data yields negative marginal utility (local optimum)

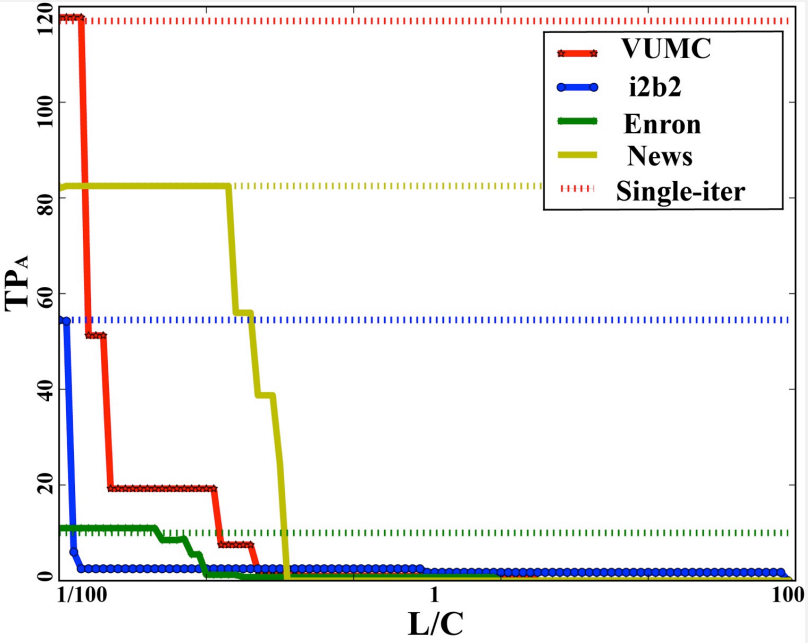
RESULTS

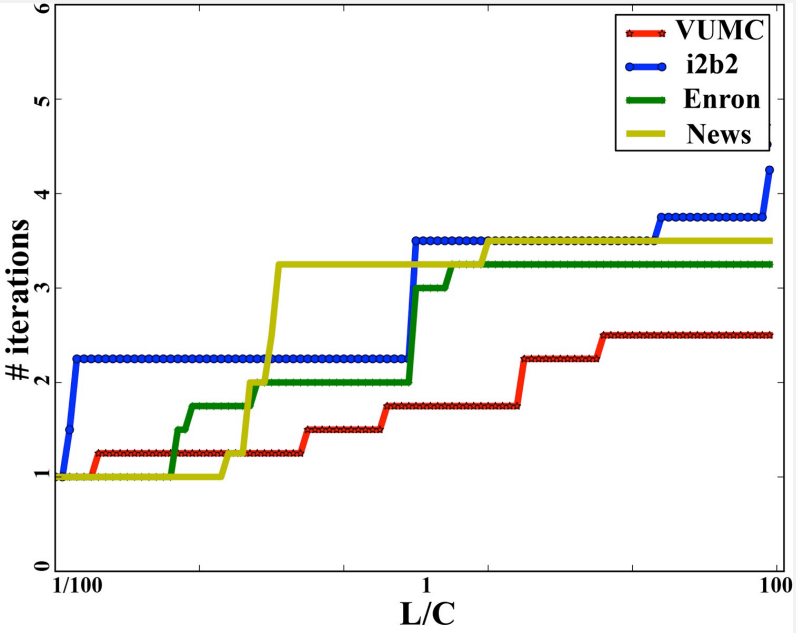
- **Theorem**: when GreedySanitize terminates (actually, *in any local optimum*),

$$TP_A \leq C/L\ (1-\alpha)n$$

  - α : fraction of identifiers in the data; TP = # of true positives for adversary
- Thus, when L >> C, the publisher will leave very few identifiers that can be found by the adversary *with a small budget*
- *With a large budget adversary can't do much better than randomly inspecting B instances*

RESULTS





GreedySanitize leaves virtually no "true positives" in the published data
(in contrast, single-iteration ML leaves a significant number of PII)

GreedySanitize number of greedy iterations is < 5 even when L >> C

ICDM, 2015

AI IN ADVERSARIAL SETTINGS

AI applied in adversarial settings

EXAMPLE:
evasion attacks on machine learning

AI used for malicious means

EXAMPLE:
automated re-identification of sanitized data

ADVERSARIAL AI @CERL

ADVERSARIAL AI @CERL

Optimal threshold selection in IDS

ADVERSARIAL AI @CERL

Choose an individualized threshold in IDS to balance FP and FN, accounting for adversarial response

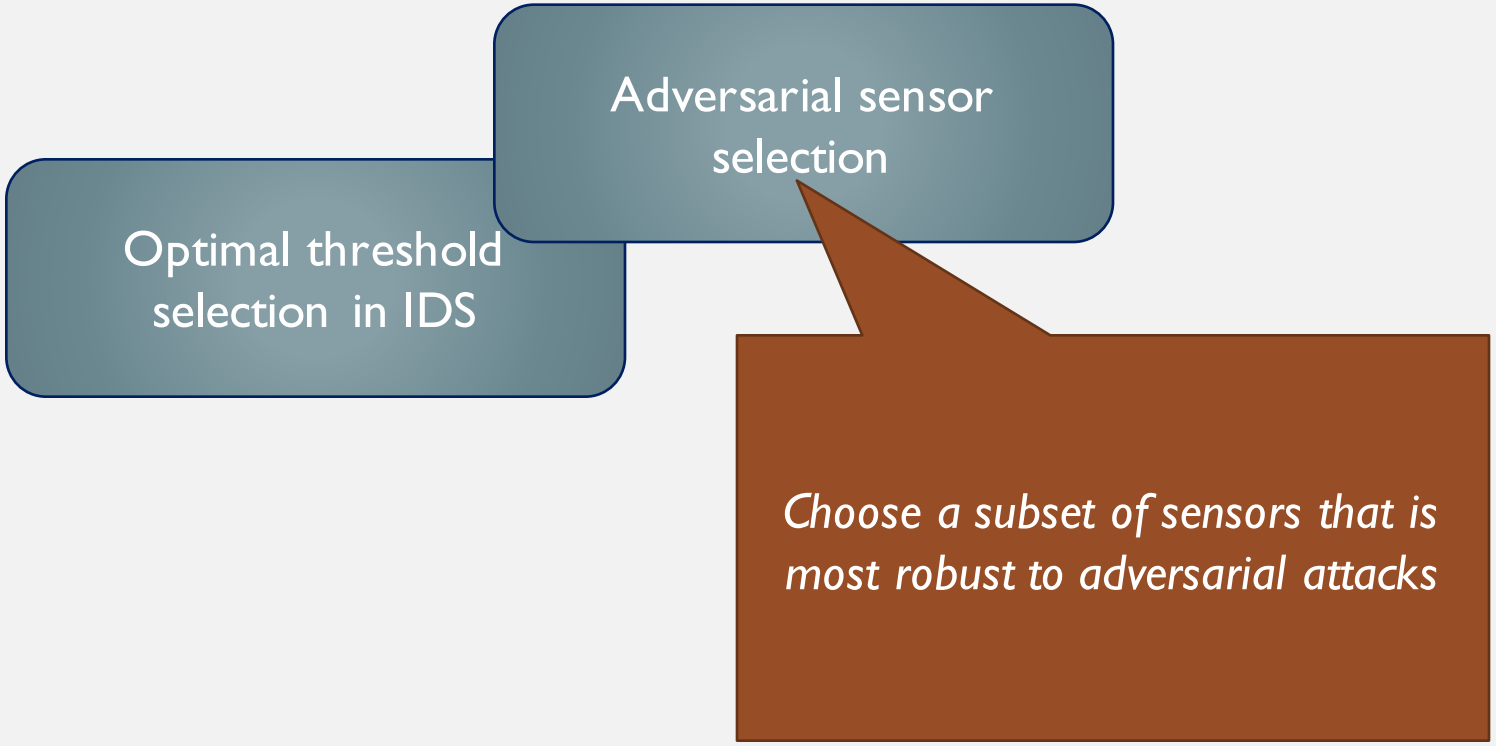
Optimal threshold selection in IDS

ADVERSARIAL AI @CERL

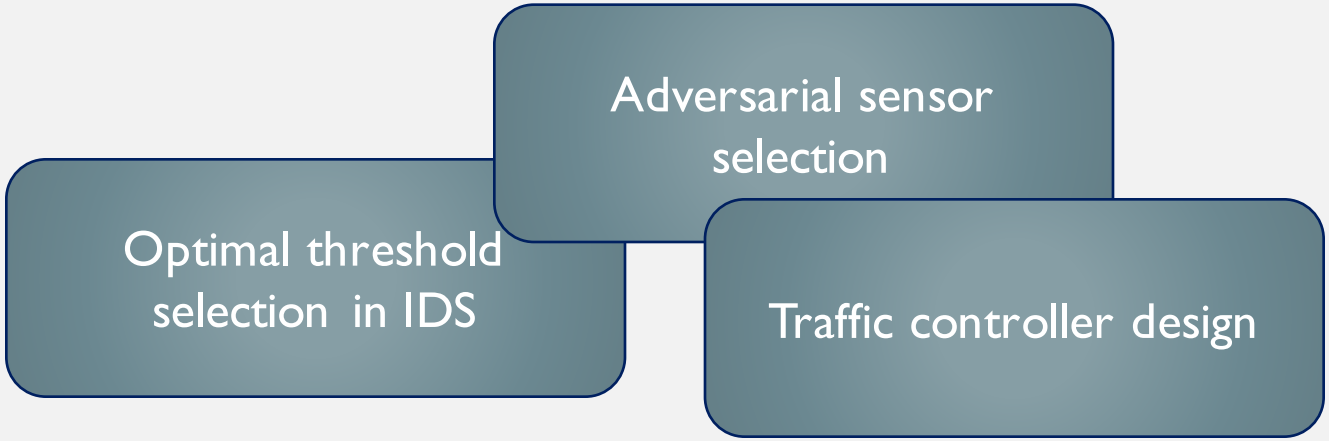
Adversarial sensor selection

Optimal threshold selection in IDS

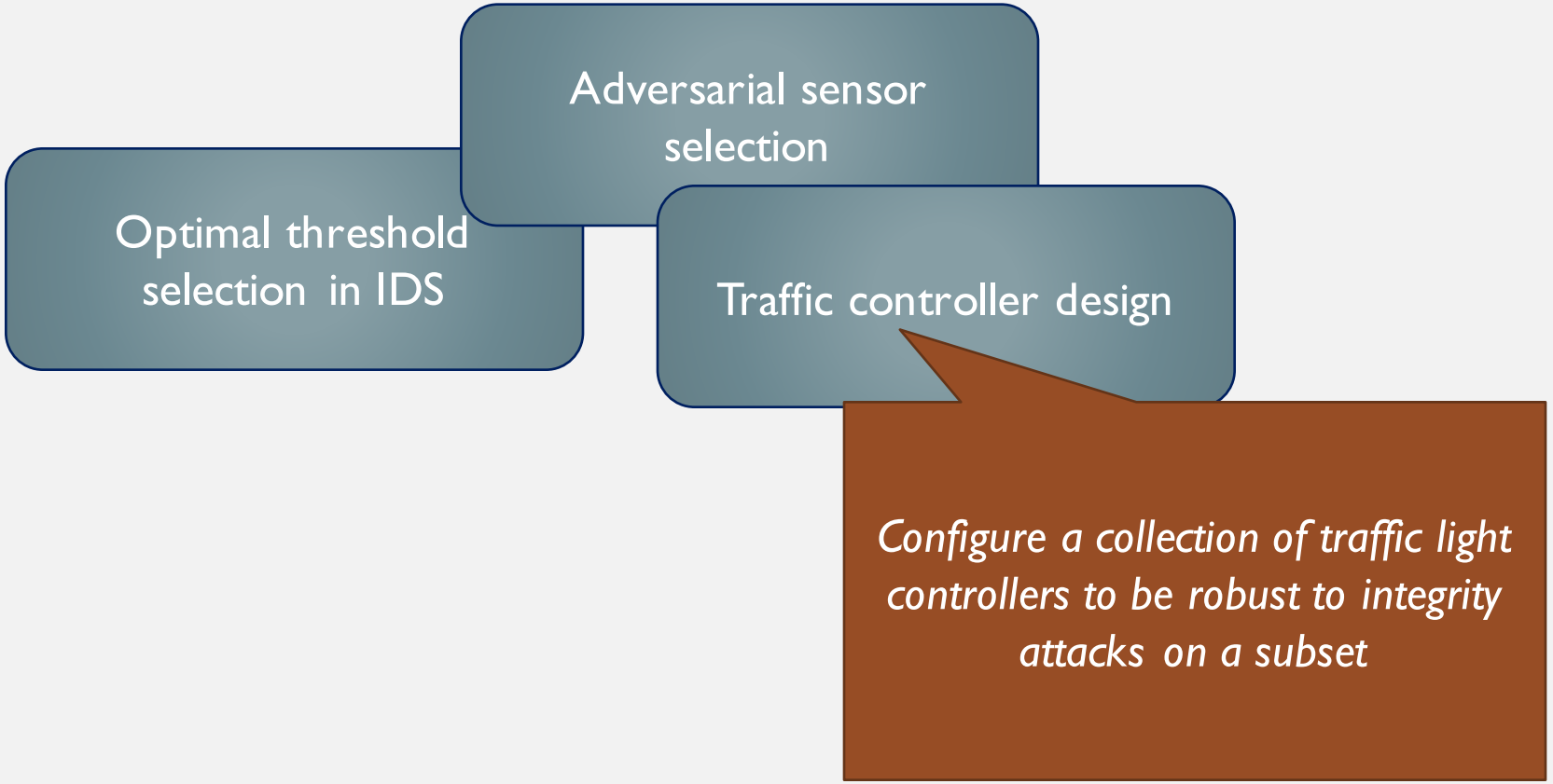
ADVERSARIAL AI @CERL



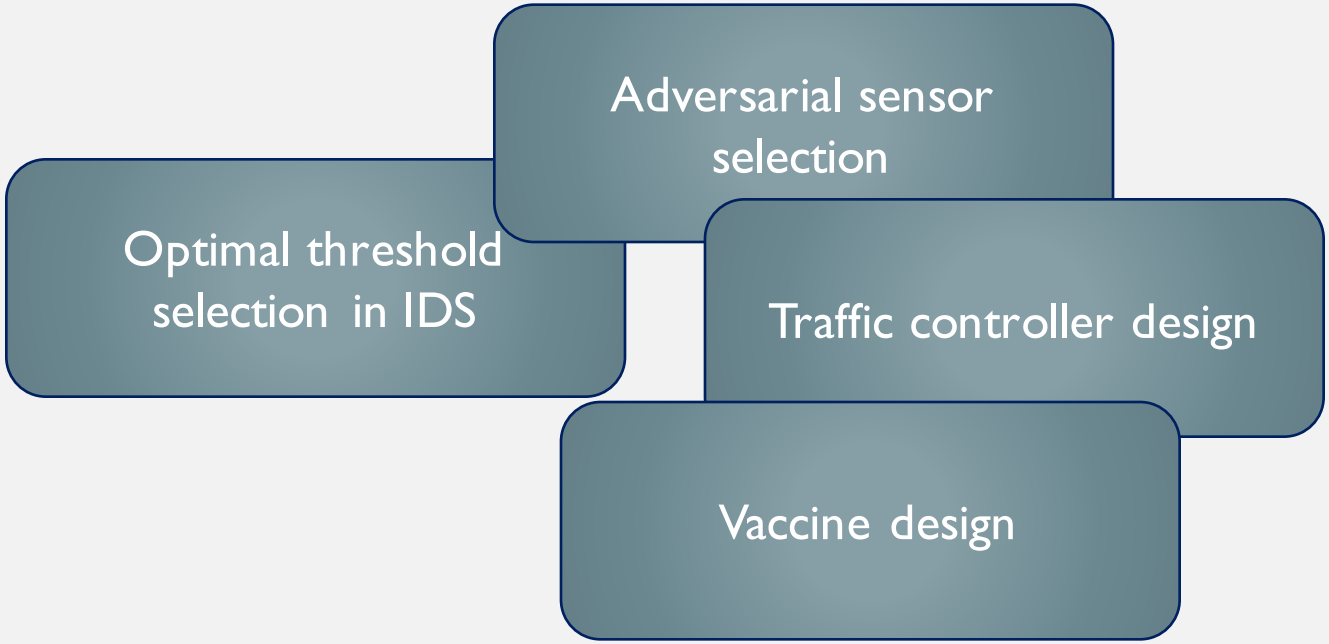
ADVERSARIAL AI @CERL



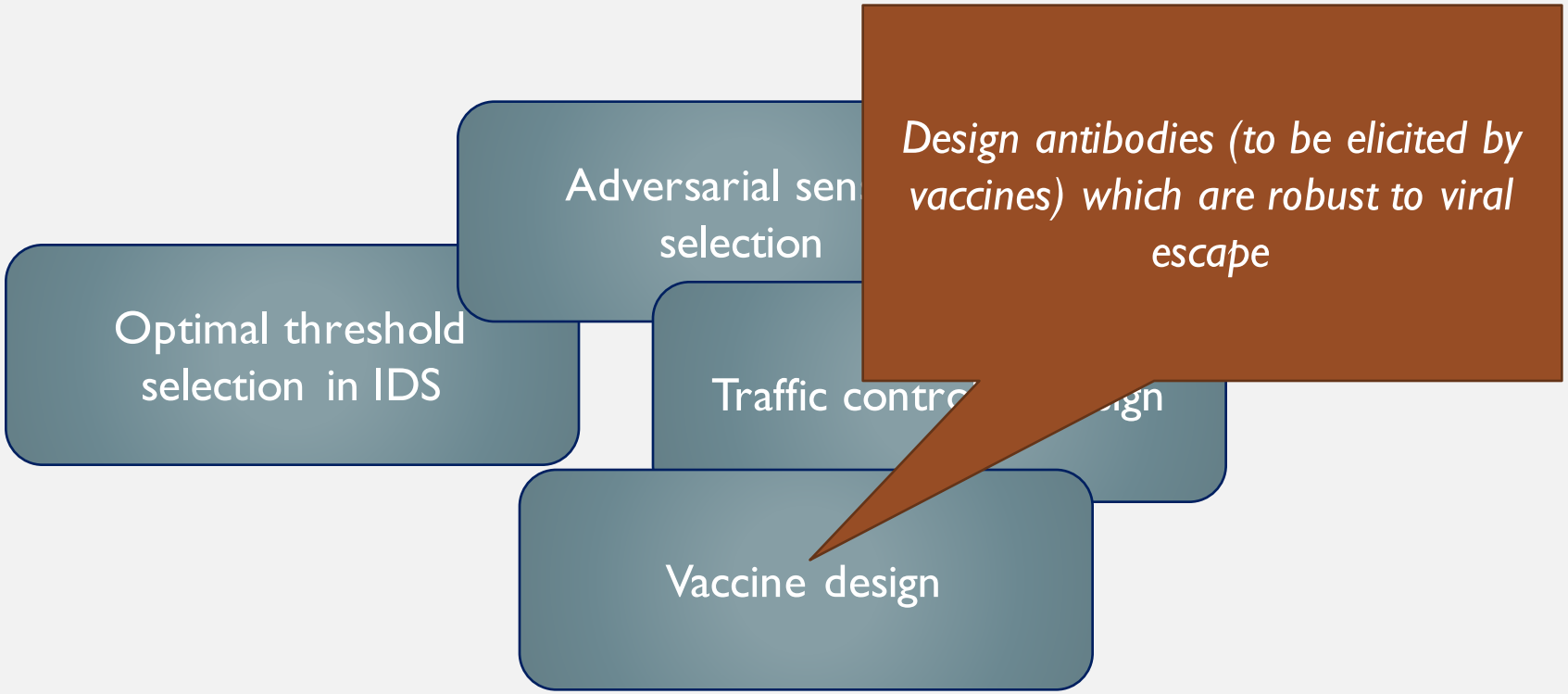
ADVERSARIAL AI @CERL



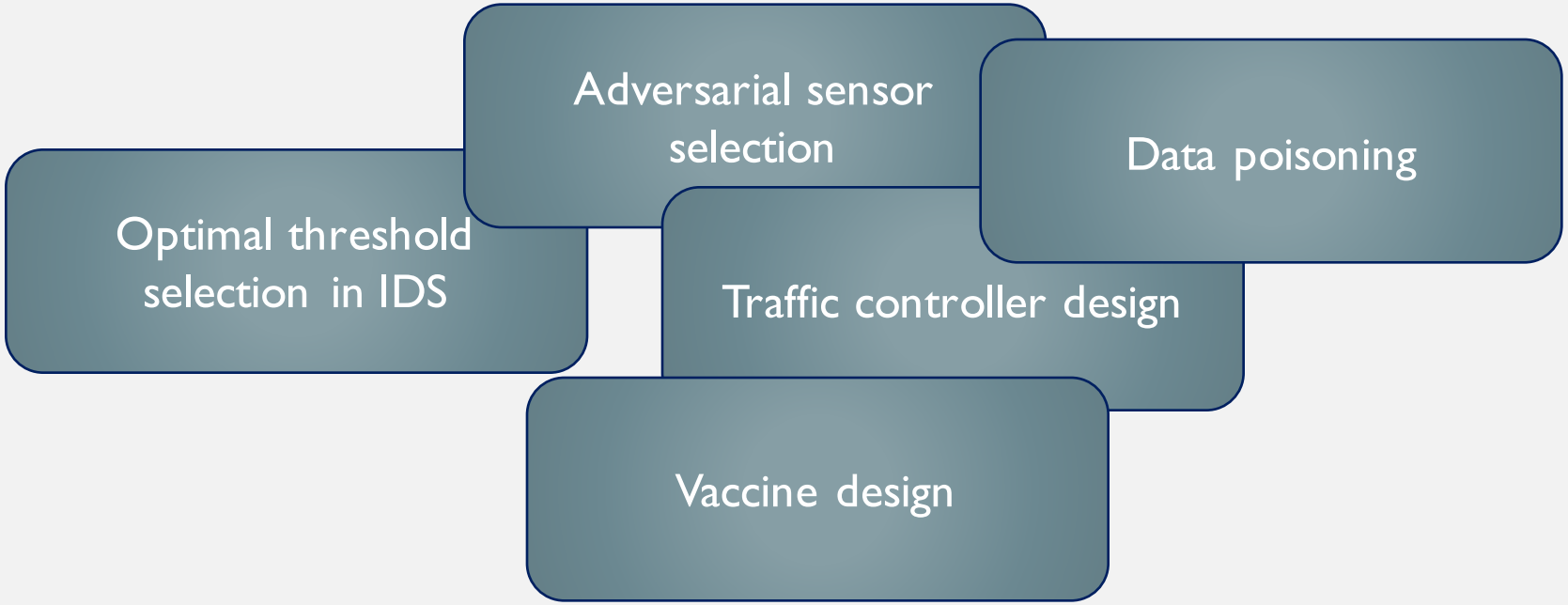
ADVERSARIAL AI @CERL



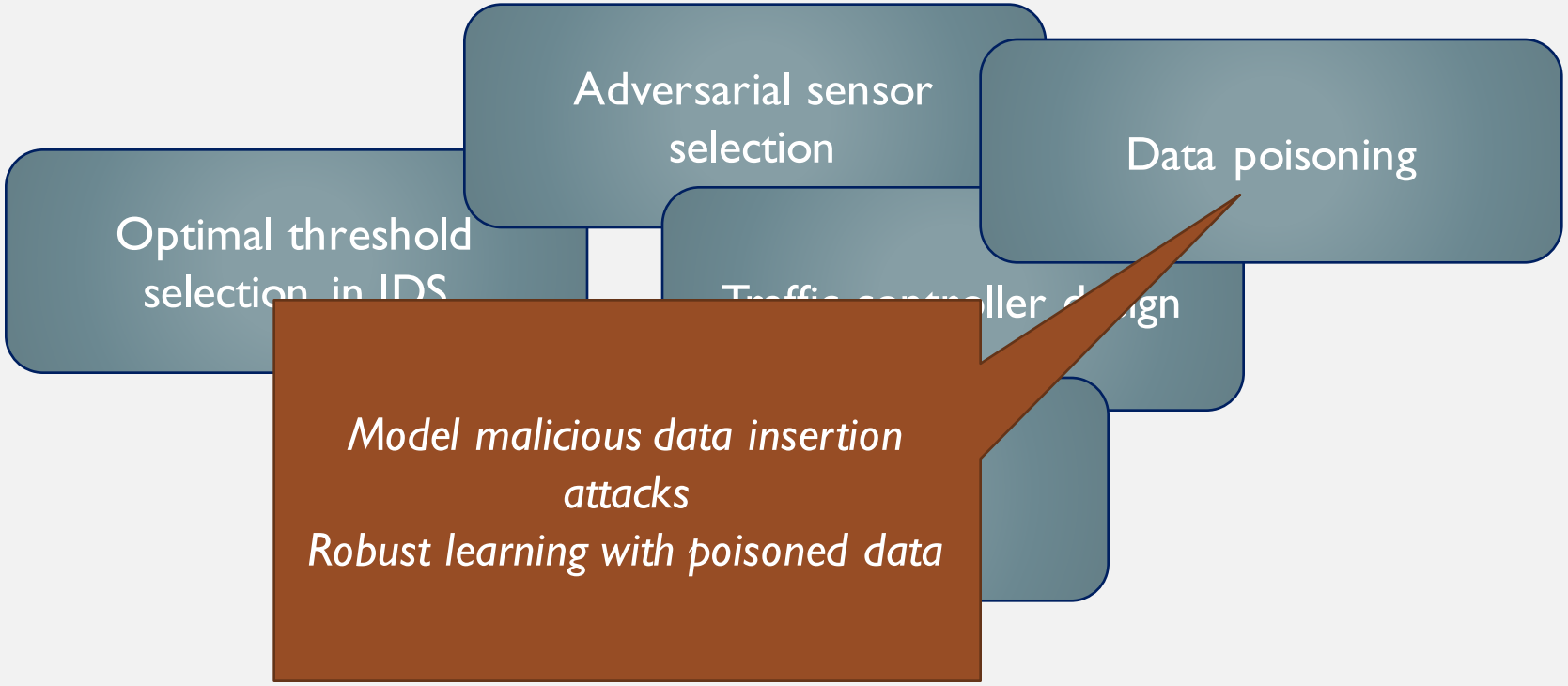
ADVERSARIAL AI @CERL



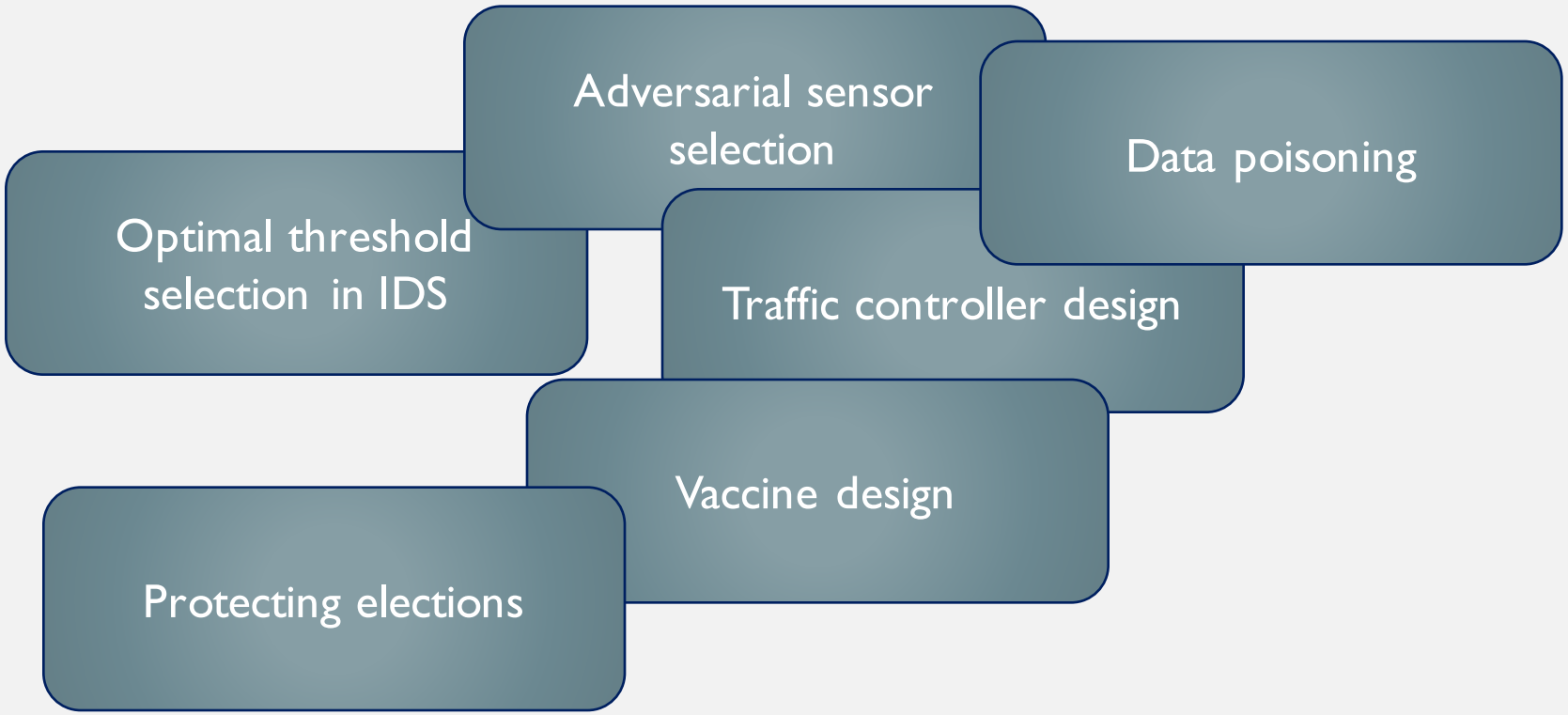
ADVERSARIAL AI @CERL



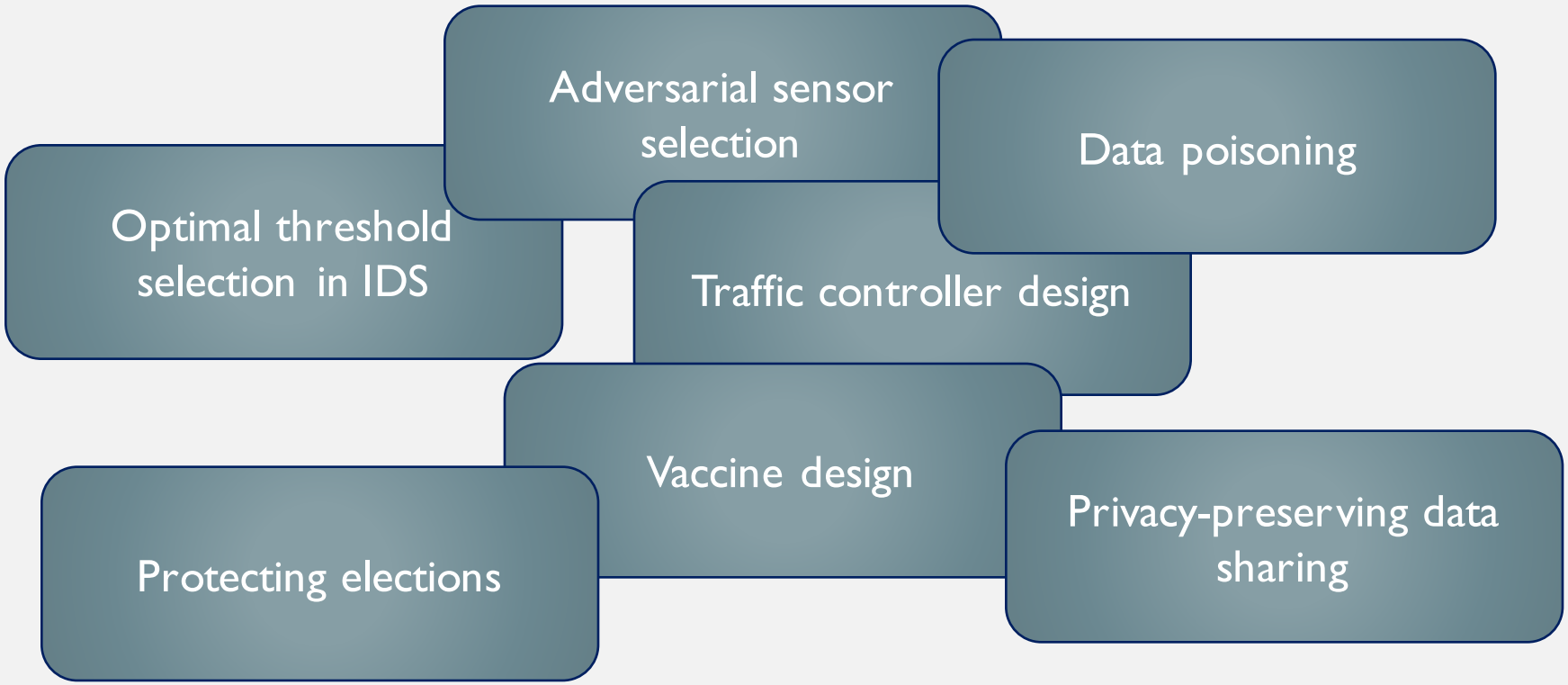
ADVERSARIAL AI @CERL



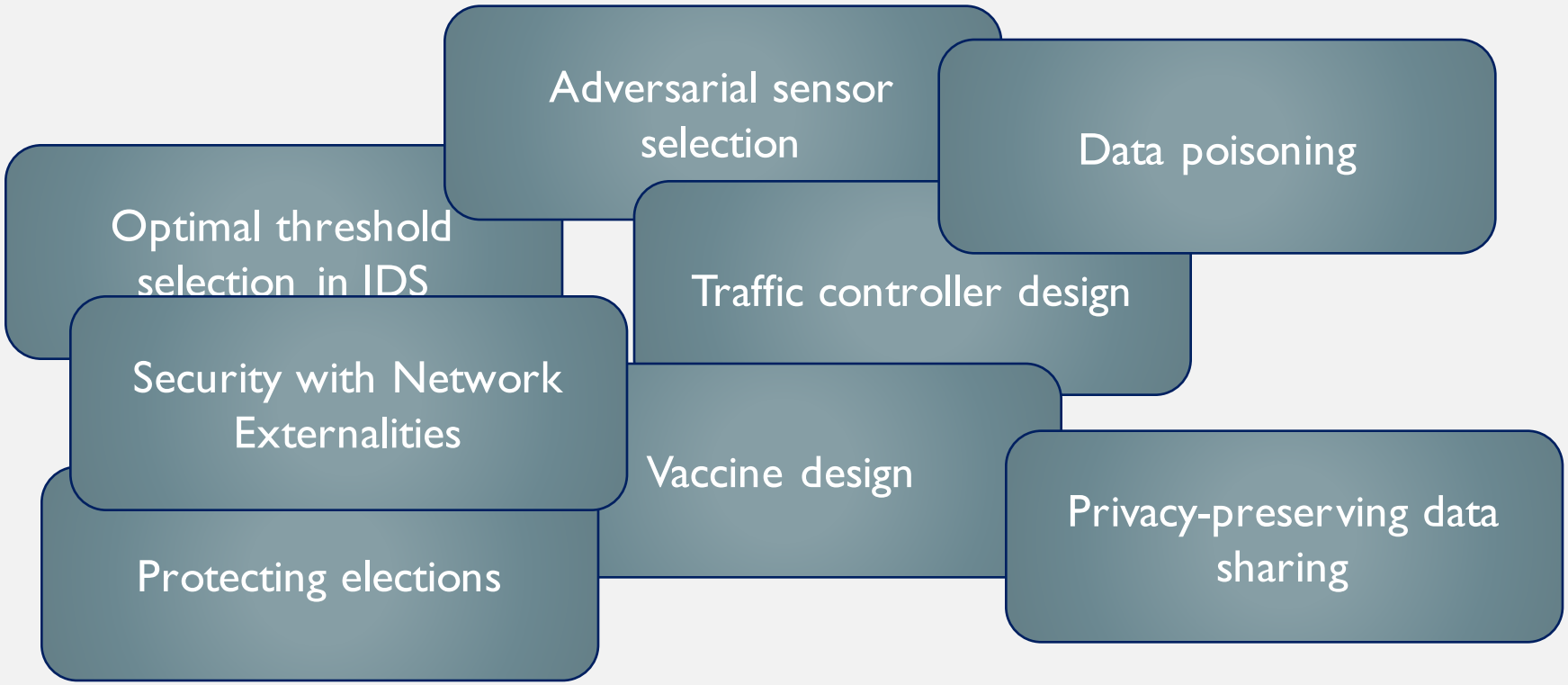
ADVERSARIAL AI @CERL



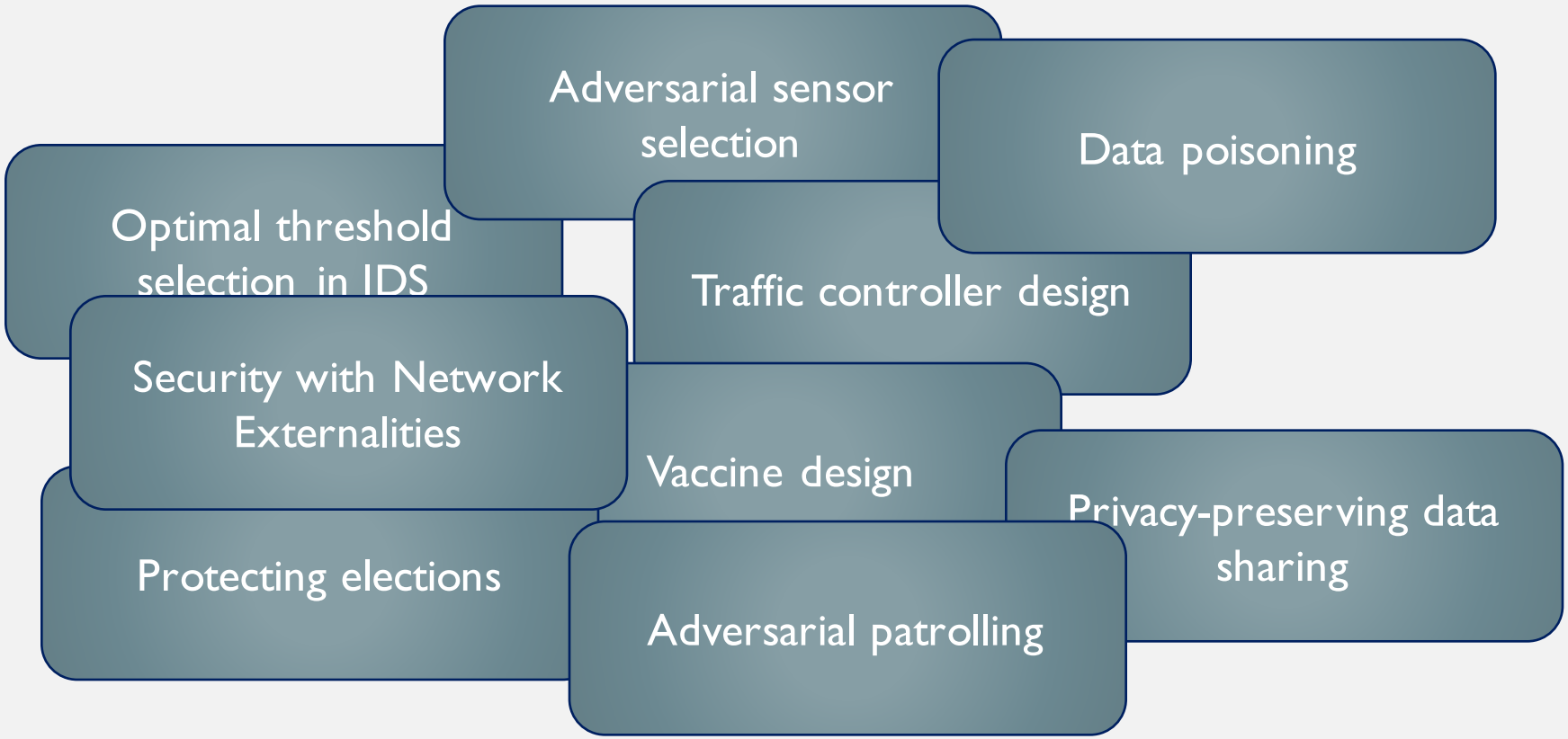
ADVERSARIAL AI @CERL



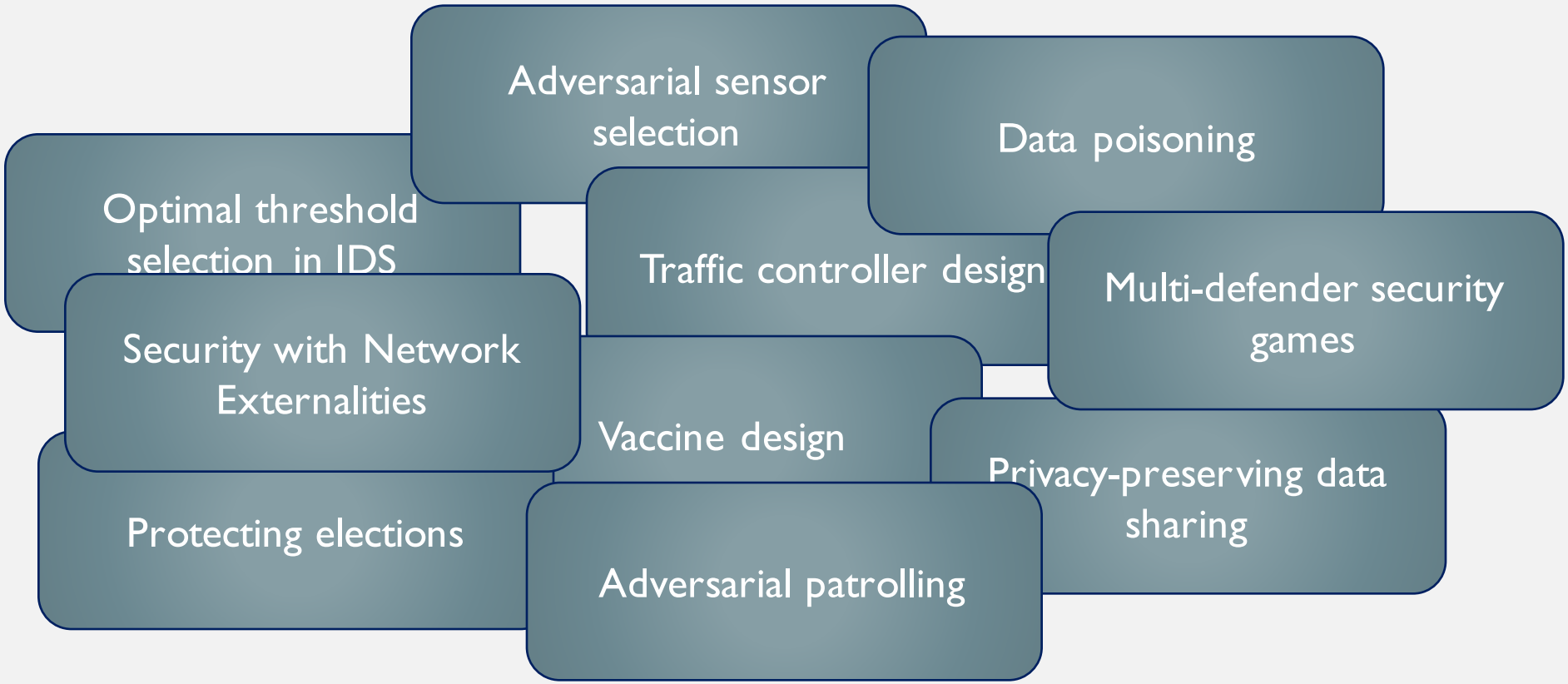
ADVERSARIAL AI @CERL



ADVERSARIAL AI @CERL



ADVERSARIAL AI @CERL



ADVERSARIAL AI

- As AI becomes increasingly important in a wide array of domains, can become increasingly manipulated or misused by malicious actors
- Broader goals:
  - Better understand vulnerabilities of AI algorithms, and "patch" these
  - Develop new algorithms for adversarial reasoning (in security and privacy)
- We've made some progress on these goals, but it is a vast area of research with a lot of as yet unexplored issues

THANKS TO COLLABORATORS, STUDENTS, & POST DOCS

- Collaborators:
  - Bo An
  - Ariel Procaccia
  - Milind Tambe
  - Long Tran-Thanh
  - Brad Malin
  - Jens Meiler
  - Ellen Clayton
  - Murat Kantarcioglu
  - Noam Hazon
  - Xenofon Koutsoukos
  - Josh Letchford
  - Steven Kimbrough
  - Howard Kunreuther
  - Satinder Singh
- Students:
  - Bo Li
  - Liyiming Ke
  - Ayan Mukhopadhyay
  - Swetasudha Panda
  - Jian Lou
- Post docs:
  - Aron Laszka
  - Waseem Abbas