Detection and Resolution of Control-Decision Anomalies in Semi-Autonomous Systems

Kang G. Shin
University of Michigan

Joint work with Daniel Chen and Noah Curran

Outline

- Motivation and Background
  - Boeing 737 MCAS
- Challenges and Approaches
  - CADD as an example
- Case Study of S-AVs
- Evaluation
- Conclusions

What’s a Semi-Autonomous Cyber-Physical System (SA-CPS)?

That can be controlled by an *autonomous controller* or a *human operator* or *both*

Human Control

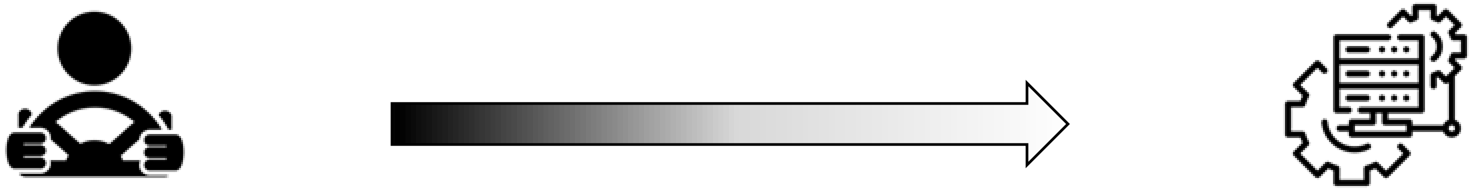
Autonomous Control

1912: Autopilot

1920’s: Autopilot

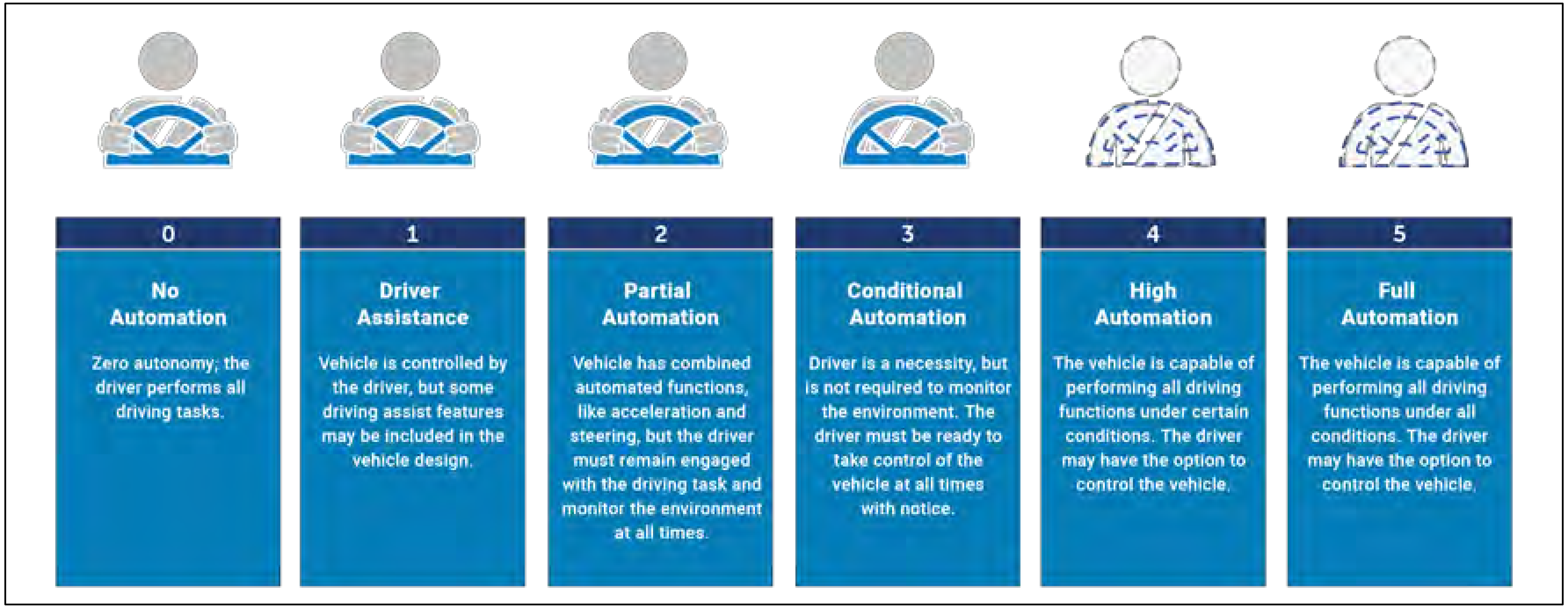
Semi-Autonomous Vehicles (S-AVs)

Vehicles controlled by an autonomous controller or a human or both



Human Control

Autonomous Control



SAE Autonomous Levels
Source: NHTSA (National Highway Traffic Safety Administration)
https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

Why Autonomy?

Safety

Eco-Friendliness

Comfort


Automated Vehicles for Safety

Overview

AV TEST Initiative — Learn about AV testing happening in the U.S.

Safety Facts

36,096


Big Fuel Savings From Autonomous Vehicles

Jeff McMahon

Follow

This article is more than 4 years old.


Transportation Research Part F

Comfort in automated driving: An analysis of preferences for different automated driving styles and their dependence on personality traits

Hanna Bellem, Barbara Thiel, Michael Schrauf, Josef F. Krems

ARTICLE INFO

ABSTRACT

https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

https://www.forbes.com/sites/jeffmcmahon/2017/04/17/big-fuel-savings-from-autonomous-vehicles/?sh=2a81009e4390

Bellem, H., Thiel, B., Schrauf, M., & Krems, J. F. (2018). Comfort in automated driving: An analysis of preferences for different automated driving styles and their dependence on personality traits. *Transportation Research Part F: Traffic Psychology and Behaviour*

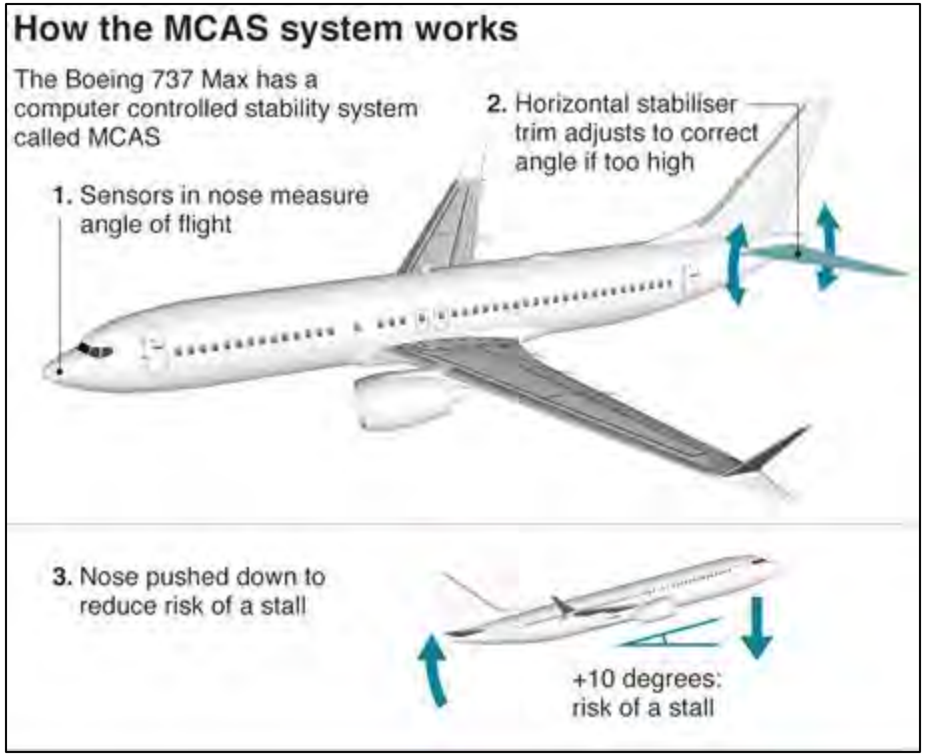
SOTA Semi-Autonomous Systems

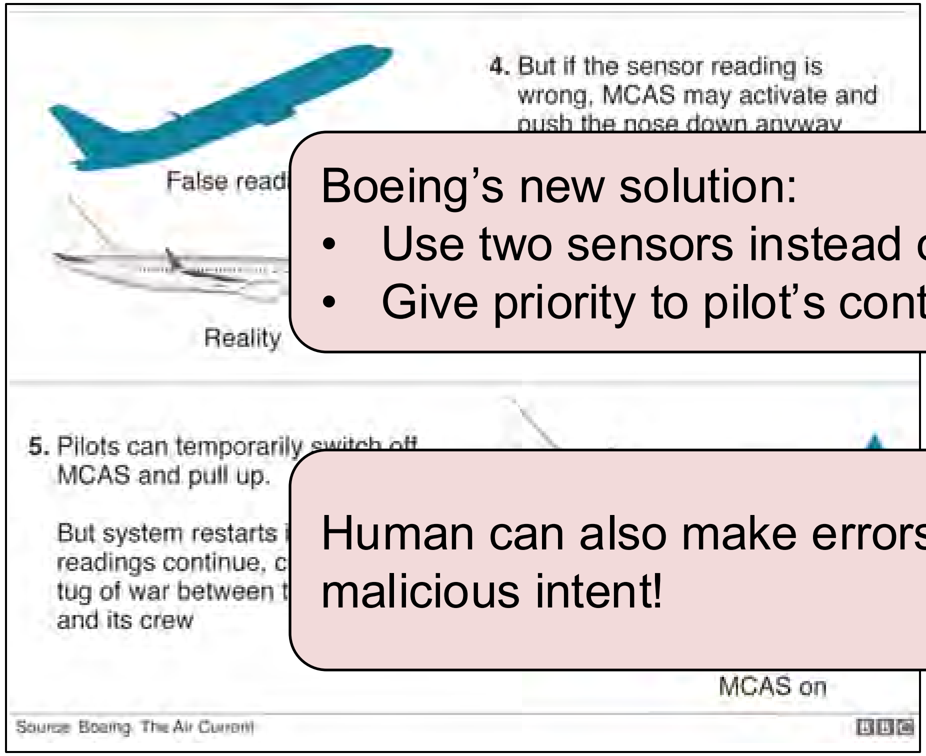
- *Static priority assignment* to control inputs under a specific condition
  - Anti-lock braking system (ABS) and previous Boeing 737 MAX
    - Autonomous control > human control
  - Adaptive cruise control (ACC)
    - Human control > autonomous control
- Assumes *trusted input data*

But *component or human failures/attacks* can lead to fatal accidents!

Static Priority Assignment Can Lead to Catastrophe!

- No perfect human-in-the-loop system!
  - Autonomous and human controllers may even "compete" with each other
    - Boeing 737 Max: MCAS (Maneuvering Characteristics Augmentation System)





Boeing's new solution:
- Use two sensors instead of one
- Give priority to pilot's control

Human can also make errors or have malicious intent!

Details of MCAS Redesign

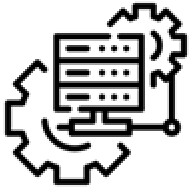
- Boeing and the FAA worked together to compile new rules for MCAS:

1. Both the left and right AoA sensor will be checked
    - If disagree by 5.5° or more *and* the flaps are up, do not activate MCAS
2. MCAS will activate *once* per sensed high-AoA/stall event
    - Lower the trim of the HS by 2.5°
3. MCAS cannot control the horizontal stabilizer more than pilots can by counteracting with the elevator
    - MCAS will not put more input on the HS than the pilot can put on the elevator
4. Flight control computer must have sensor integrity monitoring for MCAS

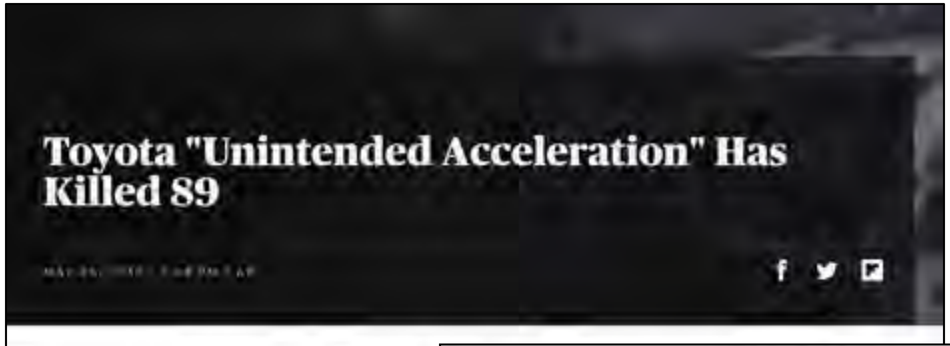
Is Boeing 737-MAX Still Safe? Analysis and Prevention of MCAS-Induced Crashes

Noah Curran, Thomas Kennings and Kang G. Shin [EMSOFT'24]

Safety Threats induced by Digital Plant or Controller



System Operation (Plant)
- In response to the control input
- Design flaws/bugs
- Run-time failure/faults


Toyota "Unintended Acceleration" Has Killed 89

Unintended acceleration in Toyota vehicl
people over the past decade, upgrading t
massive recalls, the government said Tue

The National Highway Traffic Safety Adm
received more than 6,200 complaints inv
The reports include 89 deaths and 57 inju
had been suspected of being connected t

Toyota Motor Corp. has recalled more tha
because of problems with gas pedals, floo
a record $16.4 million fine for its slow res
hundreds of state and federal lawsuits.


The New York Times

In Ethiopia Crash, Faulty Sensor on Boeing 737 Max Is Suspected

The flight data recorders on both planes survived the crashes, allowing investigators to begin piecing together what caused the disasters.


Tesla Model 3 driver again dies in crash with trailer, Autopilot not yet ruled out


Germanwings Flight 9525

From Wikipedia, the free encyclopedia

Germanwings Flight 9525[a][b] was a scheduled international passenger flight from Barcelona–El Prat Airport in Spain to Düsseldorf Airport in Germany. The flight was operated by Germanwings, a low-cost carrier owned by the German airline Lufthansa. On 24 March 2015, the aircraft, an Airbus A320-211, crashed 100 kilometres (62 mi) north-west of Nice in the French Alps. All 144 passengers and six crew members were killed.[2][3] It was Germanwings' first fatal crash in the 18-year history of the company.

The investigation determined that the crash was caused deliberately by the co-pilot, Andreas Lubitz, who had previously been treated for suicidal tendencies and declared "unfit to work" by his doctor. Lubitz kept this information from his employer and instead reported for duty. Shortly after reaching cruise altitude and while the captain was out of the cockpit, he locked the cockpit door and initiated a controlled descent that continued until the aircraft impacted a mountainside.

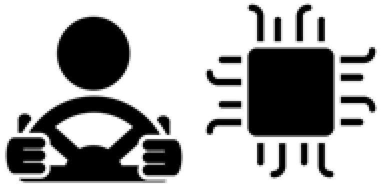

The New York Times

How a Self-Driving Uber Killed a Pedestrian in Arizona

By TROY GRIGGS and DAISUKE WAKABAYASHI UPDATED MARCH 21, 2018

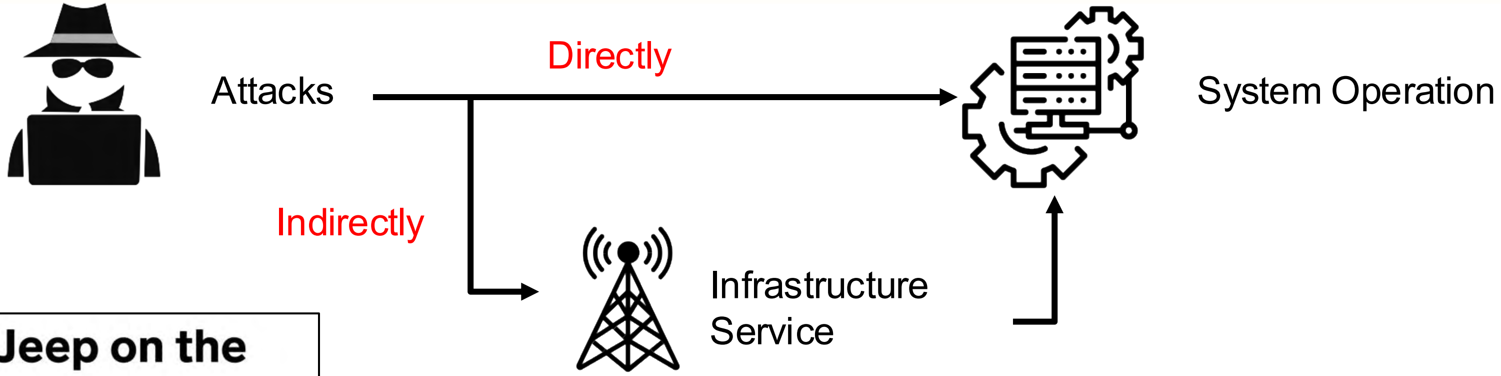
oman was struck and killed on Sunday night by an
nomous car operated by Uber in Tempe, Ariz. It was
ved to be the first pedestrian death associated with self-
ing technology.

What We Know About the Accident



Control Decision/Input (Controller)
- Incorrect human or autonomous control
- Design flaws/bugs
- Errors or malicious intents

What Makes Things Worse – Wider Attack Surces



Hackers Remotely Kill a Jeep on the Highway—With Me in It

Share

I WAS DRIVING 70 mph on the edge of downtown St. Louis when the exploit began to take hold.

Researchers Detail Spoofing Attack Against Vehicle GPS

A new paper shows that with the right amount of hardware and know-how, an attacker can spoof a vehicle's GPS system and change the route.

Researchers from Virginia Tech, the University of Electronic Science and Technology of China and Microsoft Research have come up with a way to mess with the GPS directional systems in a car. The type of spoofing attack means that a false route will supplant the one derived by the car's navigation system.

The attack is stealthy, and the driver will not realize that the attack is underway.

In their paper, "All Your GPS Are Belong To Us: Towards Stealthy Manipulation of Road Navigation Systems," the researchers outline how they did it.

Exposing Congestion Attack on Emerging Connected Vehicle based Traffic Signal Control

Qi Alfred Chen, Yucheng Yin, Yiheng Feng, Z. Morley Mao, Henry X. Liu
University of Michigan
{alfchen, yyucheng, yhfeng, zmao, henryliu}@umich.edu

***Abstract*—Connected vehicle (CV) technology will soon transform today's transportation systems by connecting vehicles and the transportation infrastructure through wireless communication. Having demonstrated the potential to greatly improve transportation mobility efficiency, such dramatically increased connectivity also opens a new door for cyber attacks. In this work, we perform the first detailed security analysis of the next-generation CV-based transportation systems. As a first step, we target the USDOT (U.S. Department of Transportation) sponsored CV-based traffic control system, which has been tested and shown high effectiveness in real road intersections. In the analysis, we target a realistic threat, namely CV data spoofing from one single attack vehicle, with the attack goal of creating traffic congestion.**

We first analyze the system design and identify data spoofing strategies that can potentially influence the traffic control. Based on the strategies, we perform vulnerability analysis by exhaustively trying all the data spoofing options for these strategies to understand the upper bound of the attack effectiveness. For the highly effective cases, we analyze the causes and find that the current signal control algorithm design and implementation choices are highly vulnerable to data spoofing attacks from even a single attack vehicle. These vulnerabilities can be exploited to completely reverse the benefit of the CV-based signal control

agency operations. These systems are currently under testing in three cities including New York City [12]. To push for a nationwide deployment, USDOT has already proposed to mandate all new light-duty vehicles to equip CV technology [11].

While having a great potential, such dramatically increased connectivity also opens a new door for cyber attacks. To ensure the security of vehicles and transportation infrastructure and the safety of drivers and pedestrians, it is highly important to understand potential security vulnerabilities so that they can be proactively addressed before nationwide deployment.

In this work, we perform the first security analysis on the next-generation CV-based transportation systems. As a first step, we target the USDOT sponsored design and implementation of a system called Intelligent Traffic Signal System (I-SIG), which performs one of the most basic urban traffic operations, traffic signal control. In this system, real-time vehicle trajectory data transmitted using the CV technology are used to intelligently control the duration and sequence of traffic signals. Such system is fully implemented and has been tested on real road intersections in Anthem, AZ, and Palo

Can we ensure a control decision is safe to execute even in the presence of component failures or malicious attacks?

Our Answer:

Context-Aware Detection and resolution of Control Anomalies (CADCA)

SOTA

- Assumes *trustworthy* input data [Fu'05, Zhang'17, Chhabra'17, etc.]
  => Input data can be anomalous

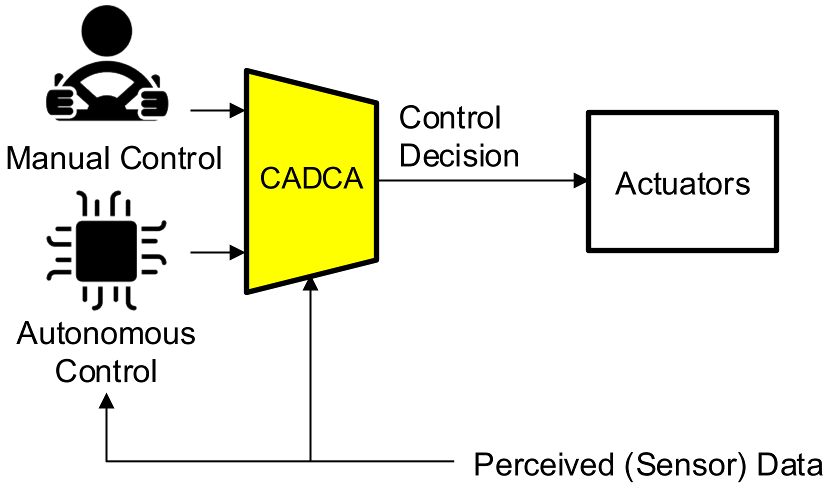
- Detects *out-of-pattern* behavior [Al-Sultan'13, Zhao'13, Shi'15, etc.]
  => Behavior consistency ≠ safety

- Assumes *complete, trusted* system observation/measurements [Shoukry'17]
  => *Limited* data availability

Need risk assessment with *limited data availability*
under *sensor faults* or even *attacks*.

CADCA is a Decision-Maker!

- Input = manual/autonomous controls + data from own sensors and others; Output = final control decision for execution by *each* ego entity
- Performs accurate risk assessment and then derives/selects safer control decisions even in the presence of
  - Failures or attacks on own sensors or false info from others
  - Imperfect controller or algorithms
  - Malicious/unsafe manual/autonomous control inputs



Challenges and Approaches of CADCA

C1: Are the received data correct/accurate?

C2: Can we infer what actually happened if data is anomalous?

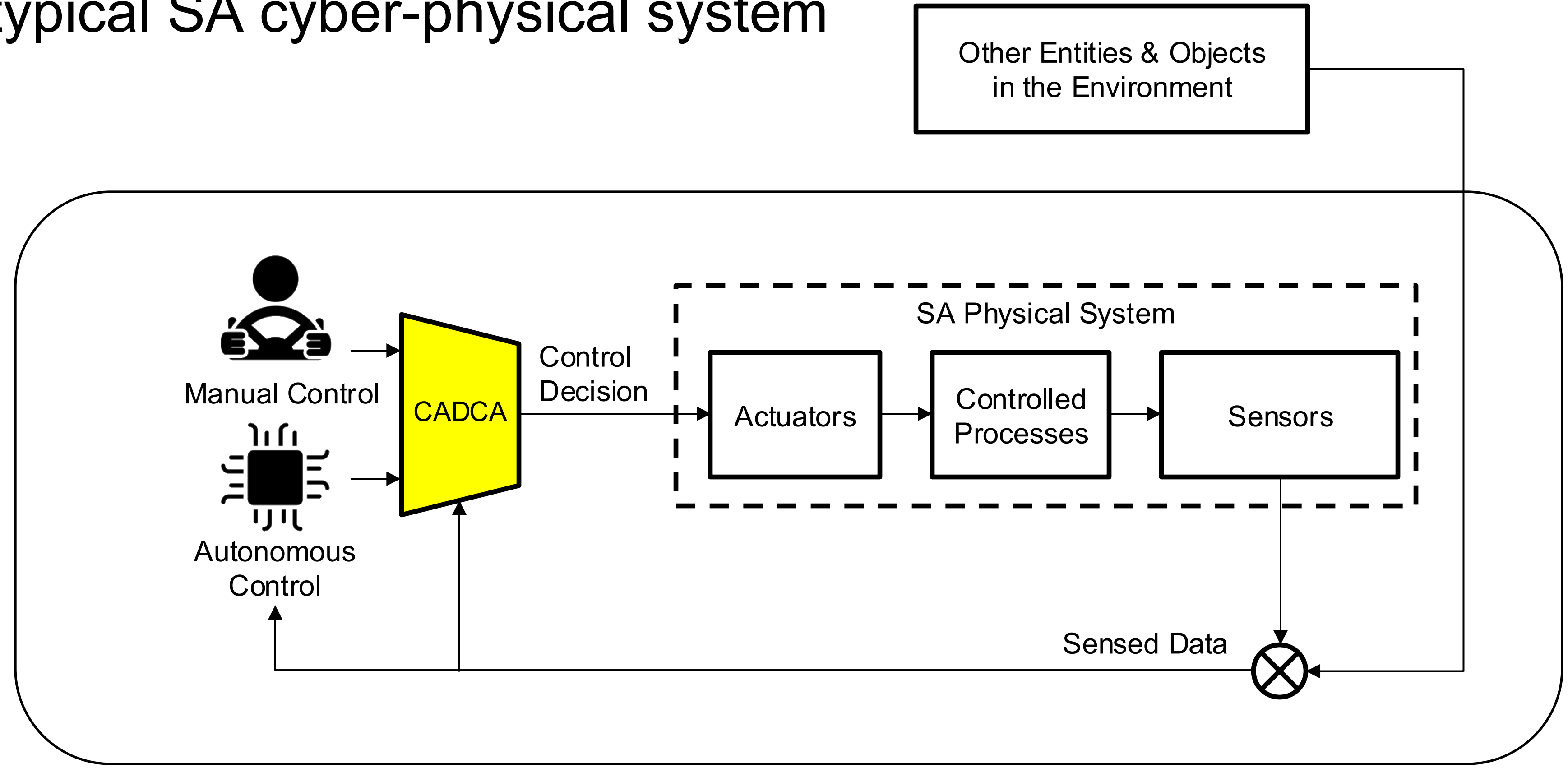
- *Different* situations can yield the *same* observation/data.
- Answer: Construct local views

C3: How to assess risks in an uncertain situation?

- Safety of a control decision depends on its operation context.
- Answer: Assess risks based on context matching

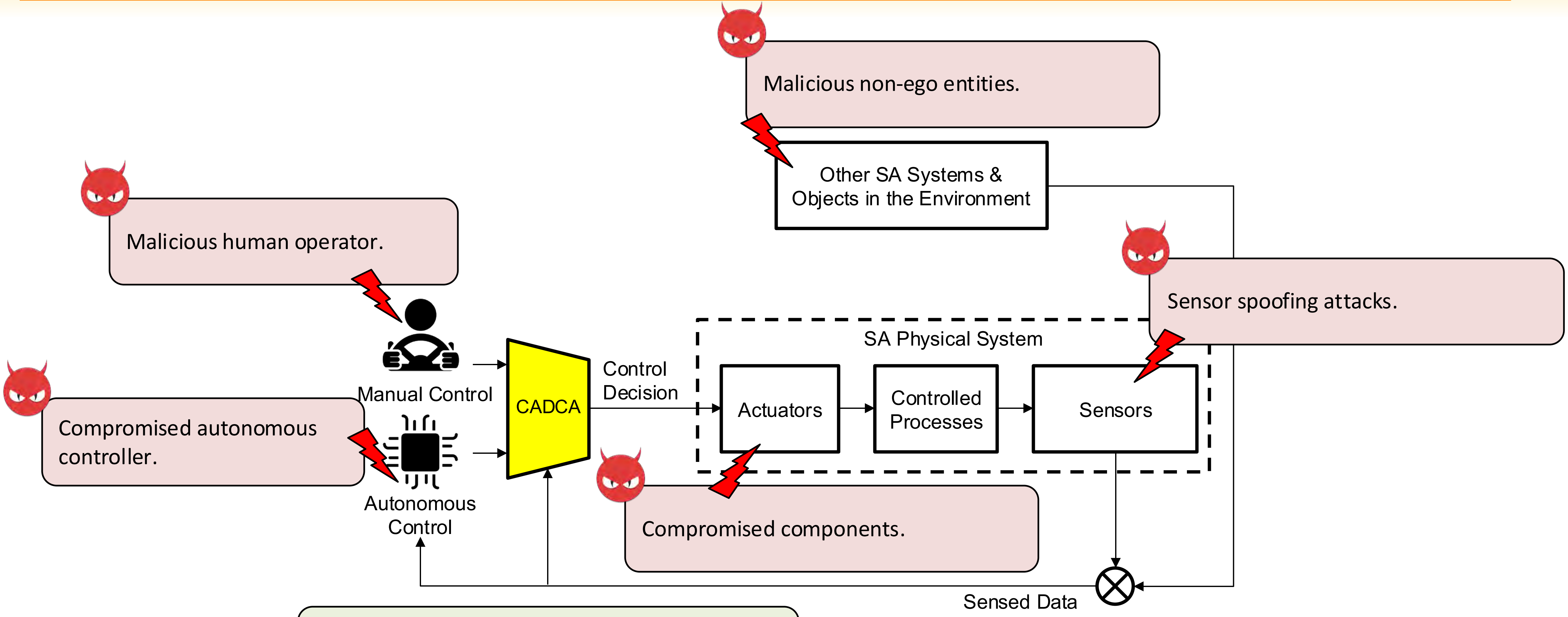
System Model

- Ego System (e.g., SAVs)
  - A typical SA cyber-physical system



Ego Semi-Autonomous System

Threat Model



CADCA's implementation can be trusted

3 Levels of Threat

Attacker's Capabilities

High

Sophisticated Threat -
Multi-Entity Attack w/ Matching Normal Data Correlation

Medium-level Threat -
Single-Entity Attack w/ Matching Normal Data Correlation

Naïve Threat -
Naïve Attack w/o Matching Normal Data Correlation

Low

Attack Scenarios

- Multiple Data
  - Matching Normal Data Correlation
    - Multi-Entity
      - Ego Anomalous (S6)
      - Ego Normal (S5)
    - Single-Entity
      - Ego Anomalous (S4)
      - Ego Normal (S3)
  - Without Matching Normal Data Correlation (S2)
- Single Data (S1)

Attack Attributes →

# of data under attack	Matching normal data correlation?	# of entities under attack	Is ego entity under attack?

CADCA's Capabilities Against 3 Levels of Threat

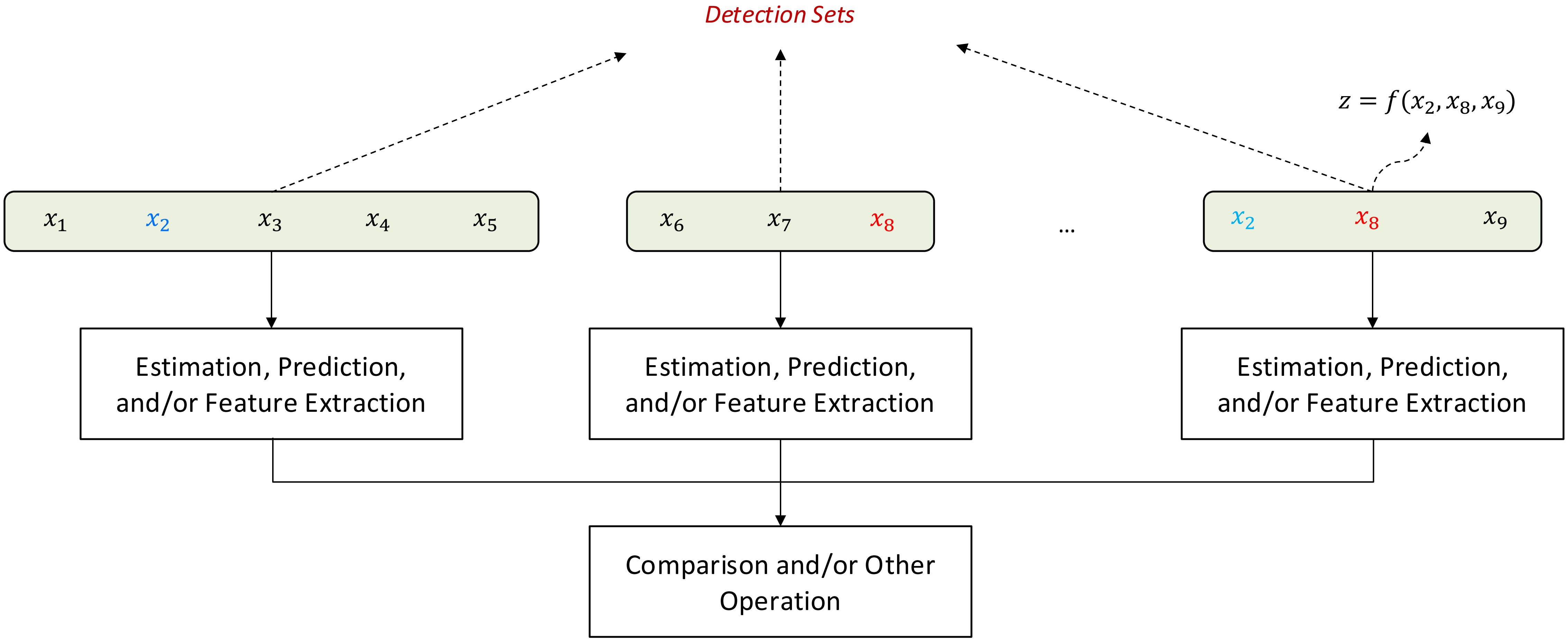
	Attacker's Capabilities	Detect Anomalies	Ensure Safety	Condition
High	**Sophisticated Threat -** Multi-Entity Attack w/ Matching Normal Data Correlation	◯	No guarantee.	Not all entities are compromised.
	Medium-level Threat - Single-Entity Attack w/ Matching Normal Data Correlation	◯	◯	Not all data from the victim are compromised.
Low	**Naïve Threat -** Naïve Attack w/o Matching Normal Data Correlation	◯	◯	

CADCA – System Design

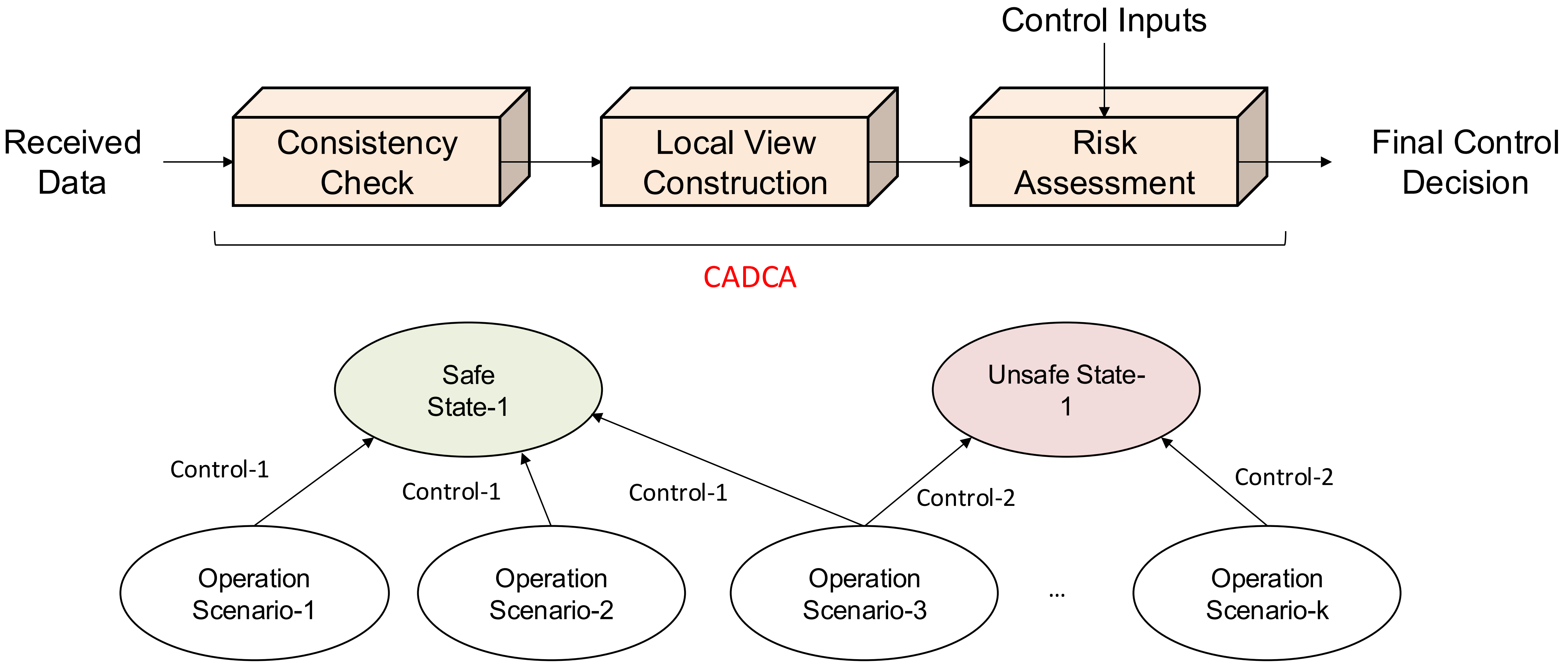
Detection/Verification Sets

Data Types of Interest	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	…

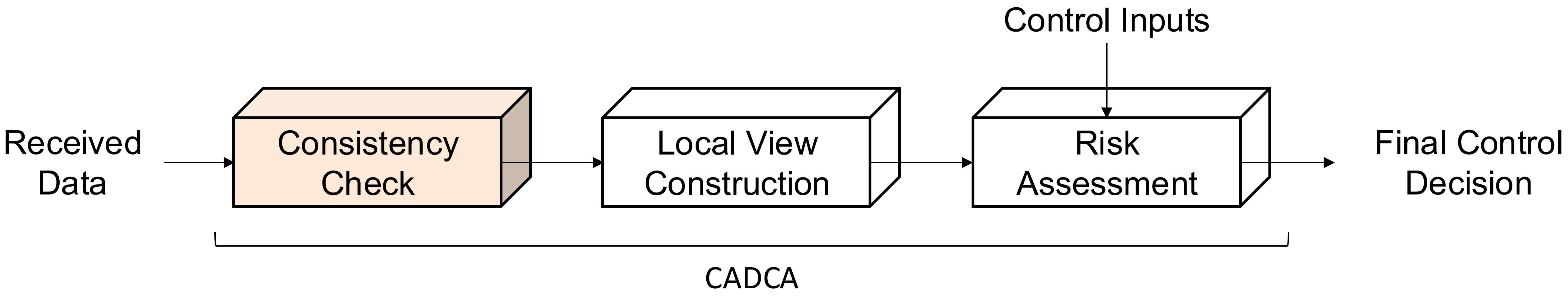
Detection/Verification Sets

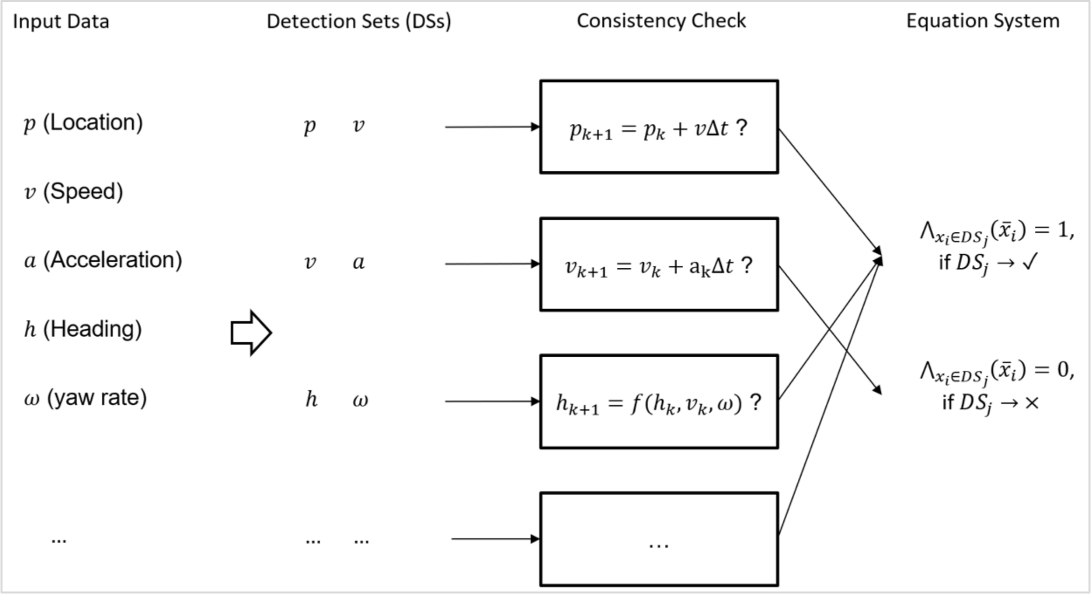


CADCA Pipeline



CADCA Pipeline – Consistency Check





CADCA Design – Consistency Check

Input Data	Detection Sets (DSs)	Consistency Check
p (Location)	p v	$p_{k+1} = p_k + v\Delta t$?
v (Speed)		
a (Acceleration)	v a	$v_{k+1} = v_k + \mathrm{a_k}\Delta t$?
h (Heading)		
ω (Yaw Rate)	h ω	$h_{k+1} = f(h_k, v_k, \omega)$?
…	… …	…

Equation System

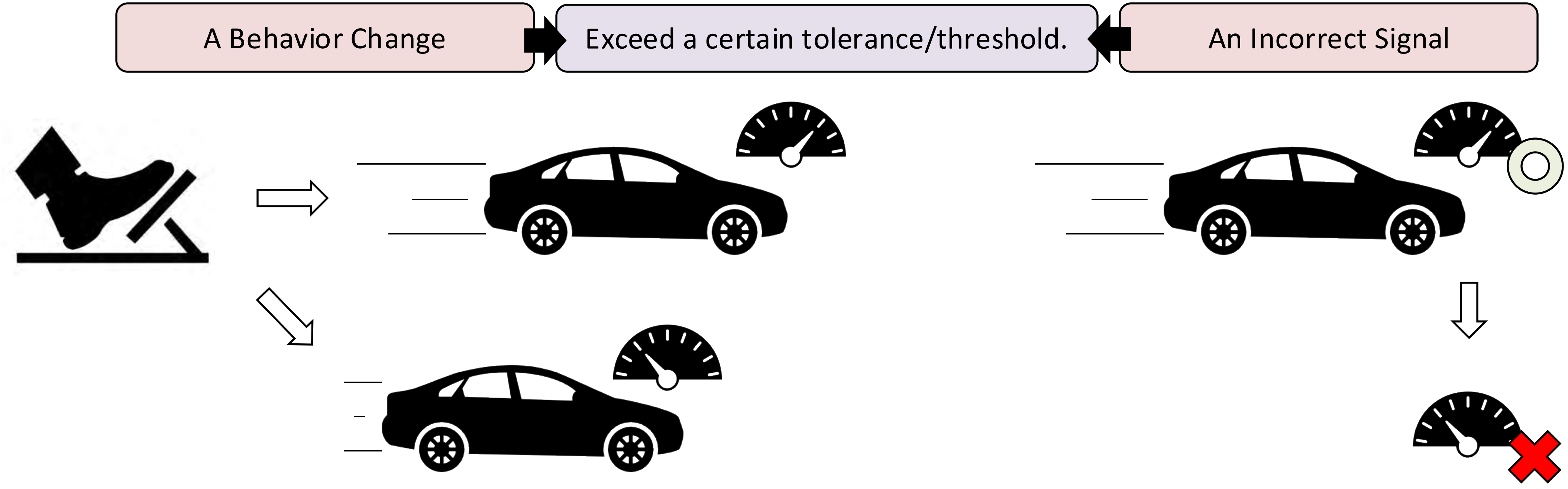
$\bigwedge_{x_i \in DS_j} (\bar{x}_i) = 1$,
if $DS_j \rightarrow \checkmark$,
where 1 indicate "normal"

$\bigwedge_{x_i \in DS_j} (\bar{x}_i) = 0$,
if $DS_j \rightarrow \times$
where 0 indicate "anomalous"

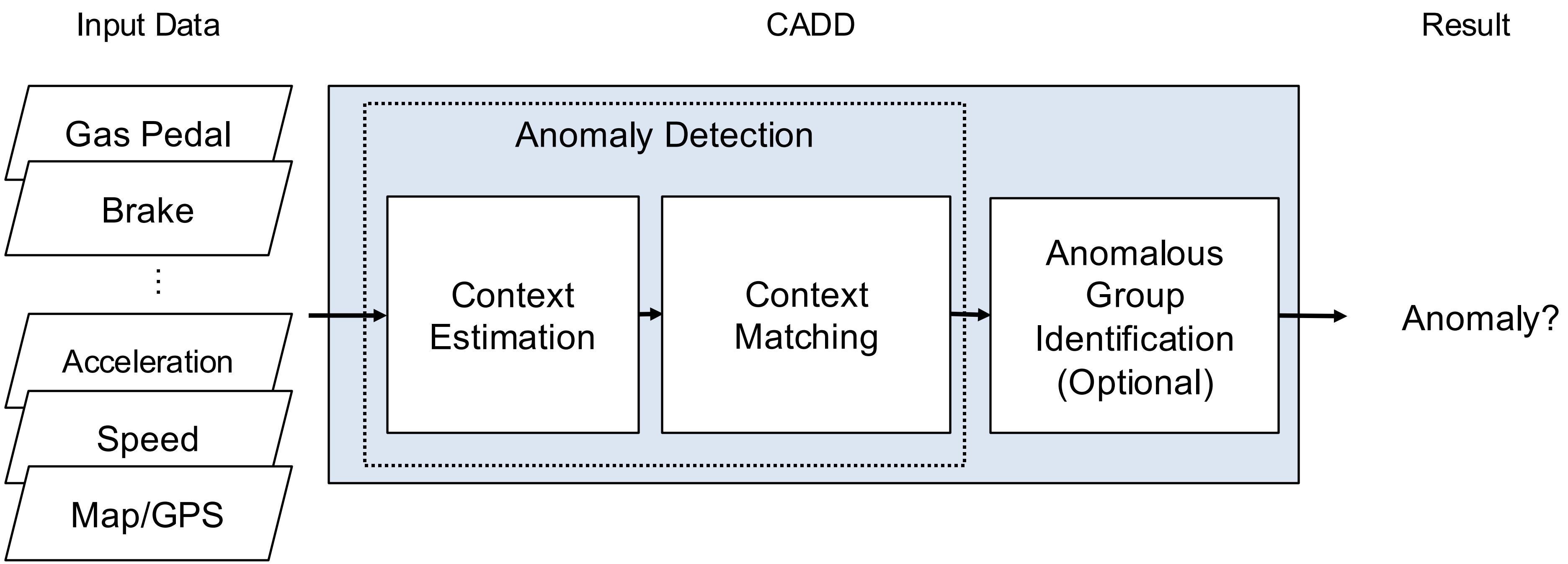
By solving the equation system, CADCA can identify which data is anomalous and then restore the normal data.

Example: Detection of SAV Anomalies

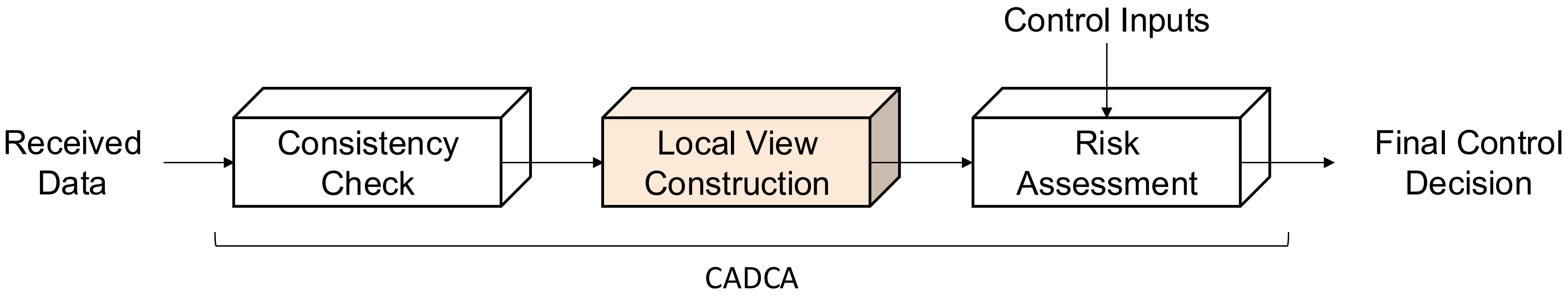
- How to efficiently detect an anomaly?
  - With limited data availability and training



– Anomaly Detection via *Context Estimation*



Construction of Local View



- *Local view* = states of the ego car and its surrounding entities
- Constructed by identifying/estimating potential operation contexts

ID	Loc.	Speed	Heading	…
0	0,20	50	270	…
1	0,0	30	90	…
…				

Example Local View

Local View Construction

Equation System

$\bigwedge_{x_i \in DS_j}(\bar{x}_i) = 1,$
if $DS_j \rightarrow \checkmark$

$\bigwedge_{x_i \in DS_j}(\bar{x}_i) = 0,$
if $DS_j \rightarrow \times$

Anomalous Data/Source Identification
(NP-Complete Fitting Set Problem)

Solution Space Algorithm (for anomalous data)

- Greedy Algorithm
- Anomalous Individual Algorithm
- Trust-Ego Algorithm
- Anomalous-Ego-Only Algorithm
- Anomalous-Ego-and-Others Algorithm

Identification Results
(operation context)

$N_G \Rightarrow normal\ set$
$A_G \Rightarrow anomalous\ set$

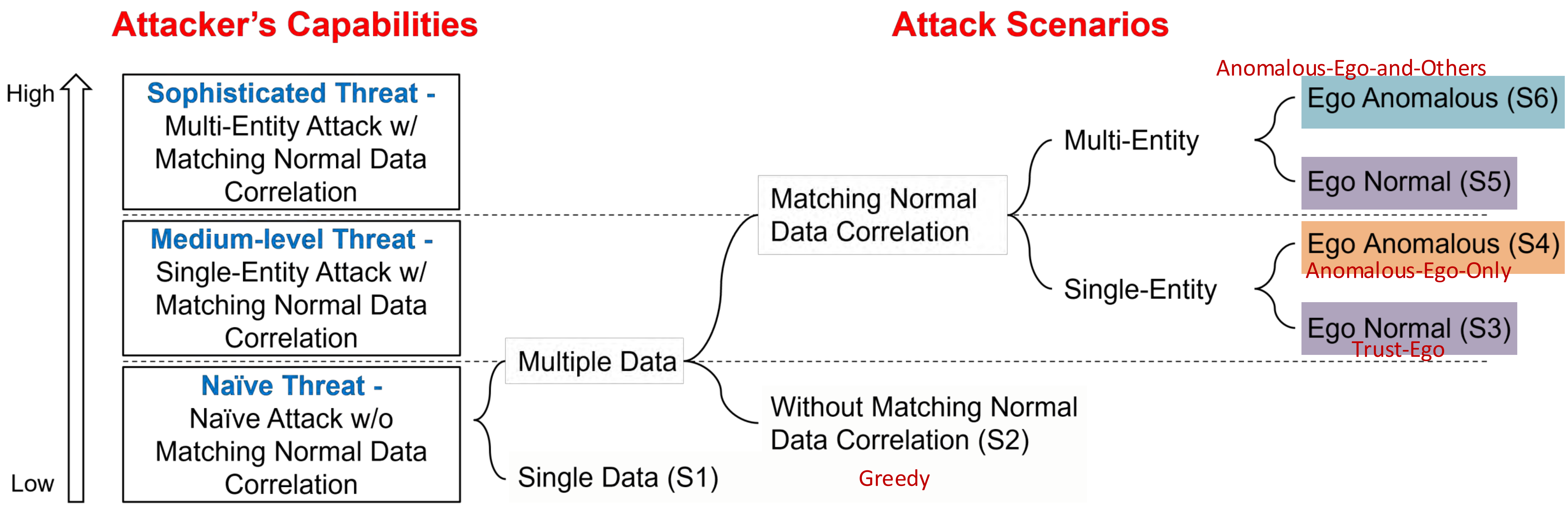
$N_{AI} \Rightarrow normal\ set$
$A_{AI} \Rightarrow anomalous\ set$

$N_{TE} \Rightarrow normal\ set$
$A_{TE} \Rightarrow anomalous\ set$

$N_{AE} \Rightarrow normal\ set$
$A_{AE} \Rightarrow anomalous\ set$

$N_{AEnO} \Rightarrow normal\ set$
$A_{AEnO} \Rightarrow anomalous\ set$

What Algorithms Cover



Attack Attributes →

# of data under attack	Matching normal data correlation?	# of entities under attack	Is ego entity under attack?

Algorithm Workflow

Consistency Check Results

Solution-Space Algorithm

Search Space of Anomalous Data L & Anomalous DSs List $L_{DS} = \{i | \forall DS_i \ni x_j, \forall x_j \in L\}$

(Greedy) Identify the data x in L that covers the most # of DSs in L_{DS}.

Put x in the result L_R and remove all DSs with x from L_{DS}.

If L_{DS} not empty

Otherwise

Final List of Anomalous Data L_R

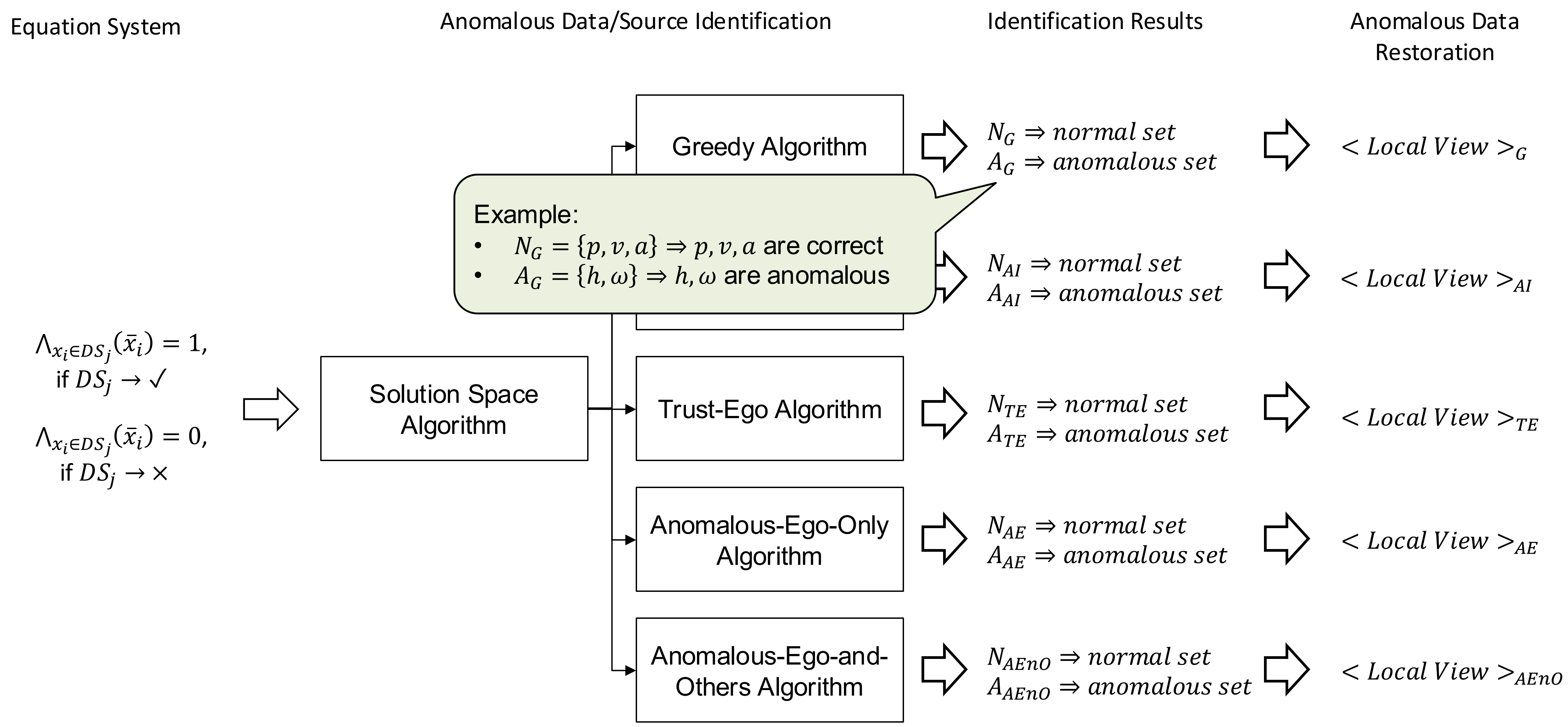
Anomalous Individual
Starts with the entity with the largest # of anomalous DSs.

Trust-Ego
Excludes ego data from L.

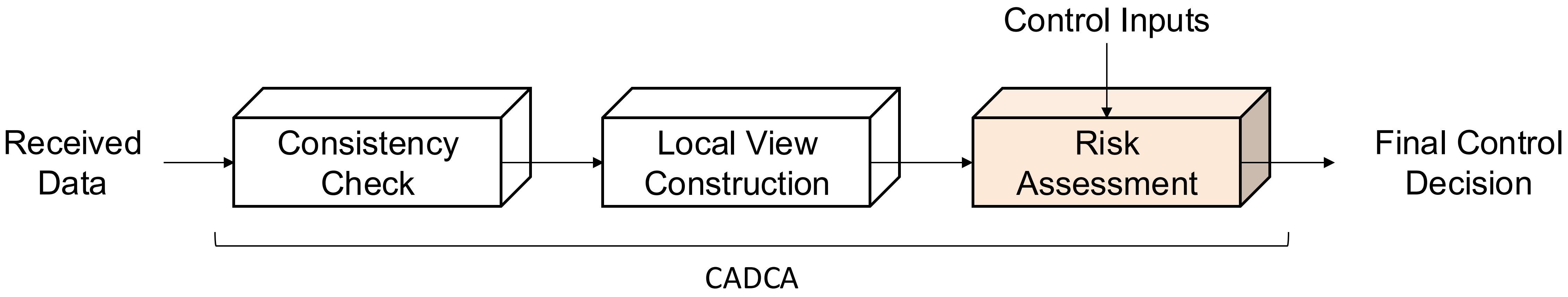
Anomalous-Ego-Only
Excludes non-ego data from L.

Anomalous-Ego-and-Others
Excludes non-ego data from L first and add them back.

Local View Construction, Cont'd



CADCA Pipeline Design - Risk Assessment



- Identify unsafe control decision

Risk Assessment Workflow

Local Views and Control Inputs

$< Local\ View >_G$ + { Control Input 1, Control Input 2, … }

$< Local\ View >_{AI}$ + { Control Input 1, Control Input 2, … }

$< Local\ View >_{TE}$ + { Control Input 1, Control Input 2, … }

$< Local\ View >_{AE}$ + { Control Input 1, Control Input 2, … }

$< Local\ View >_{AEnO}$ + { Control Input 1, Control Input 2, … }

Can be represented as a vector $< g_a, b_r, \theta_S >$.

- g_a: throttle
- b_r: brake
- θ_S: steering

Individual Risk Assessment

Each $< Local\ View,\ Control\ Input >$ pair computes

- Time-To-Unsafe_State (TTU) (e.g., Time-to-Collision)
- Time-To-Avoid-Unsafe_State (TTAU) (e.g., Time-to-Avoid-Collision)

TTU < TTAU ⇒ Safety Concern

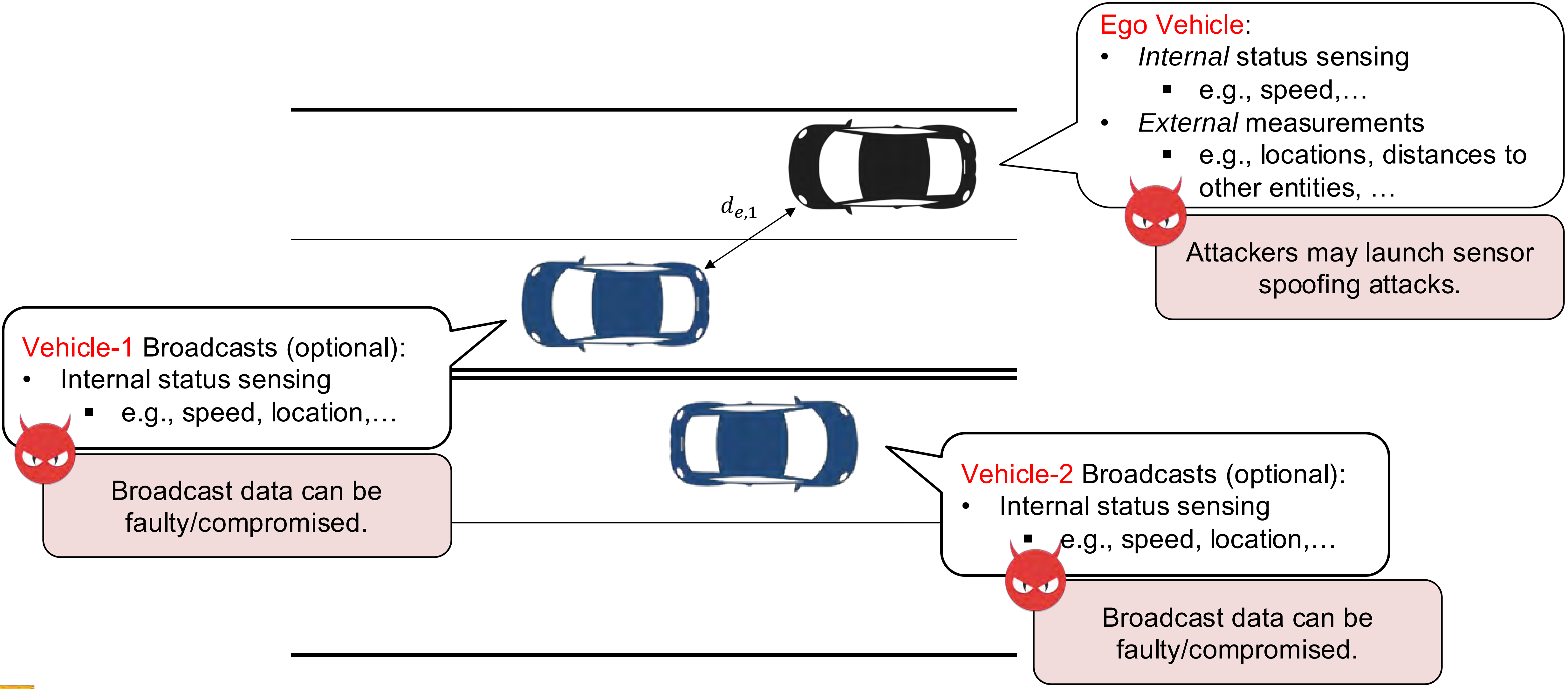
Generation of Final Control Decision

Results Aggregation

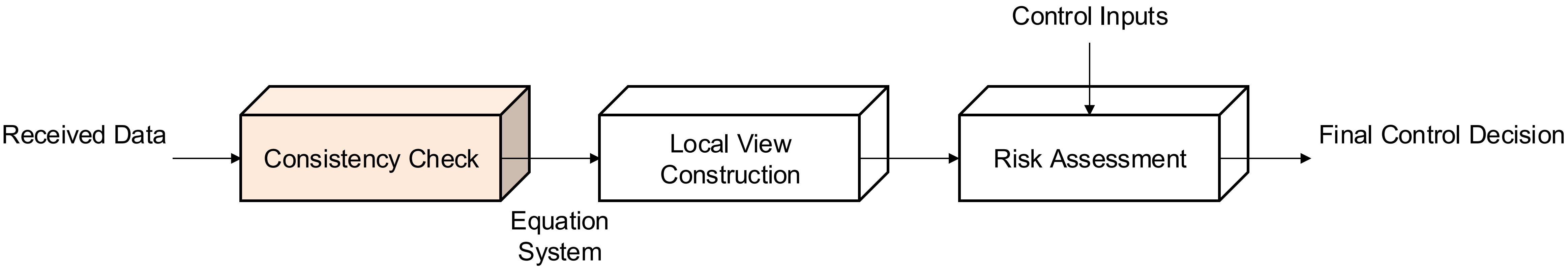
- **Common safe choice**
  → Never leads to unsafe states for all local views.
- **Most likely condition (if no common safe choice)**
  → Safe in the local view that matches the history of entity's normality.

Case Study of Semi-Autonomous Vehicles (SAVs)

Multiple SAVs on Road



Consistency Check



DS	Ego Data					Vehicle-1 Data					
	p_e	v_e	a_e	ω_e	h_e	p_1	v_1	a_1	ω_1	h_1	d_1
E,1	✓	✓	✓	✓	✓						
E,2	✓	✓									
E,3		✓	✓								
E,4				✓	✓						
E,5	✓				✓						
V1,1						✓	✓	✓	✓	✓	
V1,2						✓	✓				
V1,3							✓	✓			
V1,4									✓	✓	
V1,5						✓				✓	
V1,6	✓					✓					✓

Detection Sets (DSs) established based on the received data.

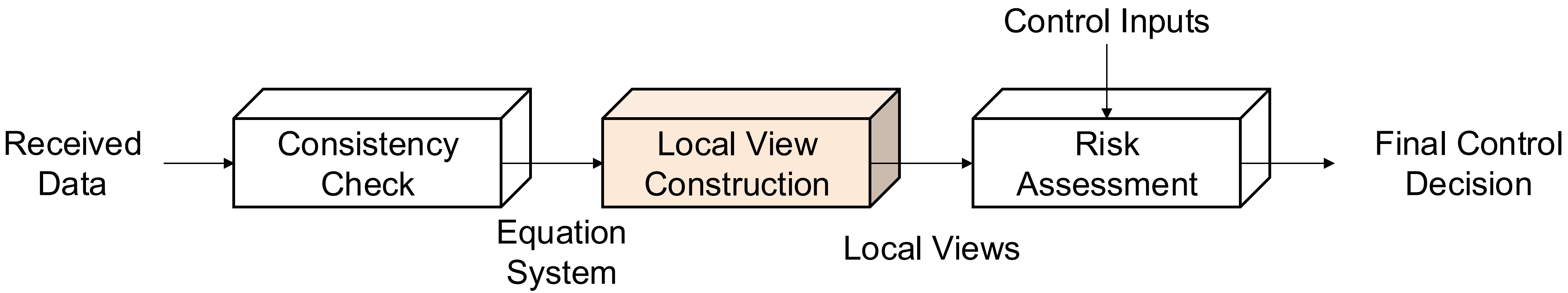
$$\sum_{\forall z_i(\tau) \in DS_x} I(z_i(\tau)) \geq 1,$$

Failure of matching data in a DS ⇨ an equation

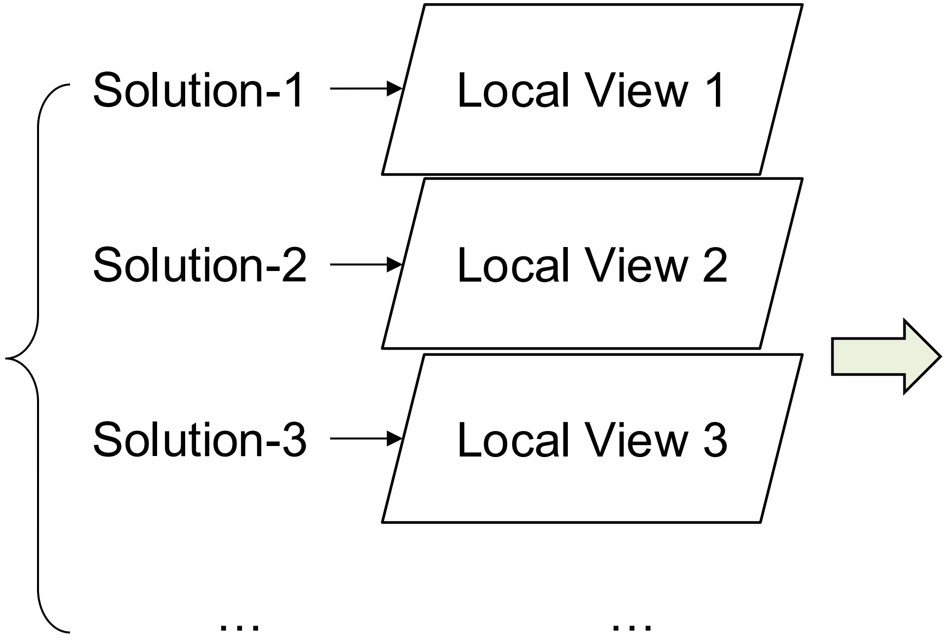
$$I(z_i) = \begin{cases} 1, & \text{if } z_i \text{ is anomalous.} \\ 0, & \text{if } z_i \text{ is normal.} \end{cases}$$

Solving equation system of data z_i's indicator function
⇨ Identification of anomalous data

Construction of Ego SAV's Local Views



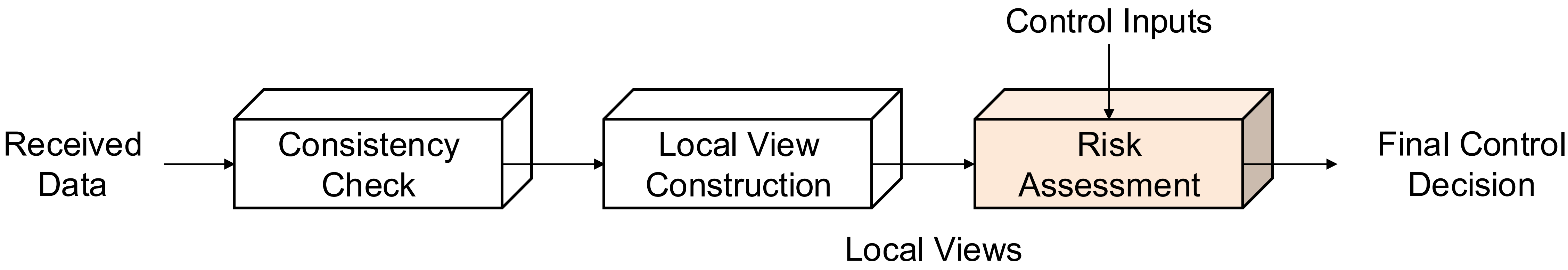
$$\sum_{\forall z_i(\tau) \in DS_x} I(z_i(\tau)) \geq 1$$



ID	Loc.	Speed	Heading	Dist.	…
0	0,20	50	20	-	…
1	0,0	30	20	20	…
…					

ID	Loc.	Speed	Heading	Dist.	…
0	0,20	50	20	-	…
1	0,10	40	20	40	…
…					

Risk Assessment



ID	Loc.	Speed	Heading	Dist.	…
0	0,20	50	20	-	…
1	0,0	30	20	20	…
…					

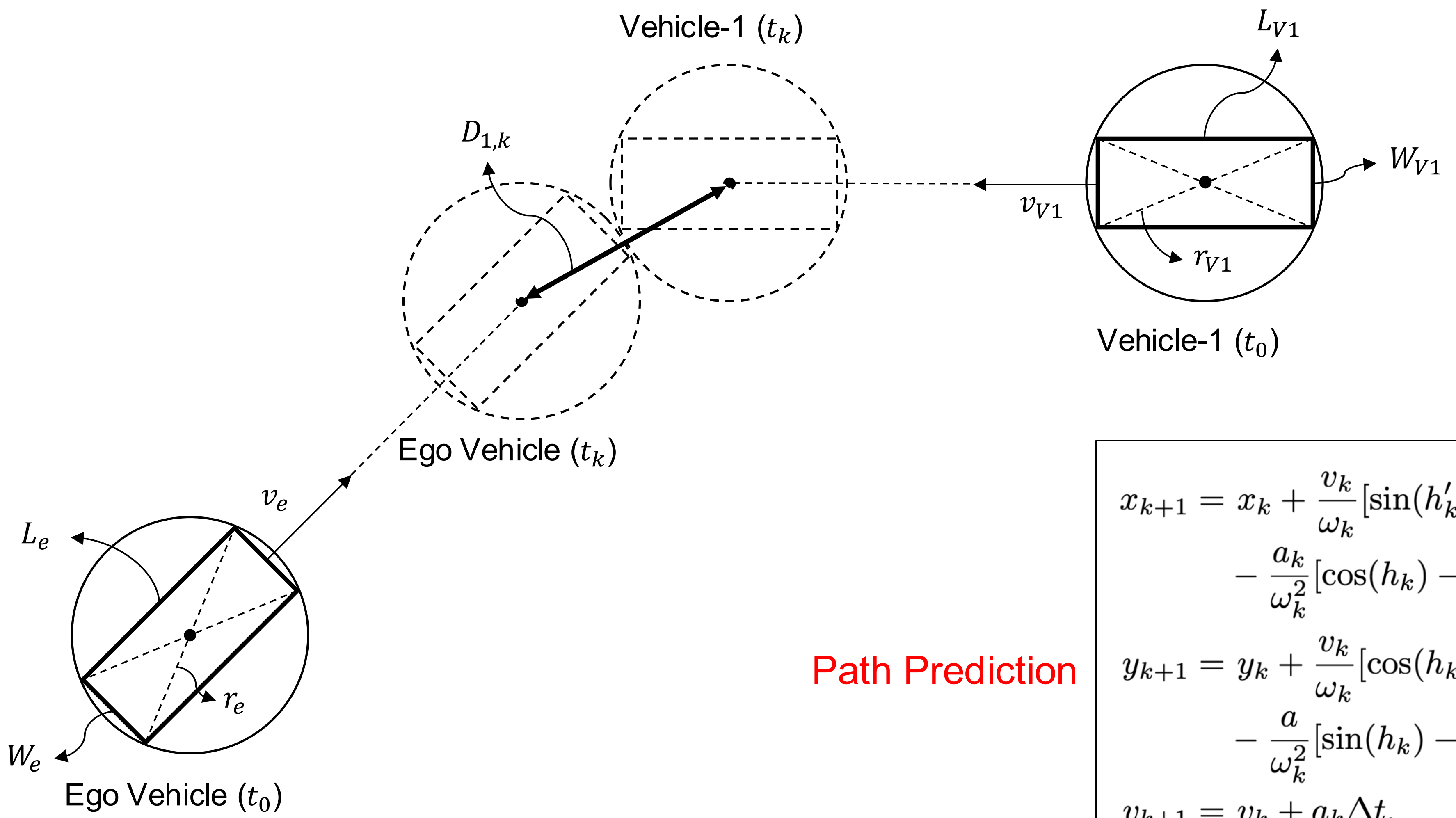
ID	Loc.	Speed	Heading	Dist.	…
0	0,20	50	20	-	…
1	0,10	30	20	20	…
…					

…

- Time-to-Collision (TTC)
- Minimum Required Response Time

Final Result Aggregation

TTC and Minimum Response Time Prediction



Path Prediction

$$
\begin{aligned}
x_{k+1} &= x_k + \frac{v_k}{\omega_k}[\sin(h'_k) - \sin(h_k)] + \frac{a_k \Delta t}{\omega_k}\sin(h'_k) \\
&\quad - \frac{a_k}{\omega_k^2}[\cos(h_k) - \cos(h'_k)], \\
y_{k+1} &= y_k + \frac{v_k}{\omega_k}[\cos(h_k) - \cos(h'_k)] - \frac{a_k \Delta t}{\omega_k}\cos(h'_k) \\
&\quad - \frac{a}{\omega_k^2}[\sin(h_k) - \sin(h'_k)], \\
v_{k+1} &= v_k + a_k \Delta t, \\
h_{k+1} &= h'_k = h_k + \omega_k \Delta t,
\end{aligned}
$$

Aggregation of Results and Selection of Control Decision

- Determine if a control decision is safe in a local view
  - Compare TTC with Threshold $T_C = \max(4.5, T_U + 2.6)$
    - 4.5s: Medium driver reaction time for smooth maneuver
    - 2.6s: Driver's reaction time in an urgent situation
    - T_U: minimum required time to avoid collision
- Control Decision Selection
  - Select the one that is safe in all local views

	Local View 1	Local View 2	Local View 3	Local View 4	…
Control-1	Safe	Safe	Safe	Safe	
Control-2	Safe	Unsafe, TTC = 1s	Safe	Safe	

  - If no control is safe in all local views
    - Choose the control with longest TTC under most probable local view

Determine Most Probable Local View

- Ego SAV records if a certain vehicle was anomalous in the past

Time →	t0	t1	t2	t3	t4	t5	t6	…
Ego	1	0	0	0	0	0	0	…
Vehicle-1	0	1	0	1	1	1	1	…
Vehicle-2	-	1	1	1	0	0	0	…
…								

Normal → 0
Anomalous → 1

	WAC
Ego	1
Vehicle-1	0.1
Vehicle-2	1
…	

- A most probable local view is selected based on
  - Weighted Anomalous Count (WAC)
    - Probability of the vehicle being anomalous
    - More weight on recent observations
  - WAC can also be computed based on *data* (not vehicle) level

Determine Most Possible Local View, cont'd

- Based on the inner product between WACs and local views
  - Need to "normalize" the results from the algorithms first
    - For fair comparison
    - Treat as vectors and compute based on inner-product

	WAC
Ego	1
Vehicle-1	0.1
Vehicle-2	1

Local View 1	Local View 1 (Normalized)
1	$1/\sqrt{2}$
0	0
1	$1/\sqrt{2}$

Local View 2	Local View 2 (Normalized)
1	1
0	0
0	0

$$1 \times \frac{1}{\sqrt{2}} + 0.1 \times 0 + 1 \times \frac{1}{\sqrt{2}} = \sqrt{2}$$

$$1 \times 1 + 0.1 \times 0 + 1 \times 0 = 1$$

Final Decision Output

- Determine the final control decision by considering the TTC of most likely local view

Most Probable/Likely Local View

	Local View 1	Local View 2	Local View 3	Local View 4	…
Control-1	Unsafe, TTC = 1.5s	Safe	Unsafe, TTC = 1s	Safe	
Control-2	Unsafe, TTC = 1s	Safe	Unsafe, TTC = 1.5s	Safe	

Conclusions

- CADCA: A Secure and Reliable Framework for General SA Systems, including
  - Autonomous and robotic vehicles;
  - Drones and airplanes; and
  - Ships and electric boats.
  - Etc.
- General System Structure + Domain-Specific Design
  - Ensure safety as the highest priority; and
  - Achieve operation smoothness and flexibility, if possible.
- More work needed on efficacy on several SA cyber-physical systems (AVs, MCAS, drones, robots,…)