

Causal Models for Accountability

Alexander Pretschner

Technische Universität München

fortiss research and technology transfer institute of the Free State of Bavaria

Bavarian Research Institute for Digital Transformation

jww Amjad Ibrahim, Severin Kacianka, Ehsan Zibaei

Shonan, June 2019

Roadmap

Accountability

Causality

- Causal models

- Notions of causality

Halpern-Pearl Causality

- Definition

- SAT-based computation

- ILP-based computation

What is Accountability?

Property of a system to help identify causes of events, and possibly assign blame.

Includes specifically monitoring (+integrity problems) and causality analyses.

Related: model-based diagnosis, runtime verification, forensics; but also fault localization, etc. Notions of causality.

Why Accountability?

Physics hard to overcome

Even for software, (design) contracts (alone) don't work - and cannot work (alone)

System boundaries of "open" (software) systems hard if not impossible to define:
Contracts define software interfaces at one (!) useful level of abstraction

Many known recurring integration faults

Hence: a priori avoidance of problems by design increasingly hard if not impossible

Side remark: Maybe contracts should be negative specifications; defect-based QA

Accountability for CPS

Capture elements common to all CPS.

Find interfaces/operations necessary for accountability mechanism.

Define “accountability” on top of them.

There are different definitions in the sciences

Provide blueprints to compare actual systems to.

NB: CPS in the European sense, as a sociotechnical system

Uber example

Root cause hypotheses over time:

1. „Couldn't have been detected“
2. Safety backup driver didn't pay attention
3. Too few LIDAR sensors
4. Crash protection was disabled to avoid false positives
5. Threshold for object detection too low: flying plastic bag
6. ...

1. Accountable: safety backup, sensors, software, Volvo, Uber? [don't expect too much! can only answer partially by now! plus, will concentrate on technical system parts!]

Things (may) go wrong: Analysis tasks

At design time:

Fault tree analysis, attack tree analyses, FMEAs, ...

At runtime:

Runtime verification, complex event/stream processing

Post mortem:

Accountability

Will focus on post-mortem analysis: understanding the why, not mitigating the what at runtime or design time.

Underlying models can be re-used though.

Is accountability new in computer science?

Certainly not: security, safety, forensics, ...

What I am interested in:

Building accountable (cyber-physical) systems

Methodology of deriving and maintaining causal models

Efficient implementation of causality

How I got into this:

Detective distributed data usage control & data provenance

Software testing and fault localization

Insider attacks

Roadmap

Accountability

Causality

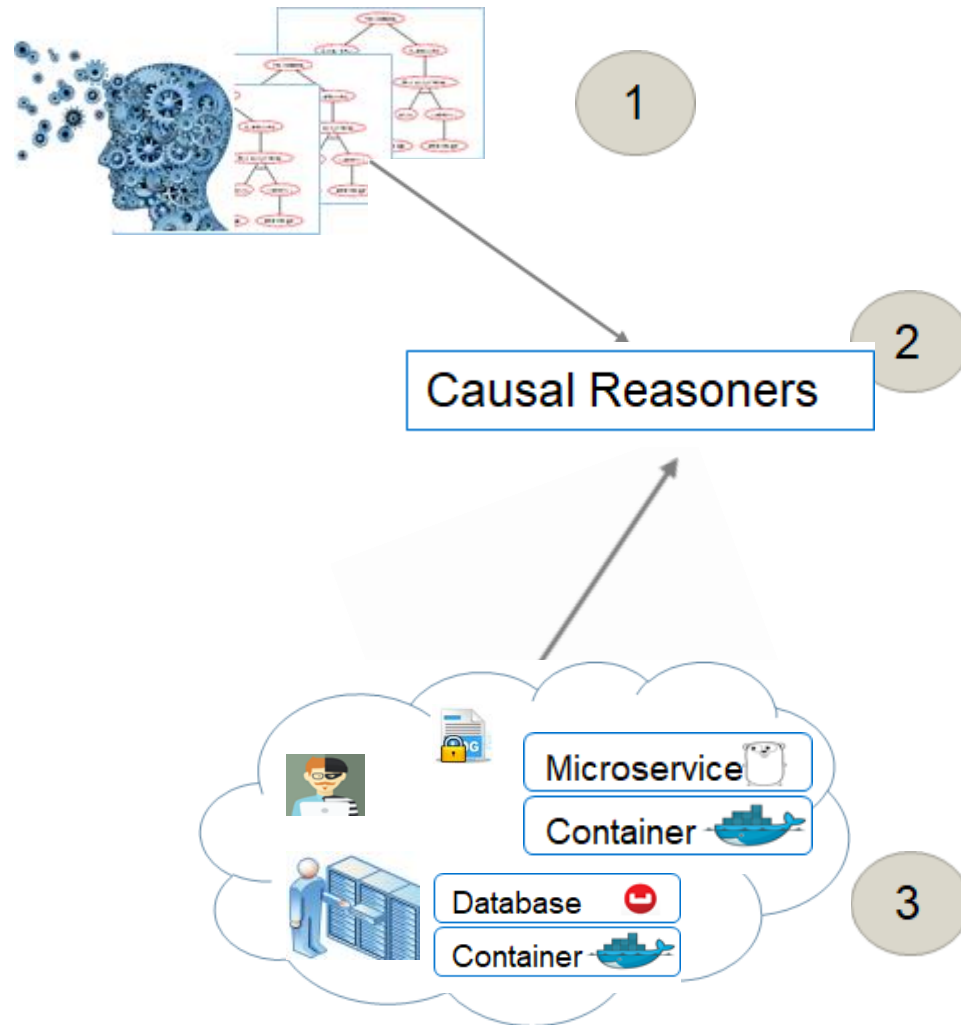
Introduction [...]

Exemplary causal models

Flavors of causal reasoning

Halpern-Pearl Causality

Approach



Causal Modeling: is the art of abstracting the causal factors and their relations.

Causal Reasoning: is the process of inferring actual causality using a **model** and a **context**

Context Setting: is the step of setting the actual values of the variables in a model

Not uncontroversial:

There must be causal models to explain “causality”.

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y x)$	Seeing	What is? How would seeing X change my belief in Y ?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y do(x), z)$	Doing Intervening	What if? What if I do X ?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x x', y')$	Imagining, Retrospection	Why? Was it X that caused Y ? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smok- ing the past 2 years?

Causal Models: Example Fault Trees

Top-down approach to quality assurance:
decompose problematic events into their causes

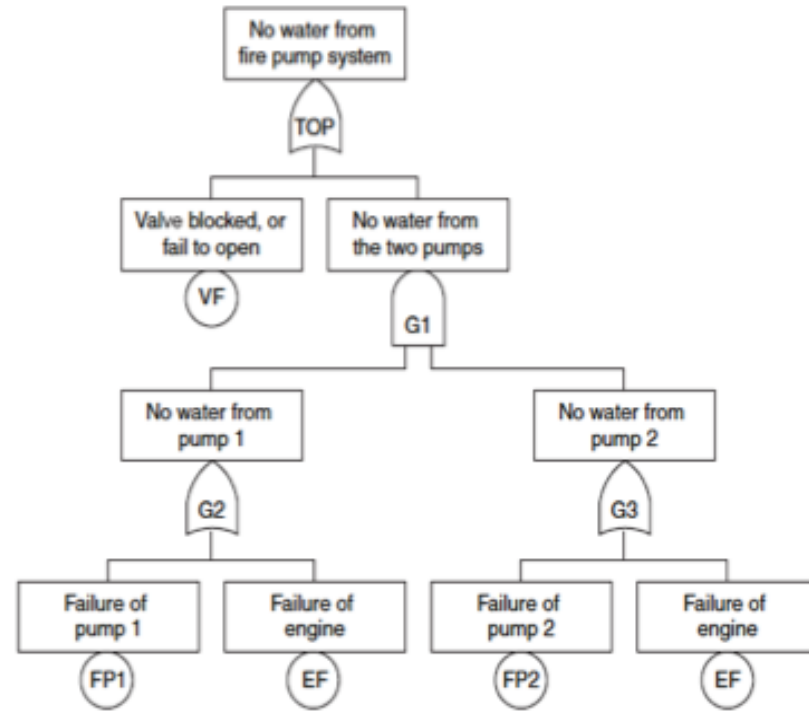
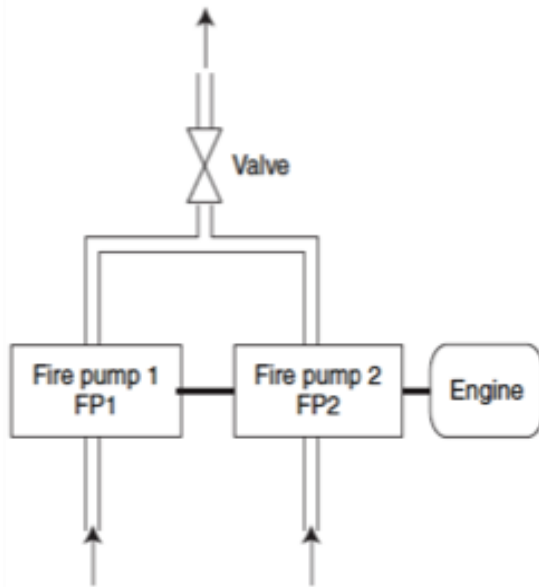
Nodes are events; edges denote „and“ and „or“ relationships. Hence propositional formulas.

Perform analysis how top-level event can happen (possibly minimal „cut sets“)

Used a-priori to identify mechanisms that prevent specific events or react accordingly

Can also be used a-posteriori

Example



Where do the fault trees come from?

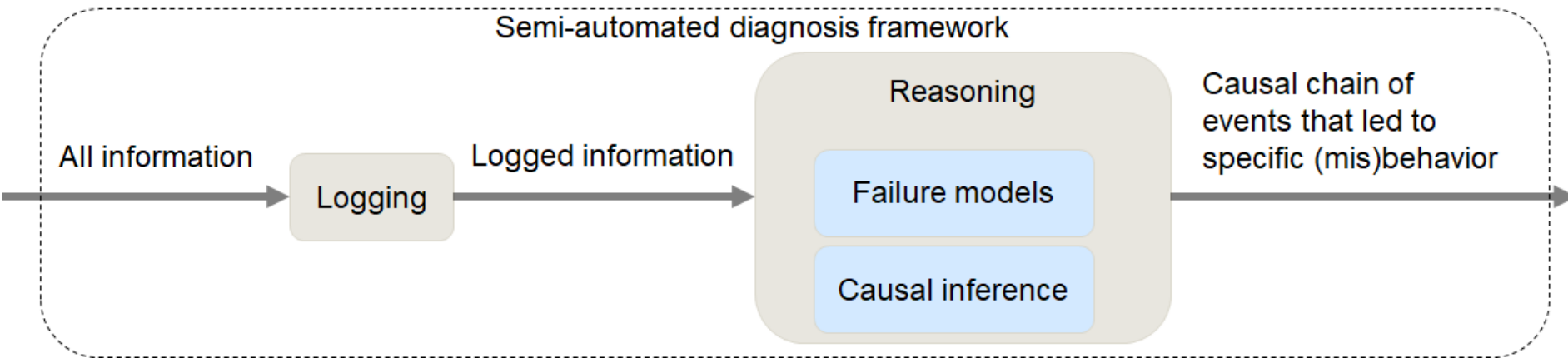
Realistically, need 4 PhDs:

mechanical, electrical, aerospace, software engineering

Hence create models?

Example: Accountability for drones

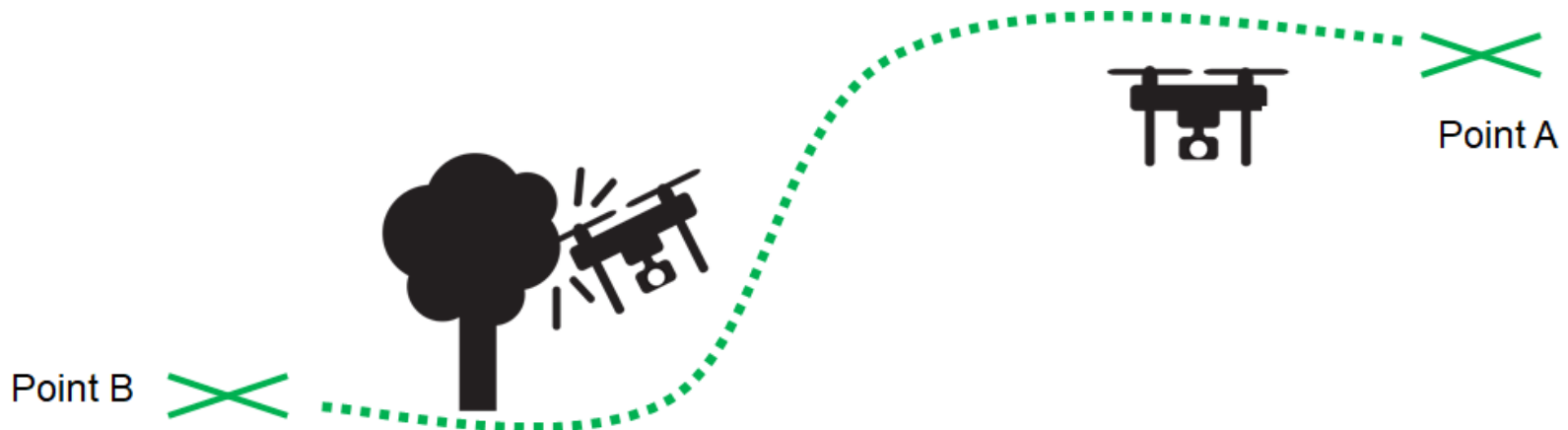
- Semi-automated diagnosis framework to diagnose the root cause of the (mis)behavior of drones
- Logging and reasoning are two main parts of the proposed framework
- We do the reasoning by having a model of the failure of the system e.g., Fault Trees and a causality algorithm e.g., Halpern and Pearl's.
- Fault tree templates created systematically;
then analysis and rule definition or machine learning on logs



(mis)Behavior of drone

- Violation of non-functional requirements:
 - Collision with objects
- Violation of functional requirements:
 - Deviation from desired destination
 - Dropping an object in a wrong location
 - Inability to maintain a communication network over an area affected by earthquake

Most of the drone's misbehavior is manifested as **wrong motion** in 3D space.



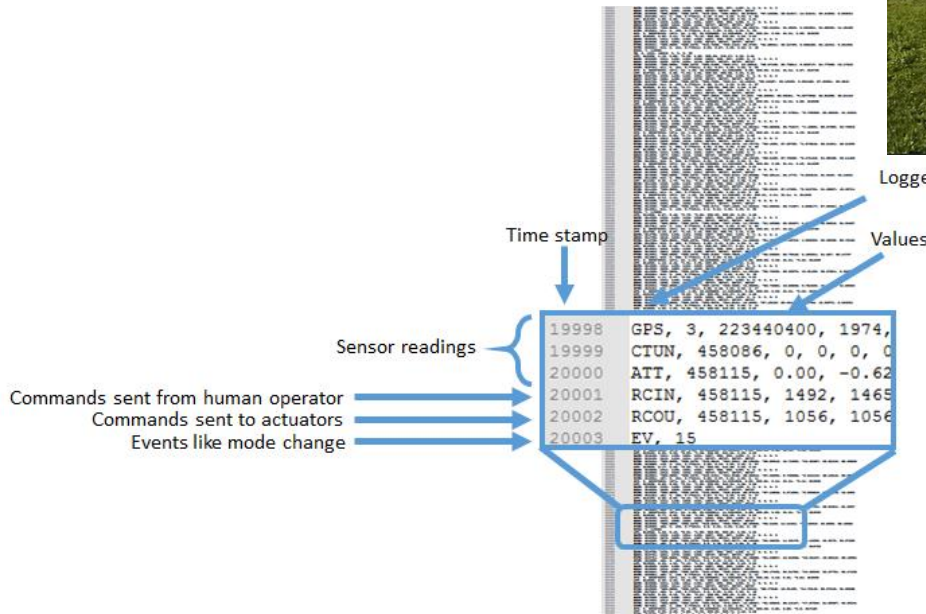
Data generated by drones

Flight logs consist of information generated by sensors, actuators and other components of the drone. These information are collections of observations generated sequentially through time (time series).

The screenshot shows a 'Status Information' window with the following sections:

- Connection:** COM(USB->Falcon)
- Packets:** IMU Raw Data, RC Data, CTRL Output, CTRL Internal-Out, GPS Data Adv, Current_Way, Debug_Data, AccTies Falcon
- Status:** UART Comm, Motors activated, Data CHKSUM, Flying, CCPS: 1000
- Flight Mode:** Height Control, GPS, Social Interface Enabled, Social Interface Active, Emergency
- Raw Sensor Data:** Gyro X, Y, Z; Mag X, Y, Z; Acc X, Y, Z; Temp (gyro); Temp (ADC); Pressure
- Calculated IMU Data:** Height, dHeight, Angle (Pitch, Roll, Yaw), Vel., Reference, Total Acc, Pseudo Velocity, Trans. Acc, Total Acc XYZ
- GPS:** Latitude, Longitude, Speed, Height, # of Sat, Status
- Fused Data:** Latitude, Longitude, Speed X, Y
- RC Data:** Pitch, Roll, Thrust, Yaw, Ser. int. sw., AUX1, AUX2
- Debug Data:** d1, d2, d3, d4
- Sensor Status:** Accelerom., Gyros, Mag fieldstr., Mag incl.
- Waypoint:** Distance to WP, Current X, Y
- Flight-Time:** 00:24
- Battery:** 12.84V

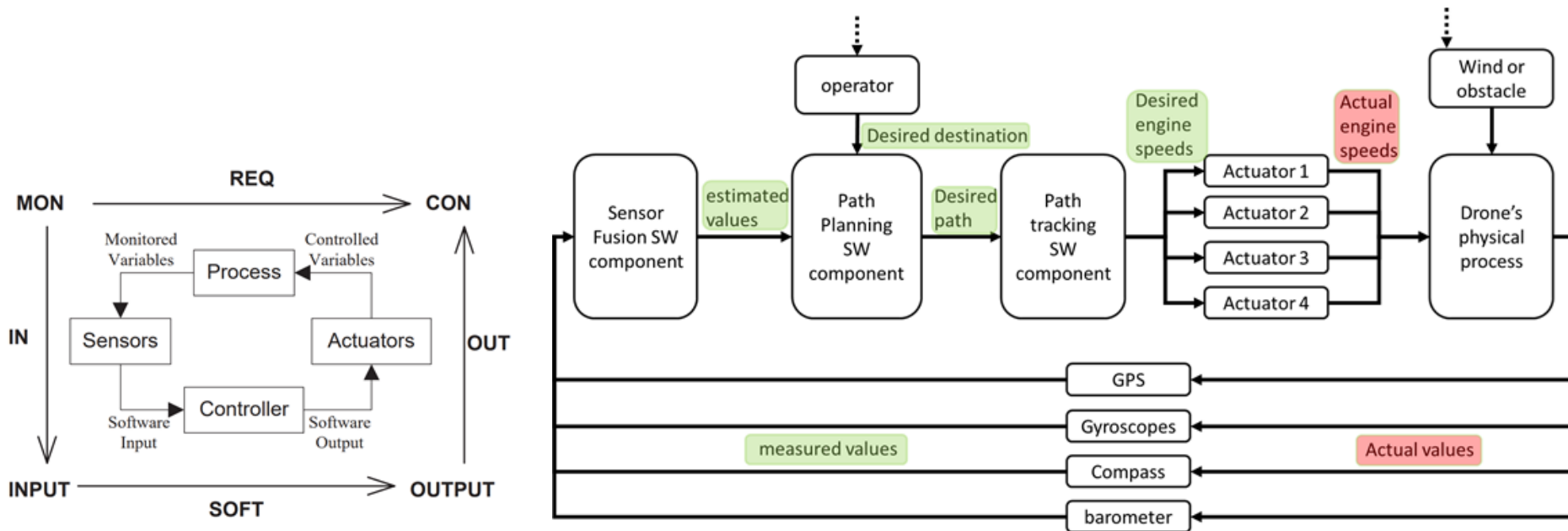
Logged components



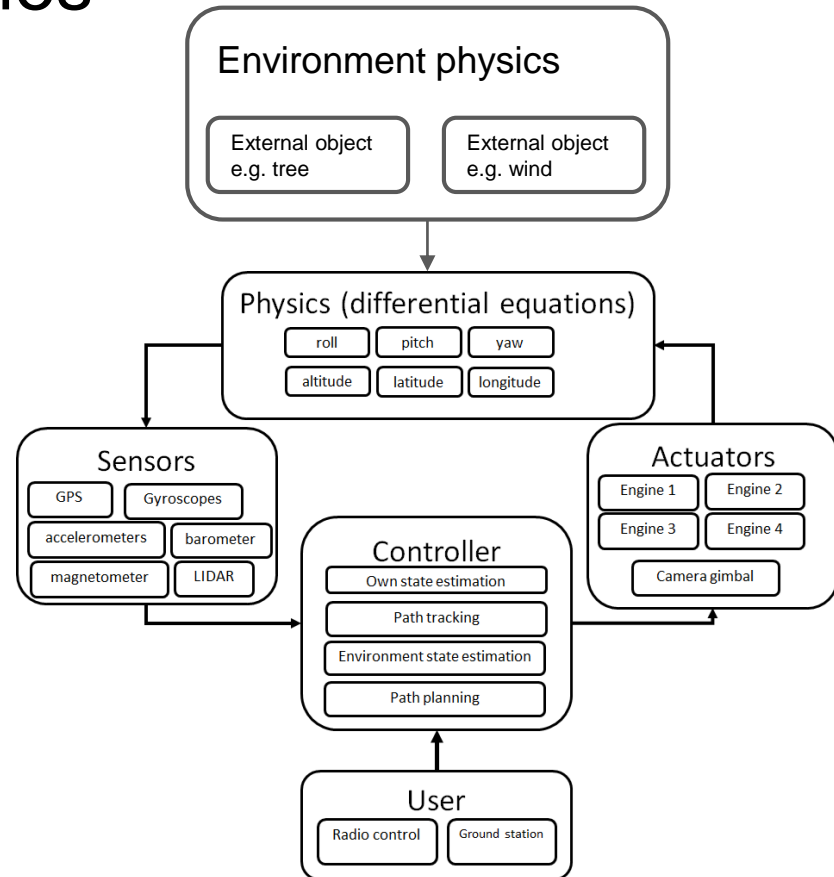
Modeling behavior and misbehavior of drone

1. Four variable model [Parnas, Madey 1995]

2. Extracting causal chains of misbehaviors
3. Building Fault trees

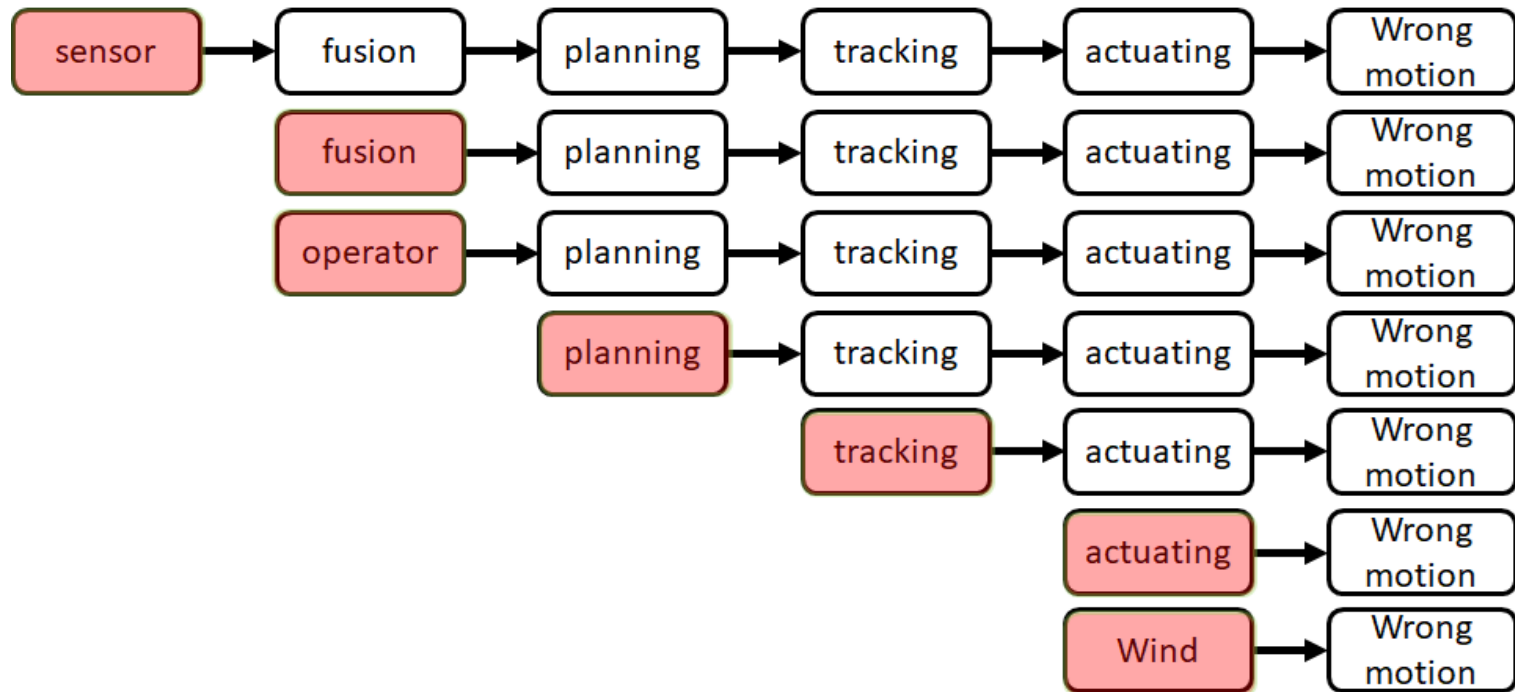


4 variable model for drones



Modeling behavior and misbehavior of drone

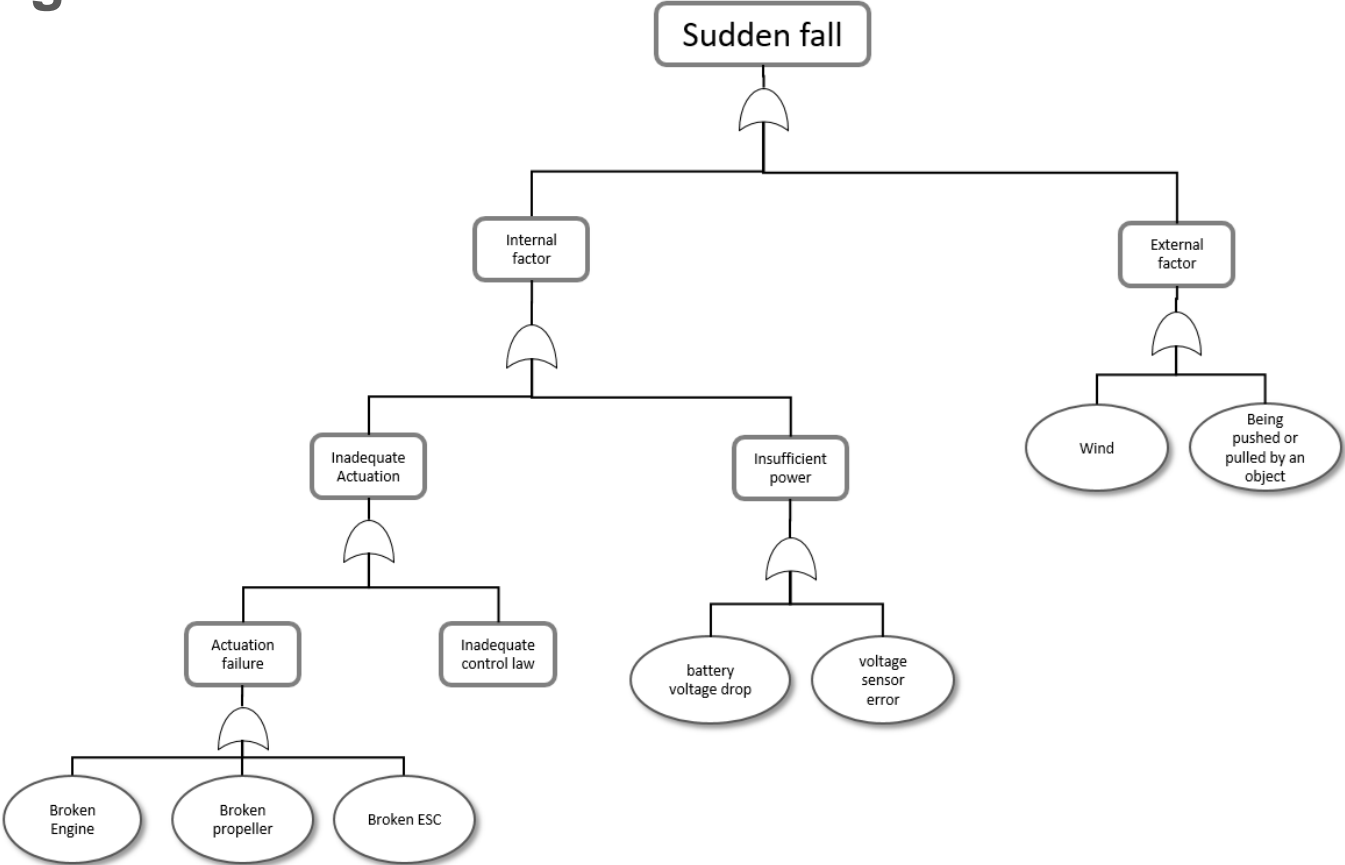
1. Four variable model
- 2. Extracting causal chains of misbehaviors**
3. Building Fault trees



Modeling behavior and misbehavior of drone

- 1. Four variable model
- 2. Extracting causal chains of misbehaviors

3. Building Fault trees



Causality analysis: deriving fault trees from four variable model

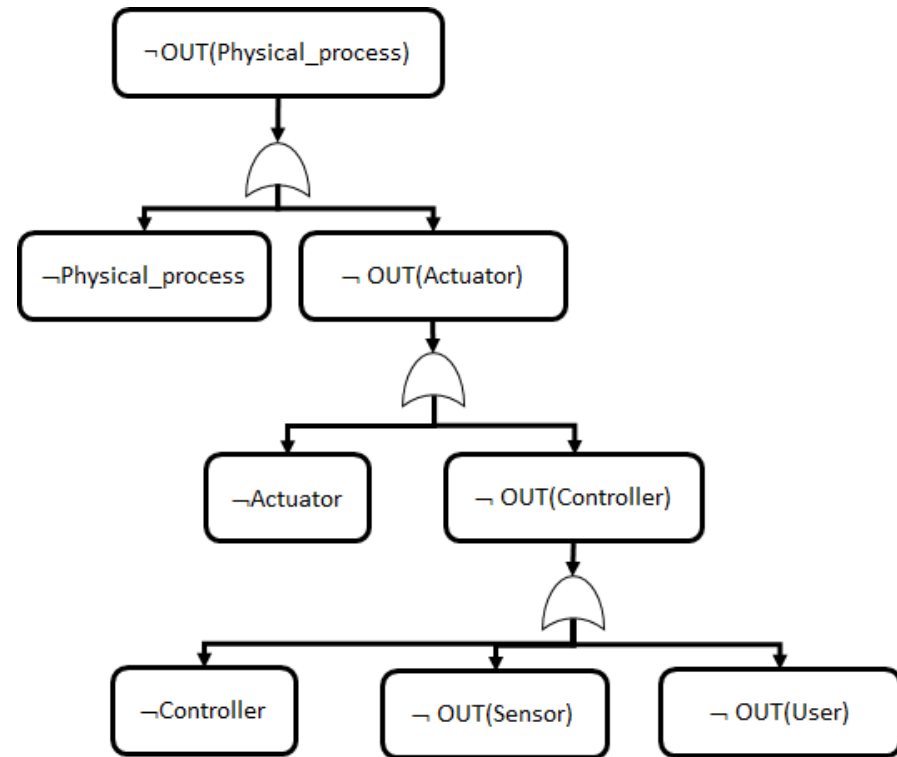
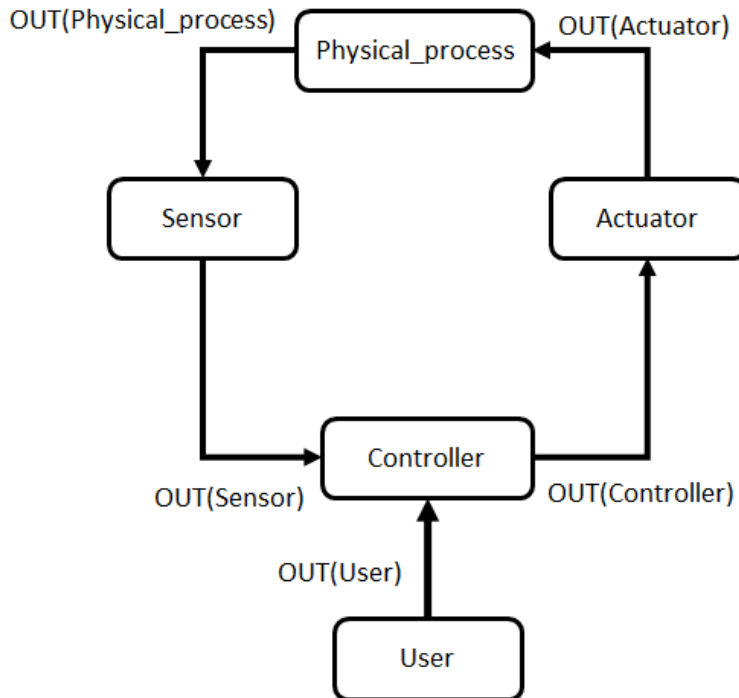
Moving backward in time windows:

$\neg \text{OUT}(\text{Physical_process}) \Rightarrow \neg \text{Physical_process} \vee \neg \text{OUT}(\text{Actuator})$

$\neg \text{OUT}(\text{Actuator}) \Rightarrow \neg \text{Actuator} \vee \neg \text{OUT}(\text{Controller})$

$\neg \text{OUT}(\text{Controller}) \Rightarrow \neg \text{Controller} \vee \neg \text{OUT}(\text{User}) \vee \neg \text{OUT}(\text{Sensor})$

First formula means that if the output of a physical process is abnormal then either the physical process itself or its input is abnormal

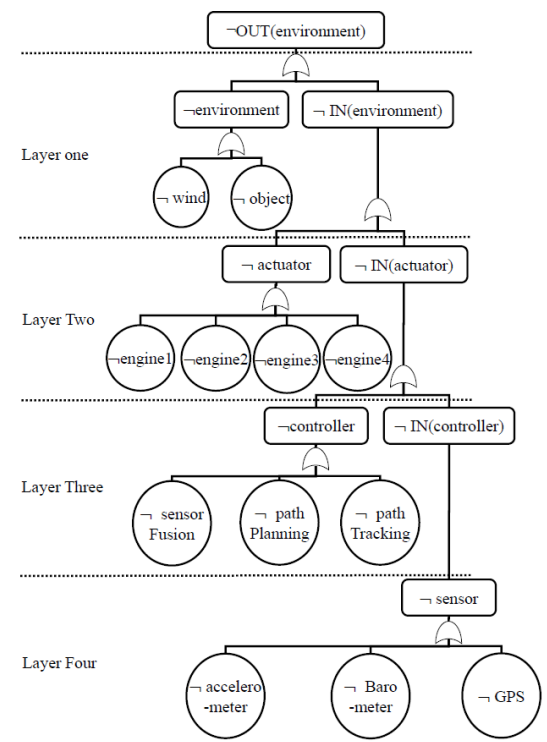
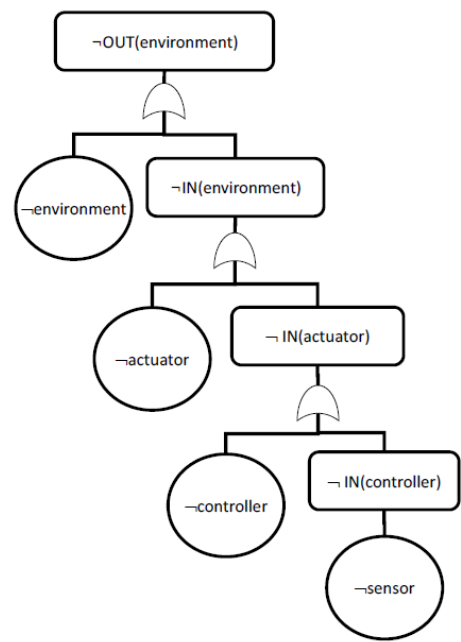
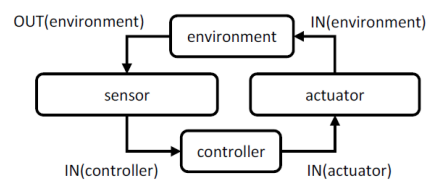


Knowledge based approach

Input: instances of each component

Output: fault tree

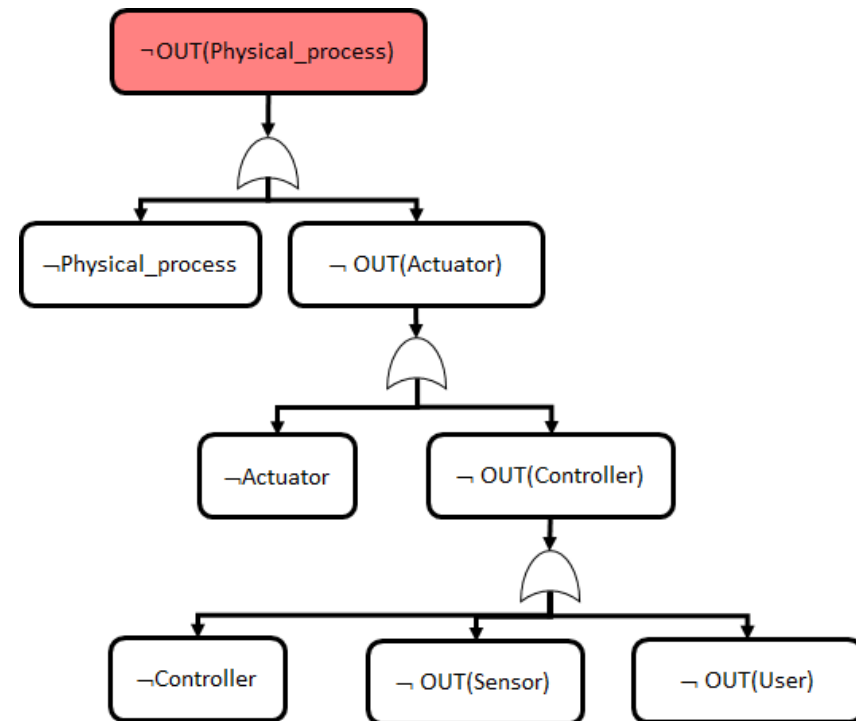
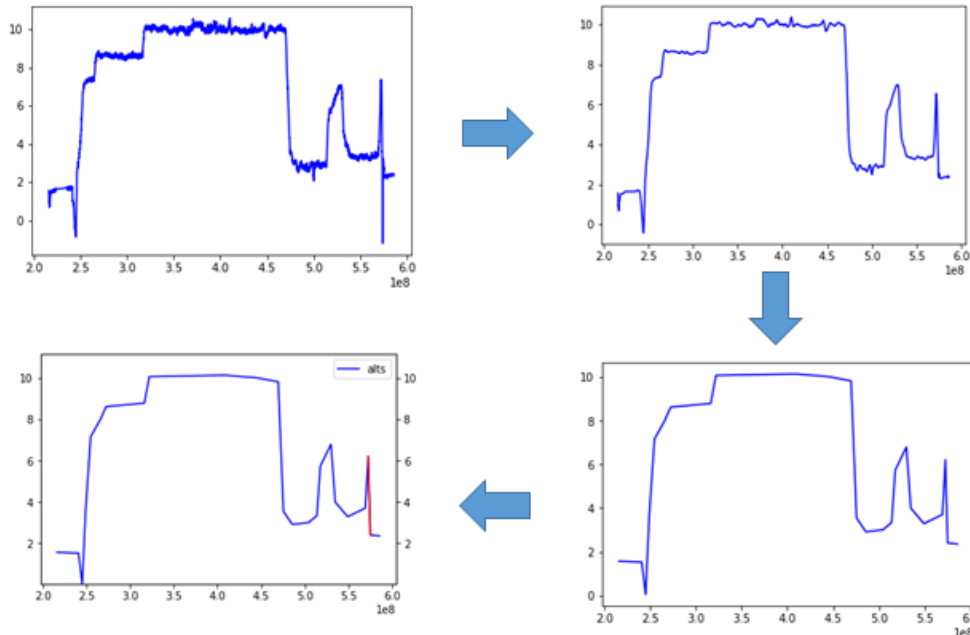
Issues: coupling of subcomponents (later)



Causality analysis: walking through the fault tree

Detecting $\neg\text{OUT}(\text{Physical_process})$, i.e. a fall from sky:

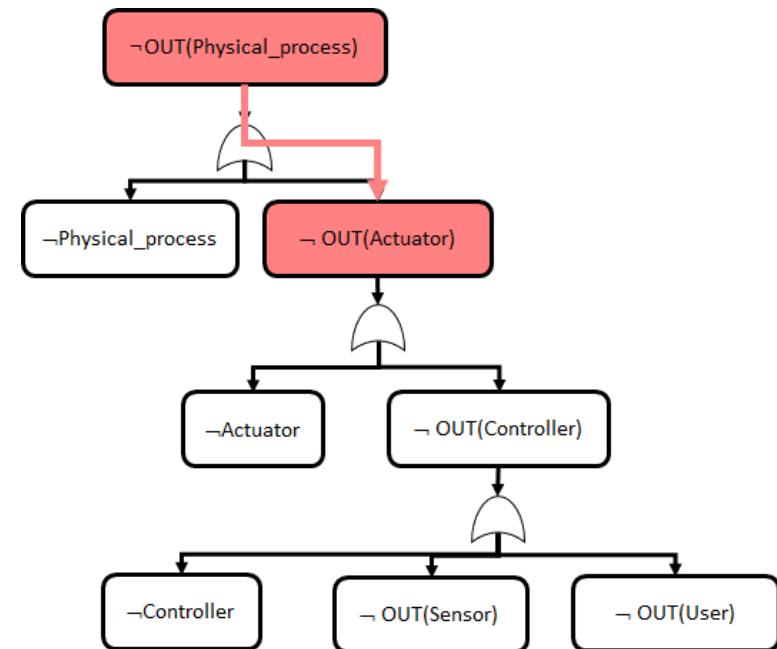
- Filtering the altitude signal from high frequencies
- Segmenting the filtered signal
- Searching for a segment with the negative slope less than threshold
- Finding the time stamp of the falling segment



Causality analysis: walking through the fault tree

Checking for \neg OUT(Actuator)

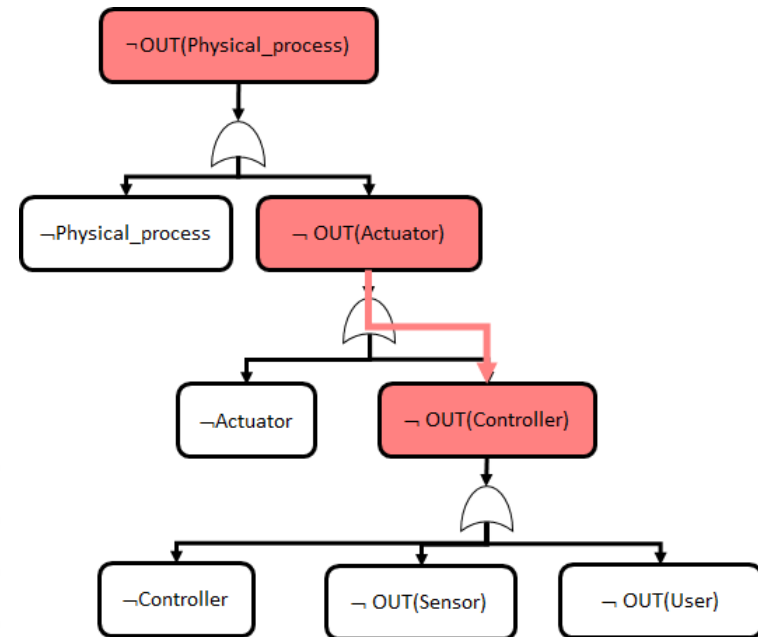
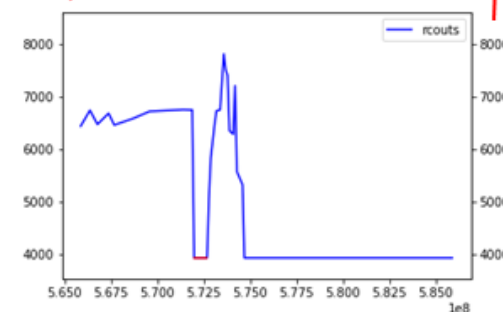
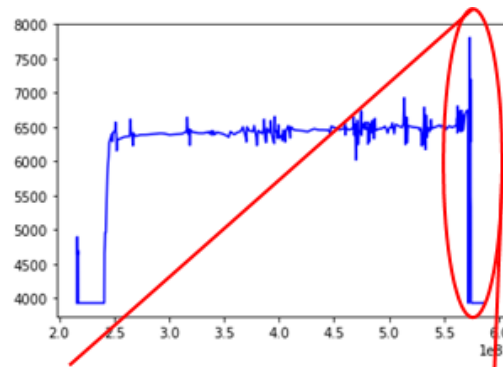
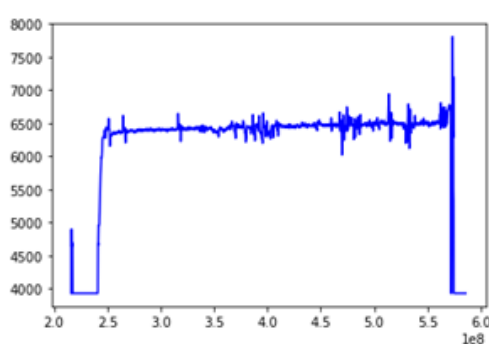
- We can assume that physical_process for example earth's gravitational force acts as expected always.
- As a result \neg Physical_Process cannot be the cause and the \neg OUT(Actuator) is the cause.
- \neg OUT(Actuator) means that output of engine contributed to the falling from the sky.



Causality analysis: walking through the fault tree

Checking for \neg OUT(Controller)

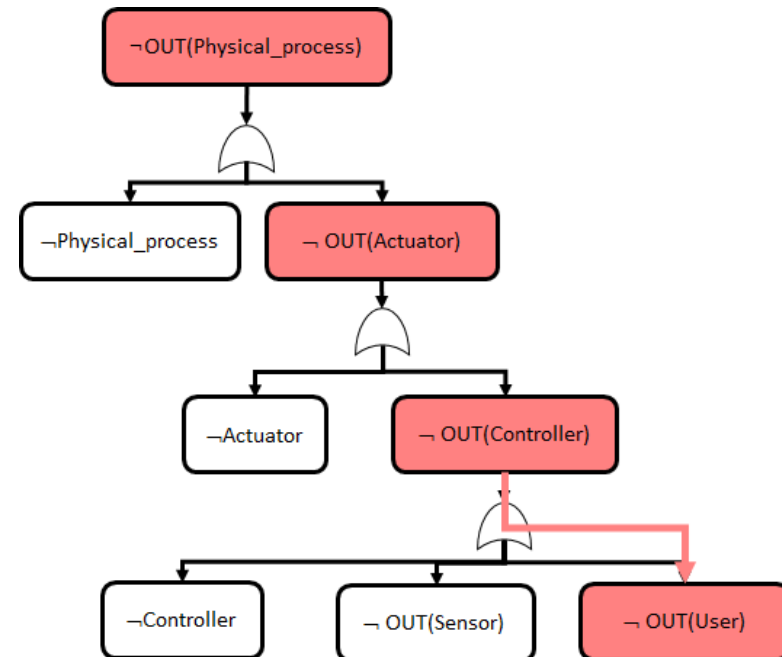
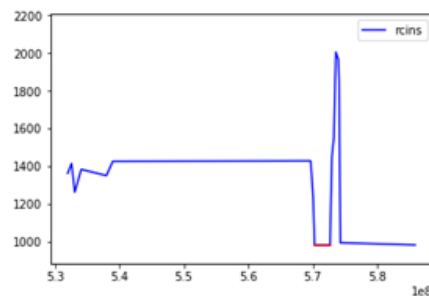
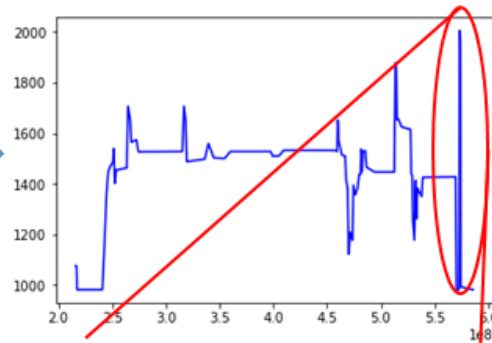
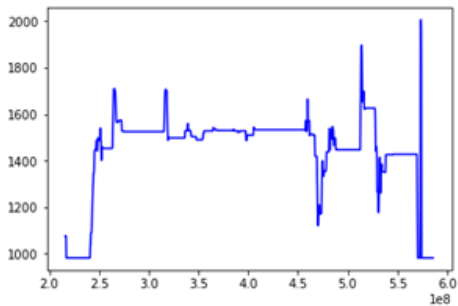
- The output of the controller is the desired engine speed
- Segmenting the desired engine speed
- Searching for a segment concurrent with the falling segment
- If the commands to the actuator are stuck at min then \neg OUT(Controller) is the cause to the \neg OUT(Actuator)
- Finding the time stamp of the minimum engine speed segment



Causality analysis: walking through the fault tree

Checking for \neg OUT(User)

- The output of the user is the radio control signal
- Segmenting the radio control signal
- If the radio control signals is stuck at min then \neg OUT(User) is the cause to the \neg OUT(Controller)
- Reached to a basic event in the fault tree. The user is the root cause for falling from the sky



Nice. And the point?

Fault tree templates can be generated

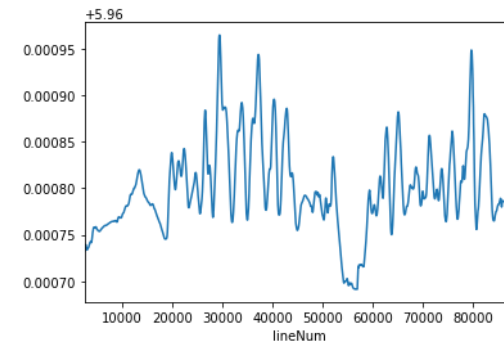
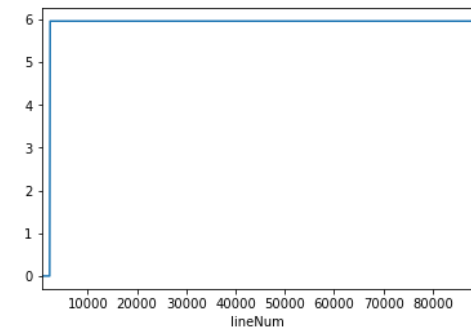
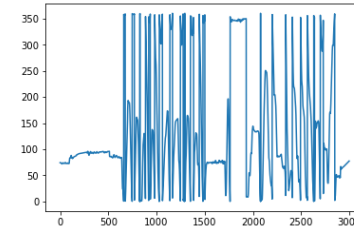
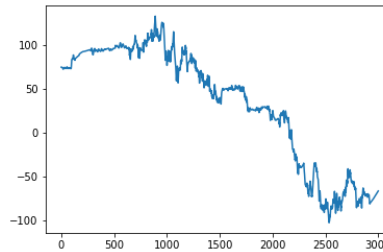
A-priori fault tree analysis often is done anyway - re-use these models!

Note that we used a model for type causality to infer actual causality (!)

Note that we didn't connect different components yet.

Data-driven Approach

- Painful!
- Format problems
- Different names in different versions
- Writing failure due to weak processing power
- Logging setting



```
FMT, 133, 75, MSG, QZ, TimeUS,Message
FMT, 134, 39, RCIN, QHHHHHHHHHHHHHH, TimeUS,C1,C2,C3,C4,C5,C6,C7,C8,C9,C10,C11,C12,C13,C14
FMT, 135, 39, RCOU, QHHHHHHHHHHHHHH, TimeUS,C1,C2,C3,C4,C5,C6,C7,C8,C9,C10,C11,C12,C13,C14
FMT, 136, 15, RSSI, Qf, TimeUS,RXRSSI
FMT, 138, 33, BARO, QffcfIf, TimeUS,Alt,Press,Temp,CRT,SMS,Offset
FMT, 139, 21, POWR, QffH, TimeUS,Vcc,VServo,Flags
CTRL, 255451202, 0.01974383, 0.00643052, 0.0338913, 0.01143248, 0.1100013
GPS, 255468462, 4, 557730600, 1939, 18, 0.62, 49.9661767, 36.0878342, 104.15, 0.4920366, 322.4314, -1.64, 1
GPA, 255468462, 0.92, 0.41, 0.56, 0.14, 1, 255468
IMU, 255471038, -0.158106, 0.1613381, 0.1375216, 0.8622345, 0.1995541, -9.276693, 0, 0, 21.74792, 1, 1
IMU2, 255471038, -0.1477846, 0.1616135, 0.1135721, 1.007267, 0.2173223, -10.417, 0, 0, 23.75, 1, 1
IMU, 255510413, -0.1393237, 0.1649317, 0.07158431, 0.7452759, 0.3478886, -7.924125, 0, 0, 21.7036, 1, 1
IMU2, 255510413, -0.1318613, 0.165016, 0.05779881, 1.029644, 0.3363889, -9.058994, 0, 0, 23.625, 1, 1
MSG, 255542884, EKF2 IMU1 ground mag anomaly, yaw re-alignedd
```

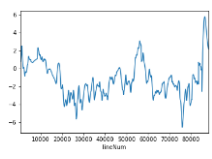
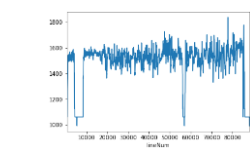
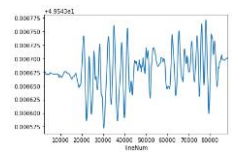
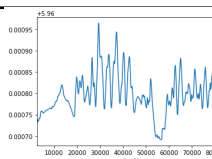
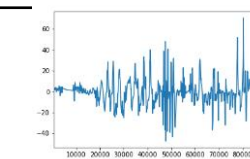
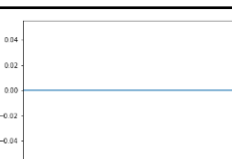
```
FMT, 134, 23, RCOU, Ihhhhhhhh, TimeMS,Chan1,Chan2,Chan3,Chan4,Chan5,Chan6,Chan7,Chan8
```

```
FMT, 136, 39, RCOU, QHHHHHHHHHHHHHH, TimeUS,C1,C2,C3,C4,C5,C6,C7,C8,C9,C10,C11,C12,C13,C14
```

Data-driven Approach

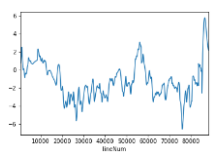
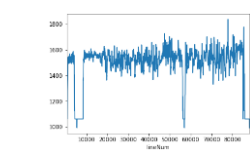
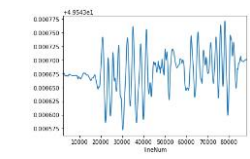
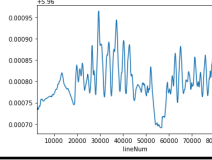
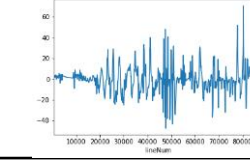
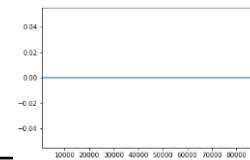
- Samples
- Variables



Flight log #	Var 1	Var 2	...	Var n
1				
...				
~2500				

Data-driven Approach (continued)

- Assumption: Causality = functional connectivity, $y = f(x)$
- The function can be logical or algebraic or differential
- Spurious correlation → “Relevant” correlation → type causality → actual causality
- Continuous analysis vs. discrete analysis: (Component, Action)
- Observational data vs. experimental data
- Type Causality : from correlation --- from regression --- from classification

Flight log #	Var 1	Var 2	...	Var n
1				
...				
~2500				

Correlation Analysis

1. Pearson, Karl [1909] proposes the correlation coefficient
2. Reichenbach [1956] proposes principle of the common cause:

“If events X and Y are correlated, then either X caused Y, Y caused X, or X and Y are joint effects of a common cause (one that renders X and Y conditionally probabilistically independent).”
3. Counterexample:
Sober [1994] shows that bread prices in Britain and sea level in Venice are correlated!
4. PC Algorithm [2000]: causal graph discovery based on conditional independence
5. Counterexample rejected!
Hoover, Kevin D. [2003]: It was because of non-stationary example!

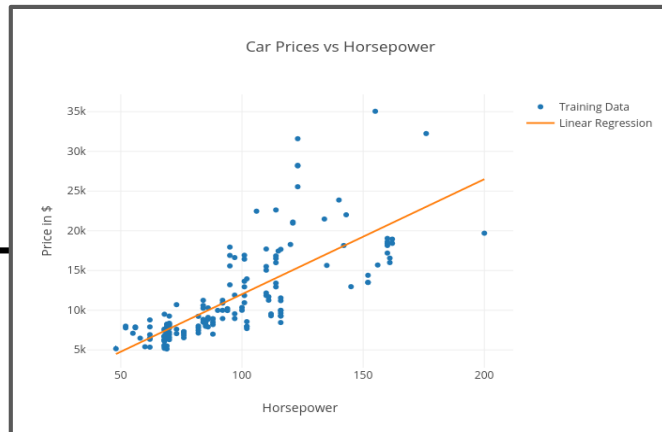
Causal graph from Autoregression

- Granger: Type causality using Autoregression

- Exogenous var: X , Endogenous var = Y
- model: $Y = a * X + e$
- Multiple regression: $Y = a_1 * X_1 + a_2 * X_2 + \dots + e$
- Vector regression:

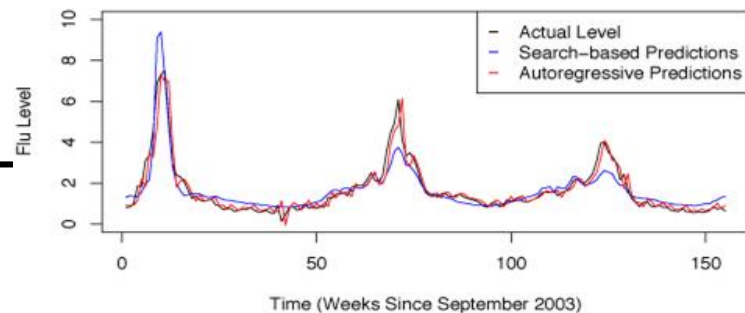
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$



- Exogenous var: $Y(t-1)$, Endogenous var = $Y(t)$
- model: $Y(t) = a * Y(t-1) + e$
- Multiple autoregression:
 $Y(t) = a_1 * Y(t-1) + a_2 * X(t-1) + \dots + e$
- Vector Autoregression (VAR):

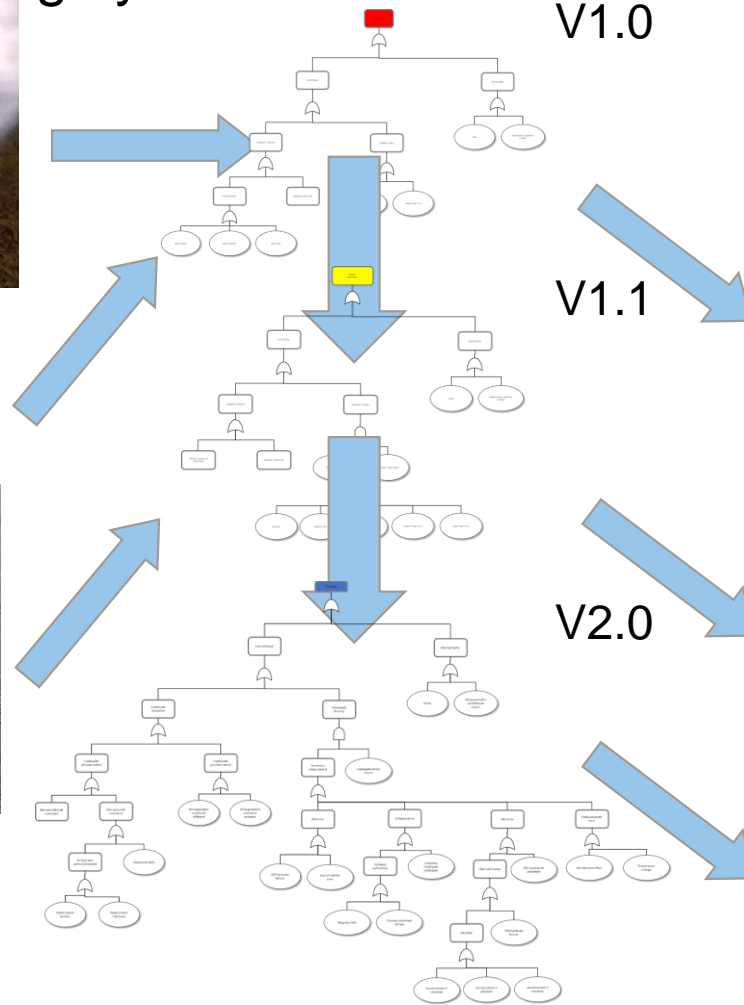
$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{y,t} \\ \varepsilon_{x,t} \end{pmatrix}$$



A constantly improving system...



...



...



...

What traditional fault trees and cut set analysis cannot do ...

Analyse for absence of event or arbitrary combinations of events; identify just one cause

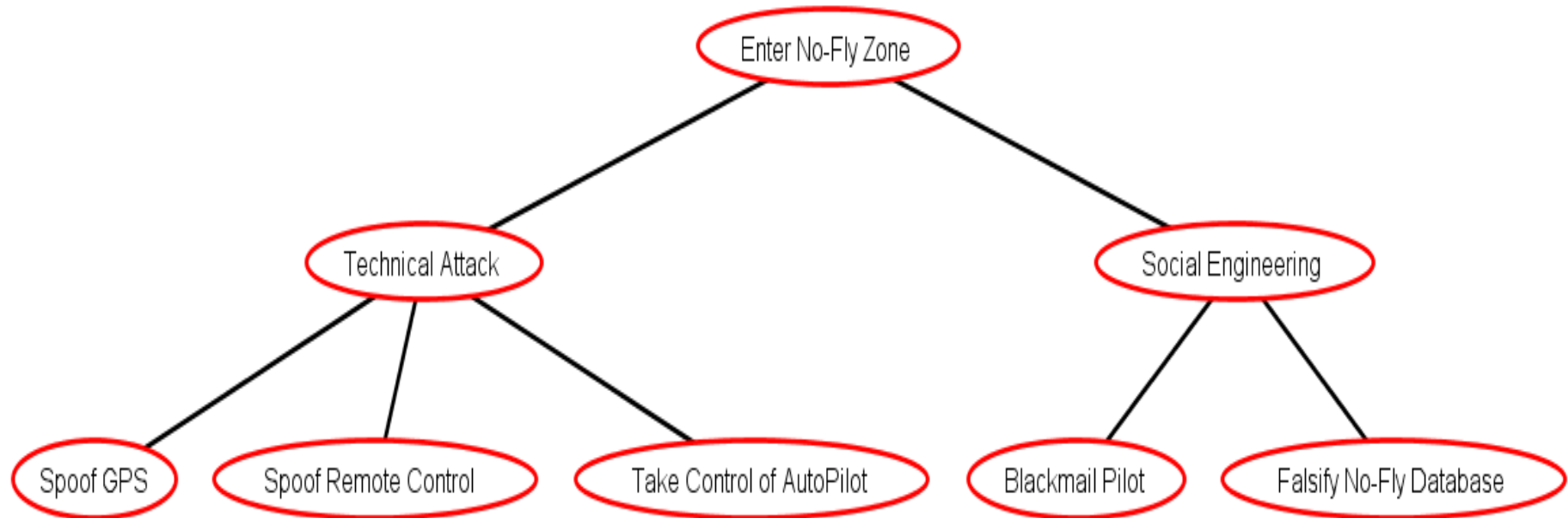
Counterfactual reasoning

[Encode temporal relationships (Reif et al.)]

[Encode preemptions (inhibitor nodes)]

Attack Trees

- Describe potential security threats and the steps necessary to successfully perform
- Hierarchical tree representation



Causal Models: Example Attack-Defense Trees

Very similar to fault trees: top level event denotes compromise of CIA of some asset. Nodes are steps in attack

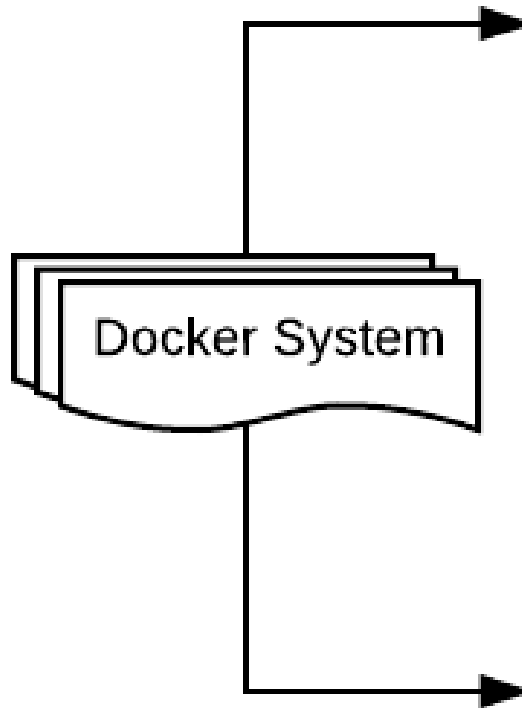
Formalization by Mauw (2005): not exactly a propositional formula

Can be used for assigning responsibility for insider attacks

Idea: Automatic Generation of Attack Graphs

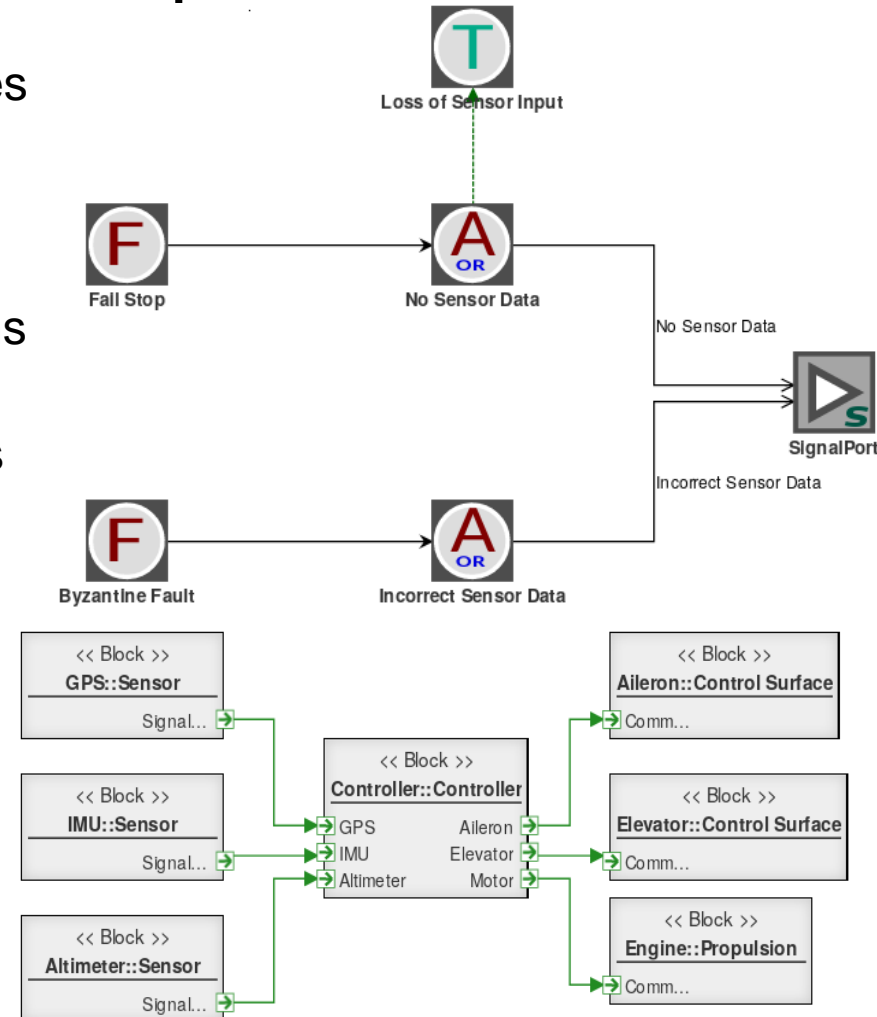
- Build on work in computer networks
- Hosts → Containers
- Physical host link → dependencies between containers
- Containers are not completely isolated
- No model checking, No graph goal
- Four level of privileges – combination of access level (User, Admin) and access mode (container, host)
 - Pre-conditions/Post-conditions from manually selected rules systems
Aksu et al.

System Overview



Timed Failure Propagation Graphs

- DAG to model the failure propagation routes for typical systems:
 - Nodes represent either failure modes or discrepancies.
 - Edges represent the cause-effect relations
- Fault identification and mitigation
- Focus on functional failures of the system's hardware and software components.

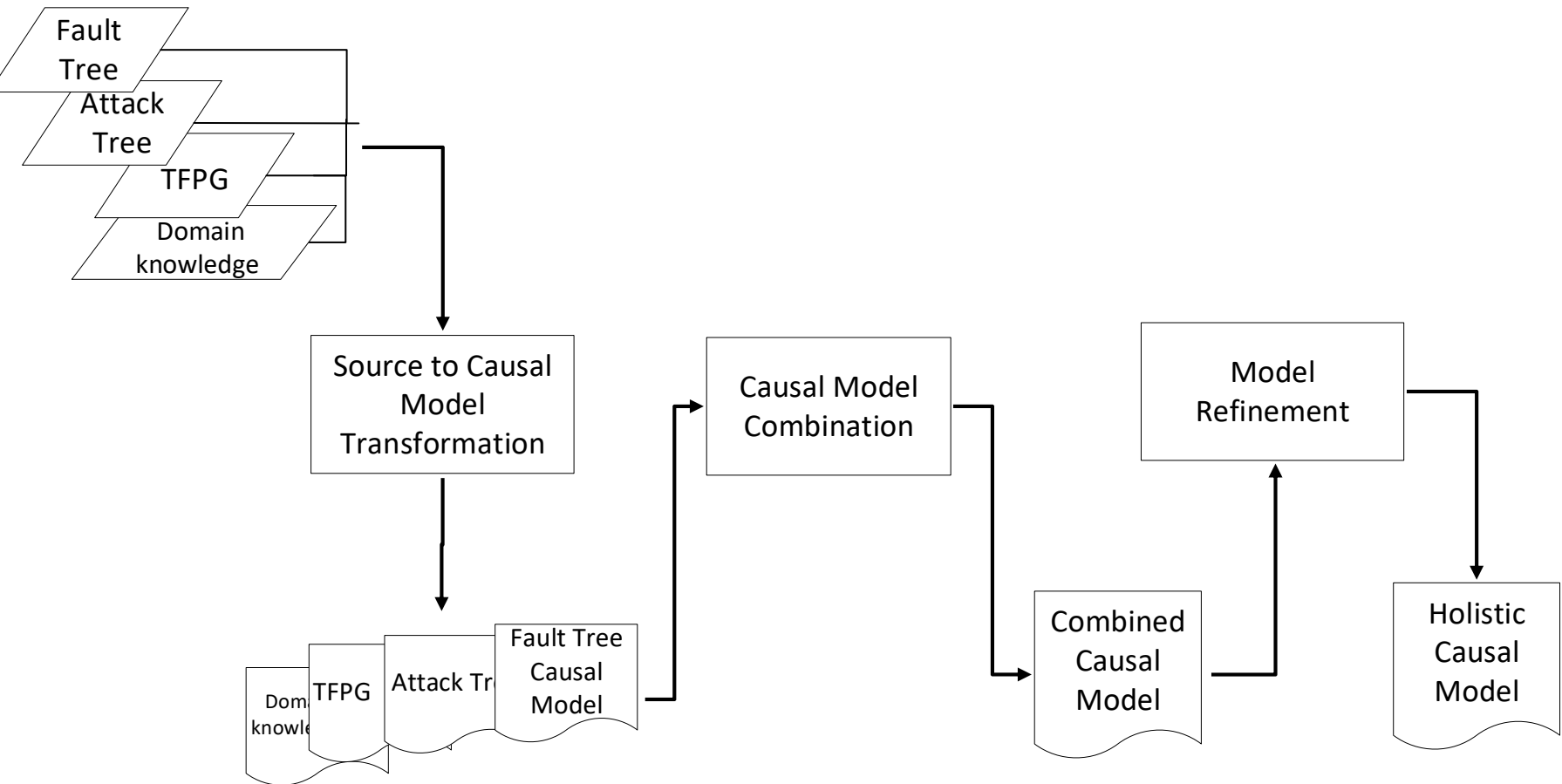


Combining Causal Models?

- Create a holistic causal model that incorporates different models focusing on different aspects, and possibly created by different teams.
 - Already existing cause-effect models (30 + models in threat modeling)
 - Used for risk assessment and mitigation during system design or run-time
 - Represent binary events, allow for logical combination
 - Acyclic
 - Can be automated

Amjad Ibrahim, Severin Kacianka, Alexander Pretschner, Charles Hartsell, Gabor Karsai:
Practical Causal Models for Cyber-Physical Systems. NASA Formal Methods 2019: 211-227

Overview of The Process

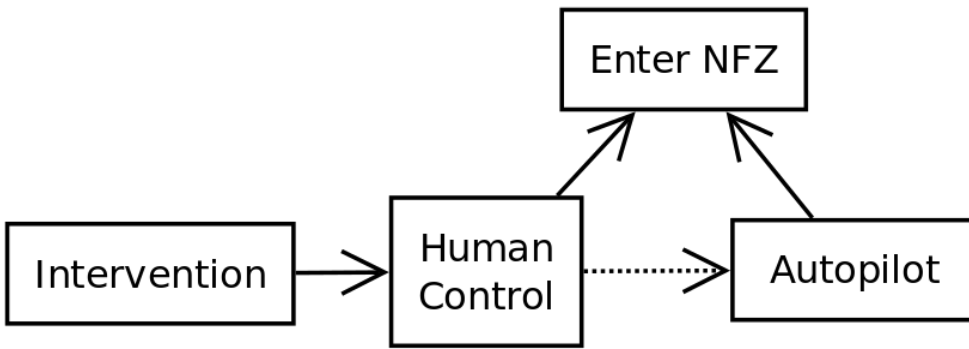


Combination

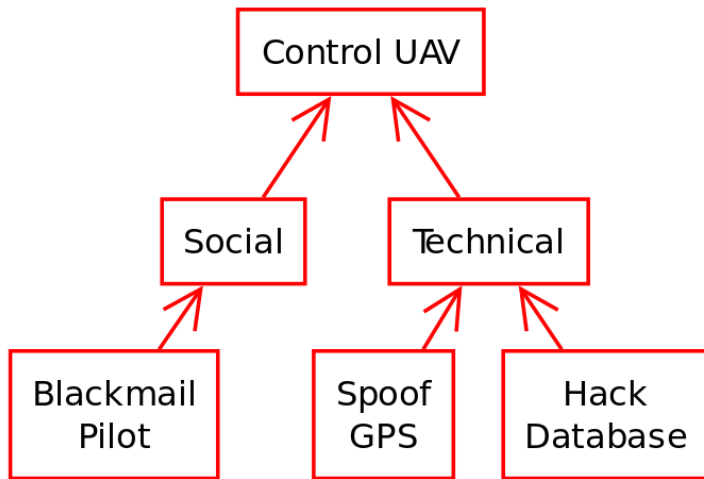
- Models are useful but have limits of their domains
- Use prior work by Alrajeh et al. (2018) and Friedberg and Halpern (2018) for combining causal models.
 - Intuition: If two models are *compatible* we can combine them
 - However: In many cases, we still need human input

Example

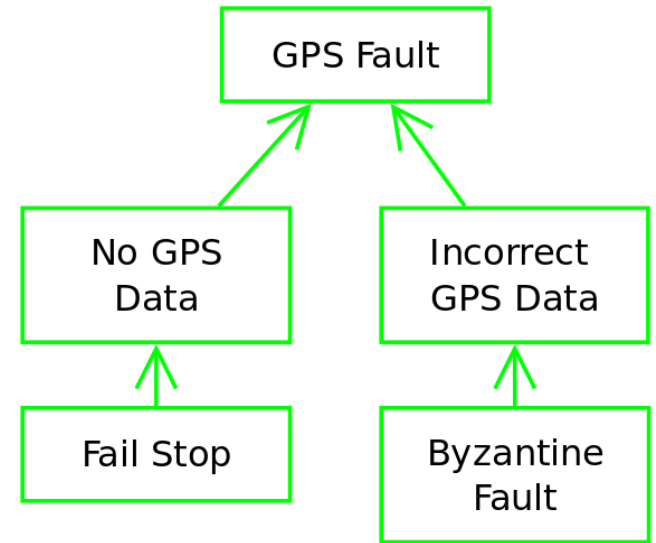
Domain Knowledge



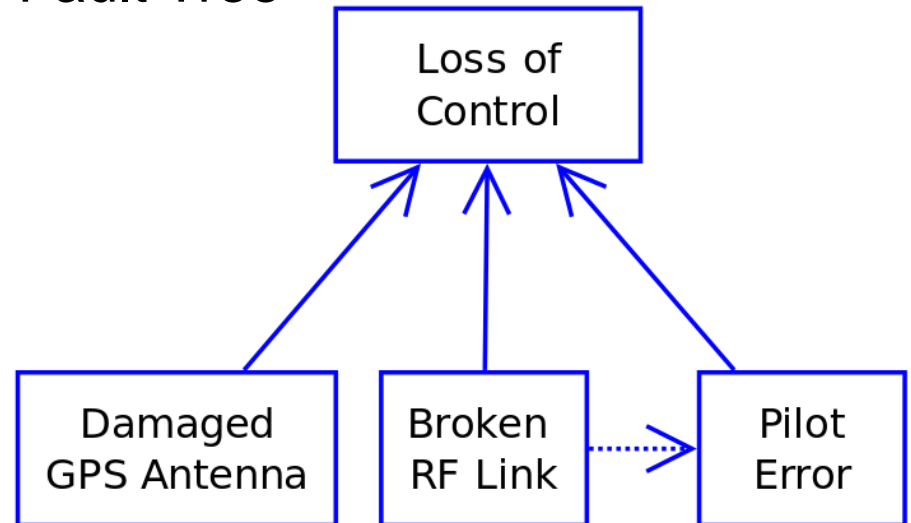
Attack Tree



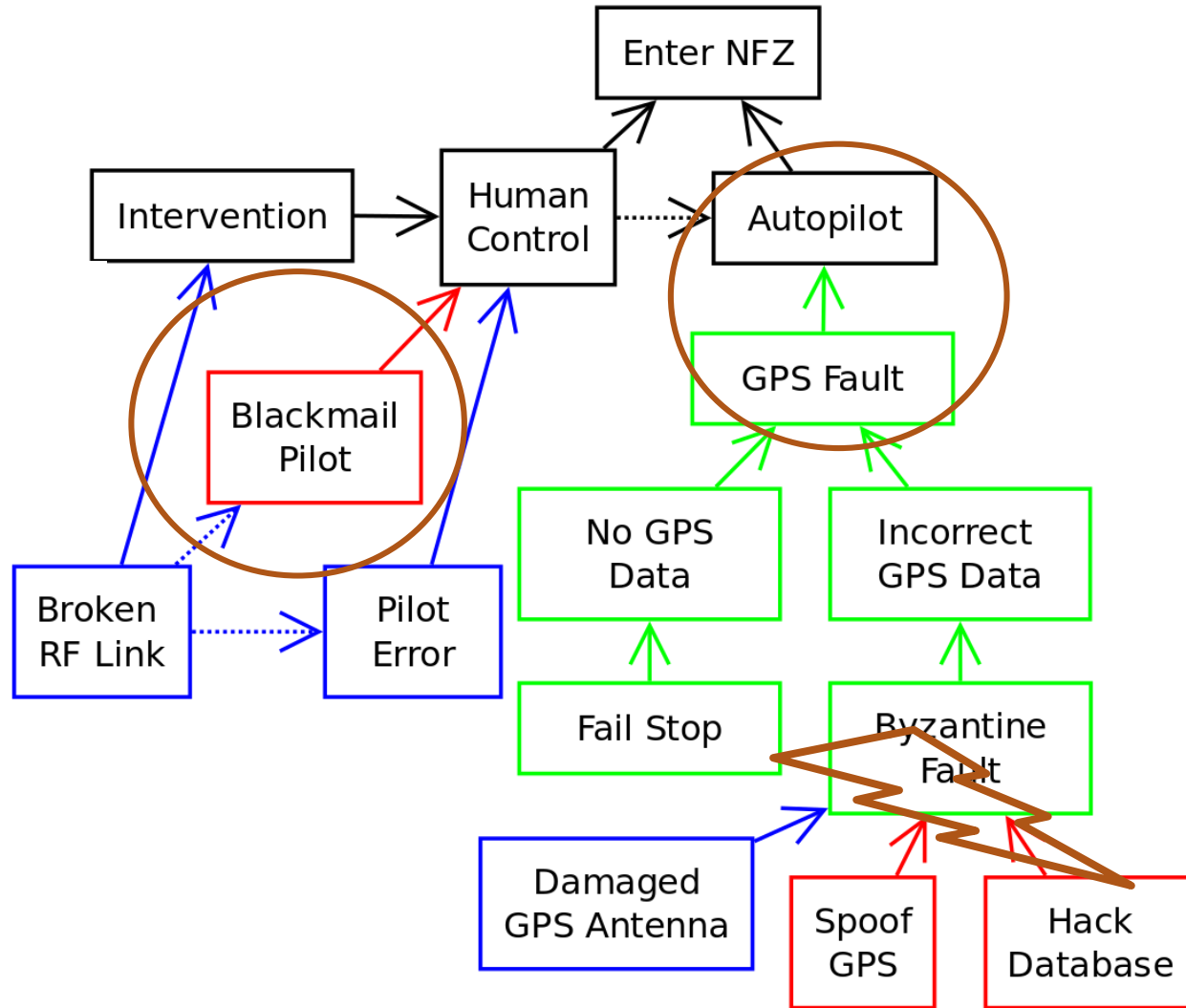
TFPG



Fault Tree



Example



Extension

Models of human behavior

Roadmap

Accountability

Causality

Introduction

Exemplary causal models

Flavors of causal reasoning

Halpern-Pearl Causality

Remember

There must be causal models to explain “causality”.

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y x)$	Seeing	What is? How would seeing X change my belief in Y ?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y do(x), z)$	Doing Intervening	What if? What if I do X ?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x x', y')$	Imagining, Retrospection	Why? Was it X that caused Y ? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smok- ing the past 2 years?

Flavors of Causality

Spectrum-Based Fault Localization

Granger Causality

Model-Based Diagnosis

Halpern-Pearl Causality

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y x)$	Seeing	What is? How would seeing X change my belief in Y ?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y do(x), z)$	Doing Intervening	What if? What if I do X ?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x x', y')$	Imagining, Retrospection	Why? Was it X that caused Y ? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Flavors of Causality

Spectrum-Based Fault Localization

Granger Causality

Model-Based Diagnosis

Halpern-Pearl Causality

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y x)$	Seeing	What is? How would seeing X change my belief in Y ?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y do(x), z)$	Doing Intervening	What if? What if I do X ?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x x', y')$	Imagining, Retrospection	Why? Was it X that caused Y ? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Flavors of Causality

Spectrum-Based Fault Localization

Granger Causality

Model-Based Diagnosis

Halpern-Pearl Causality

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y x)$	Seeing	What is? How would seeing X change my belief in Y ?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y do(x), z)$	Doing Intervening	What if? What if I do X ?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x x', y')$	Imagining, Retrospection	Why? Was it X that caused Y ? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Flavors of Causality

Spectrum-Based Fault Localization

Granger Causality

Model-Based Diagnosis

Halpern-Pearl Causality

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y x)$	Seeing	What is? How would seeing X change my belief in Y ?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y do(x), z)$	Doing Intervening	What if? What if I do X ?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x x', y')$	Imagining, Retrospection	Why? Was it X that caused Y ? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.