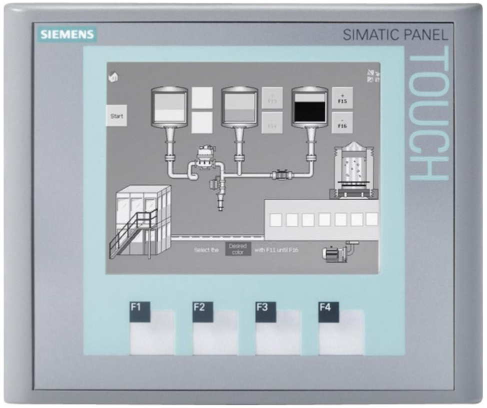
Machine Learning Based Intrusion and Anomaly Detection for SCADA

Improving current models with case specific information

Peter Prjevara - Dima van de Wouw - with the support of Deloitte

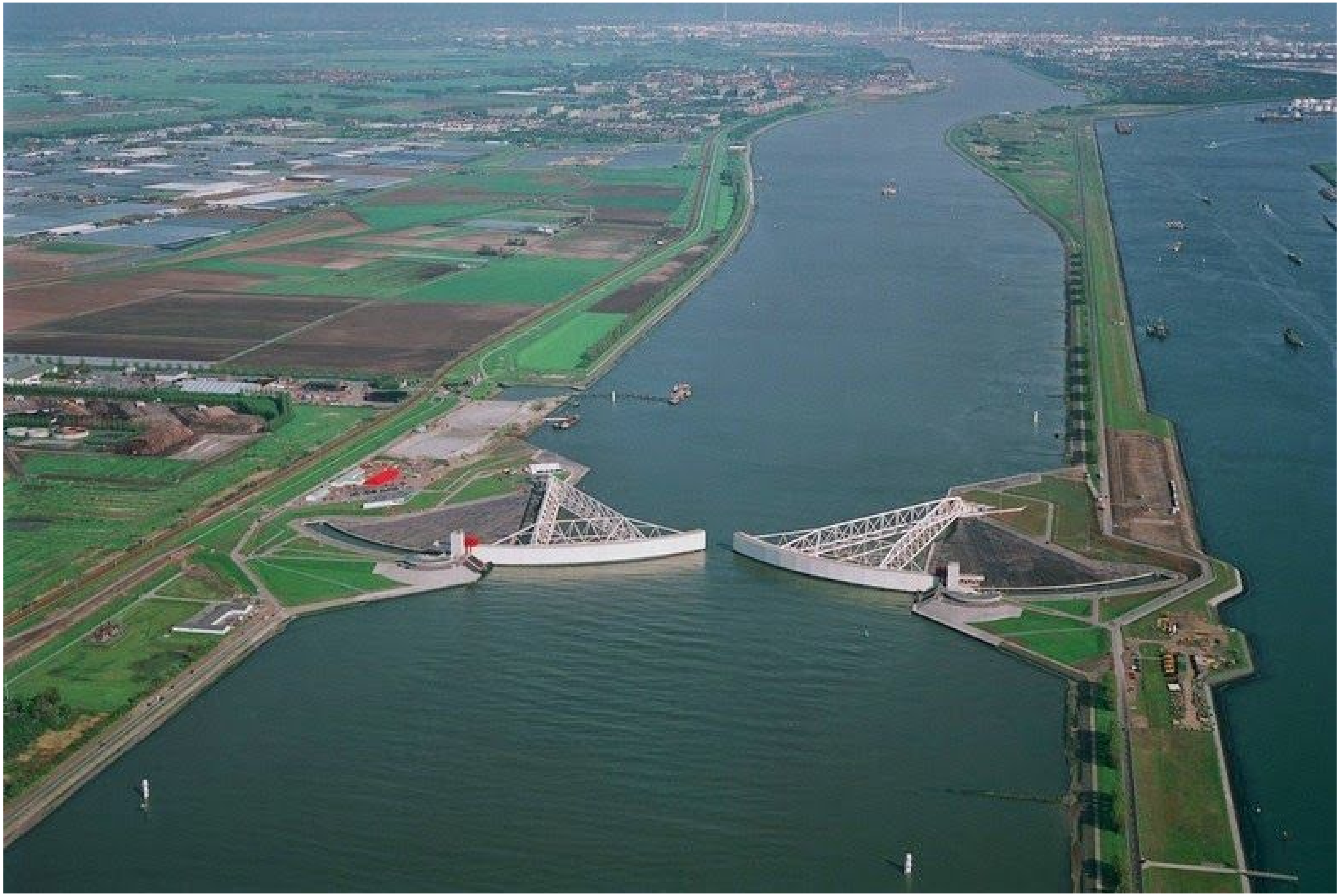
What is SCADA?

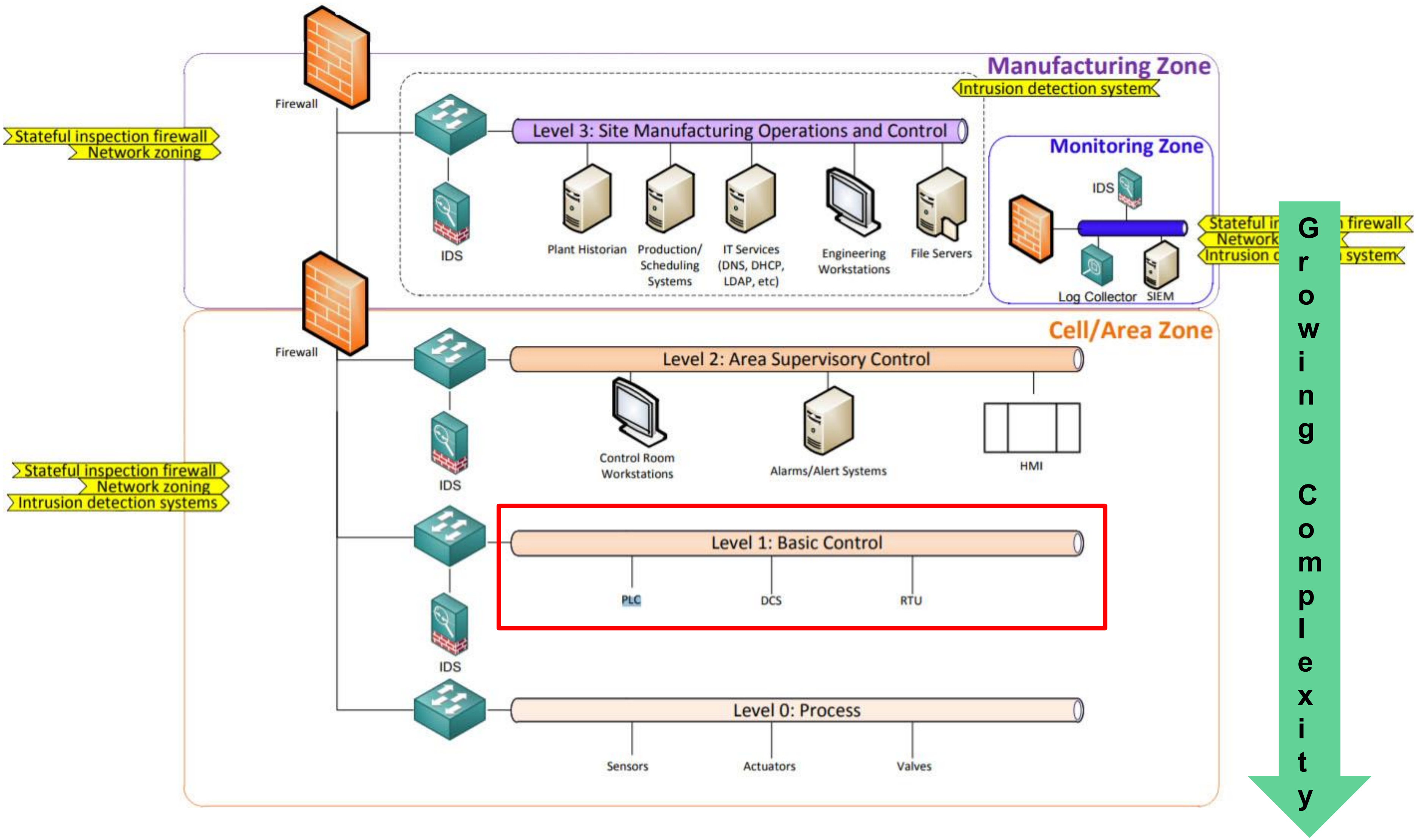
- SCADA = **S**upervisory **C**ontrol **A**nd **D**ata **A**cquisition



Bottle Filling Factory

Storm Surge Barriers





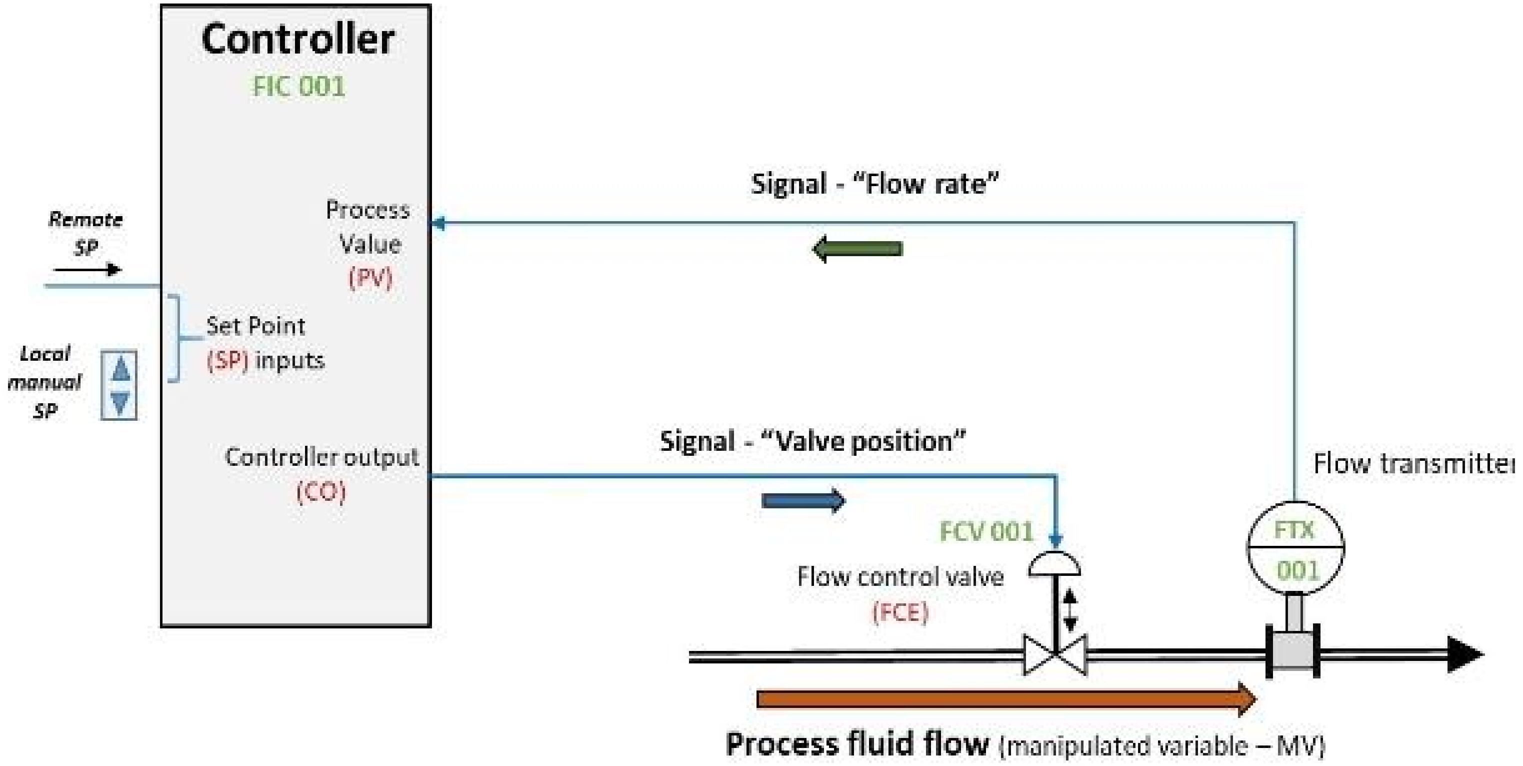
Source: Obregon, L. (2015). Secure architecture for industrial control systems. *SANS Institute InfoSec Reading Room*.

Some Terminology

- **I/O - Input / Output signals**
    - Commonly used term to refer to the signals related to the system
    - Analogue or Digital
- **Cycle Time** or **Scan Time**
    - How often the devices within the system are scanned by the control device
    - Expressed in Hz
    - Multiple different frequencies can be in SCADA simultaneously
- The **HMI -> PLC -> Device** relationship

Process Types

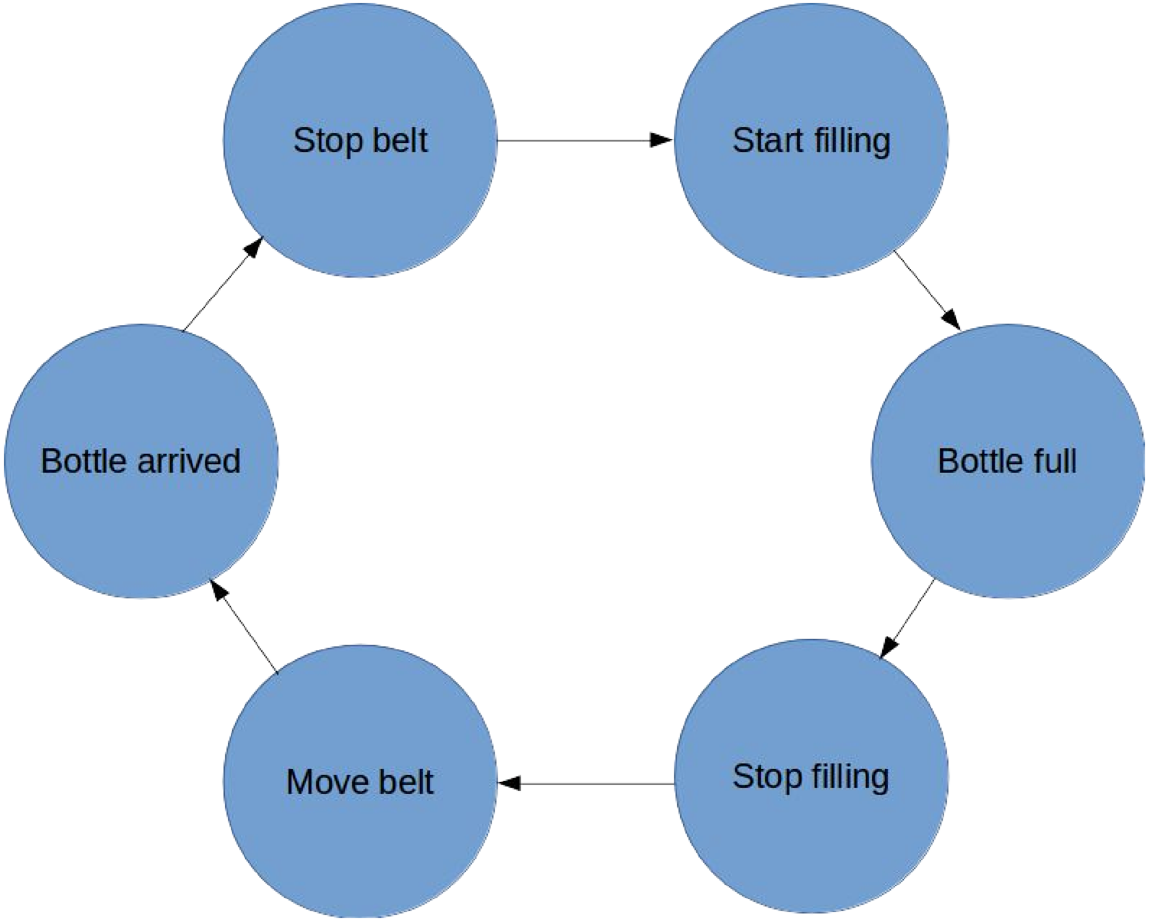
1. Self Regulating Processes



Source: https://en.wikipedia.org/wiki/Control_theory

Process Types

2. Sequential Processes



Relevance of Research

- Main priority of SCADA systems is availability
- Systems are often old
- Tailor made solutions are necessary - but how?
  - Standard security products can’t provide all encompassing protection
  - Most attacks are conducted from the **internal layers**
- **Damage can be significant**
  - Stuxnet (2010), Ukraine (2015), New York Dam (2016), Kemuri Water Company (2016)

Related Work

1. Fovino et al - State based intrusion detection system, 2010
2. Wool and Goldenberg - DFA based IDS, 2013
3. Caselli et al - Sequence aware detection, 2015
4. Wool and Goldenberg - DFA based multi layered IDS, 2017
5. Boukema and Lahaye - Comparison of ML Algorithms, 2017
6. Bengio et al - NN/HMM Hybrid - 1995
7. Alex Graves - Supervised Sequence Labelling, 2012

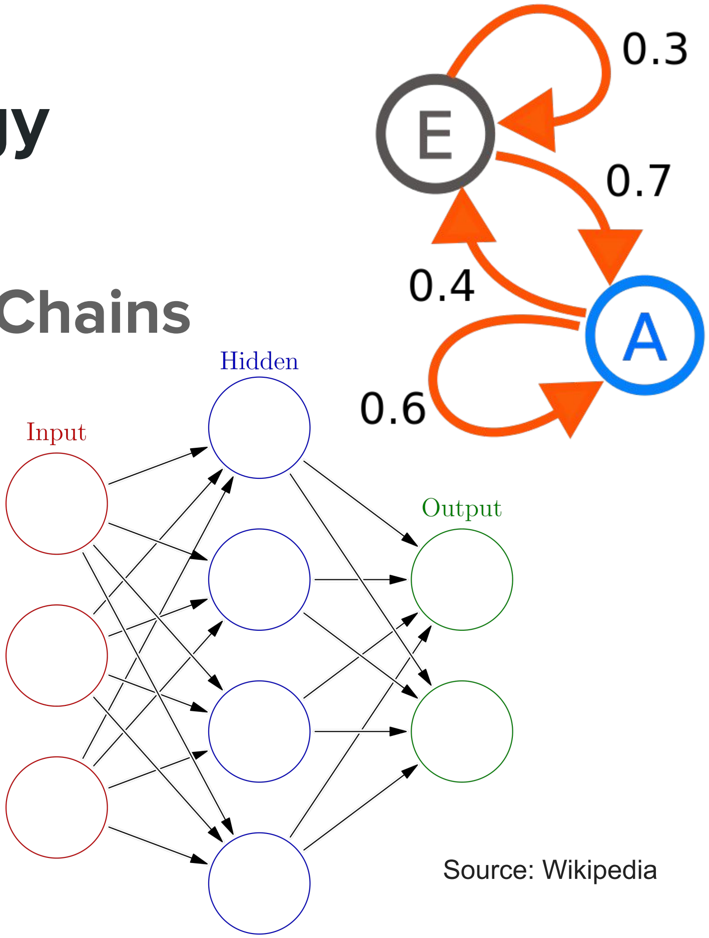
Research Questions

”What information can be used to complement the information generated by ML algorithms, to improve the efficiency and accuracy of a ML based IADS, and make it useful for sequential and non-sequential processes?”

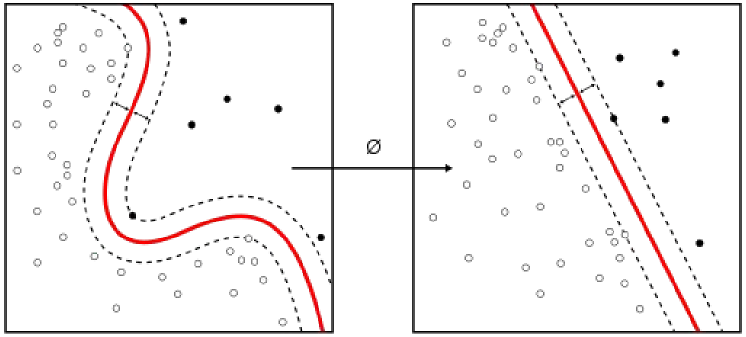
”How can this information be best combined with the ML algorithm?”

Machine Learning Terminology

- **DTMM - Discrete Time Markov Chains**
  - Similar to FSM (Finite State Machines)
- **Neural Networks**
  - LSTM - Long Short Term Memory
  - HMM - Hidden Markov Models
  - Commonly used for image recognition
- **SVM - K-Means Clustering**
- **Hybrid Models**



Source: Wikipedia



Problem Definition

- Most focus is on sequential anomaly detection
- Hybrid Machine Learning systems are effective, but not applied to the field of IDS for SCADA
- Researches are difficult to reproduce
  - test environments vary

“Finally, it is worth noting that leveraging semantic of ICS communications and parameters is a powerful way to enhance security tools’ knowledge of the environment in which they are deployed and, therefore, improve their effectiveness.” - Caselli et al, 2015

Defining the “Full Knowledge”

- From the research of Fovino et al
    - Device names, device type, possible states + fault tolerance
- Previous experience and discussions
    - I/O list should be available?
    - Known sequences, logic diagrams?
    - Length of sequences?
    - Logical groups?
    - Causal relationships?
    - Process types?

Distilled List

- Exclude what can be learned through sniffing
  - Digital or Analogue
  - Protocol
  - Sequences
  - Process types

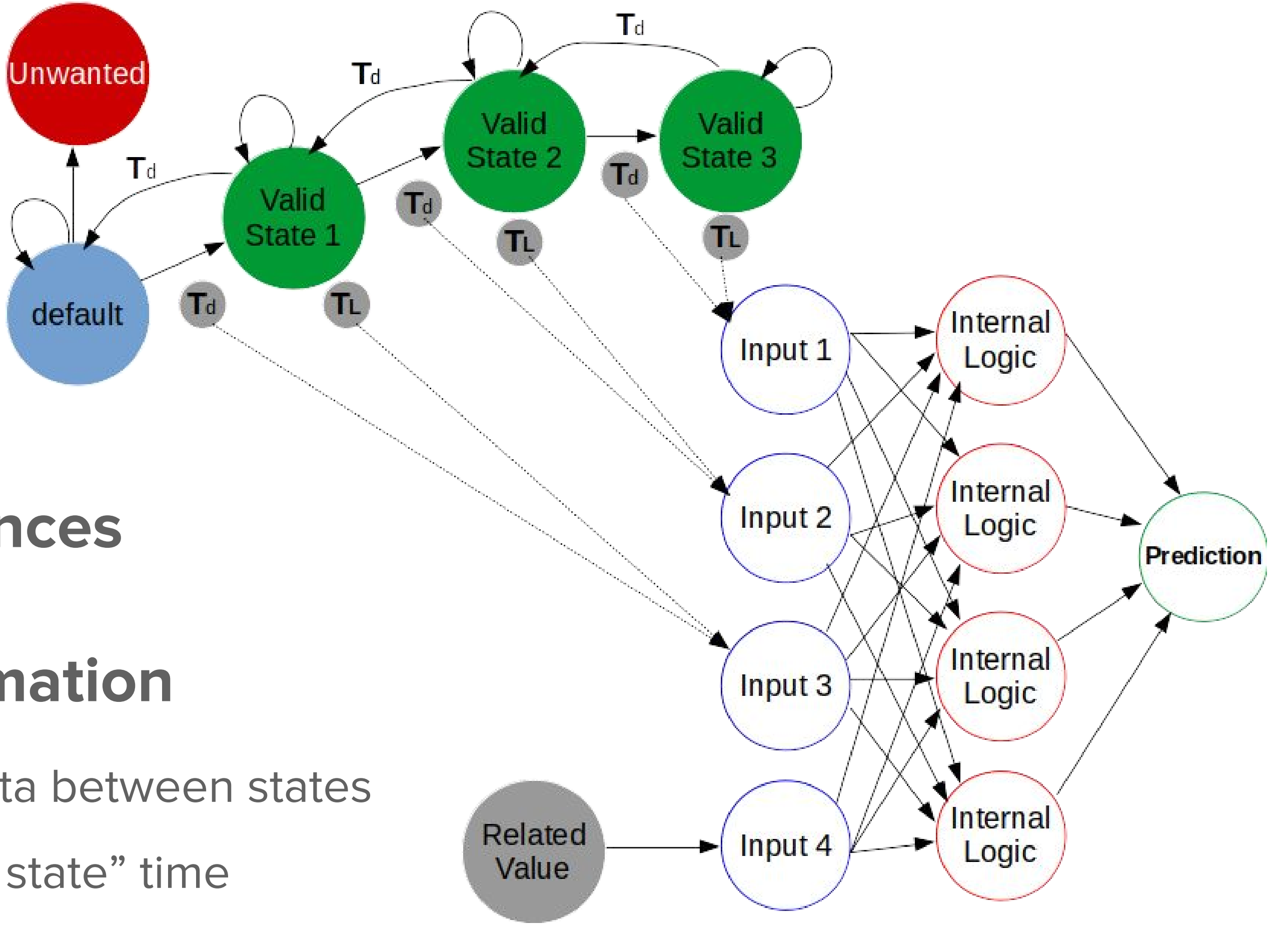
- Include what can’t be learned
  - List of signals
  - Logical Groups / Correlations
  - Error threshold for analogue
  - Age of equipment
  - **Irrelevant Information - display values, not in logic**

Improvements Proposed

- Analyse Traffic Based on Logical Relationships
- Combine Different Machine Learning Models
  - Learn the characteristics of sequential processes
  - Learn the characteristics of non-sequential processes
- Correlate the Gathered Data
- Enable Features Dynamically, As-Needed

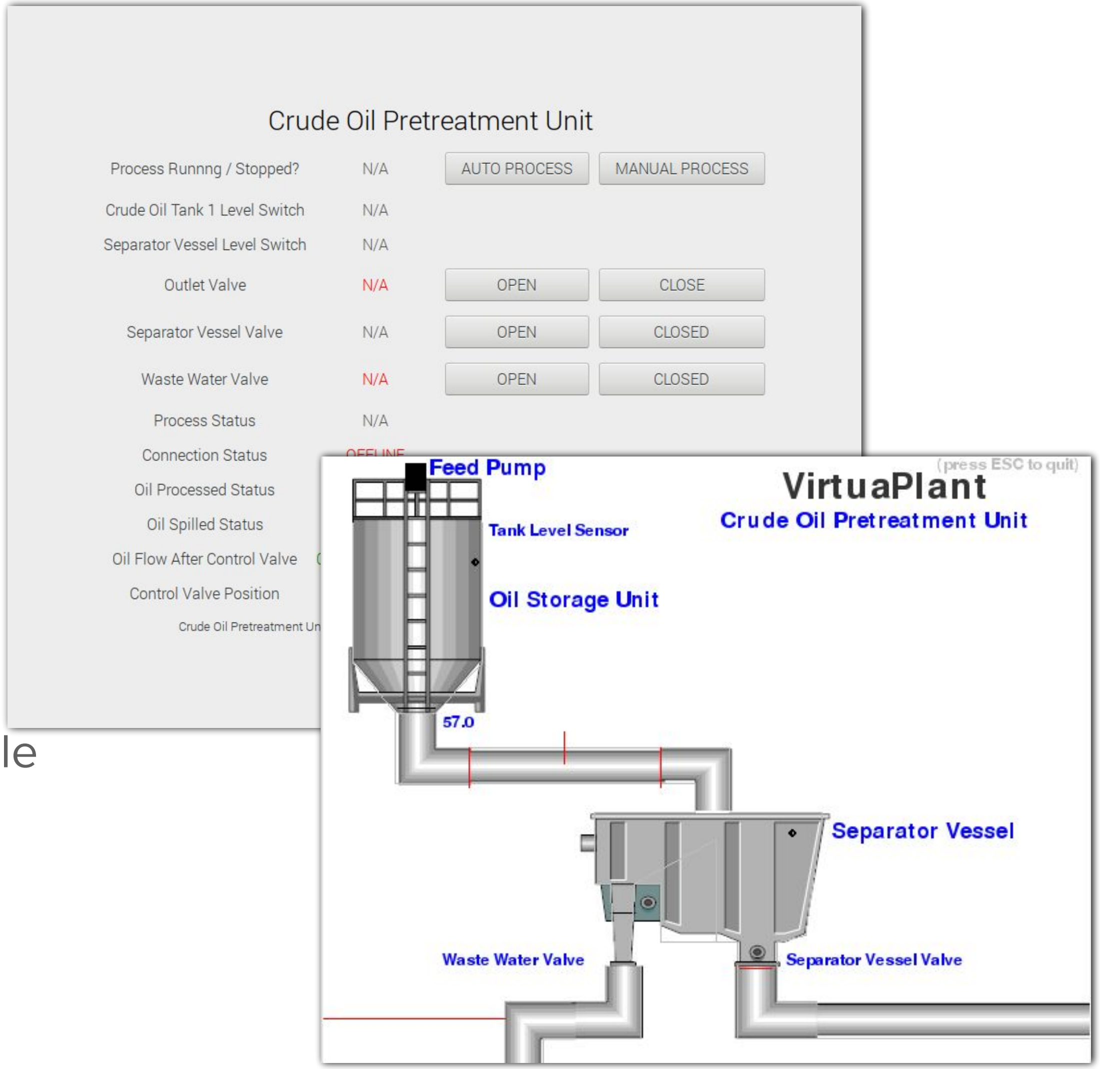
The Model

- I/O List
- Logical Groups
- Learned Sequences
- Correlate Information
  - Record Time Delta between states
  - Record “In same state” time
  - Feed this to a neural network with related information



Test Environment

- Modbus based
- Simple logic
  - Tank full -> valve opens
  - Flow control in the middle
- **Reproducible**
- **Portable**
- Other protocols ???



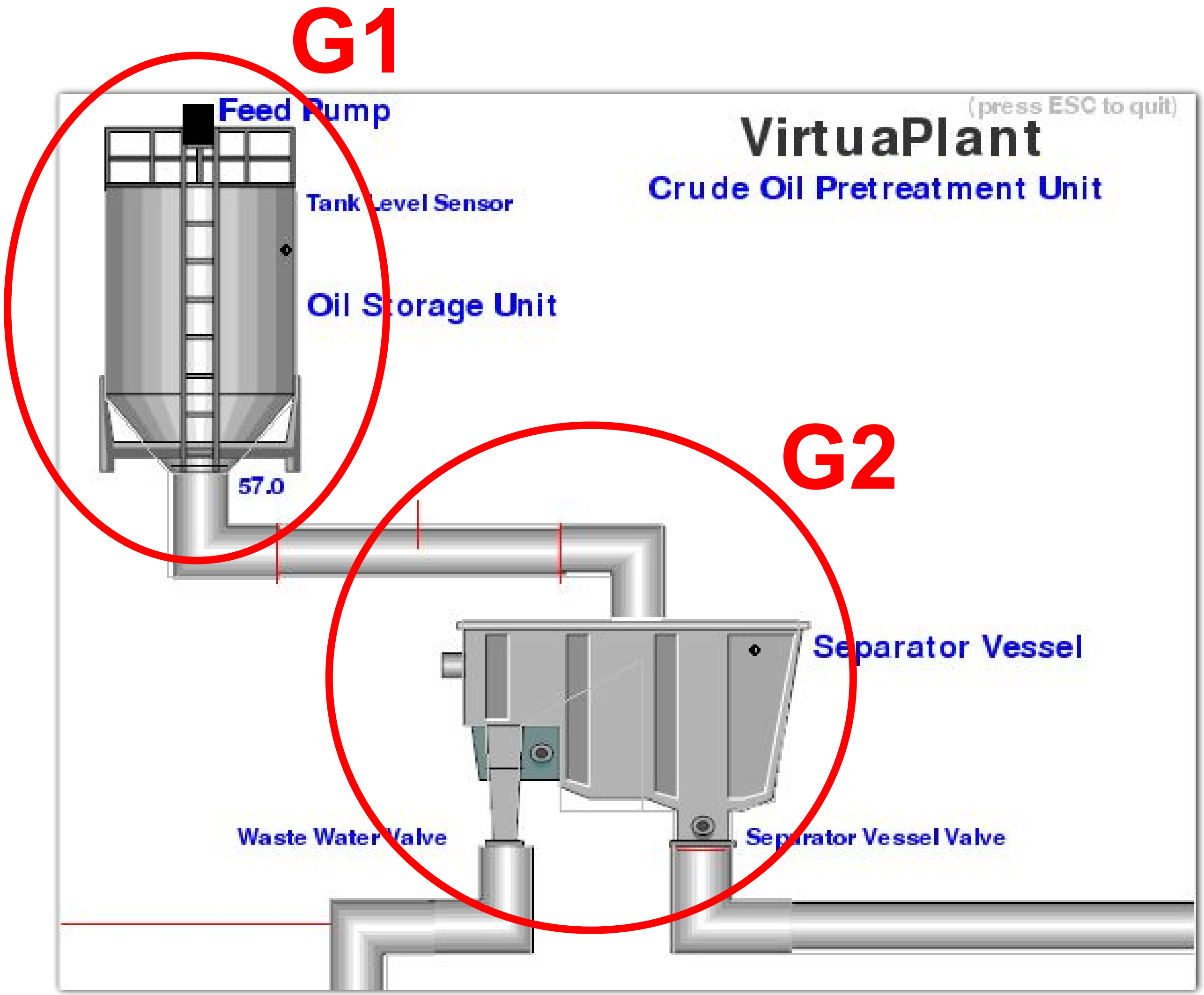
Crude Oil Pretreatment Unit

Process Runnng / Stopped?	N/A	AUTO PROCESS	MANUAL PROCESS
Crude Oil Tank 1 Level Switch	N/A		
Separator Vessel Level Switch	N/A		
Outlet Valve	N/A	OPEN	CLOSE
Separator Vessel Valve	N/A	OPEN	CLOSED
Waste Water Valve	N/A	OPEN	CLOSED
Process Status	N/A		
Connection Status	OFFLINE		
Oil Processed Status			
Oil Spilled Status			
Oil Flow After Control Valve			
Control Valve Position			

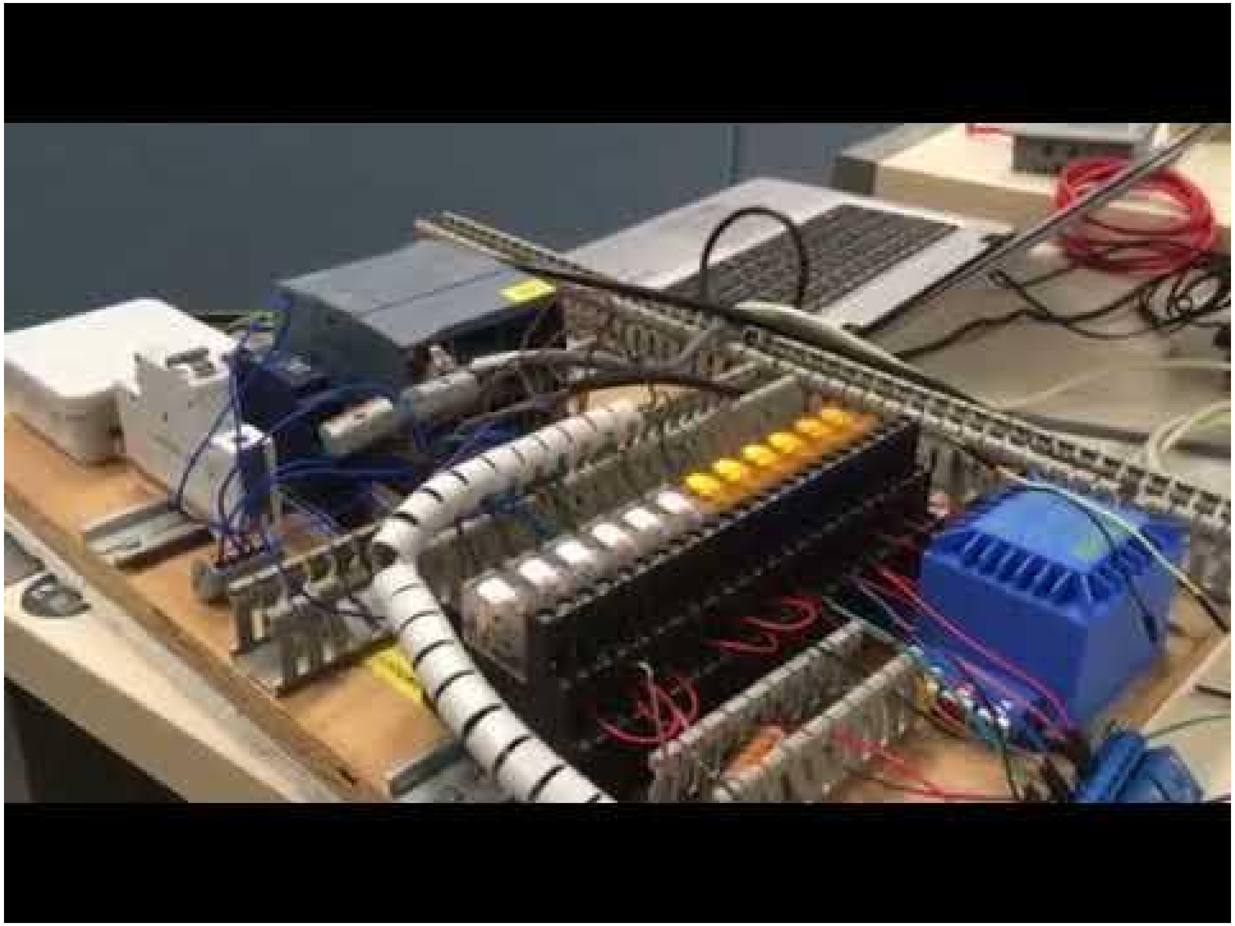
Crude Oil Pretreatment Un

Logical Groups

- Correlations:
  - G1D - G2A
  - G1A - G1D
  - G2A - G2D

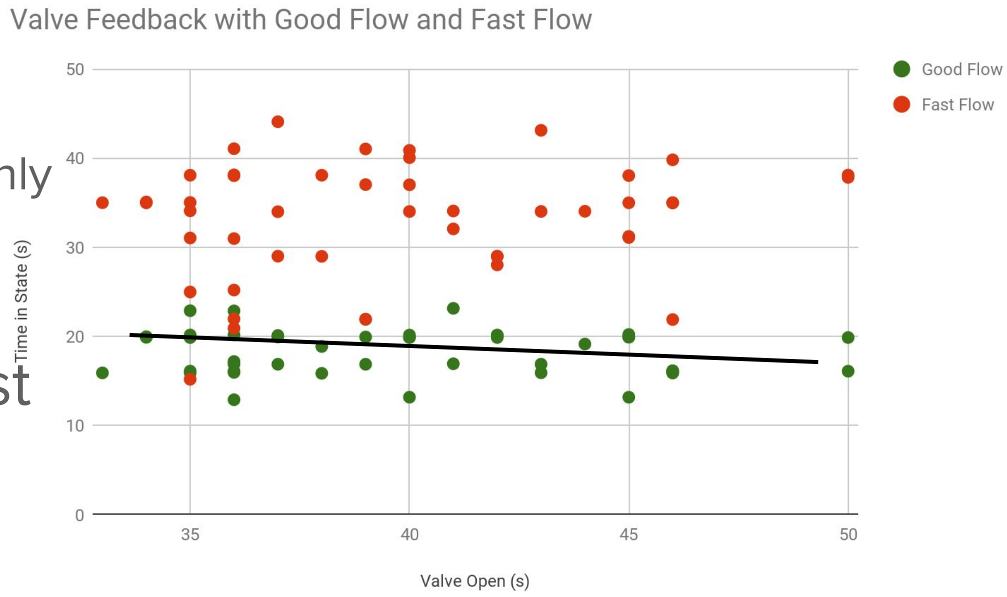


Virtuaplant on Siemens S7



Test Results Summary

- Time-In-State correlated with analogue process values
  - Time Delta between states only for condition monitoring
- SVM seems to be the best model to find outliers
- Further tests and comparisons are required



Research Results

- **Useful Information**
  - Signal list
  - Logical groups
  - Equipment age
  - Error threshold
- **Hybrid IADS Model**
  - Some contextual information required, generic otherwise
  - Covers both sequential and non-sequential processes
- **Correlation Model**
  - **Time delta** describes equipment response time
  - **Time spent in state** is more useful

Produced Artefacts

- **Modular Python API**
    - I/O Parser
    - Data Objects: ModbusObject, System State
    - Modbus Packet Dissector: based on pyshark / wireshark
    - Statechart Builder with Logical Groups
- **Portable, Expandable Test Environment**
- **Allows Easy Reproduction**
- **Allows Different Models to be tested**

Future Work

- **Performance Tests and Comparisons**
  - Experiment with LSTM, HMM and SVM
- **Find Further Correlations in Realistic Scenarios**
- **Fine Tune API for Performance**
- **Implement S7Object and Dissector - Test on Siemens S7**
- **Tests on Real Systems with Realistic Scenarios**

References

- 1st slide picture sources:
  - http://www.rainbird.com/landscape/products/flowsensors/flowSensors.htm
  - http://www.ascendant-technologies.com/direct-gas-systems/
  - https://www.conrad.com/ce/en/product/197854/Siemens-6AV6647-0AA11-3AX0-SIMATIC-KTP400-HMI-Basic-Panel-Resolution-320-x-240-pix-Interfaces-1-x-RJ45-Ethernet-for-P
  - https://uk.rs-online.com/web/p/plc-cpus/8624461/
- 2nd slide picture source:
  - https://www.shutterstock.com/video/clip-999181-stock-footage-water-bottle-factory.html
- 3nd slide picture source:
  - http://www.amusingplanet.com/2014/04/the-netherlands-impressive-storm-surge.html

Questions

???