

# Datafication

[https://github.com/Dr-AlaaKhamis/ISE518/tree/main/5\\_Datafication](https://github.com/Dr-AlaaKhamis/ISE518/tree/main/5_Datafication)

Lecture 6 – Wednesday September 10, 2025

- Datafication
- Condition Monitoring Sensors
- Data Types
- Design of Experiment (DOE)
- Data Governance

- Datafication
- Condition Monitoring Sensors
- Data Types
- Design of Experiment (DOE)
- Data Governance

- **Data explosion:** Today, around 147–181 zettabytes of data are estimated to exist globally, with projections reaching 181 ZB by 2025.
- **Recent generation:** While exact definitions vary, many estimates suggest that ~90% of this data was generated in just the past two years.
- **Digitization dominance:** Virtually all modern data is digital—already by 2014, data in digital format accounted for over 99% of all stored information.
- **Unstructured and user-generated:** Around 90% of global data is unstructured, and about 70% is user-generated (e.g., social media, videos, emails)



IoT sensors generate ~200 million TB every day



Autonomous Vehicles: 4-20 TB per day



Facebook: 350 M images uploaded per day



X community generates more than 12 terabytes of data per day

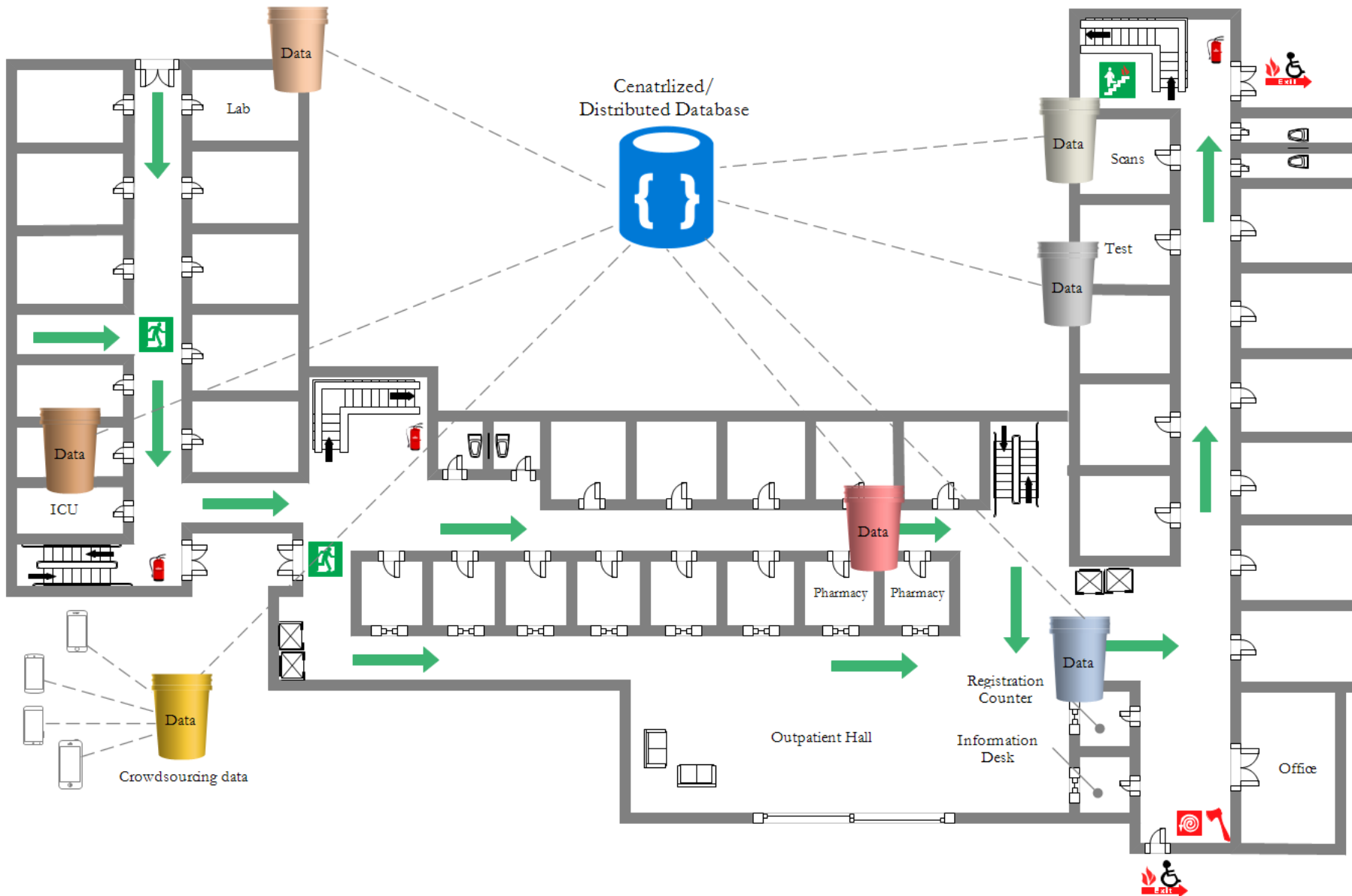


YouTube: 300 hours of video uploaded every minutes



Walmart: 2.5 Petabytes of customer data hourly

- Big Data



Data Type	Example
Visual (images/videos)	<ul style="list-style-type: none"><li>X-rays</li><li>Computerized tomography (CT or CAT scan)</li><li>Positron Emission Tomography (PET scan)</li><li>Magnetic Resonance Imaging (MRI)</li></ul>
Speech	audio or voice reports
Numerical	<ul style="list-style-type: none"><li>Patient card</li><li>Test results</li></ul>
Text	<ul style="list-style-type: none"><li>Medical reports</li><li>Reviews</li></ul>
Multimodal	Electronic Medical Records (EMR)

## • The 4 V's of Big Data

**40 ZETTABYTES**

of data will be created by 2020, an increase of 300 times from 2005



**6 BILLION PEOPLE**

have cell phones  
WORLD POPULATION: 7 BILLION



### Volume

SCALE OF DATA

**2.5 QUINTILLION BYTES**

of data are created each day



Most companies in the U.S. have at least

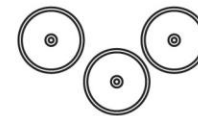
**100 TERABYTES**

of data stored



As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**



**30 BILLION  
PIECES OF CONTENT**

are shared on facebook every month



### Variety

DIFFERENT  
FORMS OF DATA

**4 BILLION +  
HOURS OF VIDEO**

are watched on You Tube each month



**4 MILLION TWEETS**

are sent per day by about 200 million monthly active users



### Velocity

ANALYSIS OF  
STREAMING DATA

The New York Stock Exchange captures

**1TB OF TRADE  
INFORMATION**

during each trading session



Modern cars have close to

**100 SENSORS**

that monitor items such as fuel level and tire pressure



**1 IN 3 BUSINESS  
LEADERS**

don't trust the information they use to make decisions



### Veracity

UNCERTAINTY  
OF DATA

**27% OF RESPONDENTS**

in one survey were unsure of how much of data was inaccurate



- Big data sizes



Byte of data: one grain of rice



Kilobyte: cup of rice



Megabyte: 8 bags of rice



Gigabyte: 3 container lorries



Terabyte: 2 container ships



Petabyte: covers Manhattan



Exabyte: covers Germany twice

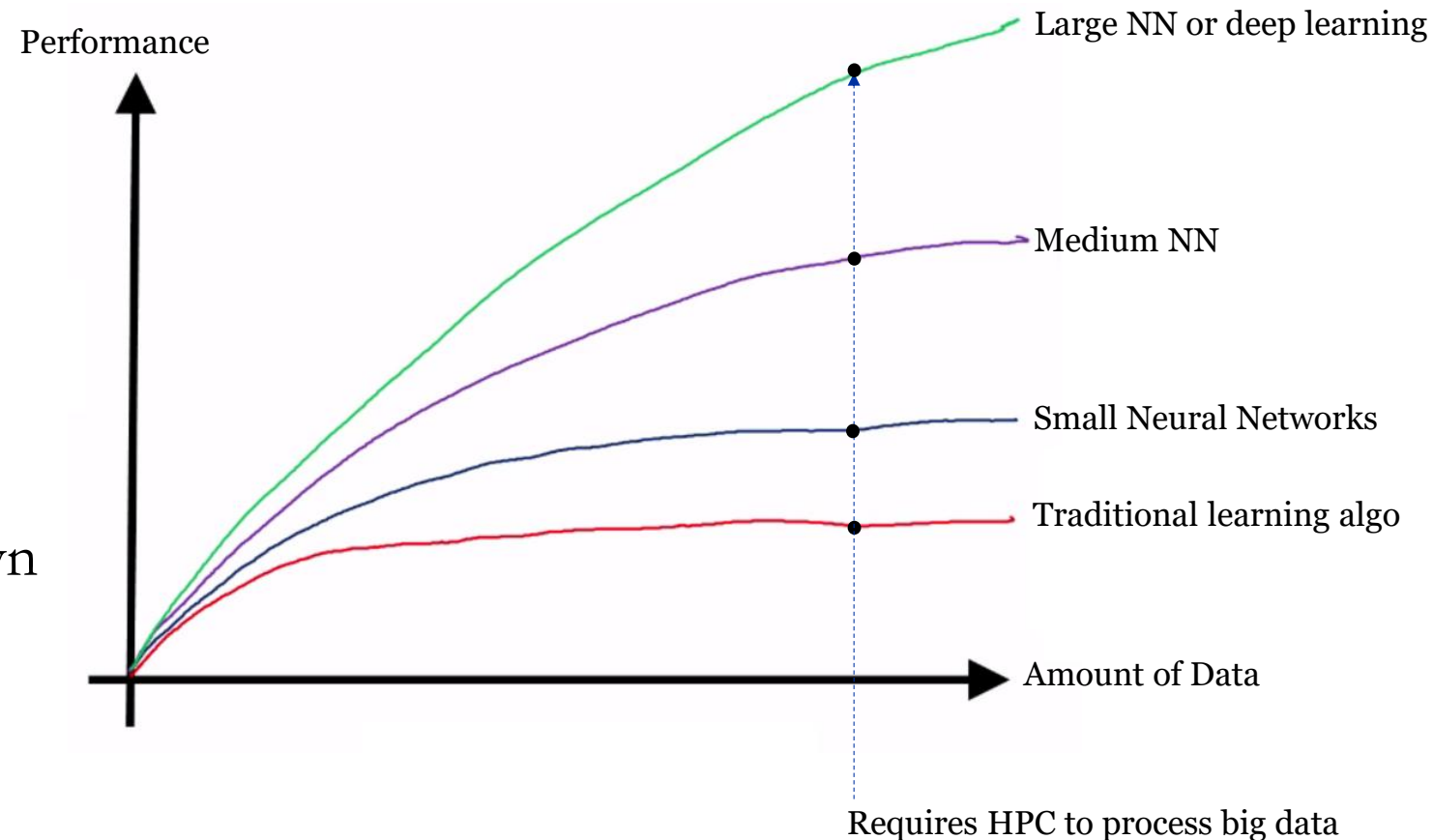


Zettabyte: fills the Pacific ocean

- **How Much Training Data is Required for Machine Learning?**

The amount of data required for machine learning depends on many factors, such as:

- **The complexity of the problem**, nominally the unknown underlying function that best relates your input variables to the output variable.
- **The complexity of the learning algorithm**, nominally the algorithm used to inductively learn the unknown underlying mapping function from specific examples.



[[Source](#)]

Source: Andrew Ng. Machine Learning Yearning. deeplearning.ai project, 2018.

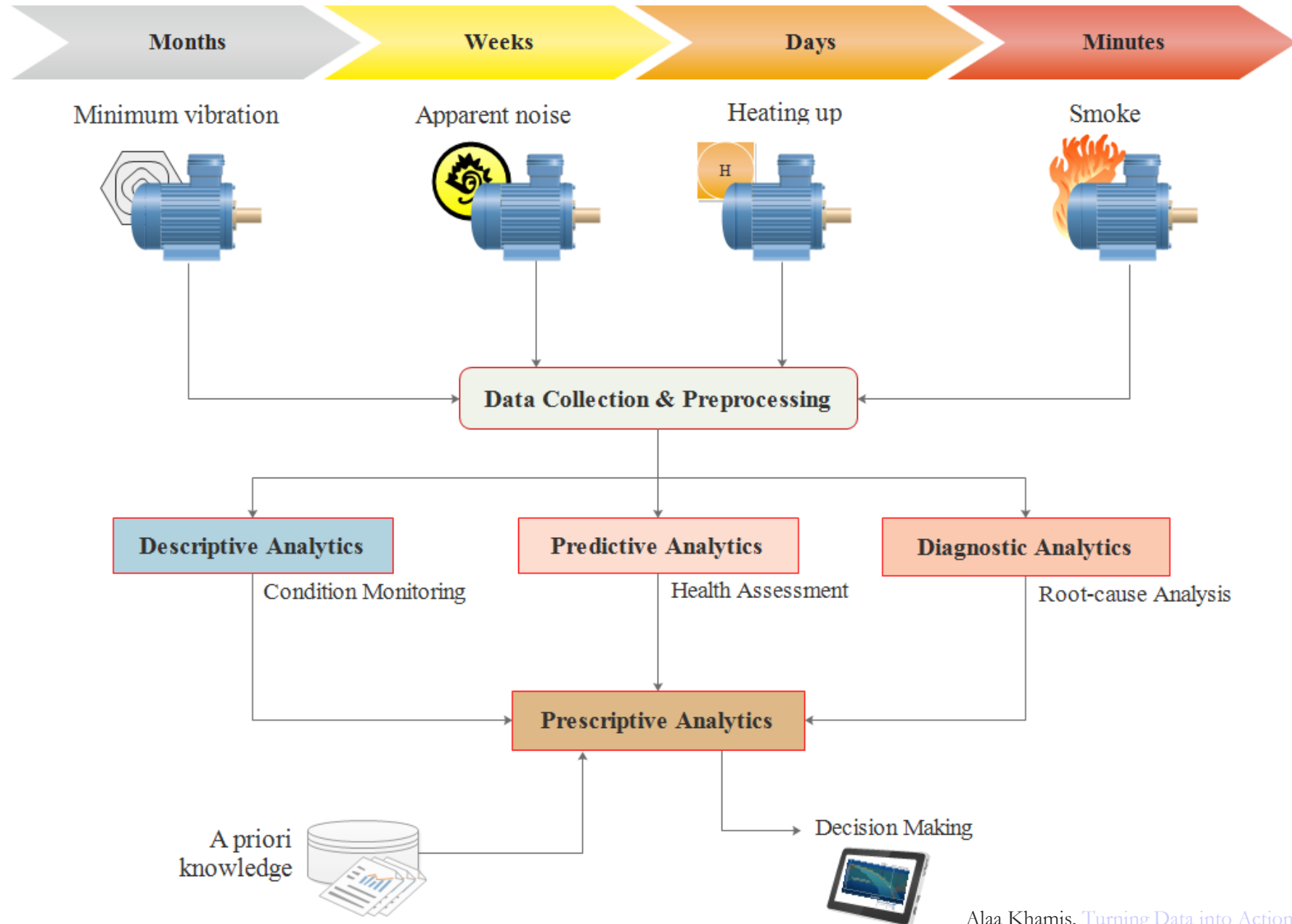


- Datafication
- **Condition Monitoring Sensors**
- Data Types
- Design of Experiment (DOE)
- Data Governance

# Condition Monitoring Sensors

10

2025 KFUPM © Alaa Khamis





Condition Monitoring Sensor Examples: power meter, non-intrusive CTs, Magnets vibration sensor, temperature sensors.



Power meter



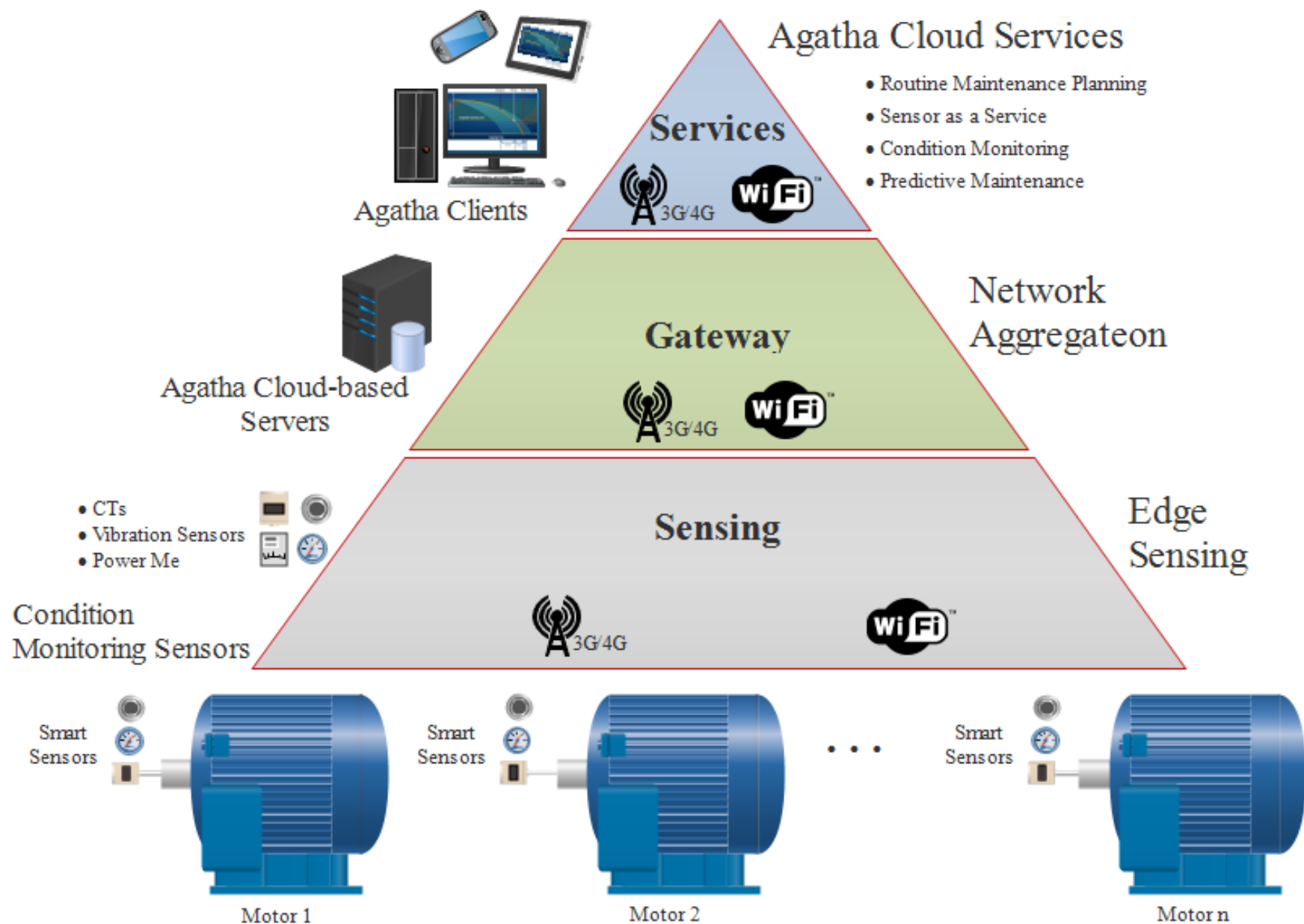
CT sensors



Vibration Sensor



Temperature Sensor

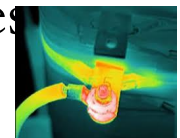
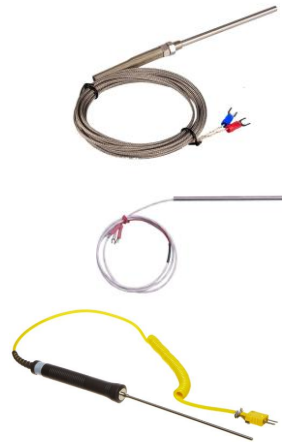


- **Non-intrusive CT (Current Transformer) Sensors:** clamp around conductors to measure current safely without cutting wires. They transmit real-time data for energy monitoring, load management, and fault detection, making them ideal for smart grids and building management.
- **Vibration Sensor:** Continuous Vibration Monitoring and Protection of Critical Equipment  
Monitors and protects 24/7
  - Operates off standard 24V loop power
  - Interfaces with plant monitoring &PI systems
  - Installs quickly and easily
  - Provides critical machine information
  - Avoids costly
  - catastrophic failures

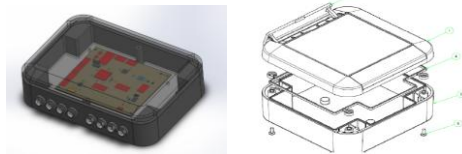
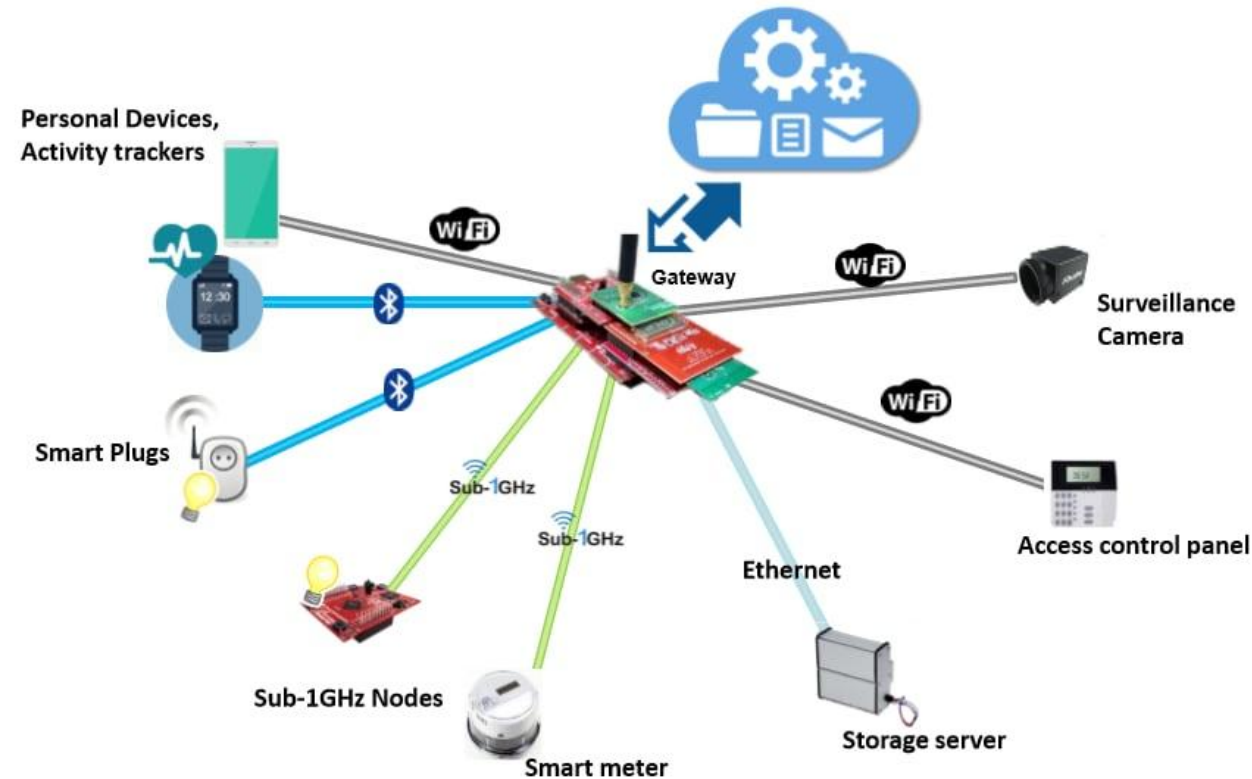




- **Power Meter:** 3 phase power meter with current and voltage transducers are mandatory to monitor the electric panels status (On/Off, normal or overload)
- **Temperature Sensors:** provide continuous, real-time data on equipment heat levels without disrupting operations. By detecting abnormal temperature rises, they help predict failures, prevent downtime, and improve safety, making them essential for motors, bearings, and other critical assets.
  - **Thermocouples:** durable, wide temperature range, common in heavy industry.
  - **RTDs** (Resistance Temperature Detectors): highly accurate and stable, used where precision is critical.
  - **Thermistors:** sensitive and fast-responding, suited for narrow-range monitoring.
  - **Infrared (IR) sensors:** non-contact, useful for moving parts or inaccessible surfaces
  - **Wireless IoT sensors:** enable remote, real-time monitoring and integration with predictive maintenance systems.

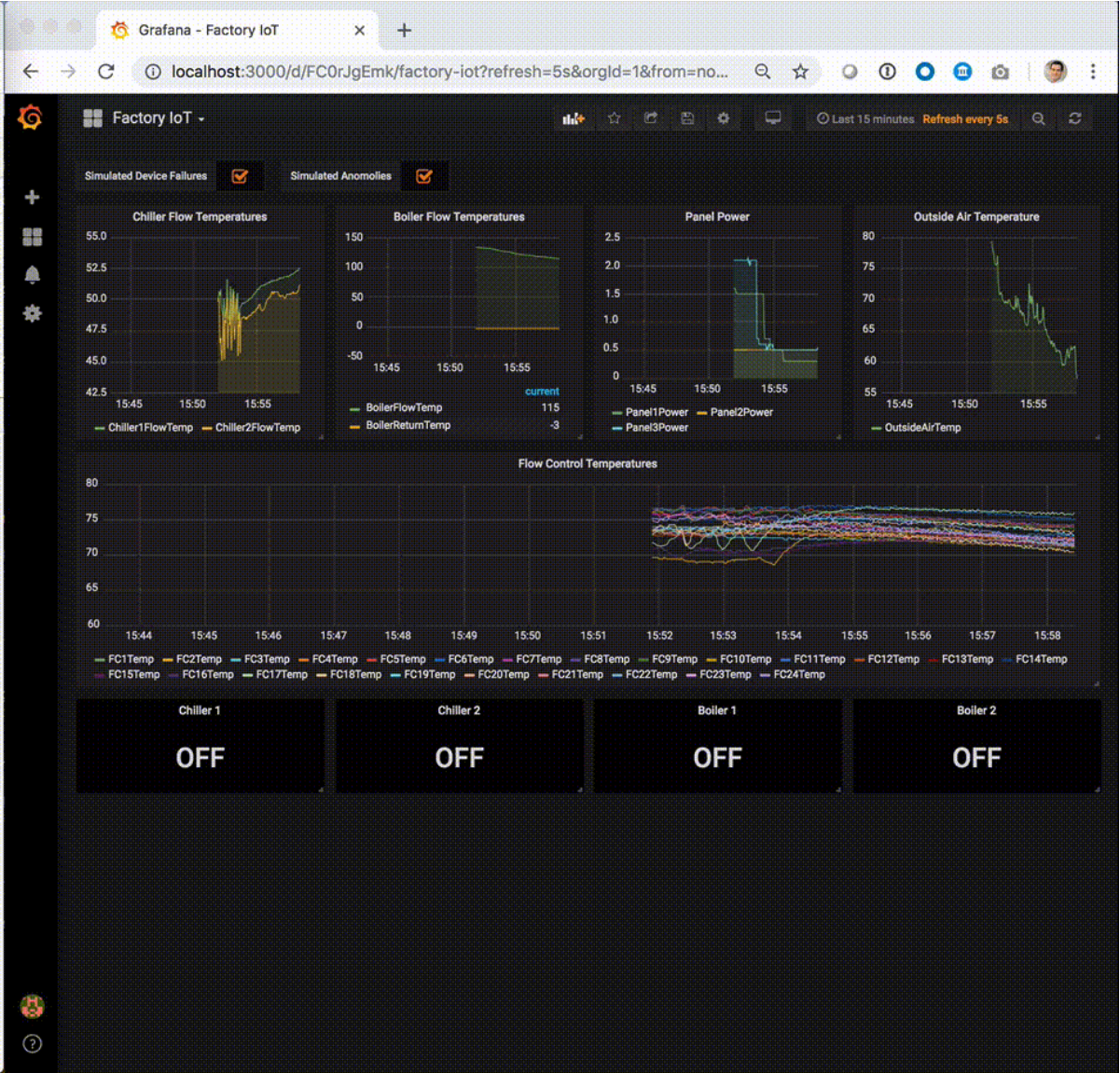
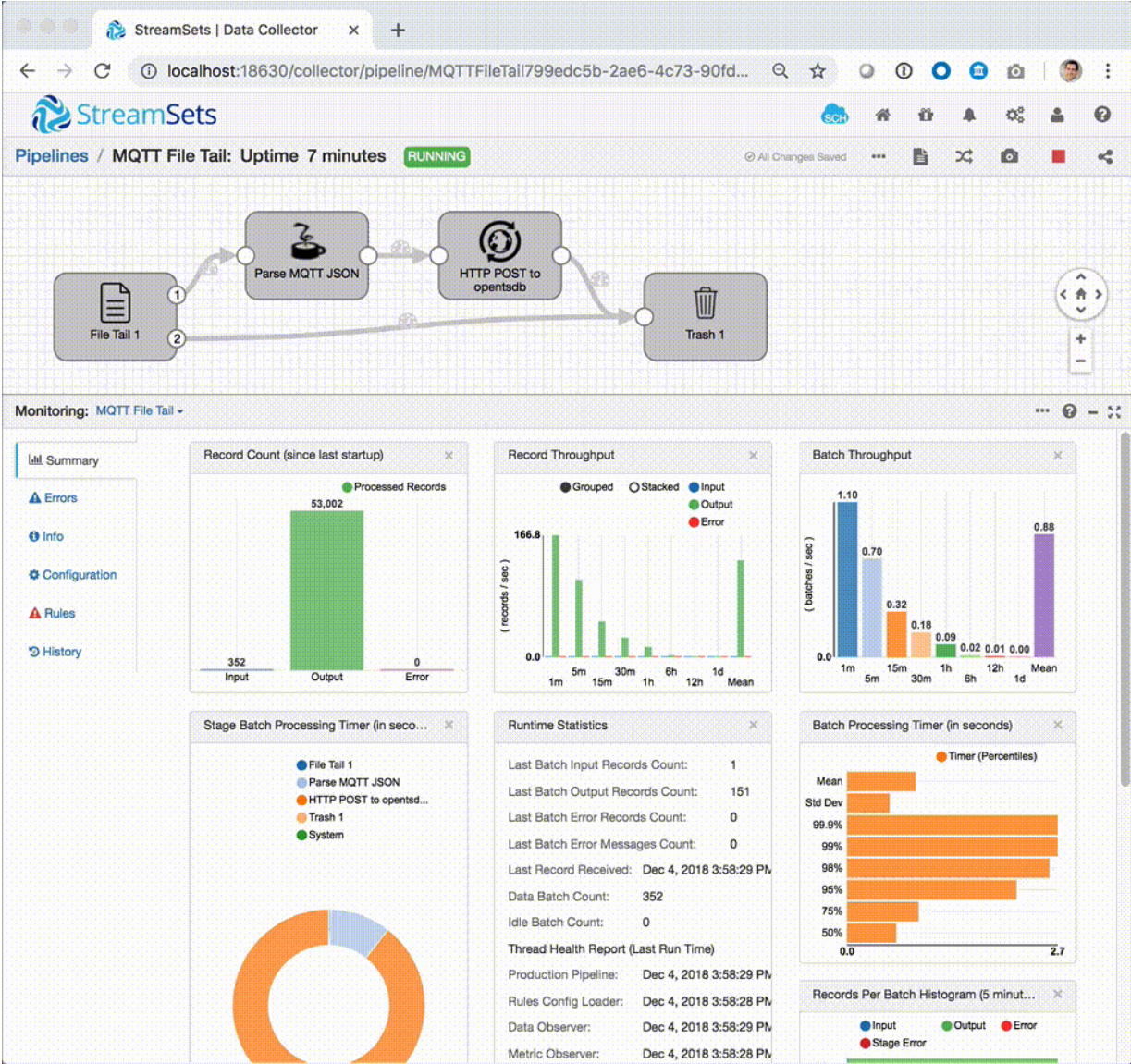


- **IoT Gateway:** Multi-communication channels to collect the data from the different sensors and direct this data to the cloud server to be analyzed. The following technical specifications are required:
  - Wi-Fi, Ethernet, RS 485, 3G/4G Connectivity
  - 12 industrial sensor Analog input (4-20 ma)
  - 8 Relay Output 220V/3A
  - Configurable over LAN and WAN
  - Secure access control
  - Sampling rate: up to 10 seconds sensors data publish rate



# Condition Monitoring Sensors

- Operational Dashboard:





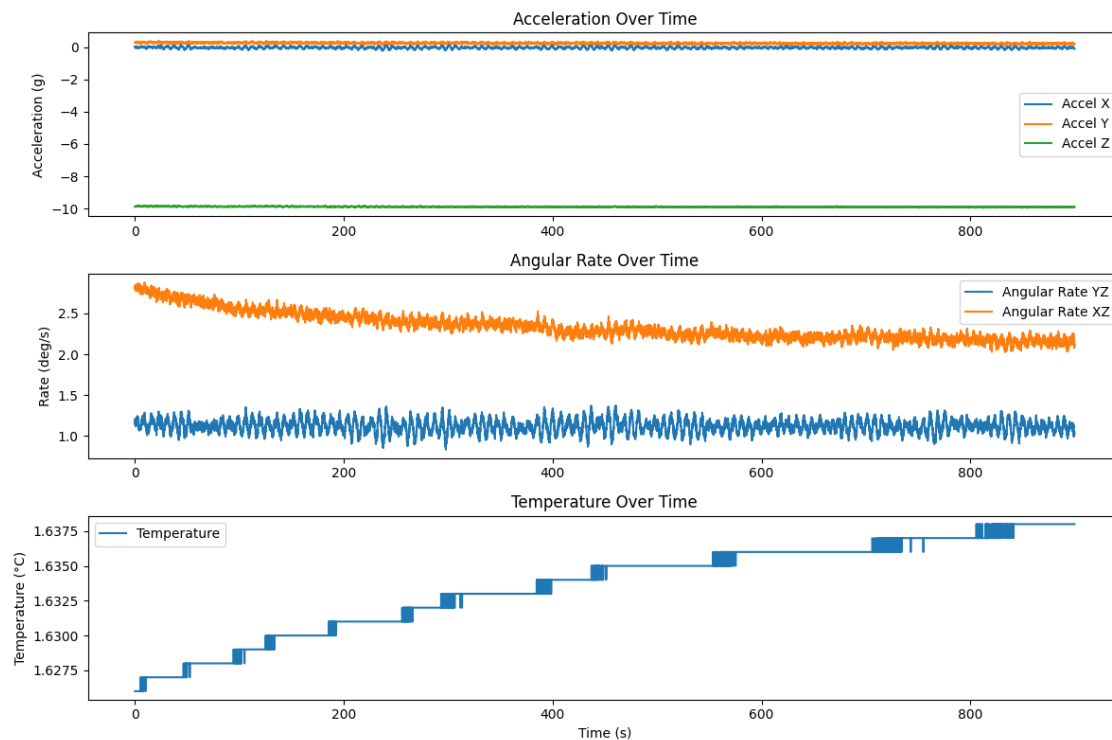
- Datafication
- Condition Monitoring Sensors
- **Data Types**
- Design of Experiment (DOE)
- Data Governance

Data Type	Example Sources
Time-Series	Vibration, temperature, pressure, flow, voltage
Acoustic/Ultrasonic	Leak detection, bearing diagnostics
Thermal/Imaging	IR cameras, visual inspections
Event/State Logs	PLC alarms, operational status
Historical Structured	CMMS data, asset metadata
Analytical/Lab	Oil/lubricant, wear debris analysis
Human/Annotation	Operator notes, RCA, inspection logs

- **Time-Series Sensor Data**

Continuously sampled measurements over time, essential for trend and anomaly detection.

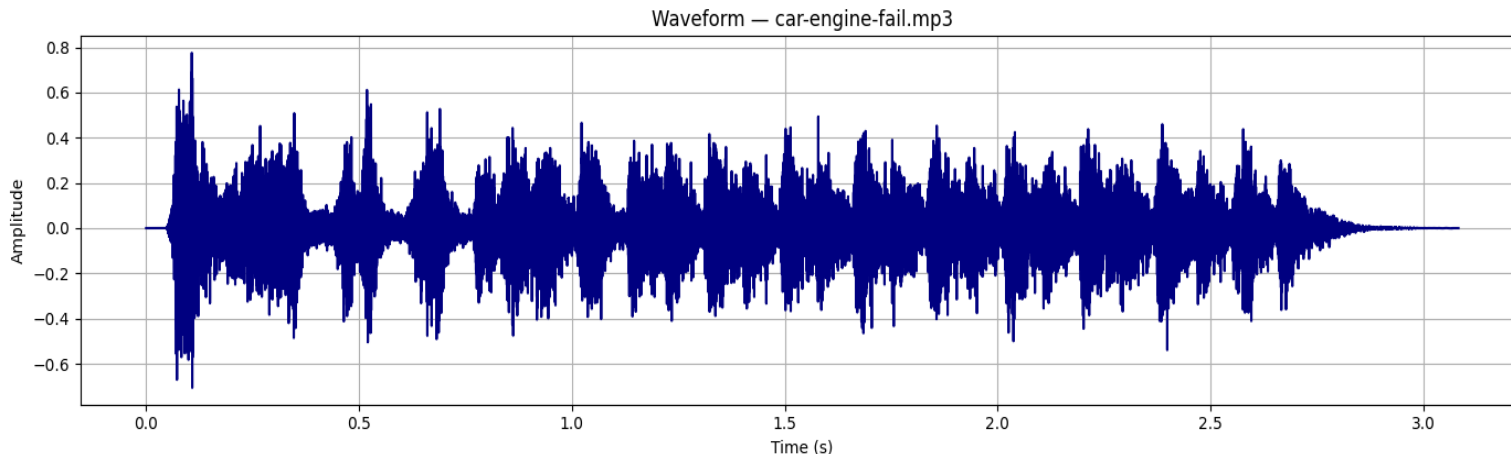
- **Vibration data** (accelerometers, velocity, displacement)
- **Temperature readings** (thermistors, RTDs)
- **Pressure data**
- **Flow rate measurements**
- **Electrical current & voltage** (e.g., motor current signature analysis)
- **RPM / speed / torque sensors**
- **Environmental data** (humidity, ambient temperature)



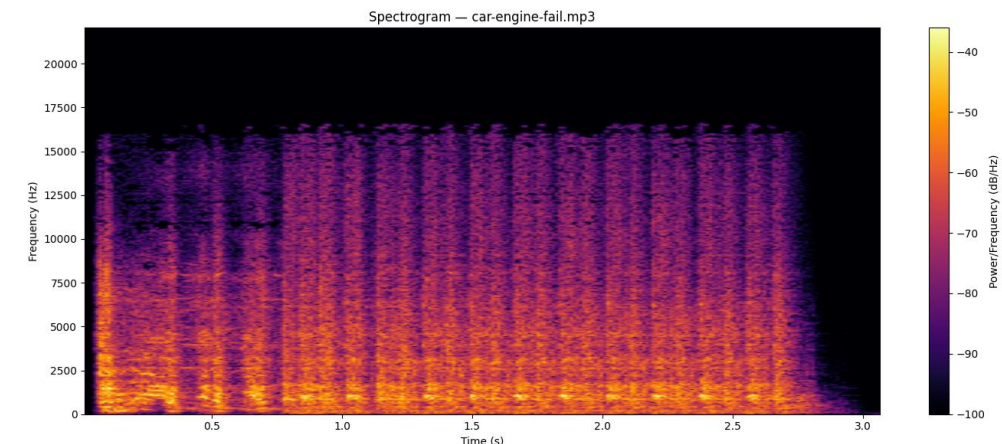
- **Acoustic & Ultrasonic Data**

High-frequency signals, often analyzed in waveform or spectrogram formats.

- Ultrasound sensor data (bearing inspection, leak detection)
- Acoustic emissions (early-stage fault detection in mechanical parts)
- Sound recordings (used with ML for diagnostics)



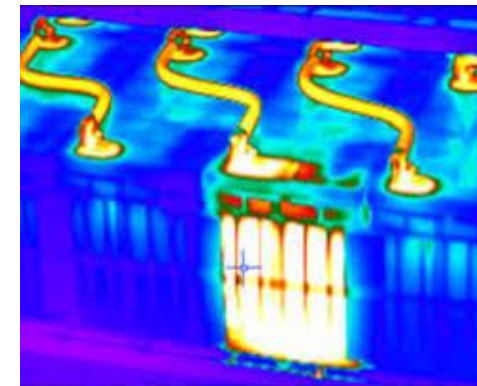
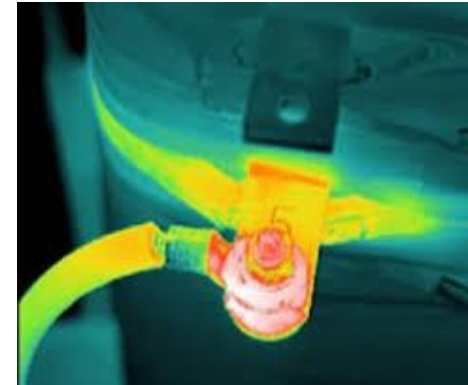
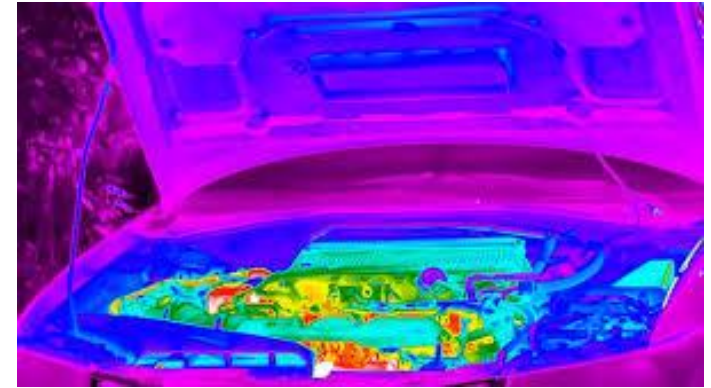
File: data/audio/car-engine-fail.mp3  
Duration: 3.08 s  
Sample Rate: 44100 Hz  
Samples: 135936  
Max Amplitude: 0.775  
Min Amplitude: -0.706  
Mean Amplitude: -0.000  
Std Deviation: 0.100



- **Thermal/Imaging Data**

Spatial or pixel-based data capturing thermal signatures or visual cues.

- Infrared thermography (IR images/videos)
- Visual inspections (RGB cameras)
- Thermal maps of components or systems



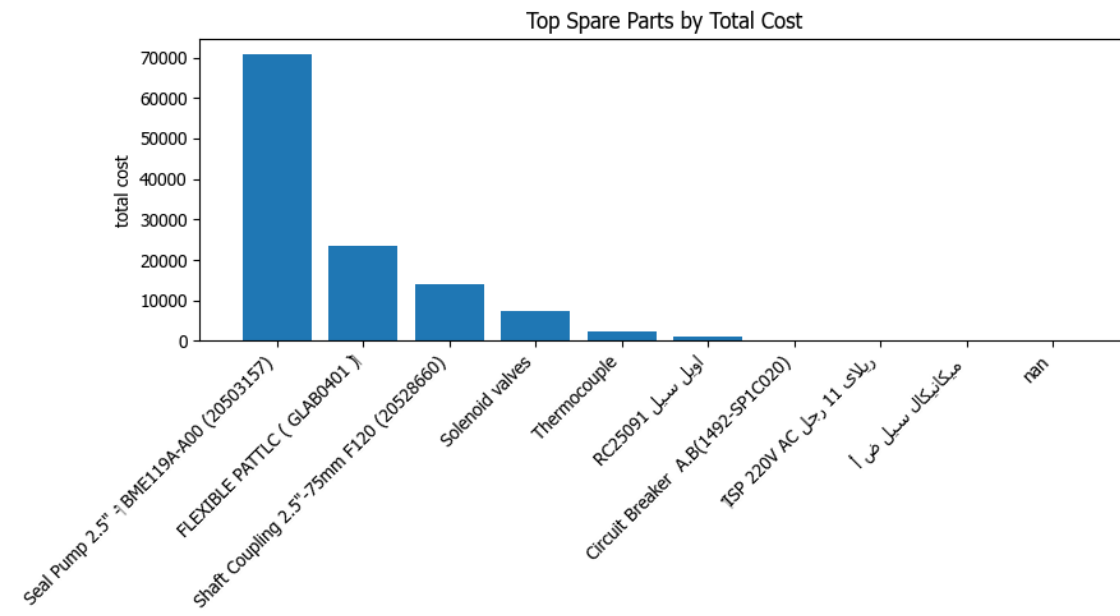
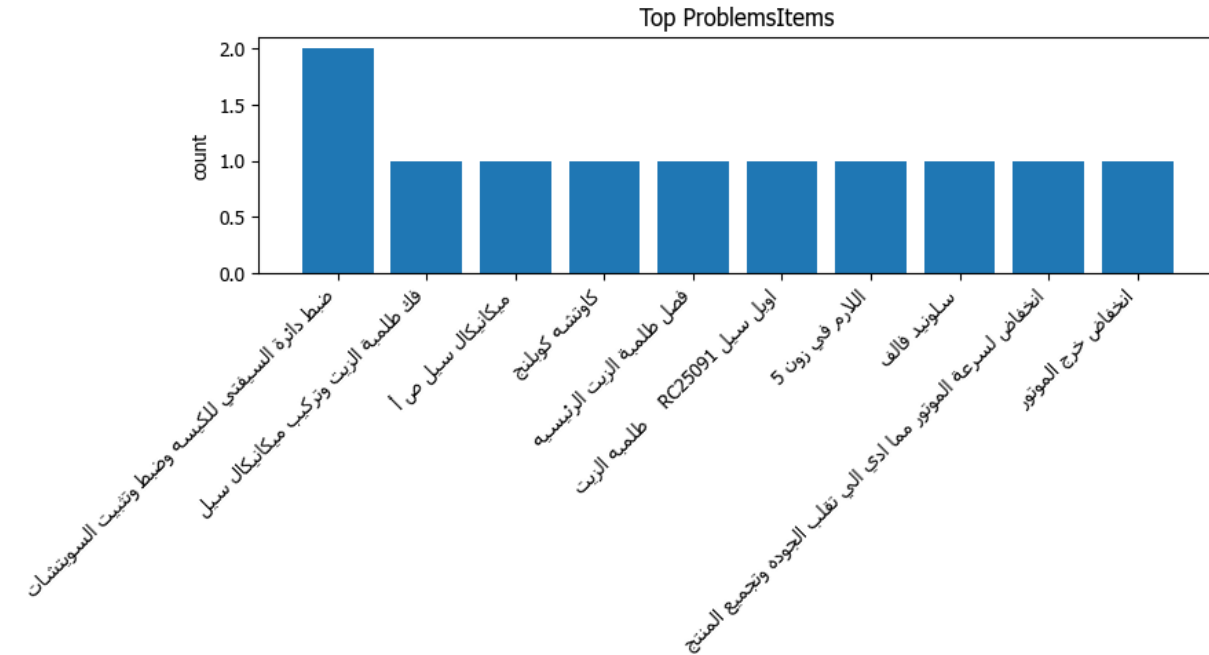
# Data Types

## • Logo/Event/State Data

Discrete logs or flags representing system state changes or events.

- PLC/DCS alarms and event logs
- Fault codes and error messages
- Start/stop events
- Operating mode/status (idle, running, error)

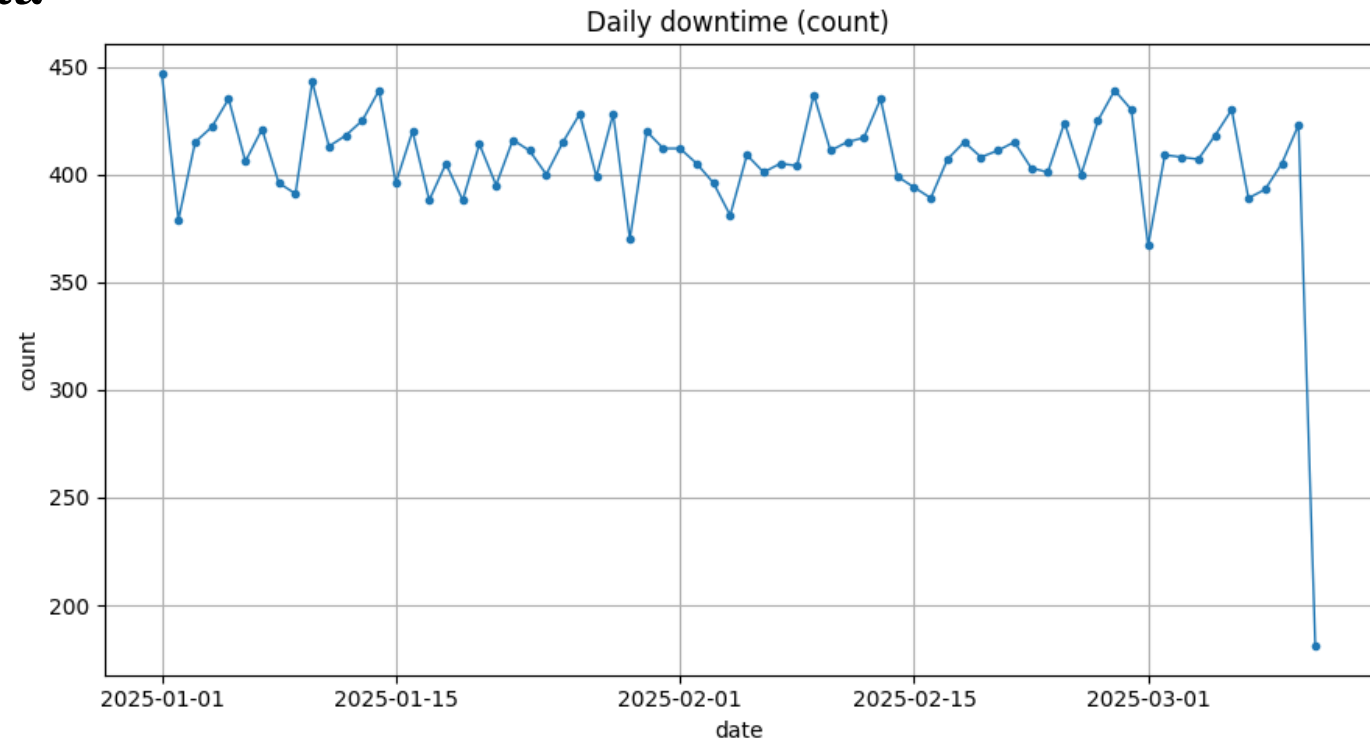
Top problem item: 'ضبط دائرة السيفتي للكيسه وضبط وتثبيت السويتشات' (2 occurrences).  
 أكثر مشكلة تكرارًا: 'ضبط دائرة السيفتي للكيسه وضبط وتثبيت السويتشات' بعدد 2.  
 Highest total cost spare part: 'Seal Pump 2.5" - BME119A-A00 (20503157)' with 70,952.33.  
 بقيمة 70,952.33 'Seal Pump 2.5" - BME119A-A00 (20503157)': أعلى تكلفة إجمالية لقطع الغيار.  
 Job completion rate: 100.0%.  
 نسبة إكمال الأعمال: 100.0%.  
 Shift with highest average NetTime: 3 (95.0 minutes).  
 الوردية ذات أعلى متوسط زمن صافي: 3 (95.0 دقيقة).



- **Structured Historical/Maintenance Data**

Tabular data, useful for failure modeling and supervised ML.

- CMMS / EAM data (work orders, failure reports, mean time between failures)
- Asset metadata (make, model, age, location, specs)
- Maintenance logs (corrective, preventive actions taken)
- Downtime records



	timestamp	machine_id	failure_type	anomaly_flag	downtime_risk	maintenance_required	machine_status	machine_status_label
2	2025-01-01 00:02:00	15	Normal	1	1.0	1	1	running
3	2025-01-01 00:03:00	43	Normal	1	1.0	1	1	running
4	2025-01-01 00:04:00	8	Vibration Issue	0	0.0	1	2	error
8	2025-01-01 00:08:00	23	Normal	1	1.0	1	1	running
13	2025-01-01 00:13:00	40	Overheating	0	0.0	1	2	error
15	2025-01-01 00:15:00	3	Normal	1	1.0	1	1	running
17	2025-01-01 00:17:00	2	Normal	1	1.0	1	1	running
18	2025-01-01 00:18:00	24	Normal	1	1.0	1	1	running
21	2025-01-01 00:21:00	38	Normal	1	1.0	1	1	running
23	2025-01-01 00:23:00	21	Vibration Issue	0	0.0	1	2	error
26	2025-01-01 00:26:00	22	Normal	0	0.0	1	1	running
27	2025-01-01 00:27:00	44	Normal	0	0.0	1	1	running



- **Analytical/Lab Data**

Intermittently collected and lab-analyzed, often stored in structured format.

- Oil/lubricant analysis (particle count, water content, viscosity)
- Wear debris analysis
- Coolant contamination analysis

**Account Information:** Specifies the account number (assigned by the lab) and contact information of the company or individual who submitted the sample for analysis.

**Component Information:** Specifies the component ID and secondary ID (provided by the customer) for equipment identification. Also specifies component type, manufacturer, model, application and sump capacity.

**Sample Information:** Identifies the location of the analysis; the data analyst's initials; and the dates the sample was taken, received and completed. Turnaround issues may be indicative of storing samples too long before shipping or shipping service problems.

**Severity Status Levels:** Indicates the status of severity of the submitted oil sample\*.

**Severity 0 (Normal)** = Oil is suitable for continued use.

**Severity 1 (Normal)** = Oil is suitable for continued use. Observe for trends in future tests.

**Severity 2 (Abnormal)** = Oil is suitable for continued use. Resample at half the normal interval.

**Severity 3 (Abnormal)** = Replace oil filter and top off system with fresh oil. Resample at half the normal interval or change oil.

**Severity 4 (Critical)** = Change oil and filter if not done when sample was taken.

\* These particular statuses are specific to engine oil samples, but often reflect the lab's recommendations for other fluids.

**Oil Analyzers INC.** Lubricant Analysis Report  
North America: 1-877-458-3315

Overall report severity based on comments.

Account Information	Component Information	Sample Information
Account Number: OILANA-1234-5678 Company Name: Contact: JOHN Q. CUSTOMER Address: 1234 MAIN STREET ANYTOWN, WI 54555 US Phone Number: 715-555-5555	Component ID: John's Truck Secondary ID: 2014 Suburban Component Type: Unleaded Gasoline Engine Manufacturer: Chevrolet Model: 5.3L Application: Transportation Sump Capacity: 7 qts	Tracking Number: 12345A67890 Lab Number: I-123456 Lab Location: Indianapolis Data Analyst: AKB Sampled: 04-Jul-2014 Received: 07-Jul-2014 Completed: 09-Jul-2014
Filter Information	Miscellaneous Information	Product Information
Filter Type: Full-Flow Micron Rating: 20		Product Manufacturer: AMSOIL Product Name: ASL SIG SIGNATURE SERIES Viscosity Grade: SAE 5W30
Comments	NEW LUBE REFERENCE - Data used for baseline reference only;	

**Filter Information:** Identifies the filter used and its micron rating.

**Miscellaneous Information:** Details additional miscellaneous information.

**Product Information:** Identifies the sample lubricant and its properties.



- Human Input / Expert Annotations

Qualitative or semi-structured data, increasingly used in supervised learning.

- Operator observations
- Manual inspection notes
- Failure root cause analysis (RCA) reports
- Expert labeling of fault types

COMPANY NAME HERE															27 POINT VISUAL INSPECTION														
123 Any Street City, State Zip Code www.YourWebsite.com PHONE: (123) 456-7890 Here ...															NAME: John Doe					CELL PHONE: (123) 456-7890									
															ADDRESS: 123 Any Street					HOME PHONE: (123) 456-7890									
															CITY: New York					STATE: NY		ZIP: 10044			WRITTEN BY: Jason Mayberry				
															YEAR: 2007		MODEL: FORD			MAKE: F-150			LICENSE PLATE: 397644						
VIN: 4 U H 8 T R 4 6 U I L 6 3 2 8 9 6															MOTOR: 5.7L V8		MILEAGE: 236,368			DATE: 11/21/2021									
NOTES		ESTIMATED COST	INSPECTED ITEMS																										
Replace Assembly Pos. & Neg.	28.50	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	1. Battery Cables																								
Replace Engine Air Filter	21.45	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	2. Air Filter																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	3. Wiper Blades																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	4. Head Lights, Turn Signals, Brake & Turn Lights																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	5. Windshield Washer Fluid Level and Operation																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	6. Radiator and Heater Hoses																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	7. Power Steering Fluid Level, Hoses & Leaks																								
Needs a Transmission Flush	189.85	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	8. Automatic Transmission Fluid Level and Condition																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	9. Automatic Transmission Fluid Leaks																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	10. Front and Rear Shocks and Struts (Oil Leaks)																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	11. Brake Fluid Level & Condition																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	12. Brake fluid Leaks																								
Front Brake Pads @ 35%	56.78	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	13. Front Brake Pads																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	14. Rear Brake Pads/Shoes																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	15. Parking Brake Operation																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	16. Engine Oil Level & Condition																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	17. Engine Oil Leaks																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	18. Lube handles, Hinges & Latches																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	19. Cooling System Levels & condition																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	20. Exhaust System																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	21. Engine Drive Belts																								
No Charge, THANK YOU!	0.00	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	22. Tire Balance & Rotation																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	23. Tires For Proper Alignment																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	24. Right Front Tire Wear																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	25. Left Front Tire Wear																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	26. Right Rear Tire Wear																								
		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	27. Left Rear Tire Wear																								

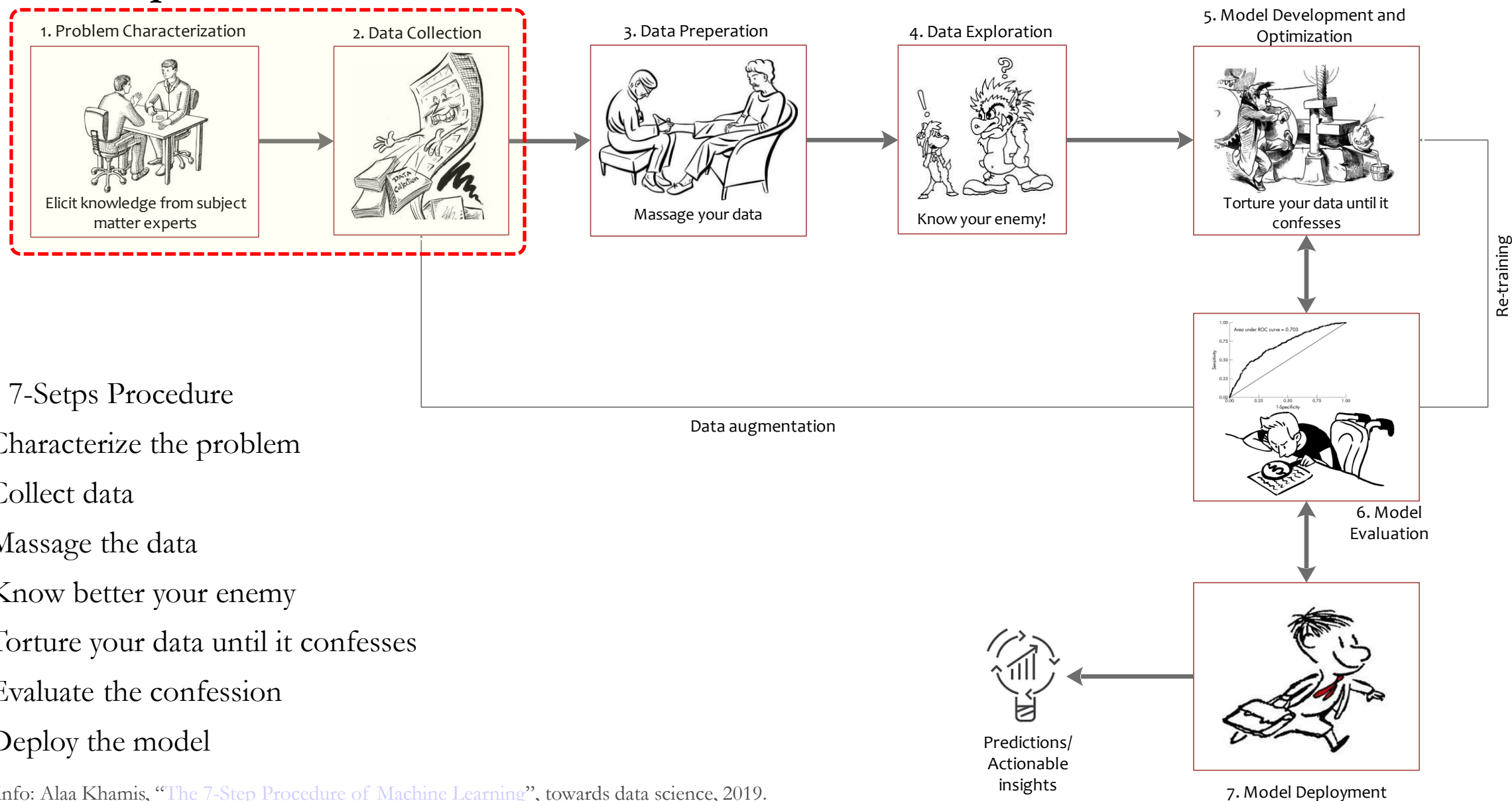
Pass

Needs Attention

Needs Immediate Attention

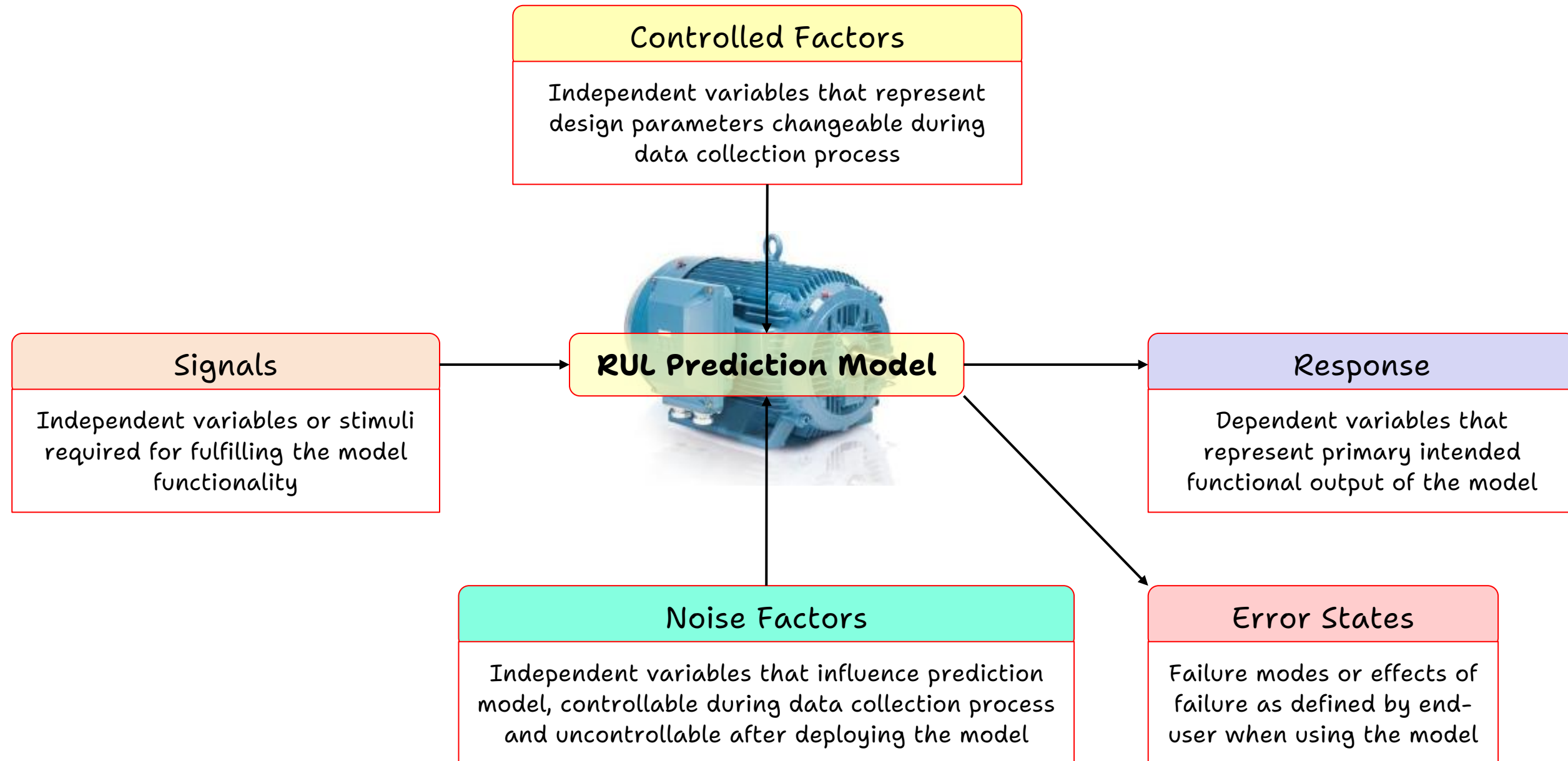
- Datafication
- Condition Monitoring Sensors
- Data Types
- **Design of Experiment (DOE)**
- Data Governance

## • The 7-Step Procedure



### The 7-Steps Procedure

1. Characterize the problem
2. Collect data
3. Massage the data
4. Know better your enemy
5. Torture your data until it confesses
6. Evaluate the confession
7. Deploy the model





# Design of Experiment (DOE)

30

2025 KFUPM © Alaa Khamis



C0: Safe Driving



C1: Text Right



C2: Phone Right



C3: Text Left



C4: Phone Left



C5: Adjusting Radio



C6: Drinking



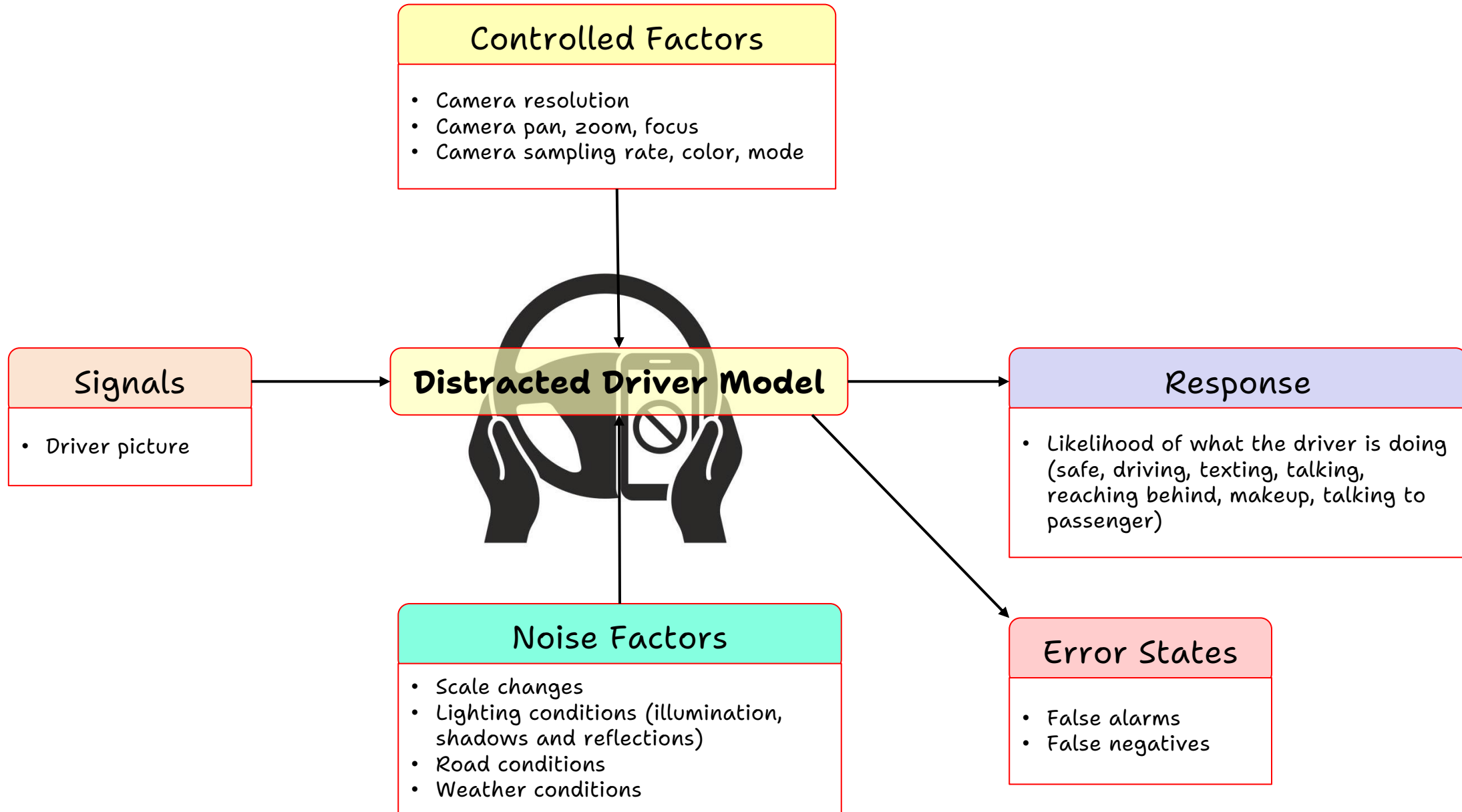
C7: Reaching Behind

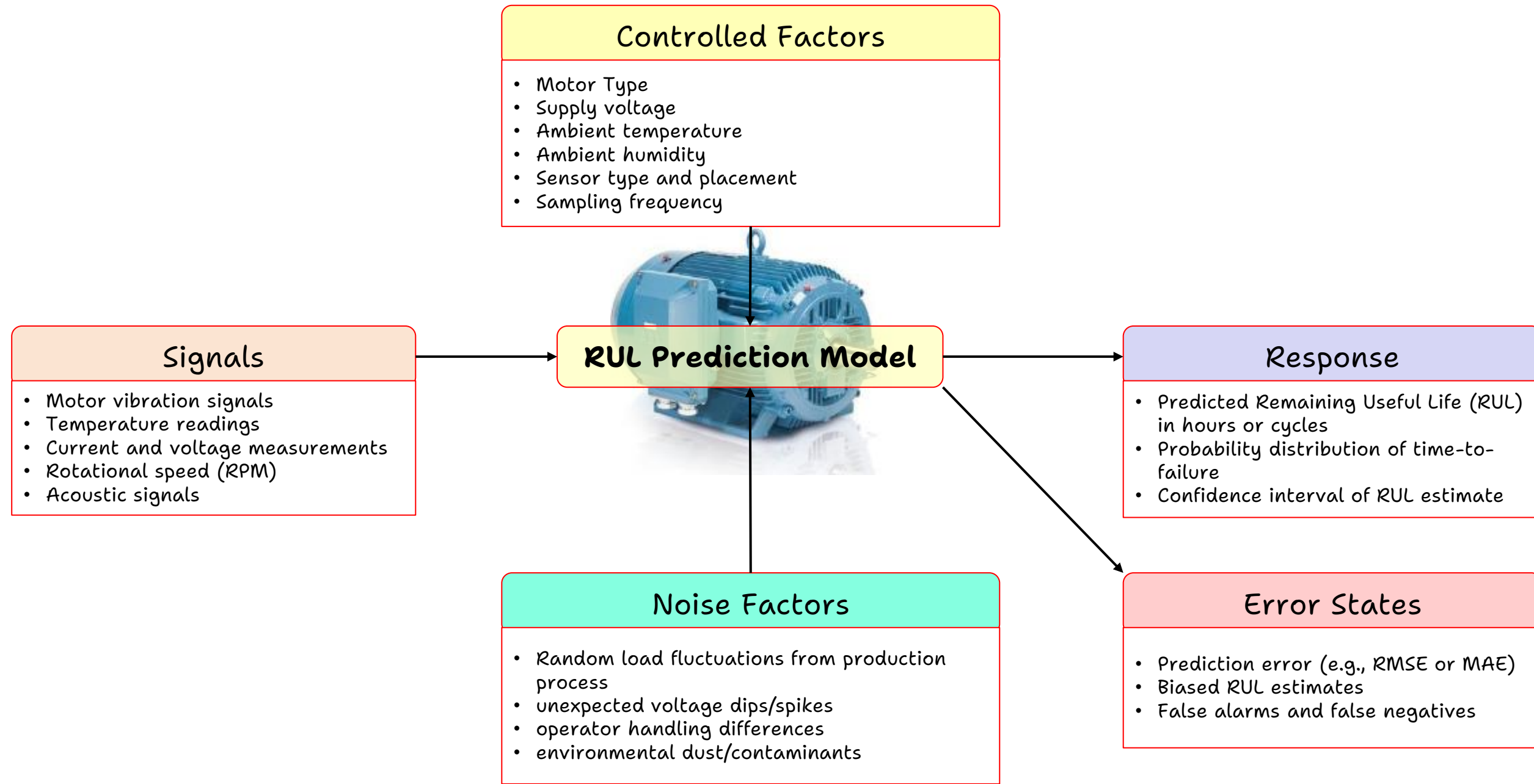


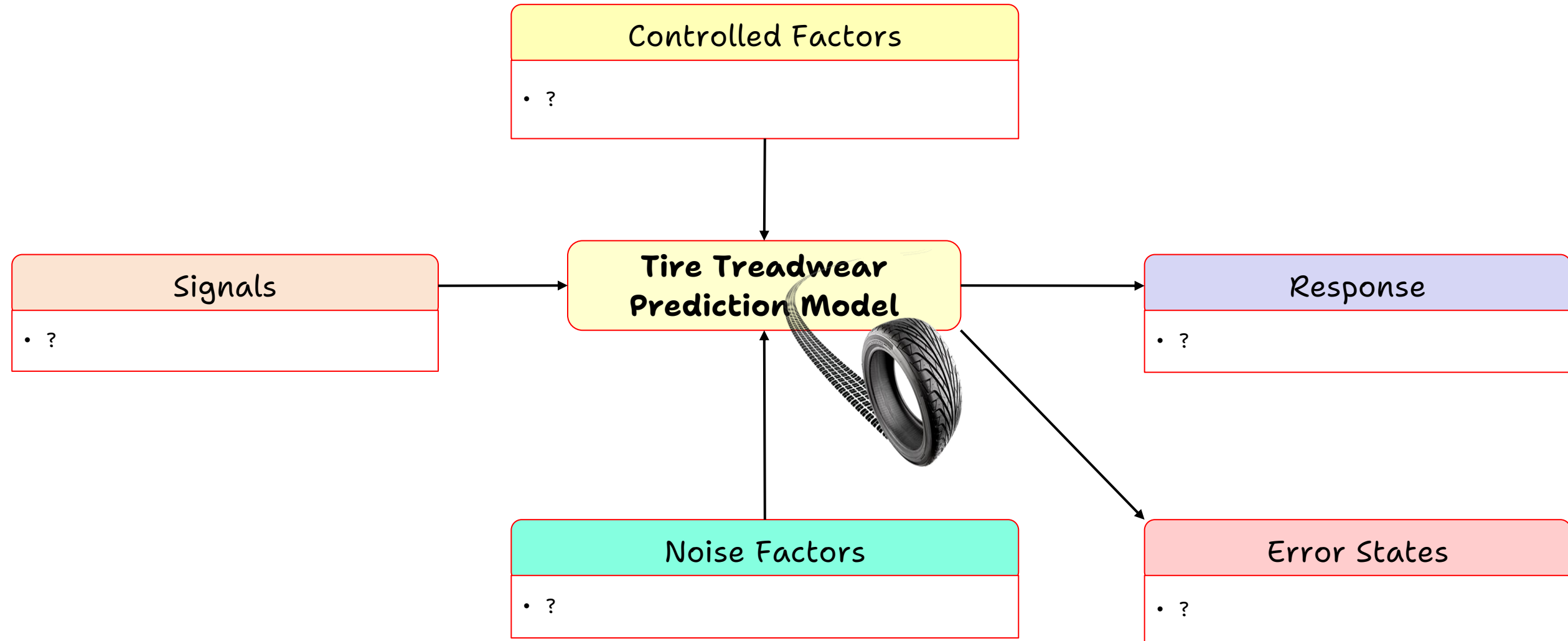
C8: Hair or Makeup



C9: Talking to Passenger









- How do I set up an experiment?

## Design of Experiment (DOE)

Full Factorial Design

Fractional Factorial Design

Pros & Cons	Full Factorial	Fractional Factorial
Pros	<ul style="list-style-type: none"><li>• All possible combinations can be covered</li><li>• Analysis is straightforward, as there is no aliasing</li></ul>	<ul style="list-style-type: none"><li>• Less memory and effort</li><li>• Less time</li><li>• Runs can be added to eliminate confounding.</li></ul>
Cons	Cost of the experiment increases as the number of factors increases	<ul style="list-style-type: none"><li>• Analysis of higher order interactions could be complex</li><li>• Confounding could mask factor and interaction effects</li></ul>

- Full Factorial Design

$$N_{full} = r \prod_{i=1}^k L_i$$

$N_{full}$  is the number of runs

$L_i$  is number of levels for factor  $i$

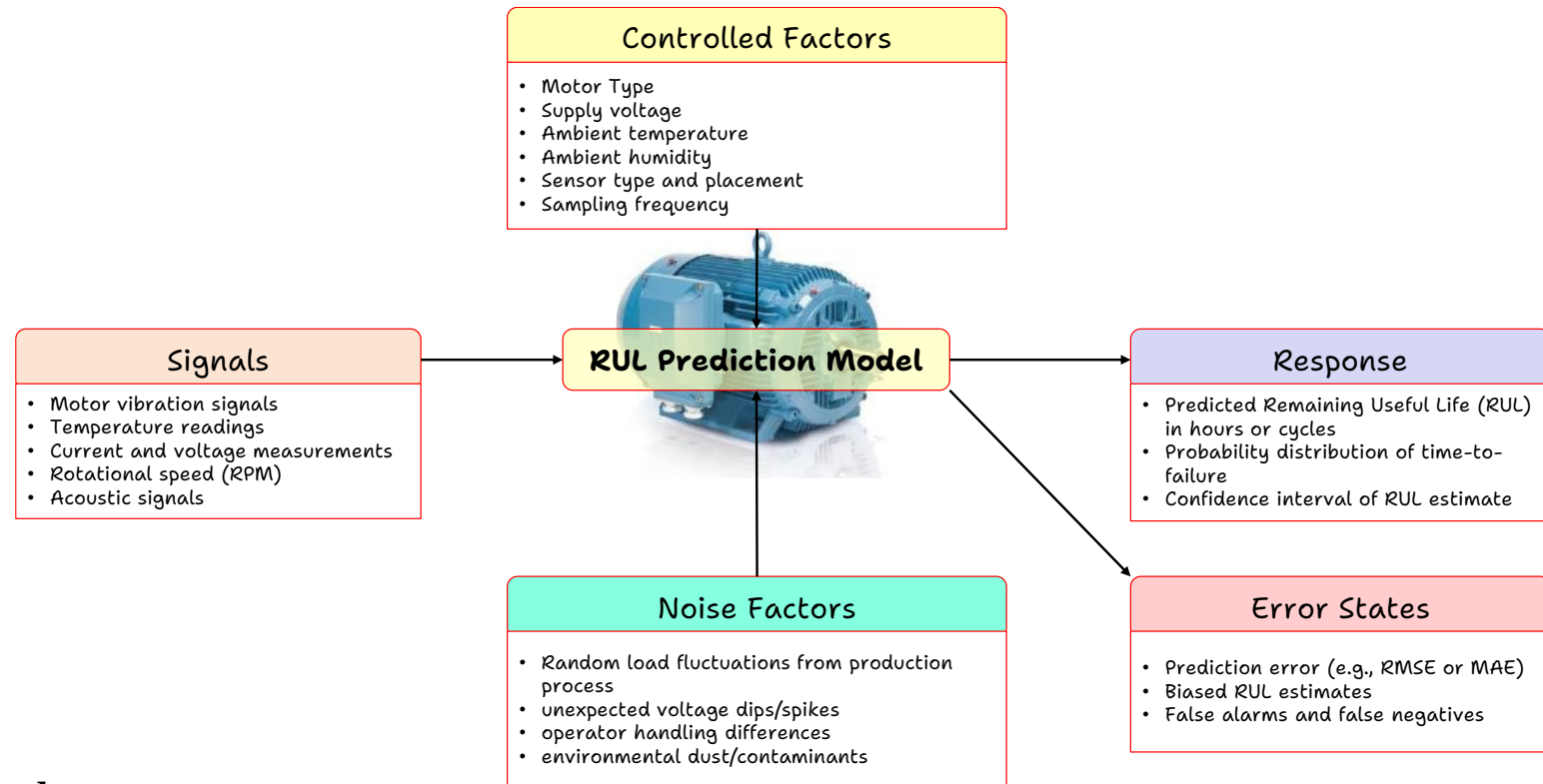
$r$  is number of replicates (optional)

## Special cases

Two-level factors only  $\Rightarrow N = r 2^k$

Three-level factors only  $\Rightarrow N = r 3^k$

**Example:** considering only 3 two-levels controlled factors and 3 two-level noise factors with no replicate:  $N = r 2^k = 2^6 = 64$  runs. If each run takes 30 minutes, the total duration to collect the data would be 1920 minutes (32 work hours or **four 8-hour full days**).



- Fractional-factorial (regular resolution) for a two-level fractional design

$$N_{frac} = r 2^{k-p}$$

$k$  is the total number of factors

$p$  is number of generators (fractionality)

$p = 0 \Rightarrow$  full,  $p = 1 \Rightarrow$  half fraction,  $p = 2 \Rightarrow$  quarter fraction, etc.

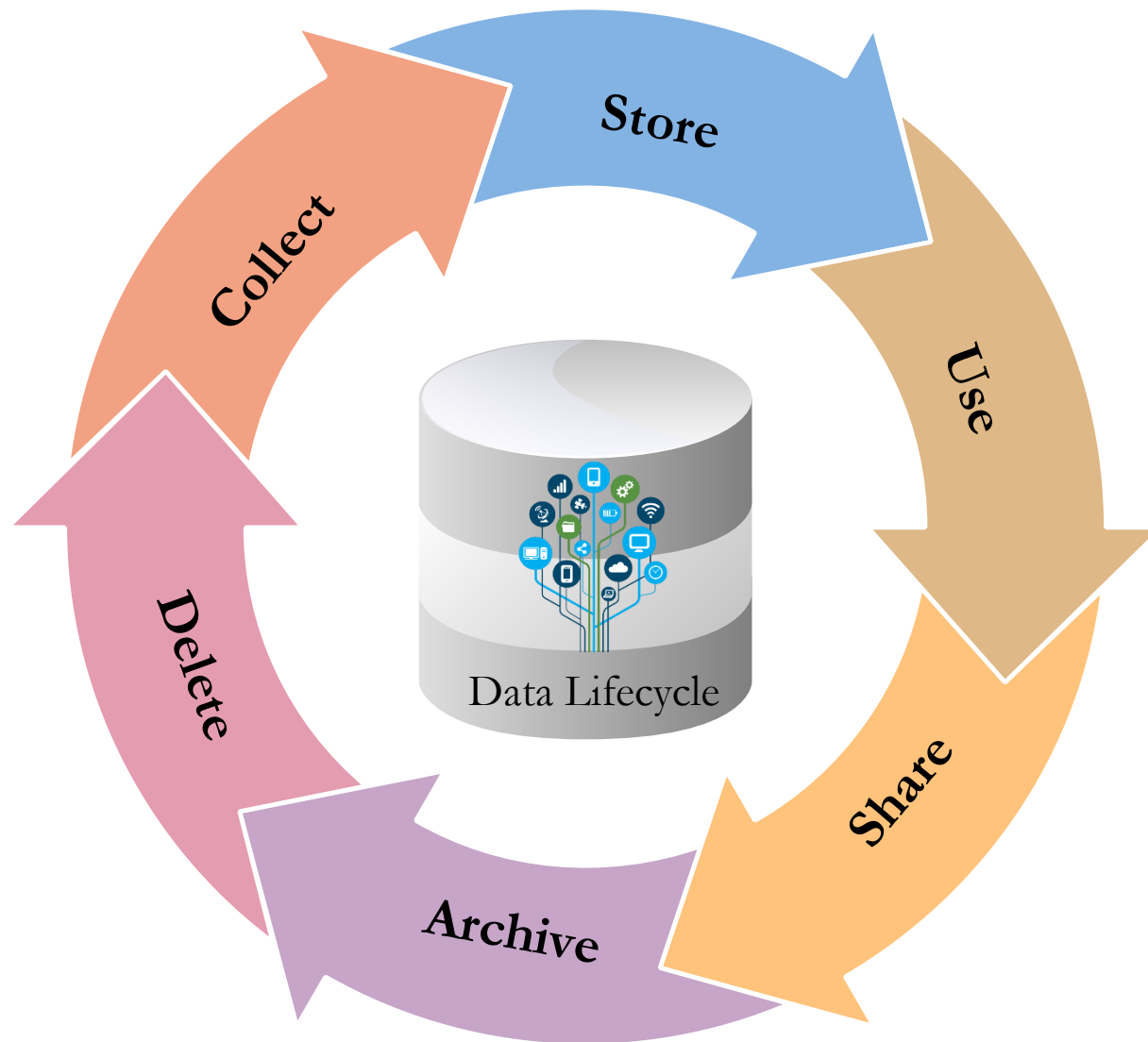
**Example:** considering only 3 two-levels controlled factors and 3 two-level noise factors with no replicate. Number of runs for quarter fraction experiment:  $N = r 2^{k-p} = 2^{6-2} = 16$  runs. If each run takes 30 minutes, the total duration to collect the data would be 480 minutes (8 hours or one full day).

- Fractional-factorial (regular resolution) for a three-level fractional design

$$N_{frac} = r 3^{k-p}$$

- Datafication
- Condition Monitoring Sensors
- Data Types
- Design of Experiment (DOE)
- **Data Governance**

- **Data Privacy, Ownership and Equity**



- What types of data are collected and shared?
- Why should the data be collected and shared?
- What are the benefits of sharing the data?
- How owns the data?
- With whom the data will be shared?
- When will the data be collected and used?
- How will the data be collected and used?
- Will anonymization and privacy masking be applied on the data?
- What will happen to the user's data or profile when the user dies?