Datafication

https://github.com/Dr-AlaaKhamis/ISE518/tree/main/5_Datafication

Lecture 6 – Wednesday September 10, 2025

Outline

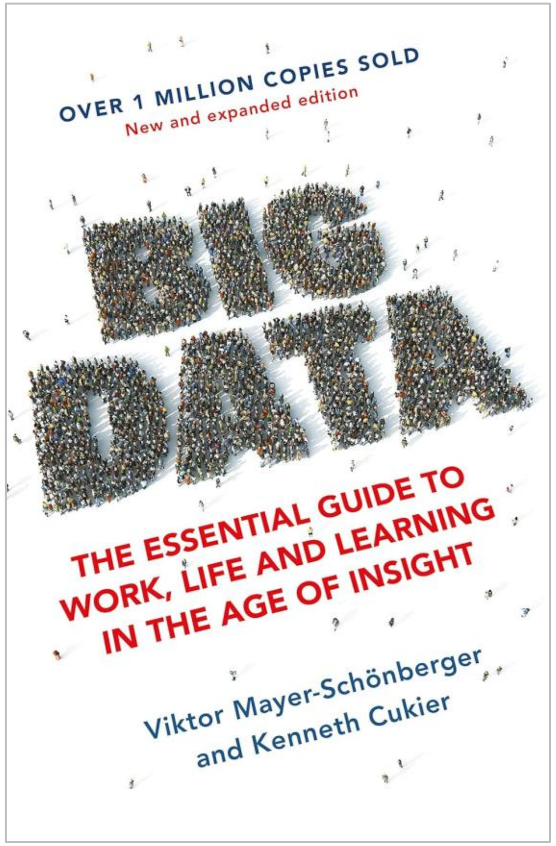
- Datafication
- Condition Monitoring Sensors
- Data Types
- Design of Experiment (DOE)
- Data Governance

Outline

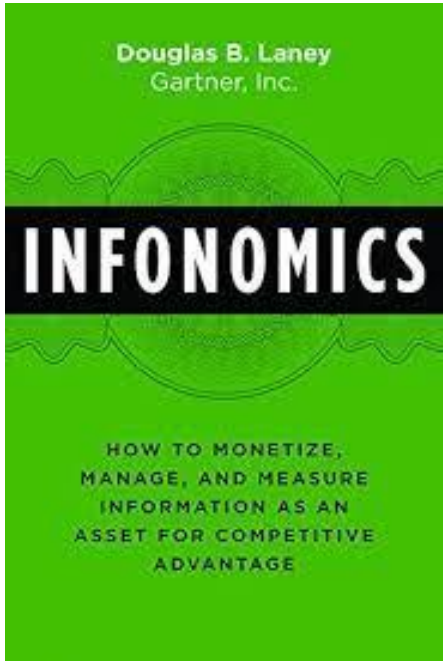
- **Datafication**
- Condition Monitoring Sensors
- Data Types
- Design of Experiment (DOE)
- Data Governance

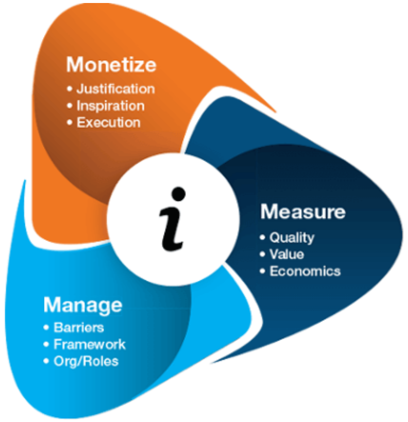
Datafication

Datafication is the broad, technology-driven process of **turning actions, interactions, objects and even thoughts** into **quantified data streams** that can be **stored, analyzed and monetized**. The term was popularized by Viktor Mayer-Schönberger and Kenneth Cukier in Big Data: A Revolution That Will Transform How We Live, Work and Think (2013).



Infonomics is "**the theory, study and discipline of asserting economic significance to information,**" applying "economic and asset-management principles and practices to the valuation, handling and deployment of information assets.





Datafication

- **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

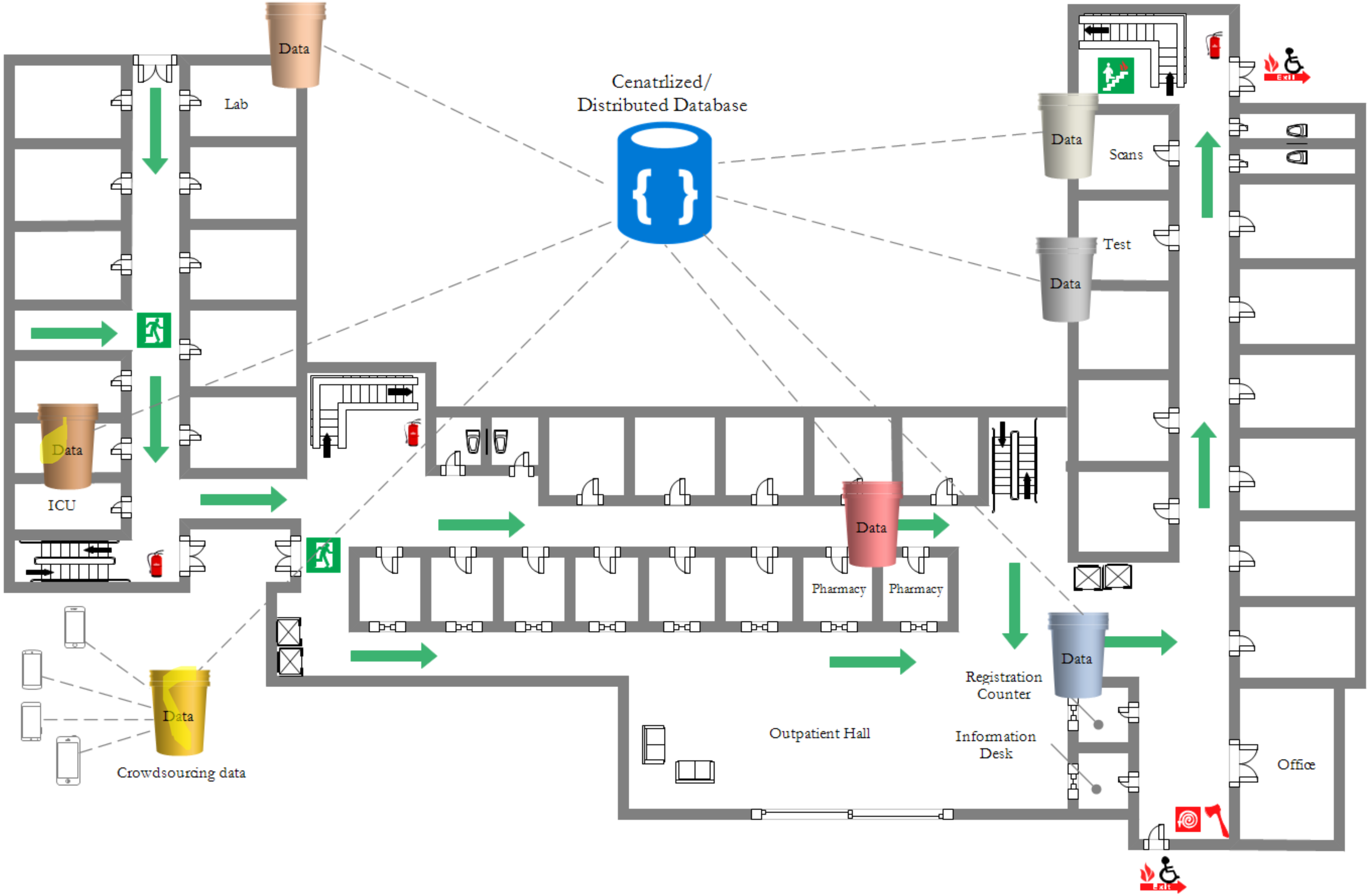
Walmart: 2.5 Petabytes of customer data hourly

Datafication

- **Big Data: Structured vs. Unstructured Data**

Industry	Structured data (rows & columns, well-defined schema)	Unstructured data (free-form text, images, sound, etc.)
Manufacturing	SCADA sensor logs (timestamp, machine ID, temperature, vibration, rpm) stored in a SQL historian	Maintenance-crew voice notes and equipment photos describing faults; PDF equipment manuals
Retail / e-commerce	Point-of-sale transactions (SKU, price, quantity, storeID, time)	Customer reviews and star ratings; product-demo videos and images
Banking / fintech	Core-bank ledger records (account #, debit, credit, balance, currency)	Chat-bot transcripts, KYC selfie images, call-center audio recordings
Healthcare	EMR/EHR vitals table (patient ID, visit date, BP, HR, lab values)	Radiology DICOM images, doctor's free-text notes, pathology slide images
Transportation / logistics	Telematics table (vehicle ID, GPS lat/long, speed, fuel level, timestamp)	Dash-cam videos, driver voice logs, shipping-label scans
Energy / utilities	Smart-meter readings (meter ID, kWh, reactive power, interval)	Drone imagery of power-line inspections, PDF regulatory filings
Media & entertainment	Subscriber database (user ID, plan tier, join date, churn flag)	Streaming-service watch-history text, movie & show video files, social-media posts about releases

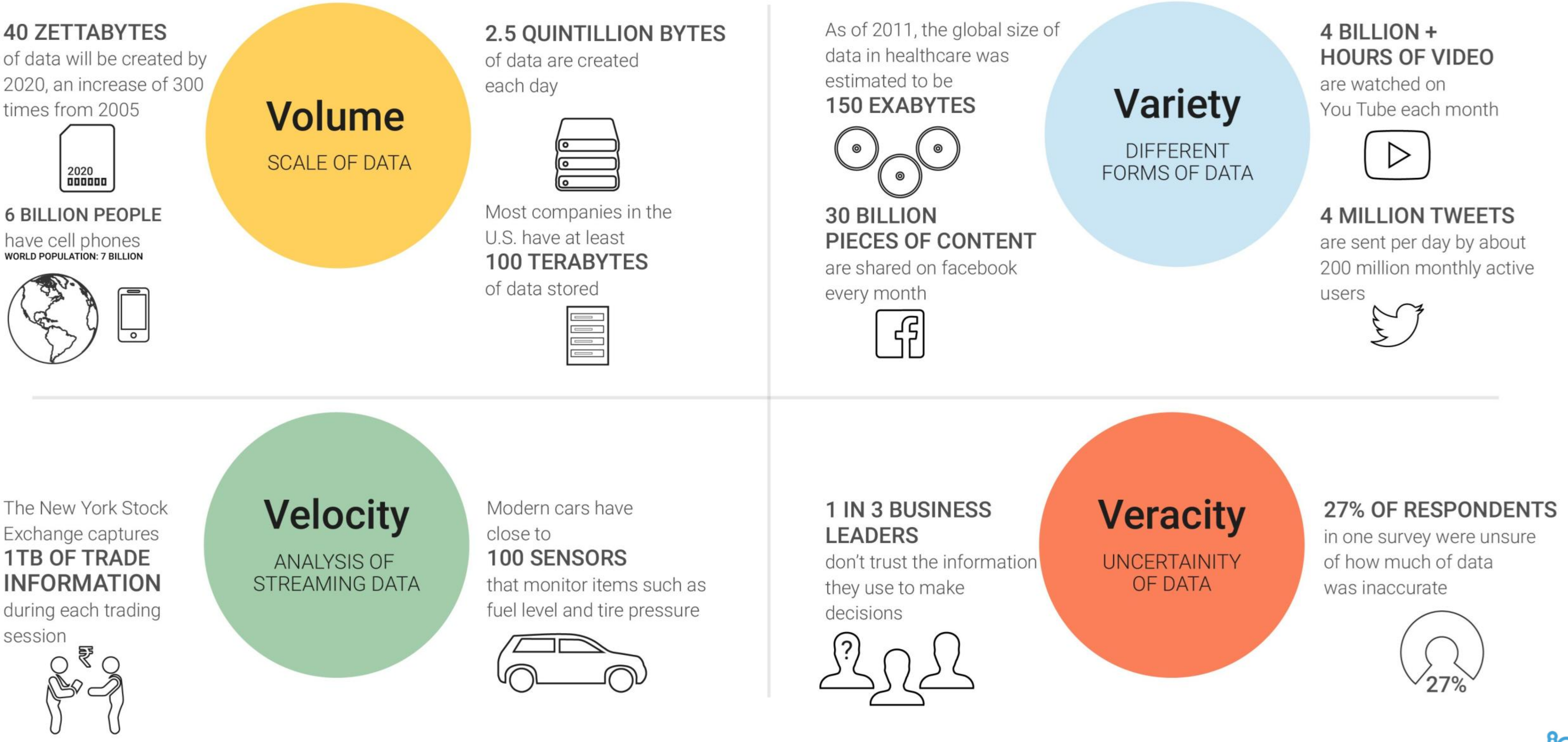
- **Big Data: Eaxmple**



Data Type	Example
Visual (images/ videos)	• X-rays • Computerized tomography (CT or CAT scan) • Positron Emission Tomography (PET scan) • Magnetic Resonance Imaging (MRI)
Speech	audio or voice reports
Numerical	• Patient card • Test results
Text	• Medical reports • Reviews
Multimodal	Electronic Medical Records (EMR)

Datafication

- **Big Data: The 4 V's**



Datafication

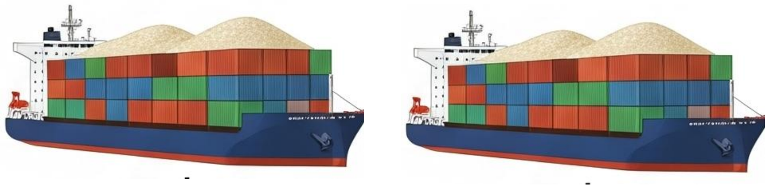
- **Big data sizes**

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

Datafication

- **How much training data is required for machine learning?**

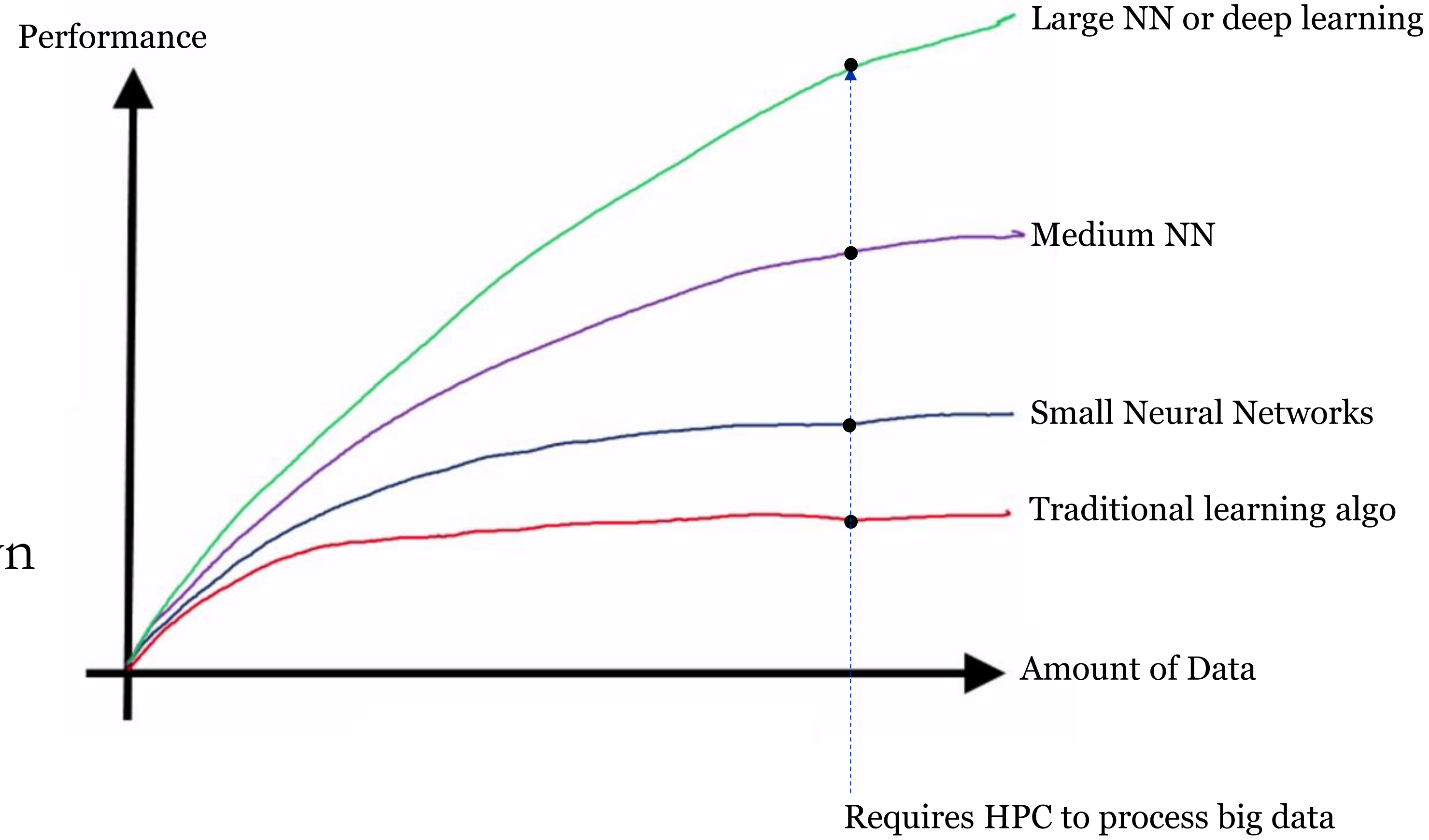
Better Data != More Data

Data Without a Sound Approach = Noise

- **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.



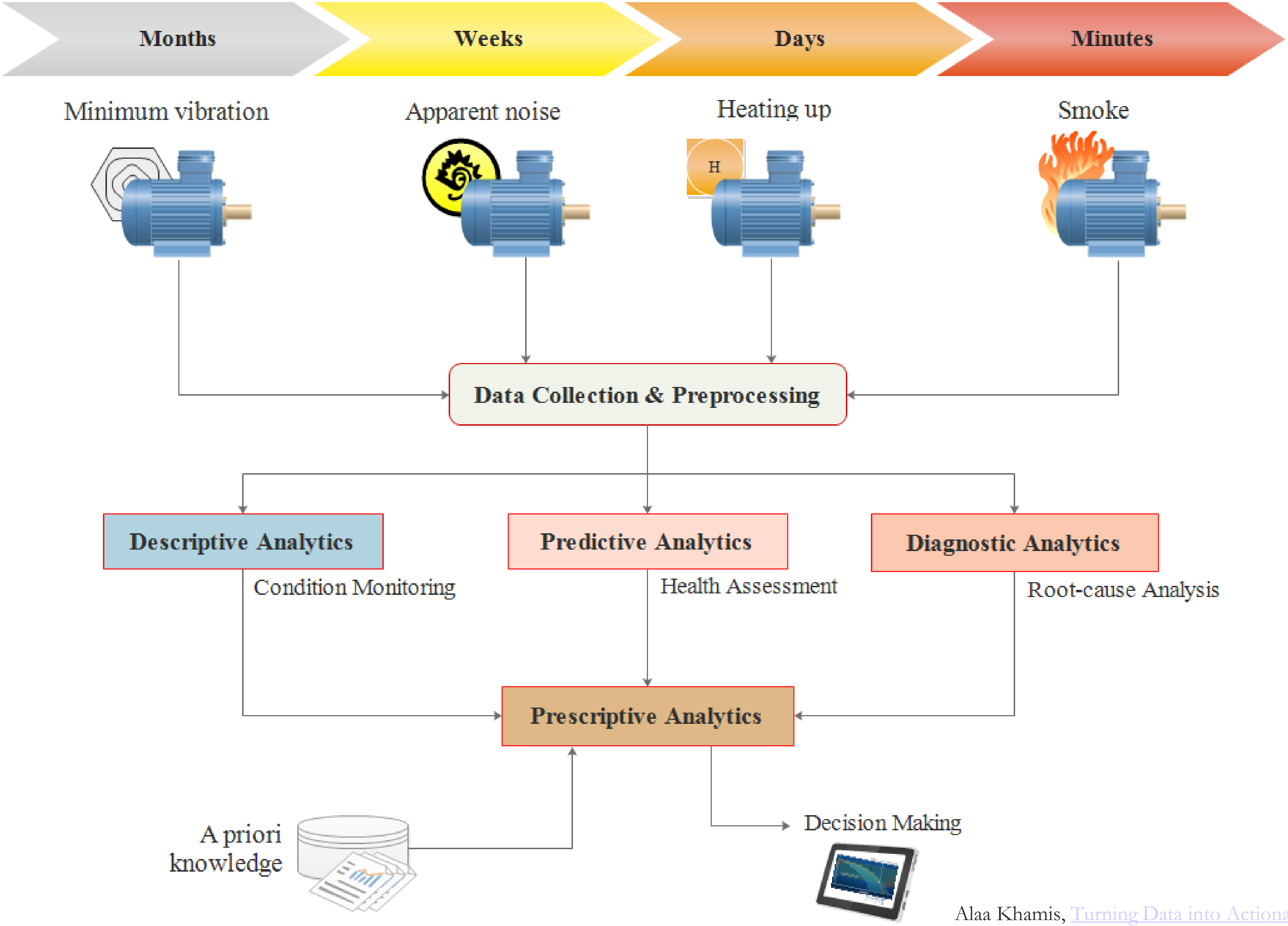
[Source]

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

Outline

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

Condition Monitoring Sensors



Alaa Khamis, Turning Data into Actionable Insights, Towards data science, March 30, 2019

Condition Monitoring Sensors

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

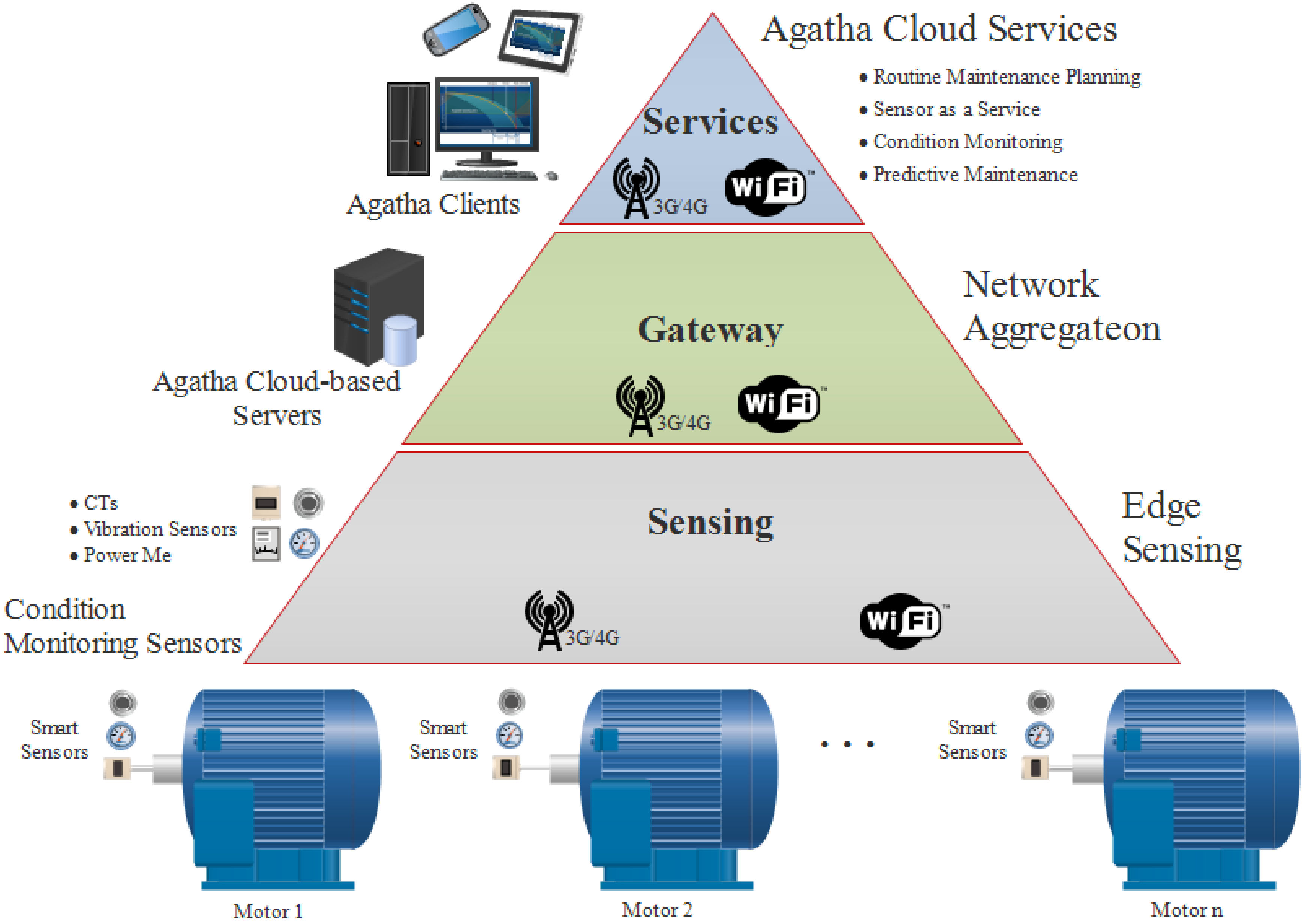
Power meter

CT sensors

Vibration Sensor

Temperature Sensor

Condition Monitoring Sensors



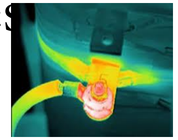
Condition Monitoring Sensors

- **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

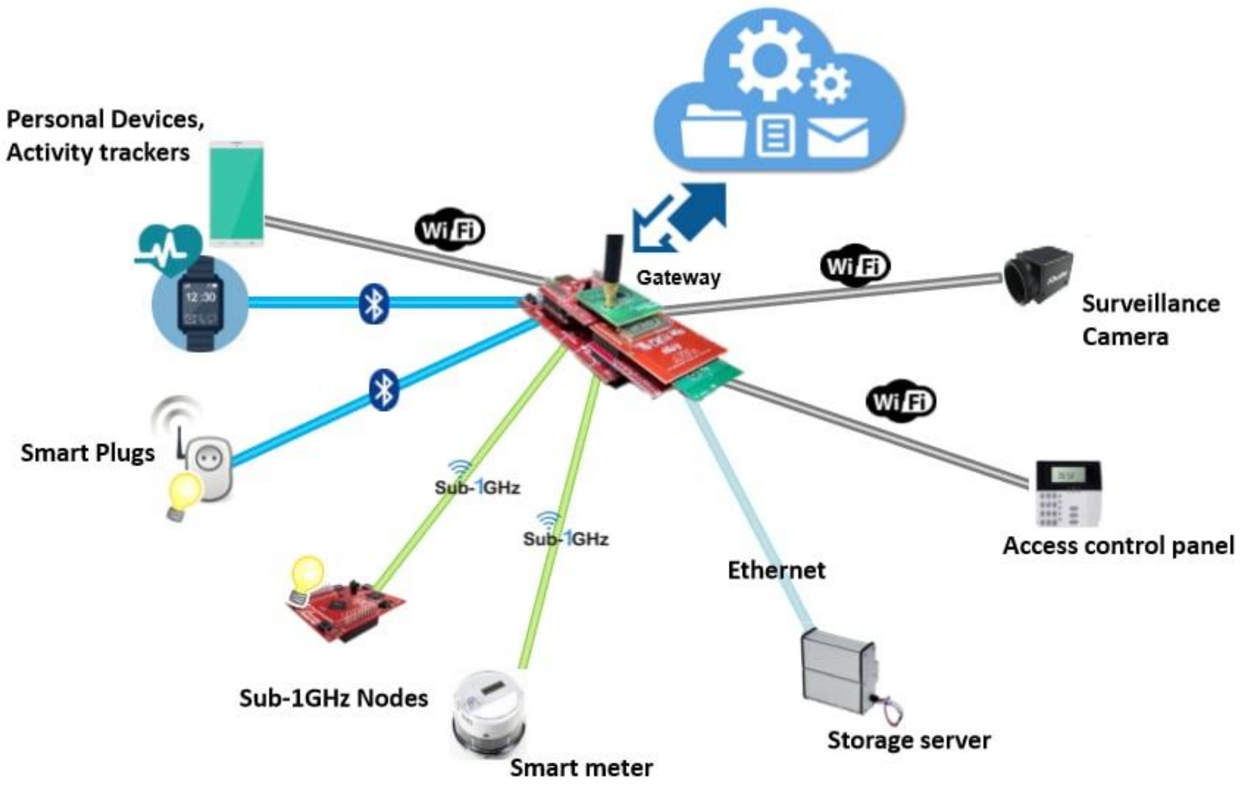
Condition Monitoring Sensors

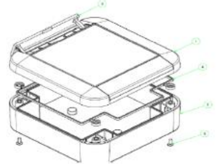
- **Power Meter:** 3 phase power meter with current and voltage transducers are mandatory to monitor the electric panels status (On/Off, normal or overload)
- **Temperature Sensors:** provide continuous, real-time data on equipment heat levels without disrupting operations. By detecting abnormal temperature 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.



Condition Monitoring Sensors

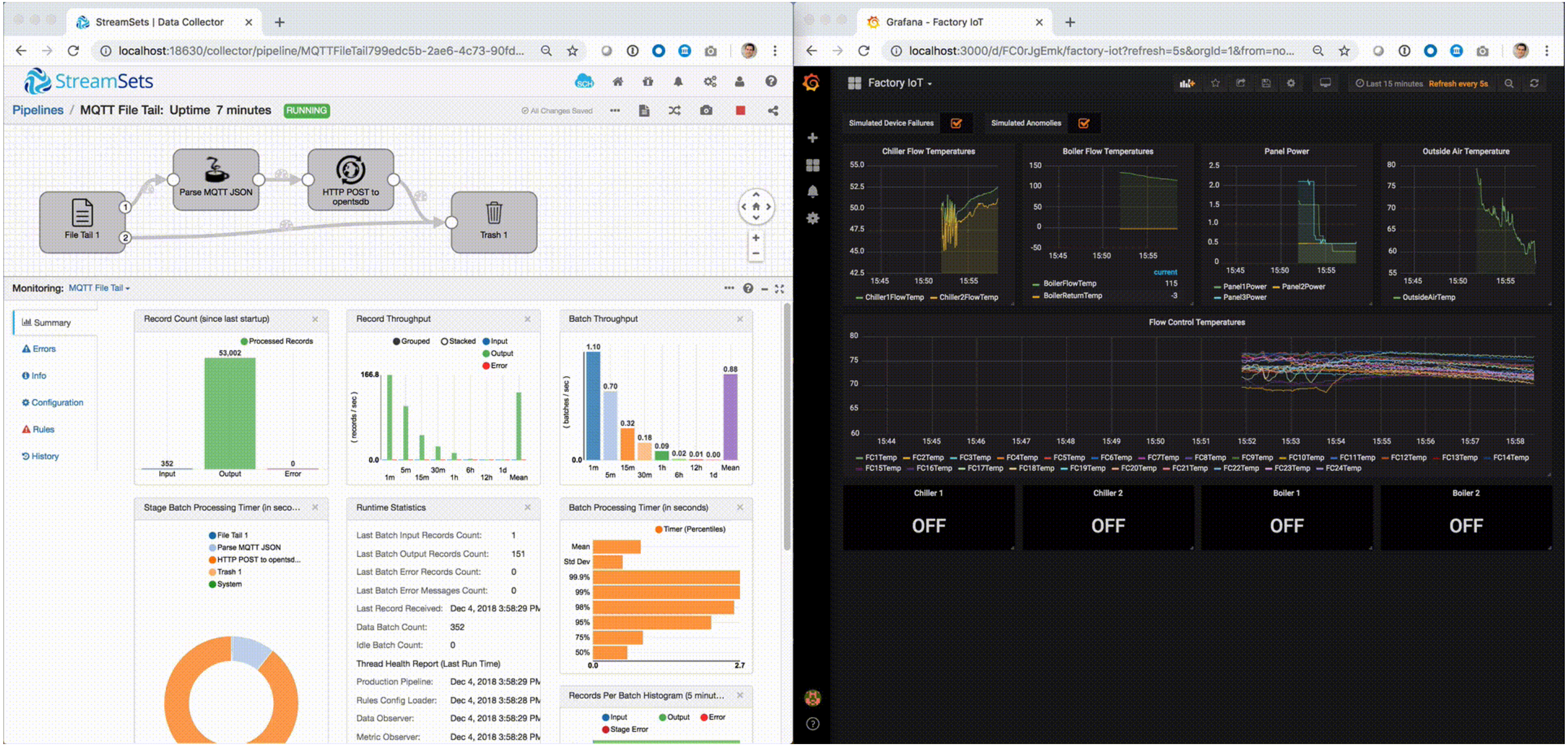
- **IoT Gateway:** Multi-communication channels to collect the data from the different sensors and direct this data to the cloud server to be analyzed. The following technical specifications are required:
  - Wi-Fi, Ethernet, RS 485, 3G/4G Connectivity
  - 12 industrial sensor Analog input (4-20 ma)
  - 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:**



Outline

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

Data Types

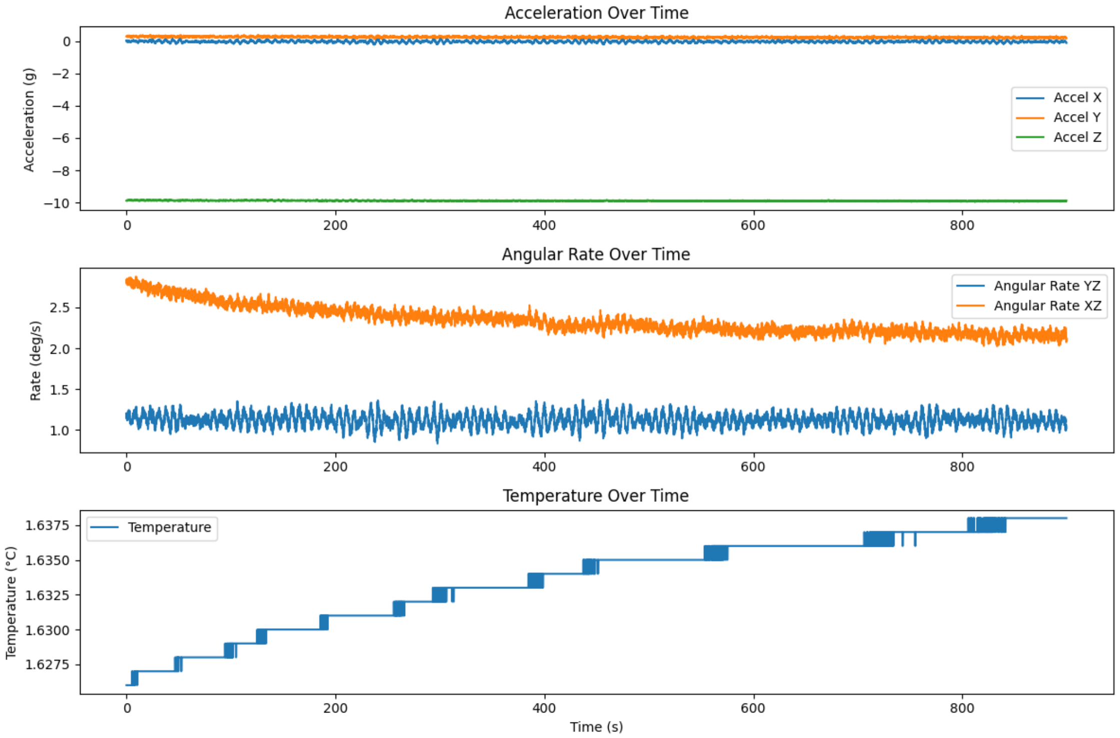
Data Type	Example Sources
Time-Series	Vibration, temperature, pressure, flow, voltage
Acoustic/Ultrasonic	Leak detection, bearing diagnostics
Thermal/Imaging	IR cameras, visual inspections
Event/State Logs	PLC alarms, operational status
Historical Structured	CMMS data, asset metadata
Analytical/Lab	Oil/lubricant,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

Data Types

- **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)



https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/5_Datafication/Time_series_data.ipynb

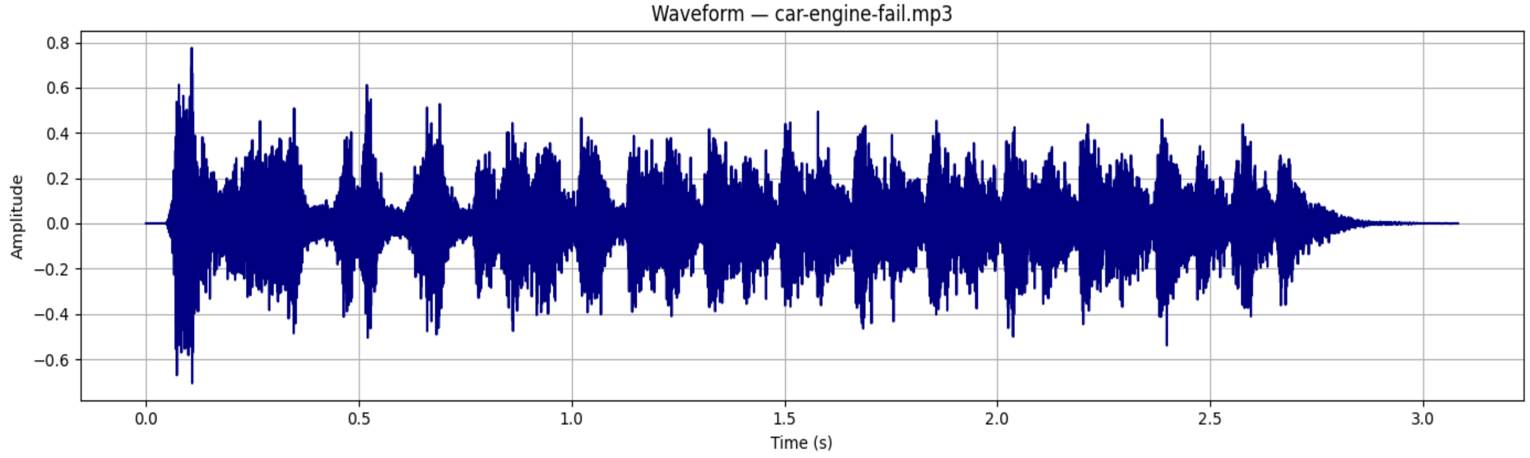
Data Types

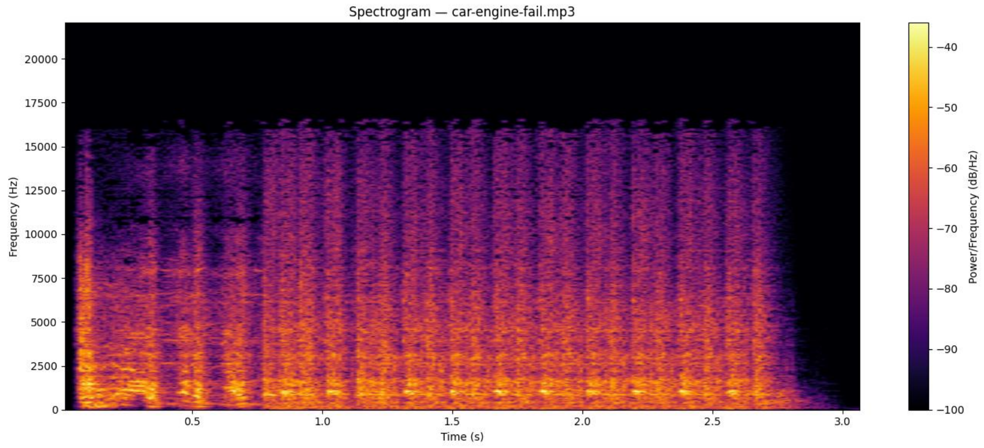
- **Acoustic & Ultrasonic Data**

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
```





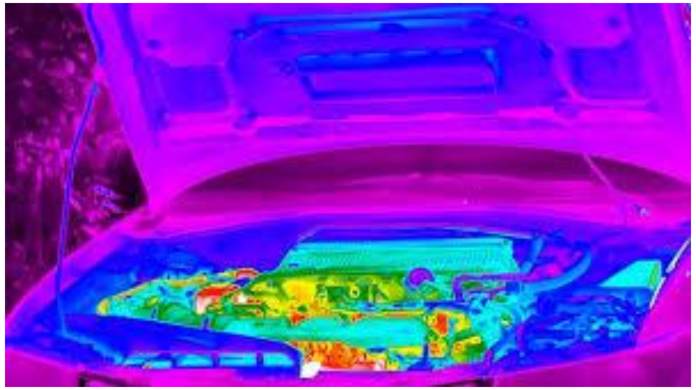
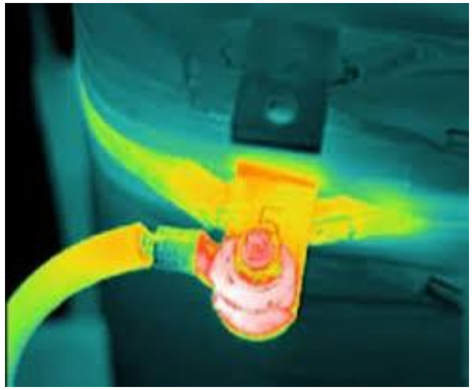
https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/5_Datafication/Acoustic_data.ipynb

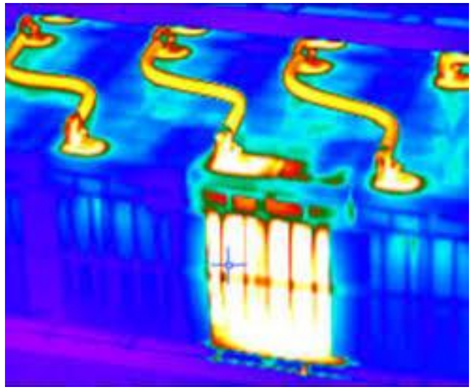
Data Types

- **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

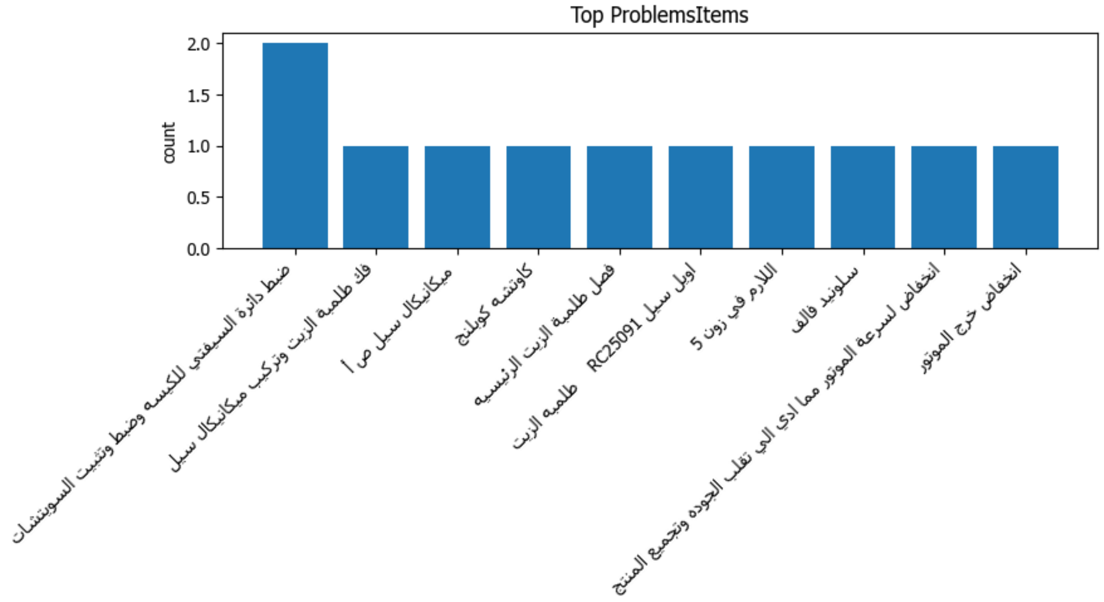


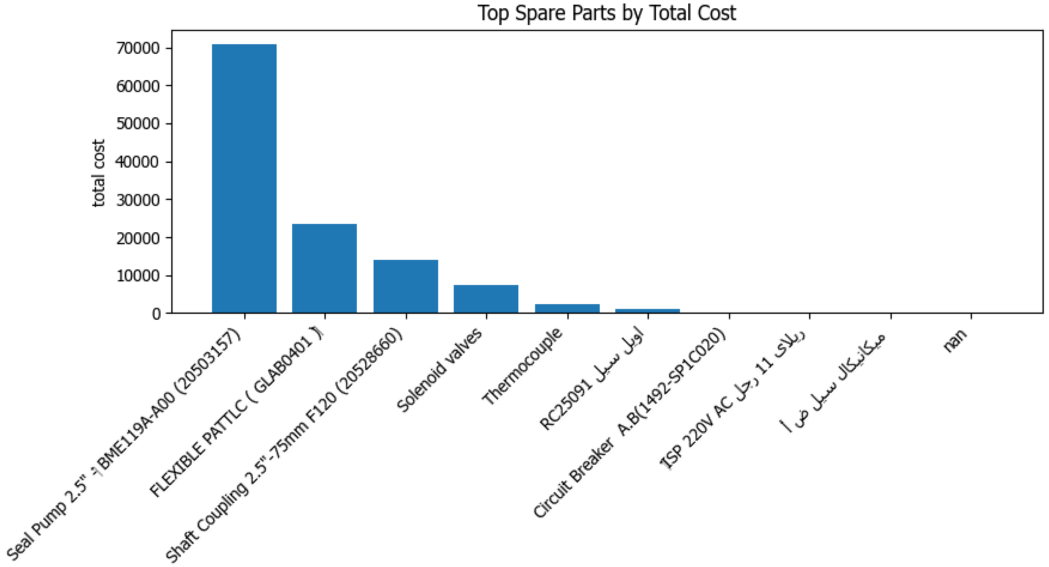
https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/5_Datafication/Infrared_Imaging_data.ipynb

- **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.
أعلى تكلفة إجمالية لقطع الغيار: 'Seal Pump 2.5" - BME119A-A00 (20503157)' بقيمة 70,952.33.
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).
```

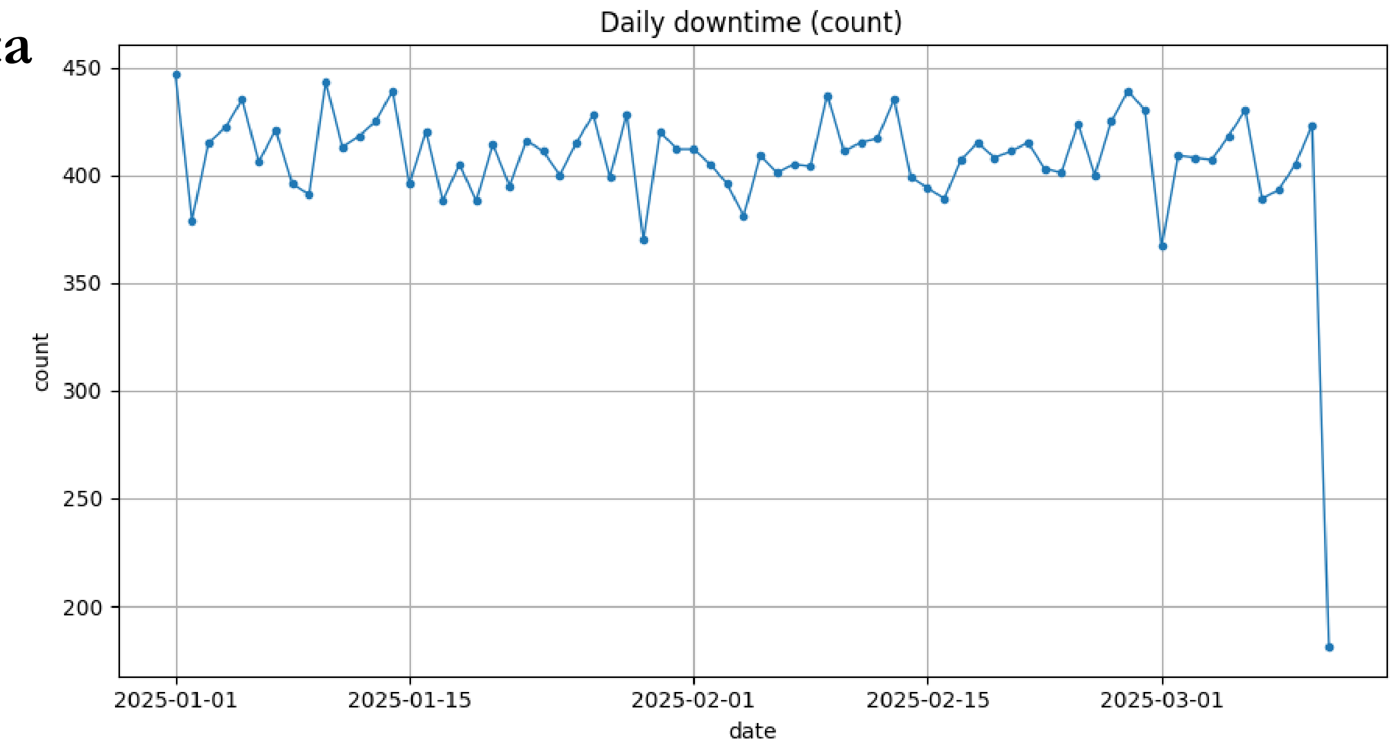
https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/5_Datafication/Log_data.ipynb

Data Types

- **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

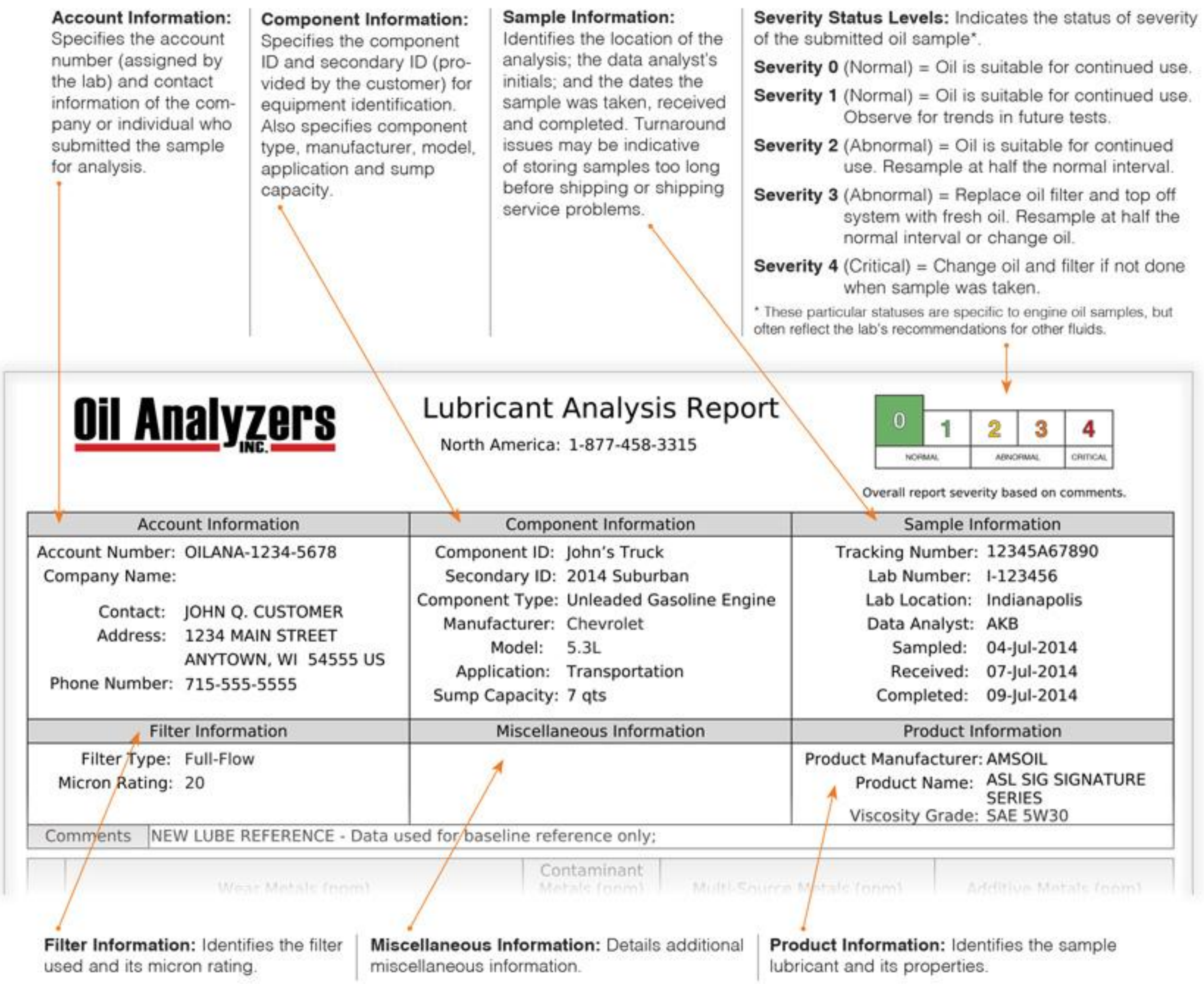
https://colab.research.google.com/github/Dr-AlaaKhamis/ISE518/blob/main/5_Datafication/Historical_data.ipynb

Data Types

- **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

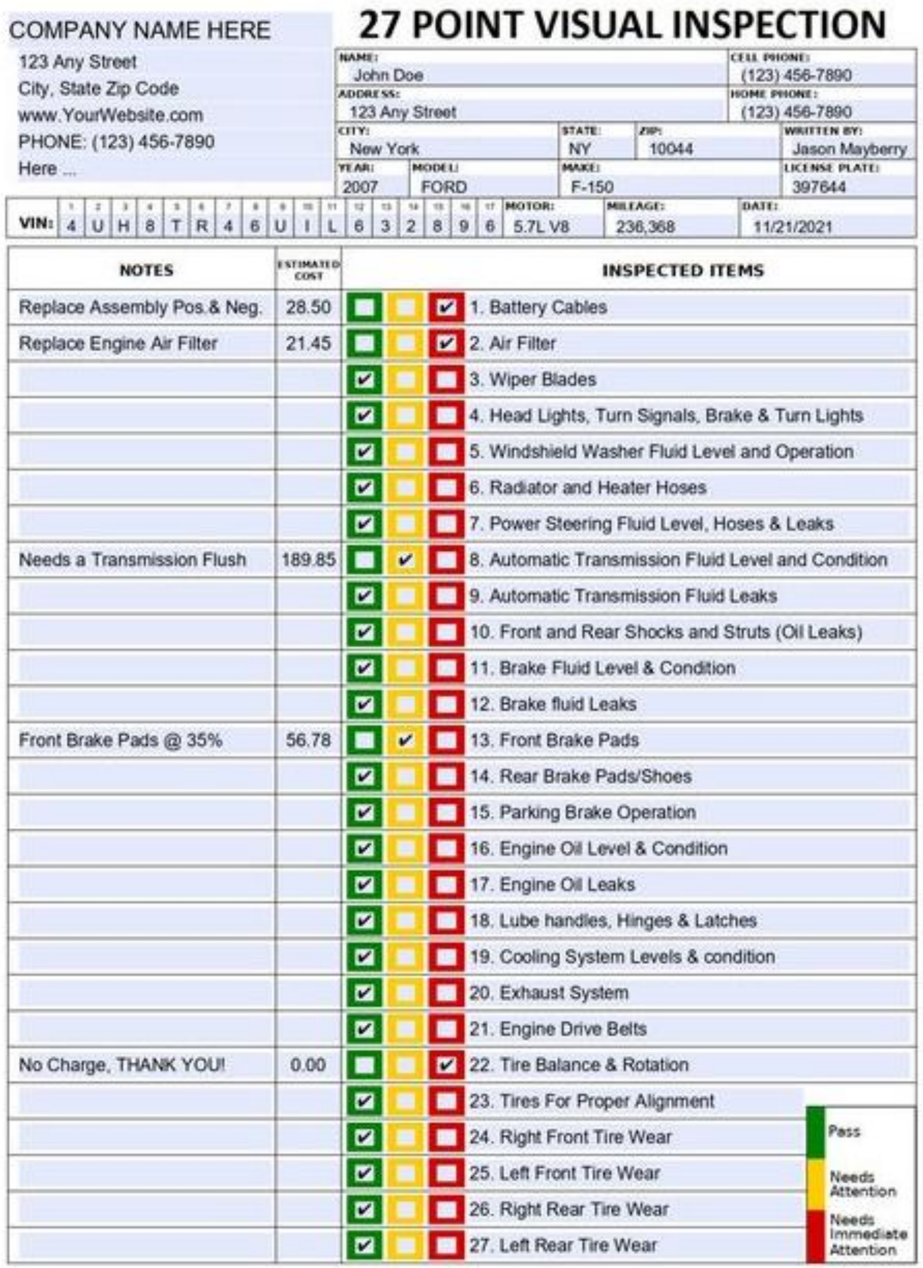


Data Types

- **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
123 Any Street
City, State Zip Code
www.YourWebsite.com
PHONE: (123) 456-7890
Here ...

27 POINT VISUAL INSPECTION

NAME:	CELL PHONE:
John Doe	(123) 456-7890
ADDRESS:	HOME PHONE:
123 Any Street	(123) 456-7890

CITY:	STATE:	ZIP:	WRITTEN BY:
New York	NY	10044	Jason Mayberry
YEAR:	MODEL:	MAKE:	LICENSE PLATE:
2007	FORD	F-150	397644

VIN:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	MOTOR:	MILEAGE:	DATE:
	4	U	H	8	T	R	4	6	U	I	L	6	3	2	8	9	6	5.7L V8	236,368	11/21/2021

NOTES	ESTIMATED COST				INSPECTED ITEMS
Replace Assembly Pos.& Neg.	28.50			✔	1. Battery Cables
Replace Engine Air Filter	21.45			✔	2. Air Filter
		✔			3. Wiper Blades
		✔			4. Head Lights, Turn Signals, Brake & Turn Lights
		✔			5. Windshield Washer Fluid Level and Operation
		✔			6. Radiator and Heater Hoses
		✔			7. Power Steering Fluid Level, Hoses & Leaks
Needs a Transmission Flush	189.85		✔		8. Automatic Transmission Fluid Level and Condition
		✔			9. Automatic Transmission Fluid Leaks
		✔			10. Front and Rear Shocks and Struts (Oil Leaks)
		✔			11. Brake Fluid Level & Condition
		✔			12. Brake fluid Leaks
Front Brake Pads @ 35%	56.78		✔		13. Front Brake Pads
		✔			14. Rear Brake Pads/Shoes
		✔			15. Parking Brake Operation
		✔			16. Engine Oil Level & Condition
		✔			17. Engine Oil Leaks
		✔			18. Lube handles, Hinges & Latches
		✔			19. Cooling System Levels & condition
		✔			20. Exhaust System
		✔			21. Engine Drive Belts
No Charge, THANK YOU!	0.00			✔	22. Tire Balance & Rotation
		✔			23. Tires For Proper Alignment
		✔			24. Right Front Tire Wear
		✔			25. Left Front Tire Wear
		✔			26. Right Rear Tire Wear
		✔			27. Left Rear Tire Wear

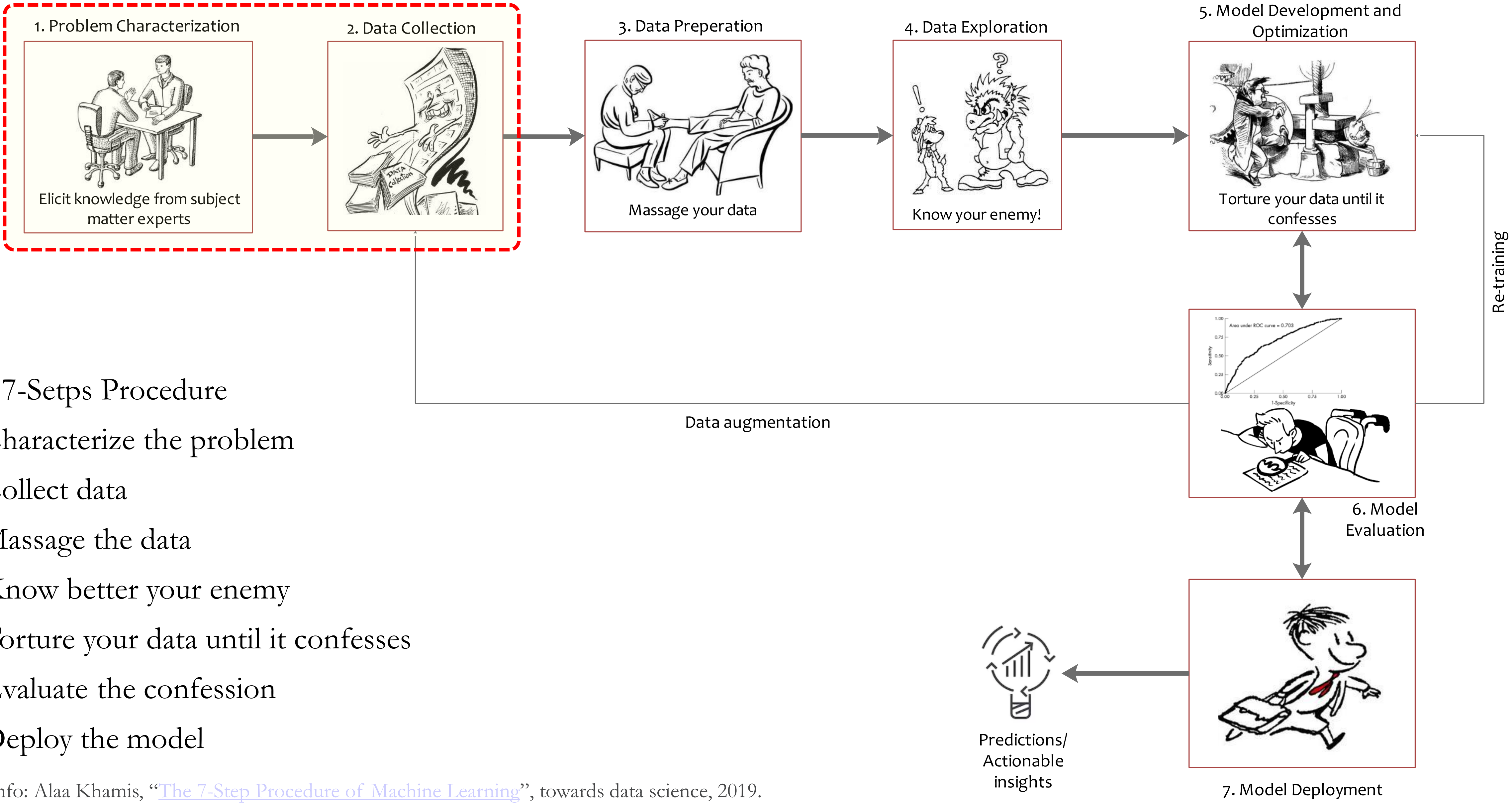
Pass
Needs Attention
Needs Immediate Attention

Outline

- Datafication
- Condition Monitoring Sensors
- Data Types
- **<u>Design of Experiment (DOE)</u>**
- Data Governance

Design of Experiment (DOE)

- **The 7-Step Procedure**



The 7-Setps Procedure

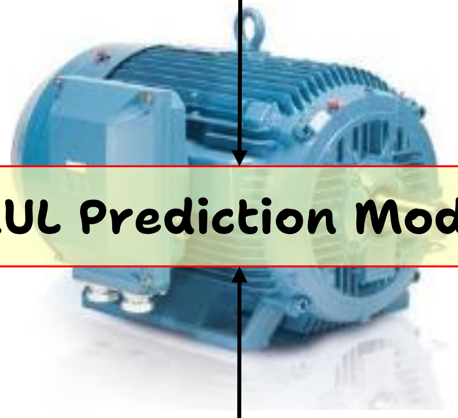
1. Characterize the problem
2. Collect data
3. Massage the data
4. Know better your enemy
5. Torture your data until it confesses
6. Evaluate the confession
7. Deploy the model

More info: Alaa Khamis, "The 7-Step Procedure of Machine Learning", towards data science, 2019.

Design of Experiment (DOE)

Controlled Factors

Independent variables that represent design parameters changeable during data collection process



RUL Prediction Model

Signals

Independent variables or stimuli required for fulfilling the model functionality

Response

Dependent variables that represent primary intended functional output of the model

Noise Factors

Independent variables that influence prediction model, controllable during data collection process and uncontrollable after deploying the model

Error States

Failure modes or effects of failure as defined by end-user when using the model

Design of Experiment (DOE)


C0: Safe Driving


C1: Text Right


C2: Phone Right


C3: Text Left


C4: Phone Left


C5: Adjusting Radio


C6: Drinking

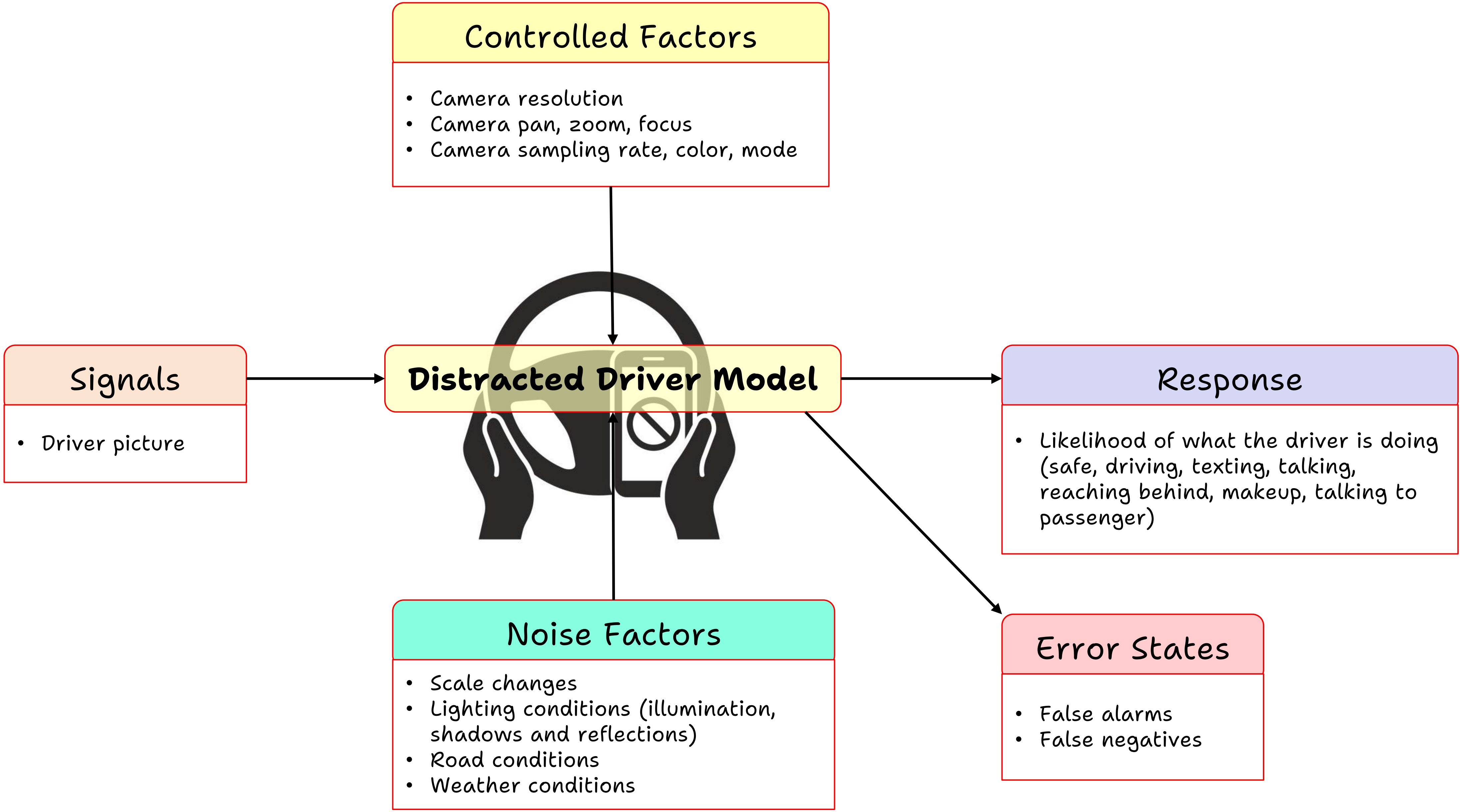

C7: Reaching Behind


C8: Hair or Makeup


C9: Talking to Passenger

More info: State Farm Distracted Driver Detect Dataset

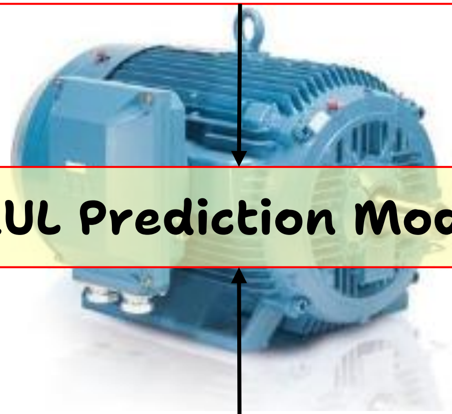
Design of Experiment (DOE)



Design of Experiment (DOE)

Controlled Factors

- Motor Type
- Supply voltage
- Ambient temperature
- Ambient humidity
- Sensor type and placement
- Sampling frequency



Signals

- Motor vibration signals
- Temperature readings
- Current and voltage measurements
- Rotational speed (RPM)
- Acoustic signals

RUL Prediction Model

Response

- Predicted Remaining Useful Life (RUL) in hours or cycles
- Probability distribution of time-to-failure
- Confidence interval of RUL estimate

Noise Factors

- Random load fluctuations from production process
- unexpected voltage dips/spikes
- operator handling differences
- environmental dust/contaminants

Error States

- Prediction error (e.g., RMSE or MAE)
- Biased RUL estimates
- False alarms and false negatives

Design of Experiment (DOE)

Controlled Factors
- ?

Signals
- ?

Tire Treadwear Prediction Model

Response
- ?

Noise Factors
- ?

Error States
- ?

Design of Experiment (DOE)

- **How do I set up an experiment?**

Design of Experiment (DOE)

Full Factorial Design

Fractional Factorial Design

Pros & Cons	Full Factorial	Fractional Factorial
Pros	• All possible combinations can be covered • Analysis is straightforward, as there is no aliasing	• Less memory and effort • Less time • Runs can be added to eliminate cofounding.
Cons	Cost of the experiment increases as the number of factors increases	• Analysis of higher order interactions could be complex • Cofounding could mask factor and interaction effects

Design of Experiment (DOE)

- **Full Factorial Design**

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

N_{full} is the number of runs

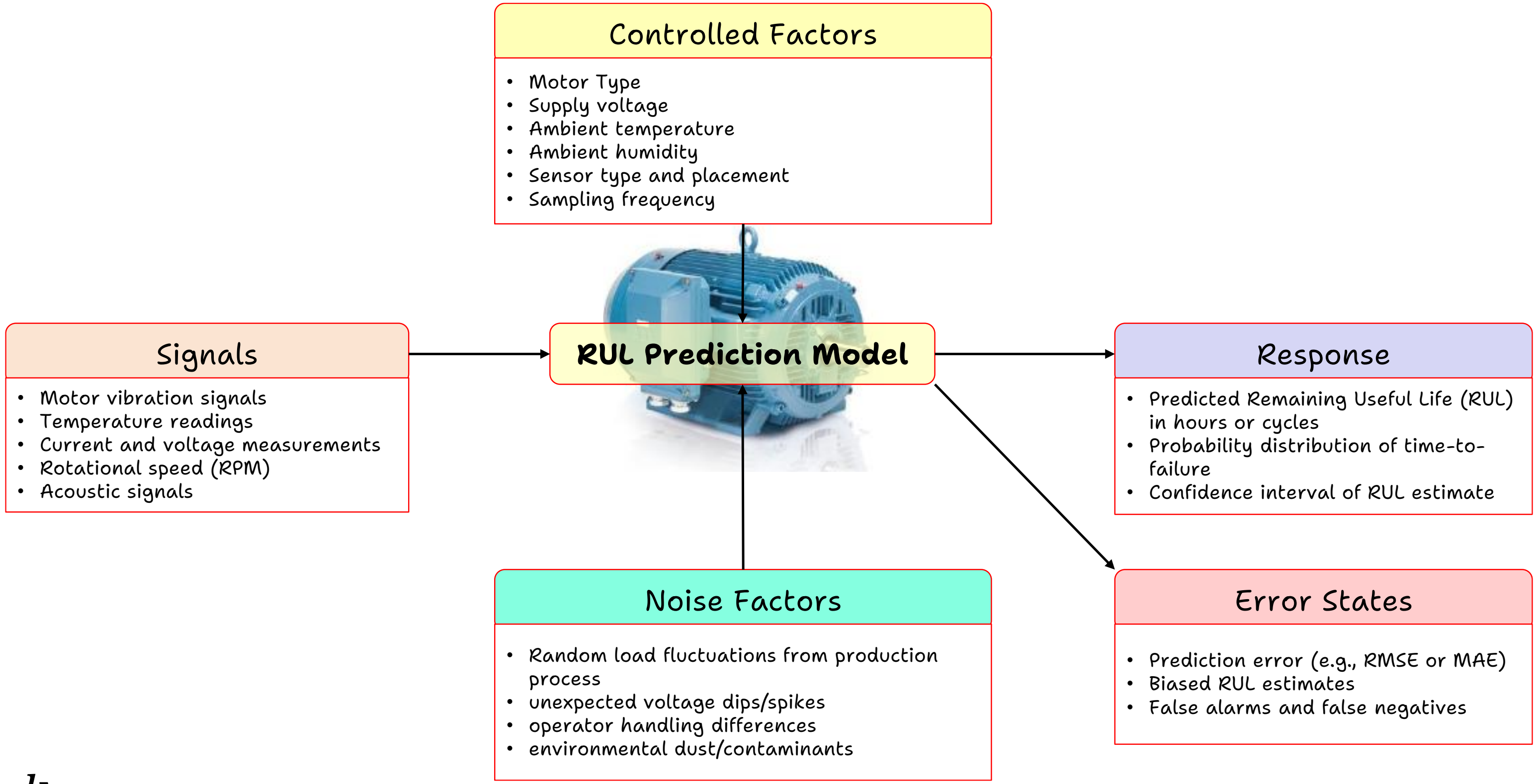
L_i is number of levels for factor $\boldsymbol{i}$

r is number of replicates (optional)

<u>Special cases</u>

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**).

Design of Experiment (DOE)

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

$$N_{frac} = r2^{k-p}$$

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**

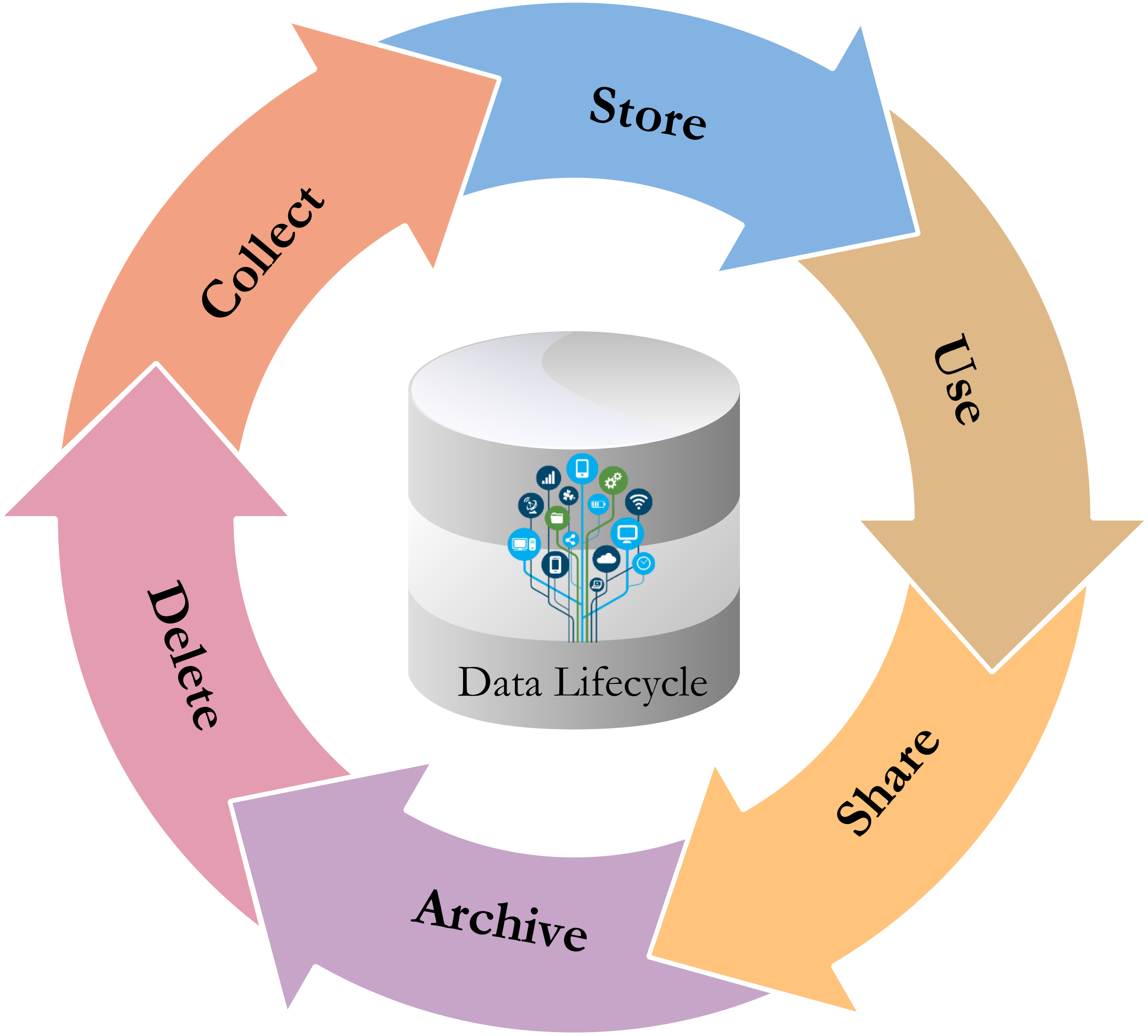
$$N_{frac} = r3^{k-p}$$

Outline

- Datafication
- Condition Monitoring Sensors
- Data Types
- Design of Experiment (DOE)
- **<u>Data Governance</u>**

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?