CSC345/M45:
Cluster Analysis: example problems

Dr. Xianghua Xie

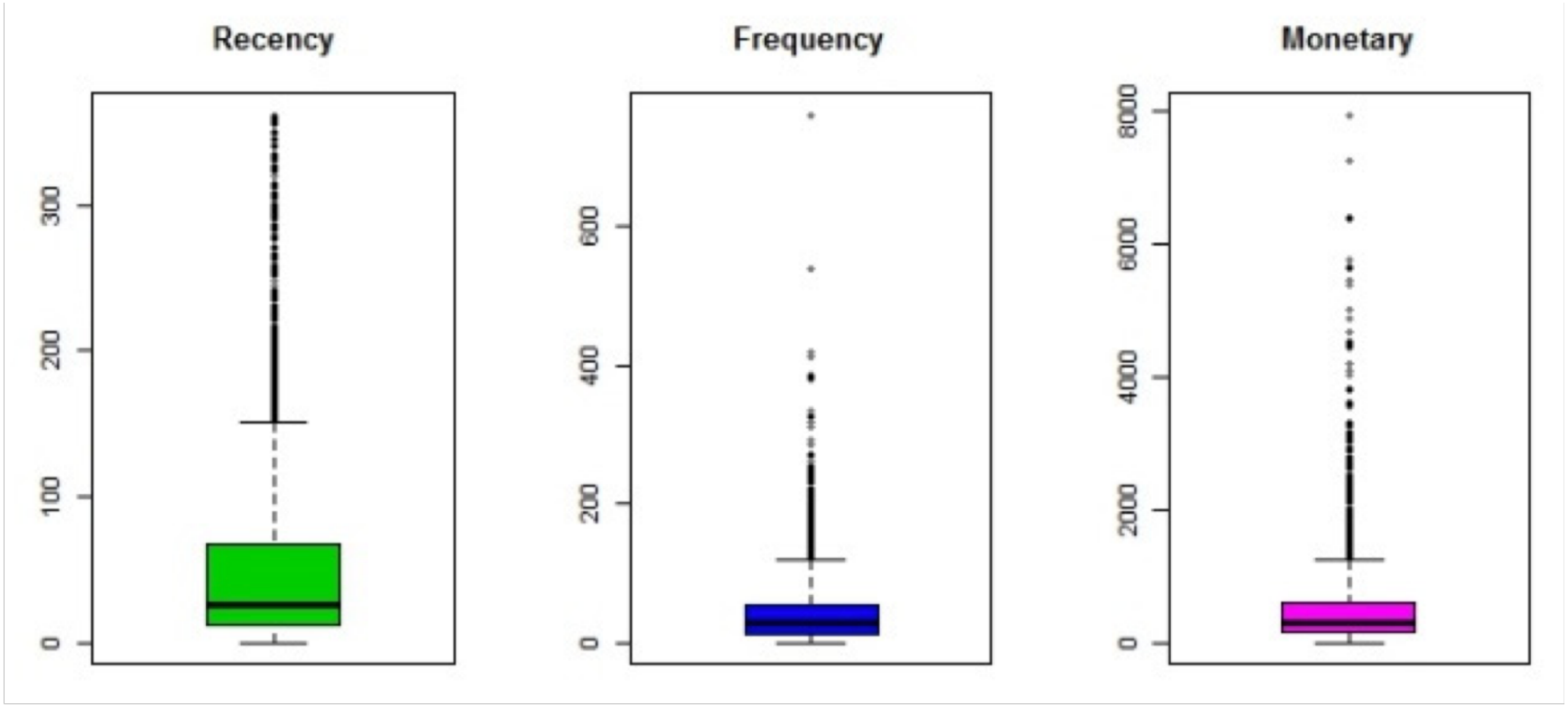
x.xie@swansea.ac.uk
http://csvision.swan.ac.uk
224 Computational Foundry, Bay Campus

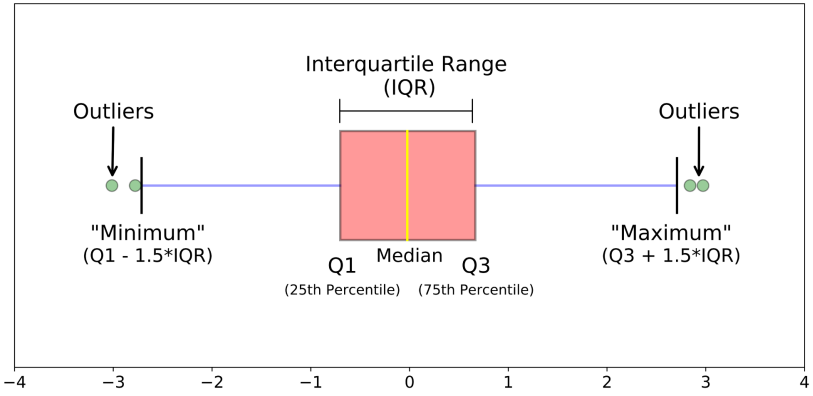
1. Consumer Value Analysis

- Consumer value analysis can be used to identify profitable customers and to develop strategies to target customers.
  - RFM is widely used to characterise customers
  - R: Recency; represents the time since the last purchase;
  - F: Frequency; denotes the number of purchases within a given period;
  - M: Monetary; means the amount of money spent in this period.

- Given RFM data, how would you identify customers with similar behaviours?

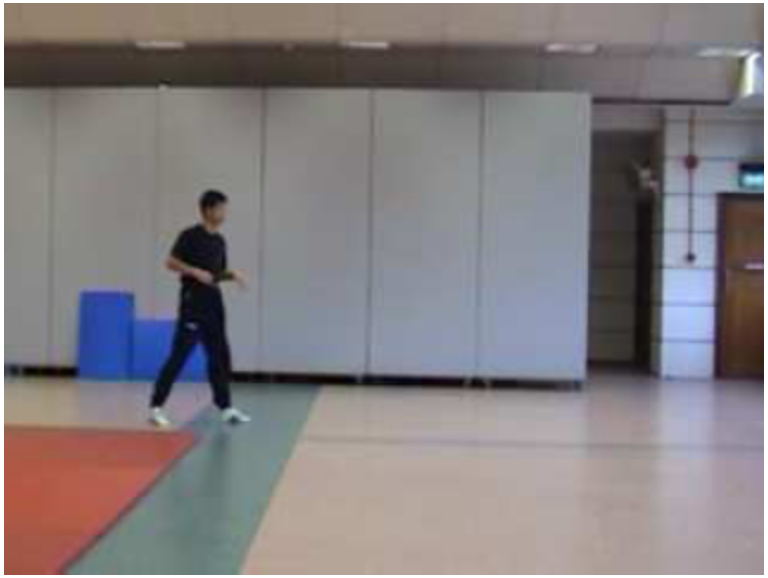
RFM data (boxplot)

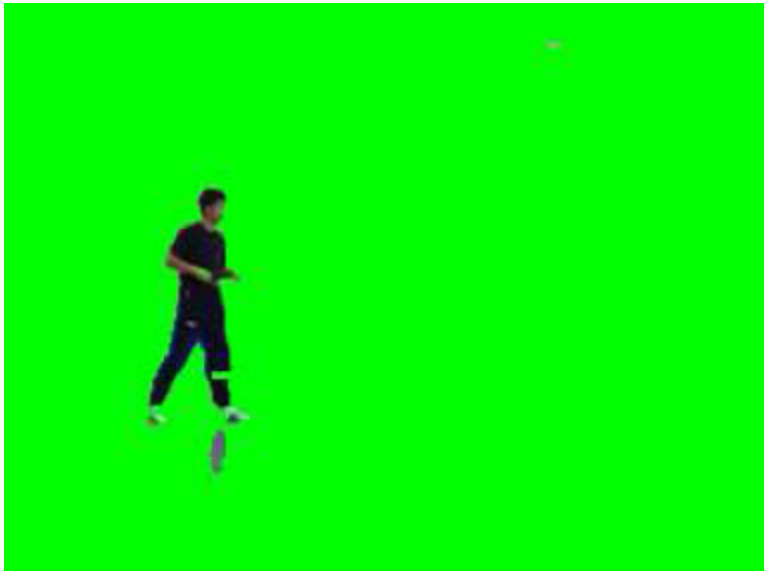




2. Foreground Segmentation (outlier detection)

- From video input to detect objects
  - Applications include security, augmented reality, virtual reality.

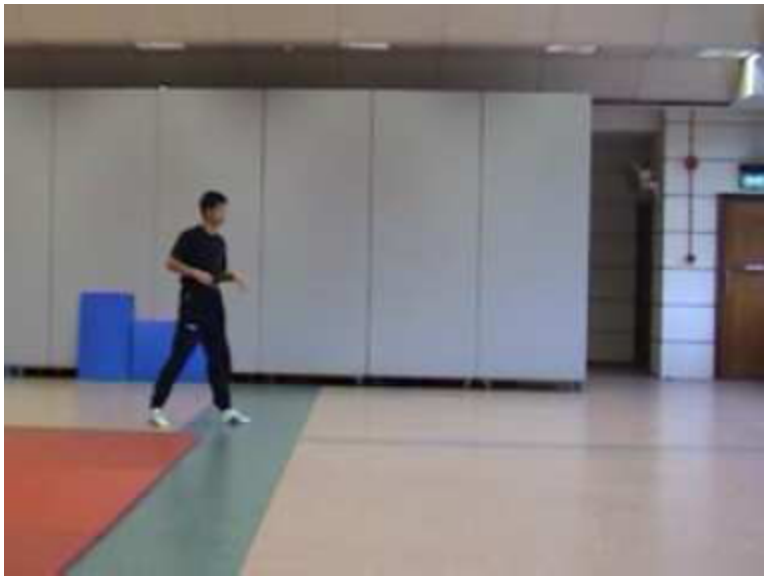


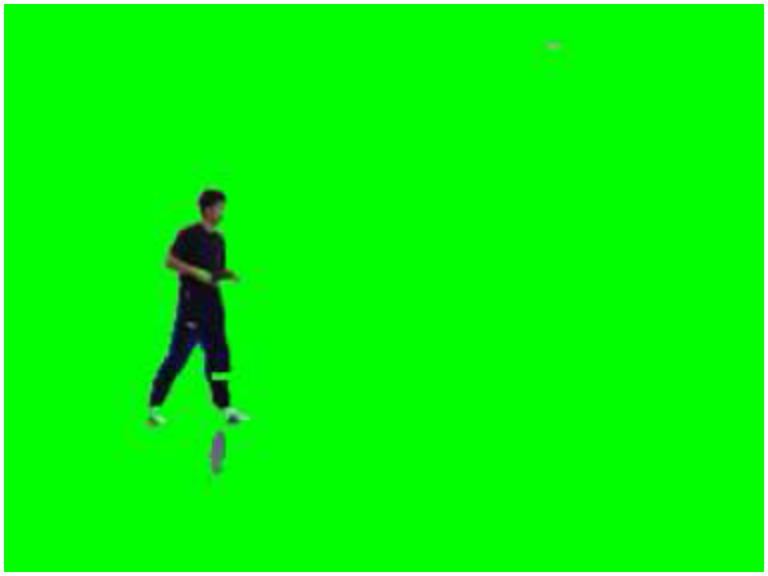


2. Foreground Segmentation (outlier detection)

- Treat this as outlier detection
  - Background pixels are modelled using GMM
  - Every pixel in new frames are evaluated using GMM
    - If the probability of the pixels below certain threshold, those pixels will be kept for display
    - Otherwise, set the pixels as background (in this example, replace them in green colour)

Which probability?



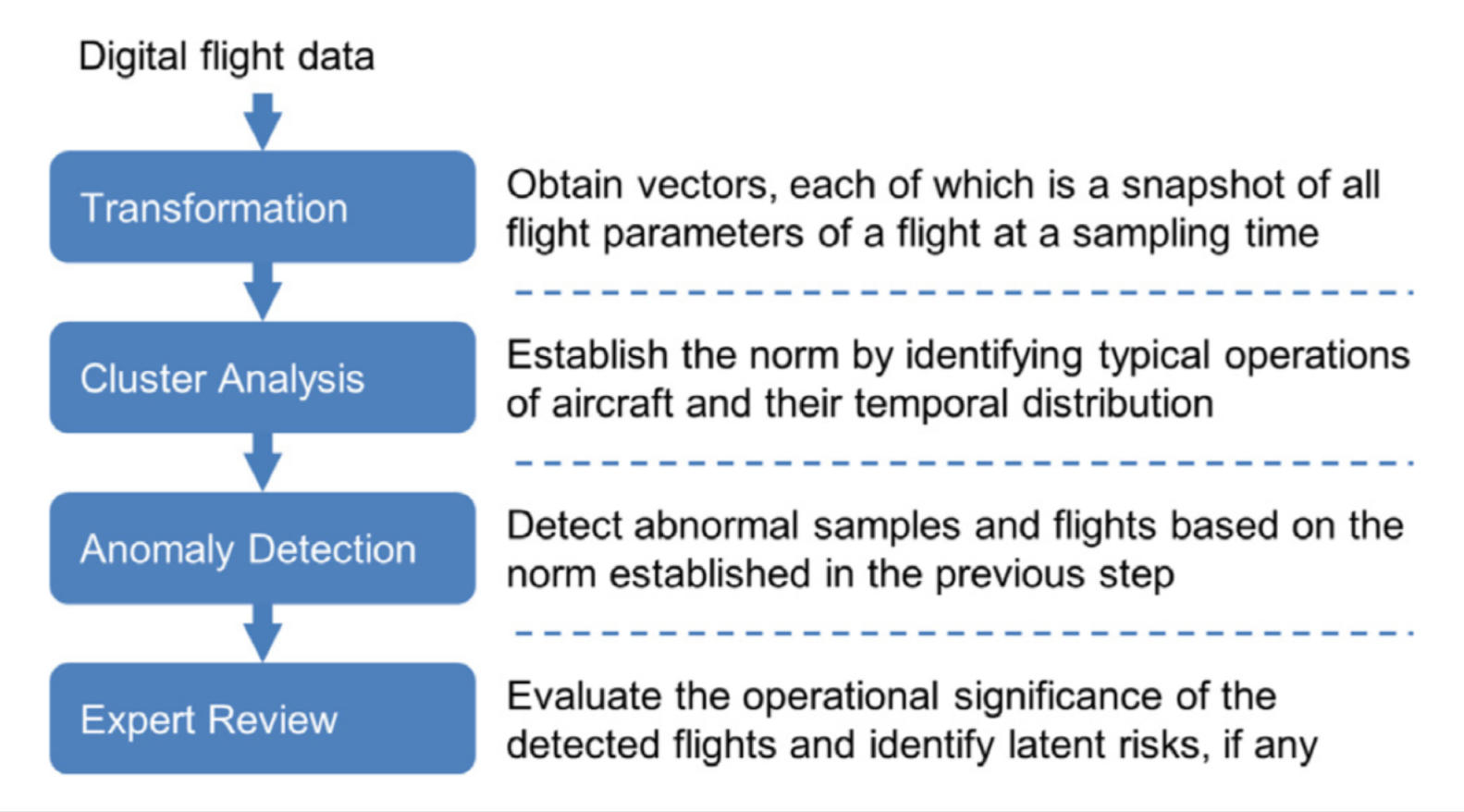


3. Anomaly detection for flight safety

- Safety remains a key element in air transportation, an industry which moves over 3 billion passengers a year.
- Proactive approach aims at continuously monitoring flight operations and identifying risks; mitigation measures are taken before incidents occur.
- The use of sensor data is critical.
  - Flight Data Recorder (FDR) or Quick Assess Recorder (QAR) onboard every aircraft
  - Thousands of technical parameters are recorded throughout a flight, airspeed, altitude, pitch, roll, engine parameters, etc.
  - Current analysis tools are based on the Exceedance Detection
  - Thresholds depend specifically on aircraft types, flight phases, air- port conditions, and flying procedures, etc.

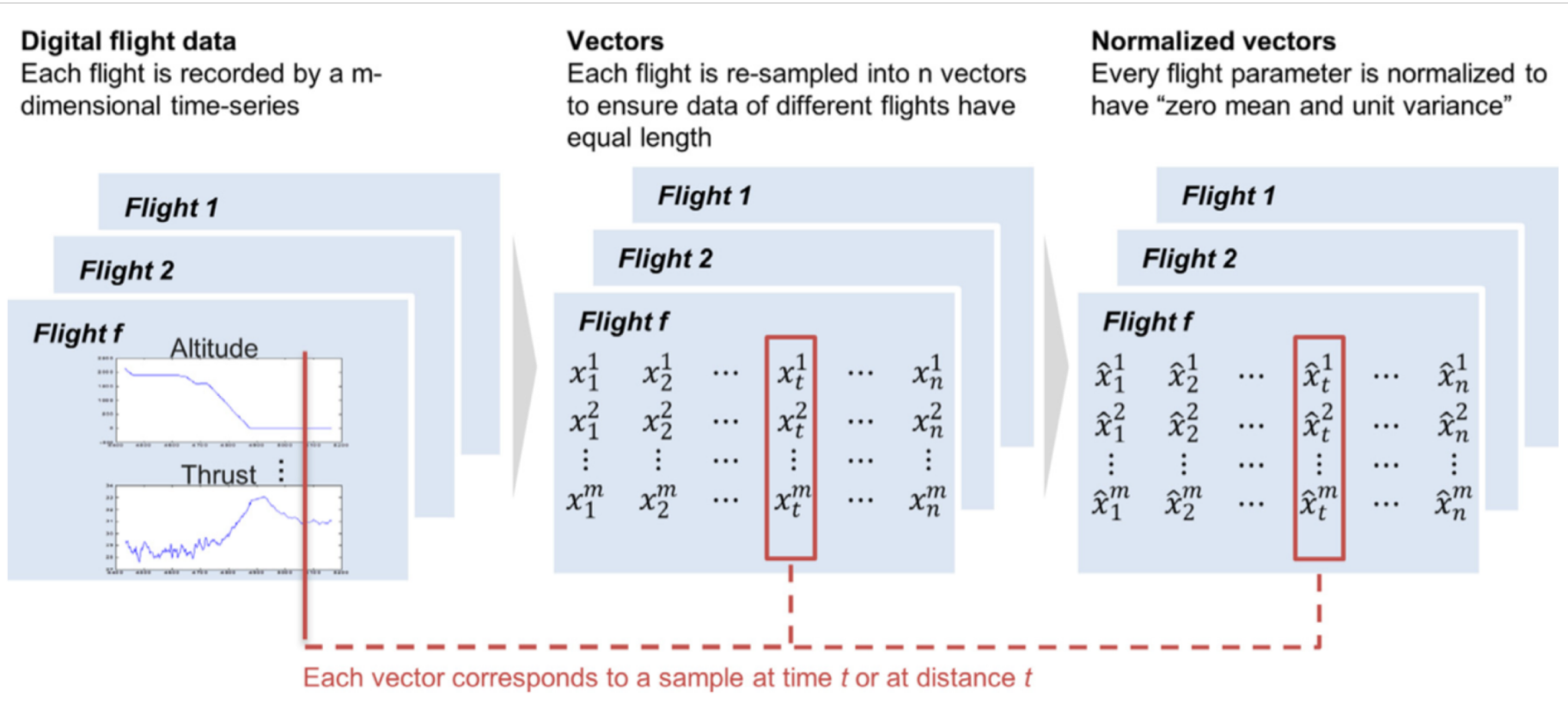
3. Anomaly detection for flight safety

- A data-driven approach



Li et al, Transportation Research, 2016

3. Anomaly detection for flight safety

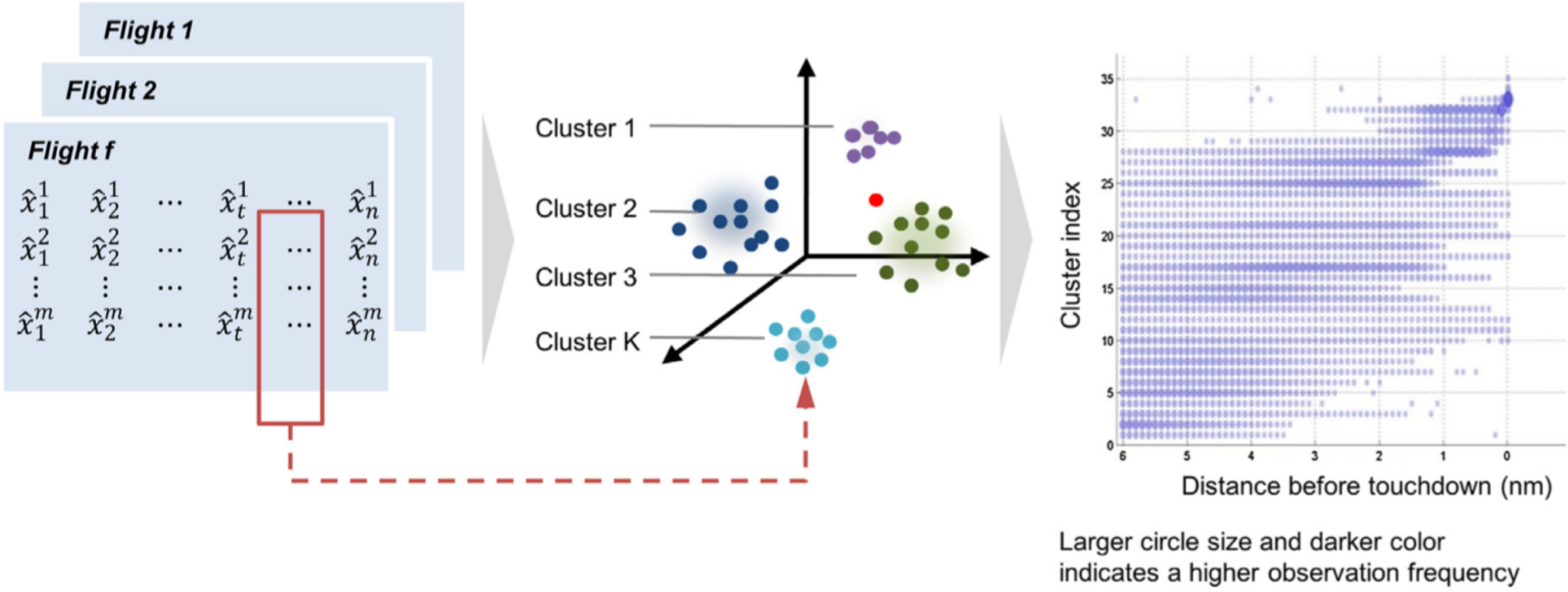


3. Anomaly detection for flight safety

Normalized vectors
Every flight parameter is normalized to have "zero mean and unit variance"

Clusters
GMM clustering is performed on normalized vectors; each cluster represent a typical operation of aircraft

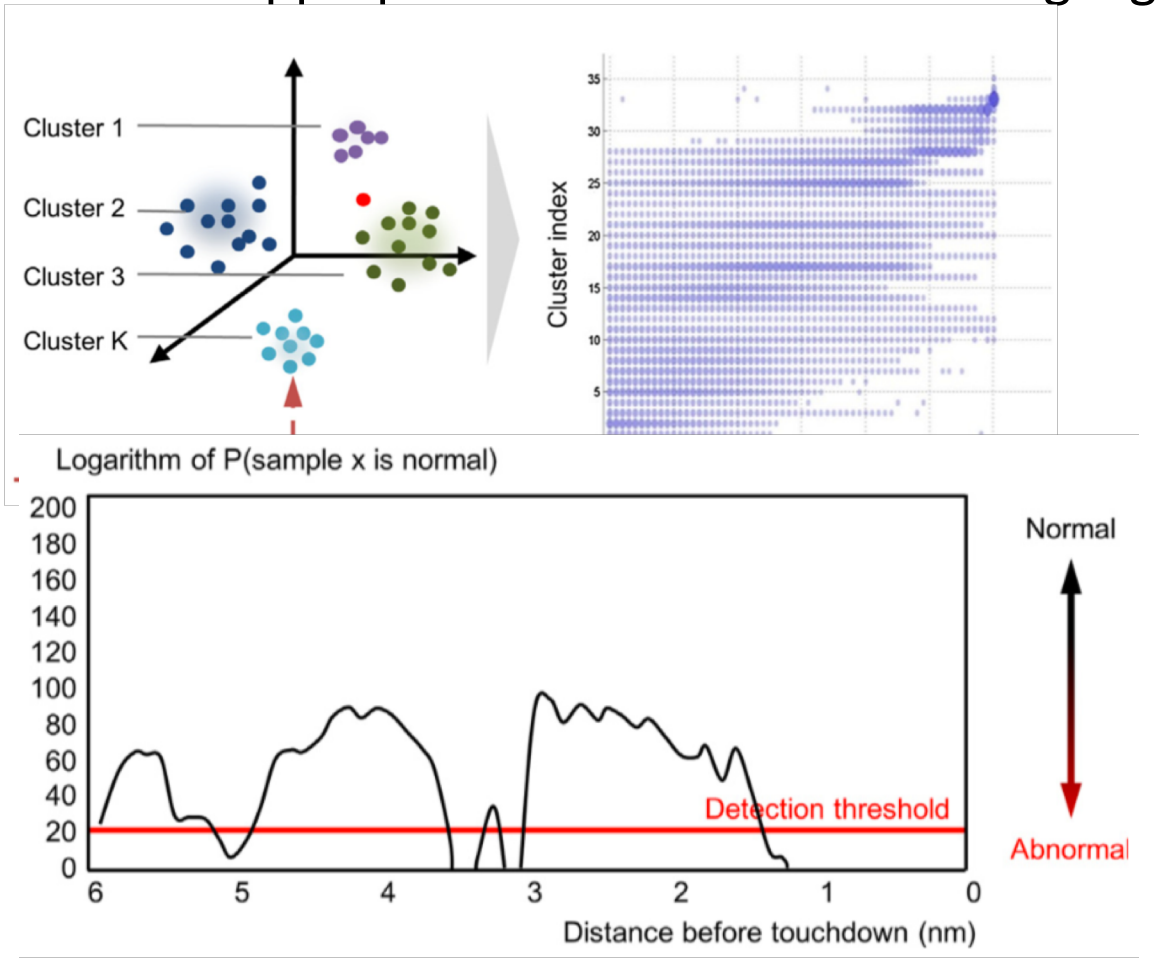
Temporal distribution of clusters
The temporal distribution of clusters is summarized by observation frequency of each cluster along the temporal reference



Larger circle size and darker color indicates a higher observation frequency

3. Anomaly detection for flight safety

- The final procedure for detecting anomalies is to compute the probability of each sample being normal, determined by
  - (1) how likely it is to belong to one of the clusters,
  - (2) the cluster is appropriate at that moment during flight.



4. Synthesis: Sampling from a distribution

- Generating new instances from GMM