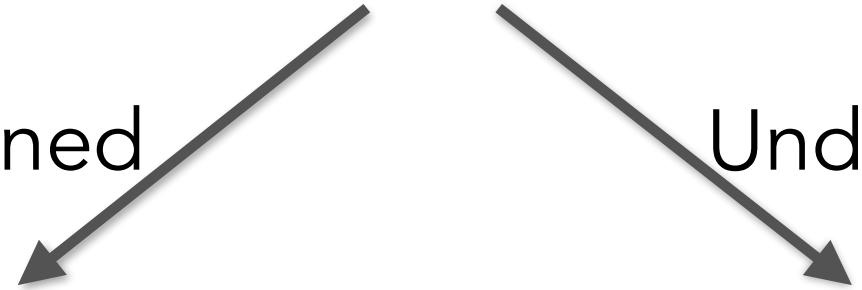
Lecture 2
Anomaly Introduction

Ref: Outlier Analysis, Charu C Agrawal

Surveillance~observing for



- For terrorists - Osama
- Intruders
- Abandoned baggage
- Objects (e.g. Red ferrari)
- Wrong way driving

- Anomaly
- Abnormality
- Unexpected
- Unusual
- Outlier detection

Anomaly Definition

- Anomaly detection is the identification of rare events.
- Anomaly detection is identification of events or observations, represented as data point, that differ significantly from majority of the data.
- An anomaly is a data point that is significantly different from rest of the data.

Intrusion in Computers

- Data: systems calls, network traffic, etc.
- Task: Detection intrusion of a malware/ virus
- Anti-virus

Credit card fraud

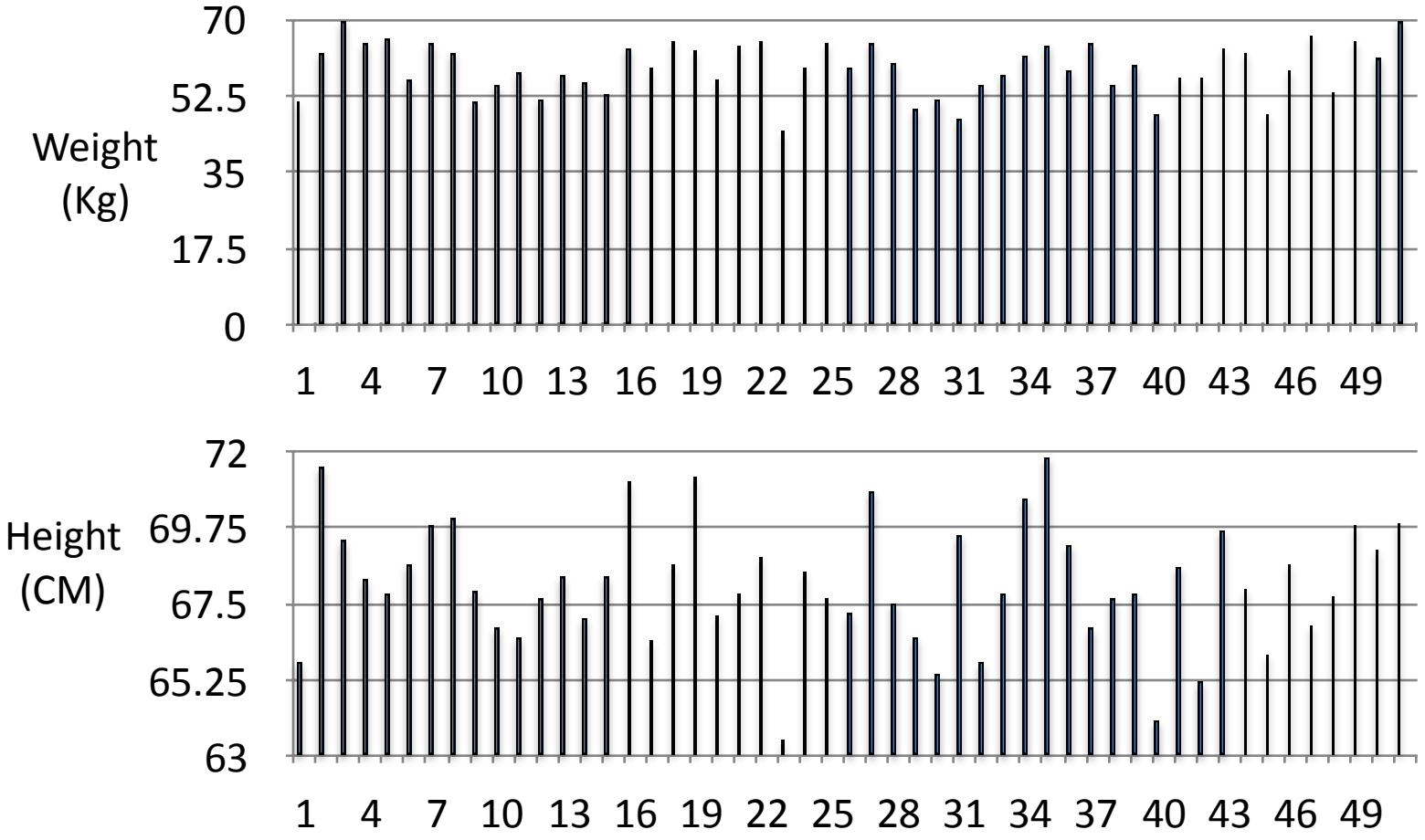
- Data: number of transactions, amount
- Task: Detect credit card theft, cloning
- Action: block, give warning

More Anomaly Applications

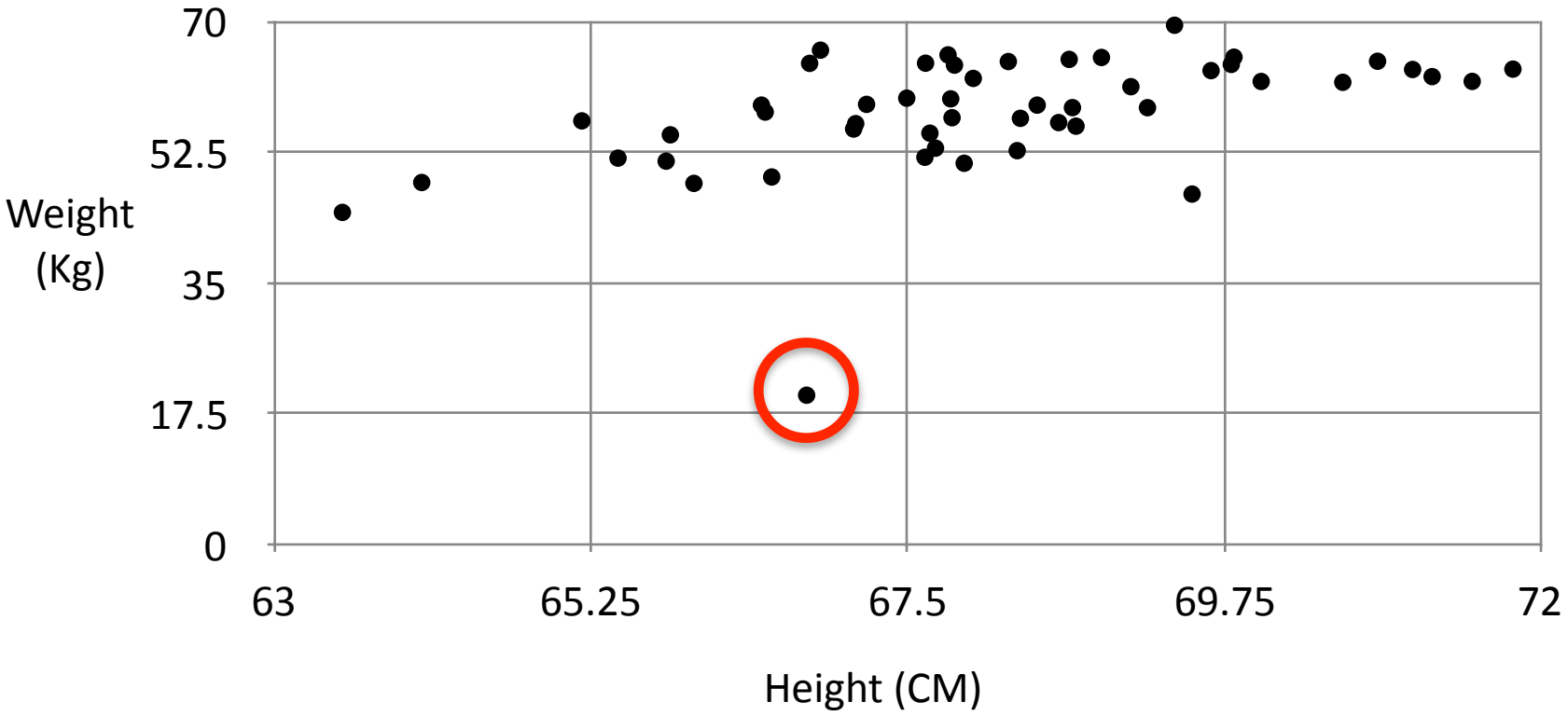
- Law enforcement
- Bank fraud detection
- Medical problem detection
- Malfunctioning equipments
- Structural defects
- Earth sciences

Problem: given a data point, how to identify whether it is normal or abnormal?

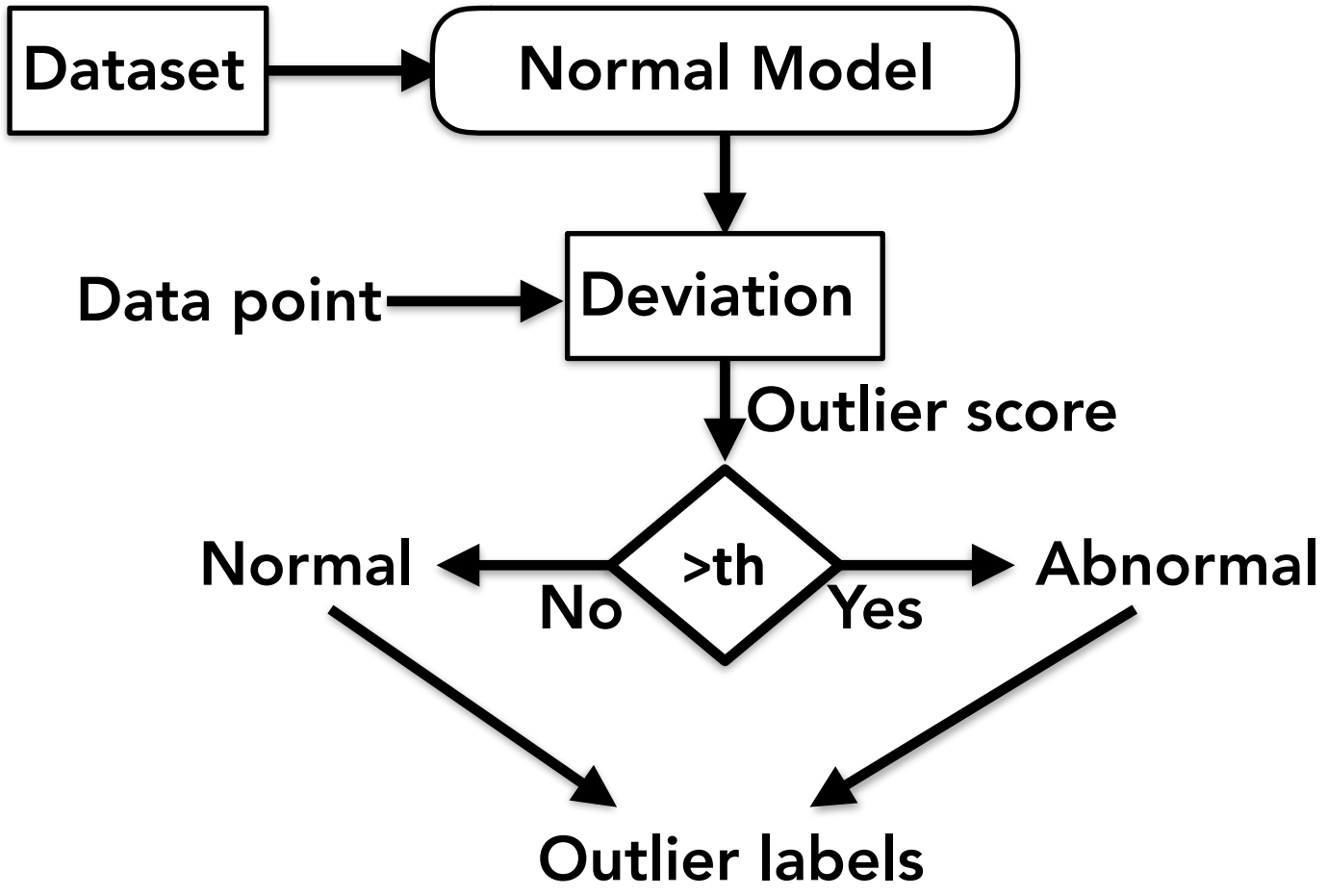
Height-Weight Data



Try Data Visualization



Typical Anomaly Model



Q: How much deviations is sufficient to call a dat point anomaly?

The deviation could be due to noise in observations also!

The deviation threshold chosen on an ad hoc basis according to application-specific criteria!

Z-value Test

Calculate Z value for i^{th} data point $Z_i = \frac{X_i - \mu}{\sigma}$

where $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2}$

Large Z value indicates anomaly, rule of thumb is $Z \geq 3$

Limitations

- Data may not be Gaussian distributed
- Sufficient samples may not be available to robustly estimate mean and standard deviation