VOLUME
Scale of data

VARIETY
Different forms of data

THE V's OF BIG DATA

VELOCITY
speed at which data is being generated & the speed at which it moves from one point to the next

VERACITY
uncertainity of data
- bias
- noise
- abnormality

VALENCE
connectedness of big data in the form of graphs

(There is also a 6th V — referred to as Value)

CHALLENGES RELATED TO VOLUME

STORAGE

ACQUISITION

DISTRIBUTION

PROCESSING

RETRIEVAL

PROCESSING BIG DATA

↑ VOLUME   ↓ PERFORMANCE   ↑ COST
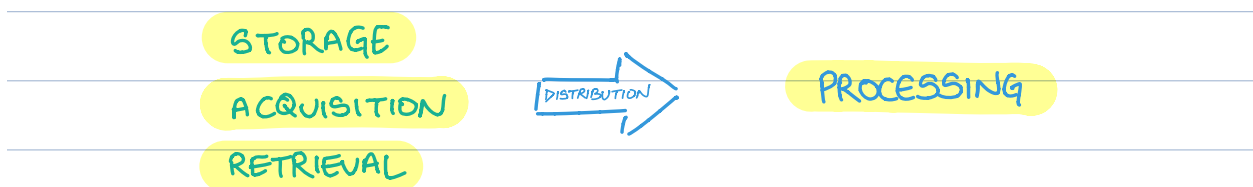
VARIETY == COMPLEXITY

AXES OF DATA VARIETY

STRUCTURAL VARIETY

formats & models

SEMANTIC VARIETY

how to interpret & operate on data

eg adult ⇒ juvenile or 8, 18, 34

MEDIA VARIETY

medium in which data gets delivered

AVAILABILITY VARIATIONS
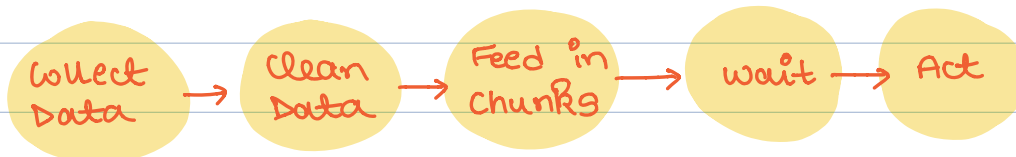
real time?
Intermittent

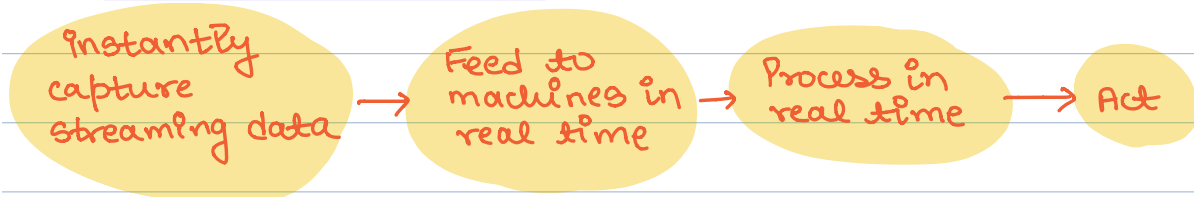# VELOCITY == SPEED

speed of **creating data**
storing data
analysing data

BATCH PROCESSING → Incomplete

Collect Data → Clean Data → Feed in Chunks → Wait → Act

REAL-TIME PROCESSING → fast

Instantly capture streaming data → Feed to machines in real time → Process in real time → Act

VERACITY == QUALITY
- VALIDITY
- VOLATILITY

Quality of data can be defined by

Accuracy of the data

Reliability of the data source

Context within Analysis

→ Uncertainity of the data increases
as we go from enterprise data to sensor data

VALENCE == CONNECTEDNESS

Data Connectivity:

Two items are connected when they're
related to each other

Valence:

Fraction of data items that are connected
out of the total no. of possible collections

VALENCE INCREASES OVER TIME
↓
MAKES DATA CONNECTIONS DENSE ⟶

more complex
data exploration
algorithms

Modelling &
Prediction of valence
changes

emergent behaviour
analysis

Group event detection