```
mestamp":"2017-06-03T18:42:18.018"

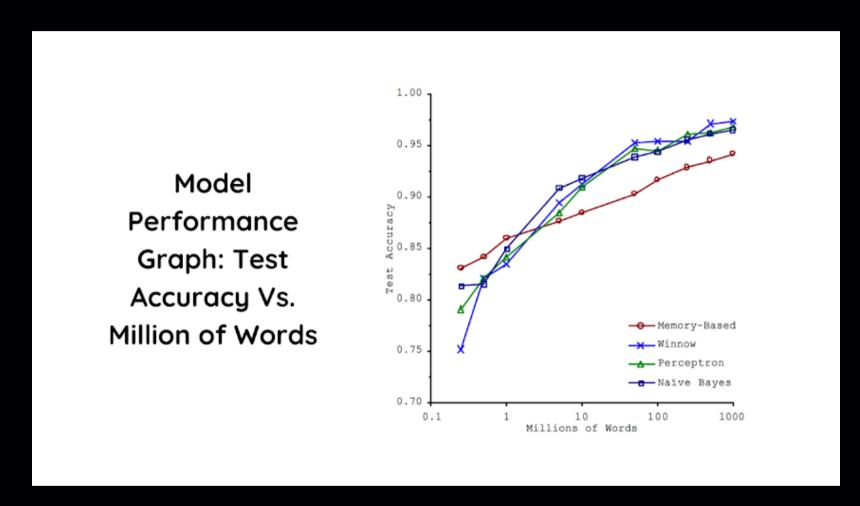
cechars":"5022", "message":"Duration Logueriar Logueria 
                       class": "com.orgmanager.handlers.Request" "webparan
                   aStartMillis":"0", "level":"INFO", "WebURL":"/app/page/report", "webPar
                      "" "789d89cb-bfa8-4e7d-8047-498454af885d", "sessionID": "144o2n620j
                           onMillis":"7"}{"timestamp":"2017-06-03T18:46:921.000", "deltaStartMillis"
                          ":"com.orgmanager.handlers.RequestHandler", "method":"handle", "requestID"
                    Chars": "10190", "message": "Duration Log", "durationMillis": "10"}{"timestam
                   RL":"/app/rest/json/file", "webParams":"file=chartdata_new.json", "class":
                         stID":"7ac6ce95-19e2-4a60-88d7-6ead86e273d1", "sessionID":"144o2n620jm9trnd3
                           onMillis":"23"}{"timestamp":"2017-06-03T18:42:18.018", "deltaStartMi
                           "COM.orgmanager.handlers.RequestHandler", "method":"handle", "request
                        hars":"5022", "message":"Duration Log", "durationMillis":"508"}{"times."
                   RL":"/app/page/analyze", "webParams":"null", "class":"com.orgmanager.l
                          ID": "8249868e-afd8-46ac-9745-839146a20f09", "sessionID": "144o2n620jms
                         LonMillis":"36"}{"timestamp":"2017-06-03T18:43:335.030","deltastartMii
             SionID": "144o2n620jm9trnd3s3n7wg0k", "webURL": "/seseee: "messee: "messee:
                       rams":"file=chartdata_new.json", "sizeChars":"48455", "agge/report".
                               "com.orgmanager.handlers.RequestHandler",
```

# Insufficient Data, what do we do?

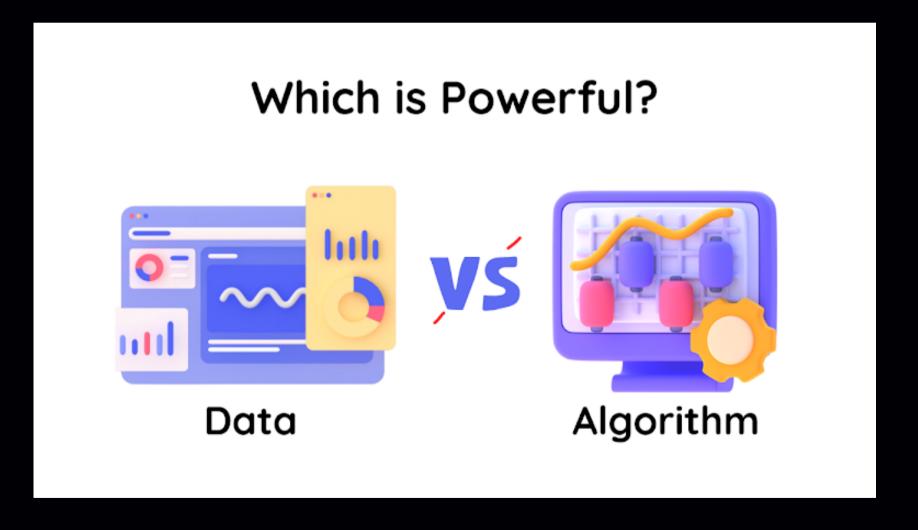
Dr. LEE MEI SIN Al Researcher



#### The Unreasonable Effectiveness of Data



The Unreasonable Effectiveness of Data - A. Haley , P. Norvig, F. Pereira 2009



# Insufficient data? What should we do?

## Panic?!

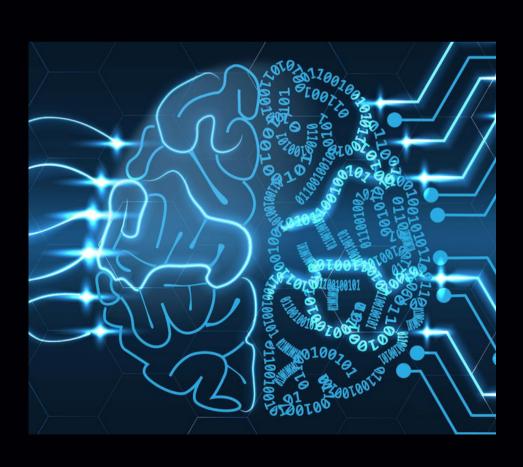
## Standard answer: Fine more data / Label more data





## Strategy #1 Use whatever we have

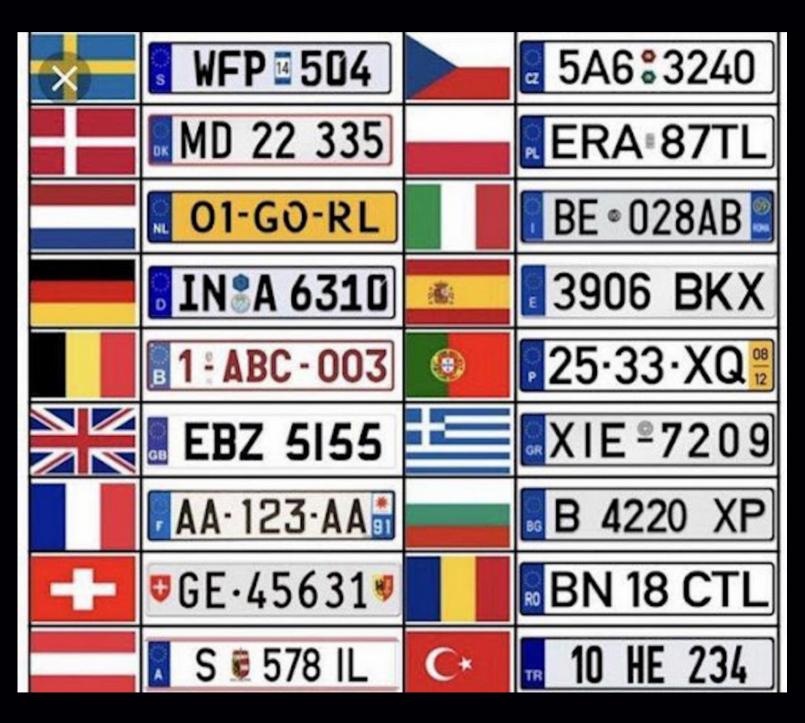




#### Use whatever we have!

Use existing data of the same task

### License Plate Recognition

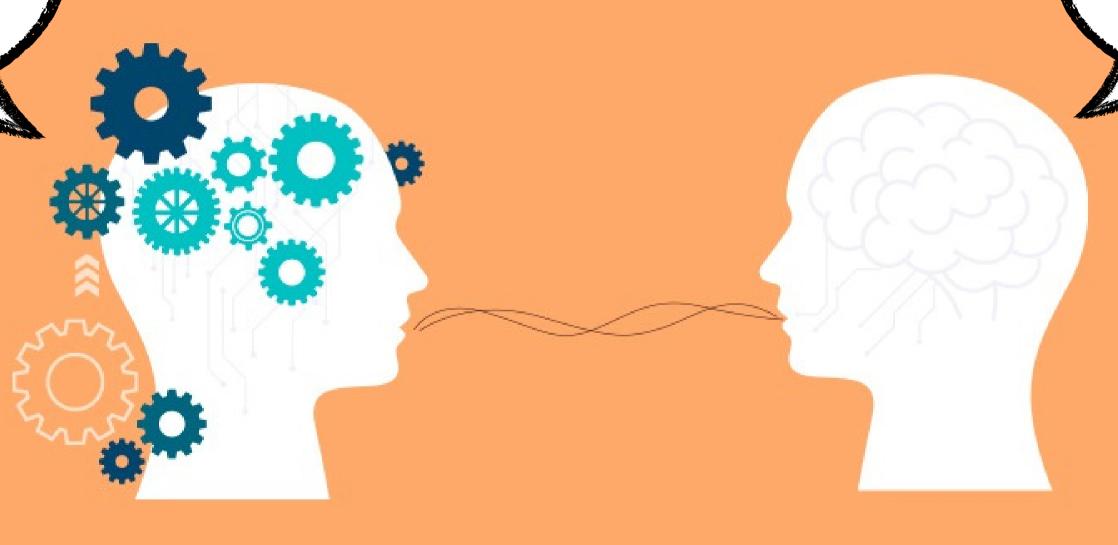


BEK 6746 PHM 2865 **PGK 936** PDN 8789 PCS 6249 **PET 1286 PGP 8181** WVL 370 PDA2961 PFW 806 PJC 8888 PDA 2961 PAC 5176 PFR8323 QPGF6878 VFI28 **PEF 8223** PLL 1828 **PHVIOOB** HP 2092 PHQ 2200 **PCR 1134 WSP 6186** PJU8499 WDK 2758 PFN 8673 PJL 8966 **PGR 5866** PGC6969 PCY 5915 PFN12 PKN 7362 PJX 4165 PGF6878 PBX 5285 PMC 5731 **HP 2809** PHA 6784 **PEK 4618** PLG 5965 PHD 555 PHD 555

Car license plate in Malaysia

Dialogue System







#### Use whatever we have!

### Use models of <u>different task</u>

## Use existing models



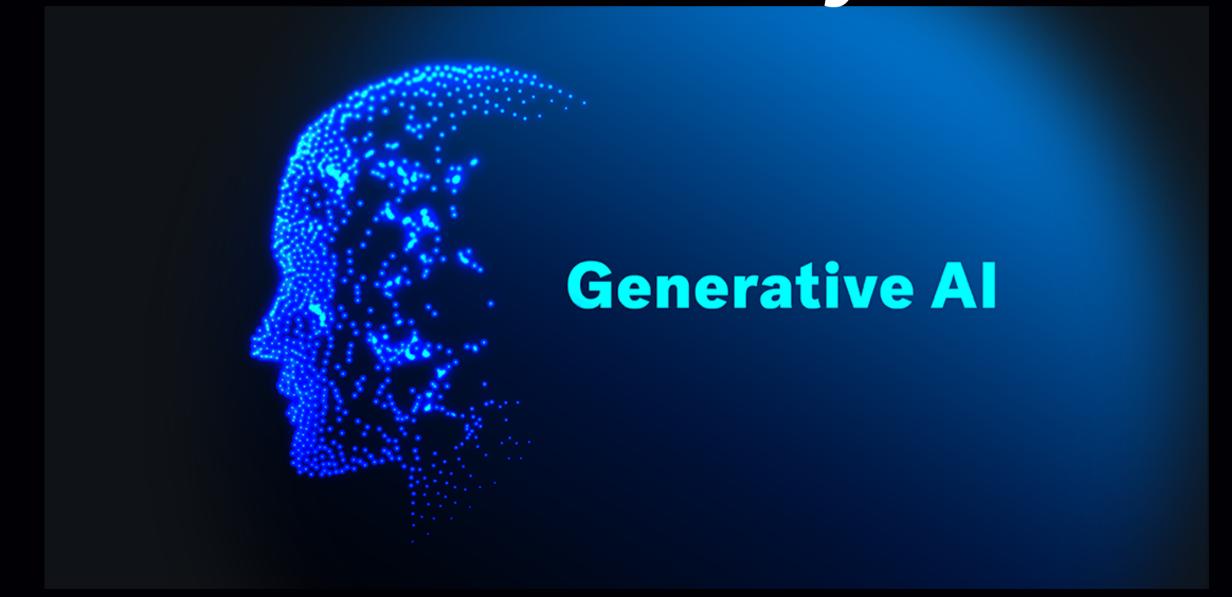




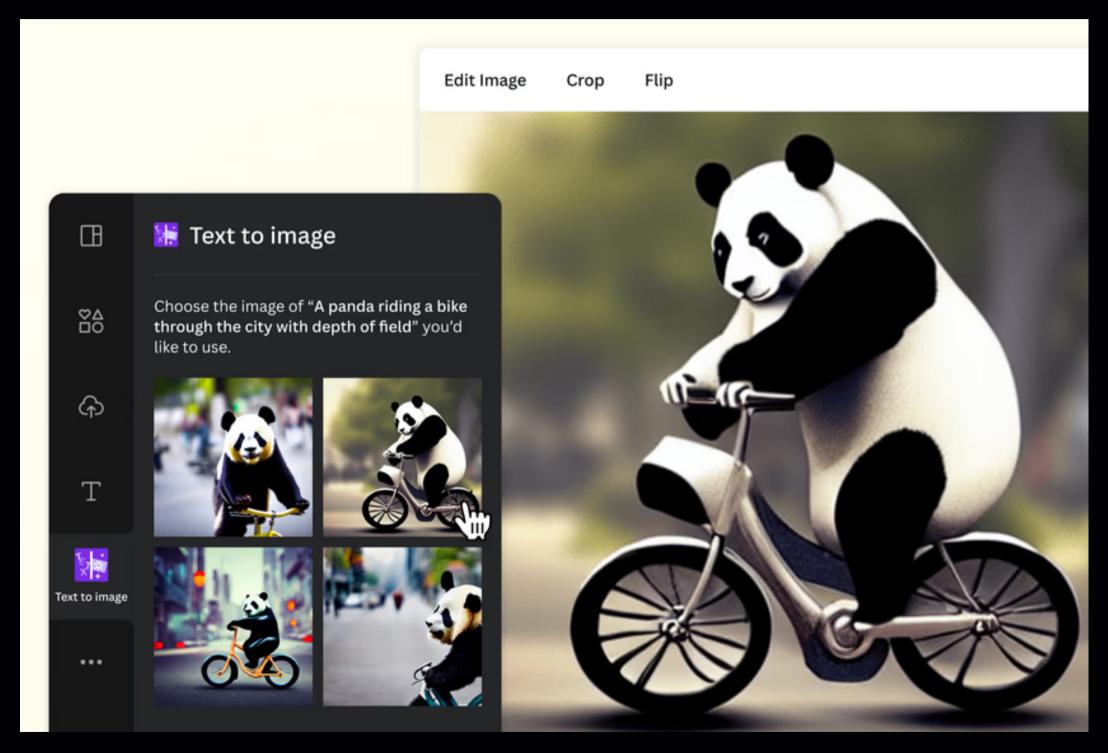
- machine translation
- paraphasing / style transfer

### Strategy #2

Use Generative Al for synthetic data



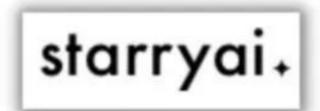
### Generate images / videos



#### Generate images / videos

#### **©Dall-E** Alternatives



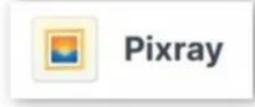


















#### Generate textual and structured data



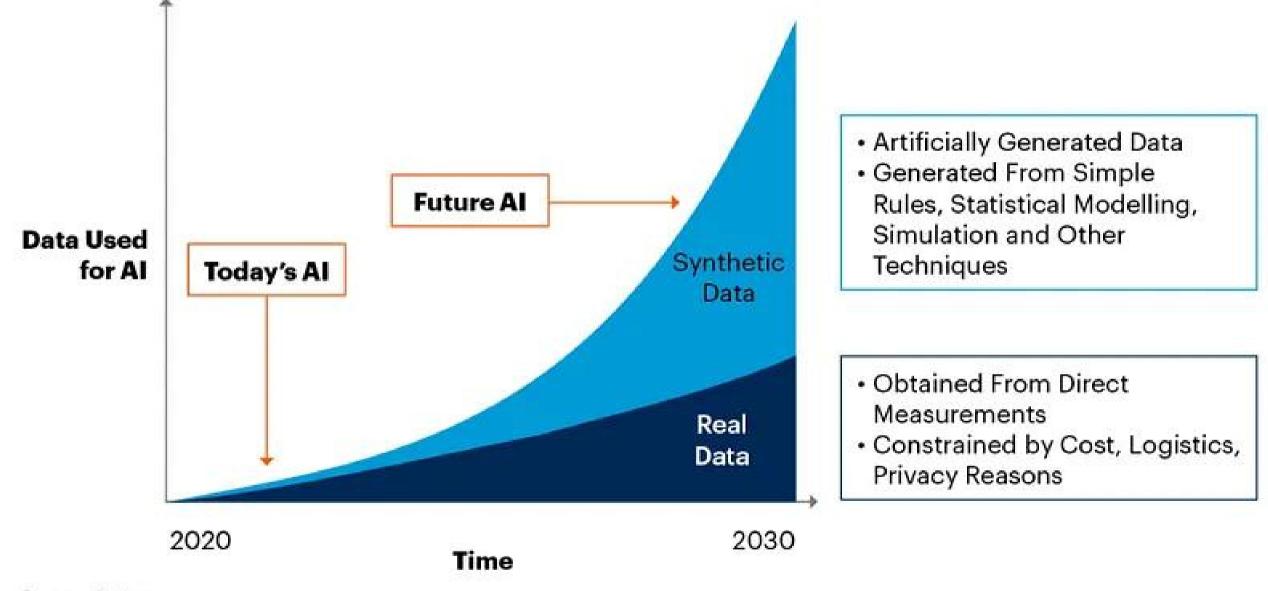
#### **Prompt:**

Create a table with columns appointment\_id (id), patient\_id (different id), chw\_id (multiple recurring ids), reported\_by\_parent (different id), reported (timestamp), child\_temperature (body temperature in celsius), child\_age\_months (integer with max value 72). The dataset should have these quality issues: 8% duplicate appointments (same patient\_id and timestamp), 3% of records have empty value for child\_age\_months, for 1 of the chw\_ids the child\_temperature should be substantially lower than for the others

#### **Model Response:**

ere is a sample table with 1000 rows that meets the specifications:							
appointment_id	patient_id	chw_id	reported_by_parent	reported	child_temperature	child_age_months	
1	P1	CHW1	RP1	2021- 01-01 10:00:00	37.2	12	
2	P2	CHW2	RP2	2021- 01-02 09:00:00	38.0	6	
3	P3	CHW3	RP3	2021- 01-03 11:00:00	36.5	18	
4	P4	CHW4	RP4	2021- 01-04	37.1	24	

#### By 2030, Synthetic Data Will Completely Overshadow Real Data in Al Models



Source: Gartner

750175\_C

#### Challenges and Potential Pitfalls

- Data quality synthetic data:
  - not reflective of actual data
  - unable to maintain correlation and dependencies between data points
- Factual inaccuracy (caused by LLM hallucinations)
- Bias



#### How to mitigate this?

- Proper Simulation of Real-World Conditions
- Validation against Real Data
- Maintaining Diversity
- Continuous Monitoring and Feedback
- Involve Subject Matter Experts



#### About the Speaker

#### CREATE. COLLABORATE. INNOVATE.



#### Dr. LEE MEI SIN

- Founding member of Women in Al Malaysia
- Al Research Fellow in Monash University
- meisin@womeninai.co
- in linkedin.com/in/meisinlee/

## Q & A