# Các mô hình ước tính phù hợp khi dữ liệu còn hạn chế

Ngô Hoàng Anh (Ecole Polytechnique - Pháp)
Hoàng Thái Nam (Beloit College - Mỹ)
Nguyễn Tuấn Khôi (University of Melbourne - Úc)

# Hiện trạng

Có nhiều dữ liệu nhưng chưa được trích xuất. Những dữ liệu hiện có phần nhiều là untidied data. Vì vậy, cần sử dụng ít dữ liệu, nhưng dự đoán nhanh.

# Yêu cầu

- Nhanh
- Độ chính xác cao
- Có ý nghĩa

# Giải pháp

- Xây dựng mô hình sử dụng tối thiểu biến số
- Độ phức tạp đủ cao nhưng không overfit

  >> Sử dụng mô hình 1 biến (1-dimensional) dựa trên số ca của các ngày trước

# Ưu điểm

- Nhanh
- Độ chính xác cao

  >> ***Đủ ý nghĩa để đưa ra giải pháp vĩ mô kịp thời***

# Nhược điểm

- Không mang ý nghĩa dịch tễ (chưa xác định nguồn nhiễm, F1-F2,...)
- Độ chính xác chưa đạt tuyệt đối (Δ=3%-5%)

  >> ***Sẽ cải thiện được bằng nhiều dữ liệu biến khác nhau***

# Tiếp cận ban đầu

- Traditional Machine Learning - Máy học truyền thống:
    - Convolutional Neural Network (CNN)
    - Long Short-Term Memory (LSTM)
    - ARIMA Family

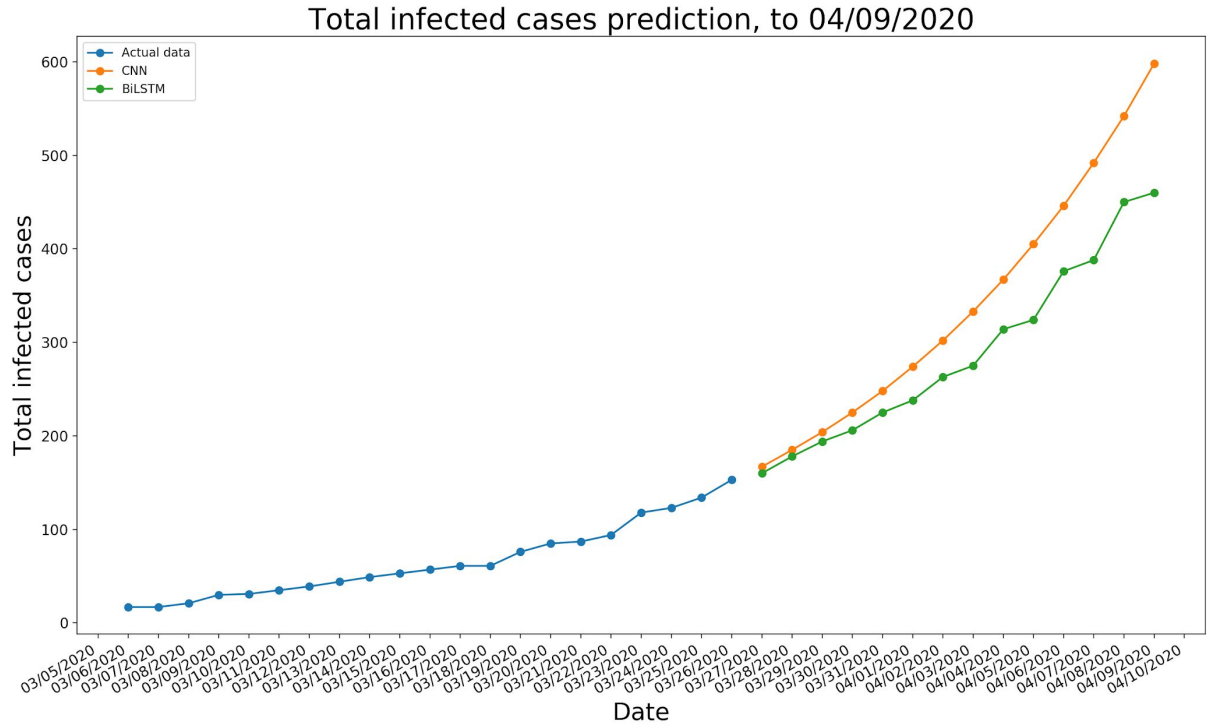
- Mô hình toán học:
    - Grey Model & Extensions: RONGBM(1.1)

# Các mô hình ML truyền thống

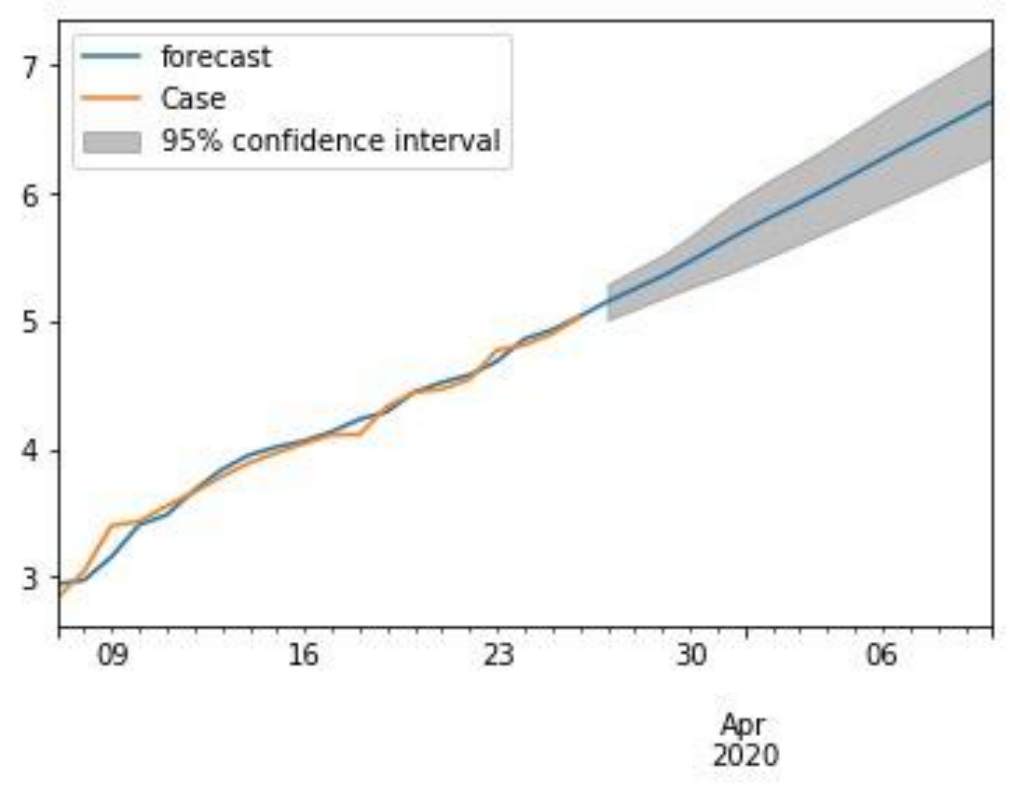
- Neural networks
  - Dựa trên mạng neural thần kinh của con người
  - CNN có nhiều mạng và neural ẩn để học tốt hơn các mạng máy học đơn thuần
  - LSTM có nhiều cell, hoạt động như bộ não của con người (có chức năng “quên" để lọc data). Biến thể Bidirectional LSTM (BiLSTM) có khả năng học từ hai chiều (input và output).
- ARIMA

  Là mô hình xác suất cơ bản dùng để benchmark các mô hình khác

# CNN + LSTM Graph



# ARIMA Graph


forecast
Case
95% confidence interval
7
6
5
4
3
09
16
23
30
06
Apr
2020

# Data

| Date | CNN | BiLSTM | ARIMA |
|---|---|---|---|
| 03/27/2020 | 167 | 160 | 186 |
| 03/28/2020 | 185 | 178 | 204 |
| 03/29/2020 | 204 | 194 | 225 |
| 03/30/2020 | 225 | 206 | 250 |
| 03/31/2020 | 248 | 225 | 281 |
| 04/01/2020 | 274 | 238 | 314 |
| 04/02/2020 | 302 | 263 | 349 |
| 04/03/2020 | 333 | 275 | 386 |
| 04/04/2020 | 367 | 314 | 429 |
| 04/05/2020 | 405 | 324 | 479 |
| 04/06/2020 | 446 | 376 | 535 |
| 04/07/2020 | 492 | 388 | 595 |
| 04/08/2020 | 542 | 450 | 662 |
| 04/09/2020 | 598 | 460 | 735 |

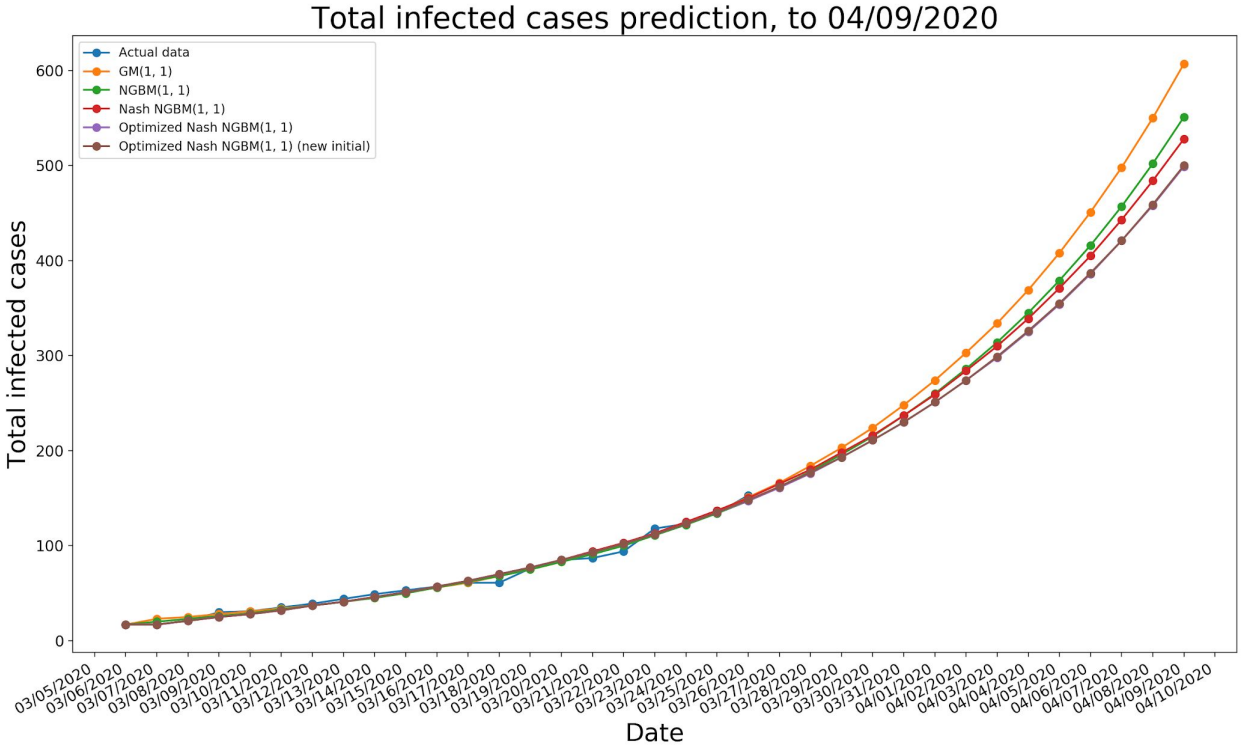
# Kết quả mô hình

- Không xuất hiện overfit khi sai số có xu hướng nhỏ dần
- Sai số lớn vào những ngày nhất định (sẽ bổ sung sau)
- CNN dự đoán trend tăng cấp số nhân, LSTM dự đoán có trend theo ngày
- ARIMA dự đoán với sai số 5%, và lệch ± 2 ngày
- Với data mới nhất (27/3), dự đoán kết quả là 160-186, kết quả đến (hiện tại) 163

# Mô hình toán học

- Grey Systems & Extensions:
  - Ý tưởng dựa trên sự discretize theo thời gian của 1 phương trình vi phân.
  - Đã được đề cập trong *A rolling optimized nonlinear grey bernoulli model ROMGBM(1,1) and application in predicting total 2019-nCoV infected cases,* HA Ngo, TN Hoang, 2020
  - Bao gồm: GM(1,1), NGBM(1,1), Optimized NGBM(1,1), Rolling optimized NGBM(1,1)

# Grey systems with extensions Graph



# Data

| Date | GM(1, 1) | NGBM(1, 1) | Nash NGBM(1, 1) | Optimized Nash NGBM(1, 1) | Optimized Nash NGBM(1, 1) (new initial) |
|---|---|---|---|---|---|
| 03/27/2020 | 166 | 162 | 165 | 161 | 162 |
| 03/28/2020 | 184 | 178 | 180 | 176 | 177 |
| 03/29/2020 | 203 | 196 | 198 | 193 | 193 |
| 03/30/2020 | 224 | 215 | 216 | 211 | 211 |
| 03/31/2020 | 248 | 237 | 237 | 230 | 230 |
| 04/01/2020 | 274 | 260 | 259 | 251 | 251 |
| 04/02/2020 | 303 | 286 | 284 | 274 | 274 |
| 04/03/2020 | 334 | 314 | 310 | 298 | 299 |
| 04/04/2020 | 369 | 345 | 339 | 325 | 326 |
| 04/05/2020 | 408 | 379 | 371 | 354 | 355 |
| 04/06/2020 | 451 | 416 | 405 | 386 | 387 |
| 04/07/2020 | 498 | 457 | 443 | 421 | 421 |
| 04/08/2020 | 550 | 502 | 484 | 458 | 459 |
| 04/09/2020 | 607 | 551 | 528 | 499 | 500 |

# Kết quả mô hình

- Không xuất hiện overfit khi sai số có xu hướng nhỏ dần
- Sai số lớn vào những ngày nhất định (4/3 và 19/3)
- Sai số trung bình nhỏ (nhỏ hơn 5% với những mô hình optimized, ngày cuối cùng có sai số khoảng 3%)
- Với data mới nhất (27/3), dự đoán kết quả là 160-166, kết quả (hiện tại ) là 163

# Dự đoán chung

Có 3 kịch bản được đưa ra theo thứ tự tệ dần, cụ thể như sau:

| Ngày\Kịch bản | Best Scenario | Average Scenario | Worst Scenario |
|---|---|---|---|
| 31/03 | 225 | 230-235 | 240 |
| 09/04 (sau 2 tuần) | 450-500 | 500-550 | 600 |

# Hướng đi tiếp theo:

- Mở rộng số lượng và phạm vi mô hình áp dụng
- Sử dụng mô hình đa biến để các mô hình mang ý nghĩa dịch tễ
- Ngoài dự đoán số ca mắc của địa phương/cả nước, có thể dùng Machine Learning để dự đoán ca bệnh (xác suất phải dùng ICU, tử vong/hồi phục, ngày xuất viện,...)