# Data Science in Techn-Socio-Economic Systems

EpidemicDatathon

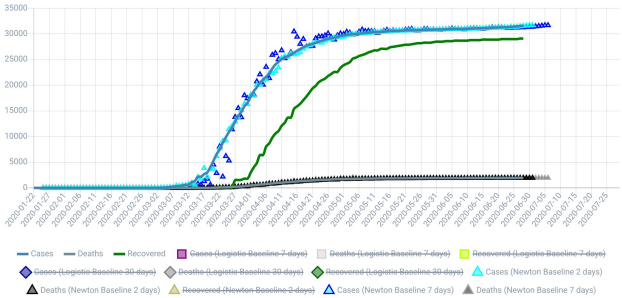
Team MAE

**Stefan Lionar***, Dharma Pranata**, Brigitta Rachel**

*ETH Zurich student, **External contributors

# Approach

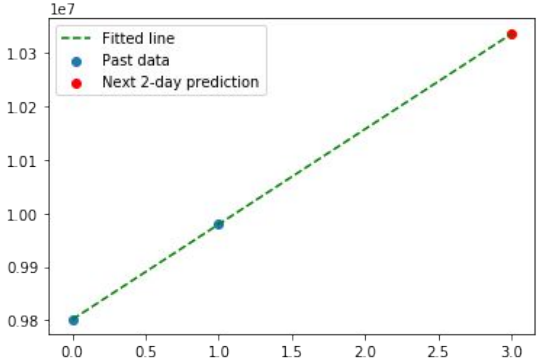
- Assumptions:
  - Dataset: JHU CSSE COVID-19
- Observations:
  - Confirmed cases, deaths, recovery have trend with shapes similar to sigmoid function

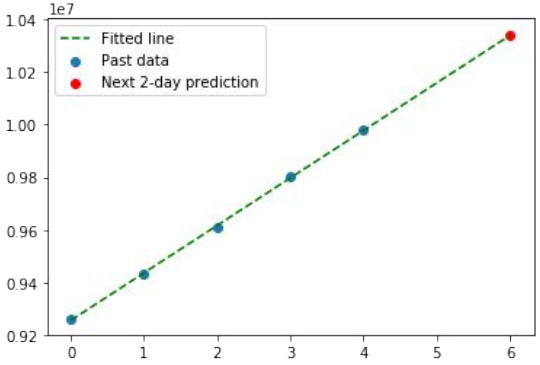


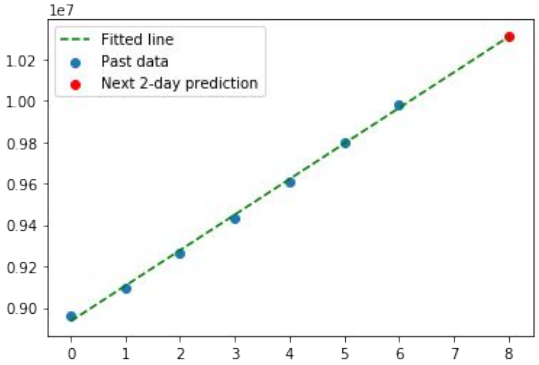

- Adjusted linear regression model:
  - Apply the same model for:
    - Confirmed cases
    - Deaths
    - Recovery

# 2-day prediction

- Linear regression with dynamic windows (2-7 days)




- Less robust towards fluctuation
- More sensitive towards gradient change

- More robust towards fluctuation
- Less sensitive towards gradient change

# 2-day prediction

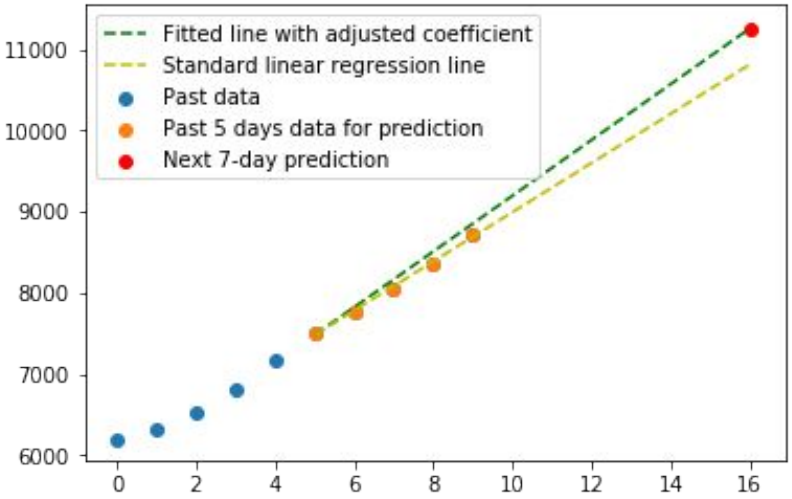
Window selection method:

- ***Window selection*** **= [2,3,4,5,6,7]**
- **For every country:**
  - **Set groundtruth from today's data**
  - **For *window* in *window selection*:**
    - **Fit past *window* data with linear regression**
    - **Collect each prediction from *window***
  - **Select *window* with the minimum Mean Absolute Error**

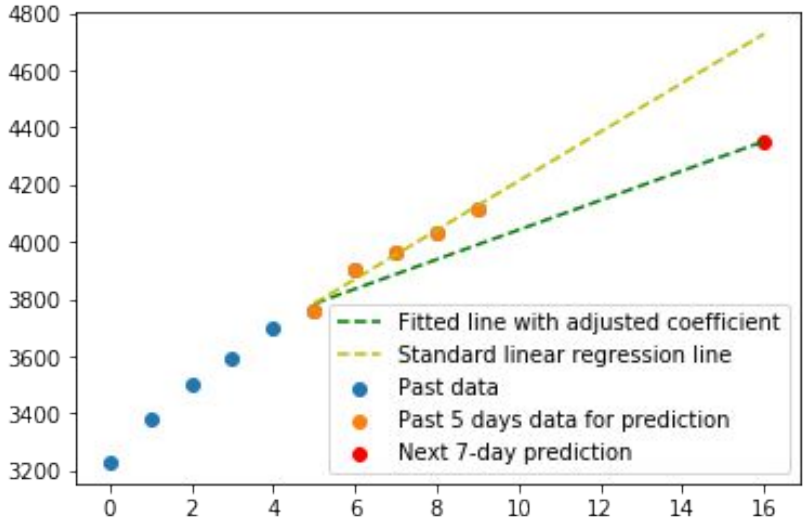
# 7-day prediction

- Adjust linear regression coefficient based on trend

Increasing

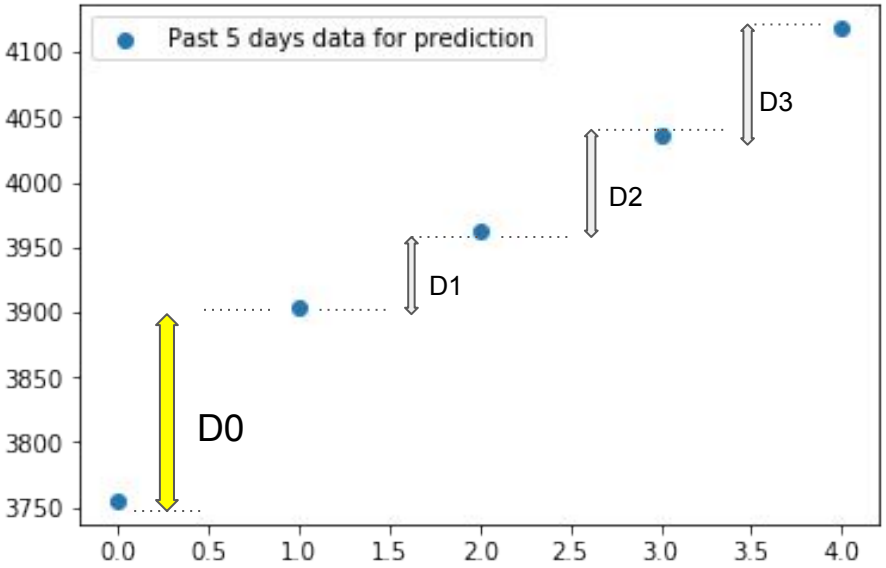



Decreasing



# 7-day prediction

## Adjusting linear regression coefficient




```
diff_ratio = ⅛ * (D1 + D2 + D3) / D0
multiplier = mul_coeff * diff_ratio**exp_coef

Fitted line:
Y = multiplier * coeff * X + intercept
```

- Use past 5 days data
- mul_coeff and exp_coeff is tuned empirically using past global cases as ground truth

OLS: y = case (confirmed case / deaths / recover)
X = time
Y = coeff *x + intercept

# 30-day prediction

- Similar approach to 7-day prediction
- Use 10 days past data with 2 days interval vs 5 days

# Results (21-27 June 2020)

## 2-day prediction

| SCORE | TEAM |
|---|---|
| 1532.1693099016484 | C.F.R.S.S. |
| 1251.7598831290466 | stayhome |
| 1062.4765106657846 | MAE |
| 1033.2791237644647 | ToBeAnMa |
| 572.5187119146683 | Quaranteed Success |
| 79.86191816073888 | GNTM_team |
| 40.347492886920755 | PandeML |
| 3.9020598929059975 | ValenciaSpain |
| 3.6922487557465598 | #SeifenBoss |
| 0.4837068243452888 | Quaranteam |

## 7-day prediction

| SCORE | TEAM |
|---|---|
| 1381.6749105327967 | stayhome |
| 1226.3876952896655 | C.F.R.S.S. |
| 984.3016250930384 | MAE |
| 718.4953536723917 | GNTM_team |
| 609.3661843714984 | ToBeAnMa |
| 135.96096510224532 | Quaranteed Success |
| 31.382075292487194 | PandeML |
| 4.997855656973526 | #SeifenBoss |
| 3.779748820904161 | ValenciaSpain |
| 0.8338339454827546 | Quaranteam |

# Conclusions

- We built models:
  - Linear regression with ability to adapt with gradient changes
  - Time series prediction for monotically increasing number of cases
- JHU CSSE COVID-19 as groundtruth:
  - Many unreported cases
  - Ability to estimate the actual cases is questionable