Modelling COVID-19

Making the best of bad data

Luiz Max Carvalho [lmax.fgv@gmail.com]

FGV EMAp

Acknowledgements

Grupo de Métodos Analíticos em Vigilância Epidemiológica

PROCC FGV EMAp Ministério da Saúde FIOCRUZ Fundação Oswaldo Cruz

https://covid-19.procc.fiocruz.br/

- Antônio Pacheco;
- Claudia Codeço;
- Daniel Villela;
- Flavio Coelho;
- Leonardo Bastos;
- Marcelo Gomes;
- Raquel Lana;
- Roberta Niquini;
- Oswaldo Gonçalves Cruz.

Non-MAVE collaborators

- Bob Kubinec (NYU Abu Dhabi);
- Nuno Faria (Oxford/Imperial College).

Outline

Data

Why is the data “bad”?

Outline

Data

Why is the data "bad"?

Models

Are all models wrong? Are some useful?

Outline

Data

Why is the data “bad”?

Models

Are all models wrong? Are some useful?

Some results

What can semi-mechanistic and non-mechanistic models do for us?

Outline

Data

Why is the data “bad”?

Models

Are all models wrong? Are some useful?

Some results

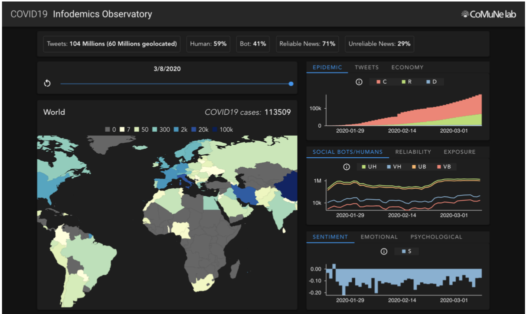
What can semi-mechanistic and non-mechanistic models do for us?

Future

What have we learned? How can we improve going forward?

COVID-19 Data

- Lots of wonderful efforts to collect and make sense of data:
    - Johns Hopkins University;
    - CoronaNet data collection project;
    - Nextstrain;
    - Covid19 Infodemics Observatory;
    - Wesley Cota’s website;
    - Brasil.IO;
    - Infogripe.



COVID19 Infodemics Observatory

Tweets: **104 Millions (60 Millions geolocated)**

Human: **59%**

Bot: **41%**

Reliable News: **71%**

Unreliable News: **29%**

3/8/2020

World

COVID19 cases: 113509

Which data, specifically?

- Cases per day;
- Deaths per day;
- Interventions: what, when, for how long;
- Tests per day;
- Mobility: Google, Apple, In Loco reports;
- Serological surveys;
- Genomic information on the virus.

Part 1: Data problems

- Inconsistent criteria;

Part 1: Data problems

- Inconsistent criteria;
- Reporting (notification) delays;

Part 1: Data problems

- Inconsistent criteria;
- Reporting (notification) delays;
- Underreporting;

Part 1: Data problems

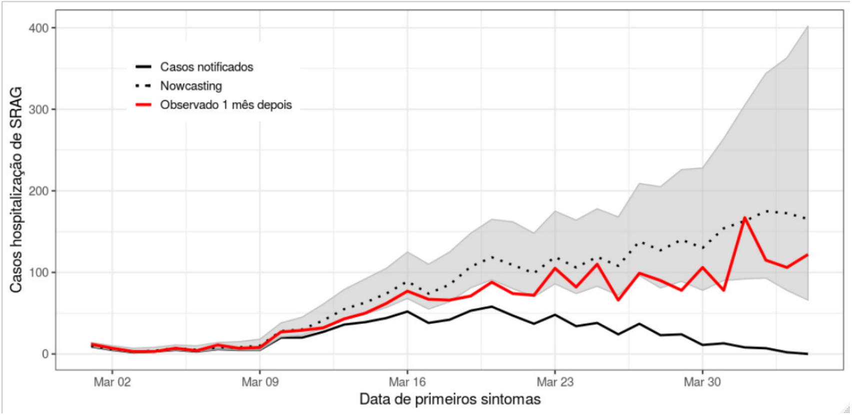
- Inconsistent criteria;
- Reporting (notification) delays;
- Underreporting;
- Completeness;

Data problems I: Inconsistent criteria

- Case definition may vary;
  Here is SRAG:
  - Fever (above 37,8 Celsius) AND;
  - Cough OR sore throat AND;
  - Difficulty breathing OR Dyspnea OR O_2 saturation below 95% AND;
  - Was hospitalised OR died with these symptoms.
- Testing is inconsistent and uncertain;
  - RT-PCR and antibody-detecting tests vary in scope and applicability;
  - Sensitivity and specificity.

Data problems II: Delay

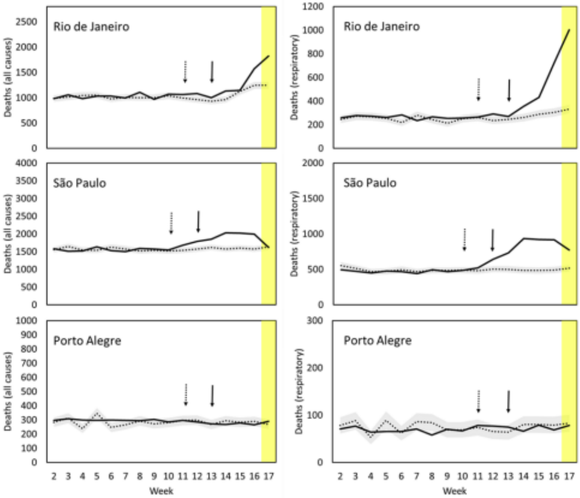
Important: delayed data exists, it just has not come into the system yet. See Bastos et al. (2019).



Data problems III: Underreporting

Cases that have not come into the system and probably won't.

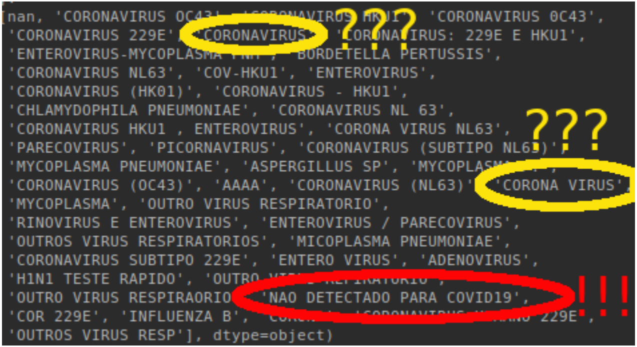
- One approach is excess mortality (e.g. Freitas et al., (2020))



Data problems IV: (In)completeness

Case data is usually incomplete/inconsistent.

- The beauty of open fields...



```
[nan, 'CORONAVIRUS OC43', 'CORONAVIRUS HKU1', 'CORONAVIRUS 0C43',
 'CORONAVIRUS 229E', 'CORONAVIRUS', 'CORONAVIRUS: 229E E HKU1',
 'ENTEROVIRUS-MYCOPLASMA PNM', 'BORDETELLA PERTUSSIS',
 'CORONAVIRUS NL63', 'COV-HKU1', 'ENTEROVIRUS',
 'CORONAVIRUS (HK01)', 'CORONAVIRUS - HKU1',
 'CHLAMYDOPHILA PNEUMONIAE', 'CORONAVIRUS NL 63',
 'CORONAVIRUS HKU1 , ENTEROVIRUS', 'CORONA VIRUS NL63',
 'PARECOVIRUS', 'PICORNAVIRUS', 'CORONAVIRUS (SUBTIPO NL63)',
 'MYCOPLASMA PNEUMONIAE', 'ASPERGILLUS SP', 'MYCOPLASM
 'CORONAVIRUS (OC43)', 'AAAA', 'CORONAVIRUS (NL63)', 'CORONA VIRUS',
 'MYCOPLASMA', 'OUTRO VIRUS RESPIRATORIO',
 'RINOVIRUS E ENTEROVIRUS', 'ENTEROVIRUS / PARECOVIRUS',
 'OUTROS VIRUS RESPIRATORIOS', 'MICOPLASMA PNEUMONIAE',
 'CORONAVIRUS SUBTIPO 229E', 'ENTERO VIRUS', 'ADENOVIRUS',
 'H1N1 TESTE RAPIDO', 'OUTRO VIRUS REPIRATORIO',
 'OUTRO VIRUS RESPIRAORIO', 'NAO DETECTADO PARA COVID19',
 'COR 229E', 'INFLUENZA B', 'CORON     'CORONAVIRUS HUMANO 229E',
 'OUTROS VIRUS RESP'], dtype=object)
```

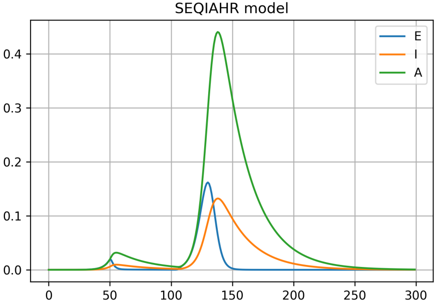
Part 2: Statistical modelling of COVID-19

First, an incomplete model taxonomy (Ronald Ross [1857-1932]):

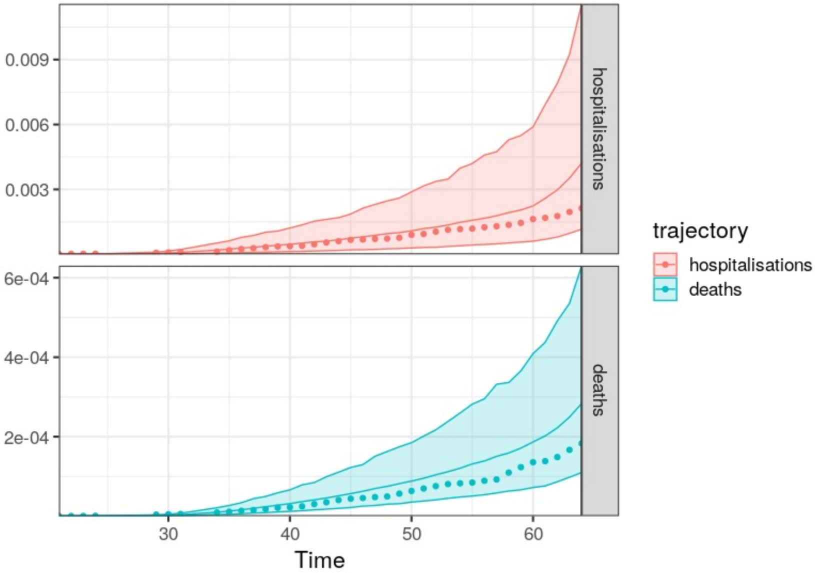
- *A priori* pathometry;
    - Representation of (disease) mechanisms;
    - Systems of differential equations;
    - "Mathematical models".
- *A posteriori* pathometry;
    - Curve fitting;
    - "Statistical models".

A (fancy) SEIR model

$$
\begin{aligned}
\frac{dS}{dt} &= -\lambda[(1-\chi)S],\\
\frac{dE}{dt} &= \lambda[(1-\chi)S] - \alpha E,\\
\frac{dI}{dt} &= (1-p)\alpha E - \delta I - \phi I,\\
\frac{dA}{dt} &= p\alpha E - \gamma A,\\
\frac{dH}{dt} &= \phi I - (\rho+\mu)H,\\
\frac{dR}{dt} &= \delta I + \rho H + \gamma A,\\
\lambda &:= \beta(I+A).
\end{aligned}
$$



A (fancy) SEIR model – fitted to data



Limitations of (deterministic) mathematical models

- Hard to fit to data;
- Usually quite sensitive to “minor” features of the data;
- Projections can be thrown off as a result;
- Hard to integrate various sources of information.

Statistical models

- More flexible (splines, GAMs, Gaussian processes);
- More natural accommodation of stochasticity and uncertainty;
- Less insight into mechanisms;
- More difficult to simulate scenarios (but not impossible!).

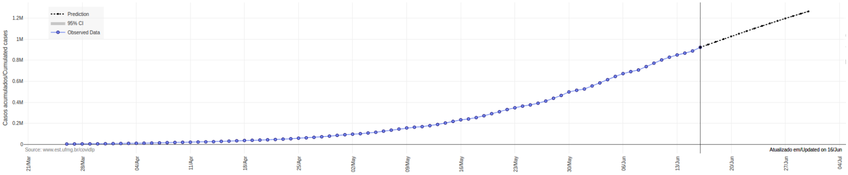
General considerations on modelling

- Meme: “All models are wrong, some are useful” (Box, 1976);
- “Many quotes are cool, few are universal”;
- Usefulness is context-dependent $\implies$ goal-oriented modelling;
- Every modelling endeavour will sacrifice “realism” for tractability, the important consideration is what to leave out *relative* to the task at hand (prediction, estimation, scenario modelling).

Related work: Statistical modelling of COVID-19 in Brazil

Many projects devoted to predicting the numbers of cases and deaths over time.

- UFMG;
- PUC-Rio;
- UFV;
- BRAM-COD (predicts ICU occupation as well);



Source: www.est.ufmg.br/covidlp

Example I: what predicts/explains infection rates?

For the first example, we will be discussing the of Kubinec & Carvalho (2020).

- If we have covariate data per state, say, can we study their effects on epidemic progression?;
- Can we make useful predictions even when the model lacks explicit dynamics?

Setup and notation

Here we will consider an *empirical*, *retrospective* model for infection rates. Consider time points $t = 1, \dots, T$ and regions $c \in C$. We will deal with

$$f_t\left(\frac{I(t)}{S(t)+R(t)}\right)$$

where $f_t : (0, \infty) \to (0, \infty)$ is a historical time trend, which we will call the “empirical” trend. This model can be seen as local linear approximation of the $I_c(t)$ curve and cannot be used for future predictions, for example.

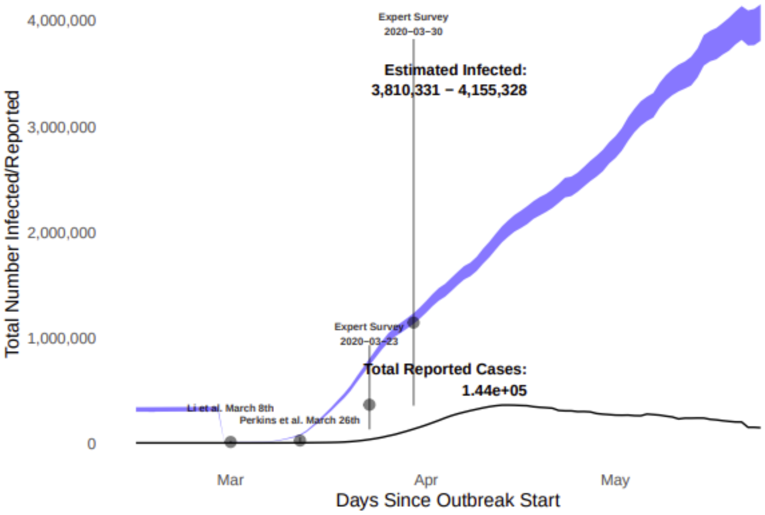
Additionally, we do not observe $I_{ct} := I_c(t)$ directly, but rather the numbers of tests, q_{ct}, and cases, a_{ct}, along with a set of covariates $\boldsymbol{X}_{ct}$.

Model details

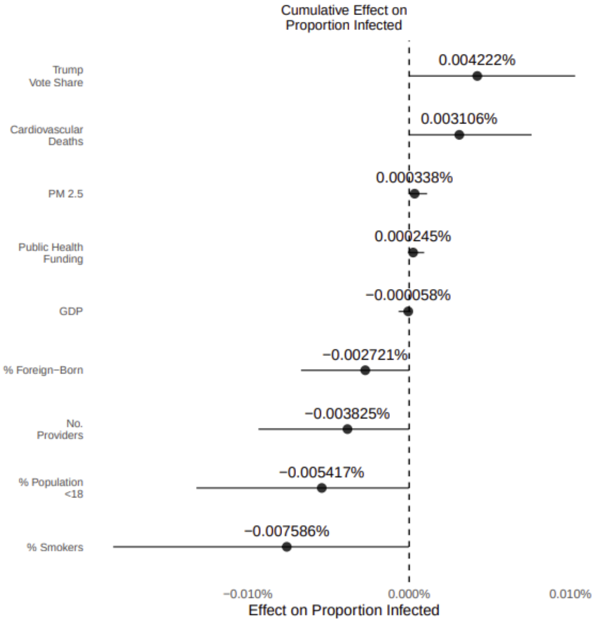
$$
\begin{aligned}
I_{ct} &\sim \text{Beta}(\alpha_1 + \beta_{O1} \sum_{c=1}^{C} \mathbf{1}(a_{ct'} > 0) + \beta_{S1} \boldsymbol{X}_{ct} + \\
&\quad \beta_{I1} t_o + \beta_{I2} t_o^2 + \beta_{I3} t_o^3, \phi) \\
q_{ct} \mid \text{pop}_c &\sim \text{Beta-Binomial}(\text{pop}_c, \text{logit}^{-1}(\alpha_2 + \beta_q I_{ct}), \phi_q), \\
a_{ct} \mid q_{ct} &\sim \text{Beta-Binomial}(q_{ct}, \text{logit}^{-1}(\alpha_3 + \beta_a I_{ct}), \phi_a).
\end{aligned}
$$

$$
\begin{aligned}
\beta_a &\sim \text{Exponential}(.1), \\
\beta_{qc} &\sim \text{Exponential}(\sigma_q), \\
\sigma_q &\sim \text{Exponential}(.1), \\
\beta_{Si} &\sim \text{Normal}(0, 2), \\
\beta_{Ii} &\sim \text{Normal}(0, 5), \\
\alpha_i &\sim \text{Normal}(0, 10), i = 1, 2, 3
\end{aligned}
$$

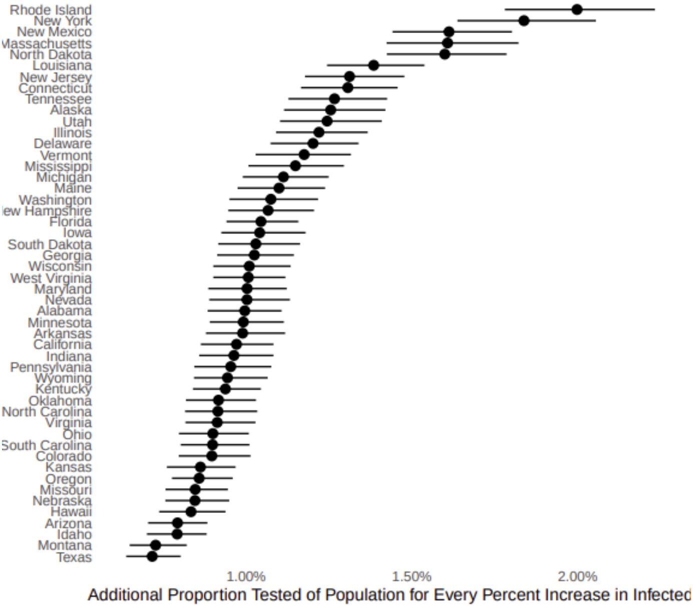
Example I results: reconstructions of I_t



Example I results: covariate effects

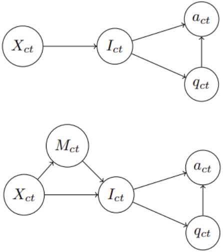


Example I results: testing rates across US states



Limitations

- Confounding structures:



- “Identifying” prior assumes constant proportion of (under)detection;
- Assumes same temporal trend for all locations, which might be unrealistic.

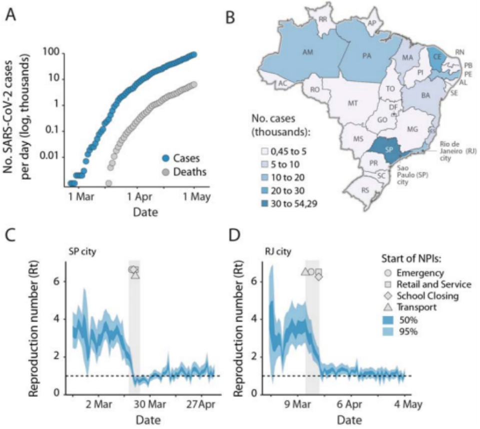
Example II: where did the virus come from?

In this example, we will look into the findings of Candido et al. (2020).

- Effective reproductive number (R_t) of COVID-19 in major cities;
- Viral dispersal across the country;

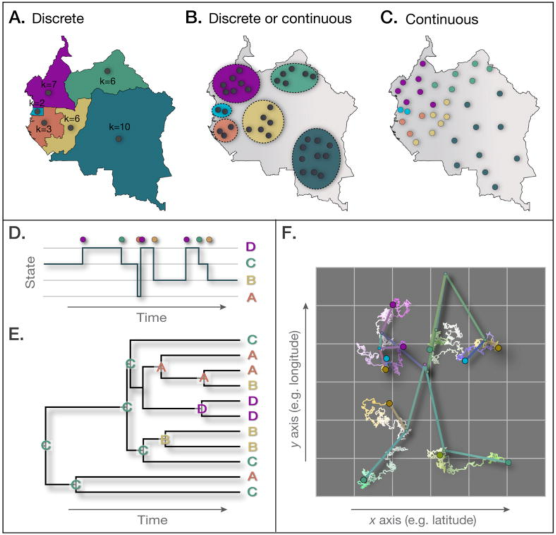
Example II: R_t

$$R_{tm} = R_{0m}\left(2\,\mathrm{logit}^{-1}\left(-\sum_{k=1}^{4}(\alpha_k + \beta_{mk})I_{ktm} + B_k\right)\right)$$

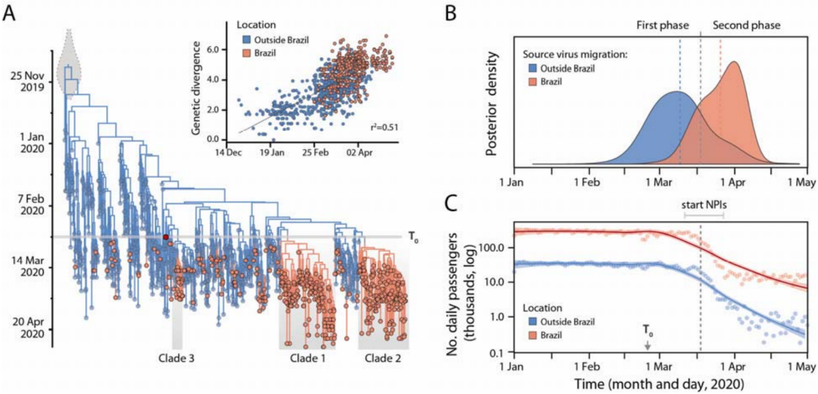


Example II: phylogeography = genomics + spatial analysis

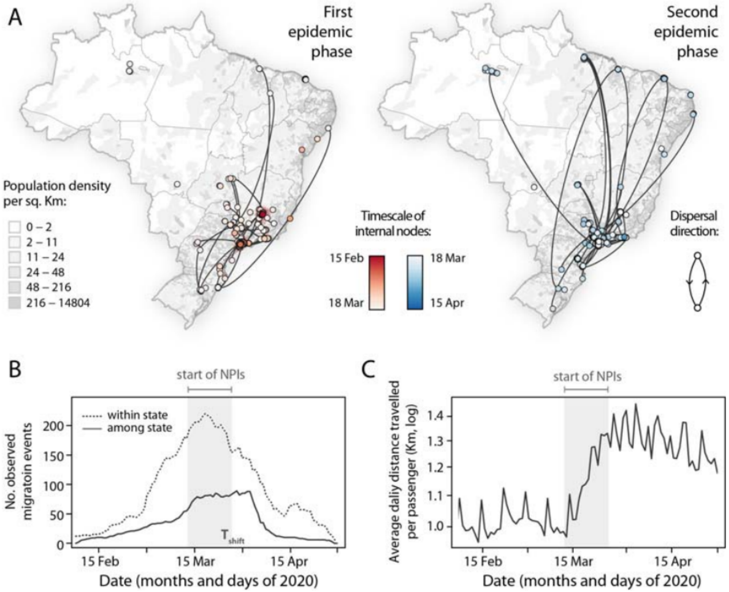
$P(t) = \exp(Qt)$ (Faria et al., 2011)



Example II: epidemic phases



Example II: dispersal patterns



Part 4: What have we learned?

After this bird's-eye tour, what have we gathered?

- The first truly "data-driven" pandemic has exposed the inadequacy of our reporting protocols;
- On the other hand, it has also shown the incredible potential for rapid sharing of data and methods and for global collaboration;
- Whilst we would like to understand *mechanisms*, semi-structured statistical models can help answer questions about
  - Factors associated with infection rates in each location;
  - Indicators of transmission risk such as R_t;
  - Disease dispersal patterns.

Practical recommendations

This is our dry-run for the end of the World. A much bigger and deadlier epidemic **will** come. How we prepare will tell whether we thrive or whither away as a species.

- Government: strengthen access to health care (Werneck & Carvalho, 2020);
- Government/funding agencies: fund surveillance programmes;
- Funding agencies: fund risk management projects targeting health crises;
- Modellers: research adequate methods for accommodating uncertainty;
- Modellers: research model-ensemble methods;

Take home

Data issues

Inconsistent criteria, delays, underreporting, incompleteness

Take home

Data issues

Inconsistent criteria, delays, underreporting, incompleteness

“Empirical” models can help answer pressing questions

Which factors are associated with infection rates? Where did the virus come from and where did it go?

Take home

Data issues

Inconsistent criteria, delays, underreporting, incompleteness

“Empirical” models can help answer pressing questions

Which factors are associated with infection rates? Where did the virus come from and where did it go?

Fund surveillance programmes

Accurate and timely monitoring are a must in ever more connected world.

THE
END