Machine Learning II
Week #1

Jan Nagler

Deep Dynamics Group
Centre for Human and Machine Intelligence (HMI)
Frankfurt School of Finance & Management

April-May 2020

Outline

Given module description
+
Covid-19 Data science / Modelling / Prediction / Mitigation

Do not record or distribute!

Guidelines

Required: Attend online lectures!
(FS asks profs not to share complete materials)

Ask questions!
In class: Raise hand (@zoom, wait until prof responds)
Questions that may disturb the flow:
Ask via gmail jan.nagler@gmail.com
Answers may be given immediately, in following lecture,
or in private communication

Assignments

Total 5 Assignments (#1-#5, max 14 points each):
You may start in class
Read assignment
Avoid useless graphs
Check before submission: Avoid pdf graphs over page limit
Avoid useless copy and paste
Answer the questions and do not leave irrelevant stuff in
Make the code your code
Only submit Python notebook pdf and code (no html)
Filenames with your name

Presentation of solutions in class,
either by prof or by randomly picked collaborator
Credits based both on
(i) submitted code,
(ii) readability (comments in notebook) and
(iii) presentation in class (if picked)

Smaller Python assignments (max 2 collaborators)
Mini projects assignments (max 4 collaborators)

Covid-19 Data science / modelling / prediction / mitigation

Secunderabad, India, April 2020

Covid-19

Biology

Physics

Mitigation

Data science

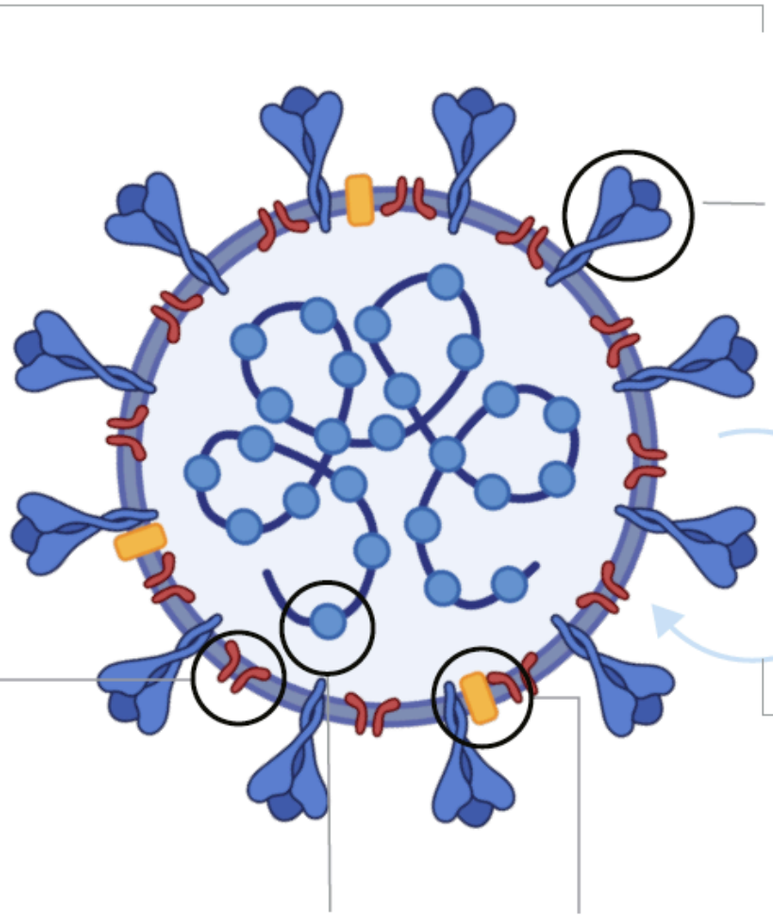
Modelling

Graded Mini projects

Biology of Covid-19

Size & Content

Diameter: ≈100 nm
Volume: ~10^6 nm^3 = 10^{-3} fL
Mass: ~10^3 MDa ≈ 1 fg



Spike trimer
Length: ≈10 nm
Copies per virion: ≈100
(300 monomers, measured for SARS-CoV-1)
Affinity to ACE2 receptor K_d: ≈1-30 nM
primed by TMPRSS2

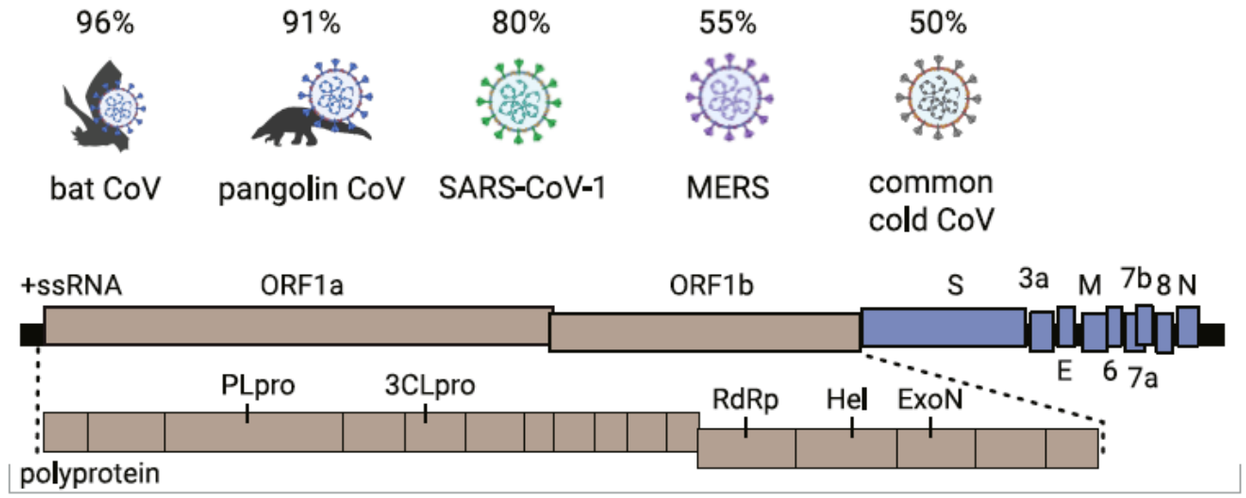
Membrane protein
≈2000 copies
(measured for SARS-CoV-1)

Nucleoprotein
≈1000 copies
(measured for SARS-CoV-1)

Envelope protein
≈20 copies
(100 monomers, measured for TGEV coronavirus)

Genome

Nucleotide identity to SARS-CoV-2



Length: ≈30kb; β-coronavirus with 10-14 ORFs (24-27 proteins)

Evolution rate: ~10^{-3} nt^{-1} yr^{-1} (measured for SARS-CoV-1)
Mutation rate: ~10^{-6} nt^{-1} $cycle^{-1}$ (measured for MHV coronavirus)

Replication Timescales

in tissue-culture
Virion entry into cell: ~10 min (measured for SARS-CoV-1)
Eclipse period: ~10 hrs (time to make intracellular virions)
Burst size: ~10^3 virions (measured for MHV coronavirus)

Sars-Covid-19 by the numbers, eLife, 2020

Biology of Covid-19

Antibody Response - Seroconversion

Antibodies appear in blood after: ≈10-20 days
Maintenance of antibody response:
≈2-3 years (measured for SARS-CoV-1)

Virus Environmental Stability

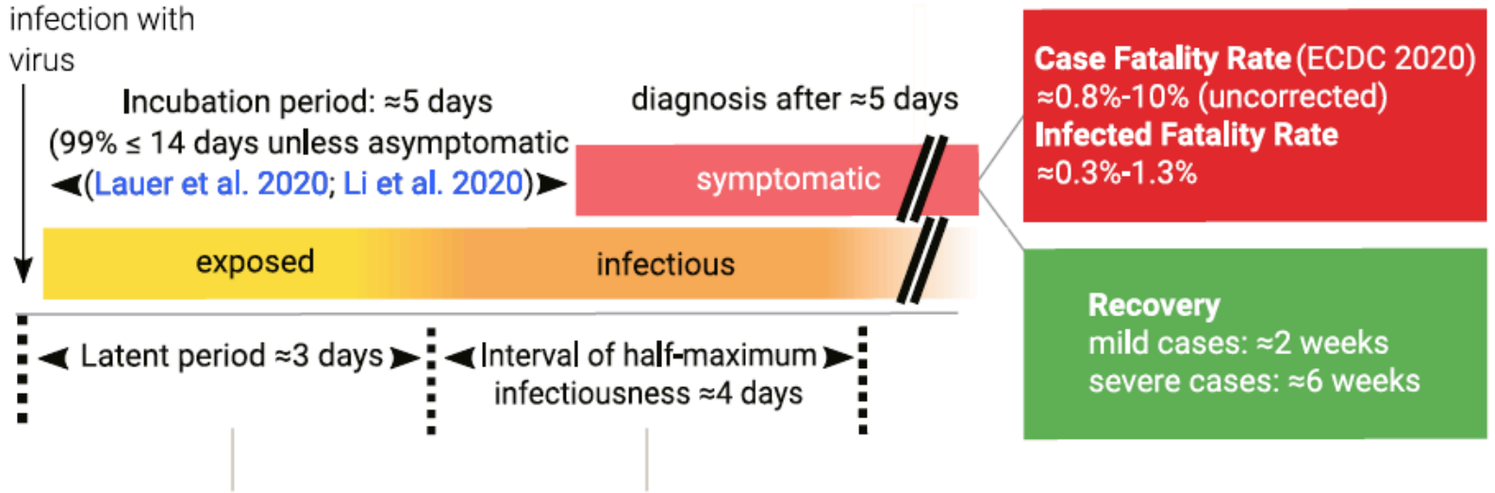
Relevance to personal safety unclear

	half-life	time to decay 1000-fold
Aerosols:	≈1 hr	≈4-24 hr
Surfaces: e.g. plastic, cardboard and metals	≈1-7 hr (van Doremalen et al. 2020)	≈4-96 hr

Based on quantifying infectious virions. Tested at 21-23°C and 40-65% relative humidity. Numbers will vary between conditions and surface types (Otter et al. 2016).
Viral RNA observed on surfaces even after a few weeks (Moriarty et al. 2020).

"Characteristic" Infection Progression in a Single Patient

Basic reproductive number R_0: typically 2-4
Varies further across space and time (Li et al. 2020; Park et al. 2020)
(number of new cases directly generated from a single case)



Inter-individual variability is substantial and not well characterized. The estimates are parameter fits for population median in China and do not describe this variability (Li et al. 2020; He et al. 2020).

Note the difference in notation between the symbol ≈, which indicates "approximately" and connotes accuracy to within a factor 2, and the symbol ~, which indicates "order of magnitude" or accuracy to within a factor of 10.

Sars-Covid-19 by the numbers, eLife, 2020

Biology of Covid-19
Phylogeny (Mutations)

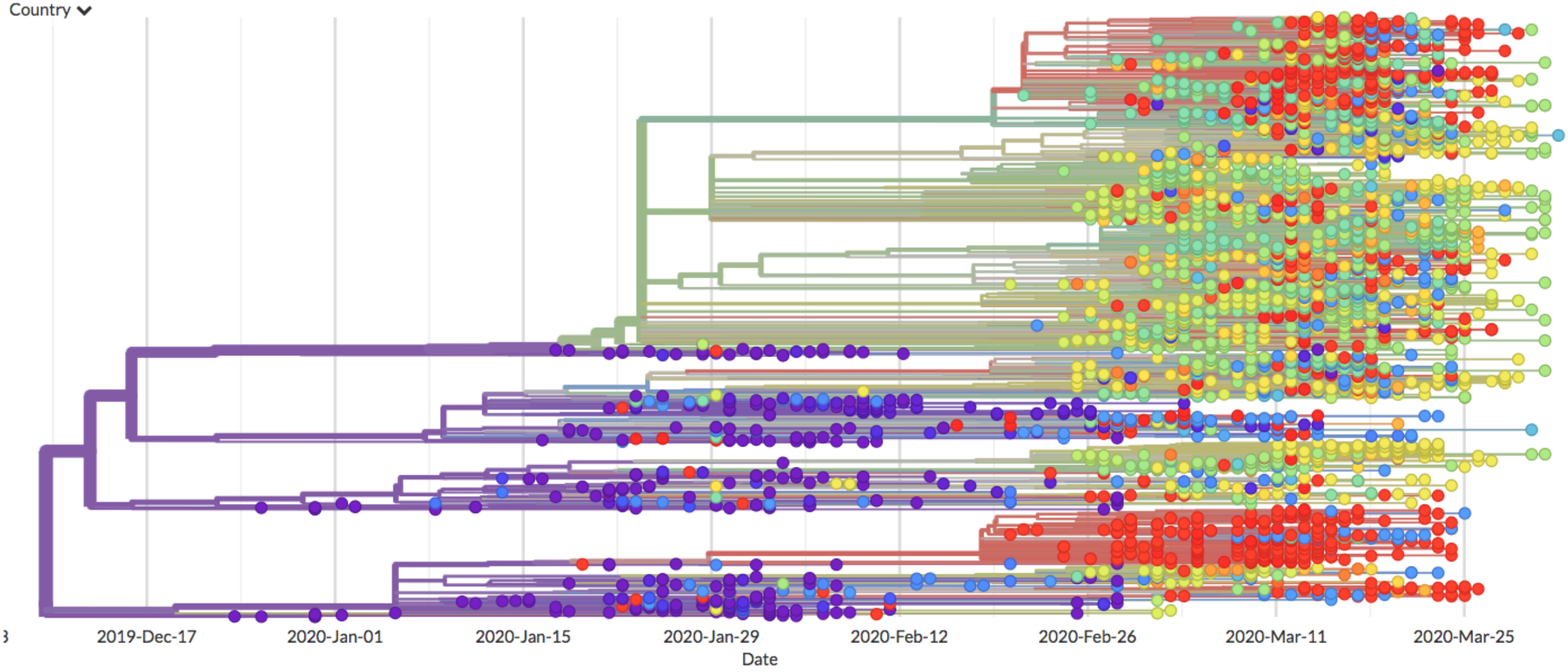
Genomic epidemiology of hCoV-19

Showing 3123 of 3123 genomes sampled between Dec 2019 and Apr 2020.

Phylogeny

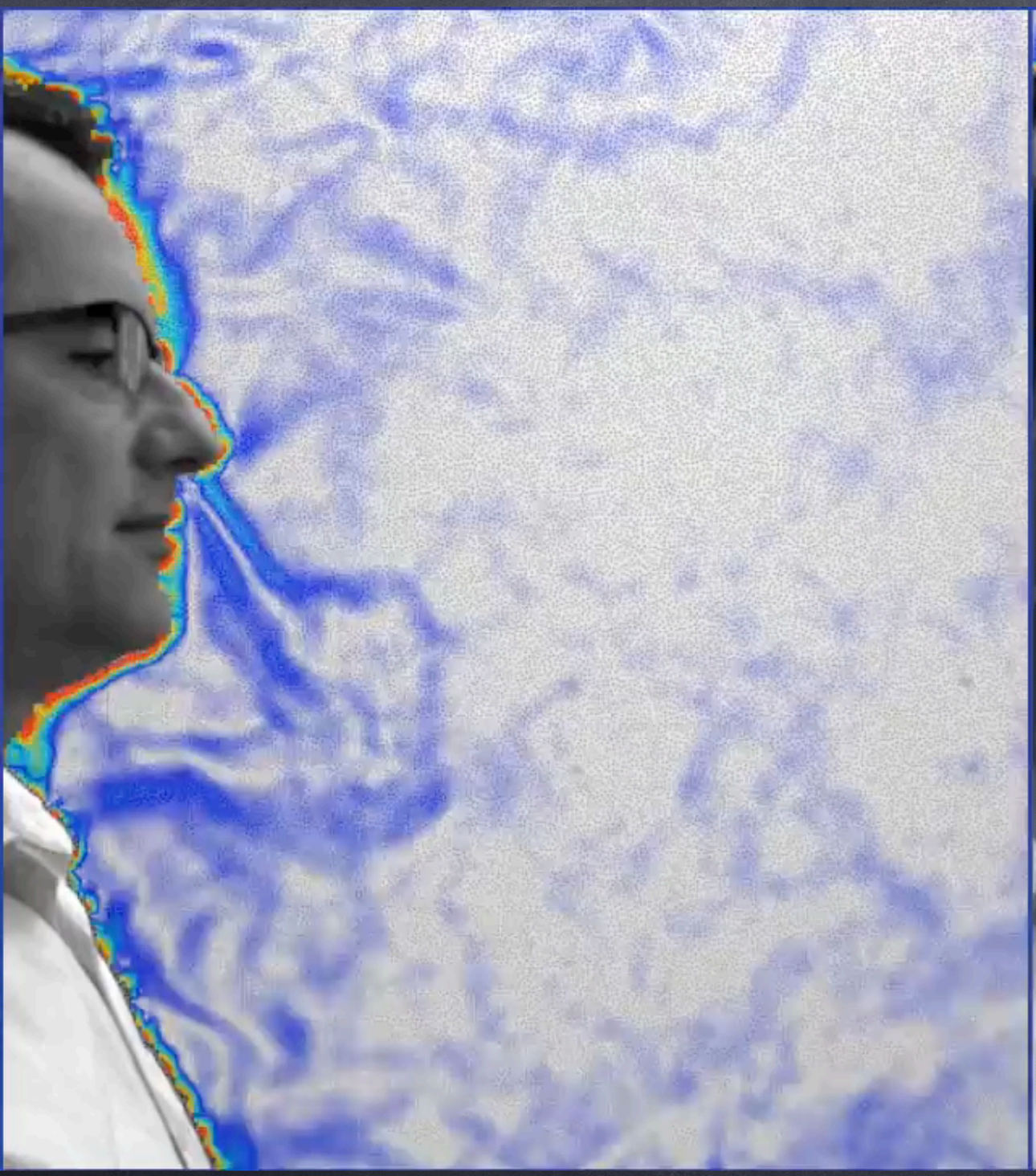
Country

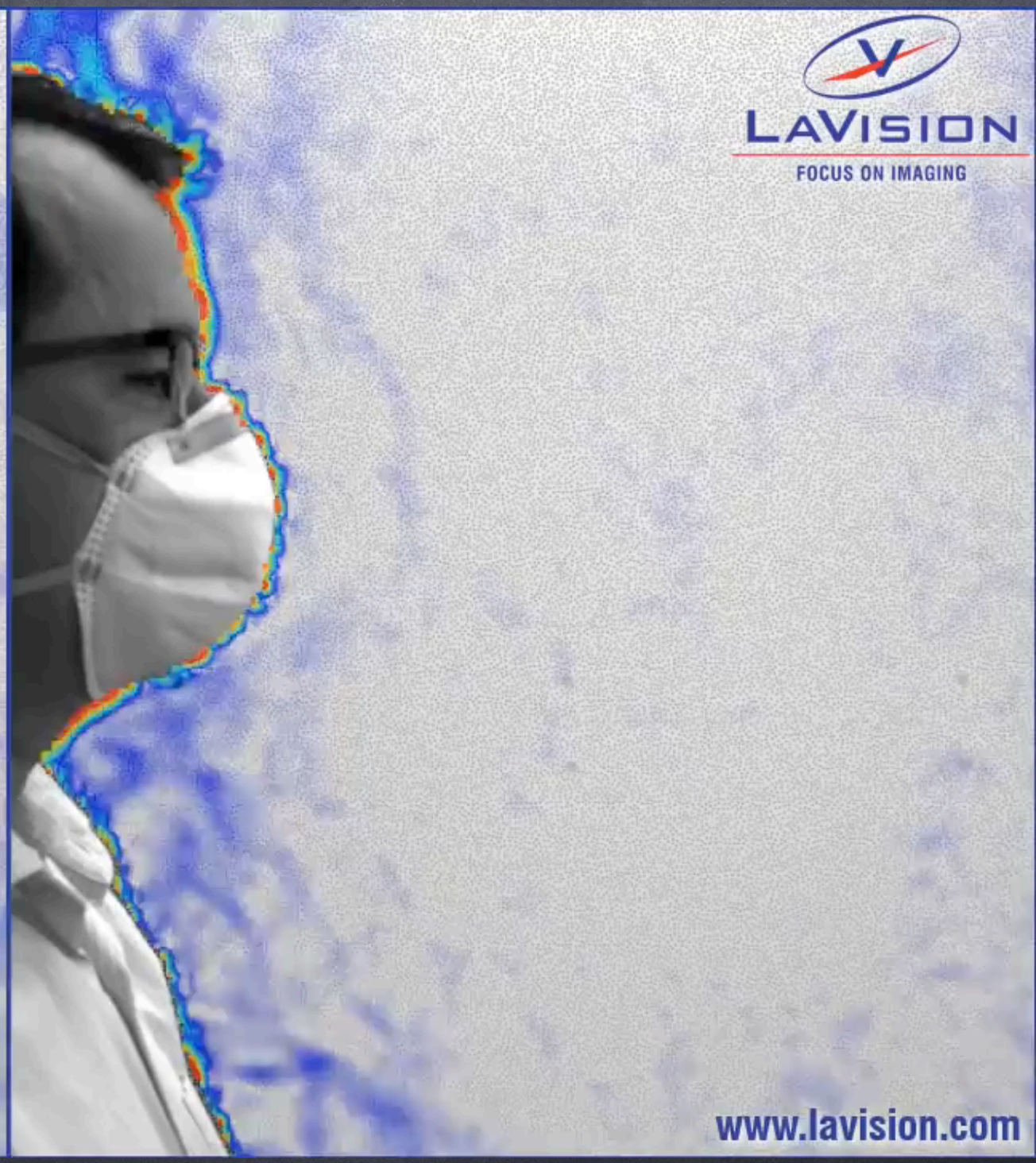
RESET LAYOUT



https://www.gisaid.org/epiflu-applications/next-hcov-19-app/

Physics of Covid-19

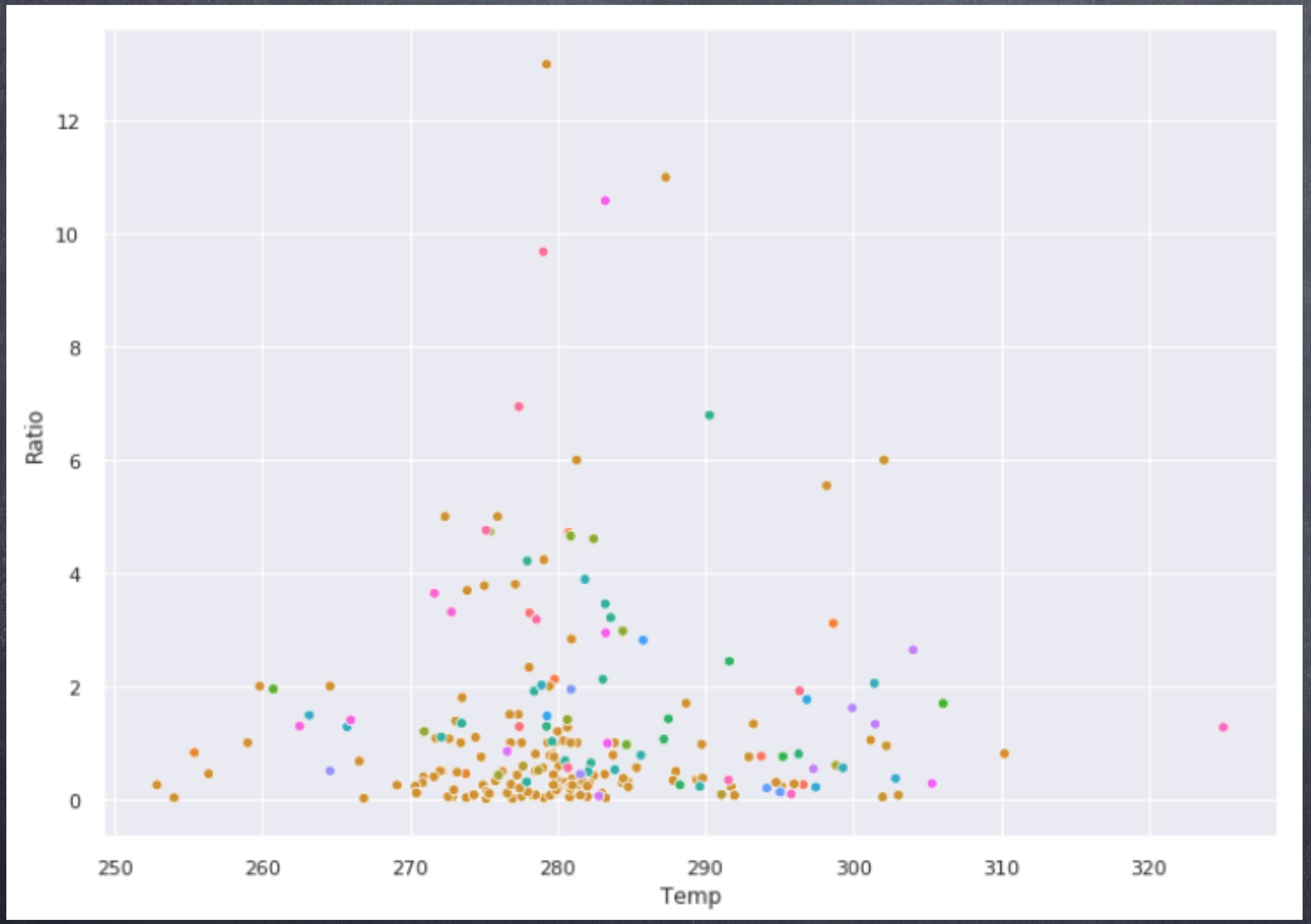




Runners + Cyclists ?

Weather & Socio-economic factors of Covid-19

(Hard data science problem)



Aguilar & Nagler (2020, preliminary)

https://www.sciencemag.org/news/2020/03/how-diseases-rise-and-fall-seasons-and-what-it-could-mean-coronaviru
utm_campaign=news_weekly_2020-03-27&et_rid=165593395&et_cid=3263545

Covid-19 (Interactive) Data Visualisation

John Hopkins data access

https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

Worldometer data access

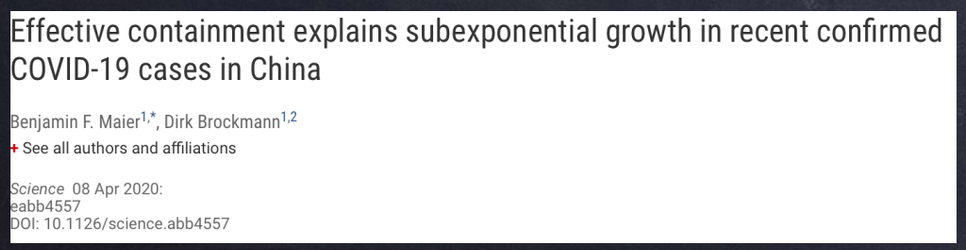
https://www.worldometers.info/coronavirus/worldwide-graphs/

YY Ahn's Trend Visualizations

https://yyahn.com/covid19/

D Brockmann's Prediction page

http://rocs.hu-berlin.de/corona/docs/forecast/results_by_country/


Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China

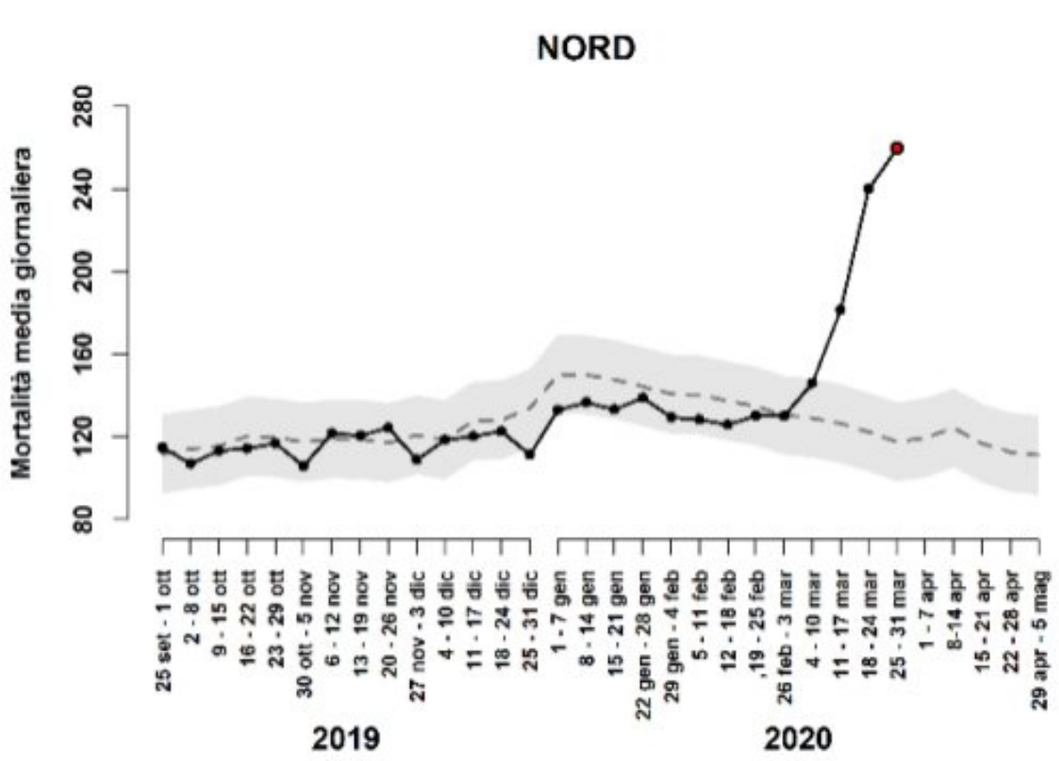
Benjamin F. Maier[1,*], Dirk Brockmann[1,2]

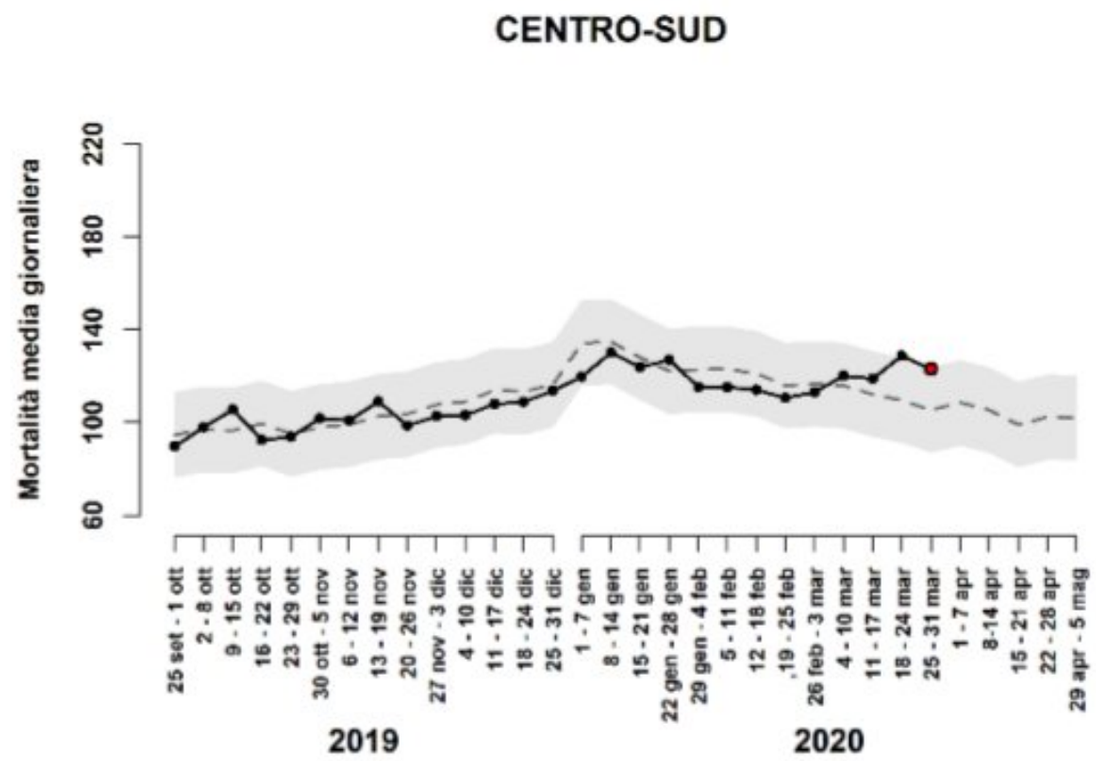
+ See all authors and affiliations

Science 08 Apr 2020:
eabb4557
DOI: 10.1126/science.abb4557

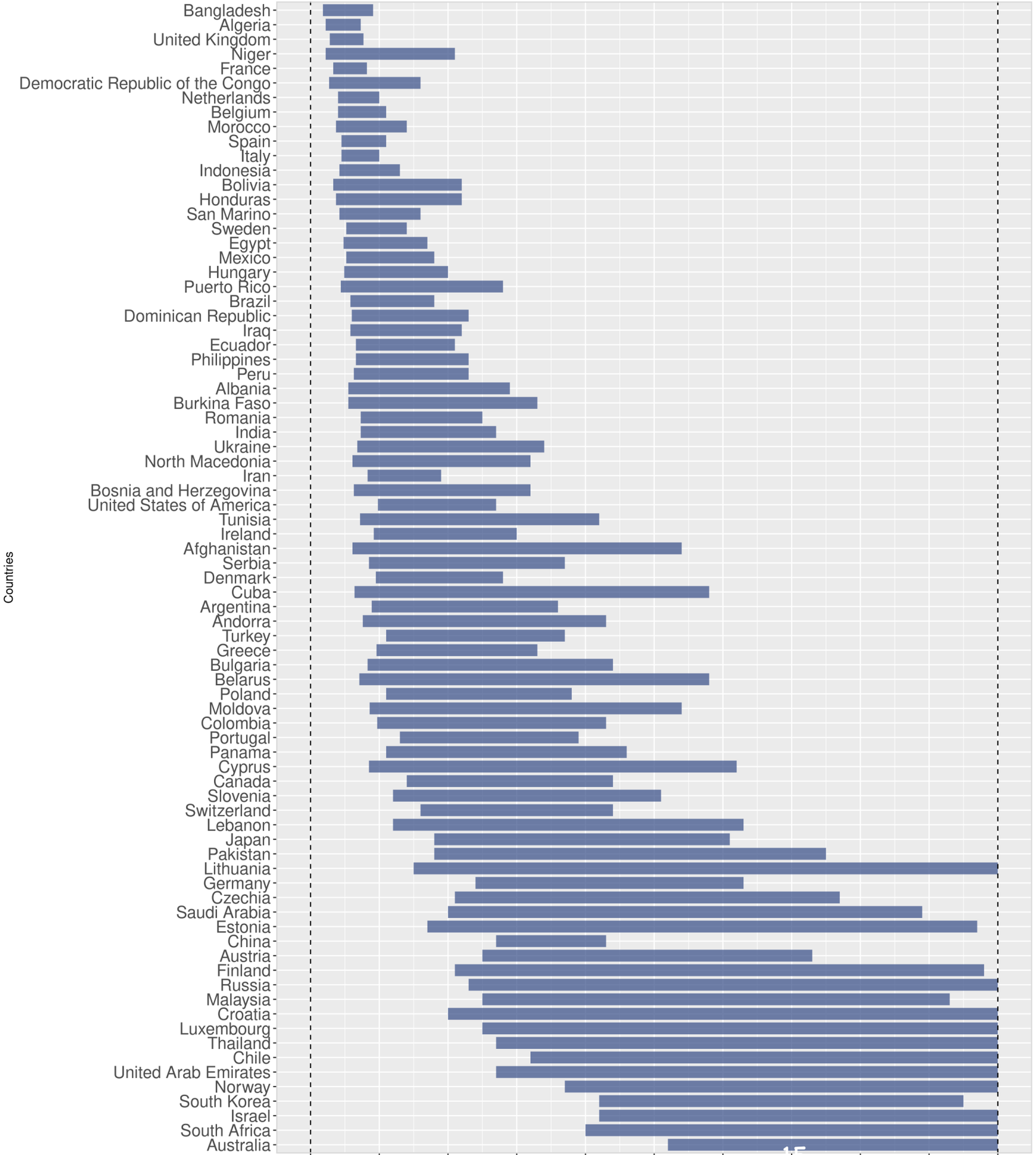
refs++ on webpages

Geographic heterogeneity of Covid-19 deaths in Italy





Numbers aggregated by a country can be very misleading
and mask the severity of covid-19 in heavily affected communities



Bias in Covid19

Percentage of reported (symptomatic) Covid-19 cases

https://cmmid.github.io/topics/covid19/severity/global_cfr_estimates.html

April 14 Covid-19 Science Review

(Highlights, see following slides)

A priest in Innsbruck, Austria, views photographs of his absent congregation. Austria eased social distancing today. JAN HETFLEISCH/GETTY IMAGES

Ending coronavirus lockdowns will be a dangerous process of trial and error

By **Kai Kupferschmidt** | Apr. 14, 2020 , 4:10 PM

https://www.sciencemag.org/news/2020/04/ending-coronavirus-lockdowns-will-be-dangerous-process-trial-and-error

Basic and effective reproduction number

Basic reproduction number R_0

Expected number of cases directly caused by one case in a population where all individuals are susceptible,
e.g., at time 0, pronounced “R nought”

Effective reproduction number R

Expected number of cases directly caused by one case in a population, time dependent

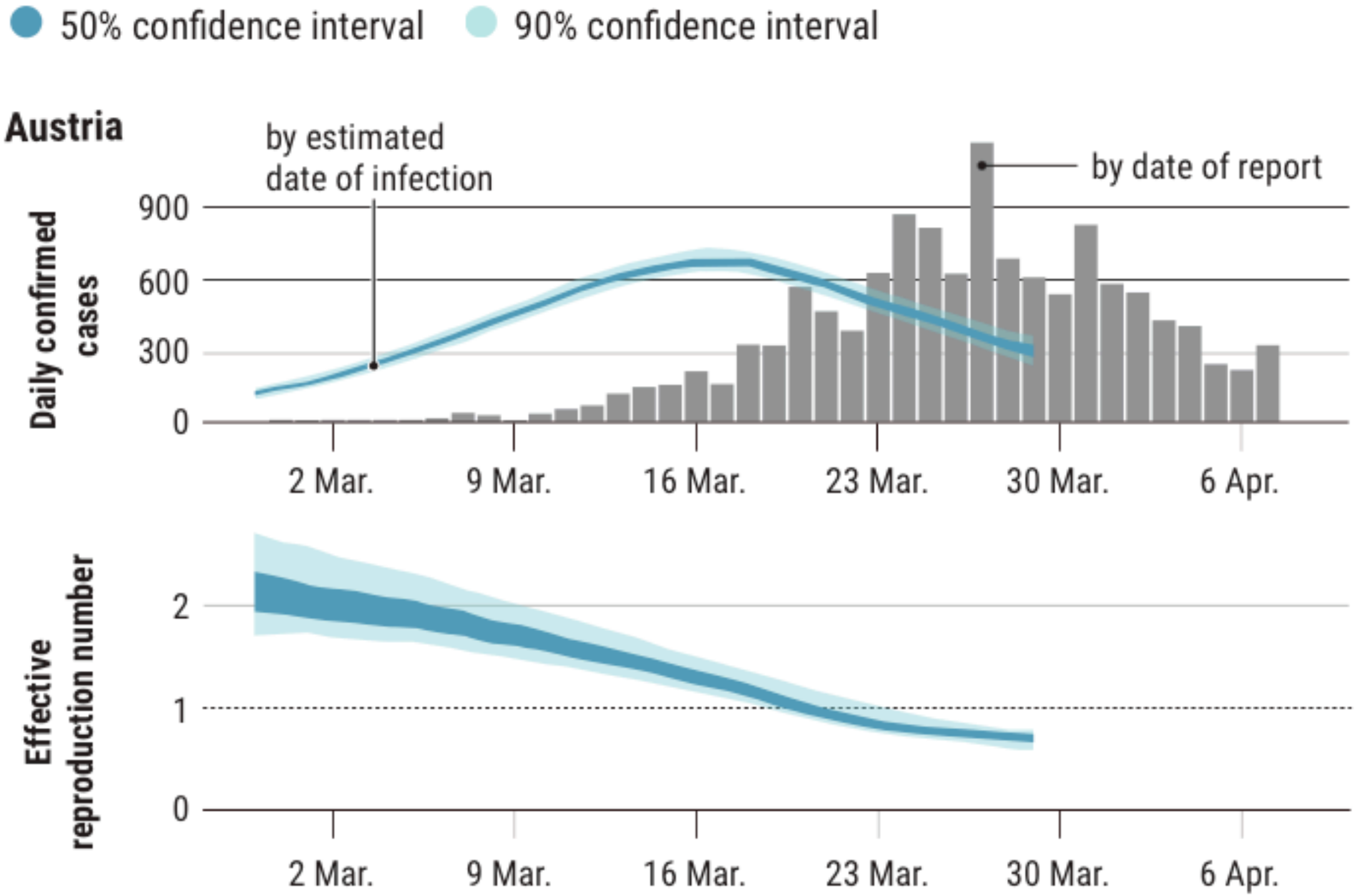
If the reproduction number is smaller than 1, the disease dies out;
$R < 1$ if it is larger than 1, there is exponential spreading $R > 1$

Mitigation

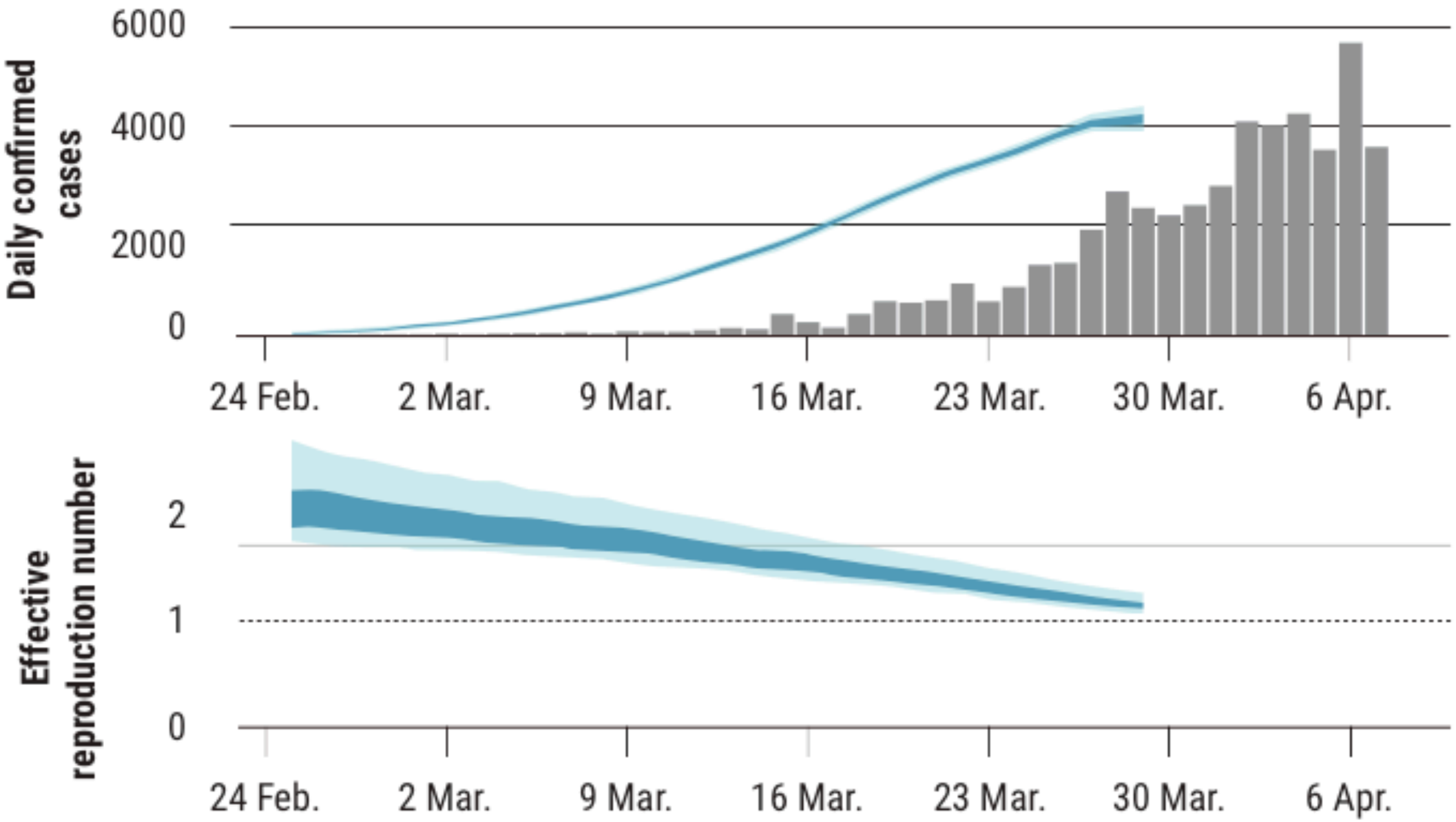
What mitigation can achieve and has achieved

The number to watch

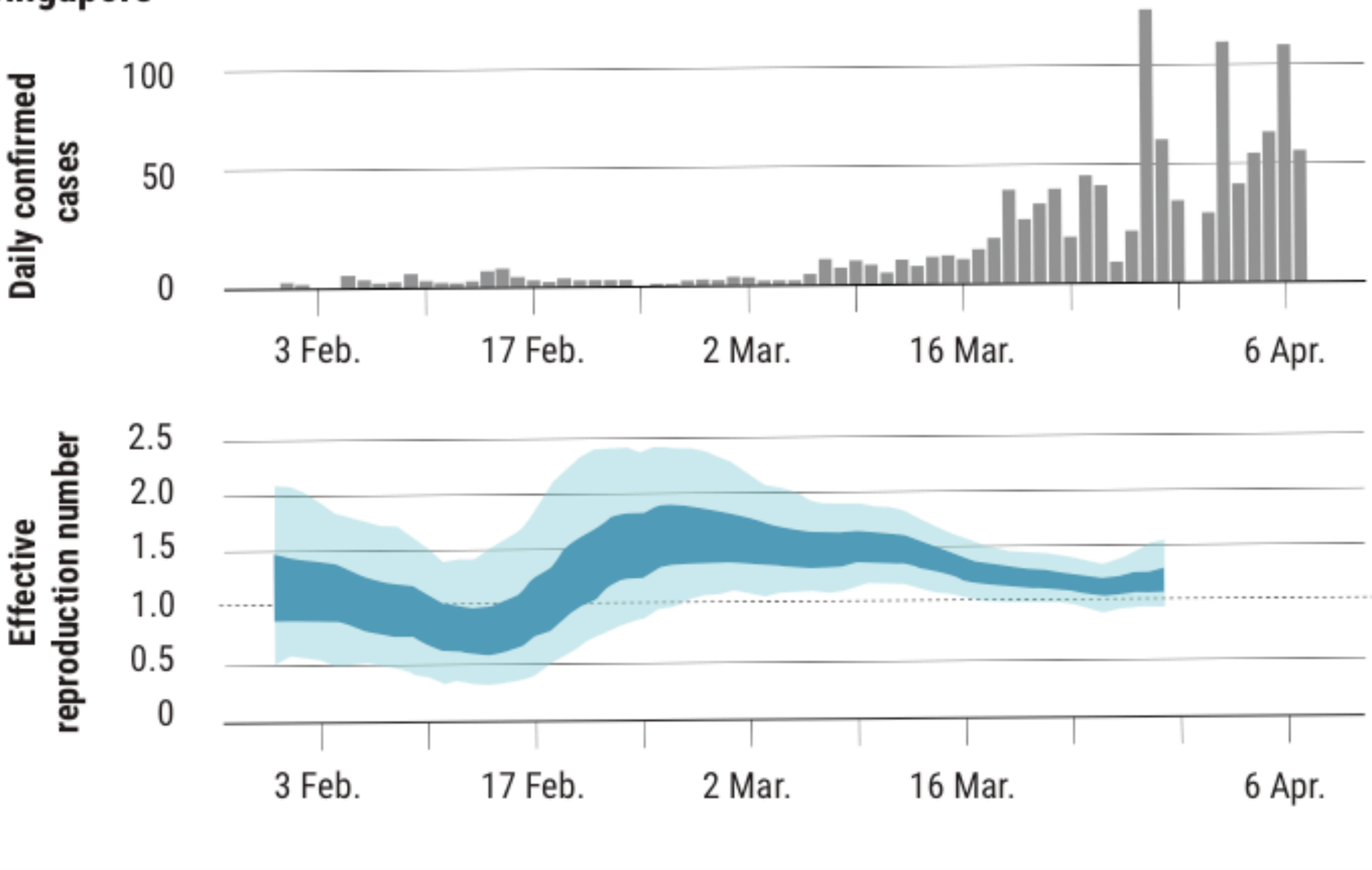
Lockdowns lower the number of new cases as well as R, the effective reproduction number. If R drops below 1, the epidemic shrinks.



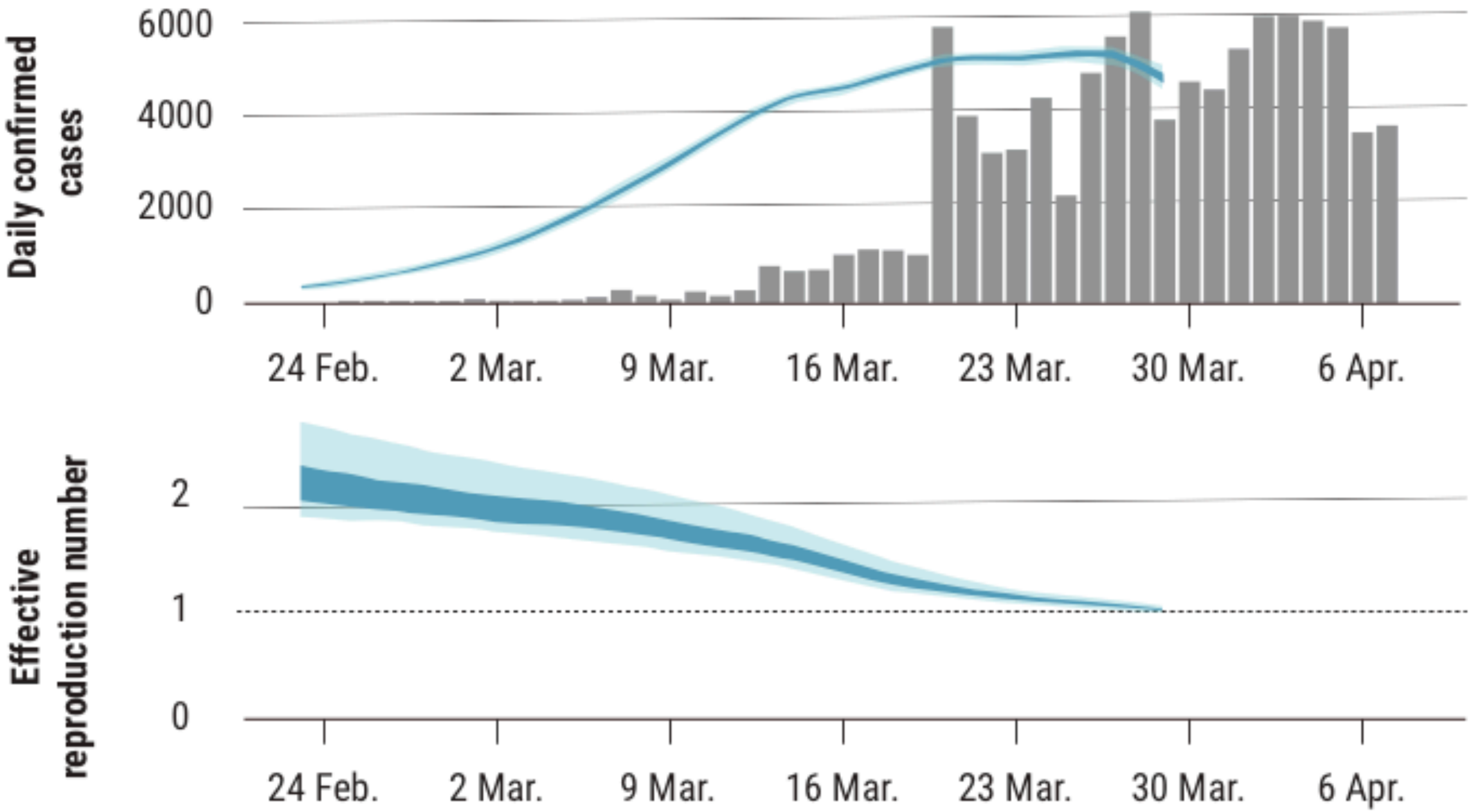
United Kingdom



Singapore


Daily confirmed cases
100
50
0
3 Feb.
17 Feb.
2 Mar.
16 Mar.
6 Apr.
Effective reproduction number
2.5
2.0
1.5
1.0
0.5
0
3 Feb.
17 Feb.
2 Mar.
16 Mar.
6 Apr.

Germany


Daily confirmed cases
6000
4000
2000
0
24 Feb.
2 Mar.
9 Mar.
16 Mar.
23 Mar.
30 Mar.
6 Apr.
Effective reproduction number
2
1
0
24 Feb.
2 Mar.
9 Mar.
16 Mar.
23 Mar.
30 Mar.
6 Apr.

Ergodicity breaking in reproduction

$$\langle R \rangle = E[R] > 1$$

If the reproduction number is larger than 1, on average,
but fluctuating in a given time window
the disease may die out!

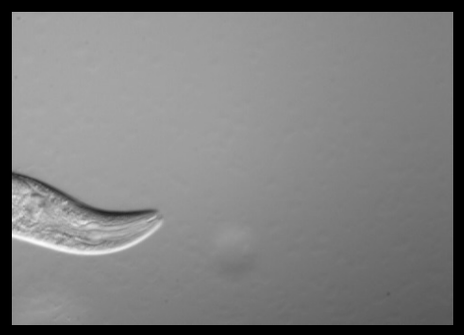
Population Growth

Change of population size

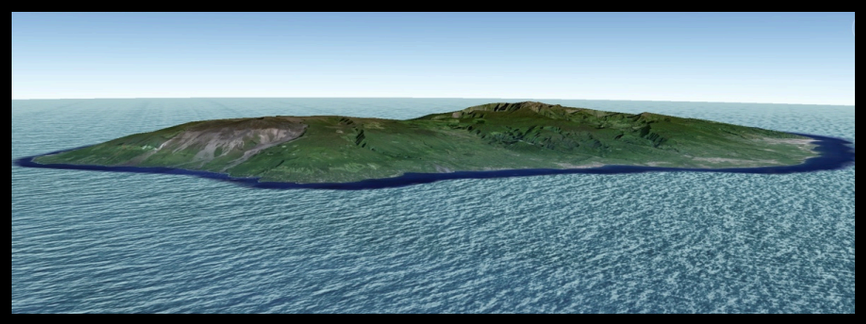
Population size

$$\frac{\mathrm{d}S}{\mathrm{d}t} = r\,S$$

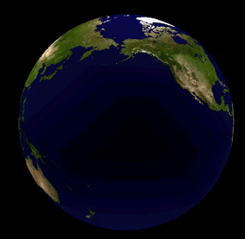
Growth factor
(fluctuating, coupling to other species)



Organisms



Ecosystems

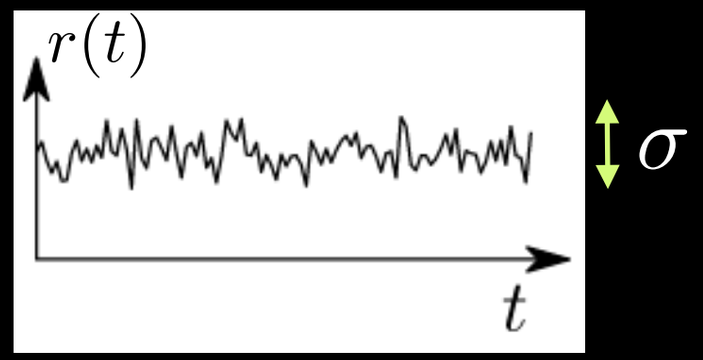


Networks

Growth is typically non-ergodic

(finance: geometric Brownian motion)

$$\frac{\mathrm{d}S}{\mathrm{d}t} = r\,S$$



→ Fluctuation mean $\mu = \langle r \rangle$ may not tell much about mean growth of S

Correct growth rate (time average) $\bar{r} = \mu - \sigma^2/2$

Not only the mean but also the variance crucially determines the long-term behavior.

Peters, Optimal leverage from non-ergodicity (2010)
Maslov & Zhang, Int. J. Theor. Appl. Finan. 01:377 (1998)
Peters & Klein, Ergodicity breaking in geometric Brownian motion, PRL (2011)

Ergodicity of (stationary) stochastic process

$$\underbrace{\lim_{T\to\infty}\frac{1}{T}\int_0^T f(\omega(t))\mathrm{d}t}_{\text{Time average of } f} = \underbrace{\int_\Omega f(\omega)P(\omega)\mathrm{d}\omega}_{\text{Expectation value of } f}$$

f observable

$\omega(=S)$ state

P pdf

Eq (1) in Peters, Nature Physics, 2019

PERSPECTIVE
https://doi.org/10.1038/s41567-019-0732-0

nature physics

Corrected: Author Correction

The ergodicity problem in economics

Ole Peters

The ergodic hypothesis is a key analytical device of equilibrium statistical mechanics. It underlies the assumption that the time average and the expectation value of an observable are the same. Where it is valid, dynamical descriptions can often be replaced with much simpler probabilistic ones — time is essentially eliminated from the models. The conditions for validity are restrictive, even more so for non-equilibrium systems. Economics typically deals with systems far from equilibrium — specifically with models of growth. It may therefore come as a surprise to learn that the prevailing formulations of economic theory — expected utility theory and its descendants — make an indiscriminate assumption of ergodicity. This is largely because foundational concepts to do with risk and randomness originated in seventeenth-century economics, predating by some 200 years the concept of ergodicity, which arose in nineteenth-century physics. In this Perspective, I argue that by carefully addressing the question of ergodicity, many puzzles besetting the current economic formalism are resolved in a natural and empirically testable way.

Ergodic theory is a forbiddingly technical branch of mathematics. Luckily, for the purpose of this discussion, we will need virtually none of the technicalities. We will call an observable ergodic if its time average equals its expectation value, that is, if it satisfies Birkhoff's equation

$$\underbrace{\lim_{T\to\infty}\frac{1}{T}\int_0^T f(\omega(t))\mathrm{d}t}_{\text{Time average of } f} = \underbrace{\int_\Omega f(\omega)P(\omega)\mathrm{d}\omega}_{\text{Expectation value of } f} \qquad (1)$$

Here, f is determined by the system's state ω. On the left-hand

Scientifically, this deserves some reflection: the models were exonerated by declaring the object of study irrational.

I stumbled on this error about a decade ago, and with my collaborators at the London Mathematical Laboratory and the Santa Fe Institute I have identified a number of long-standing puzzles or paradoxes in economics that derive from it. If we pay close attention to the ergodicity problem, natural solutions emerge. We therefore have reason to be optimistic about the future of economic theory.

This Perspective is structured as follows. I will first sketch the conceptual basis of mainstream economic theory: discounted expected utility. I will then develop our conceptually different approach, based on addressing the ergodicity problem, and estab-

Climate change and ecosystems

Fluctuations may be more important than average growth!

Clear signs of global warming will hit poorer countries first (Nature News, April 20, 2018)

The lockdown, general and special effects

Benedetta Cerruti Follow
Apr 14 · 4 min read

The normalized number of daily deaths is
insensitive to the details of the lockdown
implemented in different countries
and depends mostly on the time when the lockdown started

Bradde & Cerruti, The lockdown, general and special effects, Medium

Covid-19 Data

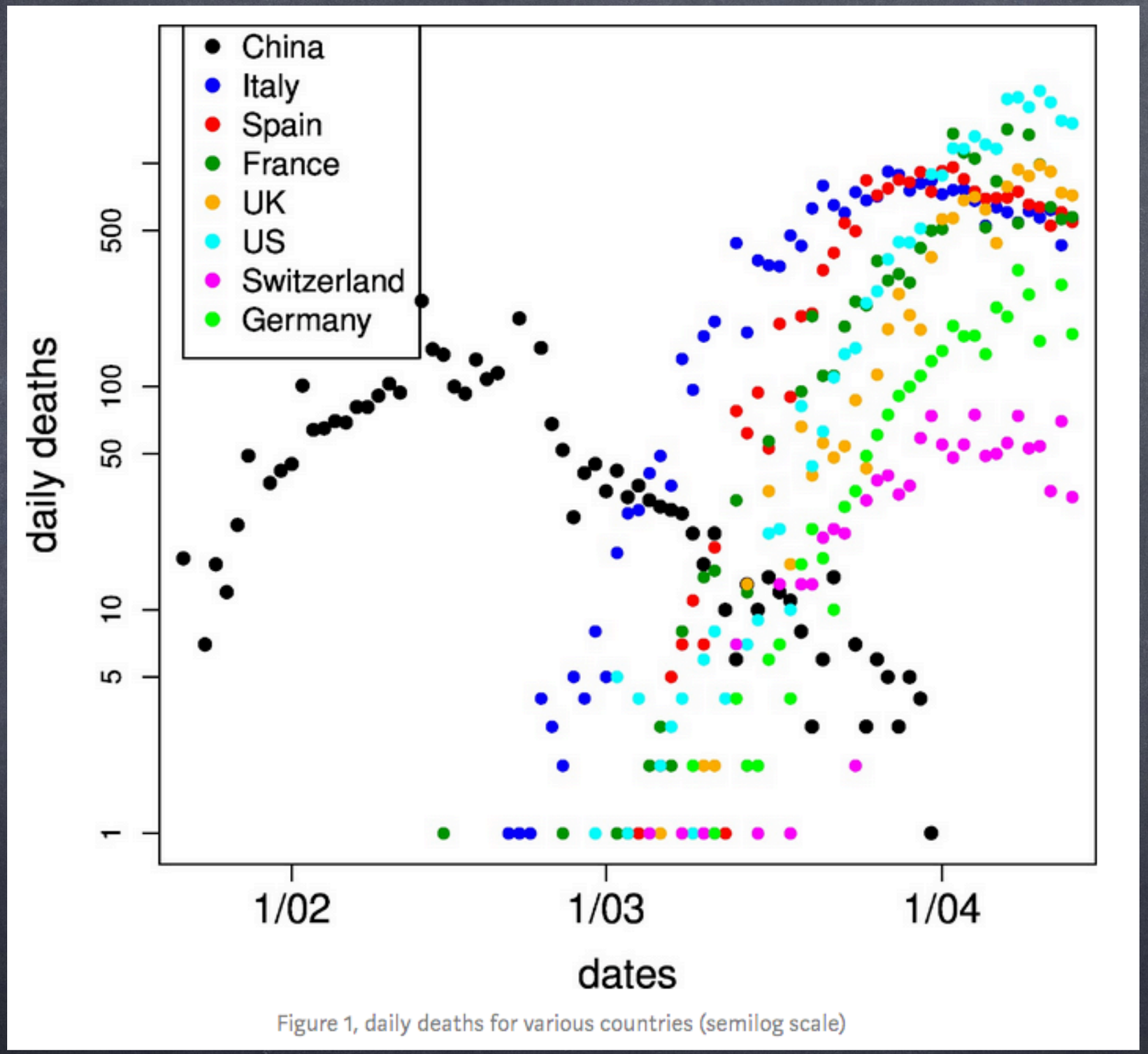


Figure 1, daily deaths for various countries (semilog scale)

Bradde & Cerruti, The lockdown, general and special effects, Medium

Covid-19 Data collapse

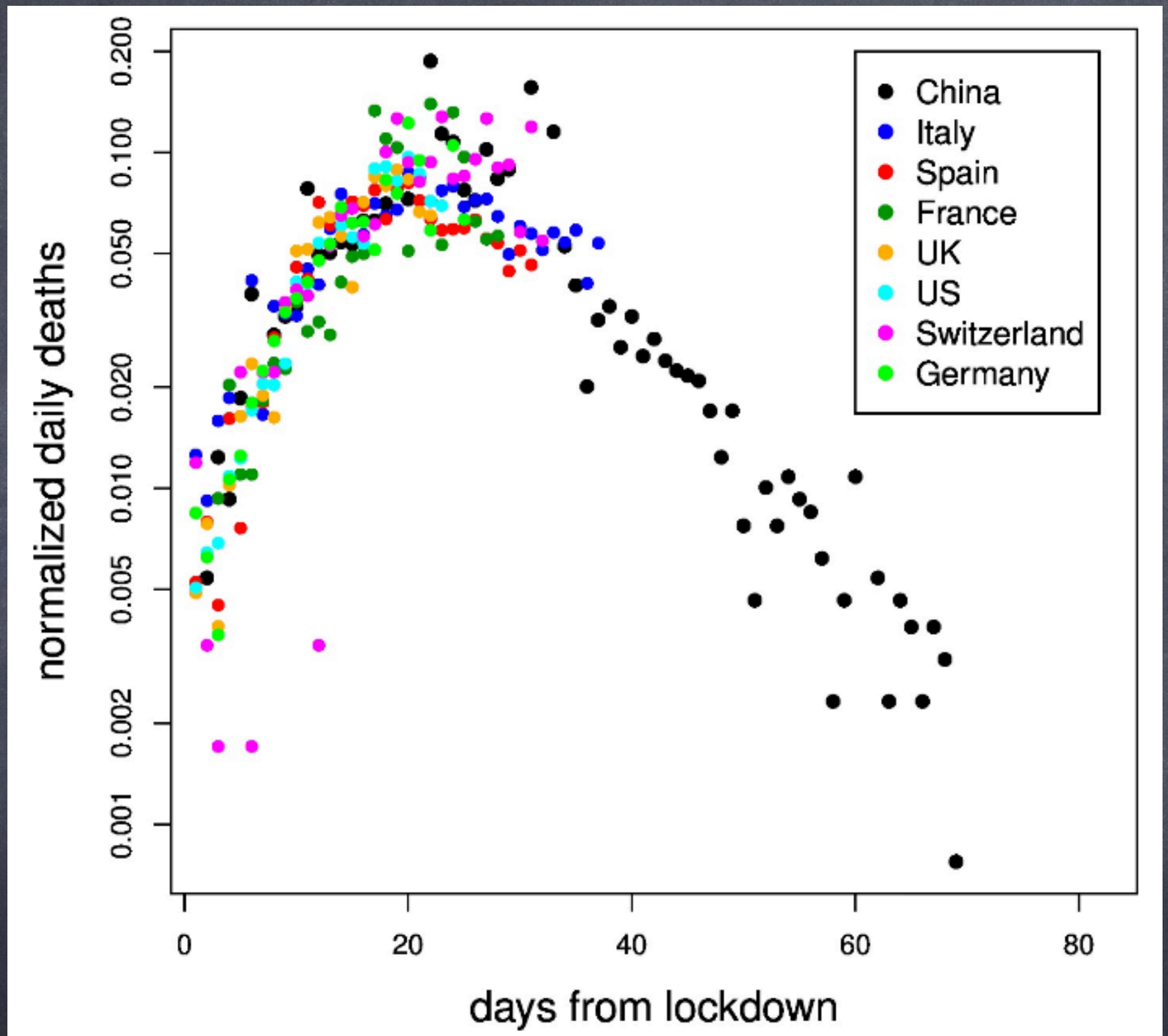
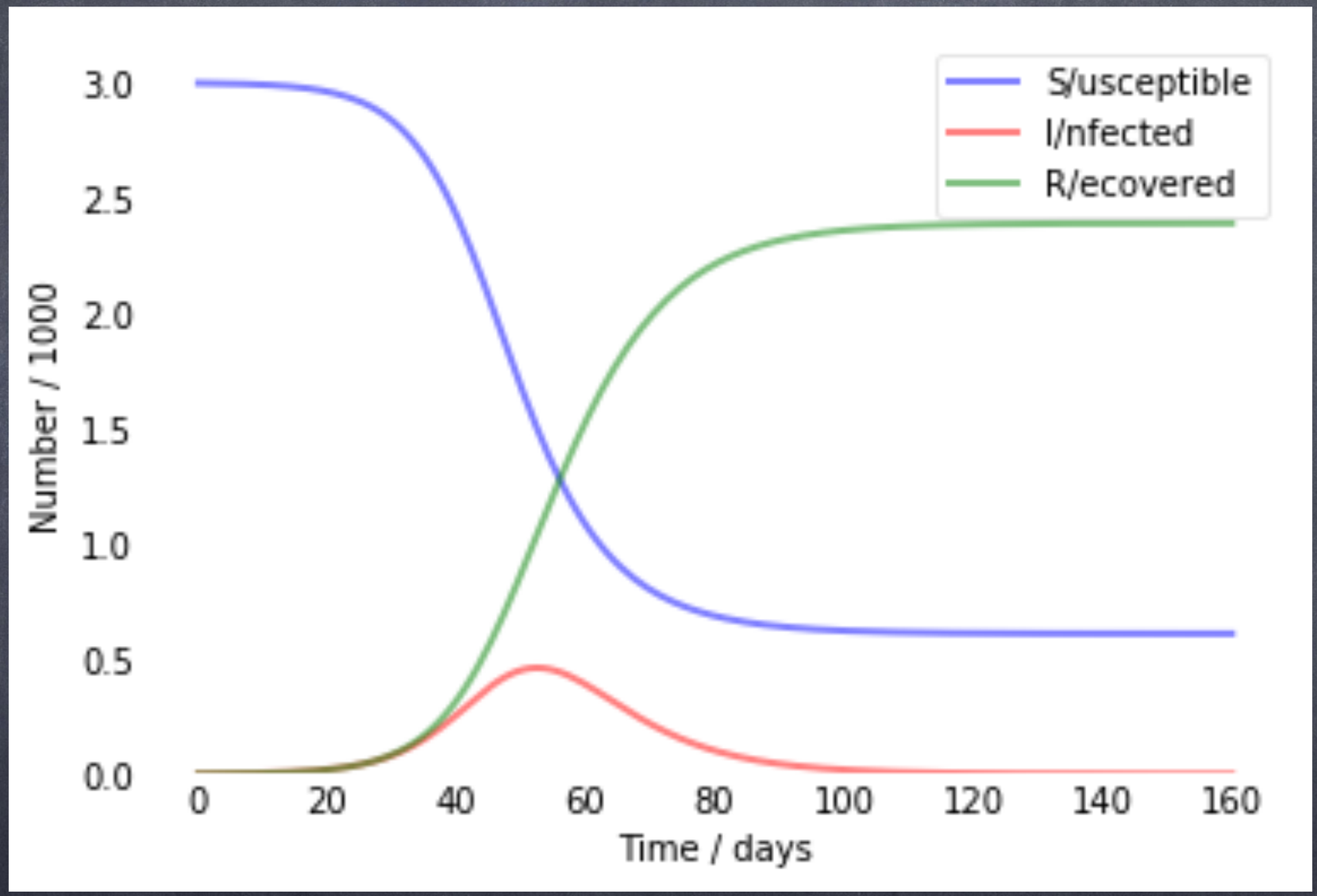


Figure 2, normalized daily deaths for various countries as a function of the time distance from lockdown (semilog scale)

Bradde & Cerruti, The lockdown, general and special effects, Medium

Covid-19 Modelling — SIR Model

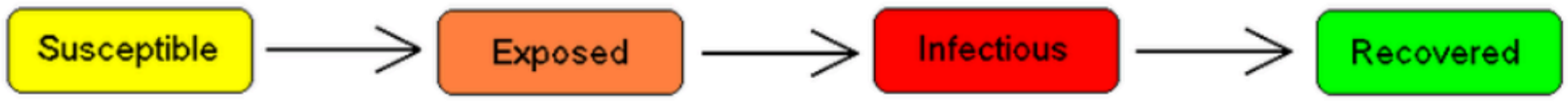


see python

Covid-19 Modelling — SEIR Model

The SEIR model [edit]

For many important infections, there is a significant incubation period during which individuals have been infected but are not yet infectious themselves. During this period the individual is in compartment *E* (for exposed).



Assuming that the incubation period is a random variable with exponential distribution with parameter a (i.e. the average incubation period is a^{-1}), and also assuming the presence of vital dynamics with birth rate Λ equal to death rate μ, we have the model:

$$\frac{dS}{dt} = \Lambda - \mu S - \beta \frac{I}{N} S$$

$$\frac{dE}{dt} = \beta \frac{I}{N} S - (\mu + a) E$$

$$\frac{dI}{dt} = aE - (\gamma + \mu) I$$

$$\frac{dR}{dt} = \gamma I - \mu R.$$

We have $S + E + I + R = N,$ but this is only constant because of the (degenerate) assumption that birth and death rates are equal; in general N is a variable.

For this model, the basic reproduction number is:

$$R_0 = \frac{a}{\mu + a} \frac{\beta}{\mu + \gamma}.$$

Compartmental models in epidemiology, Wikipedia

Covid-19 Modelling: SIRX model

D Brockmann's Prediction page

http://rocs.hu-berlin.de/corona/docs/forecast/results_by_country/

Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China

Benjamin F. Maier[1,*], Dirk Brockmann[1,2]
+ See all authors and affiliations

Science 08 Apr 2020:
eabb4557
DOI: 10.1126/science.abb4557

$$\partial_t S = -\alpha SI - \kappa_0 S \quad (1)$$

$$\partial_t I = \alpha SI - \beta I - \kappa_0 I - \kappa I \quad (2)$$

$$\partial_t R = \beta I + \kappa_0 S \quad (3)$$

$$\partial_t X = (\kappa + \kappa_0) I \quad (4)$$

a generalization of the standard SIR model, henceforth referred to as the SIR-X model. The rate parameters α and β quantify the transmission rate and the recovery rate of the standard SIR model, respectively. Additionally, the impact of containment efforts is captured by the terms proportional to the containment rate κ_0 that is effective in both I and S populations, since measures like social distancing and curfews affect the whole population alike. Infected individuals are removed at rate κ corresponding to quarantine measures that only affect symptomatic infecteds. The new compartment X quantifies symptomatic, quarantined infecteds.

Maier & Brockmann, Science 10.1126/science.abb4557 (April 8, 2020)

Covid-19 Modelling — More complex Models

Li et al, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2), Science, March 2020, DOI: 10.1126/science.abb3221

Materials and Methods

1. Model Configuration and Initialization

The transmission model incorporates information on human movement within the following metapopulation structure:

$$\frac{dS_i}{dt} = -\frac{\beta S_i I_i^r}{N_i} - \frac{\mu\beta S_i I_i^u}{N_i} + \theta\sum_j \frac{M_{ij}S_j}{N_j - I_j^r} - \theta\sum_j \frac{M_{ji}S_i}{N_i - I_i^r} \quad [1]$$

$$\frac{dE_i}{dt} = \frac{\beta S_i I_i^r}{N_i} + \frac{\mu\beta S_i I_i^u}{N_i} - \frac{E_i}{Z} + \theta\sum_j \frac{M_{ij}E_j}{N_j - I_j^r} - \theta\sum_j \frac{M_{ji}E_i}{N_i - I_i^r} \quad [2]$$

$$\frac{dI_i^r}{dt} = \alpha\frac{E_i}{Z} - \frac{I_i^r}{D} \quad [3]$$

$$\frac{dI_i^u}{dt} = (1-\alpha)\frac{E_i}{Z} - \frac{I_i^u}{D} + \theta\sum_j \frac{M_{ij}I_j^u}{N_j - I_j^r} - \theta\sum_j \frac{M_{ji}I_i^u}{N_i - I_i^r} \quad [4]$$

$$N_i = N_i + \theta\sum_j M_{ij} - \theta\sum_j M_{ji} \quad [5]$$

where S_i, E_i, I_i^r, I_i^u and N_i are the susceptible, exposed, documented infected, undocumented infected and total population in city i. Note that we define patients with symptoms severe enough to be confirmed as documented infected individuals; whereas other infected persons are defined as undocumented infected individuals. We provide a rate parameter, β, for the transmission rate due to documented infected individuals. The transmission rate due to undocumented individuals is reduced by a factor μ. In addition, α is the fraction of documented infections, Z is the average latency period and D is the average duration of infection. The effective reproduction number (R_e) is calculated as $R_e = \alpha\beta D + (1-\alpha)\mu\beta D$ (see Section 6 below for details). Spatial coupling within the model is represented by the daily number of people traveling from city j to city i (M_{ij}) and a multiplicative factor, θ, which is greater than 1 to reflect underreporting of human movement. We assume that individuals in the I_i^r group do not move between cities, though these individuals can move between cities during the latency period. A similar metapopulation model has been used to forecast the spatial transmission of influenza in the United States (*20*).

Covid-19 Lockdown models

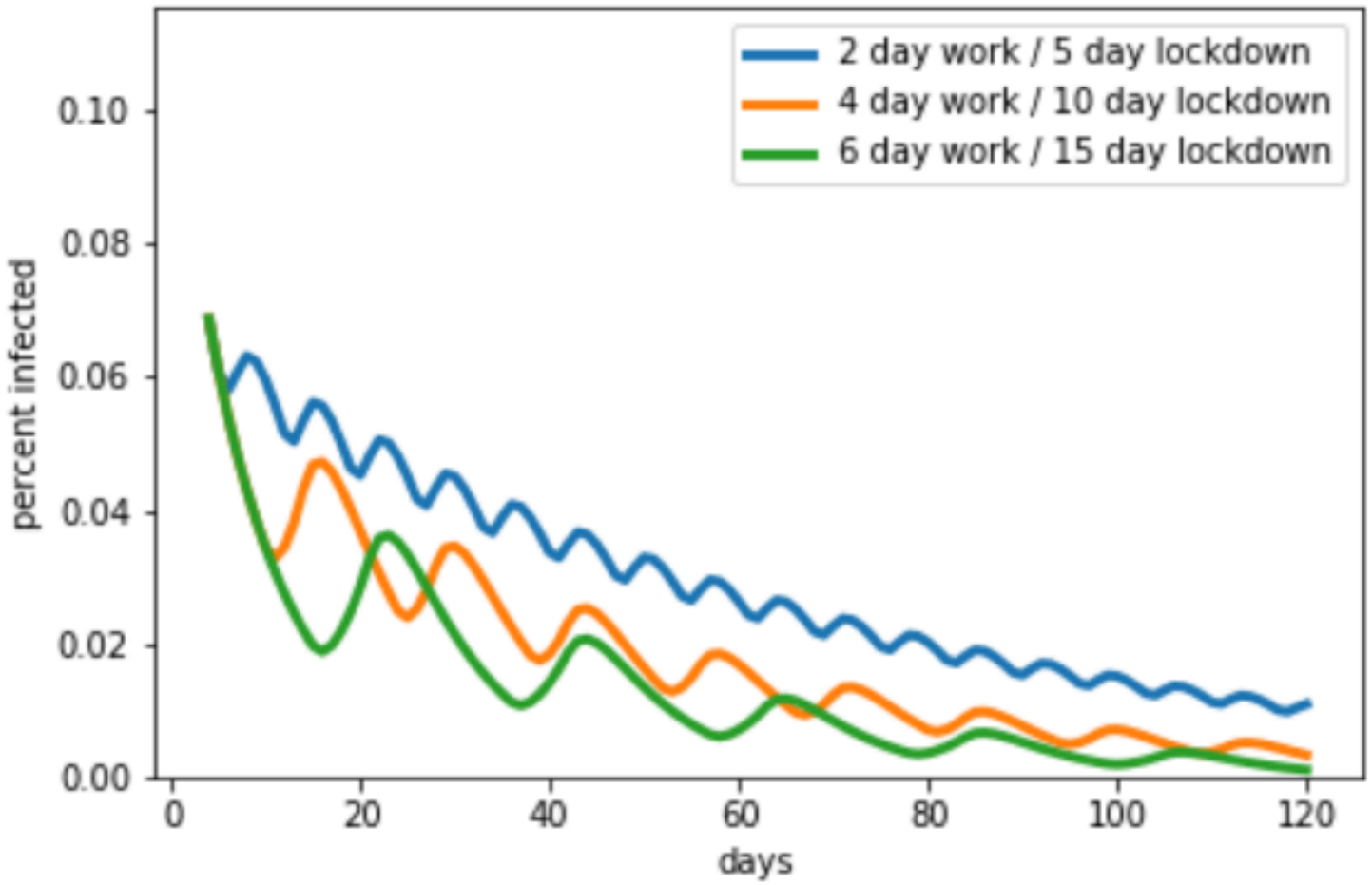


Figure 4. Infection is controlled for various schedules also in a more detailed simulation, a SEIR model calibrated for COVID19 [4]. In this simulation, the virus has a mean 5 day incubation period and 3 day infectious period. Longer schedules, such as 4-day work/ 10-day lockdown, show more rapid infection decline, because they allow expose individuals to cease becoming infectious before returning to work. Code for producing this figure is in: https://github.com/omerka-weizmann/2_day_workweek.

Barzel et al; Nagler (2020)

Karin, Korem, Dudovich, Alon, Medium & arXiv (2020)

Covid-19 Modelling — Fundamental principles

https://www.3blue1brown.com/videos-blog/simulating-an-epidemic

Machine learning for Covid-19 detection

EPFL

Home Instructions About Us

COUGHVID

Send us a recording of a cough sound and help research on COVID-19

Safe coughing instructions

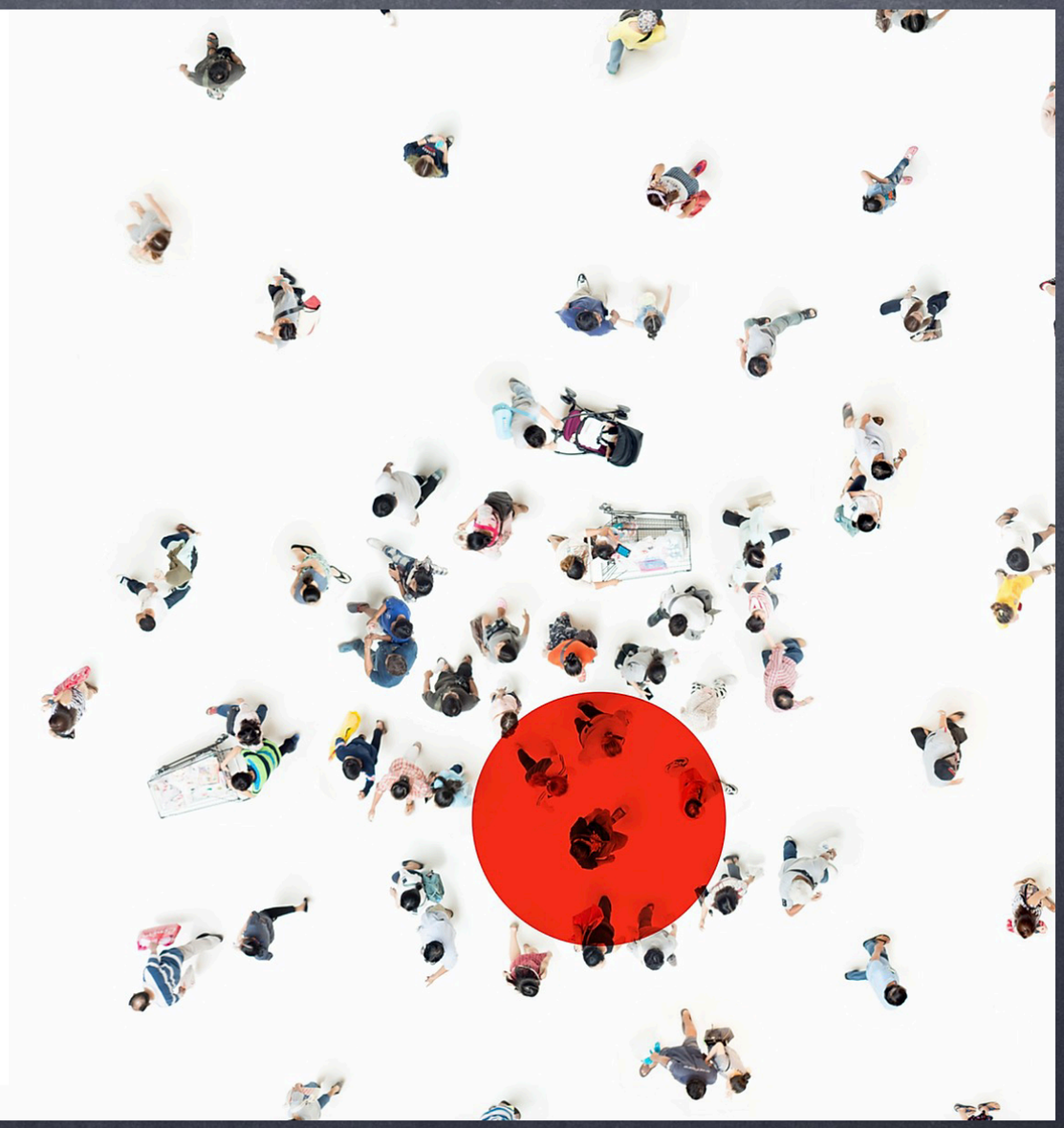
Record

EPFL, April 2020: https://coughvid.epfl.ch

Covid-19 Mitigation + Lockdown + Privacy

Pan-European Privacy-Preserving Proximity Tracing

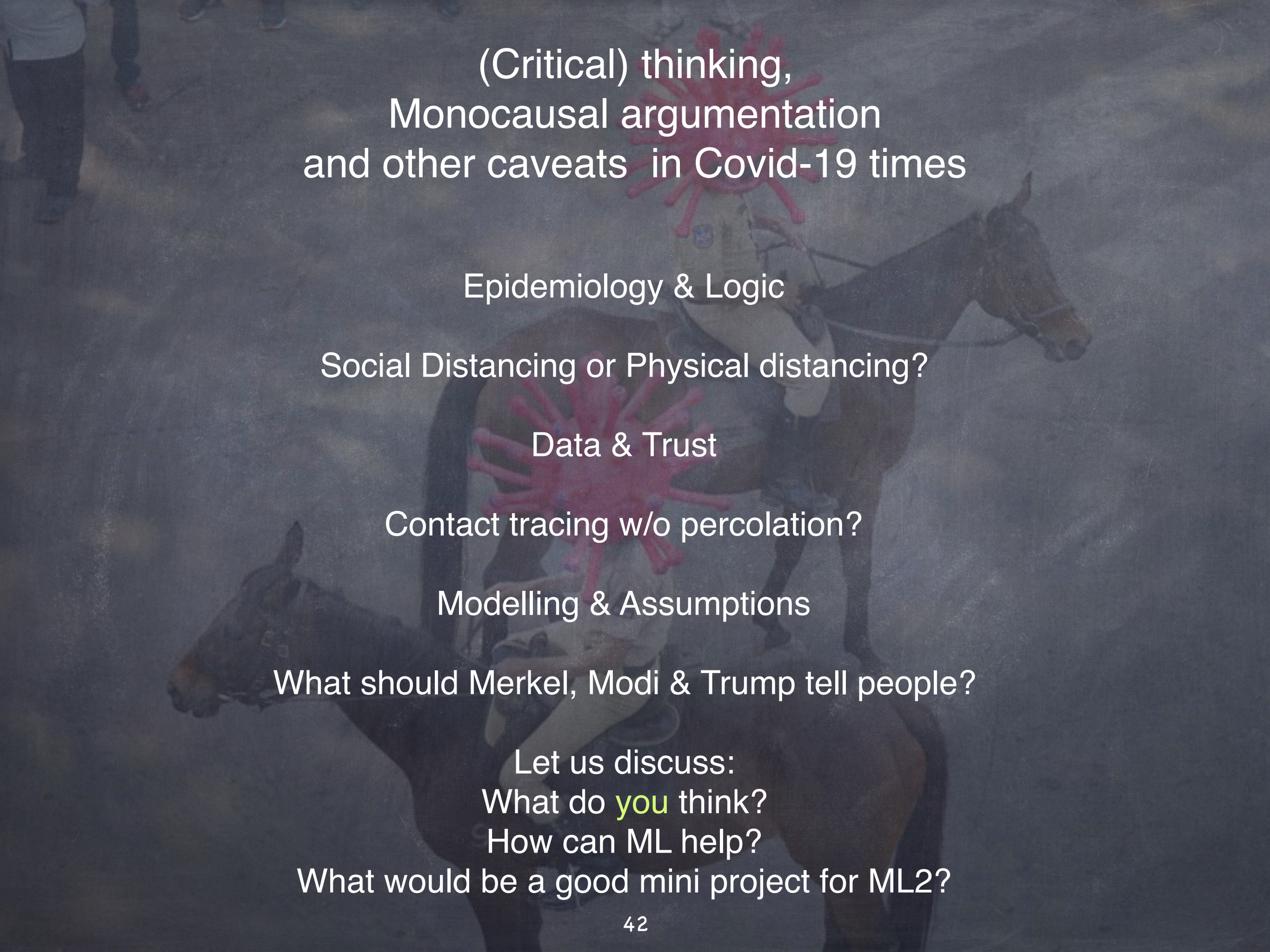
Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) makes it possible to interrupt new chains of SARS-CoV-2 transmission rapidly and effectively by informing potentially exposed people. **We are** a large and inclusive European team. **We provide** standards, technology, and services to countries and developers. **We embrace** a fully privacy-preserving approach. **We build on** well-tested, fully implemented proximity measurement and scalable backend service. **We enable** tracing of infection chains across national borders.



https://www.pepp-pt.org

Covid-19 Private automatic contact tracing

Time slot for students



(Critical) thinking, Monocausal argumentation and other caveats in Covid-19 times

Epidemiology & Logic

Social Distancing or Physical distancing?

Data & Trust

Contact tracing w/o percolation?

Modelling & Assumptions

What should Merkel, Modi & Trump tell people?

Let us discuss:
What do you think?
How can ML help?
What would be a good mini project for ML2?

Covid-19 Mini Projects (DRAFT)

Mortality rate estimation (Böttcher et al., 2020)

It’s all about representation: Clever data collapse

Variance, covariance of cases, deaths

Prediction of Cases and deaths over time

In Data we trust? Bendford’s law for data manipulation