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

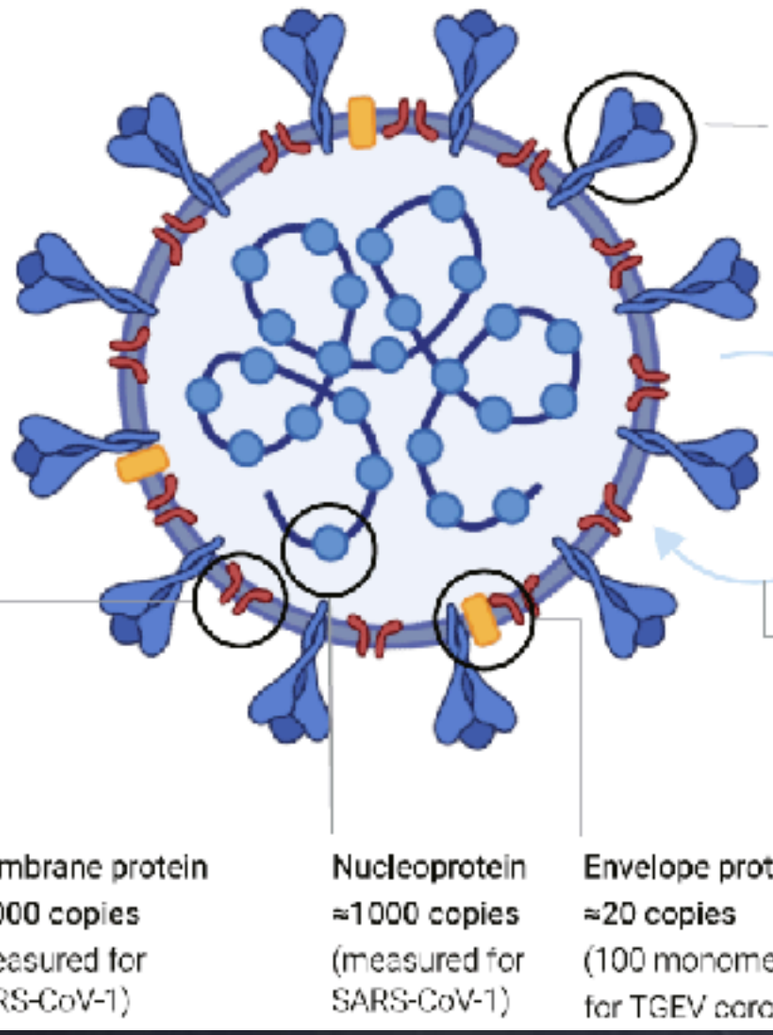
Modelling

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

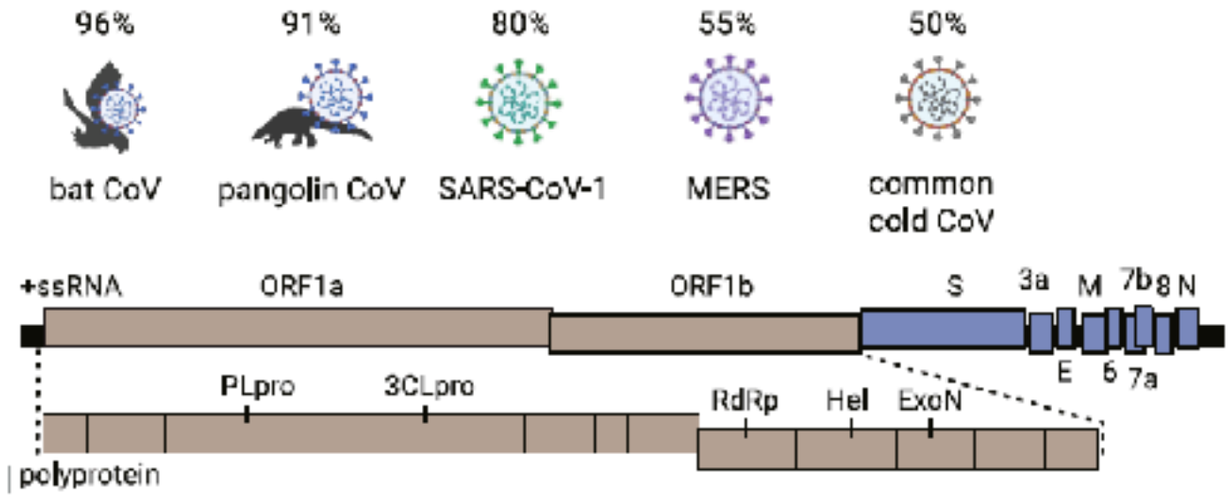
Membrane protein
≈2000 copies
(measured for
SARS-CoV-1)

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

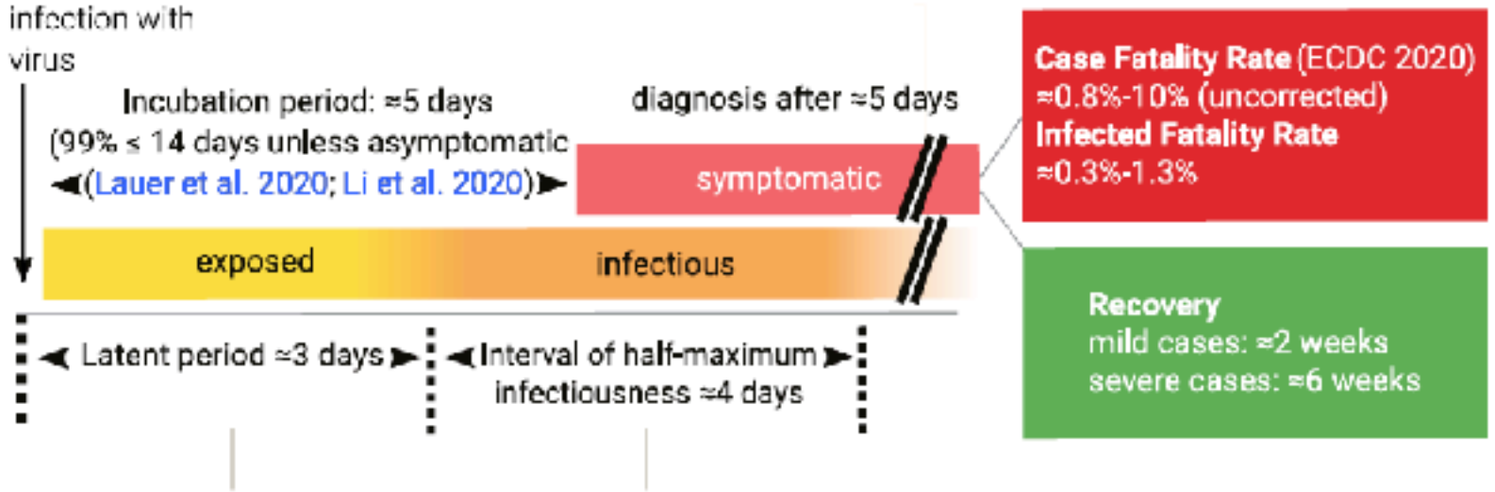
	half life	time to decay 1000-fold
Aerosols:	≈1 hr	≈4-24 hr
Surfaces: e.g. plastic, cardboard and metals	≈1-7 hr	≈4-96 hr

(van Doremalen et al. 2020)

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)

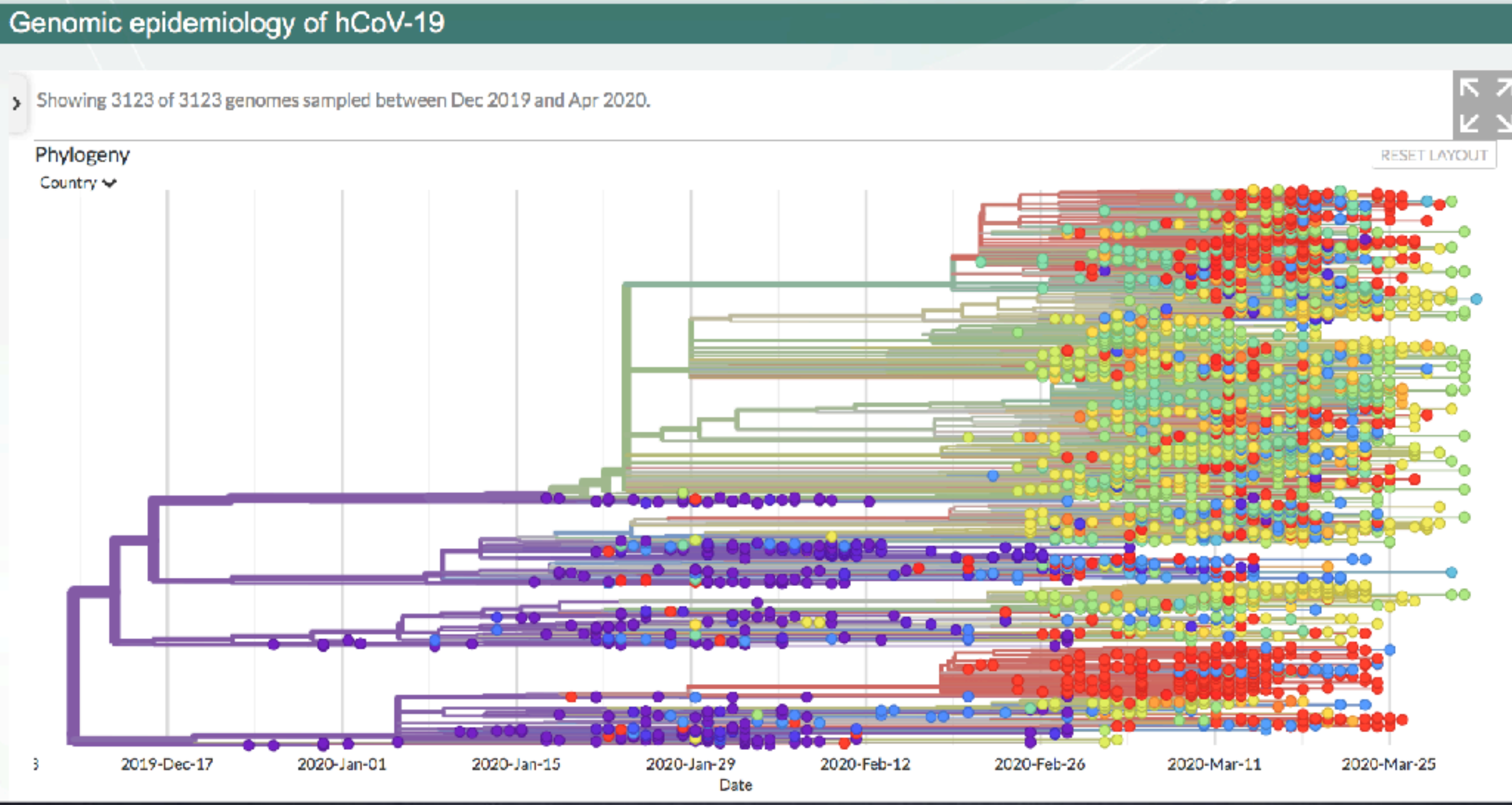


Inter-individual variability is substantial and not well characterized. The estimates are parameter fits for population median in China and do not describe this variability (Li et al. 2020; He et al. 2020).

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.

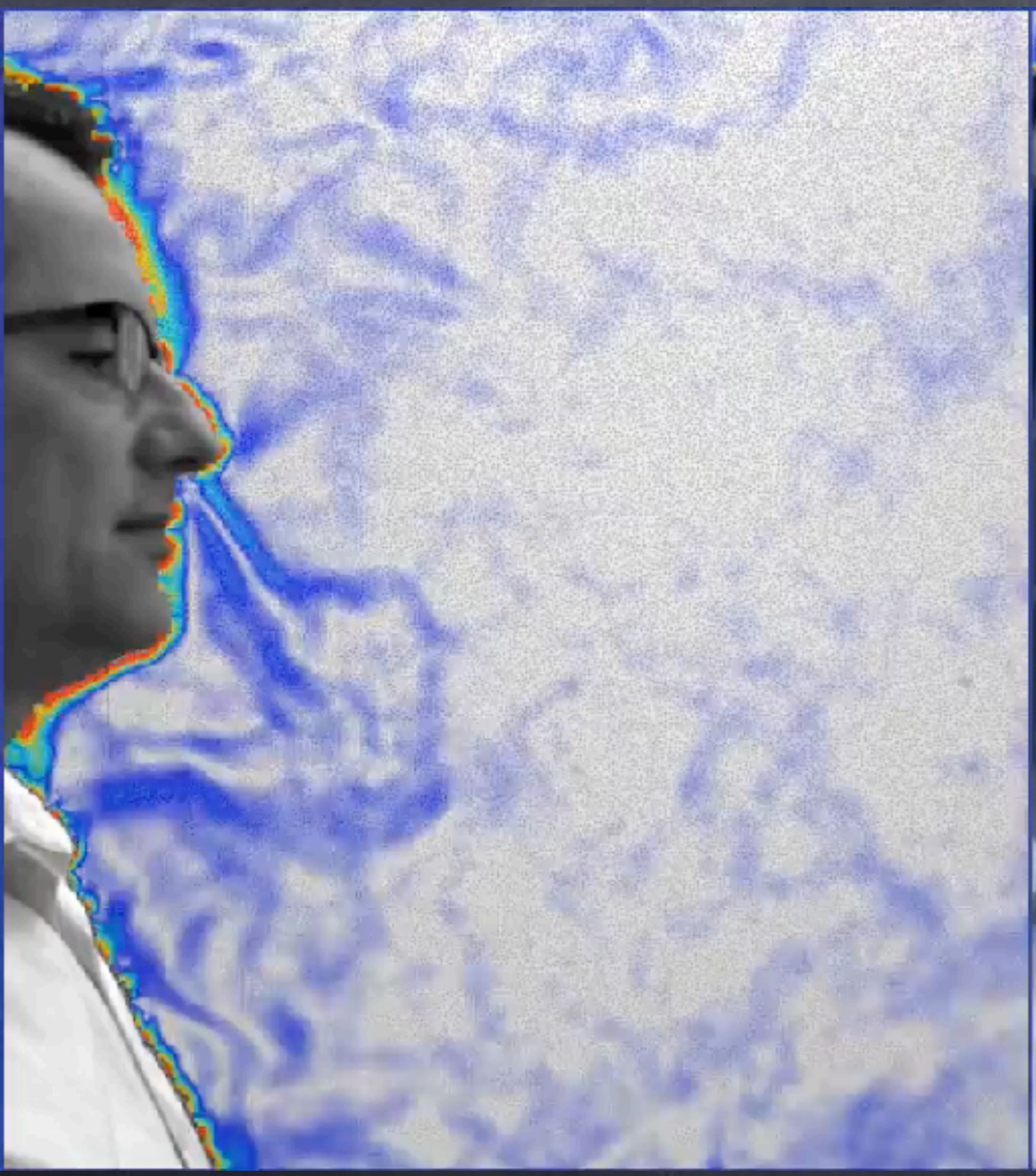
Sars-Covid-19 by the numbers, eLife, 2020

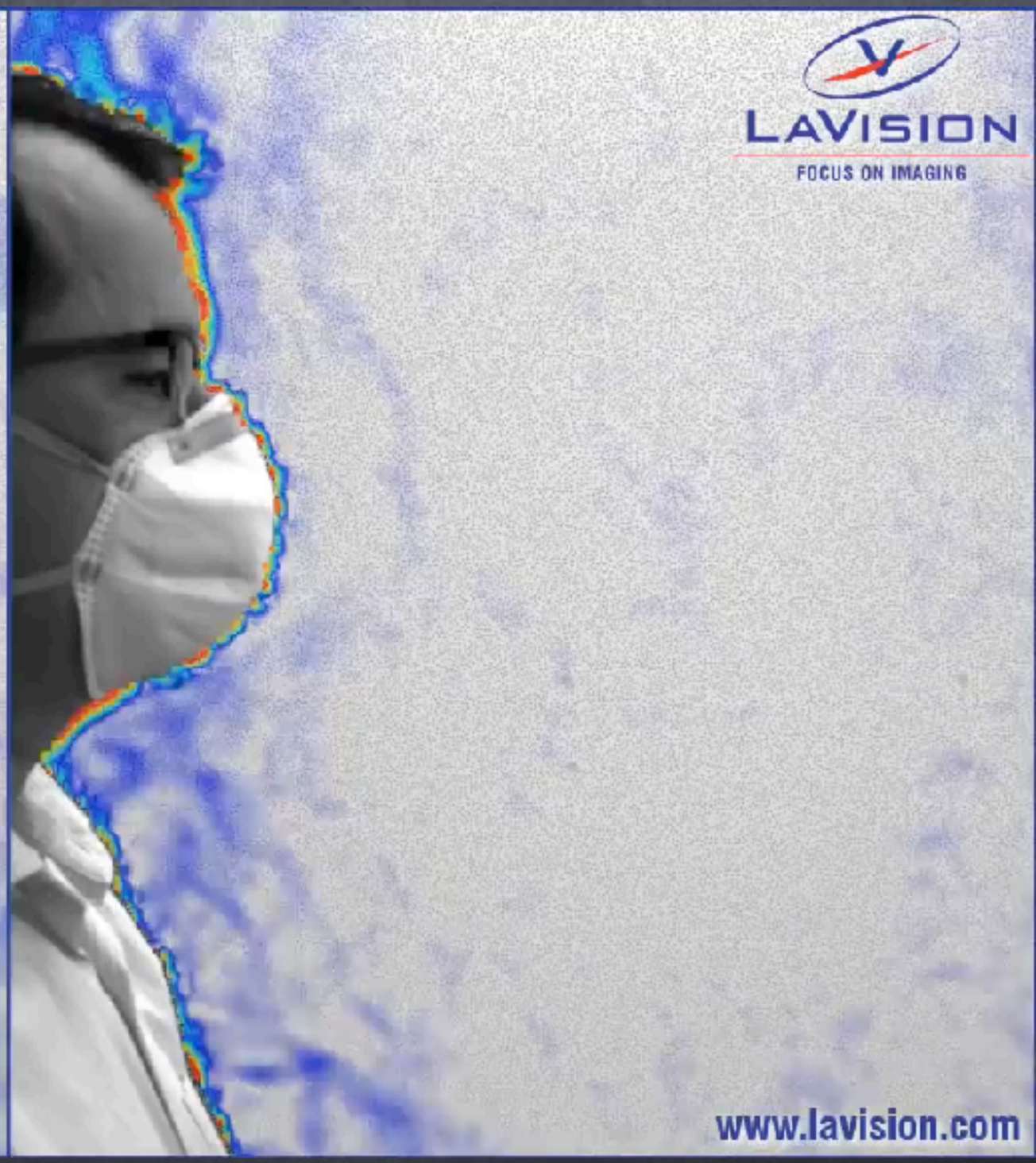
Biology of Covid-19
Phylogeny (Mutations)



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

Physics of Covid-19

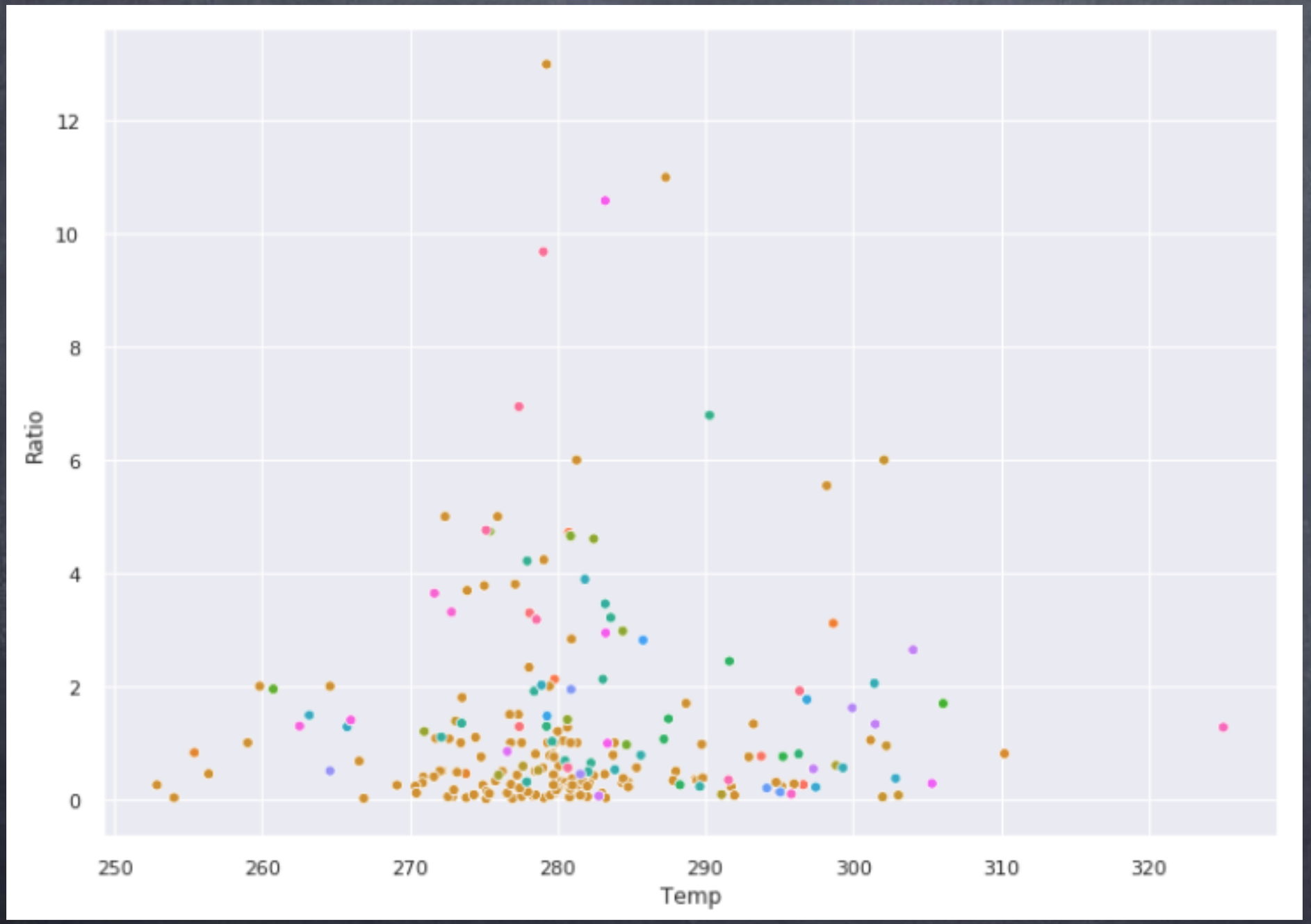




Runners + Cyclists ?

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

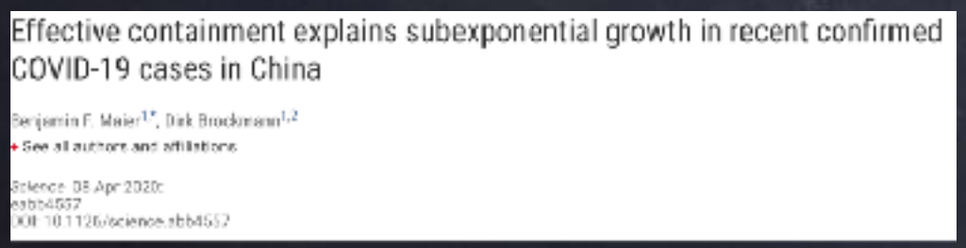
https://www.worldometers.info/coronavirus/worldwide-graphs/

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

Mitigation

What mitigation can achieve and has achieved

April 14 Covid-19 Science Review

(Mitigation 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”

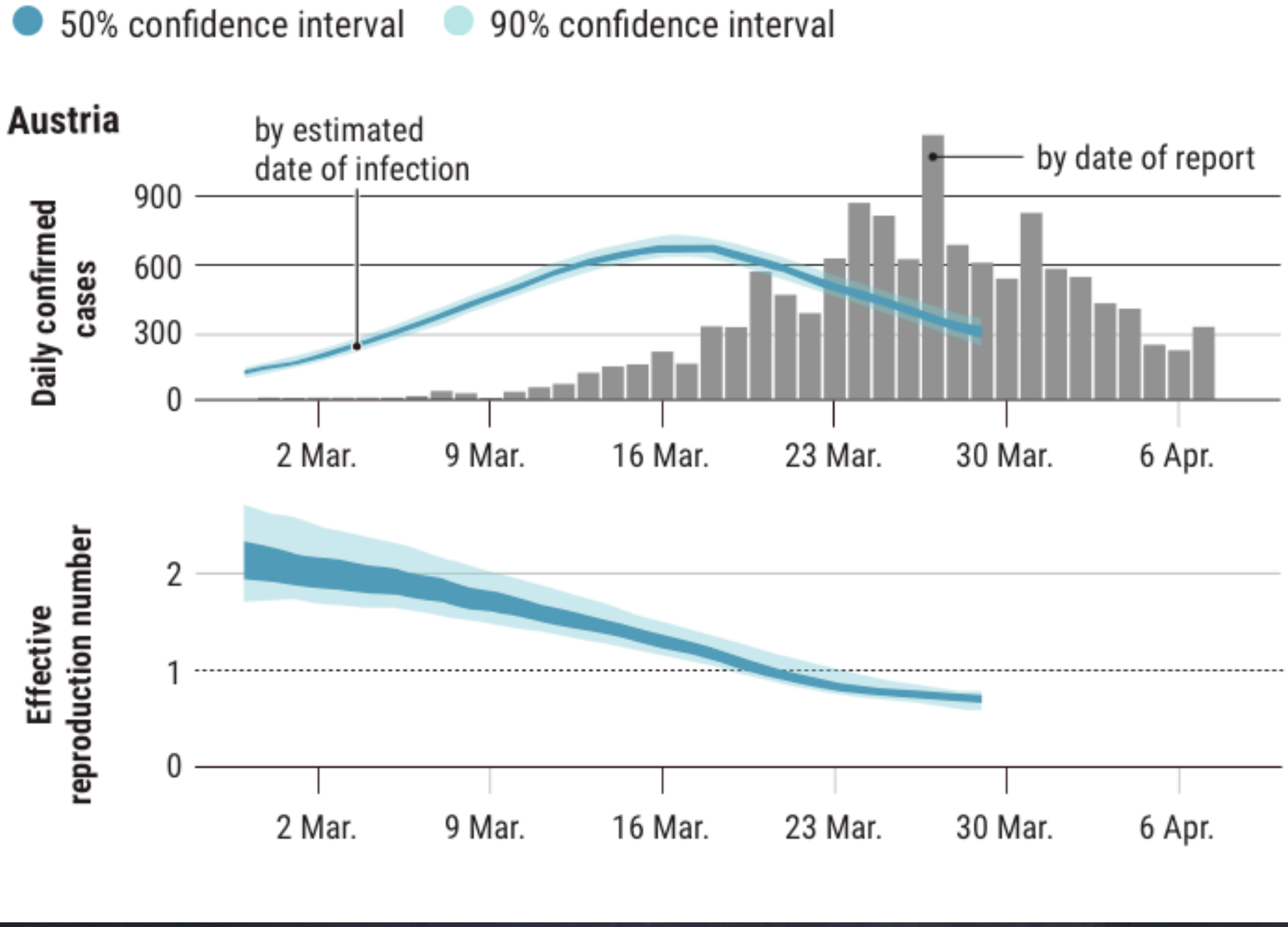
Effective reproduction number R

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

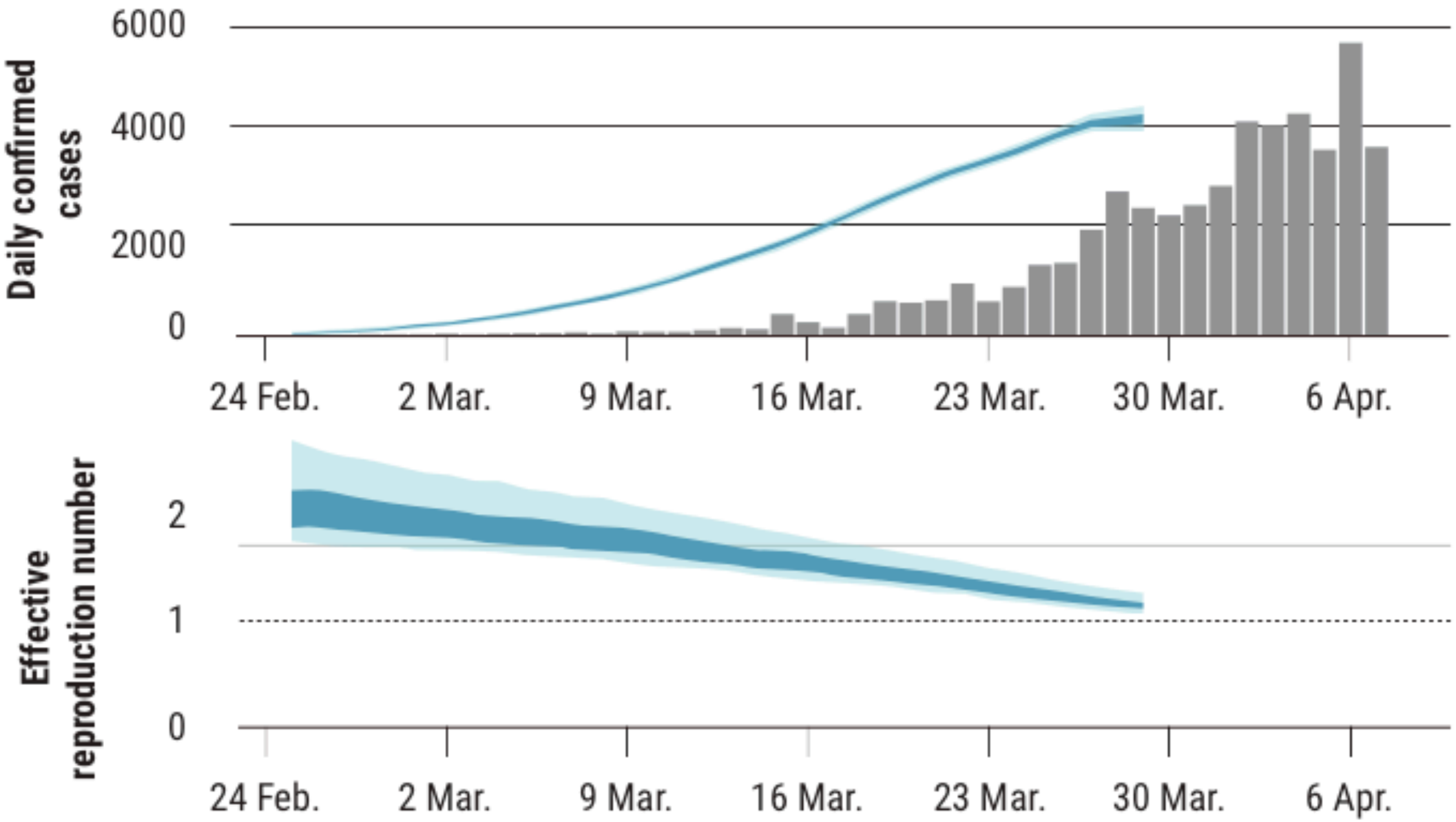
If the reproduction number is smaller than 1, the disease dies out;
$R < 1$ if it is larger than 1, there is exponential spreading $R > 1$

The number to watch

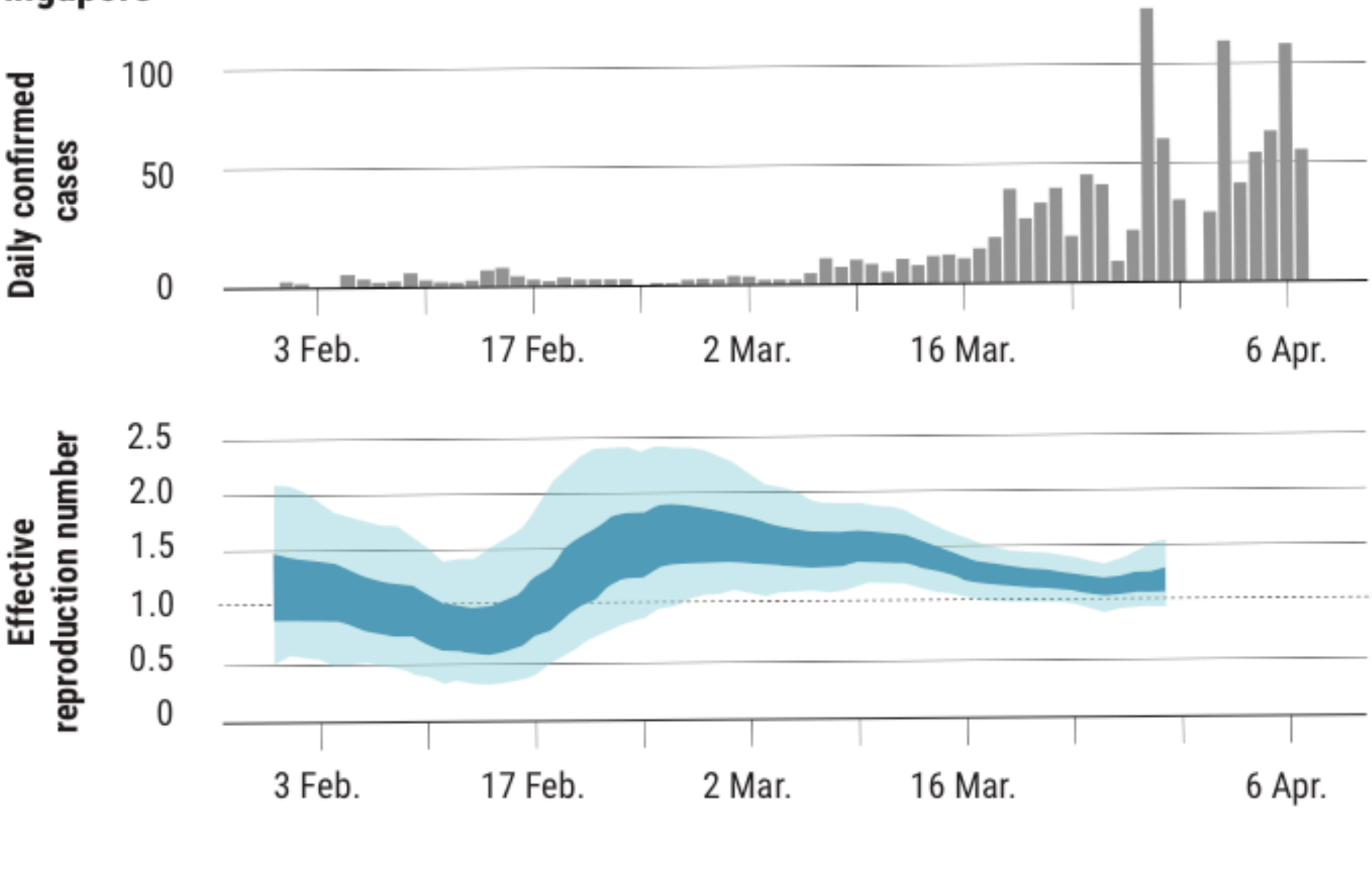
Lockdowns lower the number of new cases as well as R, the effective reproduction number. If R drops below 1, the epidemic shrinks.



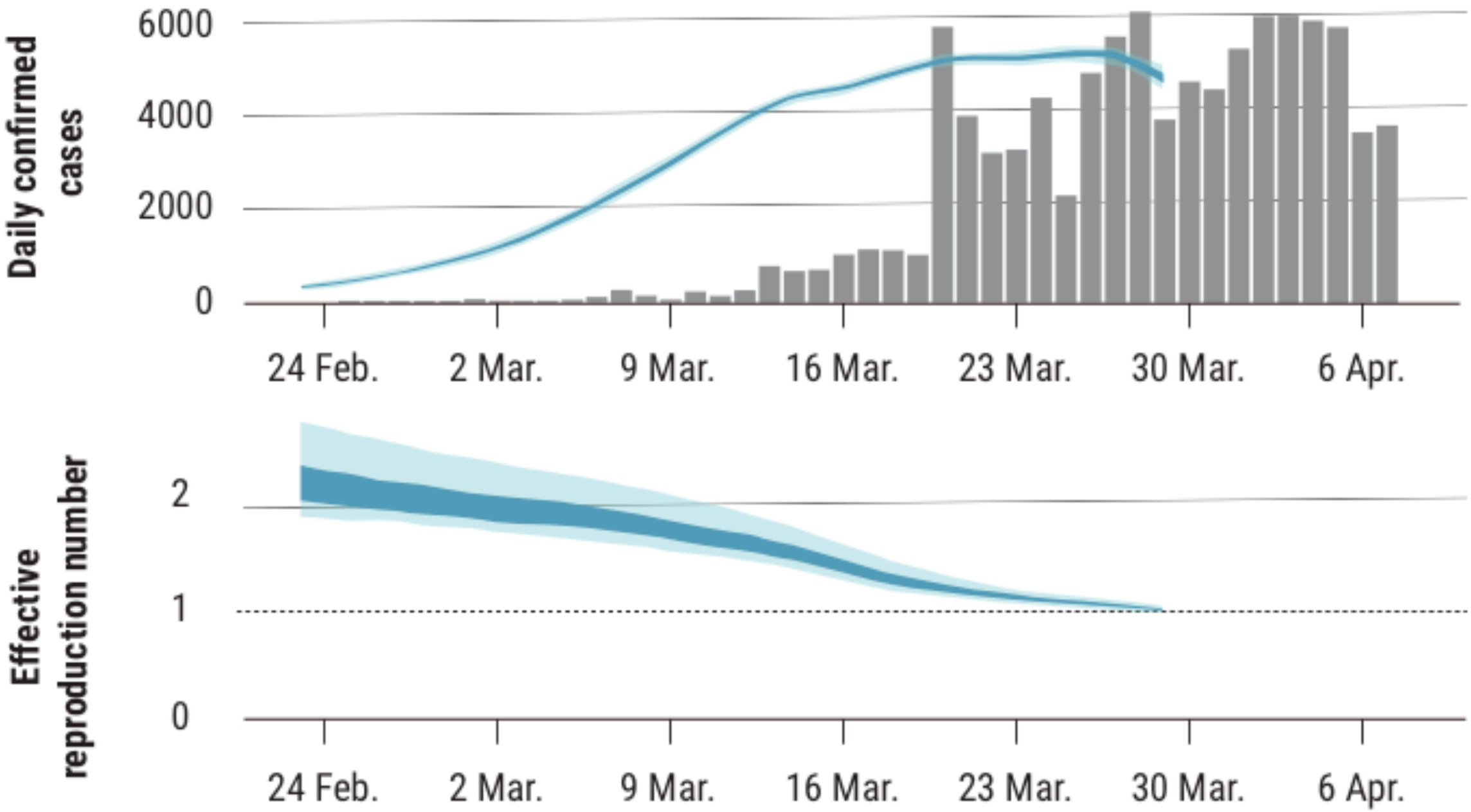
United Kingdom



Singapore

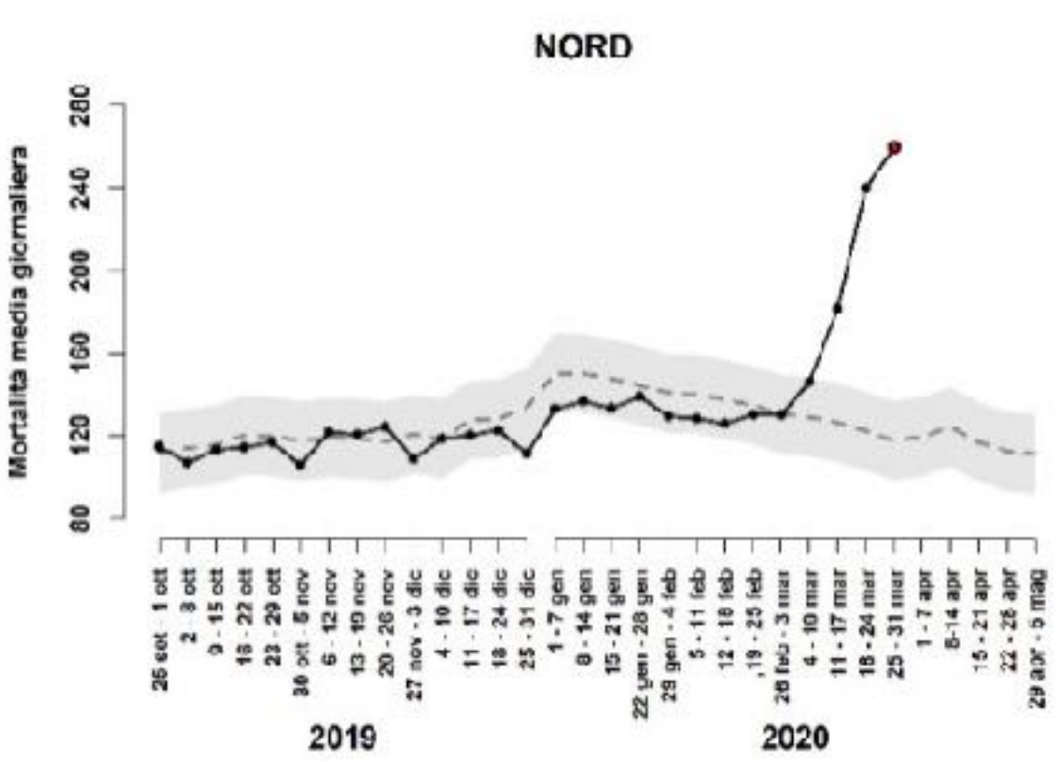

Daily confirmed cases
100
50
0
3 Feb.
17 Feb.
2 Mar.
16 Mar.
6 Apr.
Effective reproduction number
2.5
2.0
1.5
1.0
0.5
0
3 Feb.
17 Feb.
2 Mar.
16 Mar.
6 Apr.

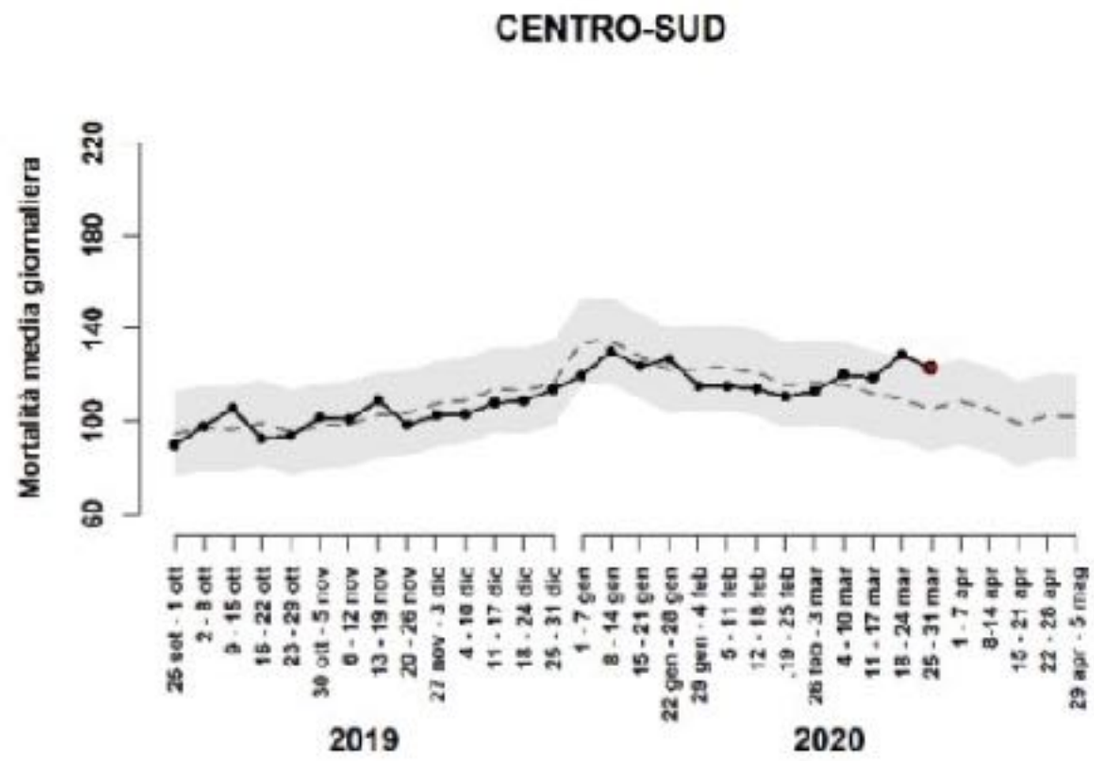
Germany


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.

Covid-19 data representation / analysis

Geographic heterogeneity of Covid-19 deaths in Italy





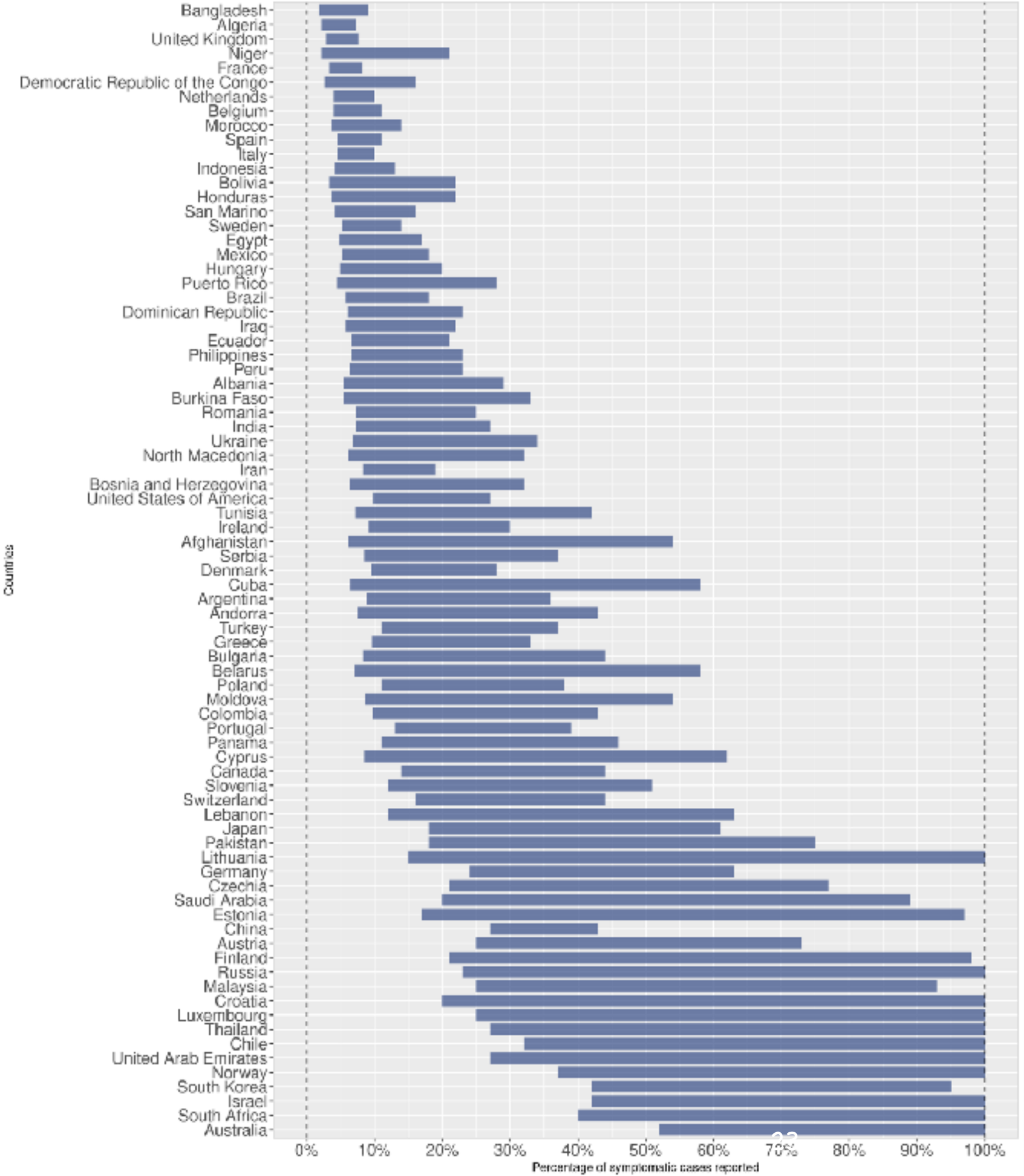
Numbers aggregated by a country can be very misleading
and mask the severity of covid-19 in heavily affected communities

Bias in Covid19

Percentage of reported (symptomatic) Covid-19 cases

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

Depending on the country, even likely up to x20 more cases than reported



SARS 2002/2003 data / analysis

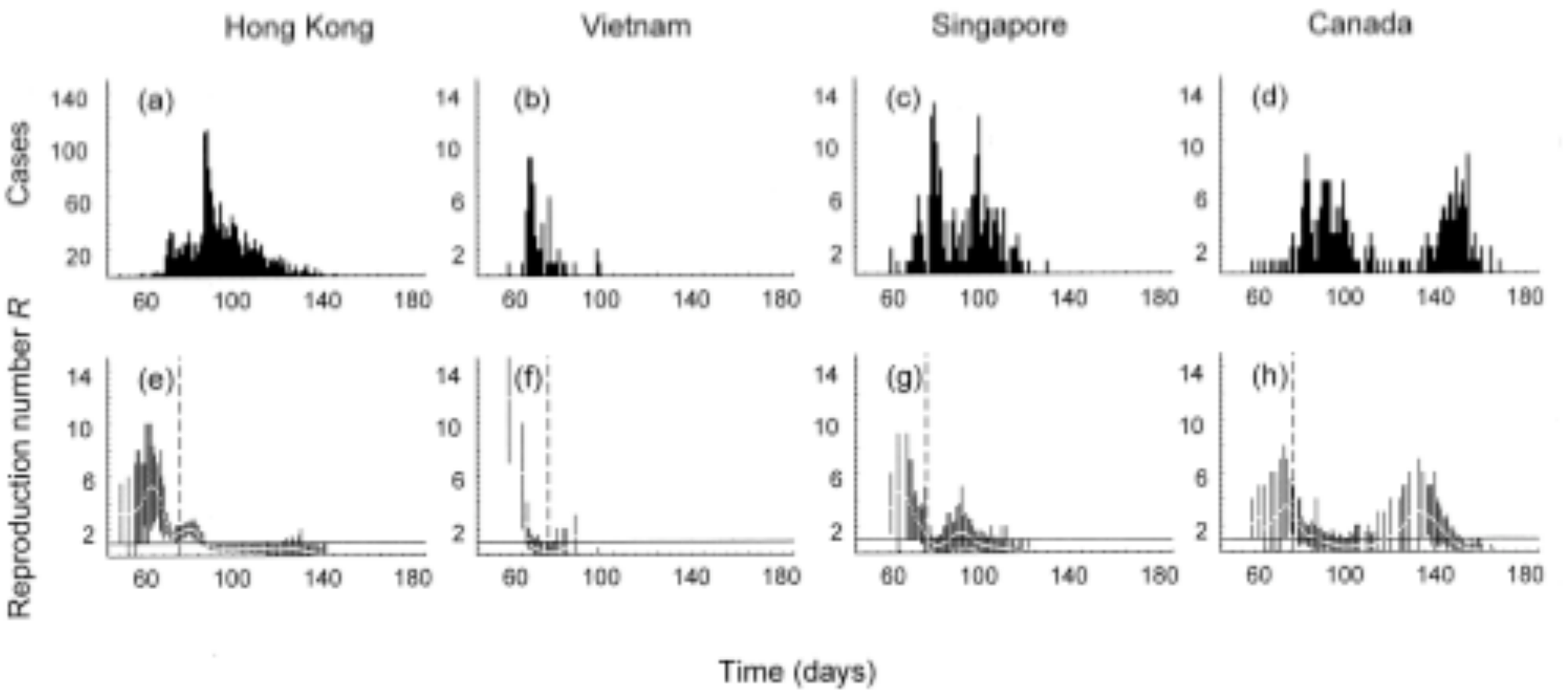


FIGURE 1. Epidemic curves (numbers of cases by date of symptom onset) for severe acute respiratory syndrome (SARS) outbreaks in *a*) Hong Kong, *b*) Vietnam, *c*) Singapore, and *d*) Canada and the corresponding effective reproduction numbers (*R*) (numbers of secondary infections generated per case, by date of symptom onset) for *e*) Hong Kong, *f*) Vietnam, *g*) Singapore, and *h*) Canada, 2003. Markers (white spaces) show mean values; accompanying vertical lines show 95% confidence intervals. The vertical dashed line indicates the issuance of the first global alert against SARS on March 12, 2003; the horizontal solid line indicates the threshold value $R = 1$, above which an epidemic will spread and below which the epidemic is controlled. Days are counted from January 1, 2003, onwards.

Canada developed a clear 2nd peak, basic reproduction number from simple fit (source: WHO)

Hidden geometry of epidemic spreading from aviation data

Based on the passenger flux fraction from airport n to m

$$P_{mn}$$

Brockmann & Helbing define totally in an ad hoc fashion the

Effective distance between airports

$$d_{mn} = [1 - \log(P_{mn})]$$

Together with mild assumptions they arrive at the conclusion

Given fixed values for epidemic parameters, our analysis shows that network and flux information are sufficient to predict the dynamics and arrival times.

Brockmann & Helbing, The Hidden Geometry of Complex, Network-Driven Contagion Phenomena, Science Vol. 342, Issue 6164, pp. 1337-1342 (2013)

Hidden geometry of epidemic spreading from aviation data

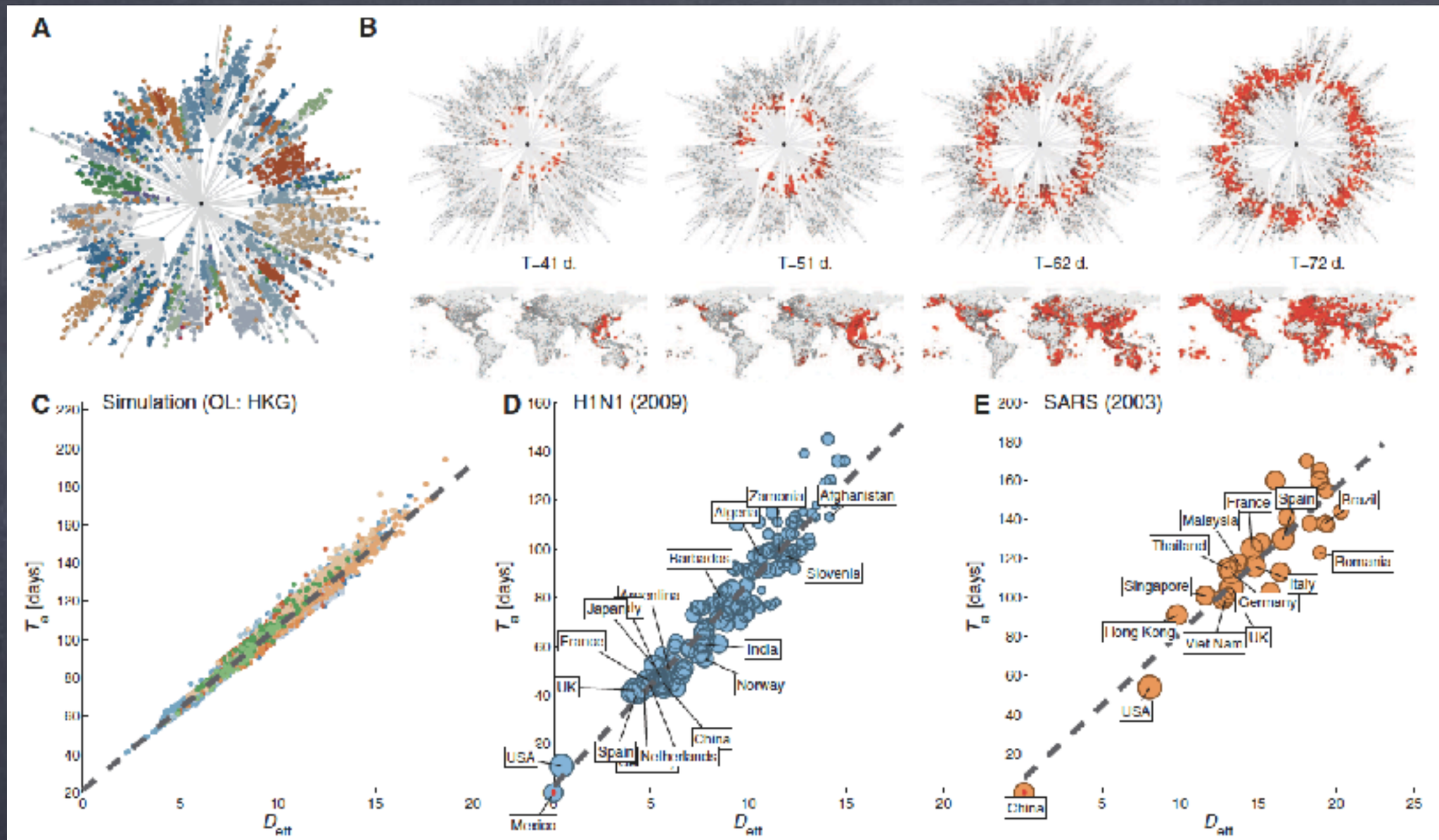


Fig. 2. Understanding global contagion phenomena using effective distance. (**A**) The structure of the shortest path tree (in gray) from Hong Kong (central node). Radial distance represents effective distance D_{eff} as defined by Eqs. 4 and 5. Nodes are colored according to the same scheme as in Fig. 1A. (**B**) The sequence (from left to right) of panels depicts the time course of a simulated model disease with initial outbreak in Hong Kong (HKG), for the same parameter set as used in Fig. 1B. Prevalence is reflected by the redness of the symbols. Each panel compares the state of the system in the conventional geographic representation (bottom) with the effective distance representation (top). The complex spatial pattern in the conventional view is equivalent to a homogeneous wave that propagates outwards at constant effective speed in the effective distance representation. (**C**) Epidemic arrival time T_a versus effective distance D_{eff} for the same simulated epidemic as in (B). In contrast to geographic distance (Fig. 1C), effective distance correlates strongly with arrival time ($R^2 = 0.973$), i.e., effective distance is an excellent predictor of arrival times. (**D** and **E**) Linear relationship between effective distance and arrival time for the 2009 H1N1 pandemic (D) and the 2003 SARS epidemic (E). The arrival time data are the same as in Fig. 1, D and E. The effective distance was computed from the projected global mobility network between countries. As in the model system, we observe a strong correlation between arrival time and effective distance.

Brockmann & Helbing, The Hidden Geometry of Complex, Network-Driven Contagion Phenomena, Science Vol. 342, Issue 6164, pp. 1337-1342 (2013)

Hidden geometry of epidemic spreading from aviation data

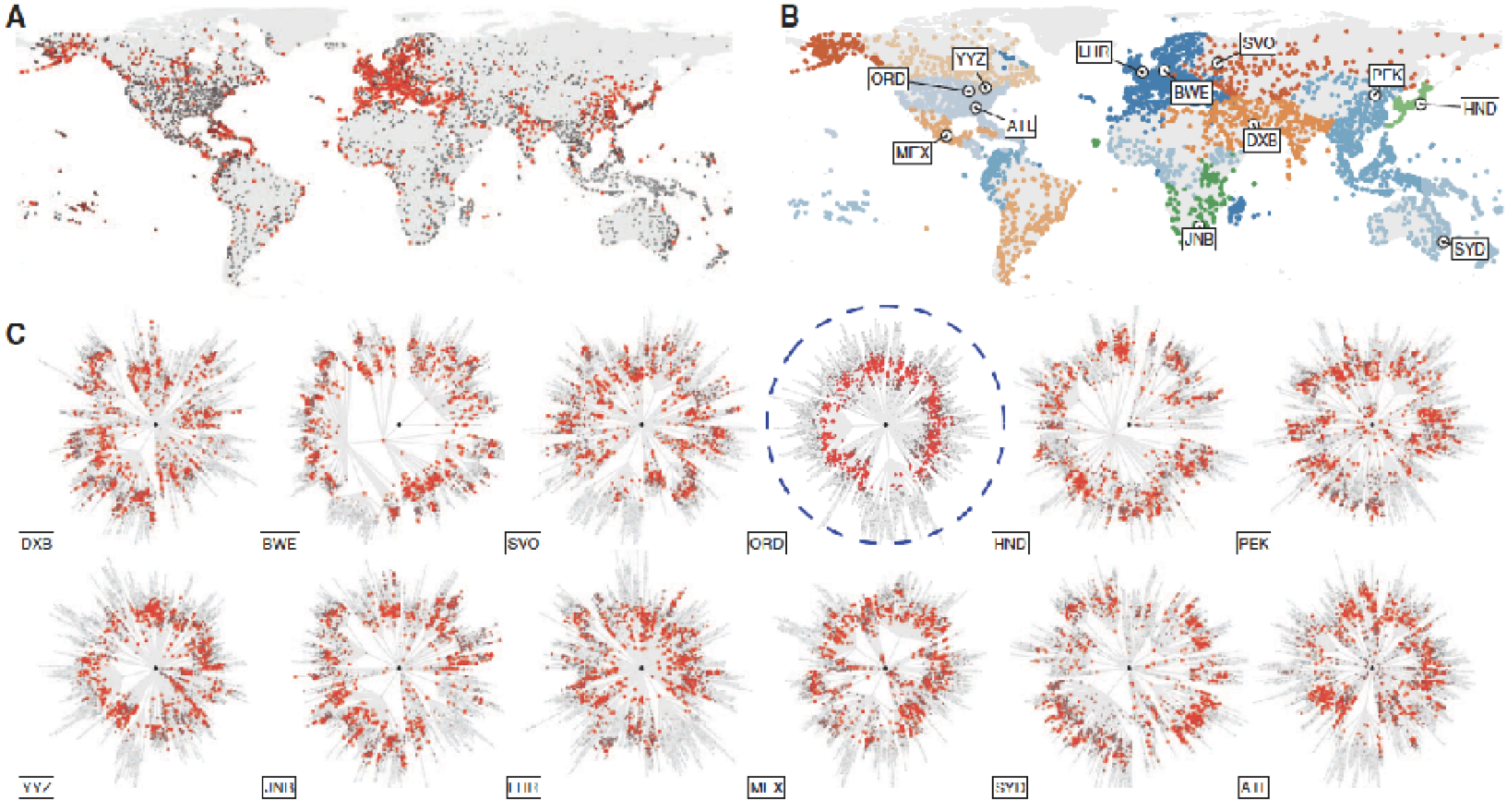


Fig. 3. Qualitative outbreak reconstruction based on effective distance. (**A**) Spatial distribution of prevalence $j_n(t)$ at time $T = 81$ days for OL Chicago (parameters $\beta = 0.28$ day^{-1}, $R_0 = 1.9$, $\gamma = 2.8 \times 10^{-3}$ day^{-1}, and $\varepsilon = 10^{-6}$). After this time, it is difficult, if not impossible, to determine the correct OL from snapshots of the dynamics. (**B**) Candidate OLs chosen from different geographic regions. (**C**) Panels depict the state of the system shown in (**A**) from the perspective of each candidate OL, using each OL's shortest path tree representation. Only the actual OL (ORD, circled in blue) produces a circular wavefront. Even for comparable North American airports [Atlanta (ATL), Toronto (YYZ), and Mexico City (MEX)], the wavefronts are not nearly as concentric. Effective distances thus permit the extraction of the correct OL, based on information on the mobility network and a single snapshot of the dynamics.

Brockmann & Helbing, The Hidden Geometry of Complex, Network-Driven Contagion Phenomena, Science Vol. 342, Issue 6164, pp. 1337-1342 (2013)

Example of questionable inference, results, conclusion in covid science

Open biased paper example

But before we need to learn a key concept!

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!

In the following we study this effect more generally
and in broader context

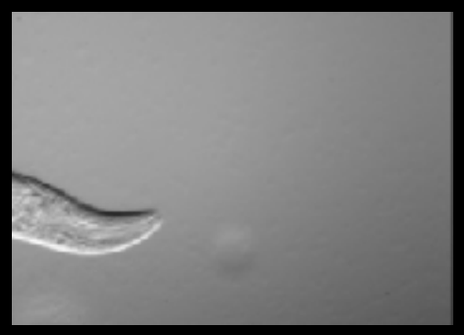
Population Growth

Change of population size

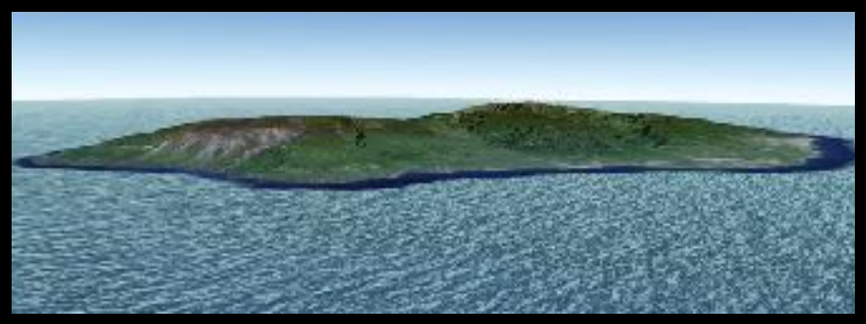
Population size

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

Growth factor
(fluctuating, coupling to other species)



Organisms



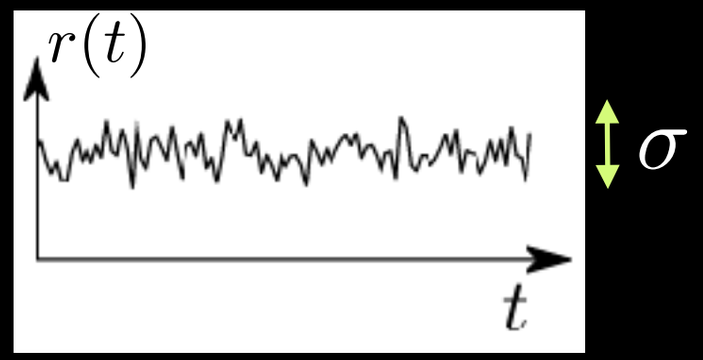
Ecosystems

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

Ergodicity breaking

iPAD presentation:

1$ example, ensemble growth

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)

Example of questionable inference, results, conclusion in covid science

Now we can open the paper!

Open biased paper example

The lockdown, general and special effects

Benedetta Cerruti Follow
Apr 14 · 4 min read

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

Covid-19 Data

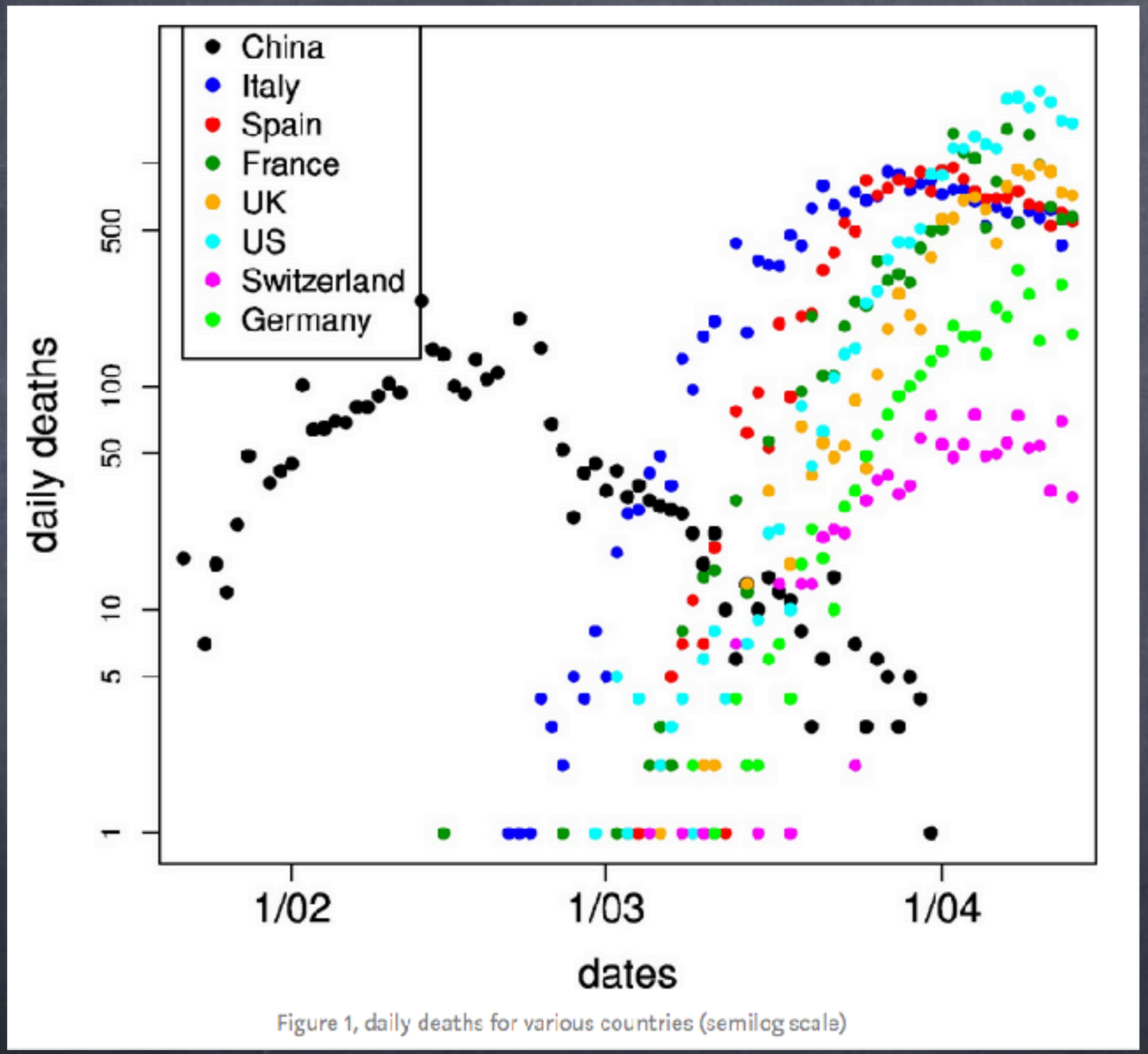


Figure 1, daily deaths for various countries (semilog scale)

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

Covid-19 Data collapse

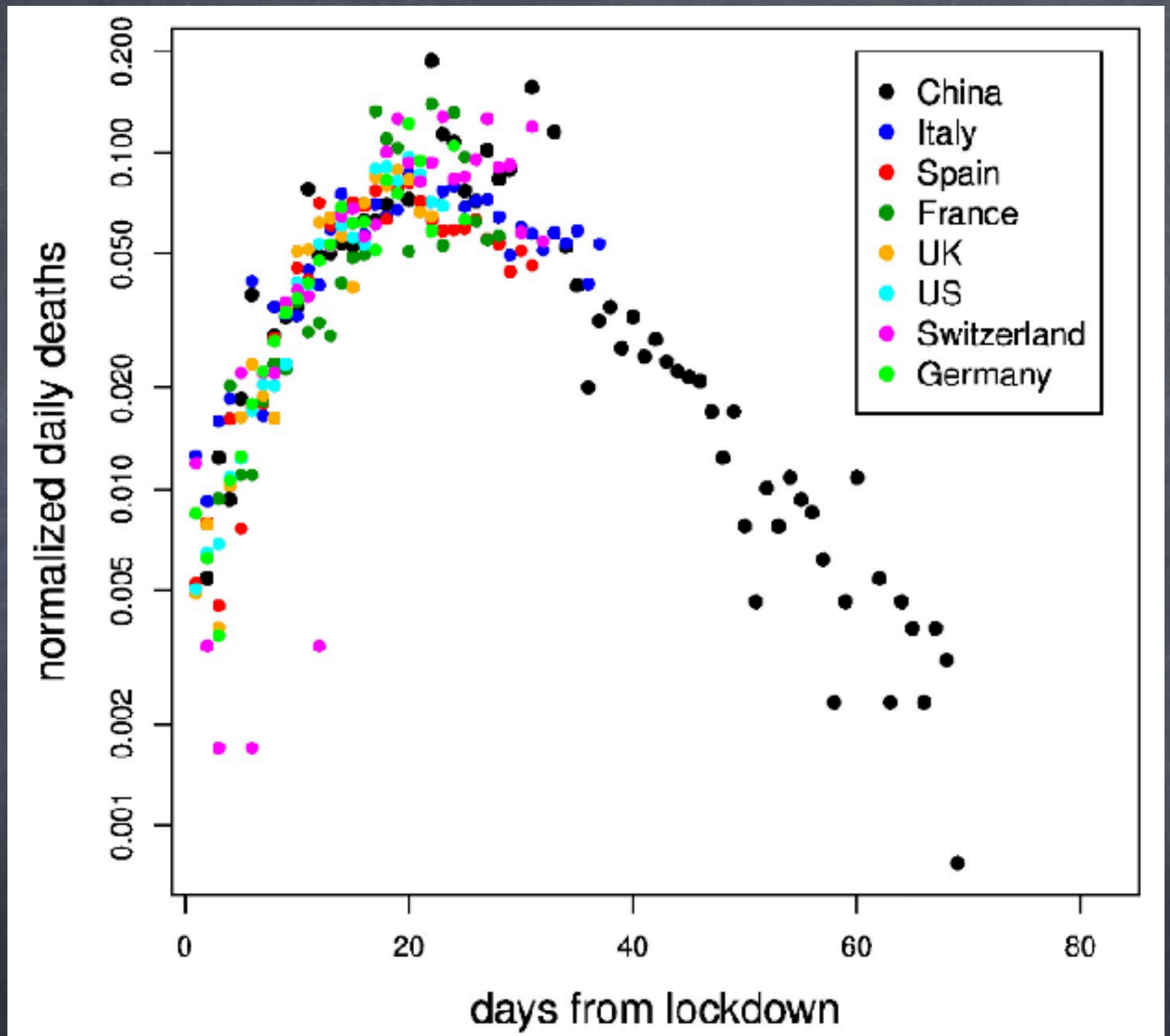
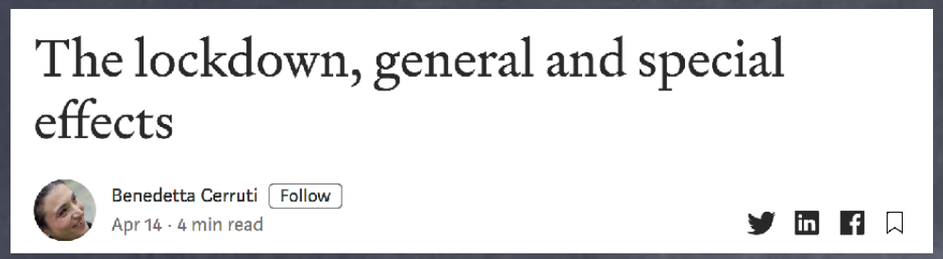


Figure 2, normalized daily deaths for various countries as a function of the time distance from lockdown (semilog scale)

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


The lockdown, general and special effects

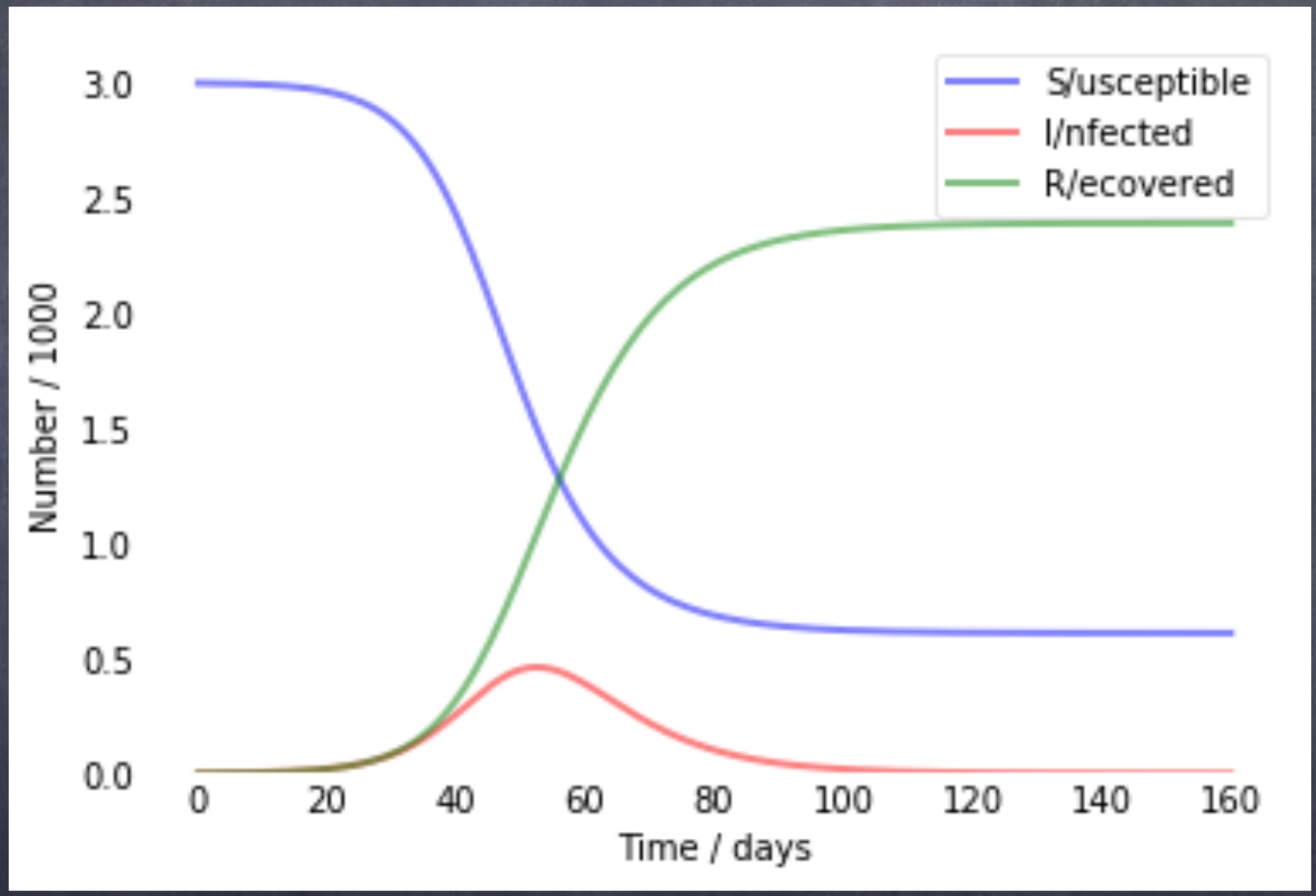
Benedetta Cerruti Follow

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 Modelling — SIR Model

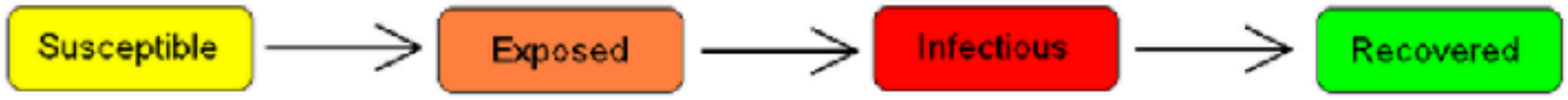


python notebook presentation

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

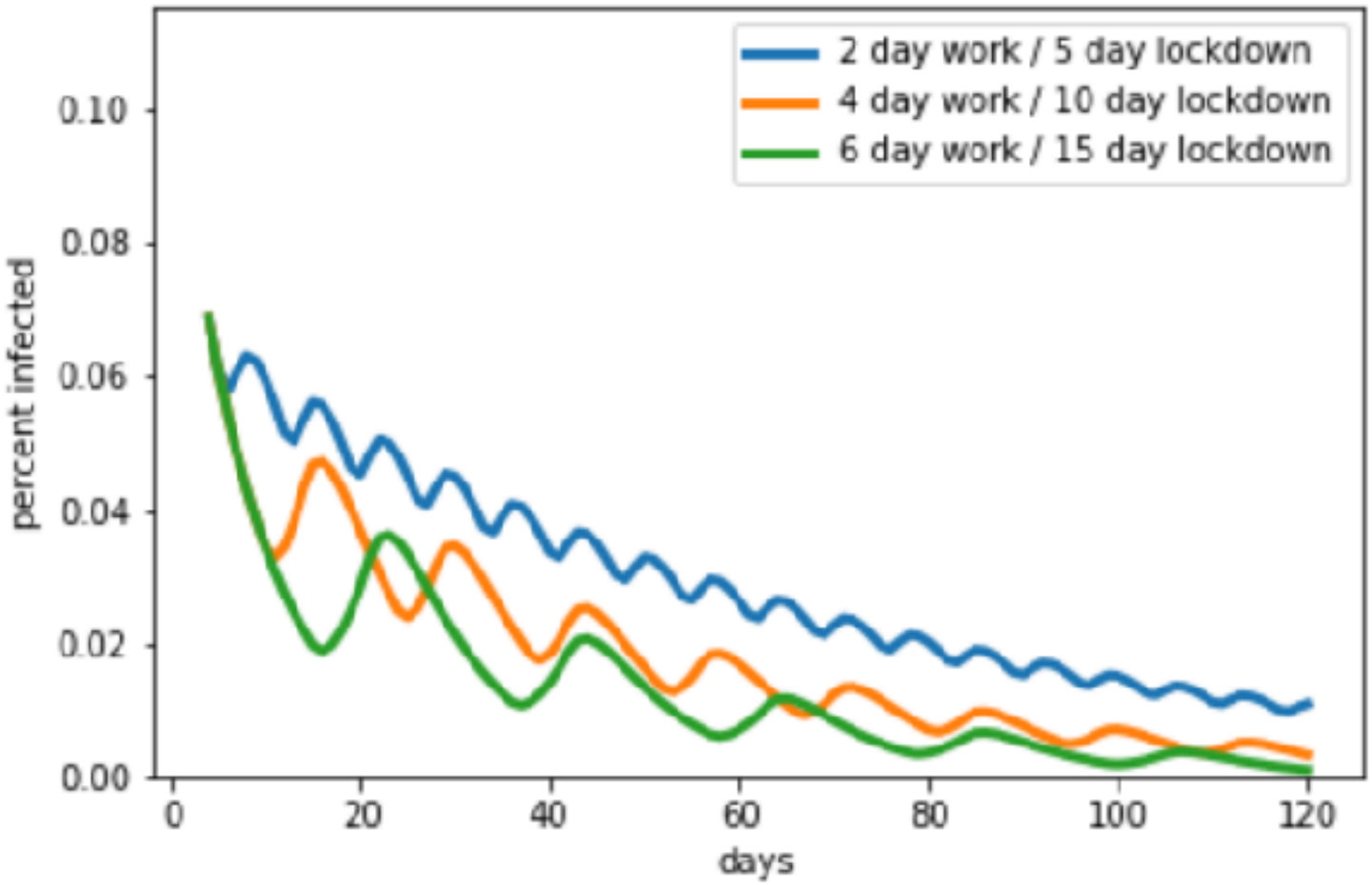


Figure 4. Infection is controlled for various schedules also in a more detailed simulation, a SEIR model calibrated for COVID19 [4]. In this simulation, the virus has a mean 5 day incubation period and 3 day infectious period. Longer schedules, such as 4-day work/ 10-day lockdown, show more rapid infection 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

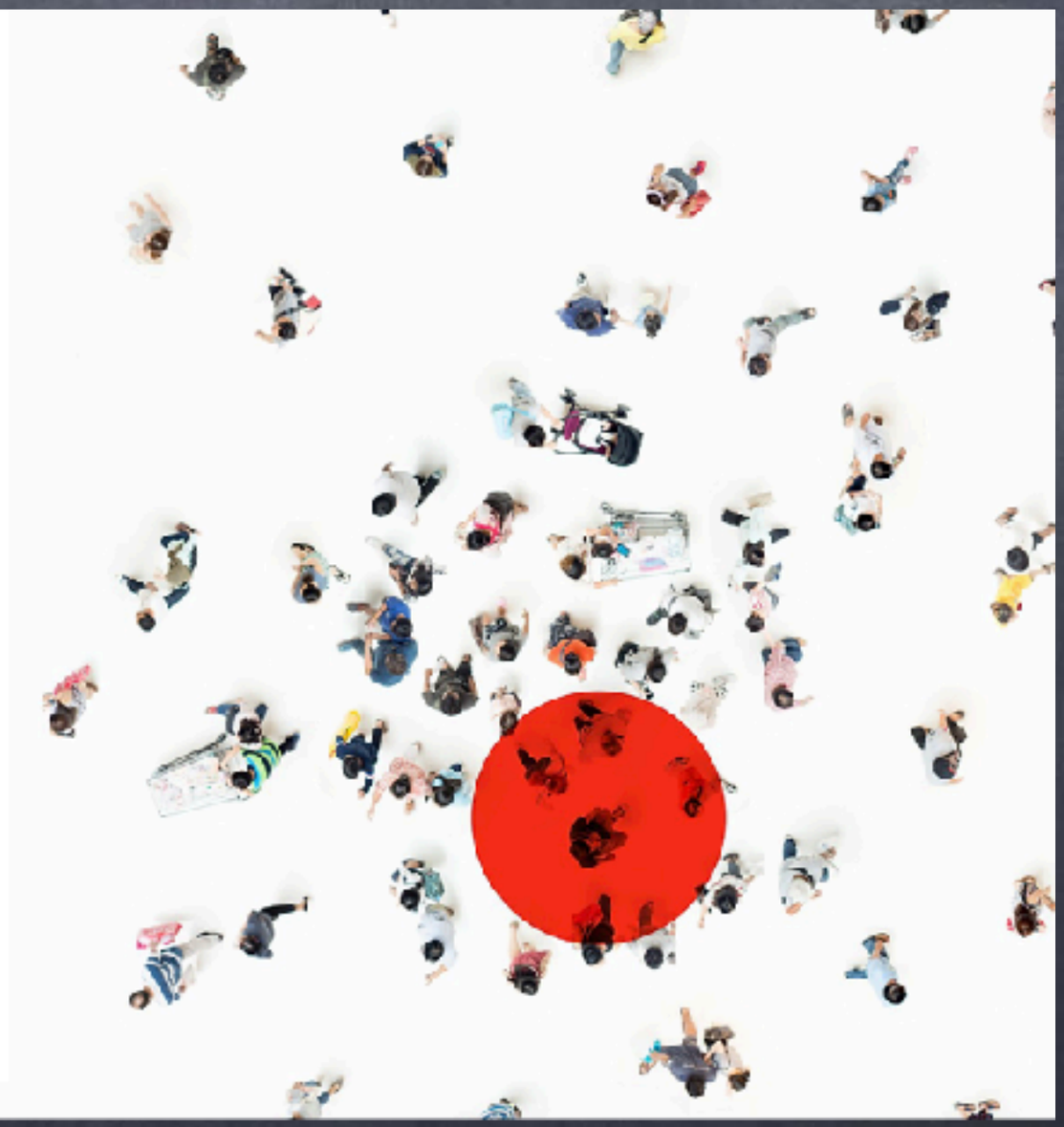
Machine learning for Covid-19 detection

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

Covid-19 Mitigation + Lockdown + Privacy

Pan-European
Privacy-Preserving
Proximity Tracing

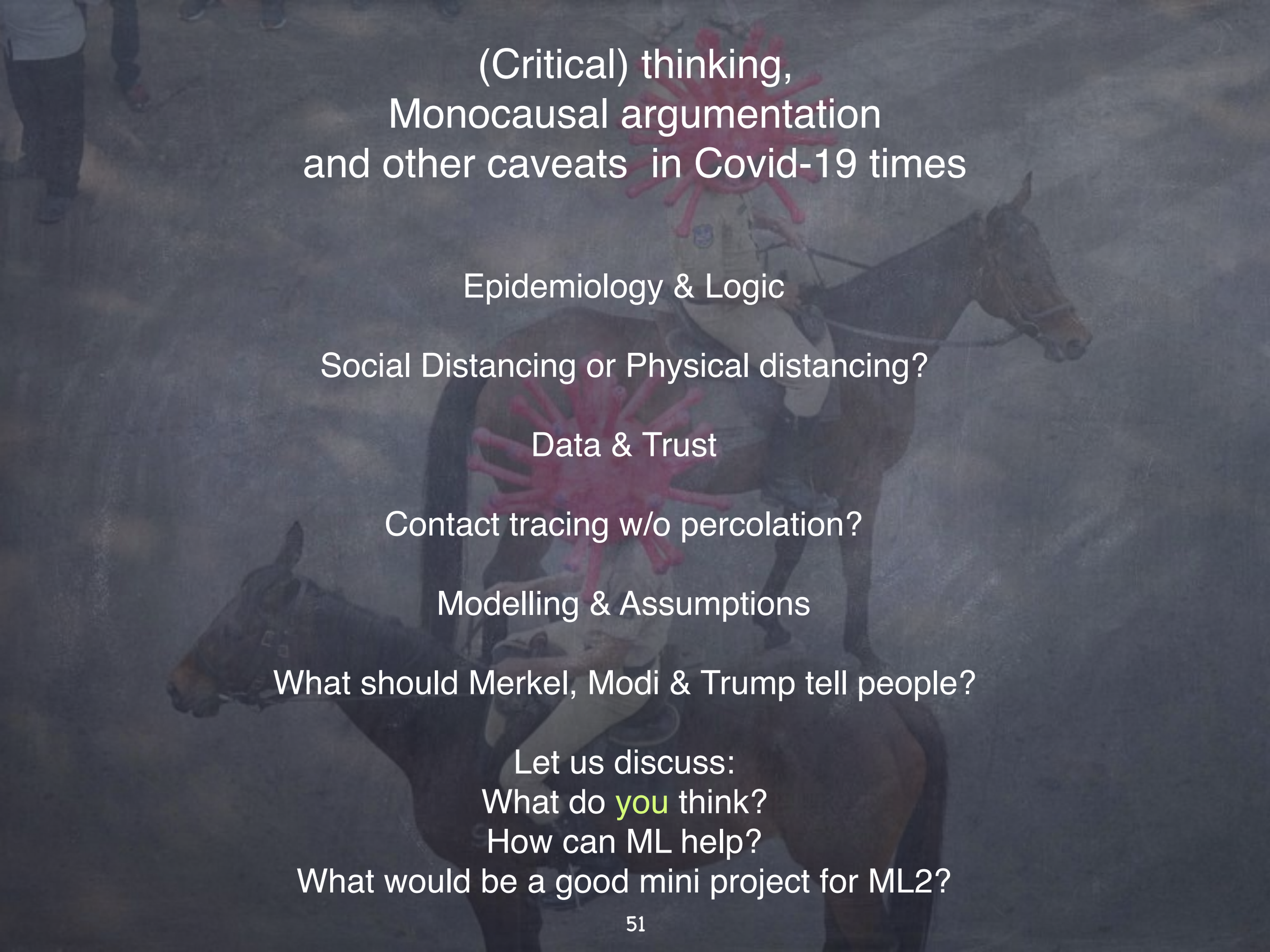
Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) makes it possible to interrupt new chains of SARS-CoV-2 transmission rapidly and effectively by informing potentially exposed people. **We are** a large and inclusive European team. **We provide** standards, technology, and services to countries 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

Mortality estimation

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

Ergodicity breaking in exponential data

insert your idea here

insert your idea here

insert your idea here

Covid-19 / Databases links

https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv

Flaxman et al, Imperial College Response Team
https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-Global-Impact-26-03-2020v2.pdf

Machine Learning II
Week #1: Part II

Jan Nagler

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

Connections

Nonlinear Correlation

Causation

Clustering

Linear Prediction

Sample size realities

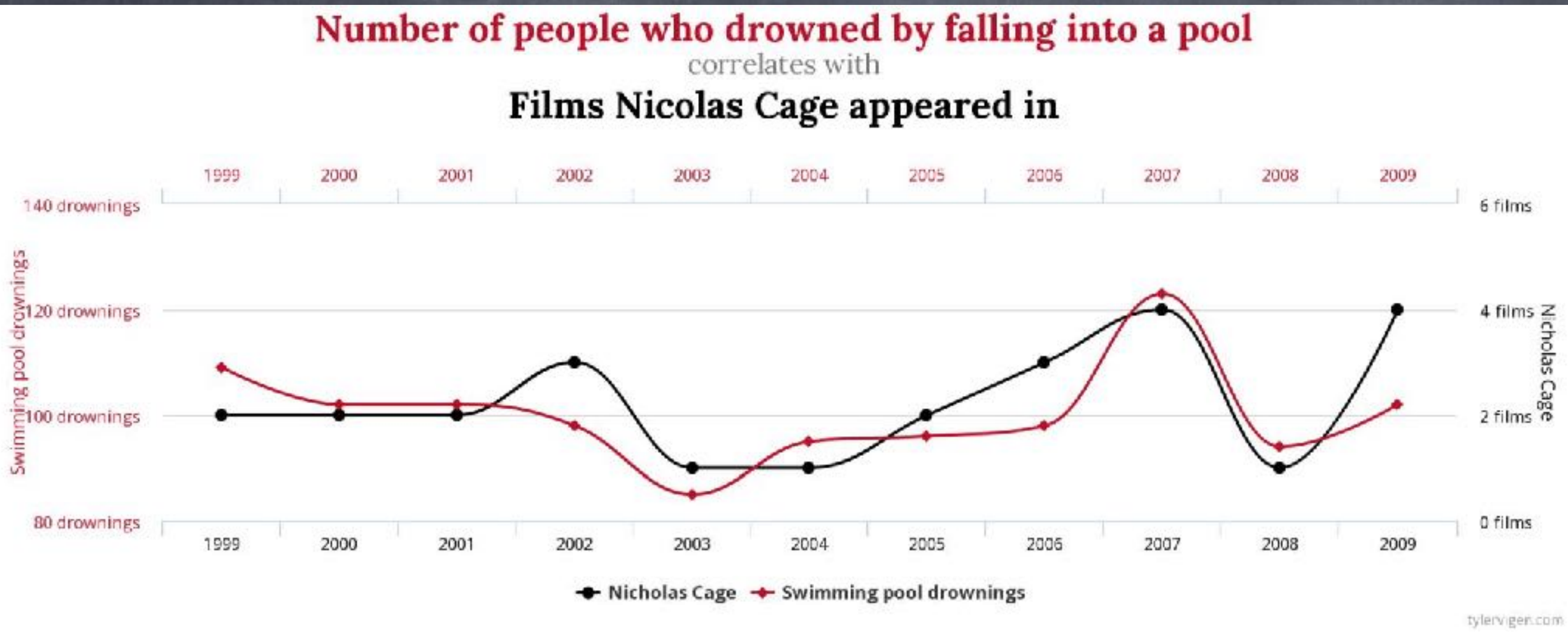
Methods
(Covariance, K-Means, PCA)

Data
(Surrogates creation)

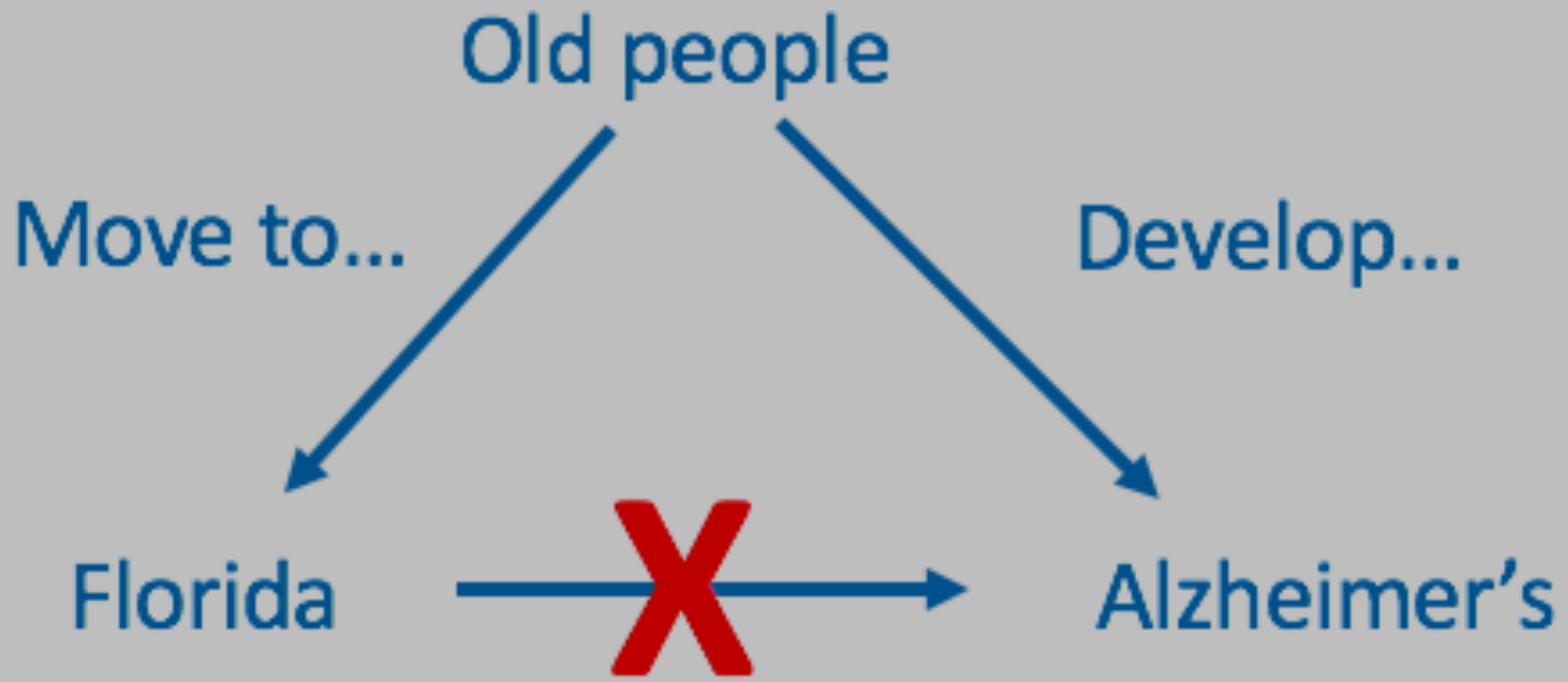
The following is to explain how covariance, correlation, clustering and PCA are connected.

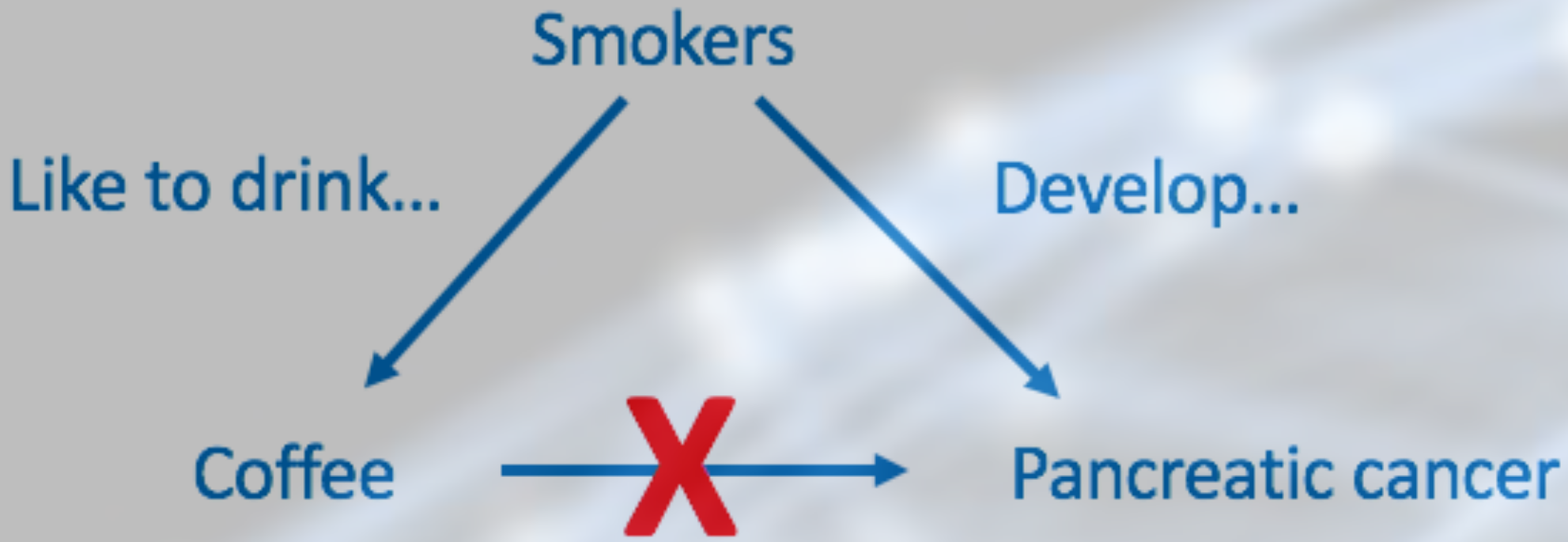
We will use concepts and insights from ML0 and ML1.

Correlation & Causation



Causality and Correlations





Anticorrelations (=neg. corr.) around year 2000

Pearson correlations = Covariance matrix (normalized) of Cryptos

Correlation between Top 15 Market Cap Currencies

@CryptoKit

Large-cap asset 3-month daily return correlation matrix (USD) 01/01/2017 - 10/04/2018

	BTC	ETH	XRP	BCH	LTC	EOS	ADA	XLM	NEO	MIOTA	XMR	DASH	TRX	XEM	ETC
BTC															
ETH	**0.45**														
XRP	0.20	0.19													
BCH	0.21	0.25	0.14												
LTC	**0.46**	**0.42**	0.27	0.23											
EOS	**0.31**	**0.31**	0.17	0.23	0.27										
ADA	0.22	0.18	0.27	0.10	0.17	0.17									
XLM	0.28	0.27	**0.50**	0.12	**0.31**	0.20	**0.31**								
NEO	0.30	**0.33**	0.13	0.15	**0.31**	0.19	0.14	0.21							
MIOTA	**0.44**	**0.39**	0.18	0.23	**0.35**	**0.32**	**0.32**	**0.34**	0.26						
XMR	**0.51**	**0.52**	0.23	0.28	**0.43**	0.28	0.26	**0.40**	0.24	**0.45**					
DASH	**0.42**	**0.45**	0.10	**0.31**	**0.37**	0.23	0.16	0.20	0.29	**0.34**	**0.56**				
TRX	0.28	0.21	0.17	0.11	0.18	0.24	0.28	0.12	0.09	0.15	0.18	0.19			
XEM	0.28	**0.35**	0.22	0.21	**0.34**	0.23	0.24	**0.32**	0.22	**0.37**	**0.32**	0.29	0.13		
ETC	**0.44**	**0.62**	0.17	**0.32**	**0.51**	0.30	**0.32**	0.29	**0.43**	**0.43**	**0.45**	**0.38**	0.20	**0.35**	

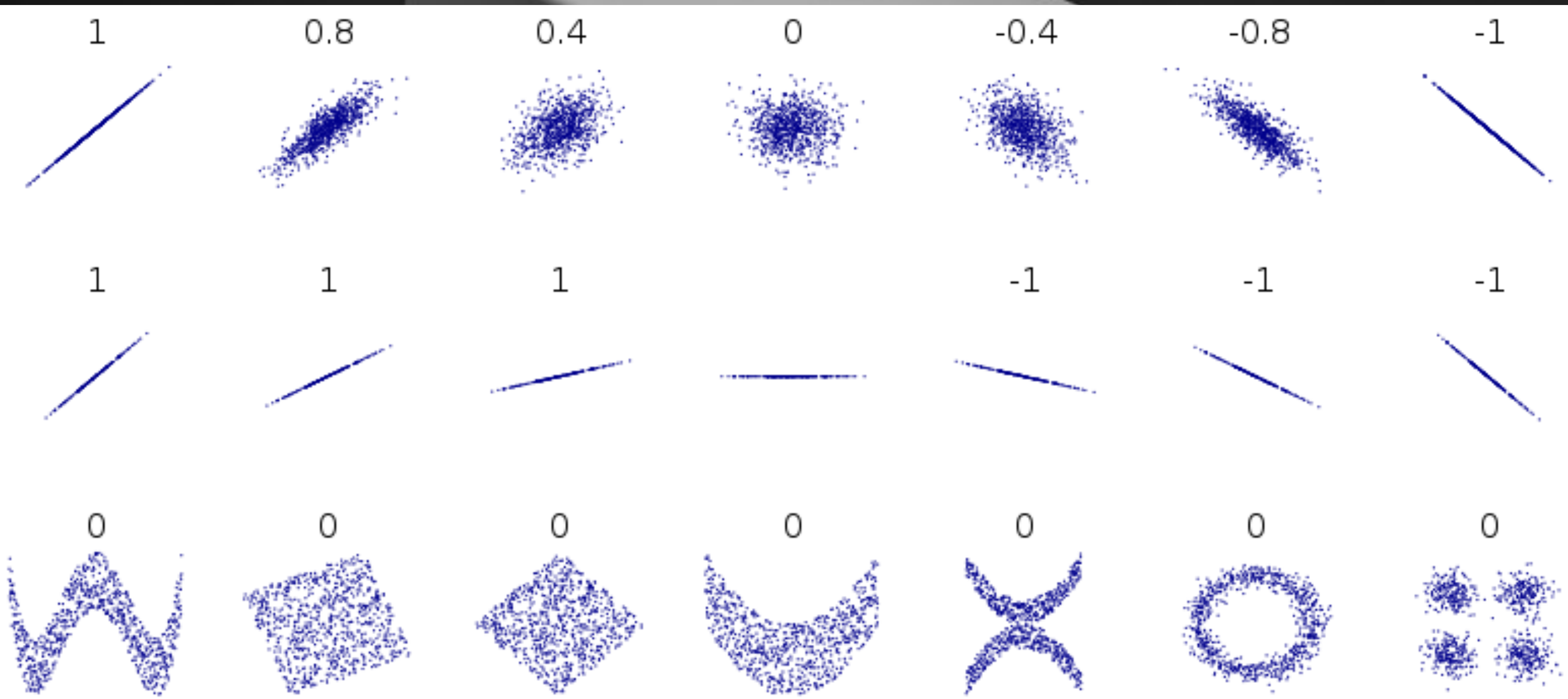
1.00	Strong Positive Relationship
0.50	Moderate Positive Relationship
0.30	Weak Positive Relationship
0.00	No Correlation
-0.30	Weak Negative
-0.50	Moderate Negative Relationship
-1.00	Strong Negative Relationship

Pearson correlations = Covariance matrix (normalized) of Cryptos

3-month daily return correlation matrix (USD) Dec 1 2018 - Mar 1 2019

	BTC	ADA	ETH	XLM	XMR	ZEC	NEO	LSK	XRP	ZRX	EOS	OMG	VET	DASH	QTUM
BTC	1														
ADA	0.908	1													
ETH	0.889	0.900	1												
XLM	0.922	0.893	0.866	1											
XMR	0.929	0.886	0.881	0.878	1										
ZEC	0.909	0.843	0.865	0.861	0.915	1									
NEO	0.873	0.875	0.885	0.864	0.858	0.874	1								
LSK	0.901	0.830	0.837	0.836	0.838	0.847	0.818	1							
XRP	0.875	0.865	0.858	0.887	0.840	0.828	0.847	0.819	1						
ZRX	0.879	0.827	0.834	0.857	0.829	0.828	0.803	0.867	0.812	1					
EOS	0.884	0.889	0.851	0.860	0.824	0.813	0.836	0.770	0.818	0.769	1				
OMG	0.843	0.822	0.846	0.802	0.829	0.839	0.755	0.837	0.740	0.847	0.751	1			
VET	0.855	0.862	0.794	0.839	0.801	0.794	0.793	0.848	0.820	0.820	0.810	0.792	1		
DASH	0.885	0.800	0.799	0.819	0.875	0.892	0.769	0.802	0.752	0.807	0.770	0.840	0.745	1	
QTUM	0.794	0.831	0.826	0.801	0.727	0.746	0.805	0.784	0.793	0.773	0.820	0.754	0.800	0.697	
LTC	0.857	0.851	0.825	0.801	0.805	0.766	0.793	0.753	0.769	0.750	0.831	0.758	0.767	0.706	0.75
ETC	0.783	0.778	0.794	0.760	0.760	0.769	0.773	0.769	0.736	0.806	0.734	0.741	0.724	0.712	0.75
DCR	0.815	0.785	0.795	0.740	0.795	0.785	0.738	0.819	0.760	0.740	0.714	0.722	0.764	0.702	0.69
NEM	0.777	0.766	0.755	0.804	0.755	0.759	0.768	0.777	0.754	0.780	0.652	0.759	0.759	0.699	0.70
IOTA	0.768	0.756	0.768	0.759	0.760	0.752	0.675	0.694	0.738	0.756	0.729	0.777	0.697	0.752	0.78
BCH	0.790	0.752	0.738	0.738	0.778	0.752	0.622	0.745	0.679	0.727	0.675	0.784	0.664	0.815	0.62
ONT	0.697	0.760	0.678	0.743	0.710	0.676	0.747	0.624	0.701	0.618	0.751	0.657	0.709	0.599	0.67
BAT	0.719	0.703	0.728	0.704	0.707	0.702	0.703	0.711	0.674	0.668	0.626	0.688	0.650	0.656	0.62
XTZ	0.696	0.714	0.720	0.711	0.680	0.709	0.724	0.682	0.669	0.666	0.696	0.635	0.605	0.656	0.62
MKR	0.643	0.662	0.698	0.646	0.650	0.642	0.718	0.651	0.618	0.554	0.642	0.566	0.575	0.541	0.64
BNB	0.697	0.652	0.635	0.681	0.667	0.644	0.664	0.643	0.614	0.548	0.643	0.560	0.598	0.569	0.58
TRX	0.615	0.722	0.633	0.628	0.614	0.599	0.646	0.617	0.589	0.583	0.567	0.624	0.693	0.505	0.57

Linear and nonlinear correlations
[number = linear (Pearson) correlation]



Recall key concepts

ML0-Examples (for how things are connected and what are the differences)
Linear correlation <-> Nonlinear correlations <-> Causation

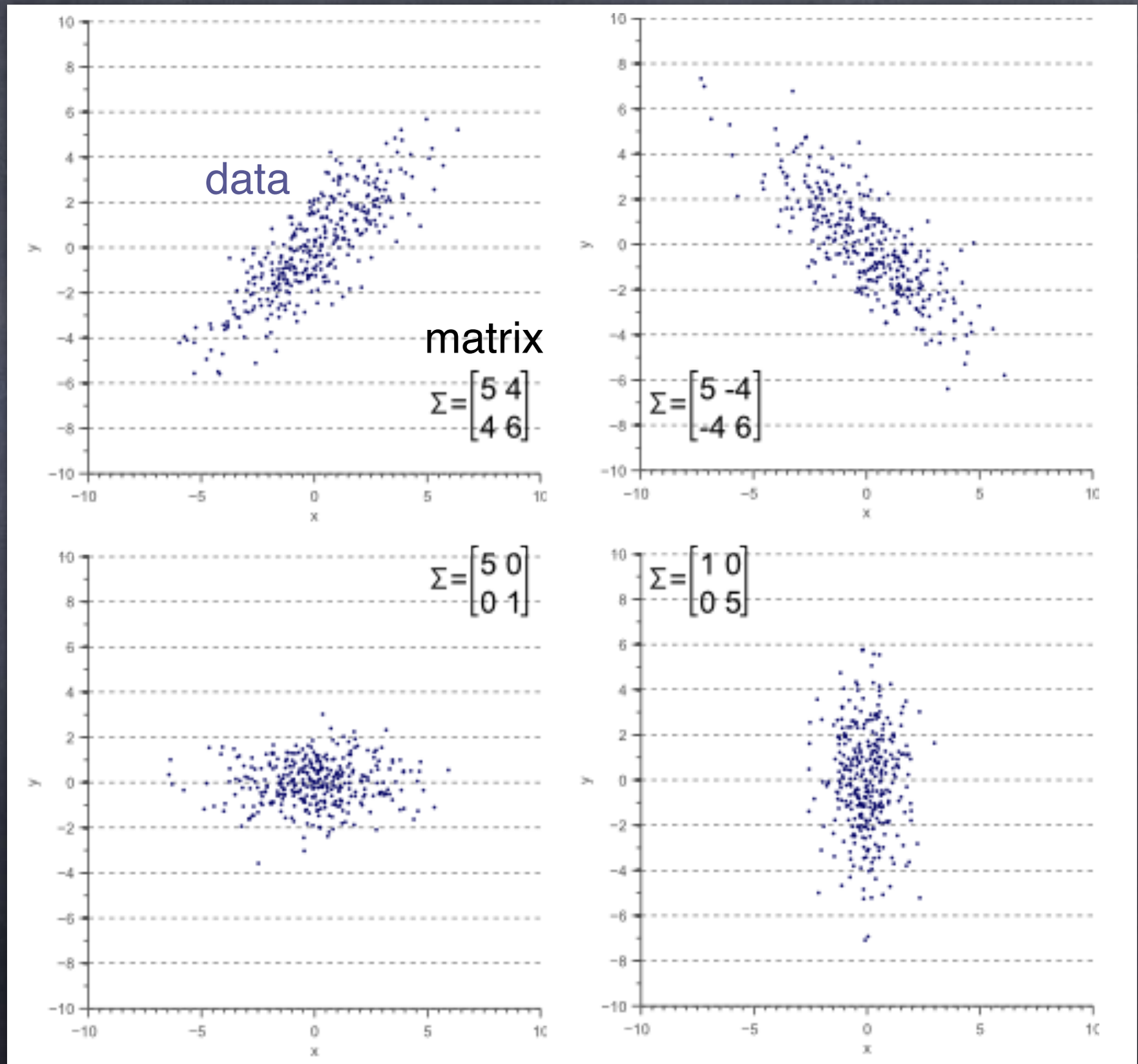
and

Covariance matrix

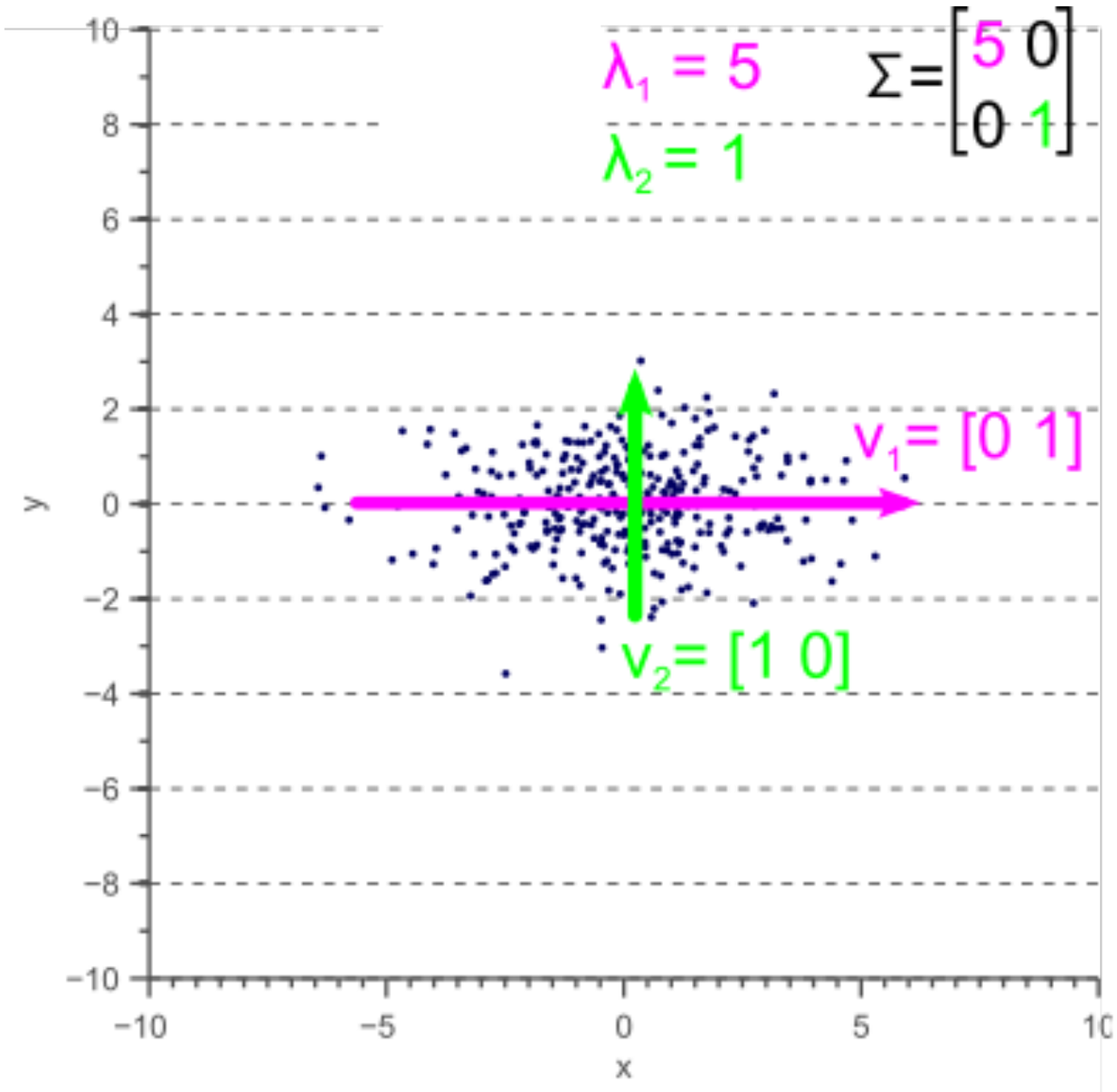
$$\Sigma = \begin{bmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{bmatrix}$$

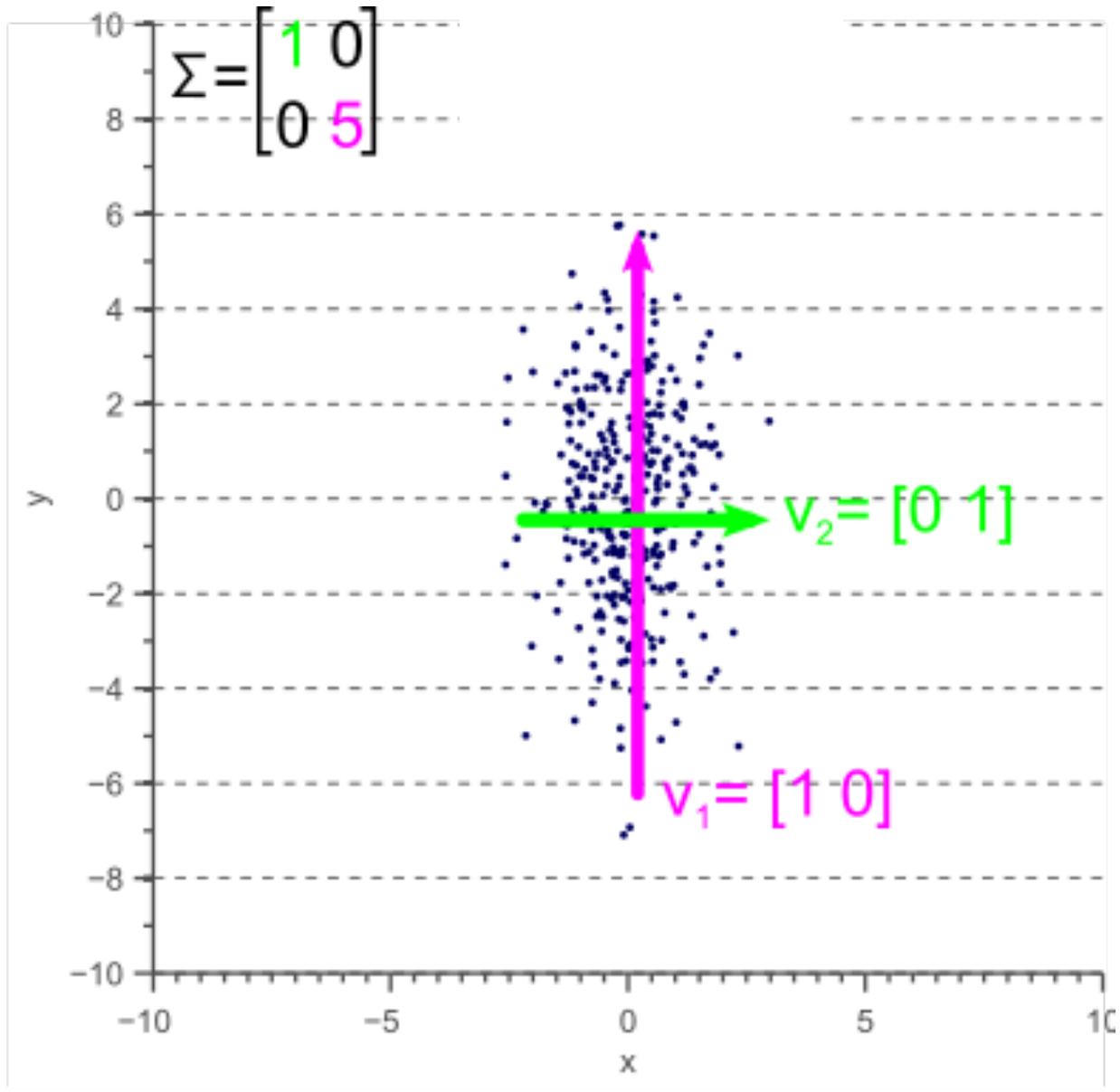
P(rincipal) C(omponent) A(nalysis)

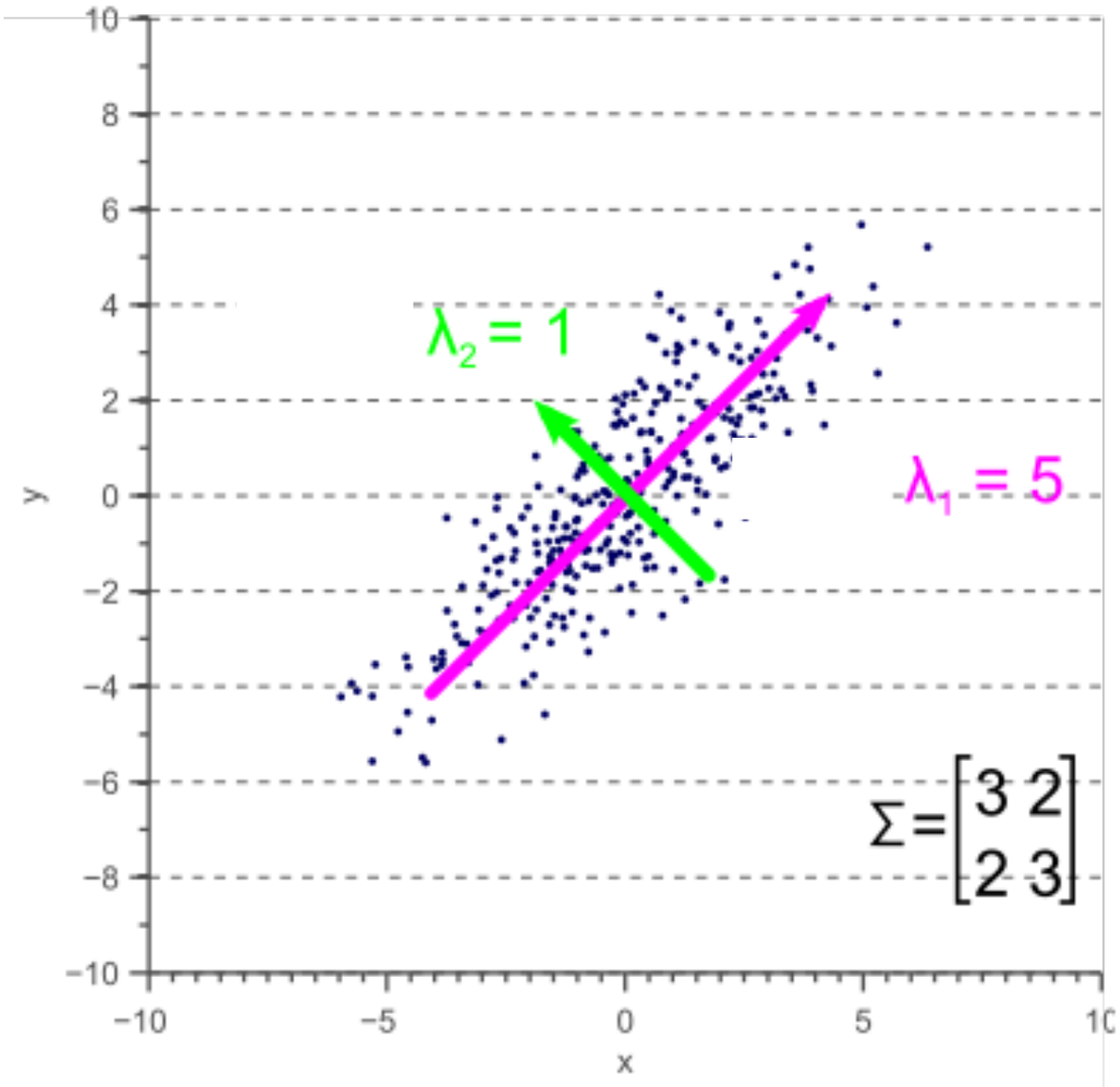
Understanding covariance

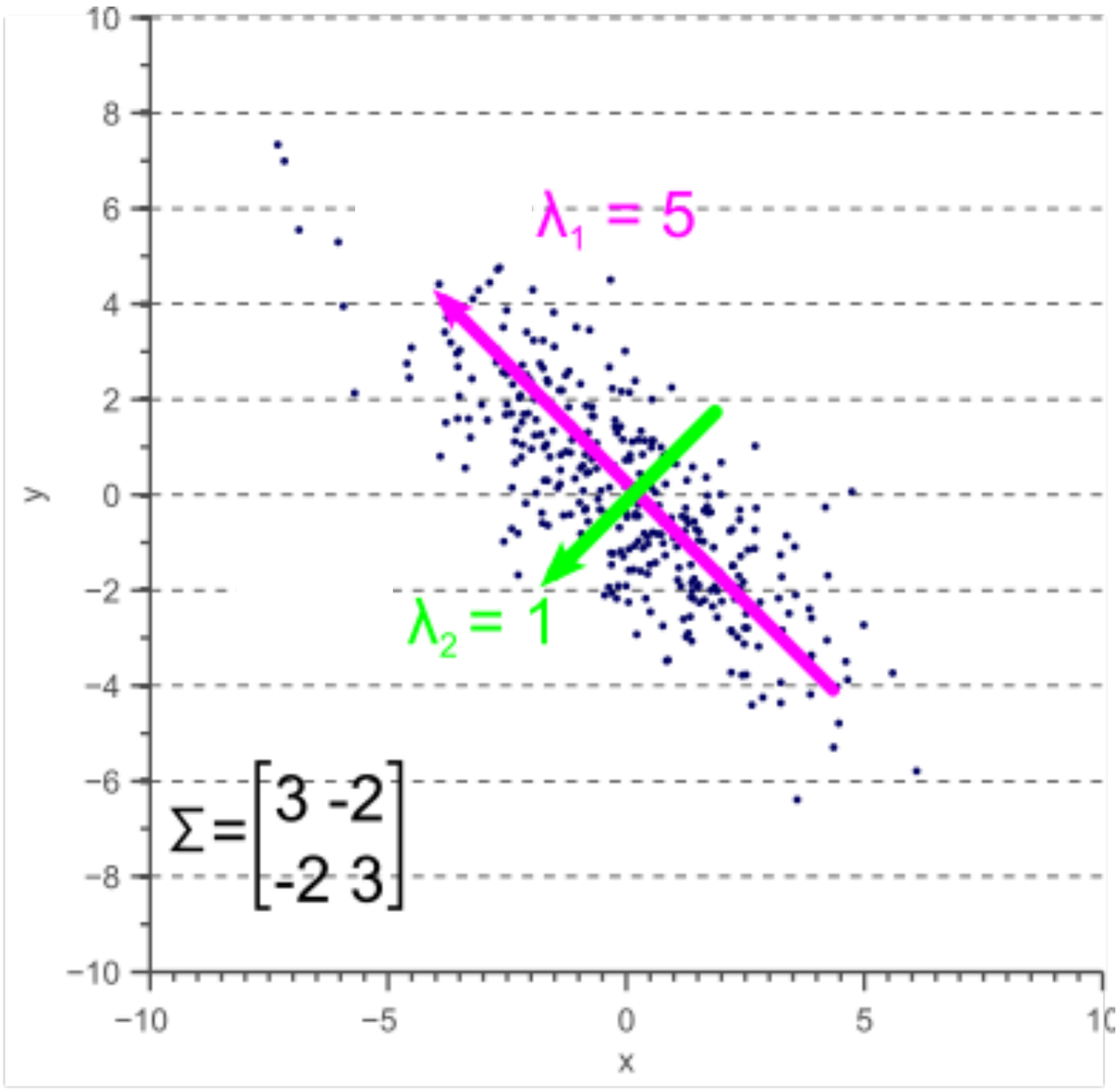


Eigenvectors and eigenvalues



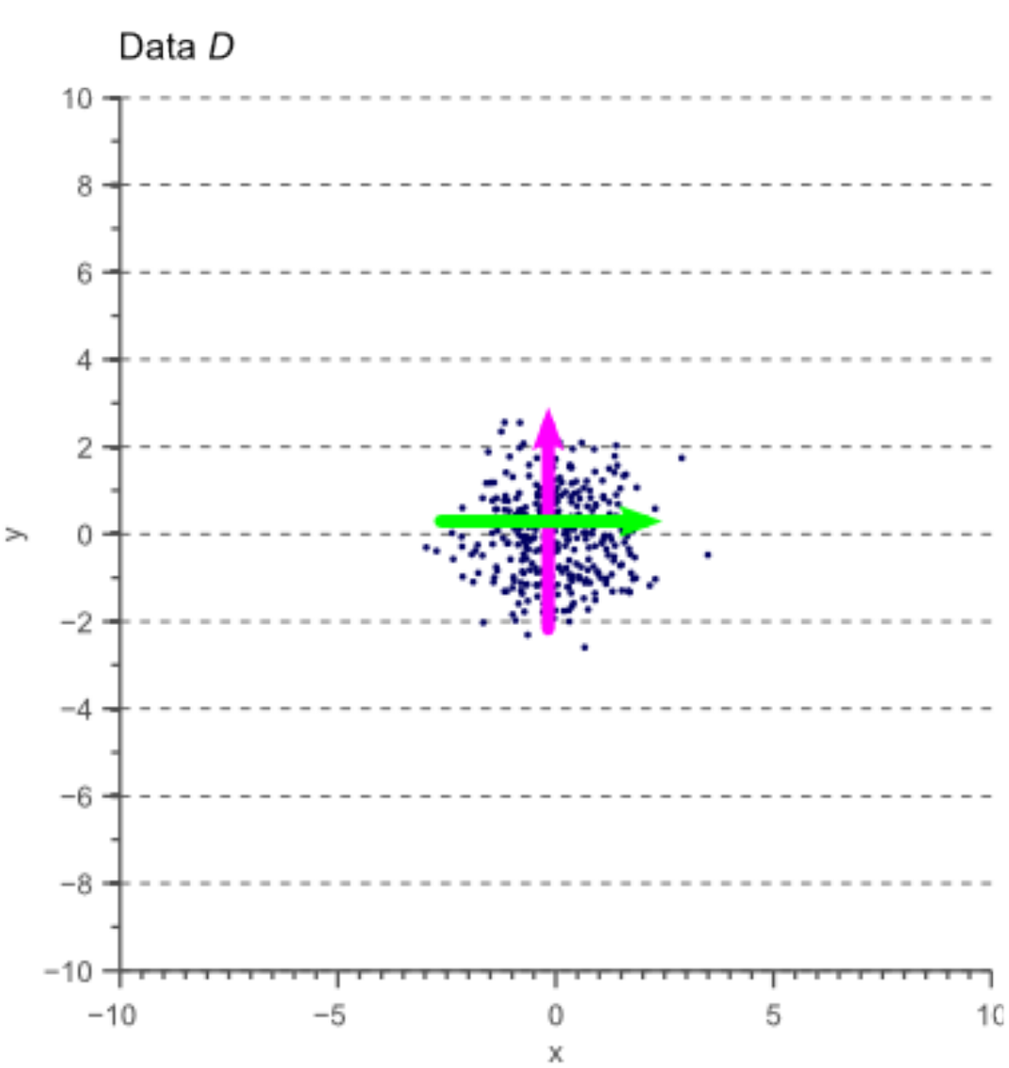



λ2 = 1
λ1 = 5
Σ=[3 2; 2 3]
y
x

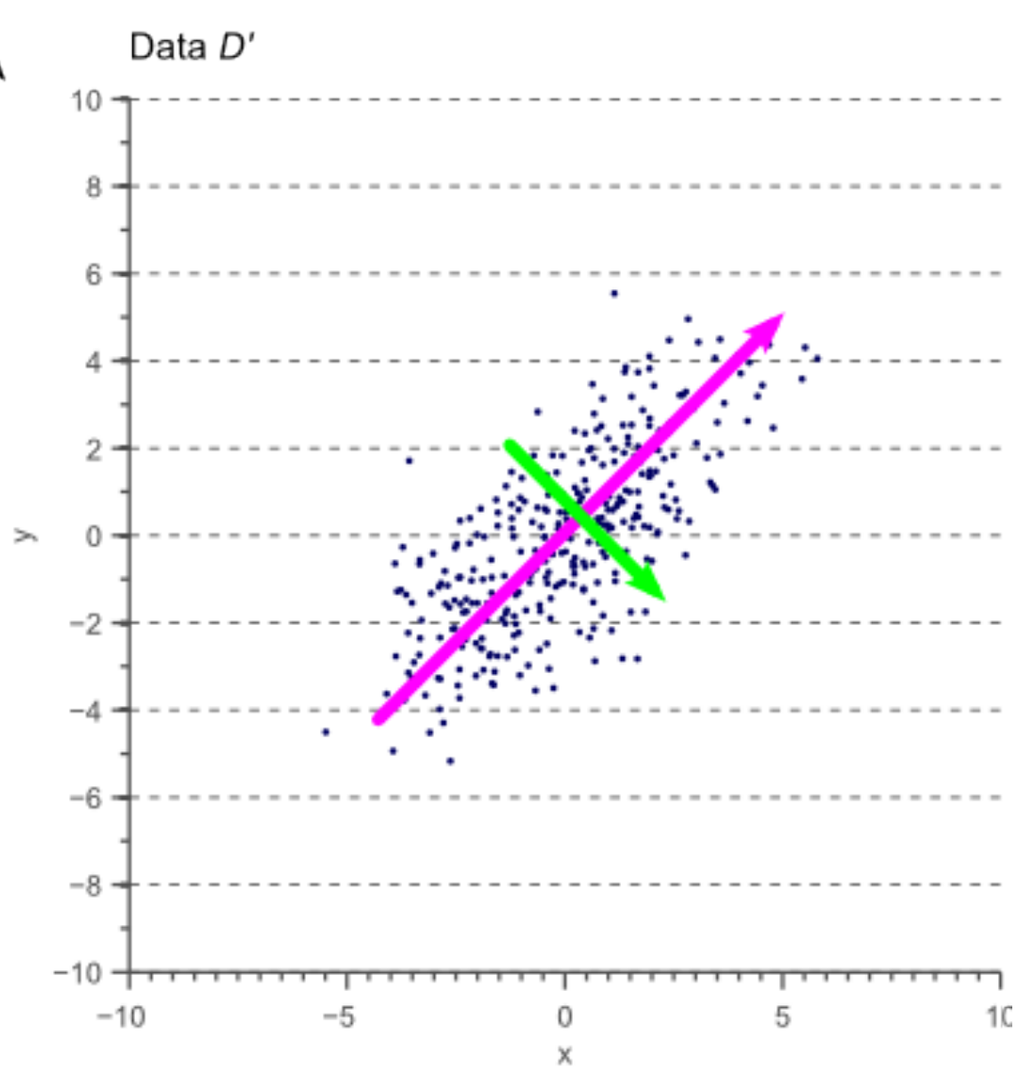

λ1 = 5
λ2 = 1
Σ=[3 -2; -2 3]
y
x

Understanding covariance is understanding PCA

Recall: Eigenvalue decomposition, apply to covariance matrix



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\Sigma' = \begin{bmatrix} 4.25 & 3.10 \\ 3.10 & 4.29 \end{bmatrix}$$

iPad presentation:

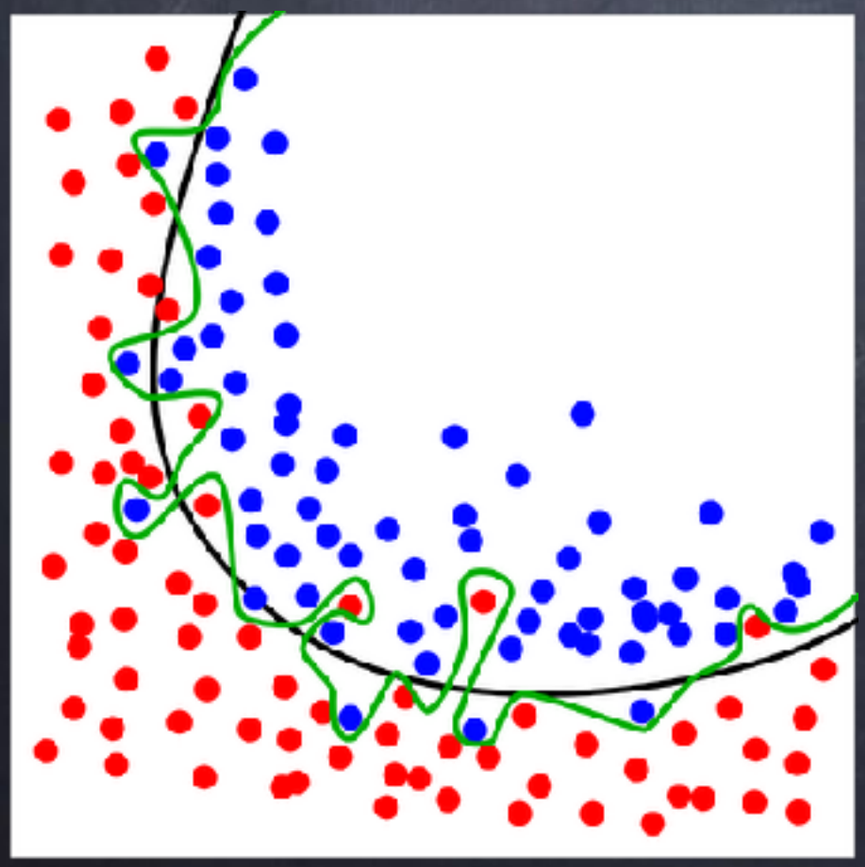
What transformation is that and what is the relation the EV decomposition/SVD?

PCA

Dimensionality reduction

(Curse of dimensionality)

Avoidance of overfitting
Avoidance of variance
Feature extraction



Methods:

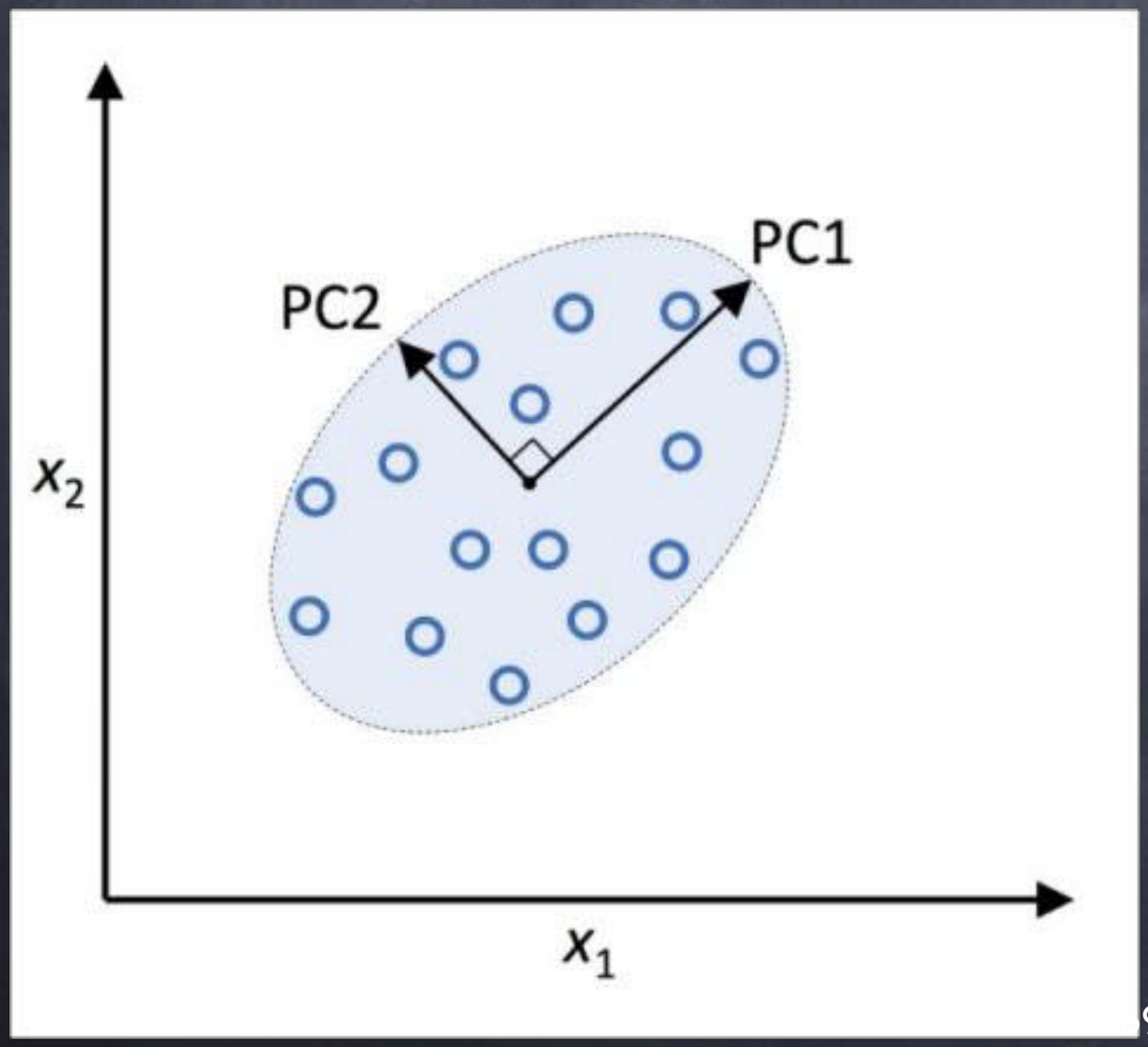
PCA

(Principal Component Analysis)

Classifier with linear decision boundary, lr/ova

P(rincipal) C(omponent) A(nalysis)

Main idea: Obtain principal components
from variance maximization,
given orthogonal feature axes



Main methodology:
Linear algebra

P(rincipal) C(omponent) A(nalysis)

Find linear transformation $W(d \times k)$

$$\mathbf{x} = (x_1, \ldots, x_d) \Rightarrow \mathbf{z} = (z_1, \ldots, z_k)$$

Objective

First principal component selected from largest possible variance,
and all consequent principal components will have the largest variance
given the constraint that these components are uncorrelated (orthogonal)
to the other principal components

Even if the input features are correlated,
resulting principal components will be orthogonal (uncorrelated).

P(rincipal) C(omponent) A(nalysis)

Standardize dataset of dim(d)

Compute covariance matrix

Decompose covariance matrix
using its eigenvectors and eigenvalues

Sort the eigenvalues by decreasing order
to rank the corresponding eigenvectors

Select k eigenvectors which correspond to the k largest eigenvalues,
where k is the dimensionality of the new feature subspace ($k \leq d$)

Construct a projection matrix W from the selected k eigenvectors.

Transform the d-dimensional input dataset X
using the projection matrix W
to obtain the new k-dimensional feature subspace

P(rincipal) C(omponent) A(nalysis)

Sample covariance (given data x)

$$\sigma_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_j^{(i)} - \bar{x}_j)(x_k^{(i)} - \bar{x}_k)$$

2-dim covariance matrix with notation $x = x_1; y = x_2$

$$\Sigma = \begin{bmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{bmatrix}$$

Eigenvalue problem

$$\mathbf{\Sigma v} = \lambda \mathbf{v}$$

The eigenvectors of the covariance matrix represent
the directions of maximum variance,
hence they are the principal components

P(rincipal) C(omponent) A(nalysis)

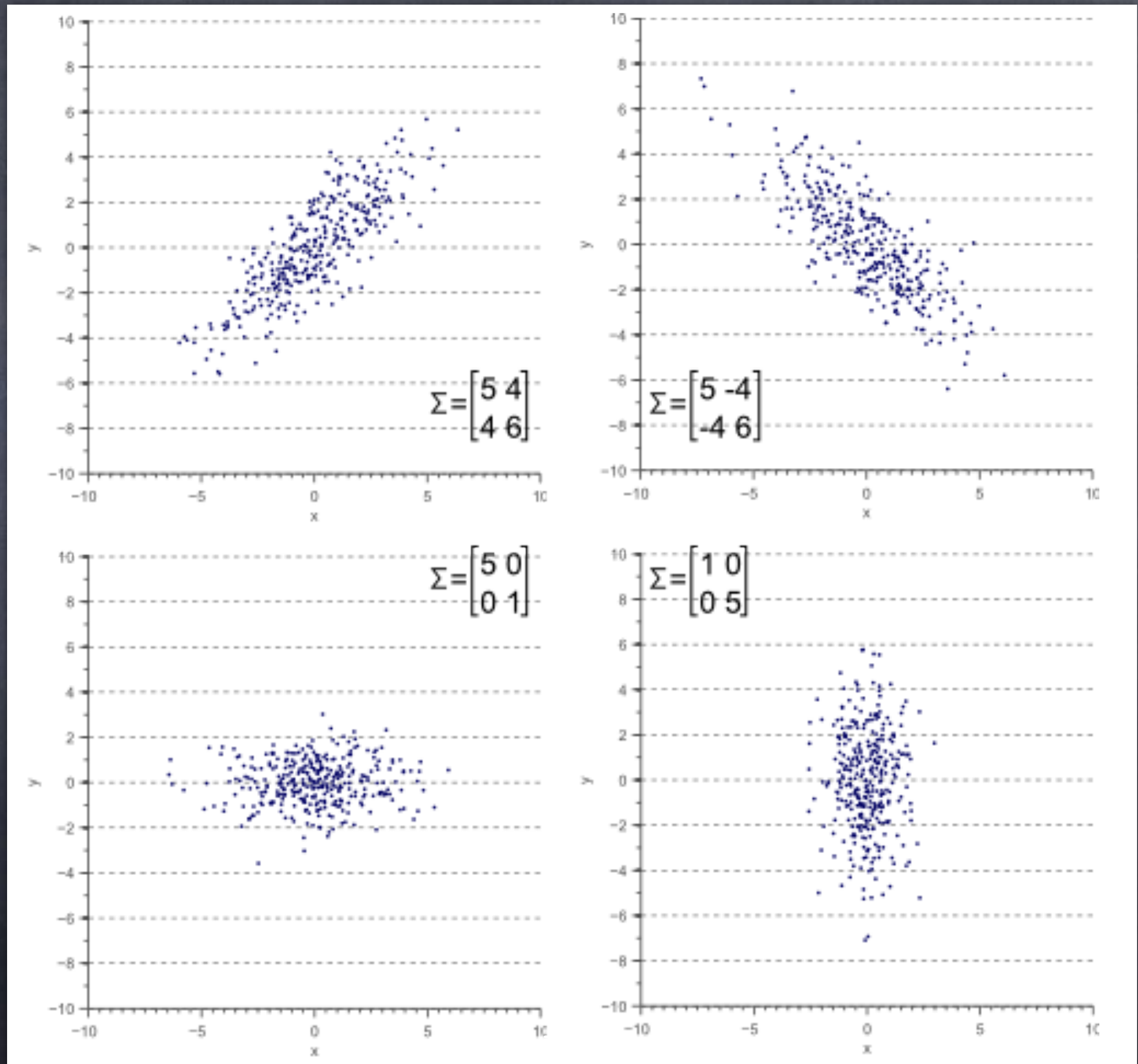
Variance-explained-ratios

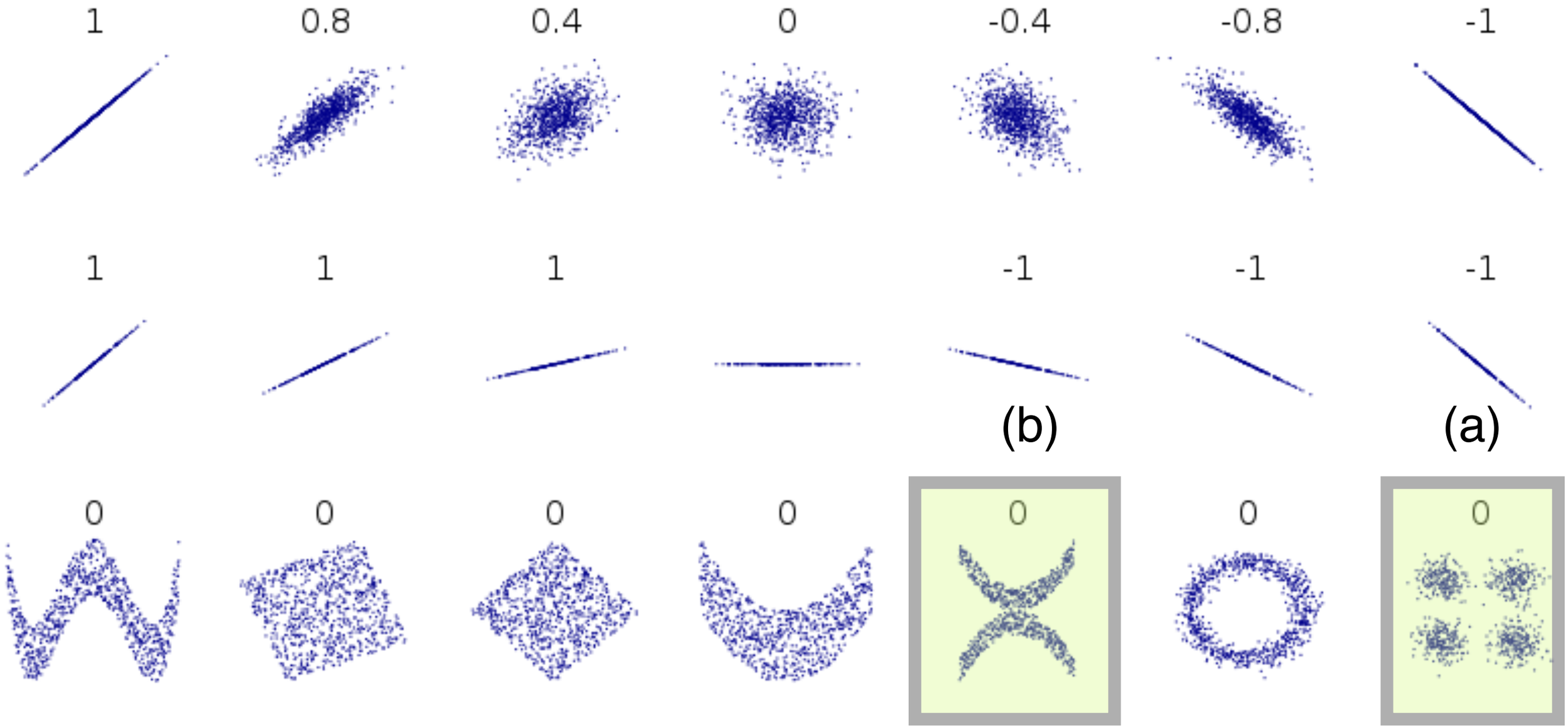
$$\frac{|\lambda_j|}{\sum_{i \leq d} |\lambda_i|}$$

python

P(rincipal) C(omponent) A(nalysis)

Understanding covariance = understanding linear correlations





Assignment 1, part II

K-Means Clustering

based on

Centroid (average)

Number of clusters, k, a priori

K-Means Clustering: Preprocessing

Standardization =
Z-score normalization =
(zero mean, unit variance)

or

Min-Max Scaling

K-Means Clustering

Number of clusters, k, a priori
Somehow (e.g., randomly) pick k centroids from the sample points as initial cluster centers

Repeat
(1) Assign each point to the nearest centroid

(2) Move the centroids to the centers of the samples that were assigned to it

Until
cluster assignments do not change or
within tolerance or
maximum number of iterations

K-Means Clustering

max_iter:
Maximum number of iterations

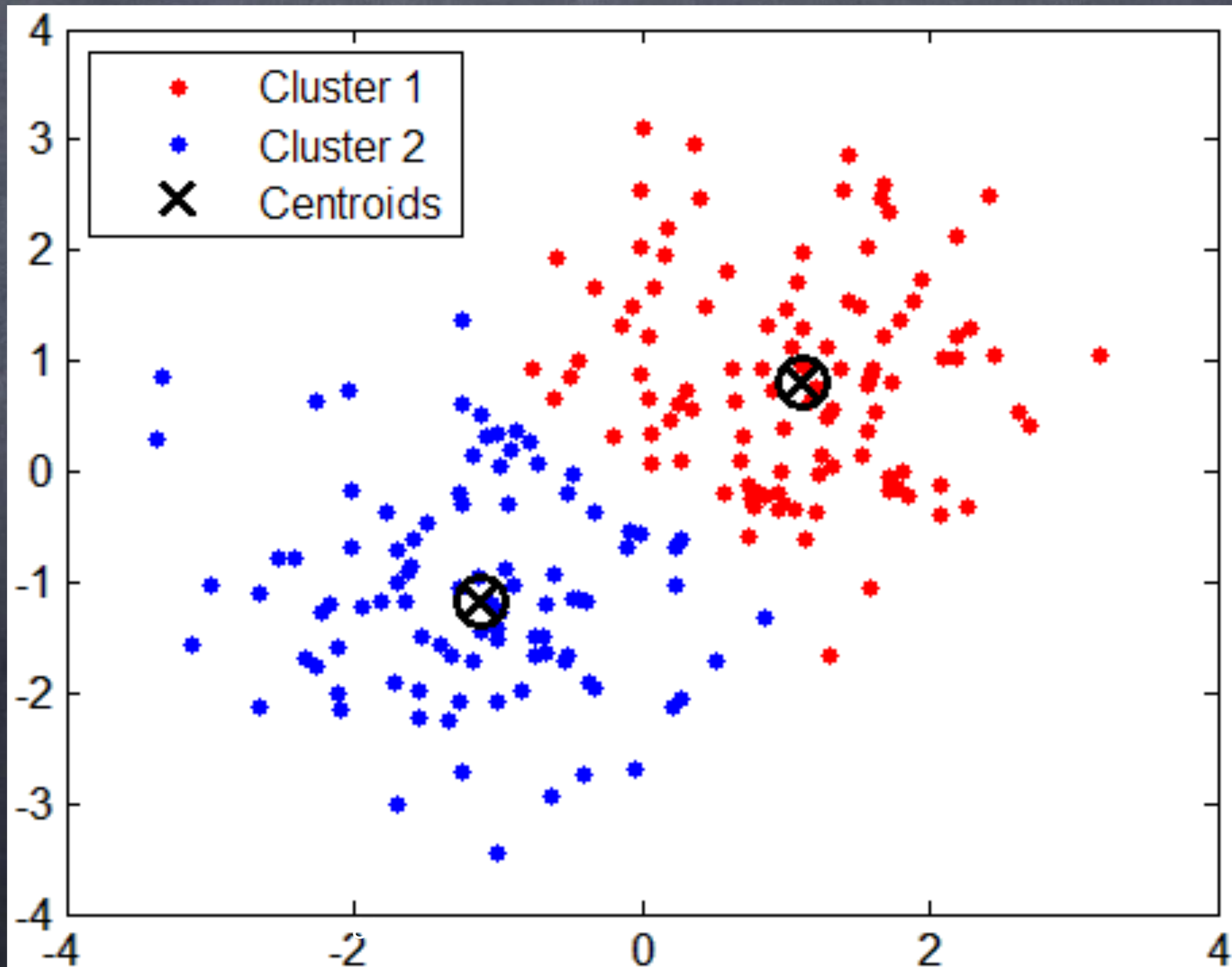
tol(erance):
Relative tolerance with regards to inertia/SSE
to declare convergence,
SSE(iter+1)/SEE(iter) < tol

Sum of squared errors (SSE)

$$\mathrm{SSE} = \sum_{i}^{n} \sum_{j}^{k} w^{(i,j)} ||\mathbf{x}^{(i)} - \mathbf{y}^{(j)}||_2^2$$

$w^{(i,j)}$ is 1 if x(i) belongs to cluster j [sitting at y(j)], otherwise 0

$$\text{SSE} = \sum_{i}^{n} \sum_{j}^{k} w^{(i,j)} ||\mathbf{x}^{(i)} - \mathbf{y}^{(j)}||_2^2$$




Clustering monster

Group green

Presentation of clustering methods

Monster(*) methods, included

Keyword guide to the cluster galaxy:

Inference/Expectation Maximization

Neigherst Neighbors

Hard / Soft Clustering

Closet point

Intercluster distance

Cohesion

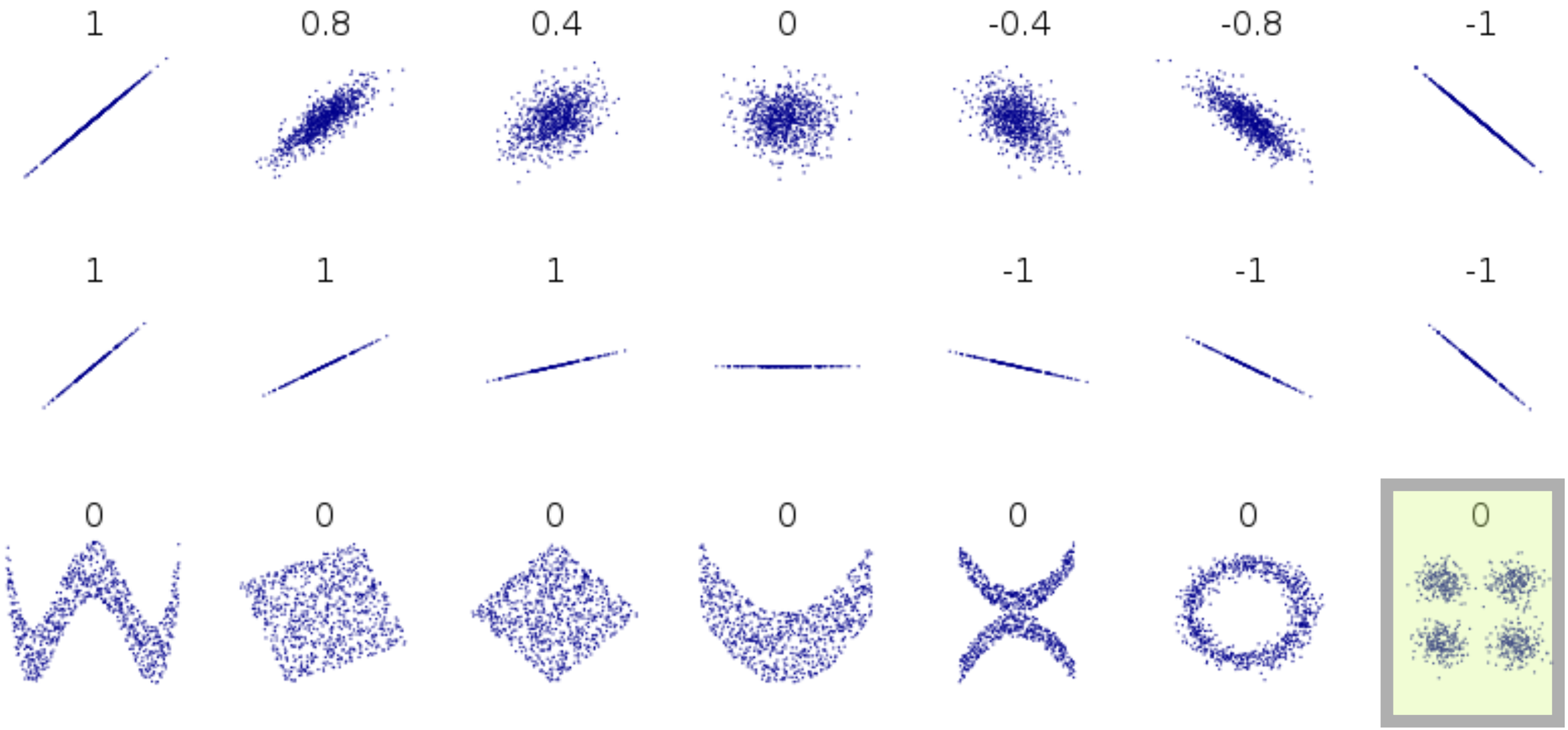
Maximal distance from “Clustroid”

Cluster diameter

Density

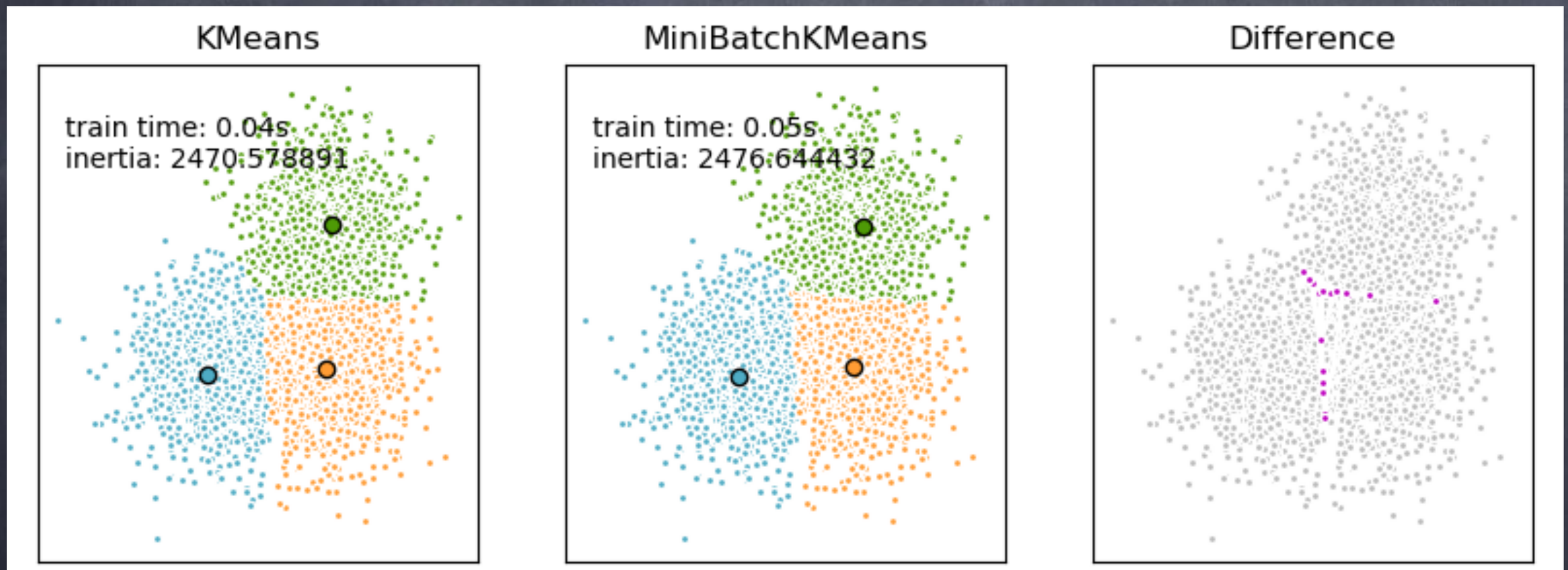
Stopping criteria

Noise points / Retained sets



user uploaded python code: cluster monster

Example: Difference (output) K-Means vs Mini Batch K-Means



Machine Learning II
Week #1 or #2…

Jan Nagler

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

(Details in lecture!)

Frankfurt, October 2019

Group blue

You are the distance group: Self-learn clustering methods and present the main ideas to the class

Group blue

The right (metric) measure:
What distance??

Euklidean? What p(ower)-norms?

Based on correlation & similarity?
Cosine, Mahalanobis distance,
Jaccard, Edit distance

Neighrest k neighbors-manifold based:
Isomap

Add cr(azy)(eative) ideas:
Randomized component pick?
How to determine the importance of components? Given data, what metric?

Study but focus:
Main ideas behind the distances

Resources:
Use web (re)sources
(skikit-learn, medium,
kdnuggets, dzone, wiki,
google)

Self-organize:
Split up the tasks

Final presentation

Be incomplete, choose what you like!

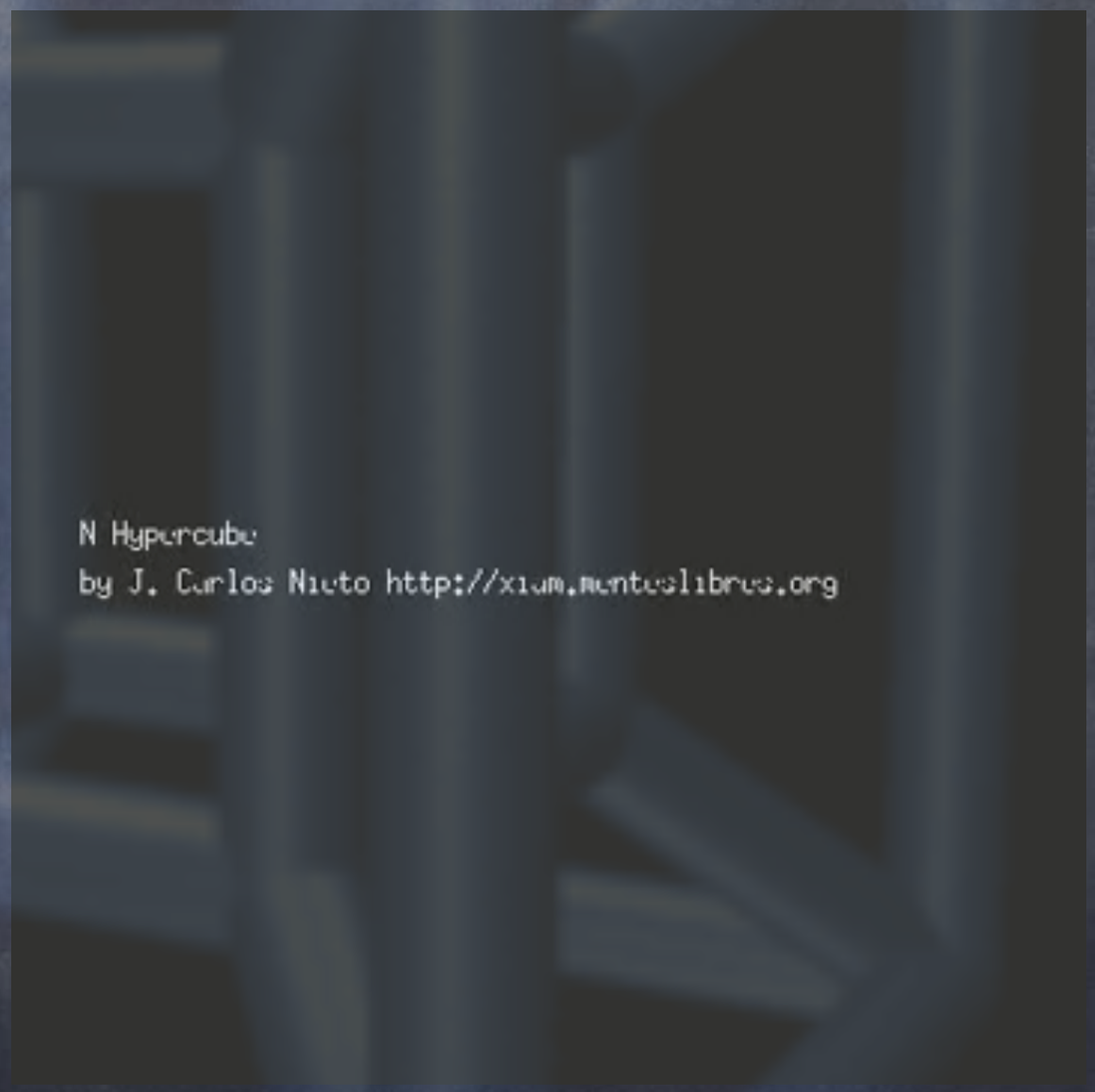
Mahalanobis distance add on:

to point cloud: $$d(\mathbf{x}) = \left[(\mathbf{x}-\mu)^T S^{-1}(\mathbf{x}-\mu)\right]^{1/2}$$

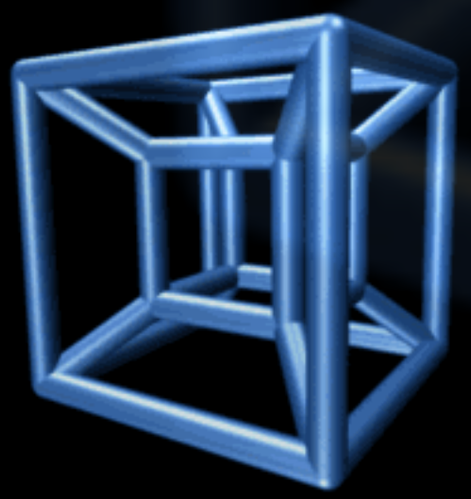
between x and y: $$d(\mathbf{x},\mathbf{y}) = \left[(\mathbf{x}-\mathbf{y})^T S^{-1}(\mathbf{x}-\mathbf{y})\right]^{1/2}$$

Consequence, if cov matrix S diagonal?

Curse of Dimensionality revisited



4d cube



Curse of Dimensionality revisited

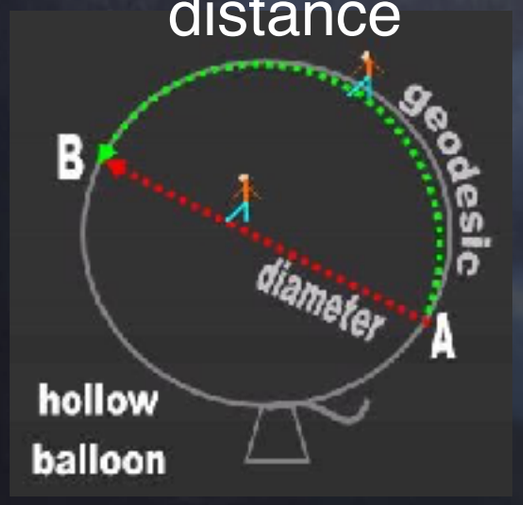
Combinatorial explosion

Vanishing volume ratio
(Hypersphere versus hypercube)

Vanishing center domain
(Distance random to corner)

Physical embedding / Meaning of orthogonality
Physical space, chemical concentrations, asset prices,
eye color, age, IQ … often relate to objects in space

good & bad & ugly distance



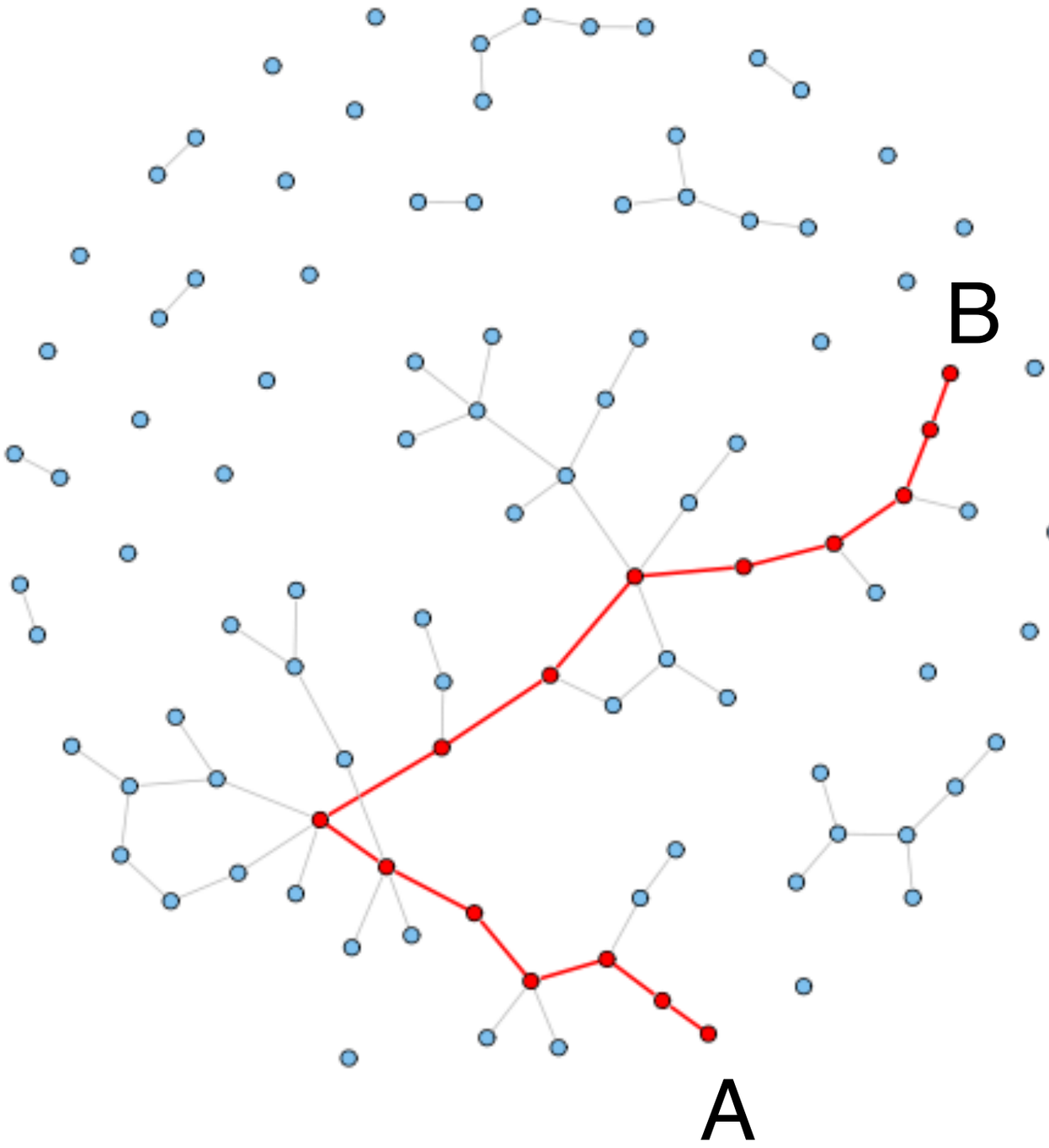
— types: nominal, ordinal, ratio, text,
image, video, audio, …

—distance: can be undefined, meaningless,
totally unknown or ambiguous but needs to be
well chosen

Curse of metric distance

Curse of data types

Network distance metrics



Network metrics

fully connected | real networks | tree

	Connectivity driven : A	Critical window : B	Utility driven : C
Pruning model