STEMinars 2021: It's Models All the Way Down!

Understanding the COVID-19 Pandemic

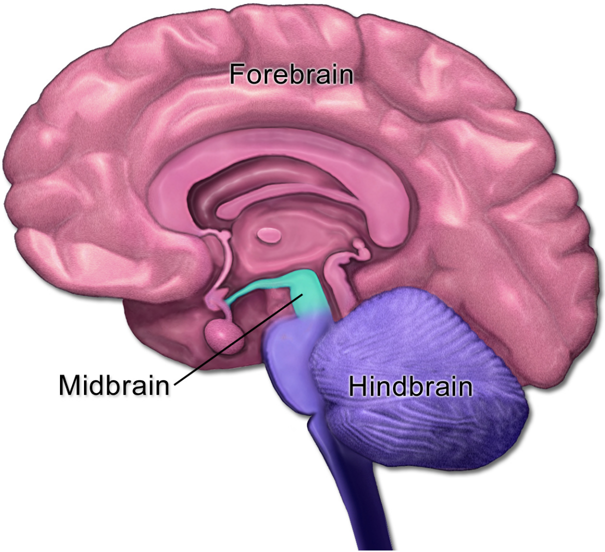
Stephen J. Turnbull

Associate Professor
Division of Policy and Planning Science
University of Tsukuba

23 September 2021

A Galaxy + . . .

A Brain = . . .



Galaxy Brain!

Introduction

- Thinking about *anything* uses models.
- Flat earth, Bertrand Russell, and turtles.
- Models lead to more models.
- Models are reusable.
- The constant model of turtles.

All models are wrong . . .

. . . but some are useful.

- Some models are useful sometimes: the constant model and weather.
- Some models are harmful: conspiracy theories.
- Some models are hard: jealous of his . . . sister?!
- But always, *some* model is necessary!

Definitions for Pandemics

endemic an attribute (such as “ill with COVID-19”) that is present at a “low” level throughout a whole population

epidemic when that “breaks out” to high levels, usually in specific subpopulation

pandemic (literally) an epidemic that covers the whole population (usually the whole world)

pandemic (practically) an epidemic that wanders from subpopulation to subpopulation

These words are from Greek. The prefixes are commonly used in words we borrow in English.

The Constant Model and COVID-19

- In some ways, COVID-19 is a lot like influenza. Although it's a new disease, many politicians and business leaders argued that we could treat it like the flu.
- They also complain that the "lockdown" orders had a big economic effect, comparing the levels of GDP (or *sector value-added*) or employment from first-quarter figures for 2019 to those for first-quarter 2020.
- This is based on the *constant model*. Why? Because the standard of comparison for "very large negative effect" is *last period's GDP* (or *employment*, etc.).
- The implied assumption is "if we had no lockdowns, the economy would work the same as last year," i.e., the constant model.

Rejecting a bad model

- OK, it's the constant model. Is that *bad*? After all, it works for the flu, and the economy does fine with the flu (even before vaccines).
- The problem is that within a few weeks **we knew** from the experience of Wuhan (China), Bergamo (Italy), and New York (US) **that COVID-19 is different from the flu.** We *didn't* have *any* vaccine, a relatively *large fraction* of the population gets sick, and relatively *many* of them get sick enough to die.
- A year later, we know about "long COVID."
- Finally, it had a great effect on the economy even before shutdown orders in those places, even if you *only* count economic losses due to sick workers and shortages from falling production, and the like.
- **The constant model is untenable in the case of COVID-19.**

Counterfactuals and Models

- To estimate the costs (economic and otherwise) of business shutdowns, **we need to evaluate a counterfactual**: “What would the level of economic activity be *with* COVID-19 but *without* the shutdown?
- How do we evaluate a counterfactual? You guessed it, I’m sure. We build a model. *We have no choice but to build a model.* **Working with counterfactuals always involves models.**
- Sometimes we can avoid explicit modeling. If we’re lucky, we have the right data and we use analogical reasoning:
  - *This* situation is like *that* situation, so the outcome *this* time should be like the outcome *that* time.
  - This is just an alternative description of the constant model.
- Building a model of the effect of COVID-19 on a national economy is hard. So hard that no serious professionals published models until 3rd quarter 2020.

Model Scope

- Part of economic modeling must be modeling the disease's *medical effects*, which is not something we can do here.
- But there's a component of that model that we can at least get started on: the *epidemiology* of the virus (i.e., the scientific study of how it spreads).

  We decompose the problem and model the parts we can understand.
- Working on just that part is called "limiting the model's *scope*."

Modeling Strategy

- How do we build good models? It's simple:
- **We keep asking "why?" and what it means for our model.**

Simplest epidemiological model

- What's a better model? From the worst-hit cities, the SARS-CoV-2 virus has *faster than linear growth*, varying from place to place (and responding to policy), with an estimated *doubling time* of 2–7 days.
- Assuming we take no special action, and a doubling time of y days (original strain), starting from *one* infected person, we get the table on the next slide.

You may want to pause the video at the next slide to see the precise number of cases, I'm going to explain with "social" units instead of number of cases.

Simulation: Doubling Time of 7 Days

week	count	what	day	count	what	day	count
0	1	me	12	4,096		24	16,777,216
1	2	& you	13	8,192		25	33,554,432
2	4		14	16,384	UT	26	67,108,864
3	8		15	32,768		27	134,217,728
4	16		16	65,536		28	268,435,456
5	32		17	131,072	Tsukuba	29	536,870,912
6	64	class	18	262,144	+Tsuchiura	30	1,073,741,824
7	128		19	524,288		31	2,147,483,648
8	256		20	1,048,576		32	4,294,967,296
9	512	Shako	21	2,097,152	Ibaraki	33	8,589,934,592
10	1,024		22	4,194,304		34	17,179,869,184
11	2,048		23	8,388,608			

Table: Exponential growth model

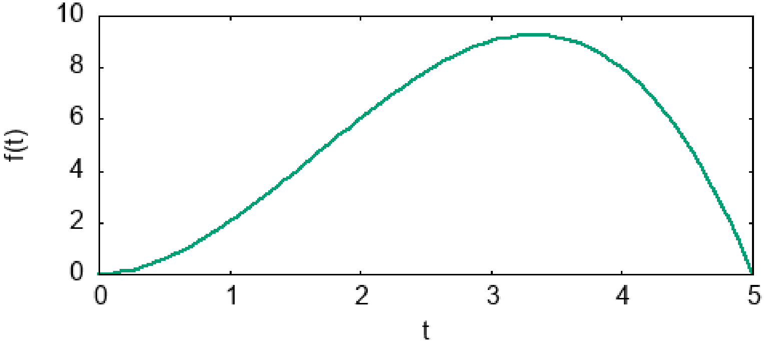
Lessons from the Doubling Model

- If I'm sick today, in a month and a half a whole class has gotten sick, in a quarter my whole university is infected, in half a year the whole Kanto area (Tokyo megalopolis), one week after that all of Japan, a month later all of Asia and Europe, in a total of eight months the **whole world**, and a week after that . . . uh, wait . . . oops.
- **Lesson #1:** You can run, but you can't hide from exponential growth.
- **Lesson #2:** Exponential growth is a bad model of an epidemic. It predicts *impossible* outcomes.

An embarrassingly bad model

- The cubic model used by the White House around May 1, 2020. t is months since Feb. 1.

$$f(t) = -0.5\, t^2(t-5)$$



Making Models Better

- The constant model is *just plain bad* because with data, it can only predict the same thing (or an average), and *without* data, it predicts anything you can imagine. But arguing about our imaginings is useless.
- The cubic model at least tries to match the early part of the curve, and avoids “crashing through the ceiling.” But it gives no reason why cases should go to zero, let alone on a particular date. Instead, it seems to be *calibrated* to give a predetermined result.
- Let’s “shift gears” to define a better mathematical model.
- As we’ll see later, the exponential growth model is not a bad model in the way that the constant model or cubic model is.

Shifting gears, *cont.*

- **To see that exponential growth is a somewhat useful model it's helpful to use different mathematics**, namely *continuous time*.
- Our calculations were done with *discrete time*, calculating only for every week. What about weekdays?
- Can we say something about hours or minutes?

Discrete time to continuous time

- A bit of thought will show that after t weeks, we have $2 \times 2 \times \cdots \times 2$ (multiply by 2, t times), or $N = 2^t$ infected individuals. So if we allow t to be fractional numbers, we can still calculate this (with a computer).
- Interesting, but it doesn't help with Lesson #2. We need to change the model's doubling time so that we can't more sick people than we have people!
- The trick: look at the relationship between the *number of infected individuals* and the *rate of increase of infected individuals*.
    - Each week we count infected individuals N.
    - A week later we'll have $2N$.
    - The *rate of increase* is $2N - N = N$ (the unit of time is weeks).

  The rate of increase is expressed in terms of N.

What is Mathematics *For*?

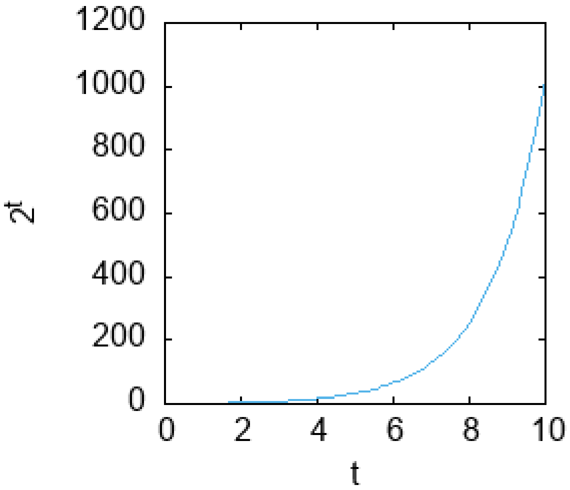
- I'm going use some math that many students haven't learned. Sorry; I need it to make a point.
- *Use calculus* (continuous time) to write $\frac{dN}{dt} = N$.
    - $\frac{dN}{dt}$ is math's way of saying "the rate of increase of N changes" at different times.
- $\frac{dN}{dt} = N$ means that the more people are infected with SARS-CoV-2, the faster (or more likely) other people get infected.
- Integrating gives *the* exponential, $N(t) = e^t$. If we know *how fast* a quantity increases at every instant of time, *integration* tells us how big it is at every instant of time.
- *With mathematics, we can* **transform** *the model to a more useful description.*

Off-topic: Exponential functions

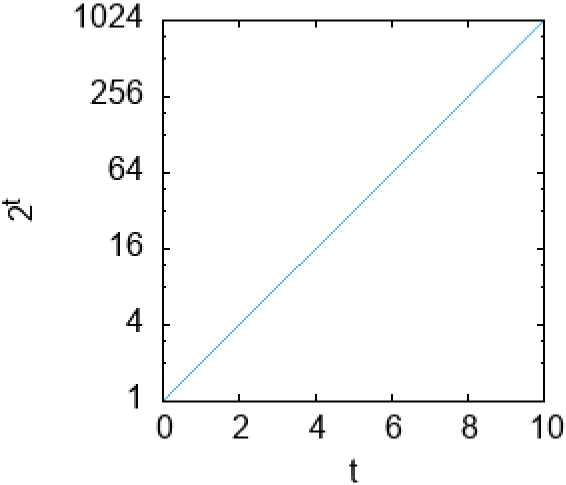
- *The* exponential e^t is *defined* by the *differential equation* $\frac{dN}{dt} = N$.
- e is an irrational number, $\approx 2.718281828459045$.
- There are many exponentials, such as 2^t (and $2^{\frac{t}{2}}$), but e^t is *the* exponential function because the differential equation is so simple.
- In fact, all exponentials can be expressed as $f(t) = Ae^{\alpha t}$, and they all have linear differential equations $\frac{df}{dt} = \alpha f(t)$.
- Although the *rate of increase* $\frac{df}{dt}$ changes over time, the *rate of growth* $\frac{df/dt}{f} = \alpha$ does not.
- Try the exponential $2^t = e^{0.693147180559945\,t}$ with some integers t!
- It's still the *same* model as Table 1. Only the math has changed.

You may want to pause the video for each of the next 4 slides.

Graph of exponential growth

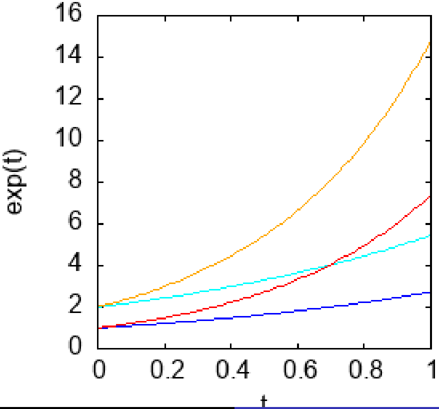


Logarithmic-scaled graph of exponential growth



More exponentials

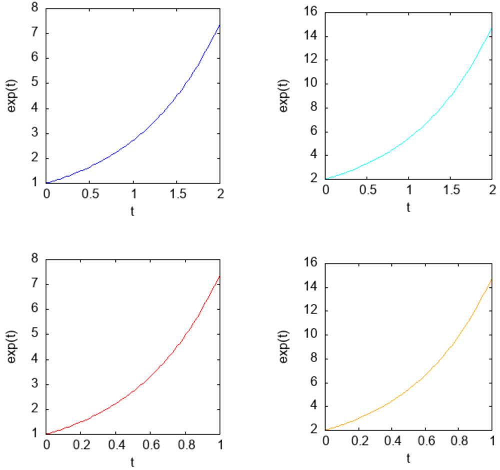
- The generic exponential function is $Ae^{\alpha t}$.
- *The* exponential function e^t has $A = \alpha = 1$.
- Here are four exponentials with $A = 1$ or $A = 2$, and $\alpha = 1$ or $\alpha = 2$.



Now, on split-screen: There's only one

Can you see the differences?

Do you know which coefficients go with which graph?

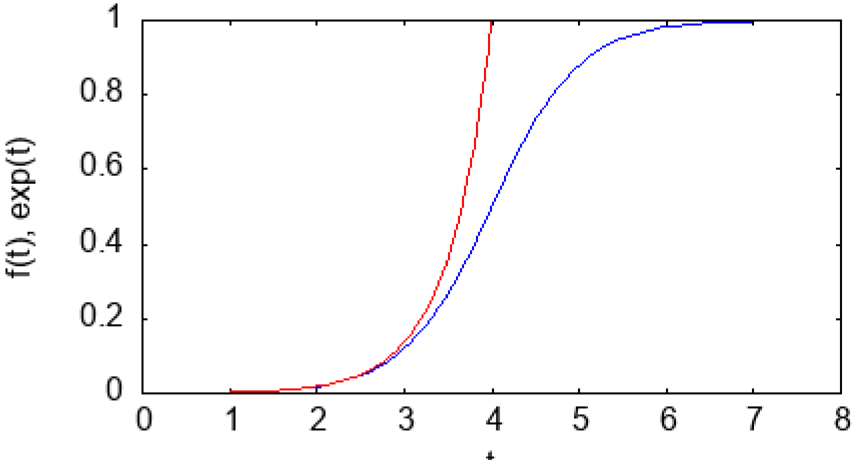


Logistic growth

- Exponential growth "crashes through" the ceiling of population because *as a model* it "ignores" the fact that only individuals who *aren't infected* can *change to* infected.
- A better model considers α as the *hazard rate*: the chance that an infected person will infect someone they meet. In the exponential model $N(t)$ infected individuals meet others, giving the rate of increase $\frac{dN}{dt} = \alpha N$.
- But already infected individuals don't "get" infected, so the chance that the individual some infected individual meets is uninfected is $\frac{\bar{N}-N}{\bar{N}}$, where $\bar{N}$ is the total population (the bar on the top symbolizes "maximum infections").
- Now the hazard rate for infections is $\alpha\frac{\bar{N}-N}{\bar{N}} = \alpha(1-\beta N)$, where $\beta = \frac{1}{\bar{N}}$, and the differential equation is $\frac{dN}{dt} = \alpha(1-\beta N)N$.
- This modification would be very tedious in discrete time. *This is what math is for, making modeling easier.*

The solution to logistic growth

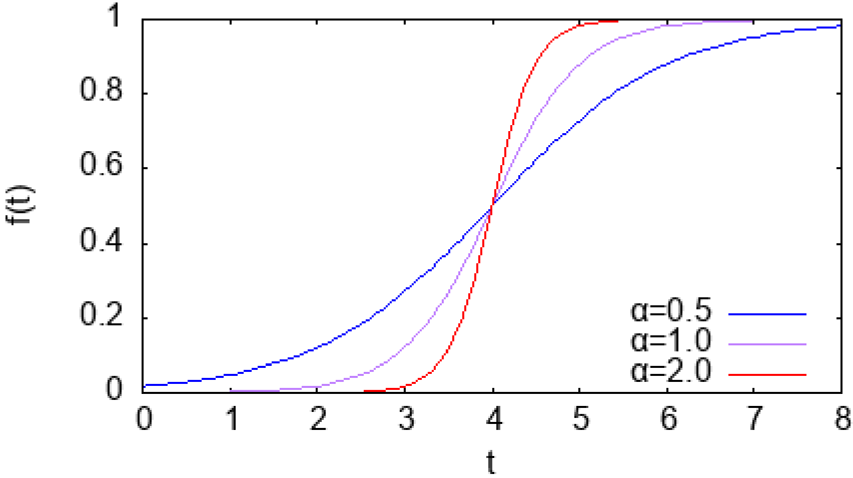
Compare logistic growth $f(t) = \frac{e^t}{e^t + e^{-t}}$ (blue) with exponential growth (red). (The exponential growth curve is rescaled to match the logistic curve.)



Policy variables and parameters

- **A policy variable is something that the authorities can change to get a better outcome.**
  - In a dynamic model like this one, we sometimes distinguish policy *parameters*, which are policy variables that do *not* change over time, from time-varying policy *variables*.
- Our hazard rate α is a sort of policy parameter. We talk about "flattening the curve" by reducing the hazard rate.
  - Why "flattening"? Because as a proportion of population, for any α the logistic curve tends to 1 as $t \to \infty$. We can't stop that but we can make the slope where $f(t) = 0.5$ smaller.

Flattening the Curve



Matching model variables to real policies

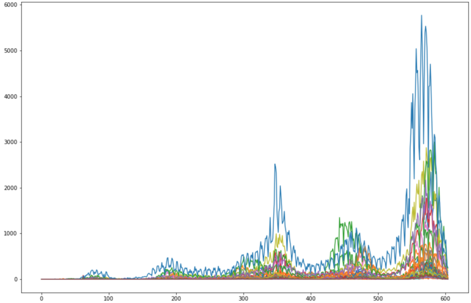
- But what are the real policies?
- Before vacccines, we didn't control the rate of infection. Instead we use *social distancing* and *filtering*.
- **For a more realistic model we can decompose a descriptive parameter $\alpha = \alpha_d \alpha_m$ into policy-relevant components** α_d, the rate of meeting people, and α_m, the leakage rate of masks.
  - Note that in this case, we can do all the math with α and substitute the decomposition later! **More realism doesn't always make the math harder.**
- We could further decompose α_m into the effectiveness *vs.* inhaling (not very good) and the effectiveness *vs.* exhaling (important).
- *The parameters of the cubic don't have such interpretations.*

Interpreting logistic growth

- Can we interpret the logistic growth model **realistically** for a disease?
- Yes! Simply define *susceptible* as “never had it,” and *immune* as “ever had it.”
- This works for measles. **But it’s not a very intuitive way to think about COVID-19.** It does *logically* generate the logistic model.
- In the discussion of hazard rate, we need to think about *why* people are removed from the susceptible population.
  You can catch the flu or a cold many times, you can be immune to measles which leads to *herd immunity* at more than 0% susceptible).

Confronting the Data

- We use models to think about everything, but especially complex phenomena.
- How do our models match up to reality?
- We *test* models by comparing them to data.
- Here are graphs of COVID-19 cases in Japan from January 2020 to date, for all 47 prefectures.



The SIR Model

- SIR is a *compartmental model*. That means the population is *partitioned* into groups called “compartments.”
- **Compartmental models are frequently a good basis for statistical analysis.**
- Sometimes our subpopulations are fixed: *e.g.*, male *vs.* female.
- Sometimes individuals move among compartments, *e.g.*, when they’re defined by age. (Age compartments are called “cohorts.”) Individuals may choose to move between compartment, as in *employment*.
- The English Wikipedia article on *Compartmental models in epidemiology* is quite good: `https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiolog`

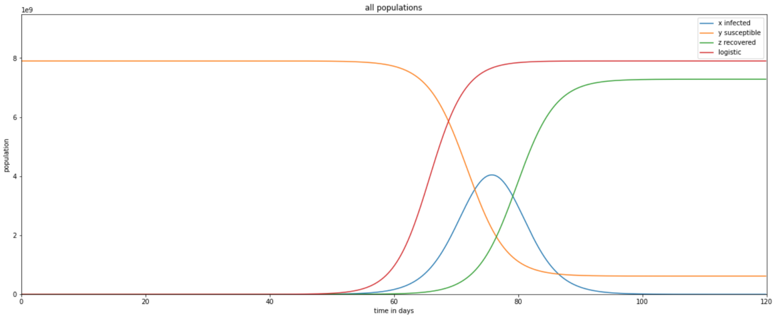
The SIR Model Equations

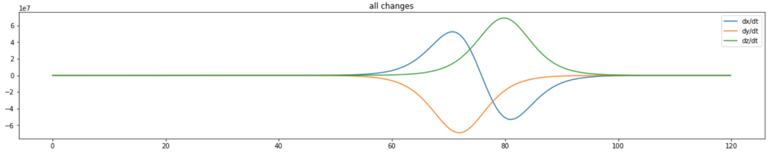
- The SIR model divides the population into three subpopulations (SIR = Susceptible, Infected, Recovered/Immune). Only the Infected population requires treatment, and only the Infected population can infect others.

$$\begin{aligned}
\frac{dN}{dt}(t) &= \alpha I(t)(1 - \frac{N(t)}{\bar{N}}) = \alpha I(t)\frac{S((t)}{\bar{N}} \\
R(t) &= N(t - 14) \\
I(t) &= N(t) - R(t) \\
S(t) &= \bar{N} - N(t)
\end{aligned}$$

- The equation $R(t) = N(t - 14)$ makes this model a *delay differential equation* model. There is no analytical integration solution, so **we use simulations**.
- Simulation is a bigger change in the modeling technology than moving from discrete doublings to the calculus of exponentials.

My Best SIR Simulation





Interpreting Logistic Growth, Part II

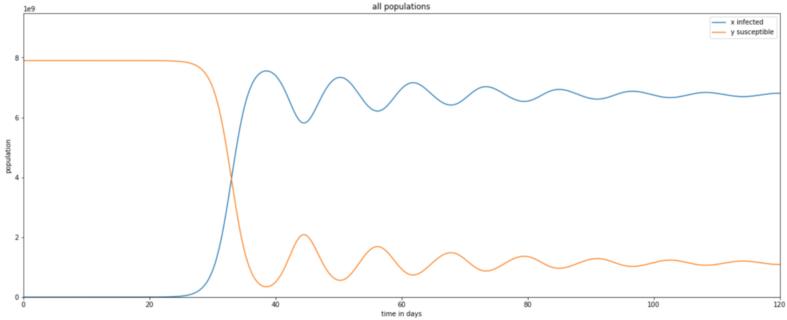
- **Note: The standard for good model is domain-specific**: "useful for public health policy." "Logically acceptable" isn't good enough!
- The logistic model's new case rate graph is similar to the SIR graph, with no skew.
- The logistic model may be the best well-founded single equation model.
- Although like the SIR model the logistic model can't explain waves, **sometimes a simple model gives a simple answer, and we're satisfied with that.** *E.g.*, the logistic model is *sufficient reason* to fear a devastating spike in cases.
- Even though cases and rate of infection don't "go through the roof," they can increase terrifyingly fast.
- SIR shows that even restricting the infectious period doesn't eliminate that.

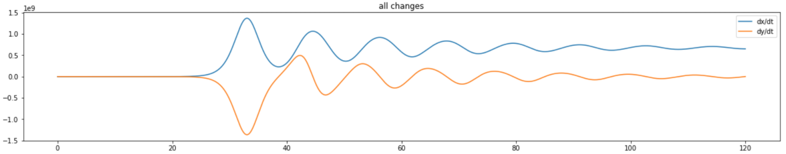
Going Beyond SIR

- *Immunity* of recovered individuals: are immediately susceptible to reinfection with the SARS-Cov-2 virus, we could use the *susceptible-infectious-susceptible* (SIS) model, which is even simpler than SIR.
- Infectious individuals may be *asymptomatic*.
  - There is an *incubation period* between the time they are infected and when they show symptoms. This leads to the *susceptible-exposed-infectious-recovered* (SEIR) model. Not very different from SIR.
  - It could also be that symptomatic individuals can recover but still be carriers (infectious).
- SARS-CoV-2 is not well understood even today, but all of the possibilities above require consideration.

None of these modifications to the compartments gives "waves."

SIS: Susceptible-Infectious-Susceptible





No Compartmental Model Is Good Enough

Public health policy requires we add other complications to the model.

- *Asymptomatic* infectious cases and *surveillance testing*.
- Infectious capacity may vary over time.
- In *pandemics* the *epidemics* are not uniform, they "break out" in *clusters*.
- *Why* people get infected (and then sick) is partly an individual characteristic. We handle that statistically by modeling it with *randomness*. We don't know enough about virus *pathology* (how it infects us), and randomness is good enough for epidemiology. (But . . .)
- In epidemiology, people are infected when they meet others. But meetings are not uniform, expressible by a single hazard rate. In fact, the probability of meeting depends on *both* individuals in an encounter. This requires a *social network model*, which generally can't be solved by algebra or calculus.

When can we stop modeling?

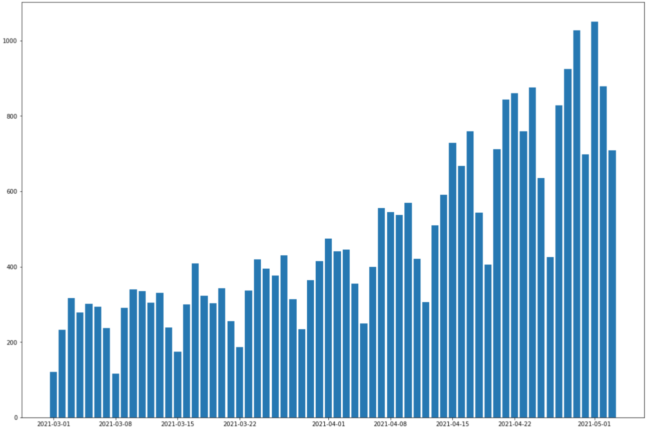
- The SIR model is good enough to give a picture that looks a lot like the graphs used to demonstrate the “flatten the curve” effect of social distancing.
- Can’t we stop there? No: for individual and economic reasons, social distancing varies greatly among individuals and across the groups they belong to.
- We need a *social network model* to *estimate* the flattening effect of a policy.
- **When do we stop modeling?** This is a pragmatic question, and the answer is “when we run out of time, budget, or motivation.”

Math Models *vs.* Japanese Data

- Unlike the *exponential* model, the number of new cases stops growing, and in fact reverses course just as steeply.
- Unlike the *logistic* model, the number of cases doesn't reach the whole population, and the new cases don't go to zero.
- Unlike *SIR*, there are repeated waves of new cases.
- Unlike *SIS*, the waves seem to be exploding, not damped.
- Unlike all the mathematical models, there seems to be a very regular high-frequency cycle around the larger movements.

A New Layer of Models

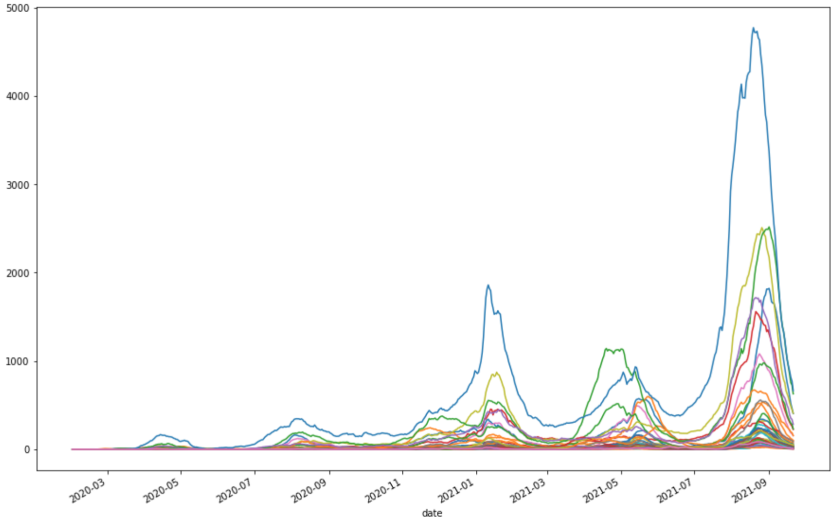
- We can't deal with *waves* today, but . . .
- . . . the high-frequency cycle clearly exists and is very regular.
- It is in fact *weekly*. **This is a statistical model.**



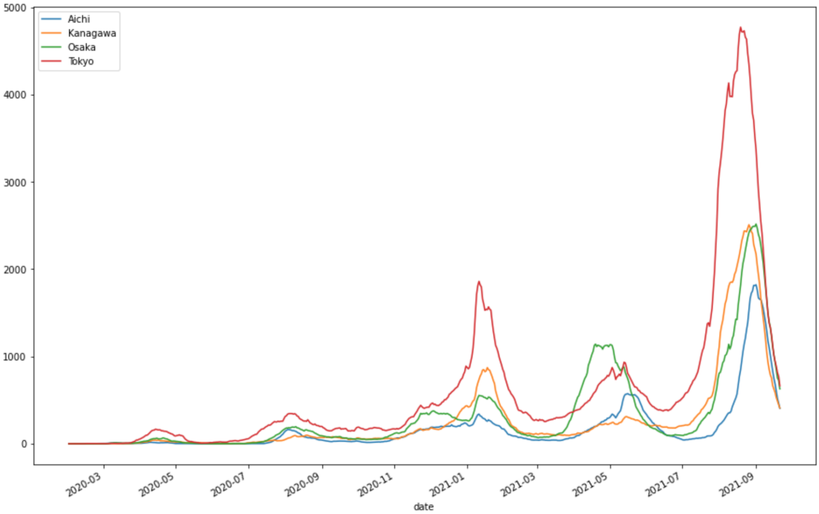
Modeling the Data: Seasonality

- Our life has weekly rhythms: weekdays *vs.* weekends, regular meetings (classes!), social, cultural, and sports events.
- People have time to go to the hospital for checkup for "worse than usual cold/flu" on weekends.
- Hospitals and labs for processing tests probably have weekly cycles, too.
- Detail explanations are unclear, but we can just take moving averages over the week: Sunday to Saturday, then Monday to Sunday, then Tuesday to Monday, and so on until we run out of data.
  - *Caveat*: it's "good enough" to smooth the graphs, but what if there is some policy variable hidden in the seasonality, such as gathering behavior?

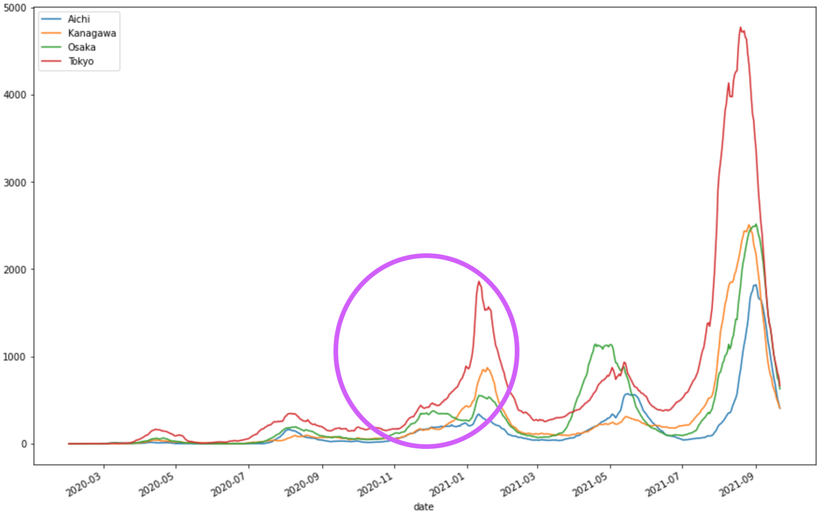
7-Day Moving Average for All Prefectures



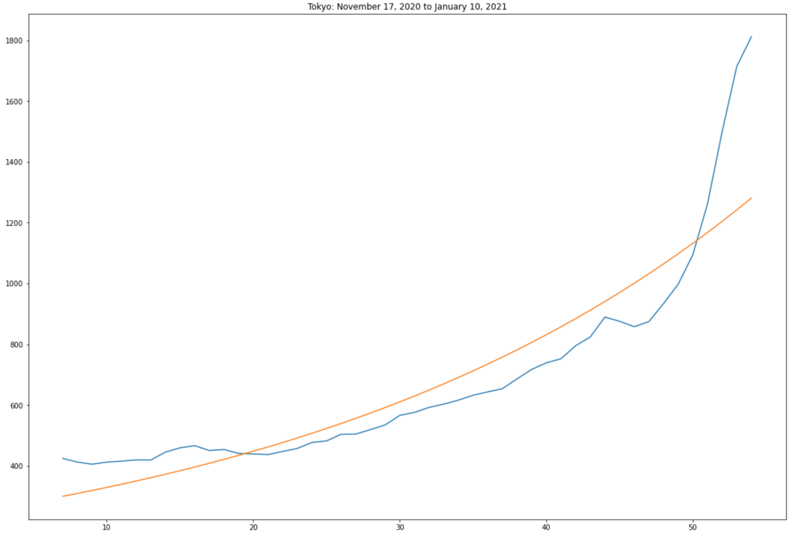
7-Day Moving Average for 4 Big Prefectures



An Exponential (?) Subsample



How Do the Exponential and Logistic Models Do?



Modeling the Data: Waves

- Why are there repeated waves?
- Even if we could get the size right (not damped), *SIS* seems implausible: we know that immunity lasts at least 6 months or so, but waves are about 4 months long.
- Perhaps prefectures are not the right unit of collection, but *social networks*. An epidemic can break out in one network, but it's difficult to jump to another.
- When it *does* make the jump, another wave starts.
- That point is a *model*.
  It's not mathematical, but I think we need to make it mathematical (and simulate it) to understand the pandemic.
  - A "simple" case: consider each prefecture a social network. The data we have allows us to simulate the pandemic "jumping" from one prefecture to another.

Coda

What we have learned:

- We think with *models*.
- This is a trivial statement, because the world contains our brains—our brains do not contain the world. What we *can* think is less complex than reality, it is a *representation*.
- Mathematics helps us transform models (it helps us *think*). It is a standard way to express certain kinds of models.
- Mathematical models like exponentials, logistic curves, and SIR help us understand *some* but *not all* aspects of the COVID-19 pandemic.
- Models come in *layers*: the *moving average* from statistics helps simplify (= create a model of) volatile data to *test* (compare to) our mathematical models.
- It's **models all the way down!**

Data Credits

COVID Data by Japanese Prefecture is public domain data organized and published frequently (not quite daily) by the Toyo Keizei publishing company.

- **source** `https://toyokeizai.net/sp/visual/tko/covid19/csv/`
- **files** `cases_total.csv`, `death_total.csv`, `demography.csv`, `effective_reproduction_number.csv`, `pcr_case_daily.csv`, `pcr_positive_daily.csv`, `pcr_tested_daily.csv`, `prefectures.csv`, `recovery_total.csv`, `severe_daily.csv`
- **format** comma separated values
  - Header strings are not quoted.
  - Numbers with embedded commas are quoted.
- **notes** prefecture.csv data is derived from prefecture press releases. All other files are derived from Ministry of Health, Labor and Welfare press releases.

Image Credits

- All images are from Wikipedia's *Wikimedia* site, and have *free public licenses* or are in the public domain.
- All graphs were generated with Python, using Pandas, Numpy, and Matplotlib.