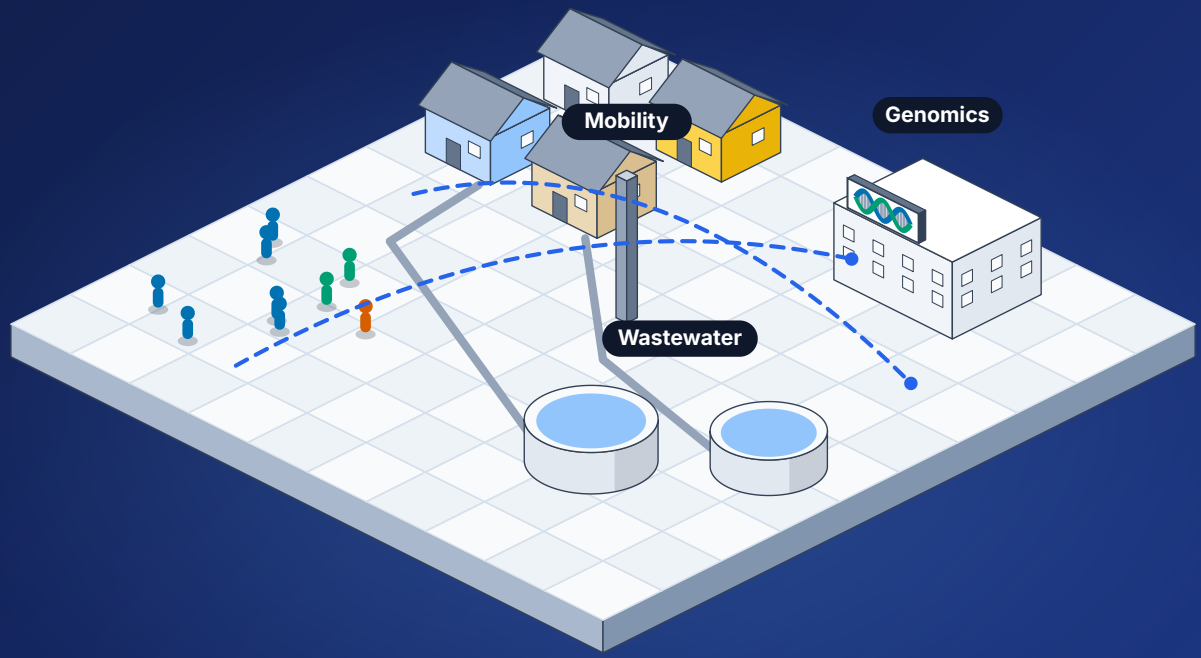
Applied Computational Mathematics Lab

EPIDEMIC MODELING · MODULE 9

Surveillance, mobility and genomic data

What the data streams see, and what they miss



3 EPISODES

1. Surveillance as measurement, and mobility data
2. Genomics: from sequences to transmission
3. Wastewater, syndromic and pharmacy signals

Based on Chapter 9 of T. Lazebnik, Epidemic Modeling: Principles, Algorithms, and Applications (Academic Press, 2026)

acml.teddylazebnik.com

BEFORE WE START

Warm-up: true or false?

CHECK YOURSELF

(a) When reported cases fall, infections are falling.

(b) The share of a random sample testing positive for antibodies estimates the share ever infected.

(c) If one variant is always three times as likely to be sequenced, its estimated growth advantage is too high.

(d) Wastewater can show an epidemic turning before reported cases do.

HOW TO USE THIS

Guess now; the module settles each one. Answers at the end.

WHY THIS MODULE

Every model in this course meets data through a measurement process. This module looks at the streams themselves: what each one measures, with what delay and bias, and how to combine them.

BEFORE WE START

This module's map

PART 1

Surveillance and mobility

Cases as a measurement of infections, serosurveys and test errors, phone-based mobility data.

PART 2

Genomes

Sequencing as an observation model, lineage frequencies, sampling bias, what trees add.

PART 3

Beyond the clinic

Wastewater, syndromic and pharmacy signals, and combining several streams.

GOALS

By the end you should be able to write the observation model behind a data stream, correct a serosurvey for test errors, judge mobility and genomic data by how they were collected, and combine streams to follow one hidden epidemic.

ONE RULE

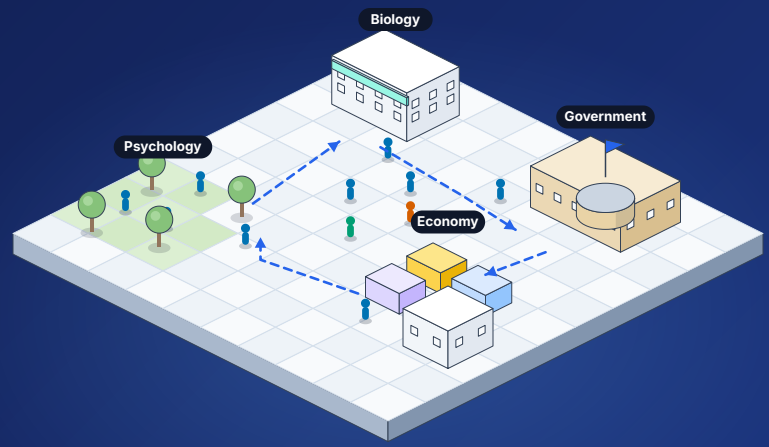
Every number in this deck comes from one **simulated city** of a million people with a known latent epidemic, seen through simulated data streams. All values are made up to show mechanisms.

Applied Computational Mathematics Lab

PART 1

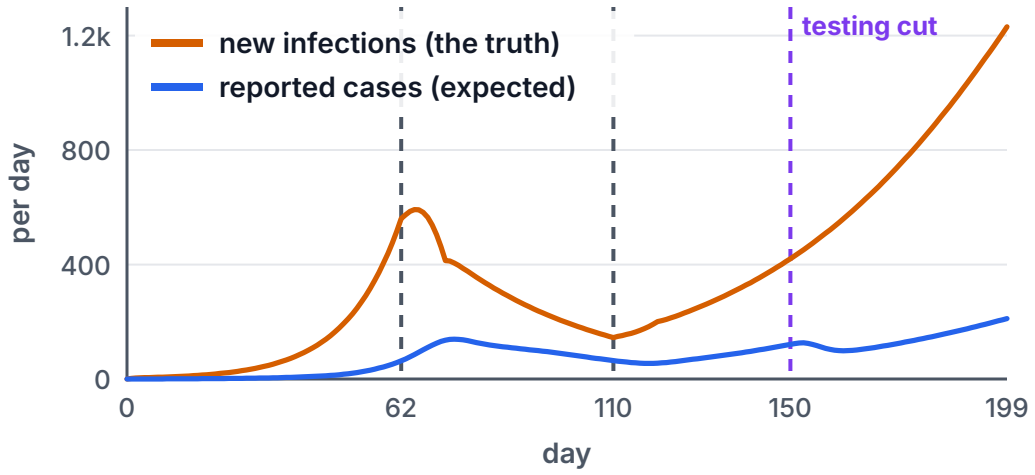
Surveillance and mobility

What a case count really measures.



PART 1 · MEASUREMENT

Cases are a measurement of infections



Our city: 71,317 infections over 200 days; lockdown from day 62, relaxation from day 110, testing cut on day 150.

$$\mathbb{E}[\text{cases}_t] = \sum_{d \geq 0} \rho_{t-d}\, I_{t-d}\, g_d$$

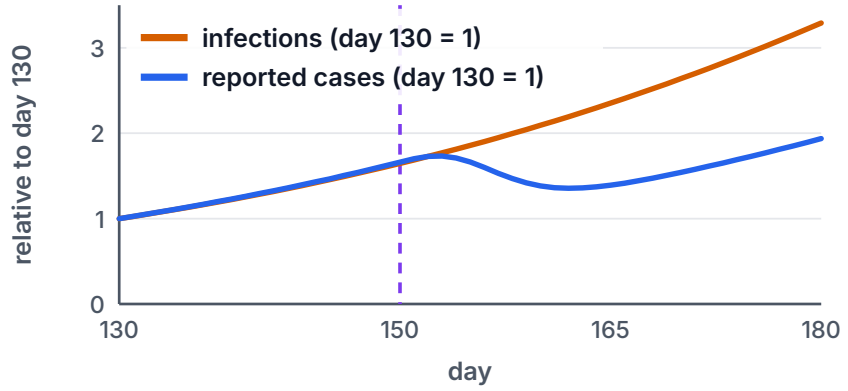
- I_t: new infections; ρ_t: the share that will be reported (testing); g_d: the delay from infection to report.
- Here the delay averages 8 days (about 5 to symptoms, 4 more to a report); the share rose from 15% to 35%, then fell to 20%.
- Overall only 22% of infections became cases, and the case peak came 9 days after the infection peak.

IN PLAIN WORDS

A model of infections must be joined to the data through this observation model; otherwise a change in testing looks like a change in transmission.

PART 1 · MEASUREMENT

When testing changes



Our city around the testing cut on day 150: infections and expected reported cases, each relative to day 130.

WHAT AN ANALYST WOULD SEE

A growth rate fitted to reported cases in the week after the cut says the epidemic is shrinking (-1.2% a day), while infections grow +2.4% a day.

THE WEEK AFTER DAY 150	GROWTH PER DAY
infections	+2.4%
reported cases	-1.2%

Our run: for 11 days after the cut, reported cases fell while infections rose.

- A change in the reporting share multiplies the cases from then on: growth rates computed from cases are wrong until the old and new shares are no longer mixed in the window.
- Signs to watch: changes in test numbers and in test positivity. Rising positivity with flat cases suggests more infections are going unseen.
- Fix it in the model: let the reporting share change with testing data, or use streams that do not depend on testing.

TRY IT

Reported cases are flat while the share reported halves over the same two weeks. What happened to infections?

Other streams, other filters

STREAM	MEASURES	TYPICAL DELAY AFTER INFECTION	MAIN DISTORTION
reported cases	infections that were tested	about a week	testing policy, access, behaviour
test positivity	share of tests positive	about a week	who gets tested
hospital admissions	severe infections	one to two weeks	age mix, capacity, criteria
deaths	fatal infections	two to four weeks	age mix, attribution, reporting
serosurveys	past infection	two weeks or more	test accuracy, waning antibodies, sampling
wastewater	shedding in a catchment	days	dilution, shedding variation

A qualitative guide: the delays are typical orders of magnitude, not estimates for any pathogen.

IN PLAIN WORDS

Every stream is the latent epidemic passed through its own filter: a share of infections, a delay, a scale, noise, and a bias. Writing that filter down is the observation model (Module 7).

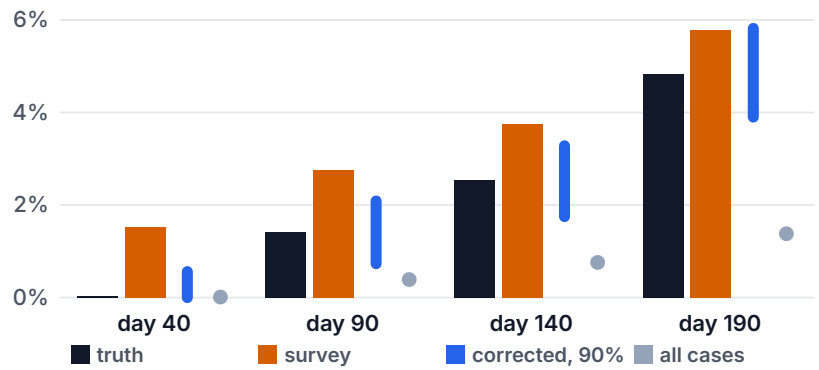
LINE LISTS HELP

Line lists (one row per case, with onset, test and report dates) let you estimate the delay distribution directly. Aggregated counts hide it, and it must then be assumed or fitted.

Serosurveys and imperfect tests

$$p_{\text{obs}} = (1 - \text{sp}) + (\text{se} + \text{sp} - 1)\, p, \qquad \hat{p} = \frac{p_{\text{obs}} + \text{sp} - 1}{\text{se} + \text{sp} - 1}$$

- Our surveys: 2,000 people, sensitivity 90%, specificity 98.5%; people become positive 14 days after infection.
- At low prevalence the false positives dominate: on day 40 a survey would show 1.5% positive when only 0.03% had been infected.
- The correction (Rogan and Gladen) removes the bias, but its uncertainty is larger than the raw binomial one, and it needs good estimates of se and sp.



Our surveys: truth, expected positive share, the range in which 90% of corrected estimates fall, and all reported cases so far (share of the population).

SURVEY DAY	TRUTH	SURVEY	CORRECTED (90%)	ALL CASES
40	0.03%	1.53%	0.0%–0.6%	0.01%
90	1.41%	2.75%	0.7%–2.1%	0.39%
140	2.54%	3.75%	1.8%–3.3%	0.76%
190	4.83%	5.78%	3.9%–5.8%	1.38%

What serosurveys can and cannot say

- On day 190 the corrected surveys centred on the truth (4.8%), while all reported cases added up to only 1.4% of the population: about 3.5 infections for every reported case.
- Antibodies wane: positivity can fall even after infections stop, so a survey measures recent enough infection, not all infection. Model the waning or use assays that last.
- Who is sampled matters: blood donors and clinic patients are not a random sample. Weighting by age and place helps; it cannot fix groups that were never sampled.
- Vaccination also creates antibodies: distinguish infection from vaccination with the right antigen, or model both.

TRY IT

A survey of 1,000 people finds 40 positive. With sensitivity 0.85 and specificity 0.97, what is the corrected estimate of the share ever infected?

WHY THEY MATTER

Serosurveys anchor the scale of an epidemic: they turn “cases” into “infections” and fix the reporting share that case data alone cannot identify (Module 8: only the product of share and size is identified from cases).

A SOBERING EXAMPLE

With specificity 98.5%, 1.5% of a never-infected population tests positive. A survey that finds 1.5% positive is therefore consistent with nobody having been infected.

PART 1 · MOBILITY

Mobility data, and who they count

THREE TOWNS (MADE UP)	PEOPLE	PHONES OF THE DATA PROVIDER
city	1,000,000	30%
suburb	500,000	30%
small town	100,000	8%

TOWN TO CITY TRIPS A DAY	
true	1,000
from phones, scaled by the national share (28.6%)	279

THE OUTBREAK STARTS IN THE SMALL TOWN	DAY THE CITY PASSES 1 IN 10,000 INFECTIOUS
true flows	27
phone flows	32

Our runs: a three-town model where residents spend the day where they travel (beta 0.3, gamma 0.1).

IN PLAIN WORDS

Mobility data describe people. Aggregate, add noise where needed, and use them only for the stated purpose.

- Two mechanisms: travel imports infection into new places, and movement changes who mixes with whom.
- Phone data count only the phones of one provider; coverage differs by place, age and income. Scaling by one national share undercounts the poorly covered towns: here five days of warning that do not exist.
- Platform mobility indices (visits by place type) measure movement, not contacts. The link between a drop in visits and a drop in transmission has to be estimated.
- How trips are defined and aggregated (pings, dwell times, zones) is part of the model, not a neutral step.

Check yourself: Part 1

CHECK YOURSELF

1. Write the expected reported cases on day t in terms of infections, the reporting share and the delay distribution.

2. A test has sensitivity 0.9 and specificity 0.99. True prevalence is 2%. What share tests positive, and what does the uncorrected survey overstate?

3. Why can scaling phone-based trips by one national market share bias a travel model? Give the direction for a town with low coverage.

TIP

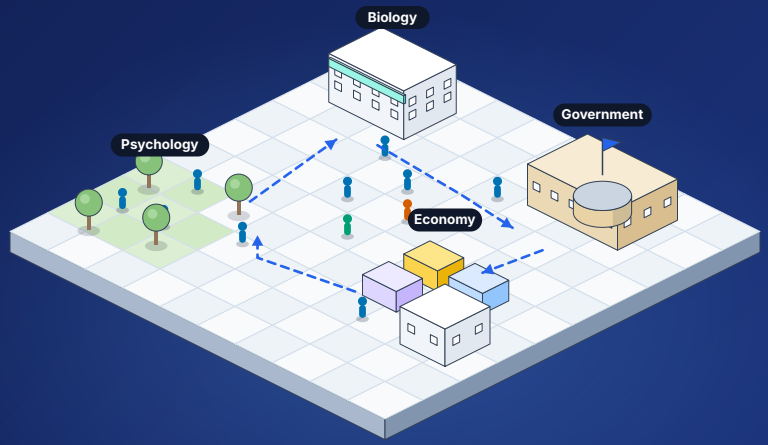
Answers are at the end of the deck. Try them before you look.

Applied
Computational
Mathematics
Lab

PART 2

Genomes

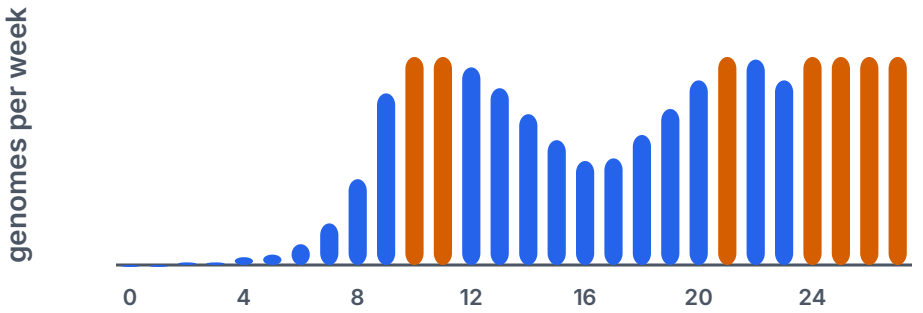
How sequences reveal variants, and how sampling shapes them.



Sequencing is an observation model

$$n_t \sim \text{Poisson}(\kappa\, \pi_t\, I_t), \qquad (m_{1,t}, \ldots, m_{V,t}) \sim \text{Multinomial}\big(n_t;\ \tilde{\pi}_{1,t}, \ldots, \tilde{\pi}_{V,t}\big)$$

- n_t: genomes sequenced; κ: lab capacity; π_t: the chance an infection is sampled; $m_{v,t}$: genomes of lineage v.
- From sample to lineage: reads, alignment to a reference, a consensus genome, and a lineage label. Low viral load or poor coverage can drop samples, and which ones drop is not random.
- In our city labs sequence 10% of cases, at most 80 a week. For 7 weeks the cap was binding; in the busiest week only 6% of cases were sequenced.



Our genomes per week (orange: at the cap). Genome counts follow capacity, not the epidemic.

IN PLAIN WORDS

Use genomes for what they measure well: the mix of lineages among sampled infections, and ancestry. Do not read epidemic size from the number of genomes.

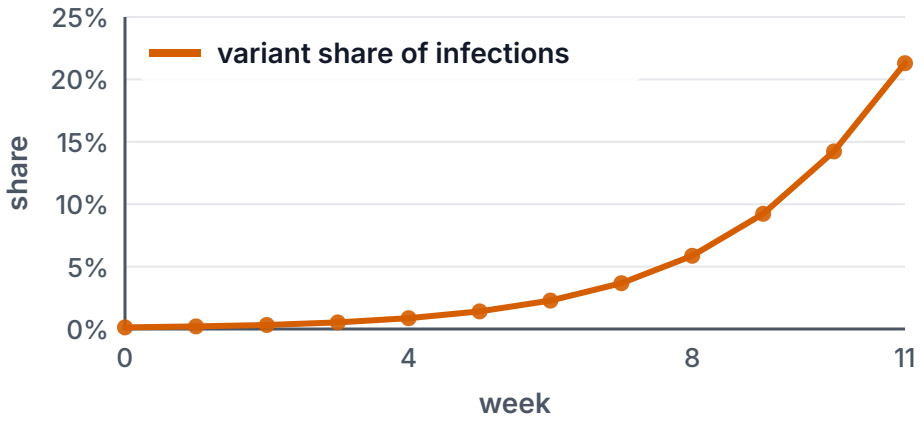
Lineage frequencies and a new variant

$$\operatorname{logit} \pi_V(t) = \operatorname{logit} \pi_V(0) + s\,t$$

- A variant growing s = 0.07 per day faster (on the log odds of its share), starting at 1 in 1,000. After 8 weeks it is still only 4.8% of infections.
- Weekly random samples of cases; we estimate s by logistic regression on the first 8 weeks, and repeat the whole experiment 1,000 times.

GENOMES A WEEK	FIRST VARIANT GENOME (MEDIAN, 10–90%)	ESTIMATED S (90% RANGE)
50	day 31 (10–45)	0.01–0.41
500	day 10 (3–17)	0.05–0.09

Our simulations: binomial sampling of the true share each week.



The true variant share over 12 weeks.

TRY IT

With 50 genomes a week and a 1% variant share, what is the chance of seeing at least one variant genome in a week?

Biased sampling

$$\tilde{\pi}_V = \frac{b\,\pi_V}{b\,\pi_V + (1-\pi_V)} \;\Rightarrow\; \operatorname{logit}\tilde{\pi}_V = \operatorname{logit}\pi_V + \log b$$

- If the variant is b times as likely to be sequenced (travelers sequenced first, targeted testing), its observed share is too high.
- A constant b only shifts the log odds: the slope, the growth advantage, is unchanged. A b that changes over time adds a trend to the log odds and biases the slope.

OUR RUNS (500 GENOMES A WEEK, 500 REPEATS)	SHARE IN WEEK 7	ESTIMATED S
random sampling	3.7% (truth)	0.071
variant sequenced 3 times as often	10.4%	0.070
the same, starting in week 4	–	0.090

Truth: s = 0.07.

IN PLAIN WORDS

Record how samples were chosen (random surveillance, travelers, outbreaks, severe cases), keep the scheme stable, or model the bias.

TRY IT

A lab starts sequencing every sample with a particular test failure (which the variant causes) halfway through. Which way will the estimated advantage move?

What family trees add

INTRODUCTIONS

Counting arrivals

Separate clusters of closely related genomes point to separate introductions; one cluster points to local spread.

DATING

When did it start?

With a molecular clock (mutations accumulating at a roughly steady rate), a tree can date when a lineage began.

PHYLODYNAMICS

Growth from trees

The branching pattern of a well-sampled tree carries information on growth and on the reproduction number.

IN PLAIN WORDS

All three depend on sampling: a tree drawn from samples of one hospital or of travelers tells the story of those samples.

WHY GENOMES

Genomic surveillance answers questions counts cannot: how many introductions, whether a resurgence is local or newly seeded, and whether a lineage, not behaviour, is driving growth.

READING A TREE

Two clusters of genomes that differ by many mutations, both appearing in the same city in the same week, are more likely two arrivals than one chain of local spread.

Check yourself: Part 2

CHECK YOURSELF

4. Why is the number of genomes per week a poor measure of epidemic size?

5. Show that a constant sampling bias b changes the intercept but not the slope of the log odds of a variant’s share.

6. Ten times more genomes a week: roughly how much narrower should the uncertainty of the estimated advantage get, and why?

TIP

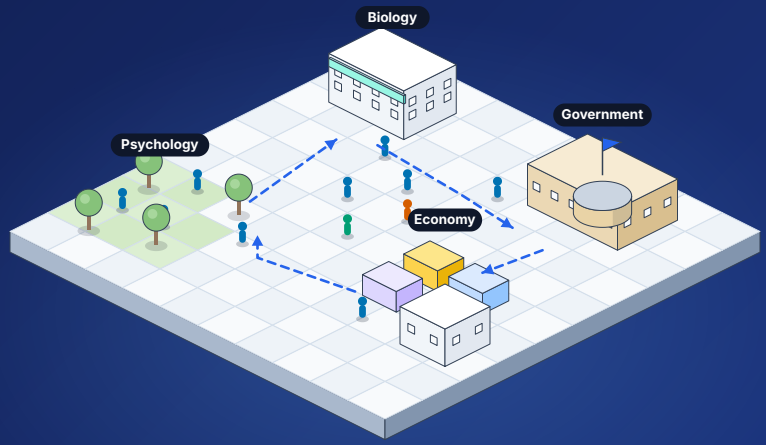
Answers are at the end of the deck.

Applied Computational Mathematics Lab

PART 3

Beyond the clinic

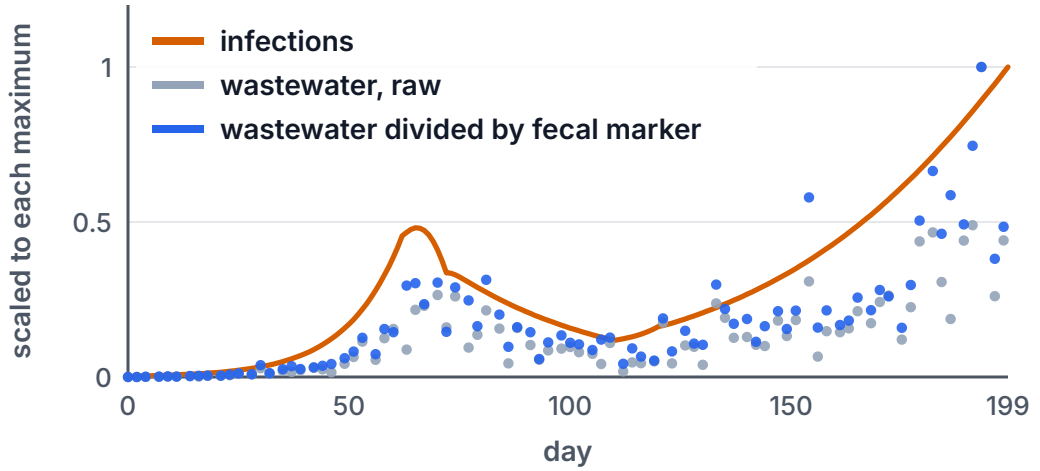
Wastewater, syndromic and pharmacy signals, combined.



Wastewater: shedding, diluted by the flow

$$c_t = \frac{q \sum_{s \geq 0} h_s \, I_{t-s}}{F_t} \times \varepsilon_t$$

- h_s: shedding s days after infection (here a made-up curve with mean 6 days); F_t: flow (more on rainy days); ε_t: measurement noise.
- Our plant: samples on Monday, Wednesday and Friday; noise with a 35% spread; 27 rainy days that dilute the sewage.
- Dividing by a fecal marker diluted the same way corrects for the flow: the error of the 14-day growth rate went from 0.042 to 0.039 per day.



Our city: infections and wastewater samples, each scaled to its maximum.

IN PLAIN WORDS

Wastewater sees everyone connected to the sewer, tested or not; it does not see people outside the catchment, and shedding varies between people and variants.

PART 3 · WASTEWATER

Earlier, and blind to testing

THE FIRST WAVE TURNS (14-DAY GROWTH NEGATIVE FOR A WEEK)	DAY
infections	72
wastewater, divided by the marker	74
wastewater, raw	77
wastewater without noise or rain	77
reported cases	83

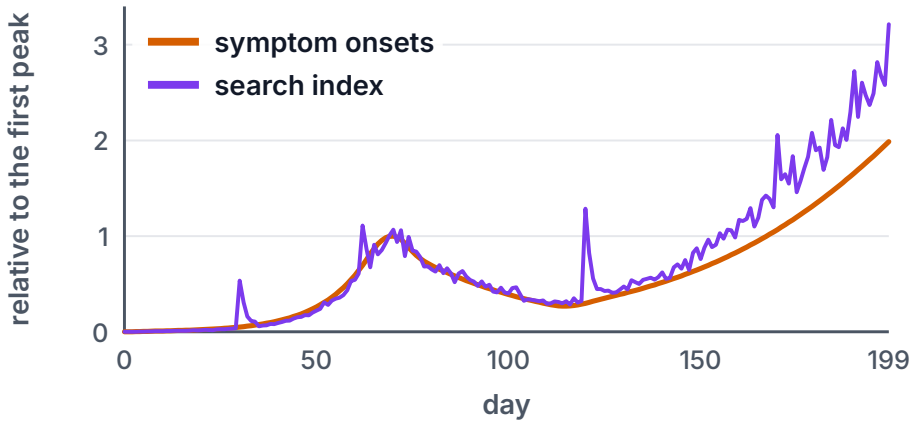
DAYS 152–165, AFTER TESTING IS CUT	GROWTH PER DAY
infections	+2.4%
wastewater (divided by the marker)	+1.7%
reported cases	+0.5%

- Wastewater lags infections by the shedding delay (here a few days) but leads cases, which wait for symptoms, a test and a report.
- Noise can move the apparent turn by days in either direction: even here the noise-free signal would have turned on day 77.
- Because it does not depend on who gets tested, wastewater kept the right sign when testing was cut, while cases nearly flattened.

TRY IT

Shedding peaks 3 days after infection and reports come 8 days after infection. Roughly how much earlier can wastewater show a turn?

Syndromic and pharmacy signals



Our search index (symptoms, media spikes and a slow drift) against symptom onsets.

SEARCH INDEX	
correlation with onsets, days 0–99	0.97
correlation with onsets, days 100–199	0.98
day 199 relative to the first peak, by the index	2.8
the same, true onsets	2.0

A high correlation can hide a drifting scale.

- Syndromic data (fever visits, influenza-like illness reports, searches) are fast but not specific, and they drift with media attention and with changes in how people search; Google Flu Trends famously overestimated the 2012–13 flu season (Lazer et al., 2014).
- Pharmacy sales and absenteeism are broad but mixed: at the first peak only 28% of our cold-medicine sales came from this epidemic; the rest came from other illnesses.

PART 3 · INTEGRATION

One hidden epidemic, many streams

$$I_t = R_t \sum_{s \geq 1} w_s \, I_{t-s}, \qquad y_t^{(k)} \sim p_k\Big(\cdot \;\Big|\; \alpha_k \sum_d g_d^{(k)} \, I_{t-d} \Big), \quad k = 1, \ldots, K$$

- One latent process (a renewal equation or a compartmental model) and one observation equation per stream: its own delay $g^{(k)}$, scale α_k and noise p_k.
- Each stream then pulls on the estimate according to its precision; a stream that changes its scale (testing) can be given a time-varying α_k.
- Fitted with the tools of Modules 7 and 8: Kalman or particle filters, or Bayesian samplers.

TRY IT

Reported cases rise 2% a day while wastewater falls 1% a day. Name two explanations, and a third stream that could tell them apart.

GROWTH-RATE ERROR PER DAY (LAG-CORRECTED)

reported cases	0.0215
wastewater (divided by the marker)	0.0384
combined (76% cases, 24% wastewater)	0.0197

Our simplest combination: an inverse-variance average of the two growth estimates, days 40–199.

IN PLAIN WORDS

No stream is the epidemic. Each is a view, with its own blind spots; good models combine the views and say how much each one is trusted.

READING THE TABLE

Here wastewater alone was the noisier estimate (three samples a week, rain, shedding noise), yet adding it with a small weight still brought the error below that of cases alone.

PART 3 · PRACTICE

Designing a surveillance mix

STREAM	TIMELINESS	SPECIFICITY	COVERAGE	COST PER SIGNAL
reported cases	medium	high	those tested	medium
hospital admissions	slow	high	severe cases	low
serosurveys	slow	high	a sample	high
genomes	medium	very high (lineage)	a sample	high
wastewater	fast	high (pathogen)	a catchment	low per person
syndromic, searches	fast	low	broad	low
pharmacy, absenteeism	fast	low	broad	low

A qualitative comparison to start a design discussion; local systems differ.

- Pair fast, unspecific streams with slow, specific ones: the first raise an alarm, the second confirm and scale it.
- Keep sampling schemes stable and documented: most of the biases in this module came from changes in who was sampled or tested.
- Protect people: aggregate, minimise, and govern access to mobility, genomic and health data.

TRY IT

Your city can afford one new stream. Which would you add to reported cases, and what question would it answer that cases cannot?

Check yourself: Part 3

CHECK YOURSELF

7. Why does dividing a wastewater concentration by a fecal marker help on rainy days?

8. A search index correlates at 0.97 with true incidence, yet overstates a second wave by 40%. How can both be true?

9. Two growth-rate estimates have error variances 1 and 4. What weights does an inverse-variance average give them, and what is the variance of the combination (if their errors are independent)?

TIP

Answers are on the next slides.

ANSWERS

Answers: warm-up and Parts 1–2

- **(a)** False: after the testing cut, cases fell for 11 days while infections rose. **(b)** False at low prevalence: false positives dominate unless corrected. **(c)** False: a constant bias shifts the share, not the advantage (a changing bias does bias it). **(d)** True: day 74 against day 83 here.
- **1.** $\mathbb{E}[\text{cases}_t] = \sum_d \rho_{t-d} I_{t-d} g_d$ (or with the share applied at report time, depending on the system).
- **2.** $p_{\text{obs}} = 0.01 + 0.89 \times 0.02 = 0.0278$: 2.8% positive against 2% infected, about 40% too high.
- **3.** Trips from a town with low coverage are scaled up too little, so its outgoing flows are undercounted: an epidemic starting there reaches other places later in the model than in reality.
- **4.** It follows lab capacity and sampling choices (here capped at 80 a week), not the number of infections. **5.** $\operatorname{logit} \tilde{\pi} = \operatorname{logit} \pi + \log b$; with logit π = a + st this is (a + log b) + st: same slope.
- **6.** About $\sqrt{10}$, so roughly 3 times narrower: the standard error of a binomial share falls like $1/\sqrt{n}$.

TRY-IT ANSWERS

Try-it answers: halving the share with flat cases means infections roughly doubled. Corrected: (0.04 + 0.97 − 1) / (0.85 + 0.97 − 1) = 1.2%. At least one variant genome in 50 at 1%: $1 - 0.99^{50}$ = 39%. Sequencing every sample with the variant's test failure from mid-way raises the observed share faster than the truth: the estimated advantage goes up.

Answers: Part 3

- **7.** Rain dilutes the virus and the marker alike; their ratio cancels the dilution, leaving the signal per unit of human waste.
- **8.** Correlation measures how well the shapes line up, not whether the scale is constant: a slow multiplicative drift keeps the correlation high while inflating later values.
- **9.** Weights 1/1 and 1/4, normalised: 0.8 and 0.2. Variance of the combination: 1 / (1 + 1/4) = 0.8, smaller than either.

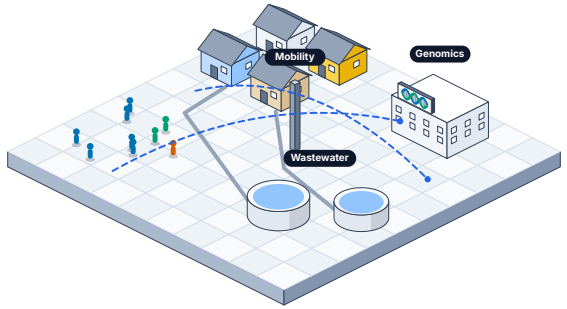
TRY-IT ANSWERS

Try-it answers: wastewater can show a turn about 5 days earlier (8 − 3), less once noise forces you to wait for confirmation. Cases up, wastewater down: more testing (a rising reporting share) or a sampling or flow problem at the plant; test positivity or hospital admissions can tell them apart. One new stream: for example wastewater (fast, independent of testing; answers whether infections are rising when testing changes) or a serosurvey (answers how many have been infected, and so the reporting share).

CREDITS

Sources and credits

- This deck accompanies Chapter 9 of T. Lazebnik, Epidemic Modeling: Principles, Algorithms, and Applications (Academic Press, 2026), the course textbook. It follows the chapter's topics in our own words; no text, figures, tables, exercises or code are copied from the book.
- The text, figures, examples and simulations were written, drawn and run for ACML; nothing is copied from the book. Numbers labelled "our run" come from our own code; counts from agencies and papers are credited where they appear.
- All numbers and curves come from our own simulated city and data streams (production/explains/sims/epi_m09_sims.py), with made-up values chosen to show mechanisms, not to describe any real pathogen, place, test or surveillance system.
- Named works, described in our own words: the Rogan–Gladen correction (Rogan and Gladen, 1978); Google Flu Trends (Ginsberg et al., 2009) and its overestimate (Lazer et al., 2014).
- Companion videos: **Epidemic Modeling**, module 9, episodes 38–40. More from the lab: acml.teddylazebnik.com



Applied Computational Mathematics Lab

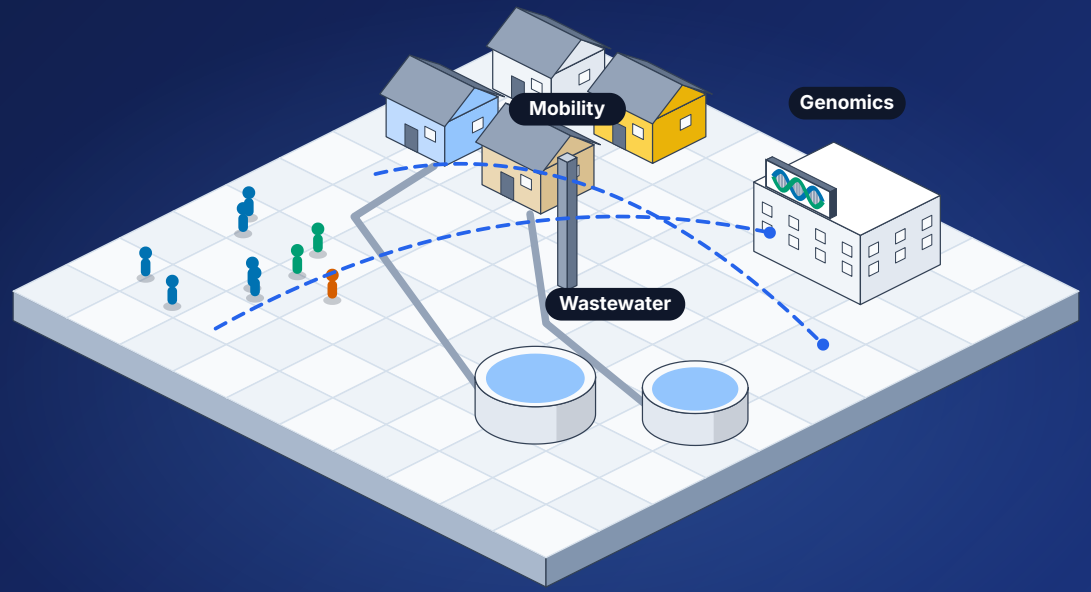
Module 9 complete

END OF MODULE 9

Surveillance, mobility and genomic data

YOU COVERED

- Surveillance as measurement, and mobility data
- Genomics: from sequences to transmission
- Wastewater, syndromic and pharmacy signals



UP NEXT · MODULE 10

Intervention policies

Interventions, optimal control and vaccination

Based on Chapter 9 of T. Lazebnik, Epidemic Modeling: Principles, Algorithms, and Applications (Academic Press, 2026)