GABRIEL SANCHEZ-MARTINEZ: Great. OK. So today's lecture is on transit service reliability. This is my penultimate lecture, because we have guest lecturers, too. So let's start out with what is reliable transit for you. What do you think of when you think of reliable transit? What does it mean for you? And I'll write a few videos on the board.

[CHALK ON BLACKBOARD]

[INAUDIBLE]

AUDIENCE: Arrival at or close to scheduled arrival.

GABRIEL SANCHEZ-MARTINEZ: OK. Low variance, OK, of-- there's a word for that. Punctuality, right? So let's use that. Just because it's hard to write on the board with chalk. It's hard to spell correctly, actually, when you're close up to the board. Emily.

AUDIENCE: Lack of disruption. So it's the same service every weekday.

GABRIEL SANCHEZ-MARTINEZ: And you kind of got two of them. Lack of disruptions is one of them and you said the same service, right? So I guess predictability. Great.

AUDIENCE: The word legibility, which is that you sort of know where transit is going and it makes sense.

GABRIEL SANCHEZ-MARTINEZ: You called it legibility?

AUDIENCE: Legibility.

GABRIEL SANCHEZ- OK. That's coining a new term there.

MARTINEZ:

AUDIENCE: Well, yeah. It's coining a new term.

GABRIEL SANCHEZ-MARTINEZ: You had one.

AUDIENCE: Oh, not in the US.

[LAUGHTER]

GABRIEL SANCHEZ-MARTINEZ: [INAUDIBLE]. I appreciate the humor, though. Any other ideas?

AUDIENCE: Maybe the quality of the ride is always the same, like tables are clean.

GABRIEL SANCHEZ-MARTINEZ: Consistent quality.

AUDIENCE: Yeah. It's always possible to enter the car. Surrounding is consistent.

GABRIEL SANCHEZ-MARTINEZ: Sufficient capacity. Actually.

AUDIENCE: I think it also has to cover all the places you want to go in like a 90-- some kind of percentile of what you want to travel, so you're not getting a car [INAUDIBLE] or something.

GABRIEL SANCHEZ-MARTINEZ: Yep. And so that includes [INAUDIBLE] coverage, which is spatial, and also span of service, right, which is the same thing but temporal.

[WRITING ON BOARD]

OK, these are some good ideas.

AUDIENCE: I have a real--

GABRIEL SANCHEZ-MARTINEZ: Sorry?

AUDIENCE: I said I have a real one now. Consistent fare structure or--

GABRIEL SANCHEZ-MARTINEZ: OK. That's usually consistent. Are you thinking of a particular example?

AUDIENCE: I think of just when places where the price keeps rising often.

GABRIEL SANCHEZ-MARTINEZ: OK. I'll say-- I don't know.

AUDIENCE: [INAUDIBLE] transfers [INAUDIBLE] confusing.

GABRIEL SANCHEZ-MARTINEZ: Good communication of fare policy, so a clear fare policy.

AUDIENCE: Sure.

GABRIEL SANCHEZ-MARTINEZ: Yeah.

[WRITING ON BOARD]

Eli.

AUDIENCE: I feel like we haven't worked in maintenance of headways and high-frequency service in here.

GABRIEL SANCHEZ-MARTINEZ: Sure.

AUDIENCE: It's kind of covered by punctuality and predictability--

GABRIEL SANCHEZ-MARTINEZ: Right. But it's a little different. And we're going to talk a lot about that today. So let's add that in. So even headways?

AUDIENCE: Yeah. You wouldn't say short headways. That's not necessarily part of the liability.

AUDIENCE: Well, it seems like we're sort of actually expanding the traditional definition for liability here.

GABRIEL SANCHEZ-MARTINEZ: Yeah, we have been. I'm not filtering too much out because-- I guess we'll stop here. The point I wanted to make is that-- I'll come back to this in lecture-- reliability means a lot of things to different people. And doing all these things can be difficult. So we're going to talk about some aspects. We're going to focus more on punctuality, lack of disruptions, predictability.

We'll talk about common equality in system. That has to do with even headways. We're not going to talk so much about fair policy, or coverage, or span of service, or legibility.

All right. So right. So impacts of unreliability, the causes of unreliability, how to measure unreliability or reliability, and then real-time control strategies. We've mentioned some of these topics before. Now let's focus on these things.

OK, so passenger impacts. What happens when service is unreliable? We have longer waiting times, right? We have the need to have a longer reliability buffer. So have we talked about reliability buffer before in this course? No? So if your trip to-- we're going to talk more about it later than we will here. But if your trip to--

[WRITING ON BOARD]

OK, so you know that running times are stochastic, right? So there's a distribution of running times. So this is running time, or in this case let's call it journey time. And this is a probability [INAUDIBLE], right? So the schedule has to have a single intrinsic value. And you can go to your journey planner, and it will say your journey takes this long. You walk this much time, you wait this much time, you arrive at this time. So you add those components up and you have a specific time.

That specific time may be close to the median, and may not. So if service is reliable, then there's little variance in this distribution, correct? So you might have-- I don't have much of

another color here. That doesn't work. So I'll just write on white as well. Maybe a dot.

So you might have one distribution of journeys that has lower variance. And therefore you need to budget less time to make your trip and arrive on time. Correct? So to the extent that the waiting is variable or your in-vehicle time is variable, then if you need to be at work or at school at 9:00 AM, that means that, even if on average it takes half an hour, you might have to budget for 45 minutes. So the extra 15 minutes of time that you budget, that's called the reliability buffer. It's a cost. You could monetize it if you wanted to. It means that people are budgeting, and in some cases wasting, that time. And we want to decrease that. So the need for trip time reliability buffer.

Higher loads. So that means, of course, that vehicles have more people in them, and therefore rides are more uncomfortable for passengers. And they are also slower, because if a vehicle has more people in it, the bus stops more often and the process is slower. We know that from earlier lectures.

We know from random incidents that from a passenger perspective, passengers will experience more crowded vehicles because they are more likely to board the vehicle where there are more people in it. It's a little bit of a chicken and egg problem. So even if, on average across the vehicles, your loads are right where they should be, most people are going to be on the more crowded vehicles. So that's that thing.

In terms of agency impacts can increase costs, making, for example, you need a longer recovery time at the terminal. And therefore, you know that that translates to higher vehicle costs, higher crew costs, reduced ridership and revenue, reduced operator morale, a public and political problem. What do we mean by that? What happens when public transportation is unreliable in terms of politics?

AUDIENCE: The public hates it. Public hates unreliability.

GABRIEL SANCHEZ-MARTINEZ: Sure.

AUDIENCE: It creates pressure. The forces against public transportation grows stronger.

GABRIEL Sure. So--

SANCHEZ-MARTINEZ:

AUDIENCE: [INAUDIBLE] funding, and then--

GABRIEL SANCHEZ-MARTINEZ: Yeah, so there's a funding, right? And we'll talk more about this in the last two lectures. So when public transportation is unreliable, sometimes governments are hesitant, or the public may be hesitant, to fund it more before they fix the problems. But fixing the problems might require some resources. So it's a vicious cycle. And then reduced effective capacity. And we'll show a real example here in Boston of that.

All right. So what causes this? We can divide causes into external and internal causes. External causes. Traffic, demand-- demand variability in particular-- incidents. So a passenger may get sick. These things are external. The agency knows that they could happen, but there's not much you can do about them.

Internal causes are in the control of the agency, at least a little more. So equipment failure, that's something that agency owns. Maybe you can't predict exactly when it will happen, but with some preventive maintenance they could reduce the chance of this happening. Insufficient resources, poor operations planning, lack of supervision and control, and human driver behavior. What do I mean by human driver behavior? How does that cause unreliability? Henry.

AUDIENCE: The way that a driver drives can vary and that affects runtime. Or maybe some are more aggressive and some are less.

GABRIEL SANCHEZ-MARTINEZ: Yes. Some drivers are slow, some drivers are fast. And then if you're running high-frequency service, they'll bunch up. And that could be the cause of bunching. Or maybe that fast driver is now running early, rolling a long headway service early. And that's really bad for waiting time, as we'll see.

So if we think about transit, and look at the whole process of planning it and delivering it as if it were a business process, we know that we start with service planning, and move on to operations planning, and then actual operations. And each of these things, the output of one is the input to the next process, and so forth.

The input to service policy, among them are is demand estimation. Then we bring in analysis of models or vehicle scheduling and crew scheduling. And during operations there's service

control. We'll talk about service control, real-time control, too, as an input to operations, to make them more reliable.

So there's also a feedback loop, because passengers might complain or somehow you have data collecting information on performance. And you can take that back to service planning or operations planning, or you could modify your service control to change the cycle. And here we have the key agents in each of these steps. So transit agency management, mostly taking care of service policy. Then you have some operational planning stuff, taking care of operations planning. And then the actual drivers and inspectors out in the field.

So I guess one thing to take away from this is that there are many decisions across all of this process that affect reliability. And you can't really have reliable service if you-- there's many places where you could-- a cause on reliability. One of them might be your vehicles break down. So no matter what you do, if your fleet is old and having difficulty with maintenance, you might have the best service control policies, and all the supervision in the world, and you still have unreliable service. So it's a hard problem.

It used to be very hard, actually, because, well, to ensure reliability you need to ensure the variability and things. So you know from data collection that if you want to measure the mean, you can calculate your sample size. Measuring the variance of something requires an even greater data collection effort. So things have gotten a lot easier with technology. And so we know that we have several data collection systems-- AVL, AFC, APC. That makes it a lot easier and cheaper to measure reliability. We'll give some examples.

Then we have scheduling systems, which make it very easy to make an adjustment in the schedule. If we see that we've not scheduled enough time for this bus, we need to add some recovery time, that's going to increase the fleet size by one. With a program or scheduling software, you can do that quickly and react.

And the best systems will actually read AVL data in, and you don't even have to manually analyze it. You can sort of use the program to analyze it. And the same program will help you schedule.

And then there's improved communications technology, which makes it easy to communicate instructions. Hey, depart a little later from this stop, or from this terminal. Don't, you know, start running without passengers. And start operating in revenue service three stops down because

we have a long gap.

So you may have heard, you can't really have control-- there are so many ways of saying things, but you can't really do something about something until you measure it. So measuring it is important. We'll talk a little bit about reliability as a performance metric. And there are different key performance indicators you can calculate for reliability, and that's the first step. You want to measure it and you want to be able to track your reliability for particular routes or system wide, and then be able to change a policy, see what the effect is.

So one thing about it is it's not the only thing that matters. Reliability is important. It would be a mistake not to measure it. Because you might have a very productive service, but you might have all of these problems that we identified. On the other hand, if you put too much focus on reliability, then it would come at a cost.

If you want to make your schedule very reliable, then you can add a lot of recovery time. And you could add many timing points and make the buses hold everywhere, and that would slow service down. It would increase journey times. Even if they would be very predictable, they would be very long, and that could deteriorate service.

So we make a distinction between reliability on low frequency service and reliability on high frequency service. What's the difference? We've talked about the waiting strategies for low and high frequency. We've talked about it several times before.

[CHUCKLES]

What's the difference between a service that runs every 20 minutes and a service that runs every five minutes with regards to your strategy to take that service?

AUDIENCE: It's more frequent and you can just walk up.

GABRIEL SANCHEZ-MARTINEZ: OK, so high frequency service is walk-up service.

AUDIENCE: No random arrivals.

GABRIEL SANCHEZ- Right. The arrivals tend to be random. They don't actually have to be for this distinction to be valid. They have to be independent of the vehicle arrivals.

MARTINEZ:

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: Yes. Yes, we'll mention that. So with low frequency service, most people will time their arrival at stops so that they are with the schedule. So if vehicles are arriving on schedule, and passengers are arriving on schedule, people wait very little time. And if there's very little variability in that, if service is incredibly punctual, people would show up the minute before and wait almost nothing.

If there's some variability, then people have to add some reliability buffer time. So even though it says it's going to come at 3:53, you might have to get there at 3:50.

OK. So you then can set-- because you want to measure reliability, people-- especially in the US, we call that on-time performance, which is a long way of saying punctuality. You could have different wait threshold windows. And we've talked about performance measurement before. We're just here, so we're going back to specifically reliability.

One way would be to say, any bus that departs this stop between one minute early and five minutes late is on time. And any bus outside of that window is not on time. And then you say, I am 90% reliable. This bus leaves that stop 90% of the time within that window. And then you could have tighter windows. You could say, not early and up to three minutes late, or not early and up to one minute late.

Of course what happens-- well, let's start with the first one. What's wrong with the first one? What's different about the first one? You allow early departures. What's bad about early departures when you have a long headway service?

AUDIENCE: People aren't expecting it and they miss their bus.

GABRIEL SANCHEZ-MARTINEZ: Right. So if people are really timing their arrival at the stops according to the schedule, and it's 20-minute headway service, bus leaves early. You arrived a minute early. Now you have to wait 20 minutes. It's really bad. Really bad.

This thing is a little-- real-time apps help with this, but not everyone has one. And if service tends to be reliable, and every now and then it isn't, then people might still have-- they're trying to show up to work on time. So there's not much the app can do unless there's a good warning ahead of time, this bus is going to come early. You'd better be there. So you have to

rush to get ready. OK. So it's good to not allow early departures.

What happens when you move to tighter bounds on the late side? Two things happen.

AUDIENCE: Harder to meet it.

GABRIEL SANCHEZ-MARTINEZ: So it's harder to meet. And that means that your reported reliability comes down. So you can imagine, internally, a pressure on the agency to have a little more slack, to say that you have higher reliability. What happens in terms of performance, or in terms of waiting time, benefit for the passenger?

If the agency actually has a holding policy that prevents a bus from departing late-- more than five minutes late, more than three minutes late, more than one minute late-- from a stop in the middle of the route with people inside of it, what happens is you move to tighter windows.

AUDIENCE: The driver will be less lenient in holding the bus for people running after it?

GABRIEL SANCHEZ-MARTINEZ: Especially if you're-- right, tighter window on the early side, more holding.

AUDIENCE: Like, sorry, we've got to go.

GABRIEL SANCHEZ-MARTINEZ: Yeah. So how do you make the bus not be late? What's the strategy for that?

AUDIENCE: You schedule in more--

GABRIEL SANCHEZ-MARTINEZ: You don't schedule for median. You schedule for a higher percentile. So now it's easy, piece of cake. You schedule it for 80 percentile, only 20% of the time will you be-- OK, but what happens? That means that 80% of the time you're arriving early, so you have to hold.

So as you make that window tighter, it means that you're probably going to be setting your percentile higher. And it means that you'll be holding more often and slowing service down. And holding is very frustrating for passengers. So you shouldn't hold always and for a long time.

And in general, I guess one other comment is like any kind of window, especially on the later

side, the benefits of going from five to three to one minutes are going to be marginal in terms of waiting. And they can be substantial in terms of how often you'll have to hold it and how slow the service will be.

AUDIENCE: I'm sorry. Why is holding more often if you're--

GABRIEL SANCHEZ-MARTINEZ: If you have a tighter window.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: So if you have a tighter window-- if your vehicle must depart between zero minutes early and one minute late, it's a very tight window, right? So how can you guarantee that that vehicle departs that stop in the window? You have to schedule it such that it almost always arrives-- you have to allow a lot of time on the schedule so that almost always it can arrive, and it's not arriving one minute late.

AUDIENCE: Ah, all right. [INAUDIBLE] time. So you're--

[INTERPOSING VOICES]

GABRIEL SANCHEZ-MARTINEZ: So you control the late side by scheduling more time. And then what happens on the early side? On the early side, if you arrive early, you--

AUDIENCE: You just sit there.

GABRIEL SANCHEZ-MARTINEZ: You have to sit there. OK. So in long headway service, or low frequency service, there is a little interaction between successive vehicles, so bunching is not a normal thing. So we'll talk about bunching in the context of high frequency service.

And then, of course, real time information is changing this. So now you have an app. And your app, as you're doing your breakfast, phone vibrates and says, here's the time. Here's my predicted time. So you rush. Maybe you'll skip something on breakfast or you'll-- I don't know.

So that is very poorly understood from a research standpoint. We don't really know how exactly that happens, what percentage of people are actually changing when they would arrive based on the app. And we don't really know what the implications are. Because you can't just

say-- imagine if everybody had an app, and everybody used that app. And you're running 20-minute headway service. Now you say, well, everybody is going to show up to time the arrival at the stop. So let's not worry at all about reliability. It doesn't matter when vehicles come because people know and they time their arrivals. So now we can speed up service.

But there is some sense of convenience of knowing that it runs every 20 minutes. The availability of service every 20 minutes is a factor that wouldn't be considered if you did that. So in some way, it is still important even if people can time their arrival at stops, to keep service as scheduled when it's long headway. And we don't really-- how do you measure that? How you quantify that convenience, that coverage? All questions for researchers.

AUDIENCE: So is there-- for this time window, is there a defined threshold to define something that is a very tight time frame? Like median, long, and tight time frame for this? And when does it matter? Because obviously, tight is different in the mornings when you have [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: This is-- that decisions are up to the agency?

AUDIENCE: It's up to the agency.

GABRIEL SANCHEZ-MARTINEZ: Yeah. And usually there's a trade-off where you want to set a goal that is attainable. You don't want to be reporting 30% reliability. So it's a bar that moves according to the abilities of the agency, usually. That's what happens in practice.

If there are no more questions on low frequency service, let's move to high frequency service. A little more interesting. Do you have a question?

AUDIENCE: Yeah. What's the cut off there between high frequency and low frequency?

GABRIEL SANCHEZ-MARTINEZ: Good question. So what do you think?

[INTERPOSING VOICES]

That is a good point. He said it depends which country you're in. That's a very legitimate observation. So it's not just US and non-US. It's--

[INTERPOSING VOICES]

AUDIENCE: Urban, rural.

GABRIEL SANCHEZ-MARTINEZ: Here in the US, here in Boston, what would you consider to be high frequency service?

AUDIENCE: Up to 10-minute headways.

GABRIEL SANCHEZ-MARTINEZ: 10 minutes?

AUDIENCE: Up to 10 minutes.

AUDIENCE: Yeah, certainly not more than that.

GABRIEL SANCHEZ-MARTINEZ: If it's 12 minutes, do you look at the schedule?

AUDIENCE: Yeah.

AUDIENCE: Yes.

GABRIEL SANCHEZ-MARTINEZ: OK. All right.

AUDIENCE: [INAUDIBLE] divides it as 15 and [INAUDIBLE] going up to 16 and 17.

GABRIEL SANCHEZ-MARTINEZ: So yeah, some agencies say 15. In London, I think it's 12. I think that's the cutoff. Now, if you go to Chile and you say that you're running 10 minute, they think of that as long headway service-- very long. Because they have buses that run every minute. So that's high frequency, and seven minutes is low frequency.

So that's a distinction, at least from an operational standpoint, that people make differently in every country. But in terms of waiting time strategy, I think it still holds that it's somewhere between 10 and 15. Maybe between 8 and 15. We don't know. Obviously, there's a

continuum. It's not like everybody switches at exactly some threshold. It depends on the person, and what technology they have, and what service they're taking.

Between those 8 and 15 minutes, or even 12 minutes, there's going to be a transition from one strategy to the other. And you know that when you're well in the high frequency zone, your models of random, or at least independent, arrivals of passengers at stops hold. Whereas as you move towards 15 and longer, they stop holding and you shouldn't be measuring waiting time like this.

AUDIENCE: I was asking that because, even with 10 minutes, which is considered high frequency, I still want to schedule myself for when I'm going to get to the bus stop. I'm not just like, oh, let me just walk to the bus stop and I'll wait.

GABRIEL SANCHEZ-MARTINEZ: So for 10 minutes, do you actually consult the schedule?

AUDIENCE: Well, I consult the app for 10 minutes.

GABRIEL SANCHEZ-MARTINEZ: But do you plan to take the 9:10 train? Or-- you see, there's a distinction, like the day before, you know I'm going to take the 9:10

AUDIENCE: If they were running clock-based, consistent 10-minute headways, then I would.

GABRIEL SANCHEZ-MARTINEZ: Then you would show up.

AUDIENCE: Yeah. Yeah. But they're not.

GABRIEL SANCHEZ-MARTINEZ: So for you, 10 is right there on the fence between the two, it seems. More on the scheduled side, on the high frequency side, perhaps.

AUDIENCE: Yeah.

GABRIEL OK. But let's say that it's five minutes. So now we're clear on that vague area. So now you're in

SANCHEZ-MARTINEZ: a zone where most people will time their arrivals at stops-- sorry, will not time their arrivals at stops. They will show up when convenient. And they expect to wait about half of the headway.

You know that if the headways are variable, you're going to have that waiting time increasing. So there's going to be a reliability buffer time calculation that you can do. And it's going to be the coefficient variation of headway, will sort of affect. That's the factor that sort of increases the expected waiting time.

Punctuality is not as critical for the passengers. They don't really care if every train is five minutes late, as long as they come in every five minutes. It doesn't really matter that they're all five minutes late.

So really, what matters for a passenger is bunching and long gaps. We've talked about that before. We know that in high frequency service, there is a lot of vehicle interaction. We can maybe review it quickly. Think of this as a bus route. And here's a vehicle. Each of these are vehicles. And there's a random arrival process. Passengers are arriving randomly at the different stops.

So what happens when vehicles are not evenly spaced in time, not in space. So what happens when they're not evenly spaced?

AUDIENCE: The one that's close to the next one picks up less?

GABRIEL SANCHEZ-MARTINEZ: OK, so this vehicle has a smaller gap ahead of it. And therefore, because people are arriving at some rate, in that time, for that rate, fewer passengers will be taking this bus. And that means that real times are shorter. Vehicle catches up.

So now you have a bunch. Eventually this vehicle will catch up to this one. If this vehicle, the one ahead of it, has a long headway ahead of it, the opposite happens. More people than average arrive, and then that vehicle has many more people in it. The loads increase. It slows down. So those two effects work together to create bunching. And we've talked about that.

And with bunching, the counterpart is long gaps. You usually have a bunch and then a long gap. Or a long gap preceding a bunch. So you'll hear passengers complaining, I've been waiting 15 minutes, and now three buses arrive, as if that were ironic. But that's actually exactly why it happens.

AUDIENCE: So what's the best strategy to deal with bunching?

GABRIEL SANCHEZ-MARTINEZ: We'll talk about many of them. Yeah. OK, so one observation is that some high frequency routes have branches. So they're only high frequency in the trunk, and the individual branches are actually long headway. And there, you have to run scheduled service on the branches. And then you have to, I guess, try to run scheduled service on the trunk. There might be some hybrid strategies, but that's still the subject of research.

And of course, schedule control is much easier than headway control. You can do it without technology. You can print a schedule, and the driver can have the schedule in the vehicle. And you can tell drivers, don't leave early. And if they follow instructions, they won't leave early.

So if you want to control headways, and you want to keep the vehicles evenly spaced, now you need all this technology infrastructure to detect where vehicles are, to predict the running times between them, and to communicate instructions to those vehicles so that the headways can be adjusted.

Here's a little reliability buffer time. So let's talk about how to actually calculate it in a high frequency, closed fare system. So think of the London Tube. You tap in the station, and then you tap out when you leave.

So you can measure the probability distribution of travel time, or of journey time, from tap-in time to tap-out time. And we define the reliability buffer time as the difference between the 95th percentile and the 50th percentile of that journey time. So again, if this distribution were much narrower, then people would not have to budget that difference as they plan to make a trip, to arrive at a certain time at work or school.

Something interesting about this is that London has somewhat recently started refunding fares if people's journeys exceed 30 minutes of the expected. So they're not doing it by percentile, but they-- so here's an application in fare policy, which is not the most obvious one. But you can certainly measure this reliability of buffer time. Yes, question.

AUDIENCE: Do you know how much it costs them?

GABRIEL SANCHEZ-MARTINEZ: No. I don't think much. Well, it is a lot in total.

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: I don't remember what percent of passengers. I think the percent is quite low, except if there's a big disruption it might be high. But because of the sheer size of the network, I'm sure that that adds up, too. Yeah.

AUDIENCE: It's a good marketing strategy.

GABRIEL SANCHEZ-MARTINEZ: Yeah.

AUDIENCE: Builds confidence.

GABRIEL SANCHEZ-MARTINEZ: London, they don't talk so much passengers. They talk about their customers. And they want to have customer service. And so they like to have that internal mentality.

AUDIENCE: Well, how do they know someone wasn't dwelling at a station as a passenger?

GABRIEL SANCHEZ-MARTINEZ: Because they can look at other people doing that same trip. So if somebody wants to hang out, they're not going to get a free trip just because they wanted to.

AUDIENCE: It's interesting. In the daily metro, to prevent crowding, if you spend too much time in the system, they actually give you a penalty when you tap out. And so some of the journeys are so long, people actually have to rush to get out, or else they suffer a [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: Not very good reliability. We can add that to the list. Yeah.

AUDIENCE: How do they determine the expected journey time? Is it median--

GABRIEL SANCHEZ-MARTINEZ: They look at this-- so they measure this. And I'll show you the next slide. Before we move on to the next slide, any other questions on this one? I saw several hands raised. OK.

So they look at good days and bad days, to answer your question. So they pick some baseline of days that have no major disruptions, service is running well. It's the amount that they

expect. And then they'll measure reliability buffer time on those days. So that is an amount that they say is sort of inherent to the operation and not much under their control, unless they invest in better signaling systems or things like this.

Anything in excess of that, if you look at, now, all days together, will be something that they could have avoided. That's the idea. This distinction between the baseline buffer time and the access buffer time. So if that x as buffer time exceeds some amount, you can then say, well, we have a problem. Passengers were taking much longer than we scheduled for, or that we thought they could take, and therefore let's refund the fare, because it was bad service.

AUDIENCE: This is for [INAUDIBLE] on the Victoria Line?

GABRIEL SANCHEZ-MARTINEZ: Yes. Northbound? You mean this graph?

AUDIENCE: Mhm.

GABRIEL SANCHEZ-MARTINEZ: Yeah. So northbound, southbound, yeah. And the reference is right here. David Uniman, 2009, so you can check it out in the library if you're interested.

OK, so what happens when it's low frequency. Now people are not arriving necessarily to randomly or independent now. They might actually be scheduling their arrival to meet a particular train. But of course, trains may not be running on schedule.

So you have this space-time diagram here. We have time on the horizontal and space, or the stations, on the vertical. And you see that there is some degree of variability in timing. And if these vehicles were on time, maybe they would be more evenly spaced.

So we have taps. Here, there's a touch in at 8:00. And so what we do is that we look at what's the next train? When does that tap in aboard, and that's how much they waited. You're comparing to scheduled departure, which is in this case 8:06.

So because the train-- I don't know if I'm explaining this clearly. So you can match the tap-in time to the closest scheduled departure that follows that tap in. And that's the train that the person intends to board. And that's how much the person intends-- the difference between those two times is the amount of time that the person intends to wait or expects to wait. Any

waiting time beyond that is excess waiting time.

And then you're going to do the same for the in-vehicle time. That's where [INAUDIBLE] and that's how you get reliability buffer time. So you're now considering the waiting strategy is different. And here's the reference. 2010, another thesis.

For bus, there's another challenge. Well, there's two challenges. The first one is we need to distinguish between the performance of the contractor, if it's a private operator and you're paying the bus company to run service for you, and the performance as the passenger sees it. So as it says on the bottom, if service is unreliable, the passenger doesn't care why it's unreliable. But the operator needs to know if it's the fault of the operator or if it was an external reason that was not under the control of the operator. Why? Why is that important? Henry.

AUDIENCE: If there is traffic, then there could likely be traffic downstream, as well. And you would continue to be on the line for the remainder of the trip.

GABRIEL SANCHEZ-MARTINEZ: OK. But you're thinking of a specific trip. Because I'm thinking more generally. So why would the agency want to distinguish between bad performance that can be assigned or blamed on the operator versus bad performance that is exogenous.

AUDIENCE: Because if the bad service is due to some act of God, then the agency maybe can't hold the contractor accountable. But if the contract and service provider did something wrong, then there can be some kind of punishment.

GABRIEL SANCHEZ-MARTINEZ: OK, yes. And in this case, a punishment under the contract might be a penalty. So they might withhold some-- if there's a performance bonus, maybe the operator does not get that bonus when there might be actually provisions for a penalty. Or a discount on the payment.

And it doesn't have to be an act of God. It can be normal traffic. And the variability in travel times that is expected because of traffic that is not under the control of the operator. But what if the operator is not maintaining his buses and some vehicles have to be put out of service? So the operator drops trips and now performance is not as good. That is very clearly the fault of the operator, and you want to penalize the operator. So that's one challenge.

So you have to measure reliability, but you might have to assign its sources. You might have to calculate the sources of those and sort of assign cost codes to whatever you measure. And there might be a mix.

The other challenge is that typically, buses are only tap on and not tap off. So you have to-- journey time includes the in-vehicle portion. So now you have to figure out where this person alights to figure out the in-vehicle component.

And waiting time is another portion of this, and the passenger doesn't tap in until they have finished waiting. So you have to go back and calculate, how long did they wait? So those are two minor challenges.

Here's a strategy. For destinations, we have ODX. So we can look at their next tap in. We can do trip chaining, as we discussed previously, and infer the destination for that person. And for some people that will fail, but we can scale up in probability. We can say, well, if this person didn't have a next tap, let's look at the distribution of passengers who do have an next tap and let's assume that these other people have destinations that are in proportion-- distributed in the same way as people who have destinations inferred.

And then from AVL, we can sort of do the same trick that we did for rail. We can assume that there is some process by which passengers arrive at stops. So if it's high-frequency bus service, we say it's some arrival rate. If it's low frequency service, we say they arrive on time to match the schedule. And then we can compare that with actual vehicle times.

So if we have an assumption about when the passenger arrived, then we can calculate the difference and measure the waiting time. So it's more stochastic because we don't know exactly when this person arrived. But in probability, you know that in aggregate, across all passengers, any one person arrived between this bus and that bus. So yeah. Questions about this process for measuring journey time reliability? Comments?

AUDIENCE: One man complained about waiting time, it's like with the bus. Like, oh, with the scheduled bus, if the person only waits one minute, that's great. But if I'm waiting at my desk for eight minutes, I actually was--

GABRIEL SANCHEZ-MARTINEZ: So that it goes back to the discussion about apps that we had just a few minutes ago, that you can't just say that because there are apps, then we can just run service unreliably if it's long headway.

AUDIENCE: Yeah.

GABRIEL SANCHEZ- So if it's high frequency, you probably won't care. You'll just still go. Or you'll be timing it, but you would have arrived randomly, anyway, to some extent, i.e. your intention to begin the trip

MARTINEZ: is random, or at least independent of the schedule. But if it's long headway service, then yes, you maybe stay at your desk or get a cup of coffee. But you're still inconvenienced. And so there's a penalty for it.

AUDIENCE: Right.

GABRIEL SANCHEZ-MARTINEZ: And we need to figure out how to measure it. And one way is to assume that people still arrive, or want to arrive, on time. And that's what we've been doing because that's what we used to do, and kind of works. But it's slightly different.

AUDIENCE: So how do you use the stochastic process to estimate wait time?

GABRIEL SANCHEZ-MARTINEZ: Yeah. So let's-- another timeline. So in this case it's a timeline, not a space line. So let's say that there's a tap at that time. And the buses are arriving here, here, here.

Sorry. Well, the tap has to be when the bus arrives, or very close to it. Because the person is tapping in at the bus. So what you know for this person, assuming that this bus isn't full, is that that person arrived as late as exactly when that vehicle arrived, and as early as this vehicle.

So you know that the arrival time of that person and the time at which they began waiting, is a stochastic variable with a uniform distribution on that range. And you know the mean of that distribution is half the headway. So if the bus arrivals are high frequency and the person is arriving randomly, then this is certainly true.

If it's long headway service, then you have to then consider what the scheduled times were. So now you have to do a process similar to what we did for rail, where-- let me just draw it here. So if this bus was meant to arrive here, then maybe you say, well, this person must have been wanting to get on that bus. And so you made them wait some extra time.

And then if this bus was scheduled to depart here, then you know that bus was early. So that means that some of the people that end up tapping in here were probably trying to get on this bus and they missed it. Because they arrived between this and this time-- between the actual and the departed time.

So those are the people who suffer the most. And that's more difficult to calculate. You might have to assume some proportion of the people you see here. You could compare to a typical load and see if there's an excess, and assume that that excess would have been on the

previous bus, if that bus had departed on time. So you have to use more assumptions.

OK. So we understand how to measure reliability for rail tap ins and tap outs, long headway, short headway, and for bus. So strategies. So the question Eli had earlier was, what do we do about it? Now we can measure it. So if you can measure something, you do something about it. What can we do about it?

So two kinds of things. We can use preventive strategies and corrective strategies. Preventive strategies are aimed at maintaining normal service and having some robust operations plans that can handle unreliability without having a domino effect, where everything cascades into unreliability. And that it reduces the probability that problems occur.

Corrective strategies are you know, you're monitoring the system, you see unreliability, and you do something to the system at that time to fix it, to make it more reliable and to minimize the impact on passengers. So examples of preventive strategies are having reserve fleet of drivers and vehicles. If a driver is sick, you have another driver and you send that driver. Vehicle is not operating, you have another vehicle.

Having exclusive bus lanes. So if you have great separation, you get rid of some of the unreliability that comes from traffic. Traffic signal priority-- sorry, that should be transit signal priority. We talked about that in our previous lecture.

There are route design strategies. So we know. We've seen that longer routes are more reliable because they have more time between recovery time, between layovers, essentially. They have fewer stops. So there's less of a dwell time unreliability effect. You can make schedules that have a good amount of recovery at the end.

So you look at the percentiles and you make sure that you're not picking the average to calculate your cycle times. And you set the timing points at good percentiles. Maybe it's the median, maybe it's the 40th percentile. And of course, hiring supervisors to make sure that your people are showing up and they're leaving on time rather than whenever they want, et cetera.

So it's schedules. So two critical decisions-- the cycle time, we know about that. We know that it impacts cost and the reliability of departure time of the terminal. And you're very familiar with that, I think, by now, so I don't have to spend as much time discussing it.

Timing points is another. So you can, in the middle of a bus route, have timing points. So we know that there might be a terminal here and a terminal here. And whenever a bus arrives at the end to be, they then have a layover to catch up. We know that from assignment one, even.

But you can actually have points in the middle where buses are also not allowed to leave early. And so if you schedule the times of these in some way, then buses might hold more or hold less, and they will depart those points on time.

Where should you put those timing points, if at all?

AUDIENCE: Inter-transfers?

GABRIEL SANCHEZ-MARTINEZ: OK. So one idea is that if there's a big train station here or another bus line that also runs-- especially for low frequency service-- and people want to connect, they need to transfer, and they want to connect. You might actually have a timing point for both routes there so the vehicles meet and people can switch vehicles. So that's one idea. What else?

AUDIENCE: Maybe after a high variability section, like after [INAUDIBLE]?

GABRIEL SANCHEZ-MARTINEZ: OK. So maybe this lane has a lot of traffic and it's causing a lot of unreliability. So you want to take care of it there instead of letting it cascade into more variable headways, for example. What else?

AUDIENCE: High demand, those stations.

GABRIEL SANCHEZ-MARTINEZ: High-demand stations. OK. So what about demand? Demand is important, but what exactly?

AUDIENCE: I mean, if the bus is reliable at that stations, then the passengers--

GABRIEL SANCHEZ-MARTINEZ: What do you mean? For example, a lot of people getting off from this bus to take that station, or to enter that train station, or--

AUDIENCE: People are getting onto the bus at that station.

GABRIEL SANCHEZ- OK. So right upstream of sections where there are many boardings. That's a very important thing because that's where most people benefit. And what else? What else about passengers?

MARTINEZ: Where would you not want to do this? Harry.

AUDIENCE: When the bus is full?

GABRIEL SANCHEZ-MARTINEZ: When the bus is full. So if your load profile looks-- I don't know. A lot of people get on here. And they're mostly not getting off. And they sort of come down like that. Do you want timing points on that bus route?

You don't. You don't want timing points because most people are getting on here and they don't want to be held here and here. And very few people might be boarding in this section. So if you hold a timing point, you're going to benefit few passengers and penalize, by longer and more frustrating vehicle rides, many passengers.

So what if there's a big transfer station here, and a lot of people transfer, and then here there's another station and a lot of people get on. So then this is a good place to have a timing point, right before a lot of people board and when you don't have many people on the bus. The flow profile is low.

AUDIENCE: [INAUDIBLE] liability before that [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: Exactly. So you want to maximize the segment of the population that will benefit from the strategy and minimize the segment of the population that will be hurt by the strategy, i.e. people in the bus. OK. So that's a word about number and location of timing points.

And then the schedule at each timing point. What should you set the percentile at? Should this be the median? Should they be 80 percentile? Should they be 30 percentile?

AUDIENCE: The median realm.

GABRIEL SANCHEZ-MARTINEZ: So what happens if you-- how many vehicles will have to be held? What percent of vehicles will have to be held if you set it to 80th percentile?

AUDIENCE: Oh, 80%.

GABRIEL SANCHEZ-MARTINEZ: 80%. OK. So 80% of vehicles would be sort of getting here and waiting. That may be acceptable if, again, very few people are here and a lot of people are boarding. What if it's more of a mix? Then you want to bring that down. You dial it down. So some people say set the half cycle to somewhere between 90% and 95%-- we've done that-- and then set timing

points at 65.

I actually prefer that the timing points are below 50. Because then you can advertise to your passengers. You're essentially lying to people and saying-- you're motivating people to arrive earlier than, most often, the vehicle arrives. So their probability of arriving and missing a vehicle that departed early is lower because you're telling them, expect an early departure, essentially. Except you don't call it early. You say that's when it's scheduled to arrive.

So essentially you can control this by scheduling a certain way, and then on the other side by controlling it.

AUDIENCE: But the pushback when people are like, oh, the bus is always running late.

GABRIEL SANCHEZ-MARTINEZ: Yeah. Yeah, so it affects-- there's a balance. I'm not saying, schedule it 10th percentile. But I think a little below 50 is a good strategy. Because then you would minimize holding. You would only hold vehicles that are really early, essentially. And holding is frustrating. You don't want to be doing it all the time. So by scheduling at a lower percentile, you don't hold as often. And you don't hold for as long times.

All right. You have to be careful when you do this, when you have timing points. Because then your AVL data will come in, and you're going to say, what's the running time? Well, if people are following your instructions, the minimum running time will be whatever your timing points are. So maybe your distribution would look like this. But because you have timing points, you've chopped off the early portion. And now it looks more like this. And now your 95th percentile is more to the right than it used to be, maybe. Maybe not. Maybe it stays where it is. But you have to be careful with this. Your average will certainly move to the right. So just be careful about that.

One strategy is to distinguish holding from running, but that's a little difficult with AVL sometimes. Another strategy is to run a month without timing points, essentially instruct your drivers not to use timing points, and collect data for a month, and then use that for planning. So just the things we can do.

OK, now corrective strategies. So obviously, supervision. Having supervisors, having an operations control practice is important. There's different strategies. This is one of my favorite topics in transit. So holding-- you can do holding to schedule or to headway.

So what is holding to schedule? That's a simple one. We've mentioned it a dozen times, and

even in this lecture. You don't allow a vehicle to depart early from a timing point or from a terminal. So if a vehicle arrives a minute early at a timing point, you instruct that driver to hold one minute. Even though that bus is able to move on, that bus holds for one minute.

TSP we talked about. What's deadheading?

AUDIENCE: Just moving a bus from one place to another that's not a revenue [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: OK. So let's say that you have buses here, and this is your terminal. And there's a bus that you are about to depart. And your headway, the scheduled headway is more like this. So you were supposed to have a bus here, but you didn't. Something happened and the bus wasn't there.

And now you have two buses there instead of one. Maybe there was a bunching on the way back and the bunch arrived late, as it usually happens. So one strategy that you can use is to take the first one and instruct a deadhead to somewhere around here.

So this bus would run without its head sign on, not allowing anybody to board-- and that's why it's called a deadhead-- to this stop. And then this bus departs immediately, but it remains in service. So if that deadhead can happen quickly, maybe not even following the streets or the bus, maybe there's a faster way. And you inject it downstream. You can speed things up, you can balance the headways.

What's expressing?

AUDIENCE: Skipping stops?

GABRIEL SANCHEZ-MARTINEZ: That's one way of doing expressions. Expressing can be done from the terminal. So the distinction between expressing and deadheading is that expressing them with people inside. So in this case, you could have allowed people to board at the terminal and then said, we're going to run express.

You tell people who are boarding, we're going to run express. So if you're getting off anywhere between here and here, don't board. And the vehicle then skips stops, but with people inside. So I don't think I need to do a drawing for this one.

Expressing can also be done halfway through. So if you have this situation, then you could say these people might have boarded, not knowing that this bus was going to be expressed, but something happened. Maybe the driver was slow. Maybe there was some traffic accident and

there were delays. So you can then announce this bus is being expressed, or this train is being expressed, and we're going to skip the following three stops and arrive somewhere. So if you're getting off before the express, you have to decide, are you going to stay and then walk back? Are you going to stay and then take transit the opposite way? Or are you going to get off right now and wait for the next vehicle? So it's a higher penalty for passengers.

What about short term? We've talked about all of these, but in the context of planning. And now we're using them for real time control. So this is deadheading.

[WRITING ON BOARD]

Right. And then yeah, so short term.

[WRITING ON BOARD]

Here's the situation. You have-- this doesn't happen often, but it does happen. You have a bunch in one direction and a very long gap in the other direction. And what happens with bunching? The bus behind is usually not very full. So you stop these buses and say stop. And you tell the people here to move to the bus ahead. Get off the bus. We're in short term. We're going to terminate the trip here. And those people will have to transfer to the next bus.

If you don't allow this to happen, then it's really onerous for passengers because they have to get off and wait for the next one. So it's better if you coordinate it this way. And then this bus gets sent here and begins operating in the opposite direction.

AUDIENCE: How much of this is done automatically versus a person looking at this--

GABRIEL SANCHEZ-MARTINEZ: I don't know of any case where this is done automatically. There has been researchers here at MIT and elsewhere, including me, who have tried to write algorithms that detect these opportunities. But I don't know of any real implementation.

London buses does this, or at least they used to. I think they still do. But they really try to coordinate this. They very rarely curtail trips without coordinating in a bunch. Because then people have to wait for the next one.

AUDIENCE: [INAUDIBLE] as well.

GABRIEL Pardon?

SANCHEZ-MARTINEZ:

AUDIENCE: The train systems [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: Yeah. Yes. So in trains, it's more common. But it's usually a person in a control center determining that they want to do this.

AUDIENCE: What's stopping them from [INAUDIBLE] algorithms? Is it that the algorithms are not good enough or--

GABRIEL SANCHEZ-MARTINEZ: That's part of it. And then yeah. I think I would say that's part of it. The other is maybe a lack of interest or a lack of faith in the computer. There's a system in place. It's just, you know, it's hard.

AUDIENCE: But there are two elements to this problem. One is that you have to detect a problem in one direction. And the other thing is that you have to detect an opportunity in the other direction.

GABRIEL SANCHEZ-MARTINEZ: So this doesn't happen very often.

AUDIENCE: There are two parts here that are completely independent from each other.

GABRIEL SANCHEZ-MARTINEZ: They're not independent, though. If the route is long you actually have-- even though in a short route, in any given segment, the probability of this happening is low, if you have a long route it's much higher. It's like the point you brought up with Professor Fritz' lecture, you said, oh, the chances of a bus benefiting from this are very small. And then you said, well, what if it's a long route and there are 20 signals?

Same thing applies here. You might have a long route with many windows, and the chances that there is one window at any given time in the rush hour is actually pretty high.

AUDIENCE: And then the other thing is, as the controller-- suppose I'm the controller, I have the algorithm. And my computer starts beeping and gets all excited. I have to respond quickly to this, because that opportunity will pass.

GABRIEL And getting back to algorithms and anticipation, what's sort of neat about algorithms is that

SANCHEZ-MARTINEZ: you can try to predict that this will happen rather than waiting until it happens to react. And if you can do that early enough, you might actually instruct a bus that is departing the terminal to short term later on. And if you do that, then you can set the head sign on the bus to the short version. And people who are boarding it are going to know that it's going to be short term so you won't have to coordinate with the bus next to it. You don't have to have a bunch, in other words. So you can then minimize the problem to finding a long gap here and predicting when that gap will meet the next bus that's being dispatched. You need algorithms for that. It's very difficult to do it.

AUDIENCE: What do you tell the passengers? Do you tell the passengers, listen, guys, this bus might get short termed, just so you know.

GABRIEL SANCHEZ-MARTINEZ: No. You change the head sign. And you say this bus is running to Central Square instead of Harvard.

AUDIENCE: Yeah.

AUDIENCE: Wouldn't you know this is going to happen during commute hours? Like the red line-- I mean, I suppose for trains, the red line is so long. But if you know your [?OE?] are mainly in the downtown area, wouldn't you know beforehand that you want to just have higher frequency on this one branch?

GABRIEL SANCHEZ-MARTINEZ: That would be a service planning, short terming strategy, not a real-time control strategy. So we make the distinction. Both are short terming, but this is something that you decide in real time to bring the service that you are delivering more close to what you plan to deliver. Which is different from strategy in service planning, where you identify when we talked about rider or quarter strategies. There might be a core of your route that has higher demand and you have short terms to have frequency higher at that point. Very different reasons and horizons for planning and all sorts of things.

OK, use of reserve vehicles, this happens more with rail than bus. You could have it with bus, too, though. I know that in South America they do it with bus. So you have some buses stationed somewhere in the city near key routes. And they're just waiting there for an opportunity.

[LAUGHTER]

[CHATTER]

Right. So then someone in a control center sees a long gap and they call that bus and say, there's a big, long gap there. Rush. Go into that corridor and start operating Route X. So they can inject a vehicle. And you can do that with rail, too, if you have pocket tracks. So you might have pocket tracks somewhere and sort of insert a train. Ari knows something about that.

OK, so mostly, if you do holding, holding's the best strategy for keeping service on track. You don't usually want to do the more aggressive strategies because they hurt passengers and they require much more coordination effort. But if you have a major disruption, sometimes you need them. Holding by itself won't cut it.

Holding can be done in a bunch of different ways. You can hold to schedule adherence. We've spoken about that.

For headway, we've kind of brushed it off and said headway adherence. Well, there's actually a lot of ways you can do that. So one of them is to look at the schedule and realize that the time between scheduled trips is five minutes. So now instead of actually trying to stick to the actual times, you want to stick to a separation of five minutes between vehicles, because that's the time that was scheduled. So that would be scheduled headway adherence.

The next-- and these are in order of sophistication. So the next one is threshold headway adherence. So if you think about it, there's no reason why the scheduled difference between departure times is the best threshold that you should use for holding. There might be one that is a minute less or a minute higher that results in better performance.

Why not? What physical reason is there for the scheduled time being the best one. There isn't any. So if you optimize the threshold, you can get better performance. And usually the optimal threshold is below, it's shorter than the schedule. So you hold less often. The problem with headway adherence is that you hold too much, essentially.

So then there's headway regularity. Don't worry about the schedule at all, or about any predetermined headway. You have a separation between vehicles. And I know that if I have vehicles, I want my vehicles spaced this way. But instead, I have this situation.

So I want this vehicle to hold until it is here. Or rather, what will really happen is this vehicle will stay there. You will ask it to stay there. And you want this vehicle to move on and this vehicle

to move on to these positions. And now they're even, and then you let that vehicle go at that point. So that's even headway holding.

AUDIENCE: That seems pretty easy to implement. How often is it--

GABRIEL SANCHEZ-MARTINEZ: Now it is. It used to be very hard. So how would you implement targeted headway holding without technology?

AUDIENCE: [INAUDIBLE]

GABRIEL SANCHEZ-MARTINEZ: No, but how would you? You can't do it.

AUDIENCE: You need a supervisor [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: You need a supervisor. So in London there was actually a time when you didn't need a supervisor because you had the driver getting off the bus, and hitting a button on a post, and it starts a timer, and then the driver leaves. The next bus comes along and it sees the time. So it knows, like, ah, OK, this is when I have to leave.

So you can do that, even without wireless communication. Even in headway you can't do that. You need communication. Somebody needs to know the vehicle before, the vehicle after, that means they communicated. So it's much more of a recent strategy in terms of--

AUDIENCE: Oh, so in this strategy you're not necessarily trying to maintain a certain headway.

GABRIEL SANCHEZ-MARTINEZ: No certain headway. It's often when it emerges from the system.

AUDIENCE: OK.

GABRIEL SANCHEZ-MARTINEZ: Yep. And that's actually-- if we want to split hairs, there's different kinds of ways of doing unit headway. And let's not get into it.

Then the most sophisticated class is optimization. And there's something called rolling horizon

optimization. So what happens there is that you use a simulation model, essentially, to predict how the system would evolve if you're holding time for a certain configuration. And you can test a bunch of different headway holding configurations and pick the one that's best at a given time based on those forecasts. In other words, pick the holding time that maximizes performance across forecasts.

And so what's neat about this is that they can trade off waiting time and in-vehicle time. When you hold, the people inside the vehicle are waiting more. The winners of this strategy are people who would miss that bus and have to wait for the following bus that has a long headway ahead of it. So by evening out the headways, you make everybody wait about the same. And that minimizes average waiting time and average crowding, et cetera.

So optimization strategy can predict or keep track of how many people are inside vehicles. And it can know, well, I know this is the penalty. I know this is the benefit. Let's trade these off and let's hold the optimal amount of time such that the total time is minimized. And you can multiply by factors that take into account this utility of waiting being higher than the disutility of being inside the vehicle.

So what that does is it reduces excessive holding. So if you know that you have a full vehicle, you don't hold it, even if it's very close to the next one. Why would you hold a vehicle that is full, even if it's this one right here bunched to the next one. You don't. You don't want to hold it.

What if the vehicle isn't full, but it will be full five stops ahead, because demand is very high for some reason. And you know that. Then optimization algorithm will see that because it's predicting for next hour. So it'll be able to pick a strategy that improves performance.

One challenge is drivers. If you hold too much, then drivers might be 20 minutes late for their shift, for their meal break. What do you do about that? That's still a challenge. And that's a challenge throughout. But there is opportunity with optimization algorithms to insert constraints that take those into account so that your holding policy is something that you can implement.

AUDIENCE: It seems like [INAUDIBLE] would be to [INAUDIBLE] the algorithm. If it's taking 45 minutes to come up with this great solution, then [INAUDIBLE].

GABRIEL SANCHEZ-MARTINEZ: So this problem is [INAUDIBLE]. And if you crunched all the possibilities, you would never finish. So you need a very good algorithm to find something fast. Yeah. So there's two different ways of doing rolling horizon optimization. One way is to have constant running times in

demands within the prediction horizon. So imagine this blue window being your-- this is the time right now, and you're predicting over the next, say, hour. And that's the blue window.

What is actually happening is that your running times are going up slowly and gradually. Your demand is going up slowly, gradually because you're in the PND. If you are using static input, you have to say what the running time is for the whole horizon. So what you might do is break up the day into periods. And if you start at this time, anytime before this, you set some average. And then once you cross that threshold, you set a different average. But then you miss out on some information.

So I worked on a model that actually takes some dynamic inputs, too, to consider these transitions, essentially. And the way it works is you have these dynamic functions of running times and demand. You can see the current system state. You send that to a model. The model passes on those inputs and some holding time configuration to a performance model, which essentially forecasts what the system would do, how the vehicle would move, where passengers would board, all those things.

And then that gets sent to a cost model, which says, how much are people waiting? How much are people in vehicles? Let's add those up and figure out what the average is. And then that gets sent back to the optimization model, and the optimization model says, what if we, instead of doing that, let's turn this knob and increase that holding time and decrease this one, go again. So it can do that on a loop, essentially, until it finds optimal times.

The objective function is the average time in the system. So Wv is like extra time in the vehicle. So it's the time that a person spends at a stop in the vehicle, not moving because the vehicle is being held. So you're adding time to that person. And then Ws is waiting time at the stop. And we can multiply by some factor because we know that waiting time is more onerous.

There's a bunch of constraints. I'll skip the differences between static and dynamic. Essentially it's that running time so demand can be functions of time now.

Here's what happens. So here we test four different strategies. The first one here, TH, is threshold holding. EH is even headway holding. OS is optimization of static inputs, and OD is optimization with dynamic inputs.

And then we tested-- if we look at the bottom two sets of four, these are cases where both of the men and the running times are dynamic. So they're both transient. And we tested a case

of low crowding and a case of high crowding.

So what we see is that in all cases of low crowding-- by the way, the bottom 2/3 sets of runs are cases where maybe running times are dynamic but demand isn't, or demand is dynamic and running times aren't.

So if you look at all the cases of low crowding, the white ones here, there's very little difference in performance. So it means that you don't really need the most sophisticated strategy. And the benefit of a very sophisticated optimization strategy with dynamic inputs is not really going to bring you a lot of benefit.

If you look at the high crowding case, what you see is that there is a pretty significant benefit of even headway holding over even optimized threshold holding. But moving to the model with optimization doesn't do much for you if there's a lot of sort of dynamics and system potentially. So part of that is because the static inputs high information about when those vehicles will fill up. And when you insert that information into the system, then the optimizer can see more of what happens and can improve performance.

Here to the left is lower cost and therefore better. Yeah, there's more to parse here, but we're running out of time, so I'm going to skip a few slides.

So you have control problems, routine disturbances where you lose speed adjustments and holding lightly. Then there's short-term disruptions that might last between five and 30 minutes. And there you can apply the strategies that we've discussed. And then there's very long, much longer interruptions in service for which you need to do something major, like bringing out bus service to replace train service. And so only holding, short terming, expressing, deadheading, they're not going to cut it. You have a much bigger problem to deal with.

Here's a situation in a rail operation where there's a signal failure, or perhaps there's a suicide attempt, and there's a person on the tracks. And you have this train that's standing there and can't move. So the first thing you should do is hold the trains downstream, immediately downstream of the blockage. If you don't do that, they will run around and queue up on the other end. And the people entering these stations are going to wait a long time. and those stations are going to fill up.

The other thing you should do is these trains that are lined up here should be expressed as

soon as the blockage is cleared up so that you can move those vehicles forward. And the vehicles at the end of the queue can begin normal operation. Does that make sense?

A lot of the very sophisticated models are not used in practice for a number of reasons. Sometimes it's that the models are simplistic. Sometimes it's sort of a lack of faith in models. Often it has to do with the data in real time about where vehicles are not being that good. So actually detecting where vehicles are could be a major challenge. And you need that information to make the determination of what the optimal strategy is.

Recently our lab did a pilot of a real experiment with the Green Line holding two headways instead of the schedule. So right now they use paper sheets and they have a schedule. And Jeff Fabian, who's a student in the lab, designed an app that has all the train and driver information. And it also calculates the even headway departing time. So that departure time which would cause an even headway.

And for two weeks we had the inspectors of the terminal use that instead of the paper schedule. And what happened-- here's a picture of the app. So you have all the trains for the day. There's color coding. So green is coming up and you can see all these times are different. So 9:36, there's some extra holding there for the train that had been scheduled at 9:34 so that the headways are even coming out.

And here's what happened. The variability decreased by about 40% in some cases, if you look at the coefficient of variation. And you only look at trains that were compliant. What really happened overall was a mix of things. There was some noncompliance with headway times, some insistence in departing on time to the schedule, and then sometimes a compliance to the [INAUDIBLE] headway policy.

So here's the coefficient of variation. You know what that is. You know that lower is better and the Green Line is lower, so it's better. You can calculate the equivalent benefit that you would get by adding trains to service in terms of decreasing waiting time. You can increase frequency by adding new trains. So decrease in variability of headways was equivalent to adding 1.7 trains. Which is money.

AUDIENCE: Yeah, but if I'm not mistaken-- you can correct me if I'm wrong. There is a mechanism in this algorithm that tries to go back gradually to the schedule.

GABRIEL No, I wouldn't characterize it that way. There are constraints to prevent situations where there

SANCHEZ-MARTINEZ: aren't enough trains to be departed. And there are some constraints on how much holding there can be, et cetera. So yeah, there's an intention not to have too much of a cascading holding throughout the day. But it's not so much as a programmed linear trend towards schedule or anything like that.

AUDIENCE: Because you're going off schedule, does--

GABRIEL SANCHEZ-MARTINEZ: Slightly.

AUDIENCE: Slightly. So would you ever run into issues with communicating with the public?

GABRIEL SANCHEZ-MARTINEZ: Green Line, especially in peaks, high frequency, not running very reliably, anyway. So I don't know anyone that looks at the schedule. Maybe very early in the morning, but I don't know. Maybe late at night, but I don't know.

AUDIENCE: I've totally used Green Line's schedule late at night.

GABRIEL SANCHEZ-MARTINEZ: Late at night. Yeah, late at night, I think it would happen. But I don't think that during the day it would happen. And what happened with this late at night is that the schedule is reliable enough that the even headway times are the scheduled times, anyway. So there's no issue with that.

AUDIENCE: So would there be an issue in terms of applying this to a more robust system?

GABRIEL SANCHEZ-MARTINEZ: We learned a lot from this pilot. And I think whoever implements the next rail system will have to rejudge his thesis. Because there's so much that we learned, about compliance, about issues with data. I recommend reading it. Eli.

AUDIENCE: Can you explain how it's determining the headway that it should be?

GABRIEL SANCHEZ-MARTINEZ: It's even headway.

AUDIENCE: It's just maintaining event. It's not like optimizing.

GABRIEL No optimization. It's a very simple even headway strategy. Prediction of the time it takes this

SANCHEZ-MARTINEZ: train to reach the next one, prediction of the time it takes the trailing train to arrive at the terminal, and then using those to calculate the headways, essentially. And then determining the holding time to keep that train the middle and then adding some constraints to prevent excessive holding. Also to prevent-- if there was a long gap, you might want to not hold the second one too much. Send those together.

Here's something interesting. Here's a space-time graph. Time horizontal. Green here shows compliance and red shows non-compliance. So the first thing we see is that when there's a lot of green, those lines are nicely evenly spaced. So that's good. That's the lower coefficient to variation.

When you see red, you see some bunching later on. That doesn't happen immediately. And it's slight. Notice how slight the deviations can be. But you see the effects of the dwell time effect and the crunching happening.

The other thing that happens is that they have someone controlling at reservoir. So they have someone stationed there holding trains, keeping them evenly spaced, more or less. And which trains are being held? The ones that are not as green, right?

So you've sent a train that might be full of people. And you've decided that you somehow need it to depart a little bit earlier than the algorithm said. And then you wait till you fold to them and hold two minutes. You could have avoided that by having left a little later from the terminal.

AUDIENCE: They're never hold at Fenway? Fenway seems to--

GABRIEL SANCHEZ-MARTINEZ: No. No, just long hold times.

So that's the end. Sorry for being a few minutes late. If you have questions on holding or real-time control, let me know. Next lecture is on fare policy.