The 2021 Railroad Problem Solving Competition

Operations Research and Advanced Analytical Approaches Applied to Real-World Rail Problems

First Prize: $2,000 --- Second Prize $1,000 --- Third Prize $750

Predictive ETA Modeling

May 7, 2021

Problem Overview

Industrial supply chains and logistics operations are highly complex and require the coordination of innumerable supplies and production inputs to maintain efficiency. Keeping track of supplies as well as ensuring their secure and intact movement between origination and destination points is an art form that most organizations have yet to fully optimize and address. One of the most frequently encountered challenges in the global supply chain is tracking and monitoring in-transit products, which includes knowing where products are heading and whether they will arrive on time and in good condition. The evolution of technologies like GPS positioning, machine learning, artificial intelligence, and Blockchain have begun to disrupt logistics and supply chains across various industries. However, the rail industry has been slow to adopt these new technologies, instead relying on decades-old RFID reader technology and even manual event updates.

This year's challenge focuses on the development of predictive models to forecast the estimated time of arrival (ETA) of unit train and large multi-car shipments at their destinations. Unit trains represent the movement of an entire group of railcars as a single grouping from their origin to their destination. Unit trains in North America are typically in the range of 80 to 120+ railcars. In some cases large multi-car shipments may fall below the level of a unit train, but are still large enough that they receive special handling using a series of trains that may be carrying other traffic. In this year's challenge we will be focusing strictly unit train and multi-car movements of 30 or more cars. It should be noted that we do not have visibility to all railcars in every train, so a train that appears to be 30 cars may actually be larger then the cars not included in the provided dataset are counted.

No specific schedules exist for unit trains in North America. Instead, one knows only when a train is "launched" from its origin, and the typical time it takes to reach destination. However, the variability in the transit time can be very large due to a variety of causes, and thus the ability to update the estimated times of arrival (ETAs) for these trains becomes critical for the customers. It is the estimation of these ETAs which is the focus of this year's challenge.

As the shipments progress along their route, their location is reported to the shipper. Based on these reports (or "sightings") one can measure if the shipments are ahead or behind their historic "typical" schedule, and using appropriate modeling techniques one can update the predicted ETA for the shipments. In this challenge we will be providing highly detailed movement data for hundreds of unit trains and large multi-car shipments across a dozen different origin-destination pairs, in turn organized into three "lanes." This sighting data will be by individual railcar, but will also be grouped by train or multi-car shipment group. In total the data will comprise

millions of “sighting” or “event” data records, making this both an exercise in predictive model design and data analytics.

There are two sets of sighting events being provided. One set is based on events directly reported by the railroads, and the other is based on GPS transponders on the individual railcars. The railroad provided data arises either when a railcar passes an RFID tag reader (every North American railcar has an RFID tag on it) or when a person manually enters an “event” for the car or train. When a railcar is not moving, there are generally no railroad reported events, which can make it feel like the status of the railcar is unknown for some period of time. With each railroad provided event there is also an “event code” that tells you what happened, and a code for the railroad upon which the event occurred (shipments in this challenge in some cases travel on 2 or more railroads to reach destination).

The GPS data is generated both when the railcars are moved, and when the railcars are stationary. In the later case the GPS transponder reports its location on a fixed time interval. When moving, the GPS units report their locations based on changes in speed using an accelerometer to trigger an event. Thus, one can potentially gain additional insight as to the status of the railcar from the GPS data beyond what one can glean from the railroad provided data. However, the GPS data is strictly latitude-longitude based, does not have an “event code” or a railroad associated with it, and may have some differences against the locations reported by the railroad. In the data provided, location names have been assigned to each GPS event based on the closest named point to the latitude/longitude for the event. This results in differences in the names of locations between the railroad reported events and the GPS derived names.

Looking across all of the shipments in a train or multi-car shipment allows this event data to be combined to provide a richer dataset for the overall train. Looking at the data jointly across cars moving together may become an important element of any predictive model. However, one must keep in mind that a train can be a mile long. Many of the railroad supplied sighting events are generated when the railcar passes an RFID reader, which reads an RFID tag on the railcar. This can result in various cars in the same multi-car shipment being reported at the same location at somewhat different times. Likewise, should two GPS reports for the same train happen to have the same time, they could have somewhat different latitude-longitude values due to the physical separation distance between the cars in the train, which might result in different location names.

Problem Statement

The participants are expected to do the following as part of this competition:

1) **Base Predictive Model**: Develop a predictive model that will forecast the ETA of a unit train or multi-car shipment at any point in its route based on only the railroad supplied data received to that point in time (no use of the GPS data). Several specific locations in each shipment route will be identified for each O-D where the participants will be asked to provide their best ETA estimate (at a train level). These will typically be at the origin of the shipment, at a point approximately half-way to destination, and one that is somewhere between 75 and 250 miles from the shipment destination. In some cases an additional point may be requested that falls after the mid-point of the trip and before the 75-250 mile reporting point. Contestants will be expected to split the supplied data into a “training set” and a “test” set in a manner of their choosing to design and test their model.

2) **Enhanced Predictive Model:** Develop a revised predictive model that is likely similar to the above, except that the GPS data will also be used in addition to the railroad supplied data. One of the key questions is what the impact will be on the quality of the ETAs from adding the GPS information to the railroad supplied events.
3) **Special challenge:** There is also a "special challenge" that is optional. Addressing this special challenge will be taken into account in the overall competition scoring process. The challenge is to look at the case of "partial shared routes." These are cases where the shipments for more than one O-D share the same route for a substantial percentage of their overall movement, but where the routes deviate for the last 100 to 200 miles at either end. Is there any benefit from looking across different O-Ds in terms of improved ETAs? In this dataset, two of the three "lanes" will have multiple O-Ds in them. One is from North Dakota to Washington State, where there are two loading and unloading points, and thus up to 8 possible O-Ds (not all combinations will have traffic), four in each direction. The other is the lane from North Dakota to Philadelphia, where there are three loading points in North Dakota and two unloading points in Philadelphia, which could create up to 16 O-Ds (again only a subset actually have trains between them). As an example, in the case of shipments from Philadelphia to North Dakota, the O-Ds share the same route from just outside Philadelphia (there are two origin points in greater Philadelphia) to western Minnesota/eastern North Dakota, and then deviate for the last part of their move depending on the load point they have been sent to, giving them a shared route that is approximately 1,400 miles long.

Measuring the Solutions

Each entrant is expected to split the data for each O-D into a "training set" and a "test set." The idea is to build the predictive model based on the training set, and then show its effectiveness for the test set. Once the model is complete, a set of evaluation shipments will be provided to each entrant. The evaluation shipments will provide a complete set of sighting events up to, but not beyond, one of the specific points where ETAs must be generated (origin, half-way, 150 miles out, etc.). These evaluation movements will take place on dates that are interleaved throughout the date range covered by the overall dataset. In effect, the evaluation trains will be roughly every 10^{th} train in the dataset when sorted by start date. Because the frequency of the trains is relatively low, there will be few cases where other trains are operating in the same direction on the same O-D at the same time, so overlap with other trains should be minimal. RAS may elect to slightly alter the dates for the evaluation trains. Two versions of the evaluation dataset will be provided – one with GPS data and one without. No information on the actual ETAs will be provided for the evaluation shipments. The entrants will report their ETA predictions back to the competition judges for each evaluation shipment using a prescribed format. These results will then be compared to the actual, observed ETAs, and the overall results will be scored.

The exact scoring methodology is still being refined, but will likely include two measures: (a) the mean absolute error (MAE) of hours between predicted arrival and actual arrival by O-D and location in route where prediction is made, and (b) some variant of the sum of the squares of the differences (residuals) between the predicted and actual arrivals. We will also be looking at the degree any benefit is realized from the use of the GPS data (net reduction in error between the solutions using GPS and those not using GPS).

There are numerous proven examples within many industries of successful remote monitoring use cases, though it remains to be seen whether rail-related supply chain predictability can be tangibly improved through leveraging GPS location data. On a broader scale, achieving more accurate ETAs will help enhance the attractiveness of the

rail transportation mode as a whole, including driving increased supply chain efficiency and actual cost savings for rail shippers.

Shipment Data Overview

Note: The data to be used in this competition is still being refined. A small amount of representative data will be provided with the release of this problem statement. The full data set will be released no later than May 12.

The data as provided has to some extent been "packaged" and "cleaned" for this competition, and will be provided using a standard "csv" format compatible with programs such as Excel or Access. That being said, in many cases the datasets will be too large to be loaded fully into either of the above cited programs. Instead, the entrants may need to write custom code to load and manipulate the data in memory, or use other databases such as MySQL, PostgresSQL, etc.

We have structured the data into there parts: (a) Route Profiles, (b) Shipment Headers, and (c) Shipment Details. Each is described in more detail below.

As discussed earlier, the data is organized into three bi-directional lanes. The following table presents the preliminary aggregate statistics for each O-D. Please note that due on-going work to refine the data for this competition, such as actions to remove the evaluation shipments, the data each competitor receives may differ in aggregate from what is shown in this table. ***This table is strictly to give competitors an idea of the types of data they will receive, and should not be relied upon. An updated table will be provided when the final version of the data is published.***

Lane By Direction	Initial Location	Final Location	Trains >= 30 Cars	Trains >= 50 Cars	Trains >= 80 Cars	Trains >= 100 Cars
out_ARCO-FIDALGO_WA_to_EPPING-FRYBURG_ND	ARCO, WA	EPPING, ND	19	7	1	0
out_EPPING-FRYBURG_ND_to_ARCO-FIDALGO_WA	EPPING, ND	ARCO, WA	32	20	1	0
out_EPPING-FRYBURG_ND_to_ARCO-FIDALGO_WA	EPPING, ND	FIDALGO, WA	43	39	23	0
out_EPPING-FRYBURG_ND_to_ARCO-FIDALGO_WA	FRYBURG, ND	FIDALGO, WA	37	19	0	0
out_NSTRATCO_AB_to_PLAINES_IL	NSTRATHCO, AB	PLAINES, IL	156	95	70	29
out_PLAINES_IL_NSTRATHCO_AB	PLAINES, IL	NSTRATHCO, AB	92	55	37	11
out_EPPING-ELAND-TRENTON_ND_to_PHILADELP_PA	ELAND, ND	PHIESIDE, PA	12	10	10	4
out_EPPING-ELAND-TRENTON_ND_to_PHILADELP_PA	ELAND, ND	PHILADELP, PA	58	41	35	19
out_EPPING-ELAND-TRENTON_ND_to_PHILADELP_PA	ELAND, ND	PHILEASTSIDE, PA	0	0	0	0
out_PHILADELP_PA_to_EPPING-ELAND-TRENTON_ND	PHILADELP, PA	ELAND, ND	27	27	27	13
out_PHILADELP_PA_to_EPPING-ELAND-TRENTON_ND	PHILADELP, PA	EPPING, ND	33	33	29	23
out_PHILADELP_PA_to_EPPING-ELAND-TRENTON_ND	PHILADELP, PA	TRENTON, ND	118	116	108	58

The above table shows the number of trains by size for each O-D. The same train can be counted in more than one column – for example a 75 car train would be counted in both the trains with 30 or more cars, and the trains with 50 or more cars columns. This dataset spans multiple years, so these trains in some cases may operate less than monthly, and generally operate less than weekly.

Movements that share significant route segments include the moves between Philadelphia and North Dakota, and North Dakota and Washington State. These trains are shuttling crude oil between wells in North Dakota and refineries in Philadelphia and/or Washington State (some may also be exported). There are about a half-dozen highly active load points in North Dakota handling these trains. As the empty trains return to North Dakota, relatively last minute decisions are made as to exactly which load point to use for the next loaded move.

Shipment "Route Profile"

Due to a variety of factors, the locations that appear in the events for each shipment can vary between shipments. In some cases these differences represent variations in physical routes, and in some cases they are related to which RFID readers a train passed. Some of the route differences can be fairly minor, such as the exact route taken through Chicago. Others can be very large, such as the multiple routes used between Spokane, WA and the northwest corner of Washington.

It is important to note that the latitudes and longitudes can be somewhat unreliable. This is because they are applied in some cases based on a look-up process that may select the wrong instance of a place name. They can also include occasional spurious values from the GPS units. Furthermore, some locations that appear in the route are simply wrong – these data errors can have various causes including human input errors. All of this means that one cannot reliably define the route for each train based on just looking at a single shipment, and computing distances can also be challenging.

To compensate for this we have undertaken an analysis of the routes followed by each shipment for each O-D, and have identified the core route that is used the most often for each O-D. For these core routes we have used the latitude/longitude values to estimate the incremental and overall distance for each O-D. We have in turn provided these route profiles on an O-D by O-D basis for use by the participants. No mileages are shown on the detailed event data – thus these route profiles are the only source of "official" distances for this competition. You are of course free to try to do your own distance estimations, but you do need to keep in mind the risks that may arise from the errors noted above in the location reports and their associated latitude and longitude values.

A typical route profile is shown below:

O-D Origin	O-D Destination	Seq	Location	Latitude	Longitude	Cumulative Distance
ARCO, WA	EPPING, ND	1	ARCO, WA	48.8729	-122.7087	0.0
ARCO, WA	EPPING, ND	2	CUSTER, WA	48.9160	-122.6381	4.4
ARCO, WA	EPPING, ND	3	BELLINGHA, WA	48.7514	-122.4726	18.0
ARCO, WA	EPPING, ND	4	BOW, WA	48.5628	-122.3965	31.5
ARCO, WA	EPPING, ND	5	BURLINGTO, WA	48.4663	-122.3358	38.7
ARCO, WA	EPPING, ND	6	MTVERNON, WA	48.4210	-122.3133	42.0
ARCO, WA	EPPING, ND	7	EVERETT, WA	47.8118	-122.3821	84.2
ARCO, WA	EPPING, ND	8	LOWELL, WA	47.9583	-122.1935	97.6
ARCO, WA	EPPING, ND	9	SKYKOMISH, WA	47.7094	-121.3589	140.0
ARCO, WA	EPPING, ND	10	WENATCHEE, WA	47.4247	-120.3038	192.9
ARCO, WA	EPPING, ND	11	ODESSA, WA	47.3340	-118.6939	268.5
ARCO, WA	EPPING, ND	12	SPOKANE, WA	47.6558	-117.4166	332.2
ARCO, WA	EPPING, ND	13	HAUSER, ID	47.7488	-117.0091	352.2
ARCO, WA	EPPING, ND	14	SANDPOINT, ID	48.2824	-116.5462	394.8
ARCO, WA	EPPING, ND	15	TROY, MT	48.4659	-115.8909	427.4
ARCO, WA	EPPING, ND	16	WHITEFISH, MT	48.3891	-114.2302	503.7
ARCO, WA	EPPING, ND	17	SHELBY, MT	48.5083	-111.8589	612.7
ARCO, WA	EPPING, ND	18	HAVRE, MT	48.5560	-109.6600	713.4
ARCO, WA	EPPING, ND	19	GLASGOW, MT	48.1850	-106.6185	855.3
ARCO, WA	EPPING, ND	20	SNOWDEN, MT	48.0269	-104.0836	972.8
ARCO, WA	EPPING, ND	21	TRENTON, ND	48.0699	-103.8360	984.6
ARCO, WA	EPPING, ND	22	WILLISTON, ND	48.1428	-103.6332	995.2
ARCO, WA	EPPING, ND	23	EPPING, ND	48.2825	-103.3553	1011.2

North Dakota to Philadelphia Route Profile

The routing from North Dakota to Philadelphia passes through Chicago, and then proceeds to Buffalo, NY. From there it continues east through Rochester, NY and Syracuse, NY, eventually reaching Selkirk, NY. At Selkirk the shipments turn south heading to Philadelphia via a location called North Bergen, NJ (NBERGEN in the data set). Selkirk is near Albany, NY. It is over 350 miles by rail from Buffalo to North Bergen via this routing.

While we cannot explain the cause, there are very few railroad provided sighting events in the provided data between Buffalo and North Bergen, and in some cases there are none at all. On an aggregate basis there is a fairly rich set of GPS sightings along this route. We have ensured that Selkirk is in the route profiles to/from Philadelphia, but are cautioning the contestants that Selkirk will not appear with significant frequency in the detailed data, and then primarily through the GPS events and not the railroad provided events. We have made no effort to "fix" the data as we believe this may provide a prime example where the supplementary GPS data might be an effective way to improve the ETA estimates.

Shipment "Header" Data

The data is organized by lane by direction, with a separate set of files for each O-D.

For each O-D there is a "header file" and a "sighting" or detailed movement file. The header file is a summary file that we generated based on the sighting data that contains one record for each individual railcar that moved across the O-D pair. In generating the headers, we identified each unique train that was operated, and assigned each shipment to one of these trains. The header consists of the following fields:

Field Name	Data Type	Description
TRIP_ID	Number	Unique ID for each railcar trip within this lane
ASSET_ID	Number	Unique ID for the railcar (can appear on more than one trip)
TRAIN_ID	Number	Groups a set of railcars together that were moved as a single set on a specific date from the origin to the destination of the lane.
RAILCAR_ID	Number	Unique identifier for a specific railcar associated within a specific TRAIN_ID within this lane (may be gaps in the numbers due to railcars being removed from train due to mechanical or other issues)
SHIP_DATE	Date	Date upon which this shipment commenced
LEAD_CAR_ID	Number	This is a representative ASSET_ID that travels in the train and could potentially be used to trace the movement of the train
UNIT_TRAIN_ID	Text	This is a code for the unit train the shipment was placed in – its use is inconsistent and probably cannot be relied upon
COMPLETION_CODE	Text	Indicator if this specific railcar successfully moved from the lane's origin to the lane's destination (at this time only trips deemed "COMPLETE" are included in the competition dataset)
RELEASE_TIME	Date/Time	This is the date/time when the shipper "released" the shipment at origin, which means the shipper told the railroad the railcar was ready for movement.
PULL_TIME	Date/Time	This is the date/time when the railroad "pulled" or started to move the shipment from the origin.
OTHER_START_TIME	Date/Time	If neither a RELEASE_TIME or PULL_TIME are known, this is the next best available origin date/time for the shipment.
CP_TIME	Date/Time	If the customer could not receive the shipment upon its arrival at destination, then this "Constructive Placement" time represents the effective time the shipment arrived.
AP_TIME	Date/Time	This is the date/time that the shipment was "Actually Placed" at the customer's destination facility
END_OTHER_TIME	Date/Time	If neither a CP_TIME or AP_TIME are known, this is the next best available destination date/time for the shipment (typically an arrival event).

Field Name	Data Type	Description
ORIGINAL_ETA	Date/Time	The railroad moving the shipment may periodically provide ETAs for the shipments. These are often empirically derived and can both vary significantly as the shipment progresses and be quite unreliable. This field contains only the initial ETA provided at the time the railcar is pulled from the customer's origin facility.
INITIAL_LOCATION	Text	Name of the starting location (should match the lane if COMPLETION_CODE is set to COMPLETE)
FINAL_LOCATION	Text	Name of the ending location (should match the lane if COMPLETION_CODE is set to COMPLETE)
NUM_GPS_EVENTS	Number	Count of the number of event records that are from a GPS transponder for this specific shipment
NUM_RFID_EVENTS	Number	Count of the number of railroad reported events for this specific shipment

Start time/End Time (ETA): For a given trip there can be a number of events reported at the trip origin and destination. For example, there might be one time when the customer reports the shipment is ready to move (called the "release" event), and another event when the train is actually moved out of the origin (the "pull" event). If using GPS, there may be additional events. Likewise at destination you may have an "arrival" event when the railcars reach the vicinity of the customer and a second event when the railcars are actually physically placed at the customers loading/unloading facility (the "place" event). In general for this challenge the best time to use at the origin is the "pull" event, and the best time to use at the destination is either the "constructive placement" event or the "actual placement" event. Constructive placement occurs when the shipment has arrived but must be held out from the customer due to constraints such as capacity issues at the unloading facility – the railroad has done its job and it is up to the customer to determine when the actual placement occurs. However, not all shipments have pull and place/constructive placement events. In our logic, if the pull event is missing, then we set the start time to the release event, and if that is missing we try to use any other available time. At the destination, if there is no place/constructive place event, then we will use first "arrival" event, and if no arrival event we will try to use any other available time.

In this competition we will use the above logic to determine the actual time of arrival for each shipment, and any estimated time of arrival (ETA) will be measured against this derived actual time of arrival. At the origin, the start time for any transit time measurements will be based on the above logic as well.

Completion Code: If the origin of the trip details matches the lane origin and the destination of the trip details matches the lane destination, then the trip is considered complete. If there is a difference in the origin or the destination location relative to the lane, then this is noted in this field. At this time, only shipments that were deemed "COMPLETE" have been included in the dataset.

It is important to note that in the case of the shared lane between Philadelphia and the various locations in North Dakota, different shipments will have different destinations within the same dataset, but still be considered complete. This also occurs with the movements between North Dakota and Washington State.

Sample "Header" Data

A set of typical header records is shown below:

TRIP_ID	ASSET_ID	TRAIN_ID	RAILCAR_ID	SHIP_DATE	LEAD_CAR_ID	UNIT_TRAIN_ID	COMPLETION_CODE	RELEASE_TIME	PULL_TIME
56131752	415224	30	1	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:16	6/20/2016 0:04
56131754	414975	30	2	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:17	6/20/2016 0:04

56131755	414757	30	3	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:17	6/20/2016 0:04
56131756	415051	30	4	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:16	6/20/2016 0:04
56131758	414943	30	5	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:16	6/20/2016 0:04
56131760	415037	30	6	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:16	6/20/2016 0:04
56131761	414927	30	7	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:17	6/20/2016 0:04
56131762	414999	30	8	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:16	6/20/2016 0:04
56131764	415049	30	9	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:17	6/20/2016 0:04
56131765	415213	30	10	6/19/2016 0:00	414757	NORTH STRA-06/19/2016	COMPLETE	6/19/2016 19:17	6/20/2016 0:04

Continued…

TRAIN_ID	RAILCAR_ID	OTHER_START_TIME	CP_TIME	AP_TIME	END_OTHER_TIME	ORIGINAL_ETA	INITIAL_LOCATION	FINAL_LOCATION	NUM_GPS_EVENTS	NUM_RFID_EVENTS
30	1			6/23/2016 15:27	6/24/2016 5:14	6/23/2016 19:06	NSTRATHCO, AB	PLAINES, IL	28	89
30	2			6/23/2016 15:27	6/24/2016 5:14	6/23/2016 19:06	NSTRATHCO, AB	PLAINES, IL	28	89
30	3			6/23/2016 15:27	6/24/2016 5:14	6/23/2016 19:06	NSTRATHCO, AB	PLAINES, IL	29	89
30	4			6/23/2016 15:27	6/24/2016 5:14	6/23/2016 19:06	NSTRATHCO, AB	PLAINES, IL	24	89
30	5			6/23/2016 15:27	6/24/2016 5:14	1/1/1900 0:00	NSTRATHCO, AB	PLAINES, IL	30	89
30	6			6/23/2016 15:27	6/24/2016 5:14	6/23/2016 19:06	NSTRATHCO, AB	PLAINES, IL	30	89
30	7			6/23/2016 15:27	6/24/2016 5:14	6/23/2016 19:06	NSTRATHCO, AB	PLAINES, IL	4	89
30	8			6/23/2016 15:27	6/24/2016 5:14	6/23/2016 19:06	NSTRATHCO, AB	PLAINES, IL	28	89
30	9			6/23/2016 15:27	6/24/2016 5:41	6/23/2016 19:06	NSTRATHCO, AB	PLAINES, IL	30	89
30	10			6/23/2016 15:27	6/24/2016 6:03	6/23/2016 19:06	NSTRATHCO, AB	PLAINES, IL	24	89

The above example are the headers for the first 10 railcars on “Train 30” from North Strathcona, AB to Plaines, IL. Note the slight variations in the RELEASE_TIME and the variations in the number of sighting events for each trip (discussed further below).

Detailed Sighting Event Data

For each header record shipment there is a set of detailed sighting records. The sighting records consist of the following fields:

Field Name	Data Type	Description
TRIP_ID	Number	Unique ID for each railcar trip within this lane
ASSET_ID	Number	Unique ID for the railcar (can appear on more than one trip
TRAIN_ID	Number	Groups a set of railcars together that were moved as a single set on a specific date from the origin to the destination of the lane.
RAILCAR_ID	Number	Unique identifier for a specific railcar associated within a specific TRAIN_ID within this lane (may be gaps in the numbers due to railcars being removed from train due to mechanical or other issues)
SEQUENCE	Number	Orders the sighting events for this specific train/railcar combination by date/time. Note that if you filter out some records, such as the GPS records, there may be gaps in the sequence.
CURRENT_CARRIER	Text	This is the code for the current railroad the shipment is on. This field will be blank/empty for the GPS based sighting events.
LE_STATUS	Text	L = loaded shipment, E=empty shipment. All the railcars on a train should generally all have the same loaded or empty status.
LOCATION	Text	Name of the location associated with the sighting event. It is important to note that the railroad reported events are tied to a specific location on the railroad, while the GPS events are tied to locations that are derived from the latitude/longitude values. One should treat the names for GPS sightings as estimates or “best guesses” as to location.
SIGHTING_TIME	Date/Time	The date and time associated with the event.
LATITUDE	Number	The latitude of the event location in decimal degrees
LONGITUDE	Number	The longitude of the event location in decimal degrees (negative means west of the prime meridian)

EVENT_CD	Text	This is the event code associated with the sighting event. All GPS events use the code "!GP" – all other events are railroad provided events. A list of valid event codes is provided later in this document.

A typical set of sighting records from the origin location is shown below:

TRIP_ID	ASSET_ID	TRAIN_ID	RAILCAR_ID	SEQUENCE	CURRENT_CARRIER	LE_STATUS	LOCATION	SIGHTING_TIME	LATITUDE	LONGITUDE	EVENT_CD
56131752	415224	30	1	1	CN	L	NSTRATHCO, AB	6/19/2016 19:16	53.5510	-113.3600	W
56131752	415224	30	1	2		L	EAST EDMONTON, AB	6/19/2016 23:55	53.5409	-113.4937	!GP
56131752	415224	30	1	3	CN	L	NSTRATHCO, AB	6/20/2016 0:04	53.5510	-113.3600	X
56131752	415224	30	1	4	CN	L	NSTRATHCO, AB	6/20/2016 0:05	53.5510	-113.3600	P
56131752	415224	30	1	5	CN	L	CLOBAR, AB	6/20/2016 0:31	53.5202	-113.3298	A
56131752	415224	30	1	6	CN	L	CLOBAR, AB	6/20/2016 0:33	53.5202	-113.3298	P
56131752	415224	30	1	7	CN	L	DUNBAR, AB	6/20/2016 0:46	53.6543	-113.6323	A
56131752	415224	30	1	8	CN	L	DUNBAR, AB	6/20/2016 0:48	53.6543	-113.6323	P
56131752	415224	30	1	9		L	IRMA, AB	6/20/2016 5:55	52.9127	-111.2375	!GP
56131752	415224	30	1	10		L	IRMA, AB	6/20/2016 5:57	52.9127	-111.2375	!GP
56131752	415224	30	1	11	CN	L	WAINWRIGH, AB	6/20/2016 5:59	52.8431	-110.8497	A
56131752	415224	30	1	12	CN	L	WAINWRIGH, AB	6/20/2016 6:24	52.8431	-110.8497	P

Note that there are three different events at the origin (NSTRATHCO), with an interleaved East Edmonton event that came from the GPS transponder. This shows the effect of the GPS events being assigned to a place name based on the latitude/longitude values and how these place names can differ from the railroad-based sighting events. Turning to the three NSTRATHCO events, W is the "waybill release," X is the "pull" and P is the actual departure event. Based on the prior discussion, it is the X event that we would use as the start time for the trip. This is then followed by a series of A, P, and !GP events. A is for Arrival, and P is for Departure, both railroad reported events. The !GP events come from the GPS transponder, and can be at variety of locations beyond those reported by the railroad.

Here are the sighting events for the termination of this shipment:

TRIP_ID	ASSET_ID	TRAIN_ID	RAILCAR_ID	SEQUENCE	CURRENT_CARRIER	LE_STATUS	LOCATION	SIGHTING_TIME	LATITUDE	LONGITUDE	EVENT_CD
56131752	415224	30	1	90	CN	L	NORMANTOW, IL	6/23/2016 13:46	41.6816	-88.2341	P
56131752	415224	30	1	91	CN	L	RIVER, IL	6/23/2016 13:51	41.6164	-88.2038	A
56131752	415224	30	1	92	CN	L	RIVER, IL	6/23/2016 13:53	41.6164	-88.2038	P
56131752	415224	30	1	93	CN	L	TURNER, IL	6/23/2016 14:02	41.5750	-88.1419	A
56131752	415224	30	1	94	CN	L	TURNER, IL	6/23/2016 14:49	41.5750	-88.1419	P
56131752	415224	30	1	95	CN	L	JOLYD, IL	6/23/2016 14:59	41.5089	-88.0985	A
56131752	415224	30	1	96	CN	L	JOLIET, IL	6/23/2016 15:06	41.5449	-88.0796	A
56131752	415224	30	1	97	CN	L	JOLIET, IL	6/23/2016 15:07	41.5449	-88.0796	P
56131752	415224	30	1	98	CN	L	JOLYD, IL	6/23/2016 15:10	41.5089	-88.0985	P
56131752	415224	30	1	99	CN	L	PLAINES, IL	6/23/2016 15:25	41.4814	-88.1348	D
56131752	415224	30	1	100	CN	L	PLAINES, IL	6/23/2016 15:25	41.4814	-88.1348	A
56131752	415224	30	1	101	CN	L	PLAINES, IL	6/23/2016 15:27	41.4814	-88.1348	Z

In the above there are three events for the destination. As discussed earlier, the A is for an arrival event. D is an "arrival at destination" event, which is largely the same as the "A" event, but not to be confused with the placement of the train on the loading/unloading track. The Z event is the "actual placement" event, and is the one that should be used in this case to terminate the shipment and represents the actual ETA for the shipment. If a "Y" had appeared in the above, that would be for a "constructive placement," which would take precedence over the "Z" event in terms of determining the actual ETA for the shipment. Note that when we refer to the ETA, it is always for the entire route of the shipment and represents the ETA at the receiving customer's facility. This should not be confused for the estimated time of interchange (ETI) that an individual railroad may have for when it expects to hand off the railcar to the next railroad in the route.

As the railcars move across the railroad they can incur dwell (or delays) for various reasons as they sit a location for a period of time. Causes for dwell events can include changes in crews, locomotives, dispatch holds while waiting on other trains, interchange or junction delays, "bad order" events where a railcar must be removed from the train due to mechanical issues, and a variety of other causes. Some dwells will appear to be "random" and not repeated from train to train, and others will happen on a more consistent basis. If one examines the data in the above table for the location "TURNER IL" you will see that the shipment arrives at 14:02, and departs at 1449, incurring a 47 minute dwell. It is by looking at the arrival and departure events that one can detect when dwell is being incurred. However, if a location does not have an RFID reader, or only has one at one end of the facility, then one may not receive both an arrival and departure event, and the dwell may be undetected in the railroad reported event data. In these cases, you may find that the GPS data does provide some insight into dwell events that are not visible in the railroad event data.

While the above move is on a single railroad, some shipments will traverse more than one railroad. In the below example, the events surrounding the change in railroads for a different shipment that passed through Chicago on its way to Philadelphia is shown:

TRIP_ID	ASSET_ID	TRAIN_ID	RAILCAR_ID	SEQUENCE	CURRENT_CARRIER	LE_STATUS	LOCATION	SIGHTING_TIME	LATITUDE	LONGITUDE	EVENT_CD
46320947	379127	73	75	42	BNSF	L	EOLA, IL	3/19/2015 8:37	41.77472	-88.24269	P
46320947	379127	73	75	43		L	EOLA, IL	3/19/2015 8:42	41.77472	-88.24269	!GP
46320947	379127	73	75	44	BNSF	L	CICERO, IL	3/19/2015 9:21	41.84511	-87.73888	P
46320947	379127	73	75	45	BNSF	L	CICERO, IL	3/19/2015 9:22	41.84511	-87.73888	A
46320947	379127	73	75	46		L	CHGO WESTERN AVE, IL	3/19/2015 14:42	41.85635	-87.68905	!GP
46320947	379127	73	75	47	BOCT	L	CICERO, IL	3/19/2015 17:02	41.84511	-87.73888	P
46320947	379127	73	75	48		L	CROMWELL, IN	3/19/2015 20:42	41.406429	-85.611277	!GP
46320947	379127	73	75	49		L	CHGO WESTERN AVE, IL	3/19/2015 20:42	41.85635	-87.68905	!GP
46320947	379127	73	75	50		L	ATTICA JCT, OH	3/20/2015 2:42	41.08694	-82.87722	!GP
46320947	379127	73	75	51	CSXT	L	WILLARD, OH	3/20/2015 5:00	41.050833	-82.722557	A
46320947	379127	73	75	52	CSXT	L	WILLARD, OH	3/20/2015 5:06	41.050833	-82.722557	P
46320947	379127	73	75	53		L	WEST VIEW, OH	3/20/2015 8:42	41.353962	-81.906304	!GP

In the above example you will see the CURRENT_CARRIER change from BNSF to BOCT to CSXT. As with the lack of railroad sightings on the line from Buffalo to North Bergen discussed earlier, there is a lack of sightings from CSX out of Chicago as well. In the above example BNSF takes the shipment to CICERO, which is a yard in the Chicago area. It is handed off to a local terminal railroad (BOCT), which relays the shipment to CSXT. The BOCT does not leave the greater Chicago area, so we know that CSXT receives the railcar in Chicago – unfortunately this is not shown in the data! The next railroad sighting is in WILLARD, a major rail yard on CSXT that is many miles east of Chicago. This fairly clearly shows the gaps that are found in real world data, and will have to be accommodated in the modeling process. Note also the interleaving of a GPS sighting while the railcar is in CICERO as well.

In the idealized world we would see an R and a J event each time the shipment changed railroads. When a shipment is handed off from railroad A to railroad B, railroad A is supposed to issue a "J" event indicating the shipment was handed off by it to railroad B. Railroad B is supposed to issue a "R" event indicating it has received the shipment from railroad A. In general there should only be one of each of these codes, and the J should precede the R event. However, there is always noise in the data, so sometimes they are completely missing (as in the above example), sometimes the timing of the codes differs between the railroads, and sometimes the codes are duplicated. The handing off of shipments between railroads is called "interchange" and can be a major source of delays and unreliability in a trip, and thus may want to be considered in any predicative model.

Separated Traffic

As discussed earlier, in some cases traffic in a multi-car shipment may become separated. It is important to identify these situations so that the traffic that is no longer part of the main movement does not inappropriately impact the ETA estimates for the primary traffic movement. In some cases this separated traffic ends up at a different destination, in other cases it is simply delayed and arrives at a different time, and in some cases the events stop being received. In the dataset provided in for this competition, all traffic does eventually arrive at the destination, but there can be significant date/time differences for some of the shipments that originated on the same train. Three things must be examined to identify these situations. One is where the locations being reported are inconsistent with the other traffic. The second is when the sighting events have date/times that significantly deviate from the other cars. The third is if event codes are received that either indicate a problem, or are inconsistent with what would be expected (see event code list below).

One must be careful about handling of inconsistent locations. In the case of GPS sightings, the events are reported asynchronously by different cars, so the locations where these cars are reported can vary widely and could appear inconsistent at first inspection – it is more important to see if the latitude/longitude values are consistent with the expected route when looking at GPS reporting. Also keep in mind that many different latitude/longitude values from the GPS may be assigned to the same location name, and that these names might differ from the railroad's name for the same place. For railroad sighting events, the locations should be much more consistent. One might sometimes miss a reporting for an individual railcar, but one should not see a lot of cases where a location is reported for only one car in the train. ***Always remember that some latitude/longitude values for some locations are incorrect – don't let such invalid values lead you astray!***

As discussed in the introduction, there can be differences in when different railcars in the same train are reported at the same location due to a variety of reasons including the delay between when each car passes an RFID reader, and the location estimation process used for the GPS data. Thus, outliers can only be identified through use of a methodology that takes these variations into account.

Bad Latitude/Longitude Values

Some latitude/longitude values are bad. While we have attempted to correct a number of these bad values, some most certainly remain in the database. We view this as typical of large data sets, which are rarely as "clean" as one would like. Some values are simply structurally bad – in other words the location has been assigned the wrong values, so the error is consistent throughout the dataset. In other cases, mostly from the GPS events, the reported latitude/longitude values are wrong due to some kind of data capture error. These will generally be of a more random, or one-off, nature.

Spurious/Inaccurate Events

There can be spurious or inaccurate events in the datasets. These are typically cases where bad data is received for an individual railcar, though it is also possible for bad data to be received for an entire train due to data entry errors for manually reported events. The most common error is cases where a bad location is assigned to a shipment. For example, a railcar might be traversing the state of Louisiana, when in the middle of this there is a

event reported placing the car in New York state, hundreds of miles away. The subsequent events then place the railcar back in Louisiana. Part of the data analytics challenge is to recognize these spurious data events, and make sure they do not influence the outcome of the ETA predictive model.

Trains Versus Shipments

What is being predicted in this exercise is the arrival of a unit train or large, multi-car shipment. In general, all of the shipments in this challenge will involve 60 or more cars departing the origin at the same time. For the most part, one should find that the vast majority of cars on the same train have the same sequence of railroad reported events, with their times being fairly close together. Because the GPS data is issued asynchronously by each car's transponder, this data will be much more diverse for railcars on the same train. As a result, when looking at the overall events reported for the train, one should be able to compact the railroad reported events for an individual train into a single, unified set that eliminates many duplicates or near duplicates with relative ease (a near duplicate is case where events for two different cars are identical except that they are separated in time by a few minutes). For the GPS events, there will likely be fewer duplicates or near duplicates, and there may be a benefit from taking the union of all of the GPS events from the railcars within an individual train, and then simplifying the resulting set for near duplicates to the extent there are any. In both cases, one must be careful when building a train event profile from the individual shipment data to take outliers into account, and have a methodology to handle the variations in the event time reported for each railcar at each location.

We would encourage you to examine the routes for the trains in the "training data" and try to develop a standard description of the physical route of the trains in each O-D (you could start with the provided route profiles – though not every train will follow one of these profiles). This can then be used to create a way to identify deviations from the standard route, and potentially a mechanism for developing a set of statistical performance measures (dwells and running times) for the trains by route segment. As always in a large data set, not every train will follow the exact same route, there will be cases of out of order data or other data deficiencies (as noted above), and there could be individual trains that represent major outliers in terms of timing/performance. Out of order records can occur when one record has an incorrect timestamp on it – this can be caused by different reporting mechanisms using different clocks that are not fully synchronized, or due to manual data entry errors. All of this must be addressed as you create your predictive model.

As one merges the shipments together, one will need a way to both remove outliers and determine a "best" time when different shipments have different times at the same location. A number of strategies exist to do this. One is to use the median time across all of the individual shipments. Another is to use a "lead car" strategy where you always use the times from a specific car when that car has a suitable time. Arguments can also be made to use the "earliest time" as representing the "head end" of the train (providing this earliest time is not an outlier/data error). In general, using the mean time is probably not a good idea as it can be more readily skewed by outliers.

Predictive Model Logic

The competition sponsor currently uses a predictive model for estimating ETAs based solely on the railroad provided events. This model takes the following factors into account:

- Total distance for O-D
- Distance traveled to current time
- Hours in transit to current time
- Railroad provided ETA as it changes along the route (not provided in this competition -- only the railroad provided ETA at the shipment's origin is included in the competition dataset)
- Total trip distance
- Total trip duration
- Distribution of historical values from current location to destination in the following areas:
  - Elapsed time to arrival
  - Time spent in a dwell status to arrival
- Load/empty status
- Origin Departure Status (has it been departed from origin)
- Destination Arrival Status (has the train arrived at the destination – but not yet been placed)
- Train Type

In the above, Dwell is represented by the difference between a train's arrival and departure times at location (could be enhanced by GPS data). Load/empty status is a consideration when both loaded and empty traffic is moving in the same direction, which is not the case in this problem competition. While overall, train type is the same for all of the shipments in this challenge, one might see performance differences based on train size. For example, do trains that are less than 80 railcars perform differently than trains with 80 or more cars?

Modeling Considerations

There are many reasons trip duration can vary from one train to the next. One needs to think about the extent to which any predictive model tries to look at these factors and take them into account.

For example:

- We know that there are several interchanges between railroads, and that whether these interchanges have happened or not can change the ETA forecast, either due to the delays in performing the interchange, or differences in the performance of each railroad
- In some cases a multi-car shipment gets split into more than one shipment group somewhere along its route. Many of these are situations where one, or a small number, of railcars must be removed from the train due to mechanical issues requiring repair. In other cases larger groups have to be removed due to operational constraints. How does one identify when this happens? How does one ensure that the ETA remains accurate/unchanged for the shipments that are staying together?
- We may observe that there are often significant dwell times that are incurred at specific locations – whether we have already passed one of these locations or not may impact the ETA prediction.
- Do late trains only get later or try to make up time? Do the accumulated delays to date against typical performance foretell of even further delays?
- If we observe that a train has not left a location after arriving there, do we use this perceived delay to change our ETA prediction (particularly when using GPS data)?
- Do certain events happening in the route for at least some railcars impact the ETA?

- Does it appear that trains are delayed getting into the destination at certain times of day? That is, could there be prohibitions for the trains moving over certain tracks at certain times of the day due to passenger train activity?
- How does performance vary by season? How does it vary by year? What is the relative importance of near term performance versus older performance?

Event Codes

The railroad provided event codes that appear in the dataset are shown in the table below. In addition, two special codes are noted: "*W" and "!GP." The "*W" code appears to be the same as the "W" code – assume there is no difference between them. The "!GP" are the events created by the GPS transponders.

Event Sighting Code Descriptions

Code	Classification	Description
!GP	GPS Event	This is an event generated by a GPS transponder on a railcar reporting its location.
A	Arrival at an In-transit	Equipment has arrived at an in-transit railroad location other than the destination.
D	Arrival at Destination	Arrival at rail destination.
H	Equipment Delayed or Held	Equipment delayed or held.
J	Junction Delivery	Delivery from one railroad to another railroad – as reported by the delivering railroad. (This code should be followed by an R sighting code).
P	Departure	Equipment has departed from an in-transit railroad location other than the destination.
R	Junction Received	Equipment received by one railroad from another – as reported by the receiving railroad. (This should be preceded by a J sighting code).
S	Storage	Equipment is being stored (not used in current data set).
W or *W	Released	Equipment released by patron at date/time/location shown.
X	Pull	Car pulled from patron siding at date/time/location shown.
Y	Constructive Placement/Notify	Railroad notifies customer that railcar equipment is available for placement.
Z	Actual Placement	Equipment has been placed on the patron's siding.
9	Release from Hold or Miscellaneous	Date and Time a unit is reported release from hold or storage (event code "H" or "S").

As discussed earlier, when a shipment arrives at destination, but cannot be placed in the customer's facility due to capacity issues or customer not being prepared to receive the shipment, the shipment is placed into a "constructive placement" status. In effect, the railroad is storing the shipment until such time as the customer is ready to receive it. From an ETA prediction perspective, a constructive placement is the same as an actual arrival at the customer's facility, and takes precedence over an "actual placement" if both are present.

Evaluation Criteria:

The criteria that judges will use to evaluate a solution include the following:

- Feasibility of the proposed solution, it must satisfy all the given constraints.
- The quality of the solution in terms of its objective function value (in this case deviation against actual ETA values).
- The tractability of the solution approach.
- The implementation quality of the approach.
- The practical usability/reproducibility of the solution approach.
- Computational time of the proposed solution approach.
- The generalizability of the solution approach.
- The quality of the paper describing the solution approach. How clear is the explanation? Is it possible to reproduce the approach just by reading the paper?
- The quality of the presentation, to be given by three finalist teams at the Rail Applications Section Meeting at the INFORMS conference in Anaheim, CA, October 24-27, 2021.

(The *virtual* or *in-person* attendance/presentation of at least one person from each finalist team is required)

The finalists will make a presentation at the 2021 INFORMS Annual Meeting. *Aside from the previous factors, the judging panel will take into consideration the clarity of the presentation to make a final decision about the first, second and third places for the competition.* Note that being among the finalists and presenting at the Annual Meeting does not guarantee a finalist will receive first, second or third place. The decision of the judges is final.

Awards:

First Prize:	**$2,000**
Second Prize:	**$1,000**
Third Prize:	**$750**

In addition to the cash prizes, the first prize winners' contribution to this competition will also be considered for publication in the journal *Networks*. The paper still needs to go through the journal's refereeing process; however, it will receive an expedited refereeing and publication process.

Eligibility:

Any practitioners of operations research and management science who are interested in solving problems in the railroad domain using Operations Research and Analytics tools are welcome to participate. Registration is open to all with the exception of RAS officers and organizing committee members. Likewise, members of the organizing committee may NOT help nor guide any participating team.

Teams of up to three members can participate. At least one member of each finalist team must be available to present *virtually or in-person* the team's approach and results at the 2021 INFORMS Annual Meeting.

Registration:

Participation in the RAS Problem Solving Competition requires registration by **May 31, 2021**. Every team must register by the due date to participate in the contest. To register, please send the following information to **railwayapplicationssection@gmail.com** by the deadline.

- For each team member: Name, Email, Organization, Position. Do you have prior experience in problems related to railroads? (Y/N).
- Abstract (no longer than 250 words) of your proposed approach to this year's problem.
- Brief statement describing what motivated you to participate.

After submitting your registration email, you will receive an email confirming your team's successful registration and eligibility.

Can I publish?

Yes, you can. In fact, RAS encourages you to do so. Anyone can use the RAS competition problem and provided datasets in their publication. References to year-specific problem competitions are given in the URL, and as such you can reference the year-specific competition URL which will not be changed.

Important dates:

- Registration & Abstracts: Deadline is **May 31, 2021**
- Full Problem and Data Sets Release: **May 7, 2021**
- Questions and Answers Period: **May 7 – June 30, 2021**
- Participants may ask questions based on the preliminary problem description before May 7, 2021.
- Quiet Period: **July 1 – July 31, 2021** (No questions are allowed during the quiet period.)
- Release of evaluation data sets: **July 26, 2021**
- Return of model results for evaluation data sets: **August 9, 2021**
- Solution Submission: Deadline is **August 1, 2021**
  Participants may continue to work on solutions, but no additional information will be provided.
  Solution includes report on methodology, and solution data set (format of solution data set to be provided in the final problem description)
- Announcement of Finalists: **September 1, 2021**
  Finalists must give a presentation at INFORMS conference, Anaheim, CA, **October 24-27, 2021**.
- Finalists' Presentations: **October 24-27, 2021**, at INFORMS Annual Meeting, Anaheim, CA.
  Each finalist gives a presentation (15-20 minutes) on their approach.
  Judging panel ask questions.
- Winner Announced: **October 24-27, 2021**, at INFORMS Annual Meeting, Anaheim, CA.

Note: Semi-finalists may be given data for one or more additional "lanes" beyond the initial ones, and be asked to apply their solution to these lanes.

Good luck in the competition!

Problem chair: Hyeong Suk Na (South Dakota School of Mines & Technology)

Problem owner: Stephen Ecker (Trinity Rail)