Scraping and Analyzing Real-Time NJ Transit Rail Data

Pranav Badami:
Software Engineer, Data Systems @ Numina

Michael Zhang:
Data Scientist @ Yousician

Who are we?

Two data scientists masquerading as consultants,
with one transit dream.

Who are we?

Two data scientists masquerading as consultants,
with one transit dream.

One with a 15 minute walk to work.

😄

Who are we?

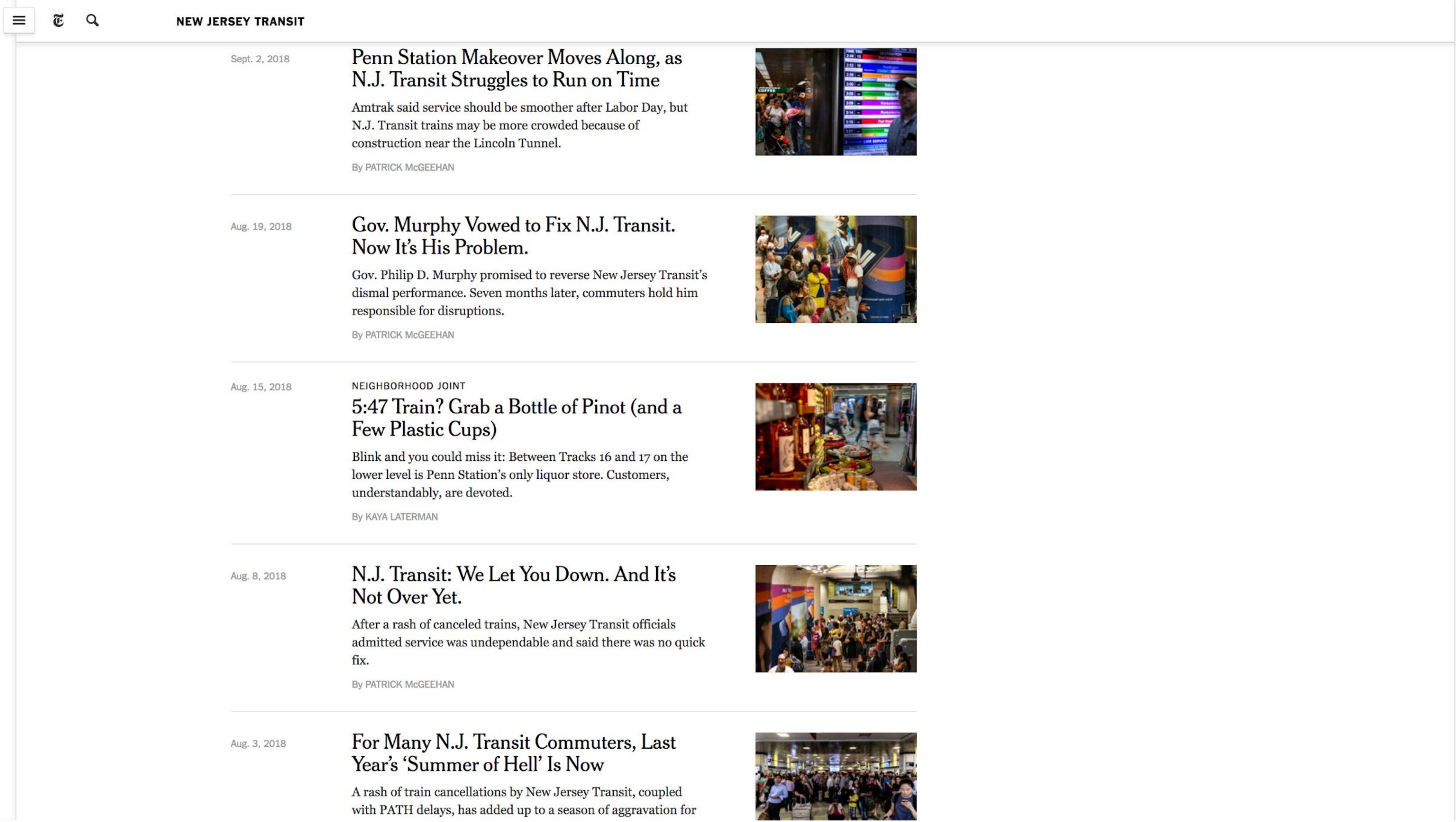
Two data scientists masquerading as consultants,
with one transit dream.

😄

One with a 15 minute walk to work.

The other, with a commute ranging anywhere from an hour to three each way,
taking NJ TRANSIT trains every day.

There's gotta be something we can do…



NEW JERSEY TRANSIT

Sept. 2, 2018

Penn Station Makeover Moves Along, as N.J. Transit Struggles to Run on Time

Amtrak said service should be smoother after Labor Day, but N.J. Transit trains may be more crowded because of construction near the Lincoln Tunnel.

By PATRICK McGEEHAN

Aug. 19, 2018

Gov. Murphy Vowed to Fix N.J. Transit. Now It's His Problem.

Gov. Philip D. Murphy promised to reverse New Jersey Transit's dismal performance. Seven months later, commuters hold him responsible for disruptions.

By PATRICK McGEEHAN

Aug. 15, 2018

NEIGHBORHOOD JOINT

5:47 Train? Grab a Bottle of Pinot (and a Few Plastic Cups)

Blink and you could miss it: Between Tracks 16 and 17 on the lower level is Penn Station's only liquor store. Customers, understandably, are devoted.

By KAYA LATERMAN

Aug. 8, 2018

N.J. Transit: We Let You Down. And It's Not Over Yet.

After a rash of canceled trains, New Jersey Transit officials admitted service was undependable and said there was no quick fix.

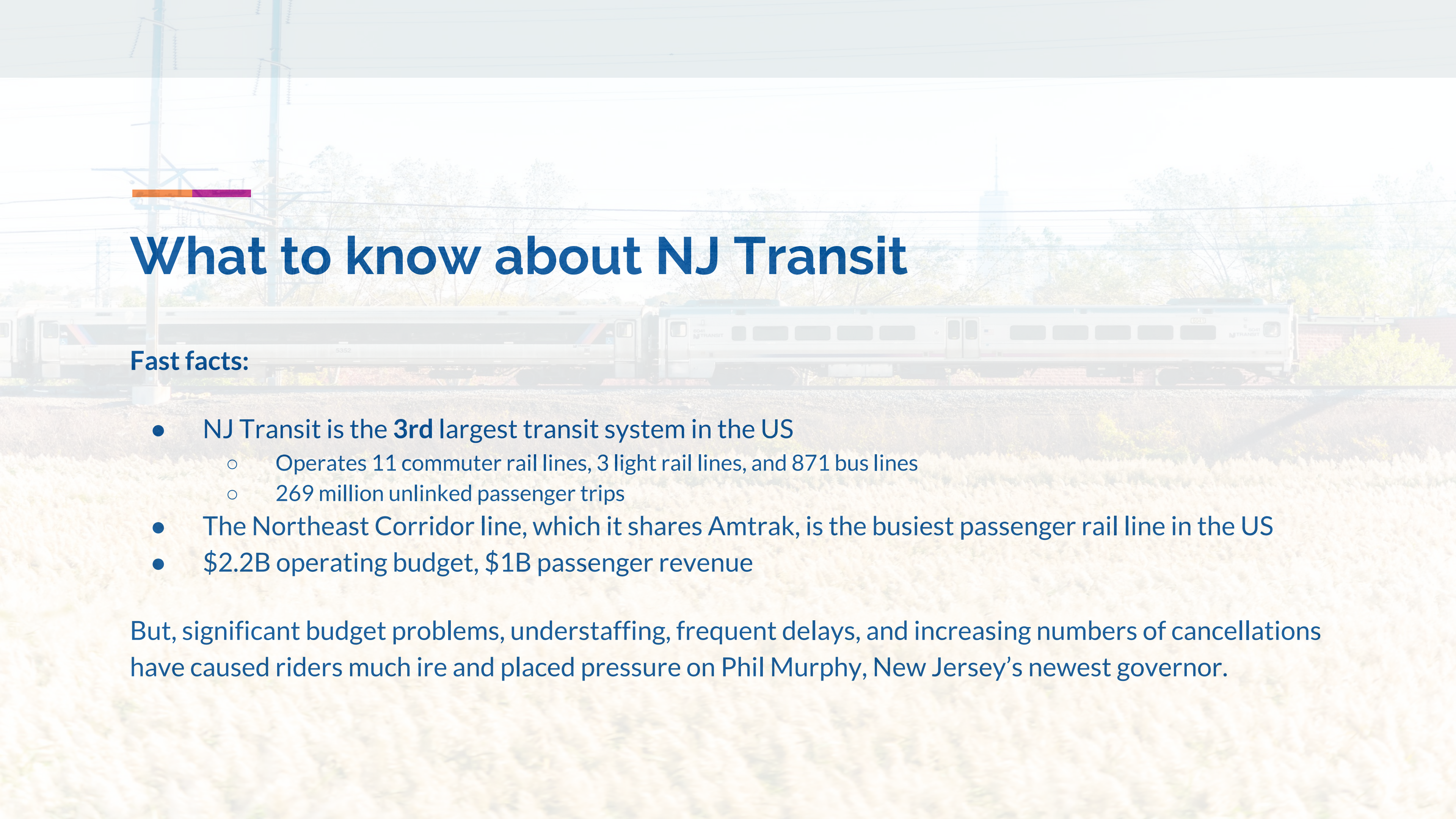
By PATRICK McGEEHAN

Aug. 3, 2018

For Many N.J. Transit Commuters, Last Year's 'Summer of Hell' Is Now

A rash of train cancellations by New Jersey Transit, coupled with PATH delays, has added up to a season of aggravation for

Source: New York Times



What to know about NJ Transit

Fast facts:

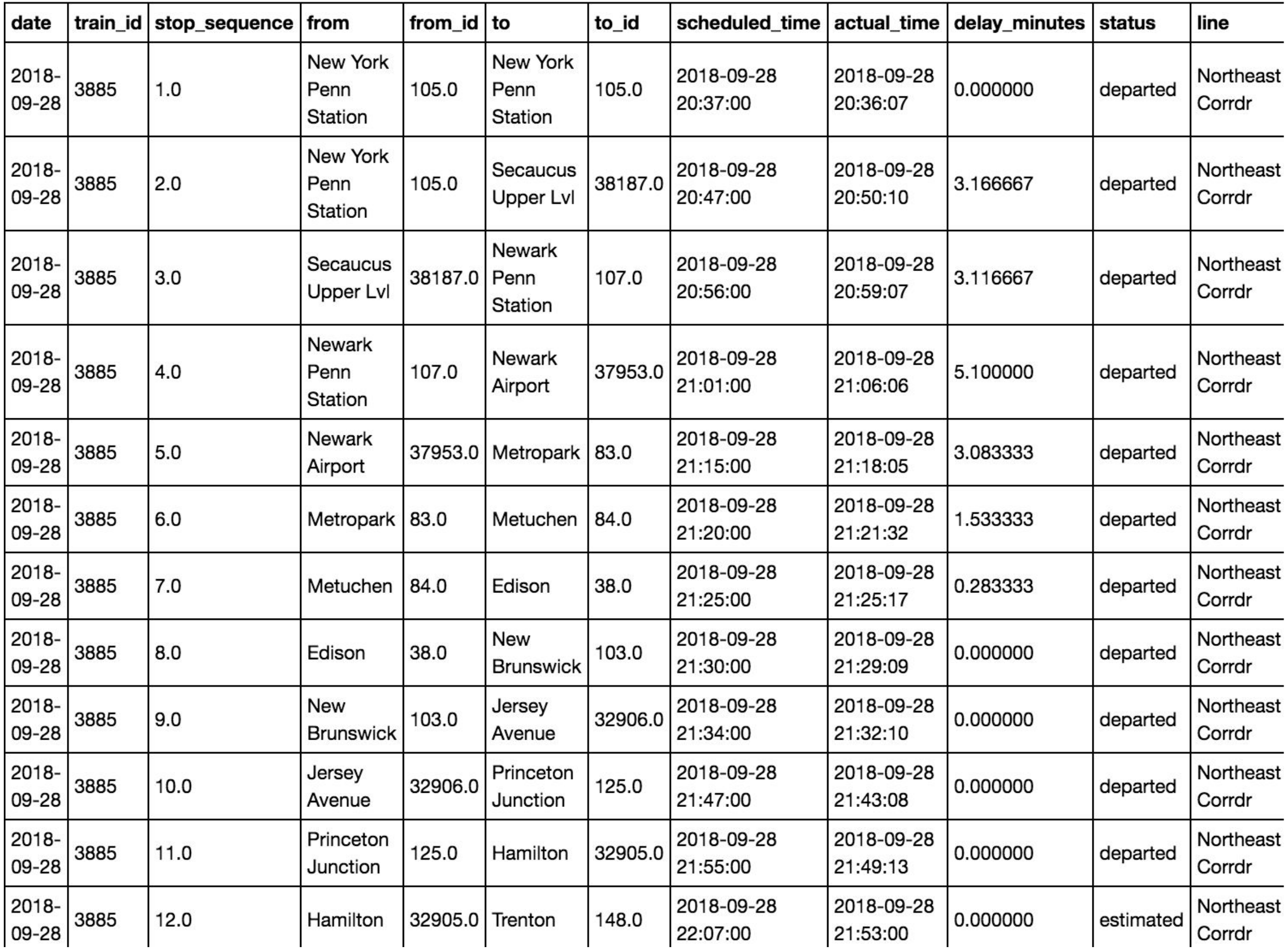
- NJ Transit is the **3rd** largest transit system in the US
  - Operates 11 commuter rail lines, 3 light rail lines, and 871 bus lines
  - 269 million unlinked passenger trips
- The Northeast Corridor line, which it shares Amtrak, is the busiest passenger rail line in the US
- $2.2B operating budget, $1B passenger revenue

But, significant budget problems, understaffing, frequent delays, and increasing numbers of cancellations have caused riders much ire and placed pressure on Phil Murphy, New Jersey’s newest governor.

What did we do?

The Data:

- From Mar 2018
- “Stop-level”
- 1 min resolution
- NJ Transit trips
- Amtrak NEC trips

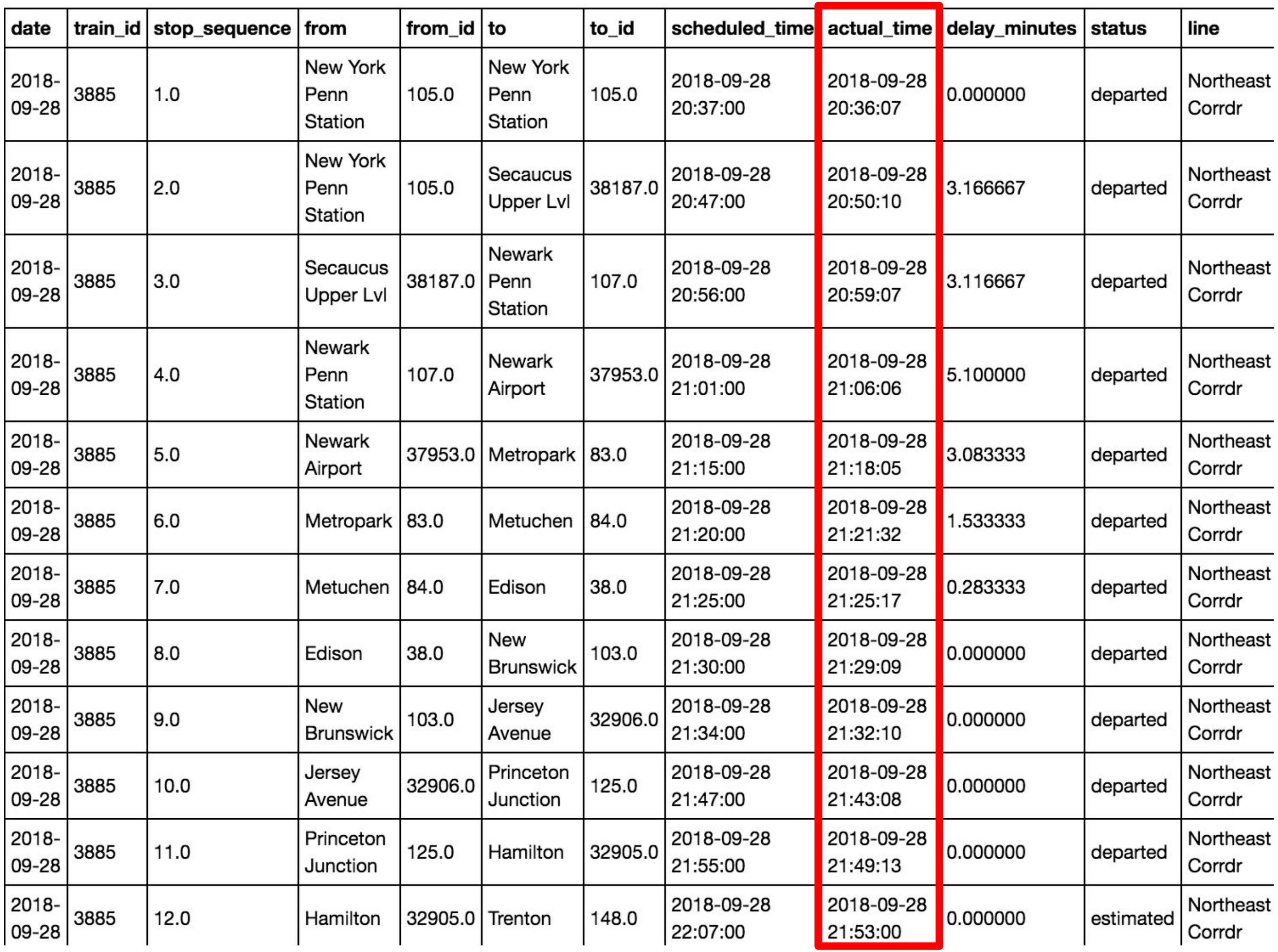


date	train_id	stop_sequence	from	from_id	to	to_id	scheduled_time	actual_time	delay_minutes	status	line
2018-09-28	3885	1.0	New York Penn Station	105.0	New York Penn Station	105.0	2018-09-28 20:37:00	2018-09-28 20:36:07	0.000000	departed	Northeast Corrdr
2018-09-28	3885	2.0	New York Penn Station	105.0	Secaucus Upper Lvl	38187.0	2018-09-28 20:47:00	2018-09-28 20:50:10	3.166667	departed	Northeast Corrdr
2018-09-28	3885	3.0	Secaucus Upper Lvl	38187.0	Newark Penn Station	107.0	2018-09-28 20:56:00	2018-09-28 20:59:07	3.116667	departed	Northeast Corrdr
2018-09-28	3885	4.0	Newark Penn Station	107.0	Newark Airport	37953.0	2018-09-28 21:01:00	2018-09-28 21:06:06	5.100000	departed	Northeast Corrdr
2018-09-28	3885	5.0	Newark Airport	37953.0	Metropark	83.0	2018-09-28 21:15:00	2018-09-28 21:18:05	3.083333	departed	Northeast Corrdr
2018-09-28	3885	6.0	Metropark	83.0	Metuchen	84.0	2018-09-28 21:20:00	2018-09-28 21:21:32	1.533333	departed	Northeast Corrdr
2018-09-28	3885	7.0	Metuchen	84.0	Edison	38.0	2018-09-28 21:25:00	2018-09-28 21:25:17	0.283333	departed	Northeast Corrdr
2018-09-28	3885	8.0	Edison	38.0	New Brunswick	103.0	2018-09-28 21:30:00	2018-09-28 21:29:09	0.000000	departed	Northeast Corrdr
2018-09-28	3885	9.0	New Brunswick	103.0	Jersey Avenue	32906.0	2018-09-28 21:34:00	2018-09-28 21:32:10	0.000000	departed	Northeast Corrdr
2018-09-28	3885	10.0	Jersey Avenue	32906.0	Princeton Junction	125.0	2018-09-28 21:47:00	2018-09-28 21:43:08	0.000000	departed	Northeast Corrdr
2018-09-28	3885	11.0	Princeton Junction	125.0	Hamilton	32905.0	2018-09-28 21:55:00	2018-09-28 21:49:13	0.000000	departed	Northeast Corrdr
2018-09-28	3885	12.0	Hamilton	32905.0	Trenton	148.0	2018-09-28 22:07:00	2018-09-28 21:53:00	0.000000	estimated	Northeast Corrdr

The Data:

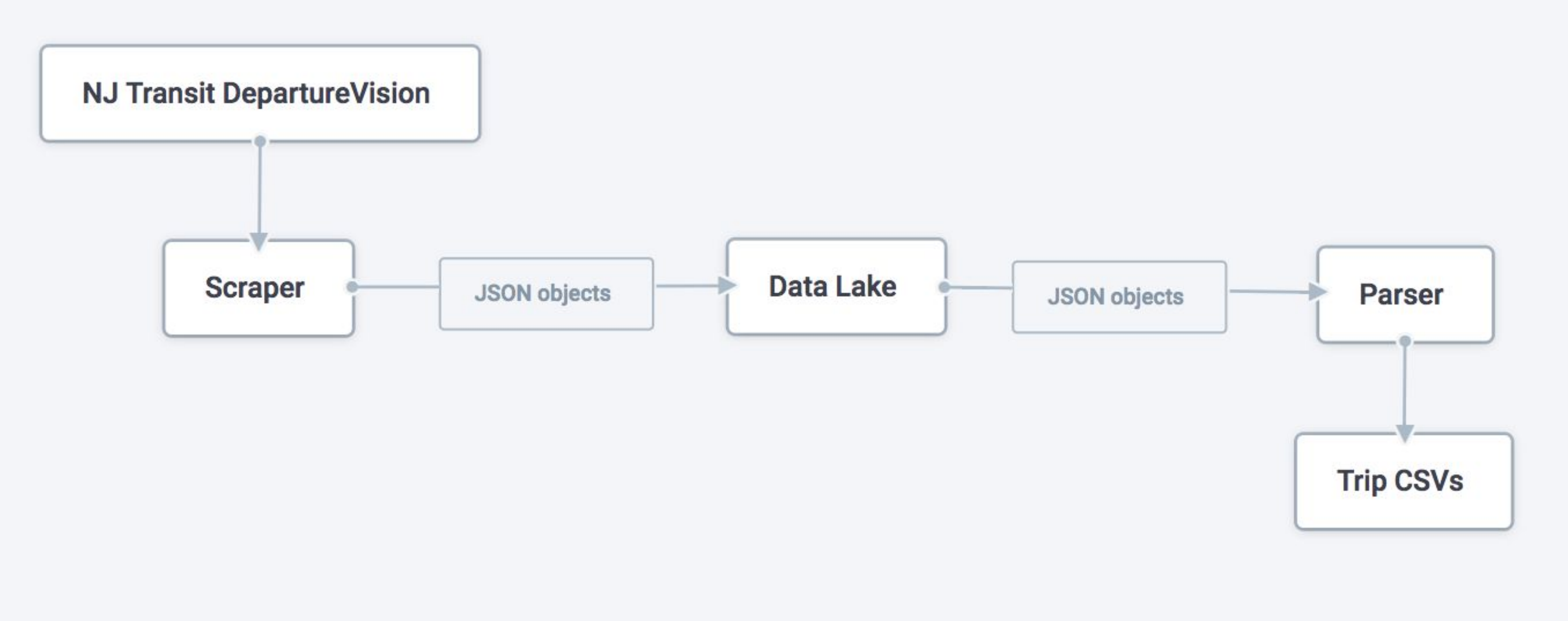
- From Mar 2018
- “Stop-level”
- 1 min resolution
- NJ Transit trips
- Amtrak NEC trips

(Not publically available!)



date	train_id	stop_sequence	from	from_id	to	to_id	scheduled_time	actual_time	delay_minutes	status	line
2018-09-28	3885	1.0	New York Penn Station	105.0	New York Penn Station	105.0	2018-09-28 20:37:00	2018-09-28 20:36:07	0.000000	departed	Northeast Corrdr
2018-09-28	3885	2.0	New York Penn Station	105.0	Secaucus Upper Lvl	38187.0	2018-09-28 20:47:00	2018-09-28 20:50:10	3.166667	departed	Northeast Corrdr
2018-09-28	3885	3.0	Secaucus Upper Lvl	38187.0	Newark Penn Station	107.0	2018-09-28 20:56:00	2018-09-28 20:59:07	3.116667	departed	Northeast Corrdr
2018-09-28	3885	4.0	Newark Penn Station	107.0	Newark Airport	37953.0	2018-09-28 21:01:00	2018-09-28 21:06:06	5.100000	departed	Northeast Corrdr
2018-09-28	3885	5.0	Newark Airport	37953.0	Metropark	83.0	2018-09-28 21:15:00	2018-09-28 21:18:05	3.083333	departed	Northeast Corrdr
2018-09-28	3885	6.0	Metropark	83.0	Metuchen	84.0	2018-09-28 21:20:00	2018-09-28 21:21:32	1.533333	departed	Northeast Corrdr
2018-09-28	3885	7.0	Metuchen	84.0	Edison	38.0	2018-09-28 21:25:00	2018-09-28 21:25:17	0.283333	departed	Northeast Corrdr
2018-09-28	3885	8.0	Edison	38.0	New Brunswick	103.0	2018-09-28 21:30:00	2018-09-28 21:29:09	0.000000	departed	Northeast Corrdr
2018-09-28	3885	9.0	New Brunswick	103.0	Jersey Avenue	32906.0	2018-09-28 21:34:00	2018-09-28 21:32:10	0.000000	departed	Northeast Corrdr
2018-09-28	3885	10.0	Jersey Avenue	32906.0	Princeton Junction	125.0	2018-09-28 21:47:00	2018-09-28 21:43:08	0.000000	departed	Northeast Corrdr
2018-09-28	3885	11.0	Princeton Junction	125.0	Hamilton	32905.0	2018-09-28 21:55:00	2018-09-28 21:49:13	0.000000	departed	Northeast Corrdr
2018-09-28	3885	12.0	Hamilton	32905.0	Trenton	148.0	2018-09-28 22:07:00	2018-09-28 21:53:00	0.000000	estimated	Northeast Corrdr

Data Pipeline



NJ Transit DepartureVision

Scraper

JSON objects

Data Lake

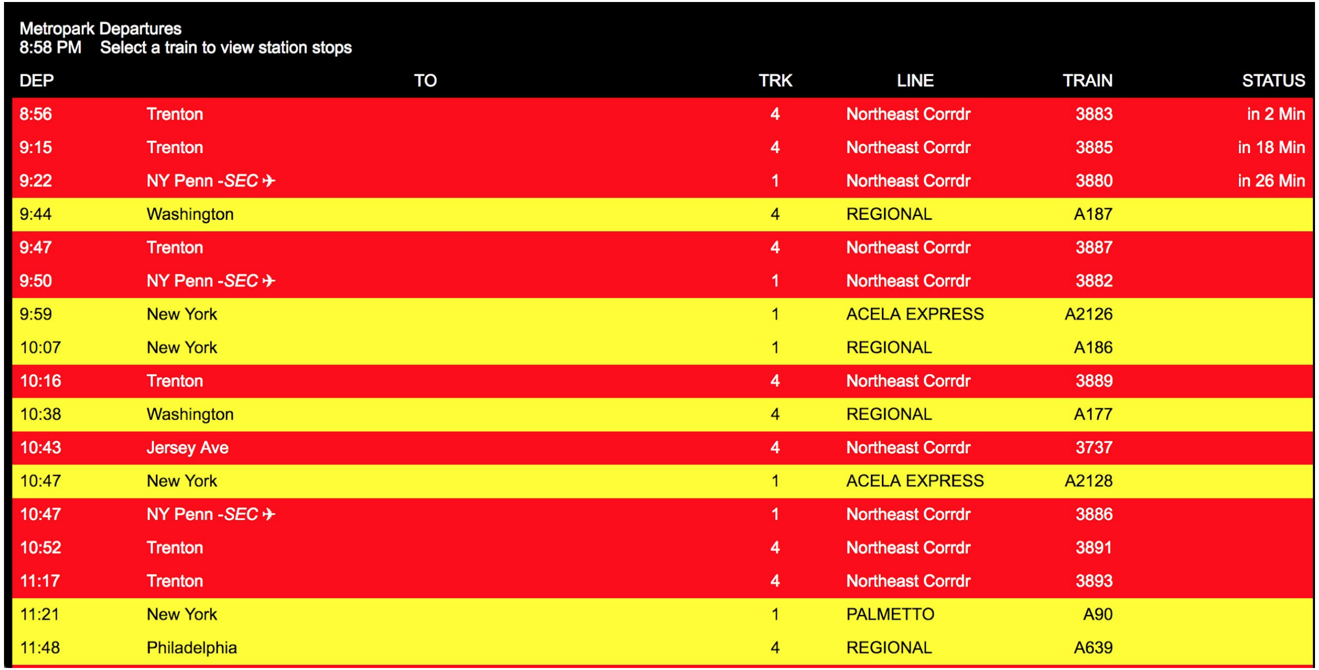
JSON objects

Parser

Trip CSVs

Scraper

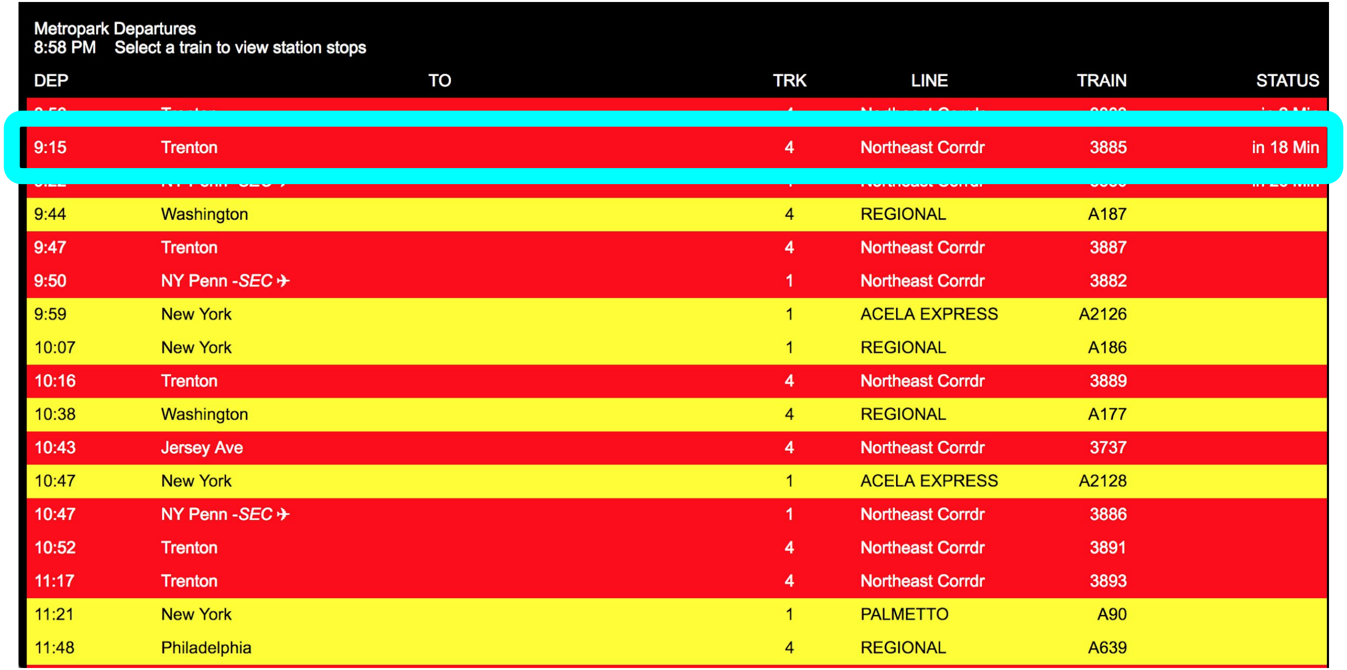
Terminal stations:
5 min to 1 hour



Metropark Departures
8:58 PM Select a train to view station stops

DEP	TO	TRK	LINE	TRAIN	STATUS
8:56	Trenton	4	Northeast Corrdr	3883	in 2 Min
9:15	Trenton	4	Northeast Corrdr	3885	in 18 Min
9:22	NY Penn -*SEC* ✈	1	Northeast Corrdr	3880	in 26 Min
9:44	Washington	4	REGIONAL	A187	
9:47	Trenton	4	Northeast Corrdr	3887	
9:50	NY Penn -*SEC* ✈	1	Northeast Corrdr	3882	
9:59	New York	1	ACELA EXPRESS	A2126	
10:07	New York	1	REGIONAL	A186	
10:16	Trenton	4	Northeast Corrdr	3889	
10:38	Washington	4	REGIONAL	A177	
10:43	Jersey Ave	4	Northeast Corrdr	3737	
10:47	New York	1	ACELA EXPRESS	A2128	
10:47	NY Penn -*SEC* ✈	1	Northeast Corrdr	3886	
10:52	Trenton	4	Northeast Corrdr	3891	
11:17	Trenton	4	Northeast Corrdr	3893	
11:21	New York	1	PALMETTO	A90	
11:48	Philadelphia	4	REGIONAL	A639	

Scraper



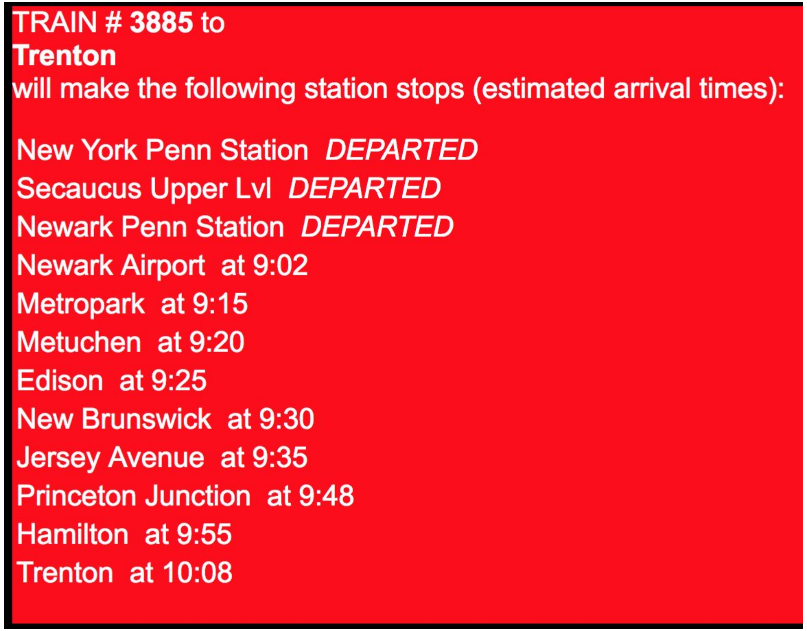
Metropark Departures
8:58 PM Select a train to view station stops

DEP	TO	TRK	LINE	TRAIN	STATUS
[illegible]	Trenton	4	Northeast Corrdr	[illegible]	[illegible]
9:15	Trenton	4	Northeast Corrdr	3885	in 18 Min
9:22	NY Penn -*SEC* ✈	1	Northeast Corrdr	[illegible]	[illegible]
9:44	Washington	4	REGIONAL	A187	
9:47	Trenton	4	Northeast Corrdr	3887	
9:50	NY Penn -*SEC* ✈	1	Northeast Corrdr	3882	
9:59	New York	1	ACELA EXPRESS	A2126	
10:07	New York	1	REGIONAL	A186	
10:16	Trenton	4	Northeast Corrdr	3889	
10:38	Washington	4	REGIONAL	A177	
10:43	Jersey Ave	4	Northeast Corrdr	3737	
10:47	New York	1	ACELA EXPRESS	A2128	
10:47	NY Penn -*SEC* ✈	1	Northeast Corrdr	3886	
10:52	Trenton	4	Northeast Corrdr	3891	
11:17	Trenton	4	Northeast Corrdr	3893	
11:21	New York	1	PALMETTO	A90	
11:48	Philadelphia	4	REGIONAL	A639	

Scraper

https://dv.njtransit.com/webdisplay/train_stops.aspx?sid=MP&train=3885

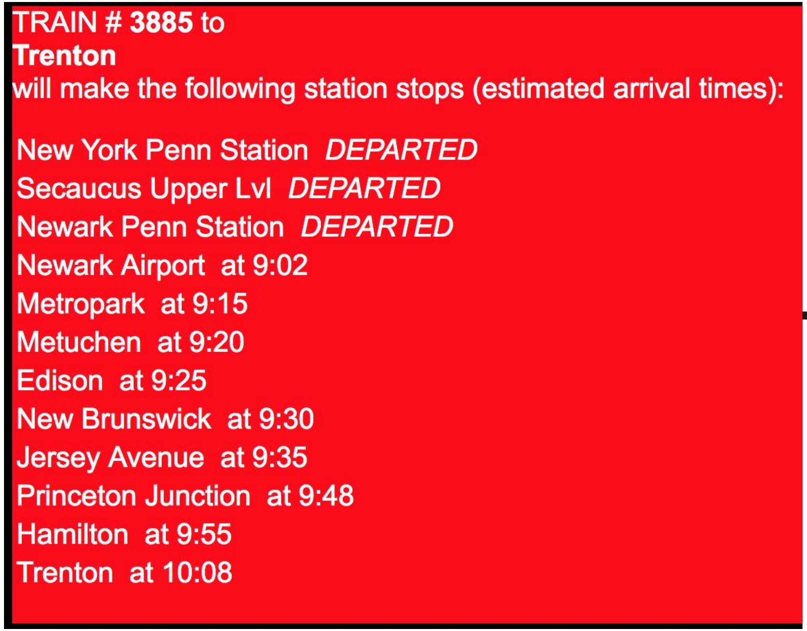
Captured at 9:00 PM



TRAIN # **3885** to
Trenton
will make the following station stops (estimated arrival times):

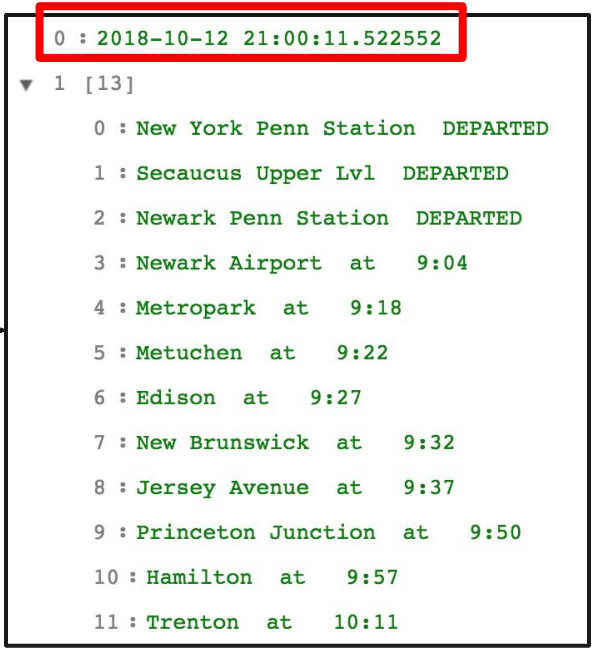
New York Penn Station *DEPARTED*
Secaucus Upper Lvl *DEPARTED*
Newark Penn Station *DEPARTED*
Newark Airport at 9:02
Metropark at 9:15
Metuchen at 9:20
Edison at 9:25
New Brunswick at 9:30
Jersey Avenue at 9:35
Princeton Junction at 9:48
Hamilton at 9:55
Trenton at 10:08

Scraper



TRAIN # **3885** to
Trenton
will make the following station stops (estimated arrival times):

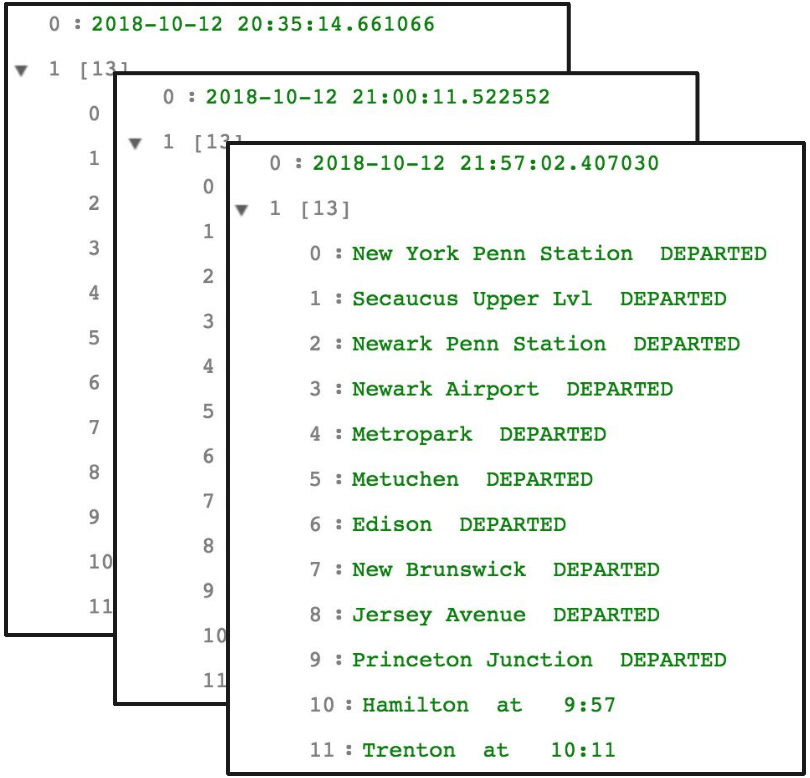
New York Penn Station *DEPARTED*
Secaucus Upper Lvl *DEPARTED*
Newark Penn Station *DEPARTED*
Newark Airport at 9:02
Metropark at 9:15
Metuchen at 9:20
Edison at 9:25
New Brunswick at 9:30
Jersey Avenue at 9:35
Princeton Junction at 9:48
Hamilton at 9:55
Trenton at 10:08



```
 0 : 2018-10-12 21:00:11.522552
▼ 1 [13]
     0 : New York Penn Station  DEPARTED
     1 : Secaucus Upper Lvl  DEPARTED
     2 : Newark Penn Station  DEPARTED
     3 : Newark Airport  at   9:04
     4 : Metropark  at   9:18
     5 : Metuchen  at   9:22
     6 : Edison  at   9:27
     7 : New Brunswick  at   9:32
     8 : Jersey Avenue  at   9:37
     9 : Princeton Junction  at   9:50
    10 : Hamilton  at   9:57
    11 : Trenton  at   10:11
```

Scraper

Every minute:

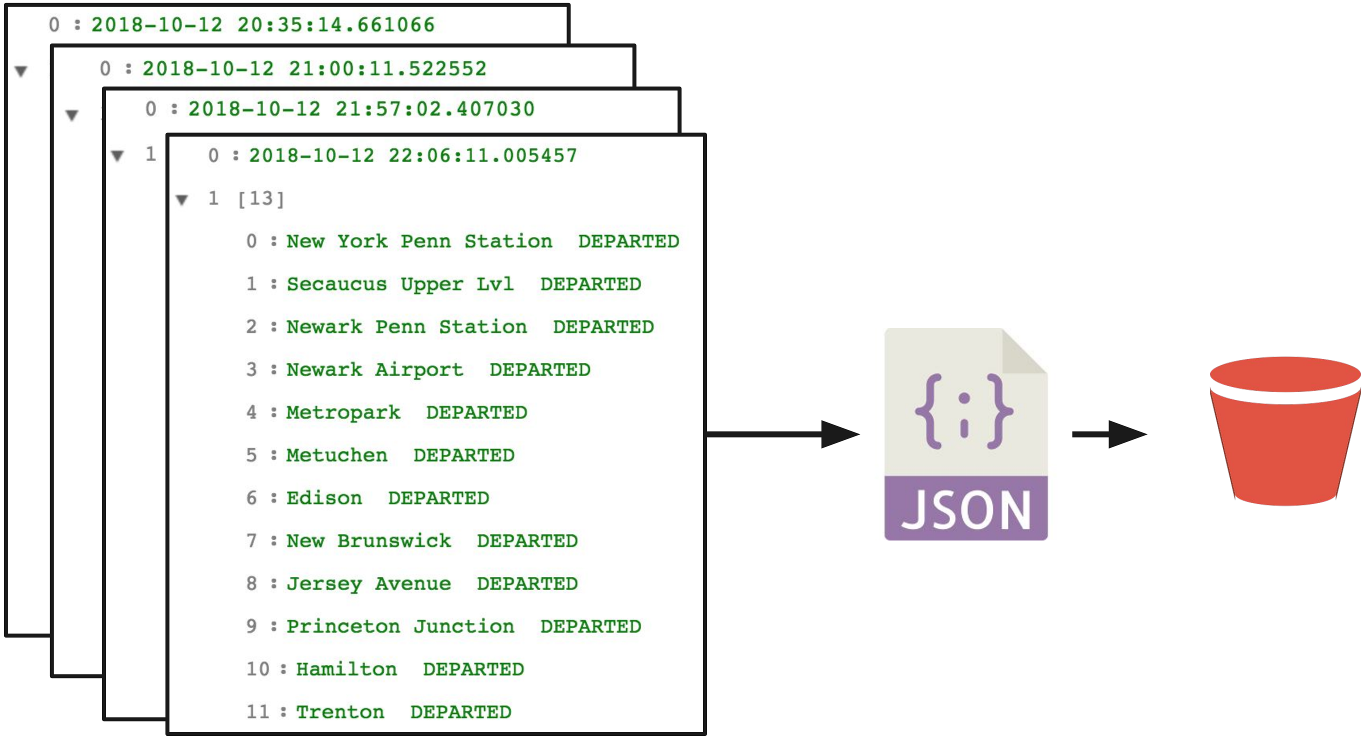


```
0 : 2018-10-12 20:35:14.661066
▼ 1 [13]
    0
    1
    2
    3
    4
    5
    6
    7
    8
    9
    10
    11
```

```
0 : 2018-10-12 21:00:11.522552
▼ 1 [13]
    0
    1
    2
    3
    4
    5
    6
    7
    8
    9
    10
    11
```

```
0 : 2018-10-12 21:57:02.407030
▼ 1 [13]
    0 : New York Penn Station  DEPARTED
    1 : Secaucus Upper Lvl  DEPARTED
    2 : Newark Penn Station  DEPARTED
    3 : Newark Airport  DEPARTED
    4 : Metropark  DEPARTED
    5 : Metuchen  DEPARTED
    6 : Edison  DEPARTED
    7 : New Brunswick  DEPARTED
    8 : Jersey Avenue  DEPARTED
    9 : Princeton Junction  DEPARTED
    10 : Hamilton  at   9:57
    11 : Trenton  at   10:11
```

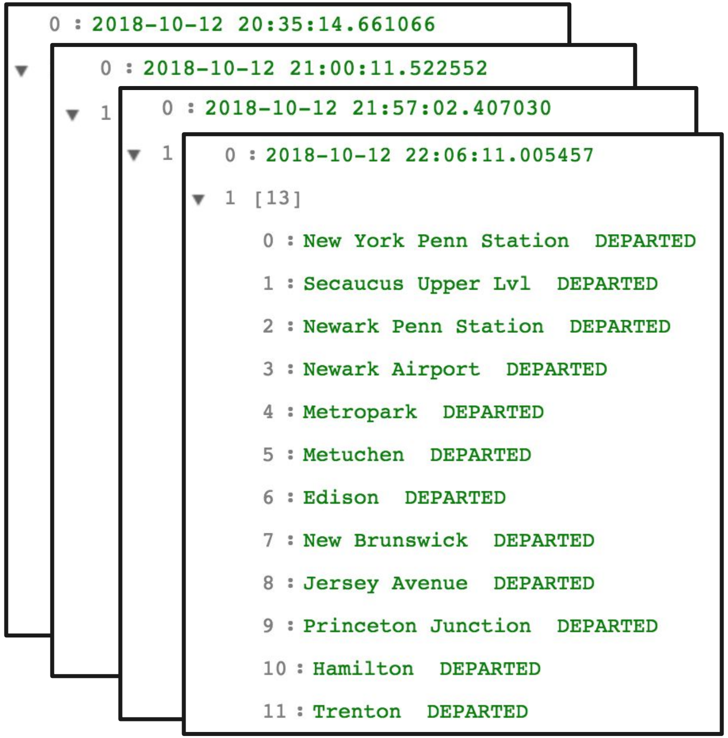
Scraper



```
0 : 2018-10-12 20:35:14.661066
▼
0 : 2018-10-12 21:00:11.522552
▼
0 : 2018-10-12 21:57:02.407030
▼ 1
0 : 2018-10-12 22:06:11.005457
▼ 1 [13]
    0 : New York Penn Station  DEPARTED
    1 : Secaucus Upper Lvl  DEPARTED
    2 : Newark Penn Station  DEPARTED
    3 : Newark Airport  DEPARTED
    4 : Metropark  DEPARTED
    5 : Metuchen  DEPARTED
    6 : Edison  DEPARTED
    7 : New Brunswick  DEPARTED
    8 : Jersey Avenue  DEPARTED
    9 : Princeton Junction  DEPARTED
    10 : Hamilton  DEPARTED
    11 : Trenton  DEPARTED
```

Parser

```
for each status_page in train_file:
    update_state_machine(status_page)
    check_next_state_machine(status_page)
```



```
0 : 2018-10-12 20:35:14.661066
▼
```

```
0 : 2018-10-12 21:00:11.522552
▼ 1
```

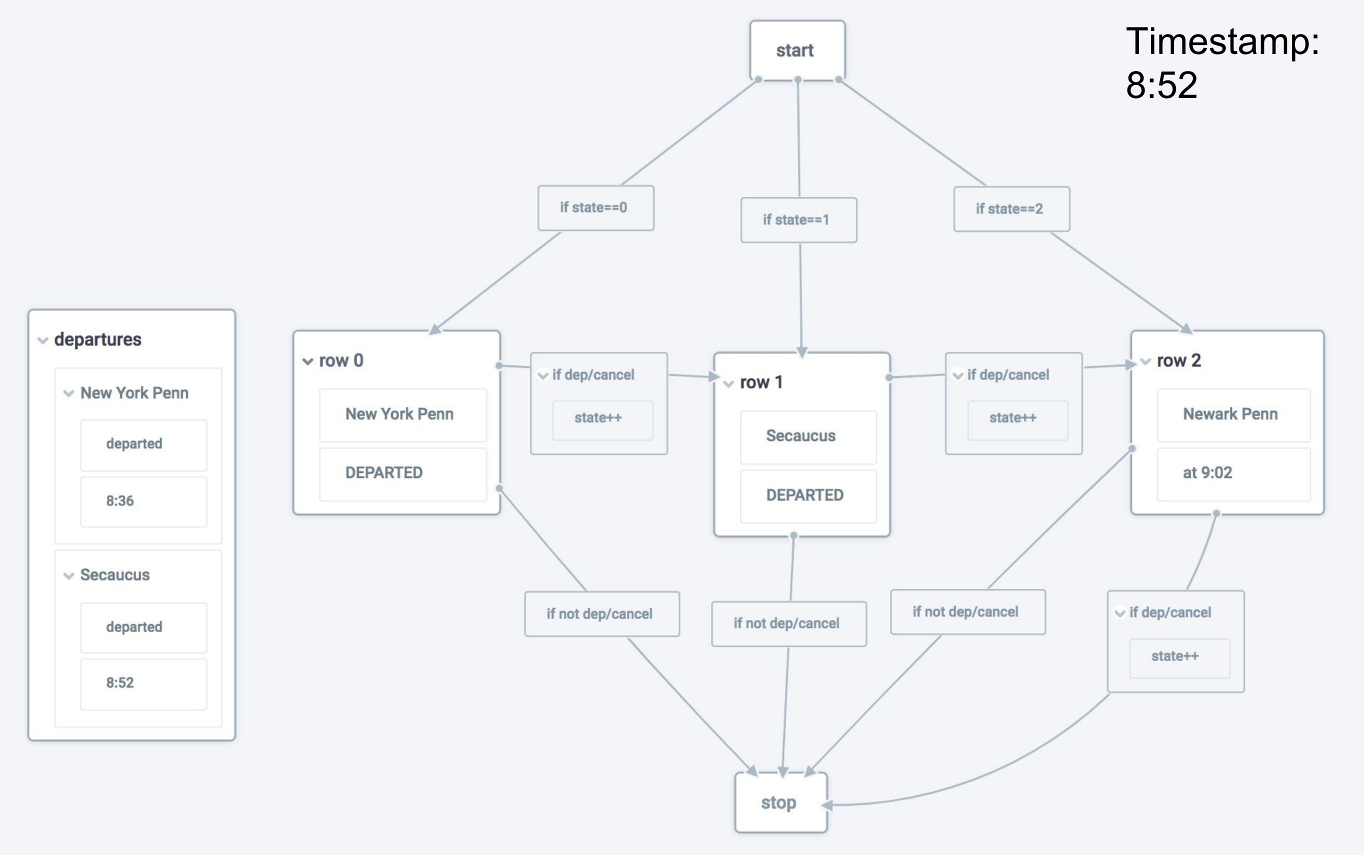
```
0 : 2018-10-12 21:57:02.407030
▼ 1
```

```
0 : 2018-10-12 22:06:11.005457
▼ 1 [13]
     0 : New York Penn Station  DEPARTED
     1 : Secaucus Upper Lvl  DEPARTED
     2 : Newark Penn Station  DEPARTED
     3 : Newark Airport  DEPARTED
     4 : Metropark  DEPARTED
     5 : Metuchen  DEPARTED
     6 : Edison  DEPARTED
     7 : New Brunswick  DEPARTED
     8 : Jersey Avenue  DEPARTED
     9 : Princeton Junction  DEPARTED
    10 : Hamilton  DEPARTED
    11 : Trenton  DEPARTED
```

Parser

check_next()

- checks to see if next row is marked departed or cancelled



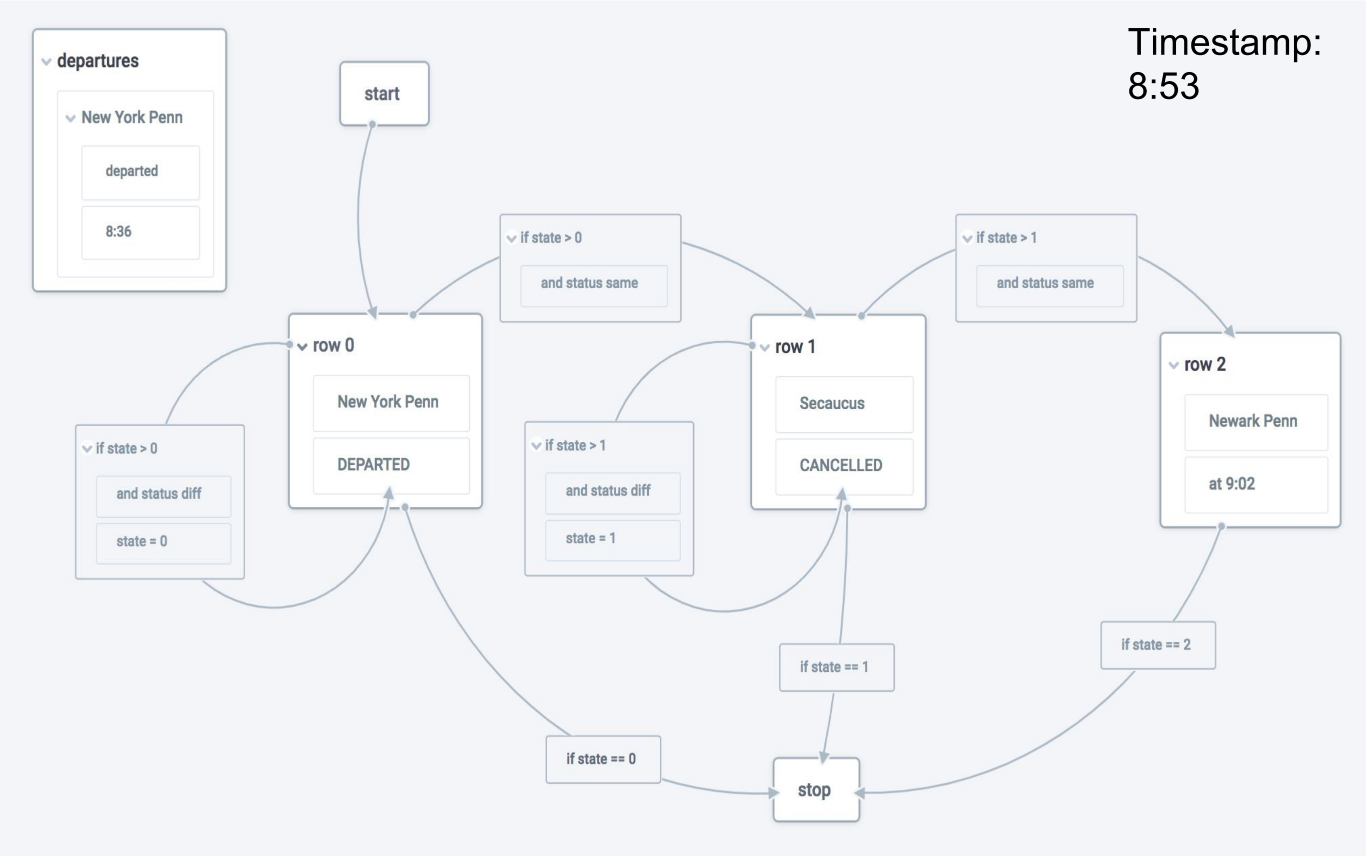
Timestamp:
8:52

Parser

update()

- starting at the first row, checks to see if status has gone from departed to not departed or cancelled

- resets state to before modified row

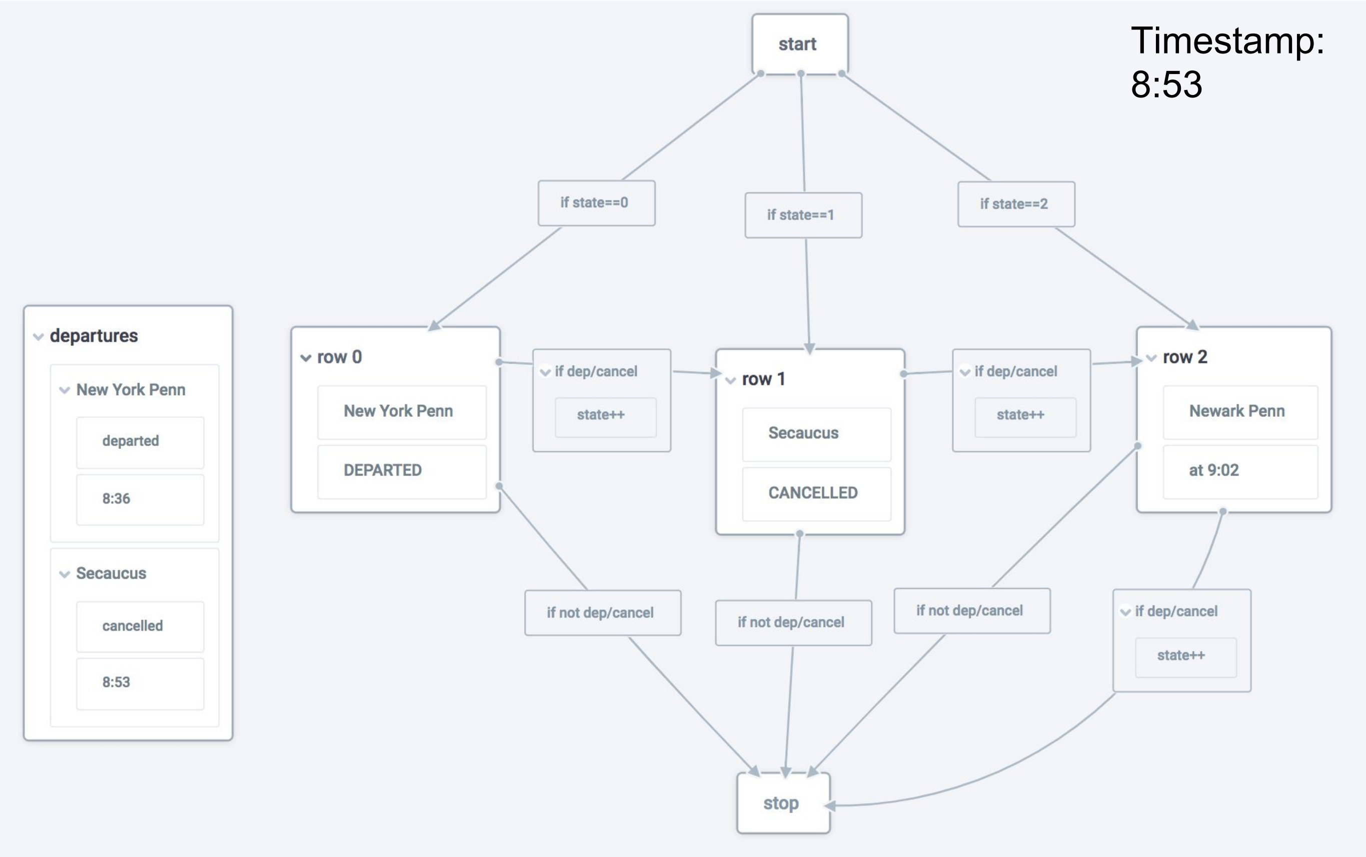


Timestamp: 8:53

Parser

check_next()

- start from state==1 again



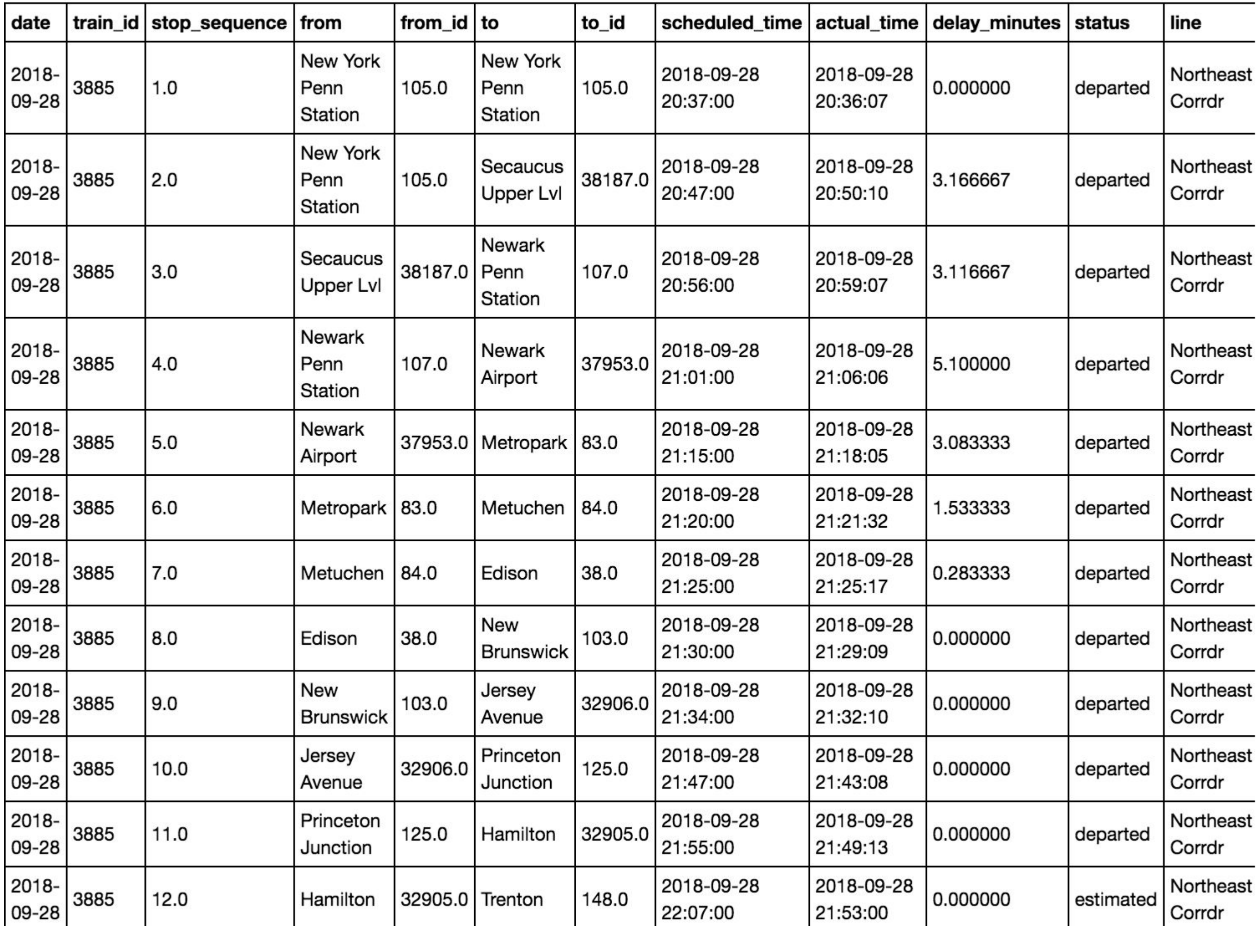
Timestamp: 8:53

Output

Join GTFS schedule

Pre-compute delay

Output monthly CSVs



date	train_id	stop_sequence	from	from_id	to	to_id	scheduled_time	actual_time	delay_minutes	status	line
2018-09-28	3885	1.0	New York Penn Station	105.0	New York Penn Station	105.0	2018-09-28 20:37:00	2018-09-28 20:36:07	0.000000	departed	Northeast Corrdr
2018-09-28	3885	2.0	New York Penn Station	105.0	Secaucus Upper Lvl	38187.0	2018-09-28 20:47:00	2018-09-28 20:50:10	3.166667	departed	Northeast Corrdr
2018-09-28	3885	3.0	Secaucus Upper Lvl	38187.0	Newark Penn Station	107.0	2018-09-28 20:56:00	2018-09-28 20:59:07	3.116667	departed	Northeast Corrdr
2018-09-28	3885	4.0	Newark Penn Station	107.0	Newark Airport	37953.0	2018-09-28 21:01:00	2018-09-28 21:06:06	5.100000	departed	Northeast Corrdr
2018-09-28	3885	5.0	Newark Airport	37953.0	Metropark	83.0	2018-09-28 21:15:00	2018-09-28 21:18:05	3.083333	departed	Northeast Corrdr
2018-09-28	3885	6.0	Metropark	83.0	Metuchen	84.0	2018-09-28 21:20:00	2018-09-28 21:21:32	1.533333	departed	Northeast Corrdr
2018-09-28	3885	7.0	Metuchen	84.0	Edison	38.0	2018-09-28 21:25:00	2018-09-28 21:25:17	0.283333	departed	Northeast Corrdr
2018-09-28	3885	8.0	Edison	38.0	New Brunswick	103.0	2018-09-28 21:30:00	2018-09-28 21:29:09	0.000000	departed	Northeast Corrdr
2018-09-28	3885	9.0	New Brunswick	103.0	Jersey Avenue	32906.0	2018-09-28 21:34:00	2018-09-28 21:32:10	0.000000	departed	Northeast Corrdr
2018-09-28	3885	10.0	Jersey Avenue	32906.0	Princeton Junction	125.0	2018-09-28 21:47:00	2018-09-28 21:43:08	0.000000	departed	Northeast Corrdr
2018-09-28	3885	11.0	Princeton Junction	125.0	Hamilton	32905.0	2018-09-28 21:55:00	2018-09-28 21:49:13	0.000000	departed	Northeast Corrdr
2018-09-28	3885	12.0	Hamilton	32905.0	Trenton	148.0	2018-09-28 22:07:00	2018-09-28 21:53:00	0.000000	estimated	Northeast Corrdr

Data quantity and quality

- Monthly CSVs for March 2018 through September 2018
  - 156,000+ trains overall, 137,000+ NJ Transit, 19,000+ Amtrak
  - Stop-level, minute resolution data

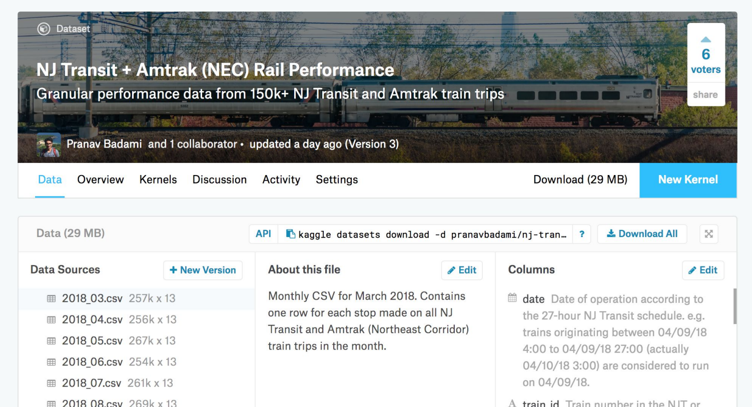
- Data from 98.6% of trains was captured correctly so far
  - Empty JSON files
  - Parsing errors

Data quantity and quality

- Monthly CSVs for March 2018 through September 2018
  - 156,000+ trains overall, 137,000+ NJ Transit, 19,000+ Amtrak
  - Stop-level, minute resolution data

- Data from 98.6% of trains was captured correctly so far
  - Empty JSON files
  - Parsing errors

- We’ve open sourced all this data on Kaggle!
  - (Scraper/parser on GitHub)



NJ Transit + Amtrak (NEC) Rail Performance

Granular performance data from 150k+ NJ Transit and Amtrak train trips

Pranav Badami and 1 collaborator • updated a day ago (Version 3)

Data (29 MB)

```
kaggle datasets download -d pranavbadami/nj-tran…
```

Data Sources

2018_03.csv	257k x 13
2018_04.csv	256k x 13
2018_05.csv	267k x 13
2018_06.csv	254k x 13
2018_07.csv	261k x 13
2018_08.csv	269k x 13

About this file

Monthly CSV for March 2018. Contains one row for each stop made on all NJ Transit and Amtrak (Northeast Corridor) train trips in the month.

Columns

date Date of operation according to the 27-hour NJ Transit schedule. e.g. trains originating between 04/09/18 4:00 to 04/09/18 27:00 (actually 04/10/18 3:00) are considered to run on 04/09/18.

train_id Train number in the NJT or

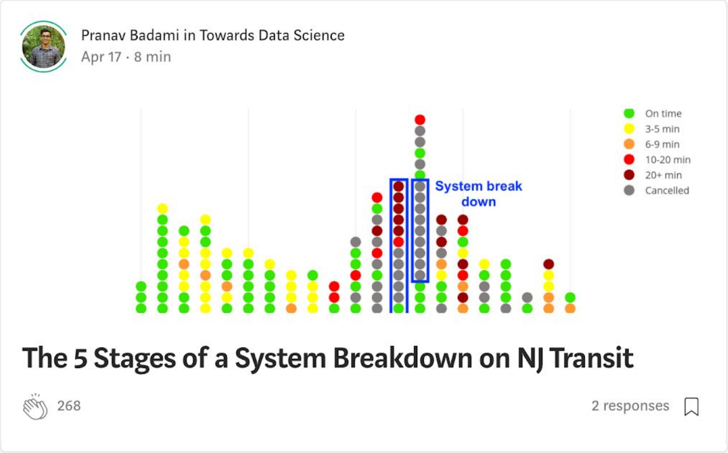
What have we done with this data?

Published on Medium

The 5 Stages of a System Breakdown on NJ Transit (Pranav)

What are the chances that NJ Transit will cause you to miss the Dinky? (Michael)

How data can help fix NJ Transit (Pranav)

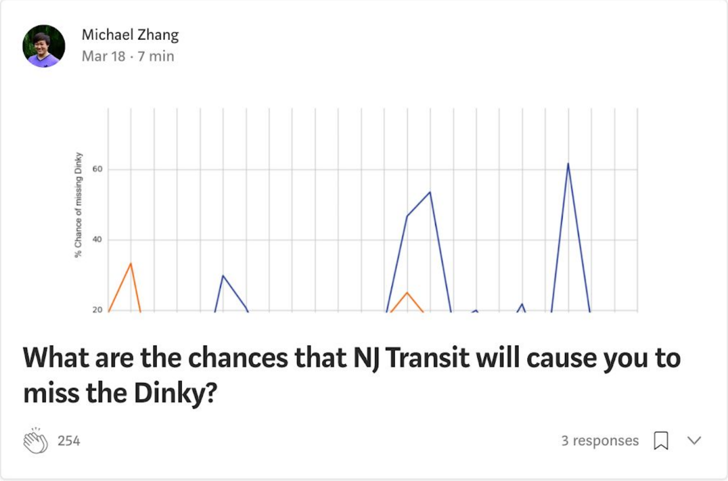


Pranav Badami in Towards Data Science
Apr 17 · 8 min

The 5 Stages of a System Breakdown on NJ Transit

268

2 responses



Michael Zhang
Mar 18 · 7 min

What are the chances that NJ Transit will cause you to miss the Dinky?

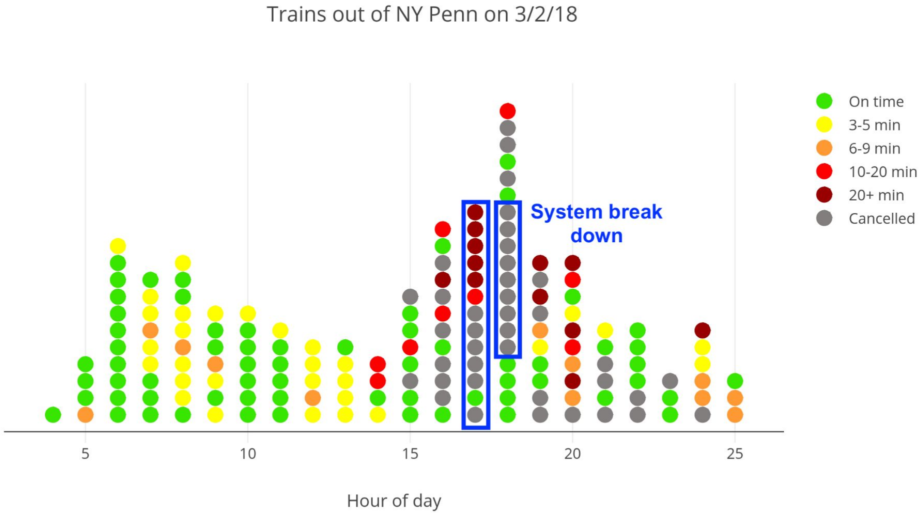
254

3 responses

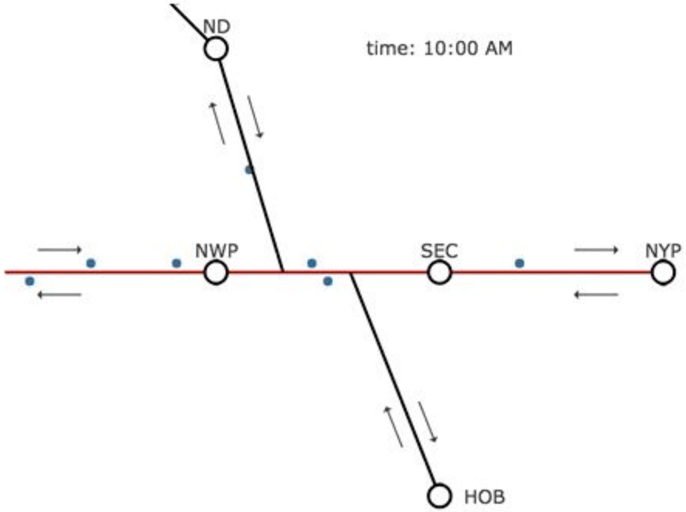
“The 5 Stages of a System Breakdown on NJ Transit”

Step 1: Nor’easter

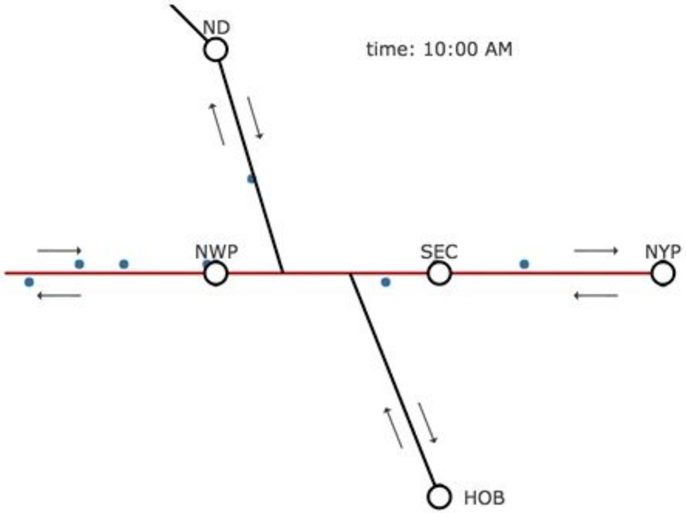
Result?



The calm before (during) the storm



Scheduled trains between Newark Penn and New York Penn (3/2/18)



Actual trains between Newark Penn and New York Penn (3/2/18)

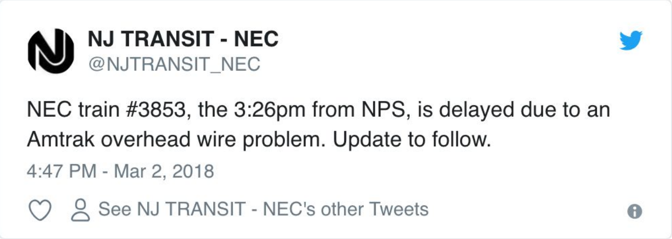
System (and communication) breakdowns

Timeline

12:00 PM: Winds start reaching over 25mph

1:00-2:00 PM(?): Wire goes down; delays

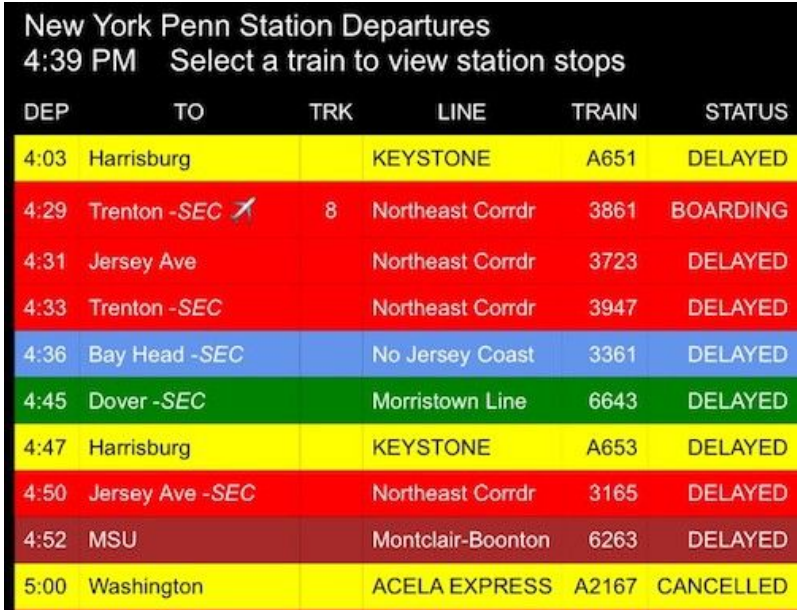
4:47 PM: Tweet about "overhead wire problem"



NJ TRANSIT - NEC
@NJTRANSIT_NEC

NEC train #3853, the 3:26pm from NPS, is delayed due to an Amtrak overhead wire problem. Update to follow.

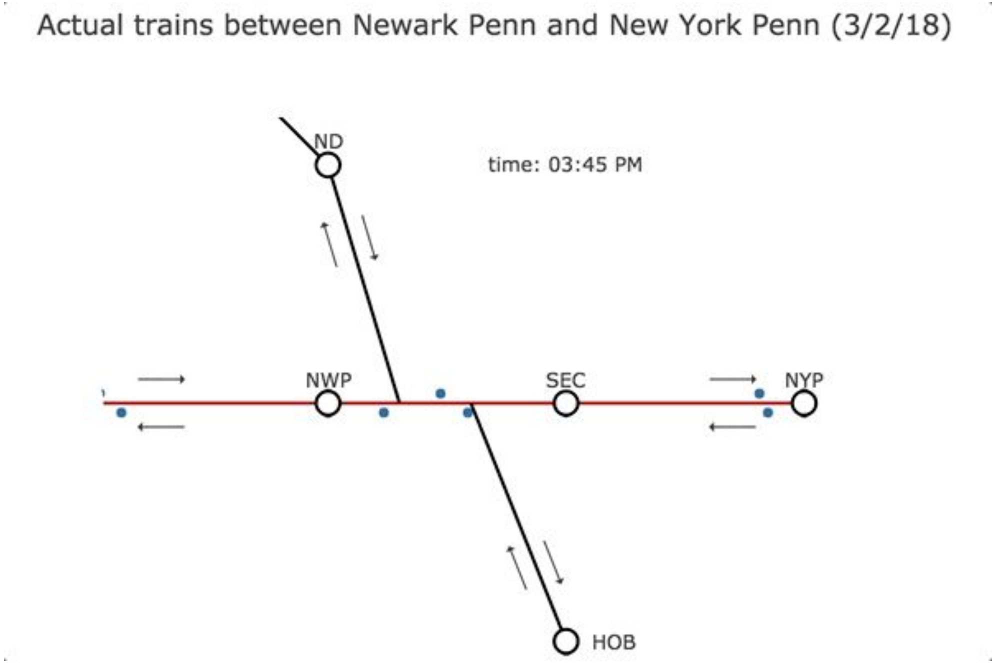
4:47 PM - Mar 2, 2018



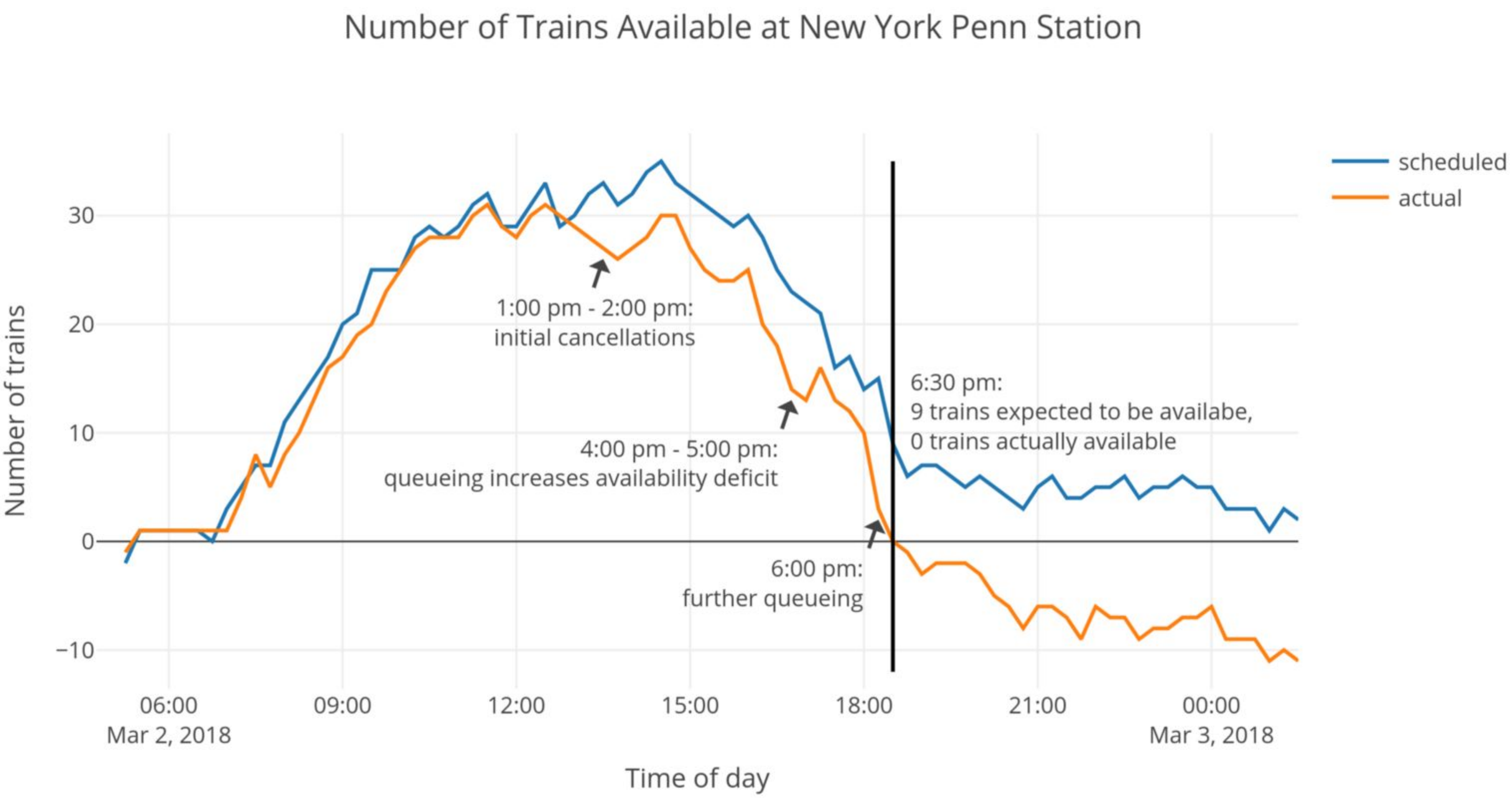
New York Penn Station Departures
4:39 PM Select a train to view station stops

DEP	TO	TRK	LINE	TRAIN	STATUS
4:03	Harrisburg		KEYSTONE	A651	DELAYED
4:29	Trenton -*SEC* ✈	8	Northeast Corrdr	3861	BOARDING
4:31	Jersey Ave		Northeast Corrdr	3723	DELAYED
4:33	Trenton -*SEC*		Northeast Corrdr	3947	DELAYED
4:36	Bay Head -*SEC*		No Jersey Coast	3361	DELAYED
4:45	Dover -*SEC*		Morristown Line	6643	DELAYED
4:47	Harrisburg		KEYSTONE	A653	DELAYED
4:50	Jersey Ave -*SEC*		Northeast Corrdr	3165	DELAYED
4:52	MSU		Montclair-Boonton	6263	DELAYED
5:00	Washington		ACELA EXPRESS	A2167	CANCELLED

Queueing begins…



Actual trains between Newark Penn and New York Penn (3/2/18)



1:00 pm - 2:00 pm:
initial cancellations

4:00 pm - 5:00 pm:
queueing increases availability deficit

6:00 pm:
further queueing

6:30 pm:
9 trains expected to be availabe,
0 trains actually available

So, what’s next?

Calls to action

Phil Murphy responds...

- NJ Transit audit: https://www.nj.gov/governor/docs/20181005NJTransitFinalReport.pdf

Analyze this data

- Kaggle!
- Incorporating external data sources
- A public facing “report card” (s/o to Bus Turnaround!)

And more articles on the way!

Thank you!

Pranav Badami

Twitter: @Pranav_Badami

Michael Zhang

Twitter: @mzhang13

Check out the data and get involved!

Kaggle:
https://www.kaggle.com/pranavbadami/nj-transit-amtrak-nec-performance

GitHub:
https://github.com/pranavbadami/njtransit

Medium:
https://medium.com/@pranavbadami
https://medium.com/@mzhang13