



# 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...*



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







# 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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd



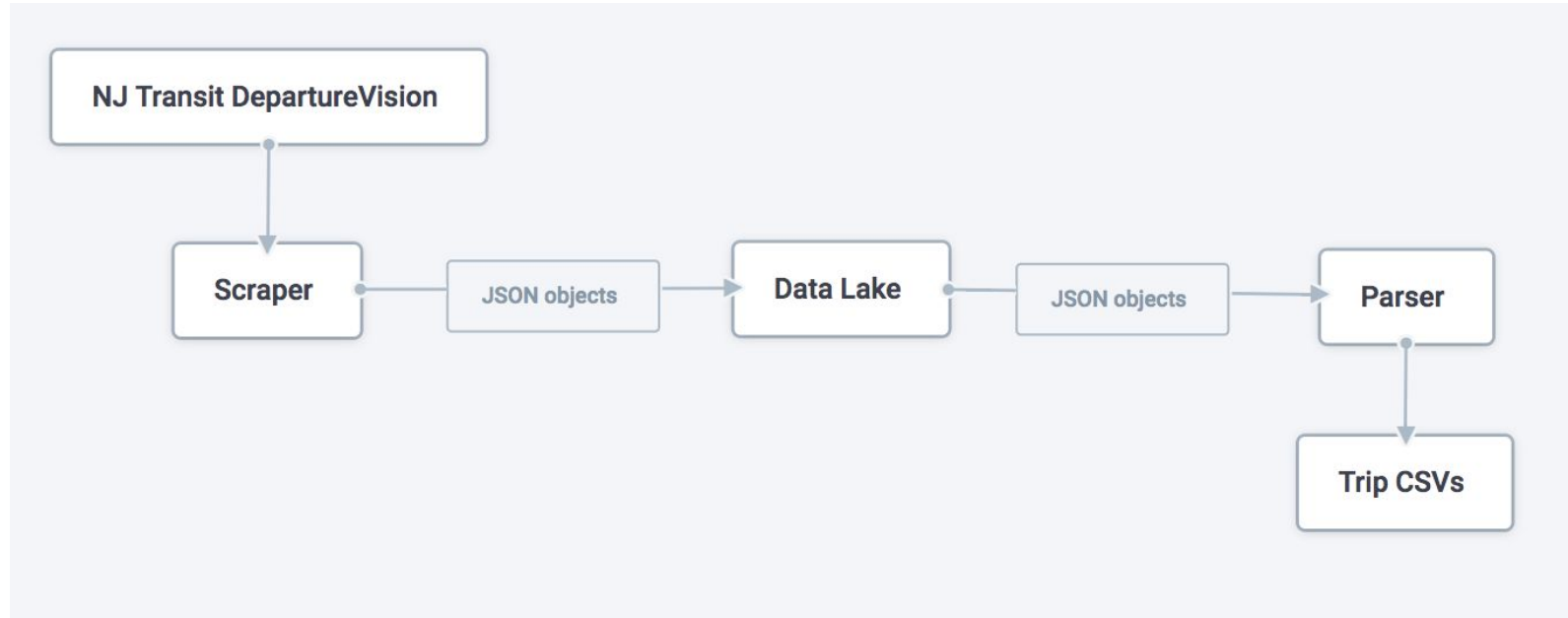
## The Data:

- 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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd

# Data Pipeline



# 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 Corldr	3883	in 2 Min
9:15	Trenton	4	Northeast Corldr	3885	in 18 Min
9:22	NY Penn -SEC →	1	Northeast Corldr	3880	in 26 Min
9:44	Washington	4	REGIONAL	A187	
9:47	Trenton	4	Northeast Corldr	3887	
9:50	NY Penn -SEC →	1	Northeast Corldr	3882	
9:59	New York	1	ACELA EXPRESS	A2126	
10:07	New York	1	REGIONAL	A186	
10:16	Trenton	4	Northeast Corldr	3889	
10:38	Washington	4	REGIONAL	A177	
10:43	Jersey Ave	4	Northeast Corldr	3737	
10:47	New York	1	ACELA EXPRESS	A2128	
10:47	NY Penn -SEC →	1	Northeast Corldr	3886	
10:52	Trenton	4	Northeast Corldr	3891	
11:17	Trenton	4	Northeast Corldr	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
8:58	Trenton	4	Northeast Corldr	3886	in 2 Min
9:15	Trenton	4	Northeast Corldr	3885	in 18 Min
9:22	NY Penn -SEC →	1	Northeast Corldr	3888	in 25 Min
9:44	Washington	4	REGIONAL	A187	
9:47	Trenton	4	Northeast Corldr	3887	
9:50	NY Penn -SEC →	1	Northeast Corldr	3882	
9:59	New York	1	ACELA EXPRESS	A2126	
10:07	New York	1	REGIONAL	A186	
10:16	Trenton	4	Northeast Corldr	3889	
10:38	Washington	4	REGIONAL	A177	
10:43	Jersey Ave	4	Northeast Corldr	3737	
10:47	New York	1	ACELA EXPRESS	A2128	
10:47	NY Penn -SEC →	1	Northeast Corldr	3886	
10:52	Trenton	4	Northeast Corldr	3891	
11:17	Trenton	4	Northeast Corldr	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](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



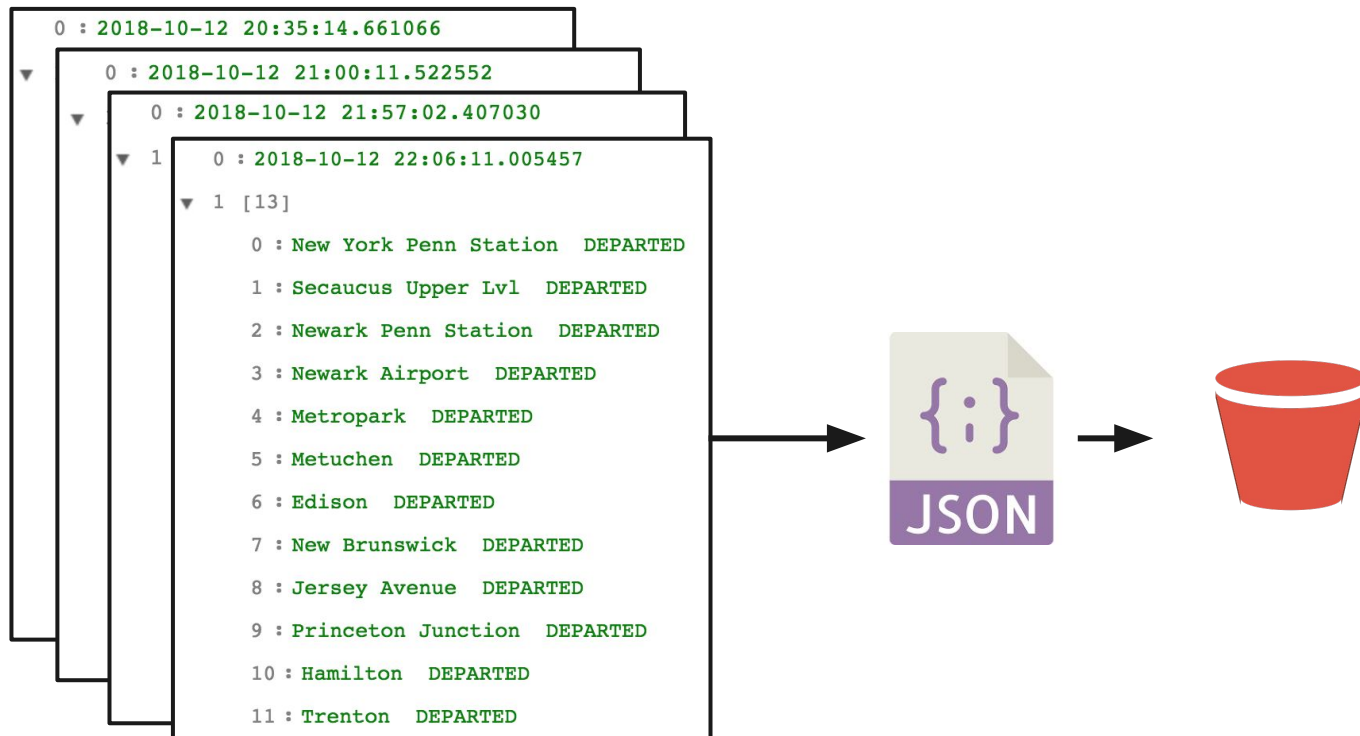
# Scraper

Every minute:

```
0 : 2018-10-12 20:35:14.661066
▼ 1 [13]
0
1 ▼ 1 [13]
2 0
3 1
4 2
5 3
6 4
7 5
8 6
9 7
10 8
11 9
12 10
13 11
14 12
15 13
16 14
17 15
18 16
19 17
20 18
21 19
22 20
23 21
24 22
25 23
26 24
27 25
28 26
29 27
30 28
31 29
32 30
33 31
34 32
35 33
36 34
37 35
38 36
39 37
40 38
41 39
42 40
43 41
44 42
45 43
46 44
47 45
48 46
49 47
50 48
51 49
52 50
53 51
54 52
55 53
56 54
57 55
58 56
59 57
60 58
61 59
62 60
63 61
64 62
65 63
66 64
67 65
68 66
69 67
70 68
71 69
72 70
73 71
74 72
75 73
76 74
77 75
78 76
79 77
80 78
81 79
82 80
83 81
84 82
85 83
86 84
87 85
88 86
89 87
90 88
91 89
92 90
93 91
94 92
95 93
96 94
97 95
98 96
99 97
100 98
101 99
102 100
103 101
104 102
105 103
106 104
107 105
108 106
109 107
110 108
111 109
112 110
113 111
114 112
115 113
116 114
117 115
118 116
119 117
120 118
121 119
122 120
123 121
124 122
125 123
126 124
127 125
128 126
129 127
130 128
131 129
132 130
133 131
134 132
135 133
136 134
137 135
138 136
139 137
140 138
141 139
142 140
143 141
144 142
145 143
146 144
147 145
148 146
149 147
150 148
151 149
152 150
153 151
154 152
155 153
156 154
157 155
158 156
159 157
160 158
161 159
162 160
163 161
164 162
165 163
166 164
167 165
168 166
169 167
170 168
171 169
172 170
173 171
174 172
175 173
176 174
177 175
178 176
179 177
180 178
181 179
182 180
183 181
184 182
185 183
186 184
187 185
188 186
189 187
190 188
191 189
192 190
193 191
194 192
195 193
196 194
197 195
198 196
199 197
200 198
201 199
202 200
203 201
204 202
205 203
206 204
207 205
208 206
209 207
210 208
211 209
212 210
213 211
214 212
215 213
216 214
217 215
218 216
219 217
220 218
221 219
222 220
223 221
224 222
225 223
226 224
227 225
228 226
229 227
230 228
231 229
232 230
233 231
234 232
235 233
236 234
237 235
238 236
239 237
240 238
241 239
242 240
243 241
244 242
245 243
246 244
247 245
248 246
249 247
250 248
251 249
252 250
253 251
254 252
255 253
256 254
257 255
258 256
259 257
260 258
261 259
262 260
263 261
264 262
265 263
266 264
267 265
268 266
269 267
270 268
271 269
272 270
273 271
274 272
275 273
276 274
277 275
278 276
279 277
280 278
281 279
282 280
283 281
284 282
285 283
286 284
287 285
288 286
289 287
290 288
291 289
292 290
293 291
294 292
295 293
296 294
297 295
298 296
299 297
300 298
301 299
302 300
303 301
304 302
305 303
306 304
307 305
308 306
309 307
310 308
311 309
312 310
313 311
314 312
315 313
316 314
317 315
318 316
319 317
320 318
321 319
322 320
323 321
324 322
325 323
326 324
327 325
328 326
329 327
330 328
331 329
332 330
333 331
334 332
335 333
336 334
337 335
338 336
339 337
340 338
341 339
342 340
343 341
344 342
345 343
346 344
347 345
348 346
349 347
350 348
351 349
352 350
353 351
354 352
355 353
356 354
357 355
358 356
359 357
360 358
361 359
362 360
363 361
364 362
365 363
366 364
367 365
368 366
369 367
370 368
371 369
372 370
373 371
374 372
375 373
376 374
377 375
378 376
379 377
380 378
381 379
382 380
383 381
384 382
385 383
386 384
387 385
388 386
389 387
390 388
391 389
392 390
393 391
394 392
395 393
396 394
397 395
398 396
399 397
400 398
401 399
402 400
403 401
404 402
405 403
406 404
407 405
408 406
409 407
410 408
411 409
412 410
413 411
414 412
415 413
416 414
417 415
418 416
419 417
420 418
421 419
422 420
423 421
424 422
425 423
426 424
427 425
428 426
429 427
430 428
431 429
432 430
433 431
434 432
435 433
436 434
437 435
438 436
439 437
440 438
441 439
442 440
443 441
444 442
445 443
446 444
447 445
448 446
449 447
450 448
451 449
452 450
453 451
454 452
455 453
456 454
457 455
458 456
459 457
460 458
461 459
462 460
463 461
464 462
465 463
466 464
467 465
468 466
469 467
470 468
471 469
472 470
473 471
474 472
475 473
476 474
477 475
478 476
479 477
480 478
481 479
482 480
483 481
484 482
485 483
486 484
487 485
488 486
489 487
490 488
491 489
492 490
493 491
494 492
495 493
496 494
497 495
498 496
499 497
500 498
501 499
502 500
503 501
504 502
505 503
506 504
507 505
508 506
509 507
510 508
511 509
512 510
513 511
514 512
515 513
516 514
517 515
518 516
519 517
520 518
521 519
522 520
523 521
524 522
525 523
526 524
527 525
528 526
529 527
530 528
531 529
532 530
533 531
534 532
535 533
536 534
537 535
538 536
539 537
540 538
541 539
542 540
543 541
544 542
545 543
546 544
547 545
548 546
549 547
550 548
551 549
552 550
553 551
554 552
555 553
556 554
557 555
558 556
559 557
560 558
561 559
562 560
563 561
564 562
565 563
566 564
567 565
568 566
569 567
570 568
571 569
572 570
573 571
574 572
575 573
576 574
577 575
578 576
579 577
580 578
581 579
582 580
583 581
584 582
585 583
586 584
587 585
588 586
589 587
590 588
591 589
592 590
593 591
594 592
595 593
596 594
597 595
598 596
599 597
600 598
601 599
602 600
603 601
604 602
605 603
606 604
607 605
608 606
609 607
610 608
611 609
612 610
613 611
614 612
615 613
616 614
617 615
618 616
619 617
620 618
621 619
622 620
623 621
624 622
625 623
626 624
627 625
628 626
629 627
630 628
631 629
632 630
633 631
634 632
635 633
636 634
637 635
638 636
639 637
640 638
641 639
642 640
643 641
644 642
645 643
646 644
647 645
648 646
649 647
650 648
651 649
652 650
653 651
654 652
655 653
656 654
657 655
658 656
659 657
660 658
661 659
662 660
663 661
664 662
665 663
666 664
667 665
668 666
669 667
670 668
671 669
672 670
673 671
674 672
675 673
676 674
677 675
678 676
679 677
680 678
681 679
682 680
683 681
684 682
685 683
686 684
687 685
688 686
689 687
690 688
691 689
692 690
693 691
694 692
695 693
696 694
697 695
698 696
699 697
700 698
701 699
702 700
703 701
704 702
705 703
706 704
707 705
708 706
709 707
710 708
711 709
712 710
713 711
714 712
715 713
716 714
717 715
718 716
719 717
720 718
721 719
722 720
723 721
724 722
725 723
726 724
727 725
728 726
729 727
730 728
731 729
732 730
733 731
734 732
735 733
736 734
737 735
738 736
739 737
740 738
741 739
742 740
743 741
744 742
745 743
746 744
747 745
748 746
749 747
750 748
751 749
752 750
753 751
754 752
755 753
756 754
757 755
758 756
759 757
760 758
761 759
762 760
763 761
764 762
765 763
766 764
767 765
768 766
769 767
770 768
771 769
772 770
773 771
774 772
775 773
776 774
777 775
778 776
779 777
780 778
781 779
782 780
783 781
784 782
785 783
786 784
787 785
788 786
789 787
790 788
791 789
792 790
793 791
794 792
795 793
796 794
797 795
798 796
799 797
800 798
801 799
802 800
803 801
804 802
805 803
806 804
807 805
808 806
809 807
810 808
811 809
812 810
813 811
814 812
815 813
816 814
817 815
818 816
819 817
820 818
821 819
822 820
823 821
824 822
825 823
826 824
827 825
828 826
829 827
830 828
831 829
832 830
833 831
834 832
835 833
836 834
837 835
838 836
839 837
840 838
841 839
842 840
843 841
844 842
845 843
846 844
847 845
848 846
849 847
850 848
851 849
852 850
853 851
854 852
855 853
856 854
857 855
858 856
859 857
860 858
861 859
862 860
863 861
864 862
865 863
866 864
867 865
868 866
869 867
870 868
871 869
872 870
873 871
874 872
875 873
876 874
877 875
878 876
879 877
880 878
881 879
882 880
883 881
884 882
885 883
886 884
887 885
888 886
889 887
890 888
891 889
892 890
893 891
894 892
895 893
896 894
897 895
898 896
899 897
900 898
901 899
902 900
903 901
904 902
905 903
906 904
907 905
908 906
909 907
910 908
911 909
912 910
913 911
914 912
915 913
916 914
917 915
918 916
919 917
920 918
921 919
922 920
923 921
924 922
925 923
926 924
927 925
928 926
929 927
930 928
931 929
932 930
933 931
934 932
935 933
936 934
937 935
938 936
939 937
940 938
941 939
942 940
943 941
944 942
945 943
946 944
947 945
948 946
949 947
950 948
951 949
952 950
953 951
954 952
955 953
956 954
957 955
958 956
959 957
960 958
961 959
962 960
963 961
964 962
965 963
966 964
967 965
968 966
969 967
970 968
971 969
972 970
973 971
974 972
975 973
976 974
977 975
978 976
979 977
980 978
981 979
982 980
983 981
984 982
985 983
986 984
987 985
988 986
989 987
990 988
991 989
992 990
993 991
994 992
995 993
996 994
997 995
998 996
999 997
1000 998
1001 999
1002 1000
1003 1001
1004 1002
1005 1003
1006 1004
1007 1005
1008 1006
1009 1007
1010 1008
1011 1009
1012 1010
1013 1011
1014 1012
1015 1013
1016 1014
1017 1015
1018 1016
1019 1017
1020 1018
1021 1019
1022 1020
1023 1021
1024 1022
1025 1023
1026 1024
1027 1025
1028 1026
1029 1027
1030 1028
1031 1029
1032 1030
1033 1031
1034 1032
1035 1033
1036 1034
1037 1035
1038 1036
1039 1037
1040 1038
1041 1039
1042 1040
1043 1041
1044 1042
1045 1043
1046 1044
1047 1045
1048 1046
1049 1047
1050 1048
1051 1049
1052 1050
1053 1051
1054 1052
1055 1053
1056 1054
1057 1055
1058 1056
1059 1057
1060 1058
1061 1059
1062 1060
1063 1061
1064 1062
1065 1063
1066 1064
1067 1065
1068 1066
1069 1067
1070 1068
1071 1069
1072 1070
1073 1071
1074 1072
1075 1073
1076 1074
1077 1075
1078 1076
1079 1077
1080 1078
1081 1079
1082 1080
1083 1081
1084 1082
1085 1083
1086 1084
1087 1085
1088 1086
1089 1087
1090 1088
1091 1089
1092 1090
1093 1091
1094 1092
1095 1093
1096 1094
1097 1095
1098 1096
1099 1097
1100 1098
1101 1099
1102 1100
1103 1101
1104 1102
1105 1103
1106 1104
1107 1105
1108 1106
1109 1107
1110 1108
1111 1109
1112 1110
1113 1111
1114 1112
1115 1113
1116 1114
1117 1115
1118 1116
1119 1117
1120 1118
1121 1119
1122 1120
1123 1121
1124 1122
1125 1123
1126 1124
1127 1125
1128 1126
1129 1127
1130 1128
1131 1129
1132 1130
1133 1131
1134 1132
1135 1133
1136 1134
1137 1135
1138 1136
1139 1137
1140 1138
1141 1139
1142 1140
1143 1141
1144 1142
1145 1143
1146 1144
1147 1145
1148 1146
1149 1147
1150 1148
1151 1149
1152 1150
1153 1151
1154 1152
1155 1153
1156 1154
1157 1155
1158 1156
1159 1157
1160 1158
1161 1159
1162 1160
1163 1161
1164 1162
1165 1163
1166 1164
1167 1165
1168 1166
1169 1167
1170 1168
1171 1169
1172 1170
1173 1171
1174 1172
1175 1173
1176 1174
1177 1175
1178 1176
1179 1177
1180 1178
1181 1179
1182 1180
1183 1181
1184 1182
1185 1183
1186 1184
1187 1185
1188 1186
1189 1187
1190 1188
1191 1189
1192 1190
1193 1191
1194 1192
1195 1193
1196 1194
1197 1195
1198 1196
1199 1197
1200 1198
1201 1199
1202 1200
1203 1201
1204 1202
1205 1203
1206 1204
1207 1205
1208 1206
1209 1207
1210 1208
1211 1209
1212 1210
1213 1211
1214 1212
1215 1213
1216 1214
1217 1215
1218 1216
1219 1217
1220 1218
1221 1219
1222 1220
1223 1221
1224 1222
1225 1223
1226 1224
1227 1225
1228 1226
1229 1227
1230 1228
1231 1229
1232 1230
1233 1231
1234 1232
1235 1233
1236 1234
1237 1235
1238 1236
1239 1237
1240 1238
1241 1239
1242 1240
1243 1241
1244 1242
1245 1243
1246 1244
1247 1245
1248 1246
1249 1247
1250 1248
1251 1249
1252 1250
1253 1251
1254 1252
1255 1253
1256 1254
1257 1255
1258 1256
1259 1257
1260 1258
1261 1259
1262 1260
1263 1261
1264 1262
1265 1263
1266 1264
1267 1265
1268 1266
1269 1267
1270 1268
1271 1269
1272 1270
1273 1271
1274 1272
1275 1273
1276 1274
1277 1275
1278 1276
1279 1277
1280 1278
1281 1279
1282 1280
1283 1281
1284 1282
1285 1283
1286 1284
1287 1285
1288 1286
1289 1287
1290 1288
1291 1289
1292 1290
1293 1291
1294 1292
1295 1293
1296 1294
1297 1295
1298 1296
1299 1297
1300 1298
1301 1299
1302 1300
1303 1301
1304 1302
1305 1303
1306 1304
1307 1305
1308 1306
1309 1307
1310 1308
1311 1309
1312 1310
1313 1311
1314 1312
1315 1313
1316 1314
1317 1315
1318 1316
1319 1317
1320 1318
1321 1319
1322 1320
1323 1321
1324 1322
1325 1323
1326 1324
1327 1325
1328 1326
1329 1327
1330 1328
1331 1329
1332 1330
1333 1331
1334 1332
1335 1333
1336 1334
1337 1335
1338 1336
1339 1337
1340 1338
1341 1339
1342 1340
1343 1341
1344 1342
1345 1343
1346 1344
1347 1345
1348 1346
1349 1347
1350 1348
1351 1349
1352 1350
1353 1351
1354 1352
1355 1353
1356 1354
1357 1355
1358 1356
1359 1357
1360 1358
1361 1359
1362 1360
1363 1361
1364 1362
1365 1363
1366 1364
1367 1365
1368 1366
1369 1367
1370 1368
1371 1369
1372 1370
1373 1371
1374 1372
1375 1373
1376 1374
1377 1375
1378 1376
1379 1377
1380 1378
1381 1379
1382 1380
1383 1381
1384 1382
1385 1383
1386 1384
1387 1385
1388 1386
1389 1387
1390 1388
1391 1389
1392 1390
1393 1391
1394 1392
1395 1393
1396 1394
1397 1395
1398 1396
1399 1397
1400 1398
1401 1399
1402 1400
1403 1401
1404 1402
1405 1403
1406 1404
1407 1405
1408 1406
1409 1407
1410 1408
1411 1409
1412 1410
1413 1411
1414 1412
1415 1413
1416 1414
1417 1415
1418 1416
1419 1417
1420 1418
1421 1419
1422 1420
1423 1421
1424 1422
1425 1423
1426 1424
1427 1425
1428 1426
1429 1427
1430 1428
1431 1429
1432 1430
1433 1431
1434 1432
1435 1433
1436 1434
1437 1435
1438 1436
1439 1437
1440 1438
1441 1439
1442 1440
1443 144
```



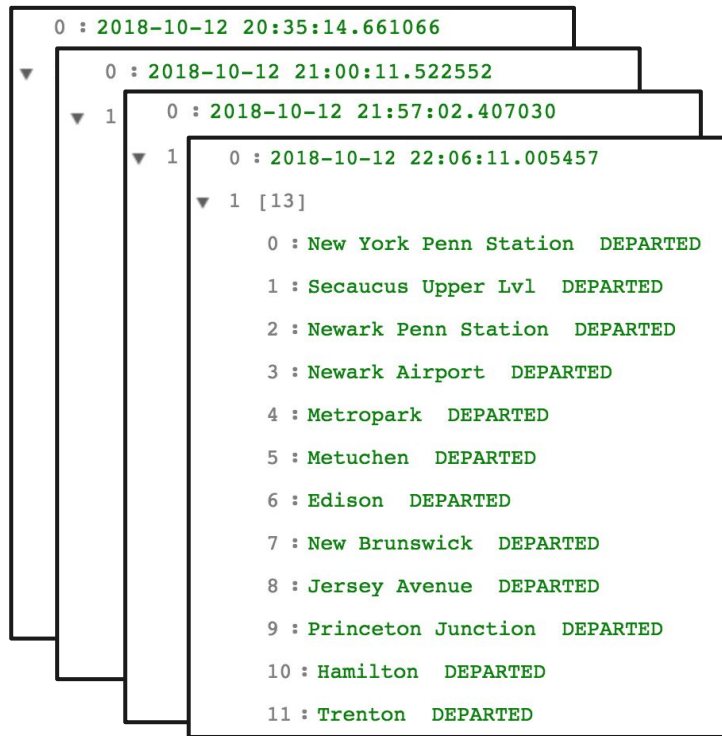
# Scraper





# Parser

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

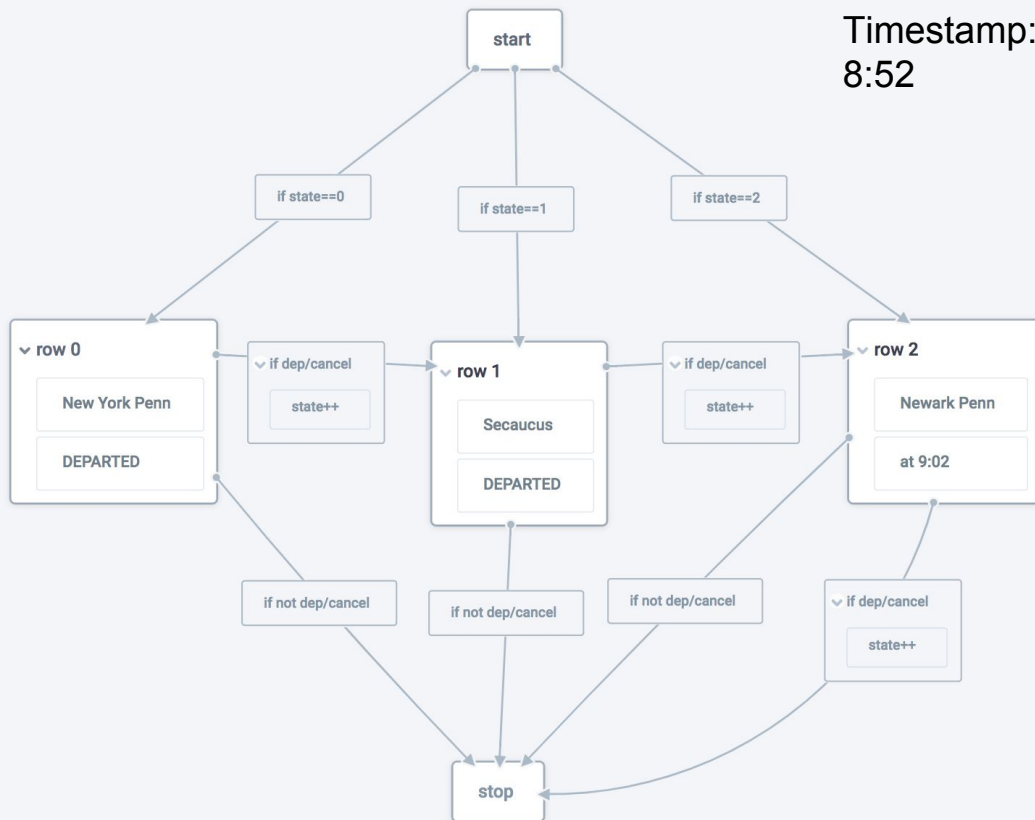


# Parser

check\_next()

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

▼ departures
▼ New York Penn
departed
8:36
▼ Secaucus
departed
8:52

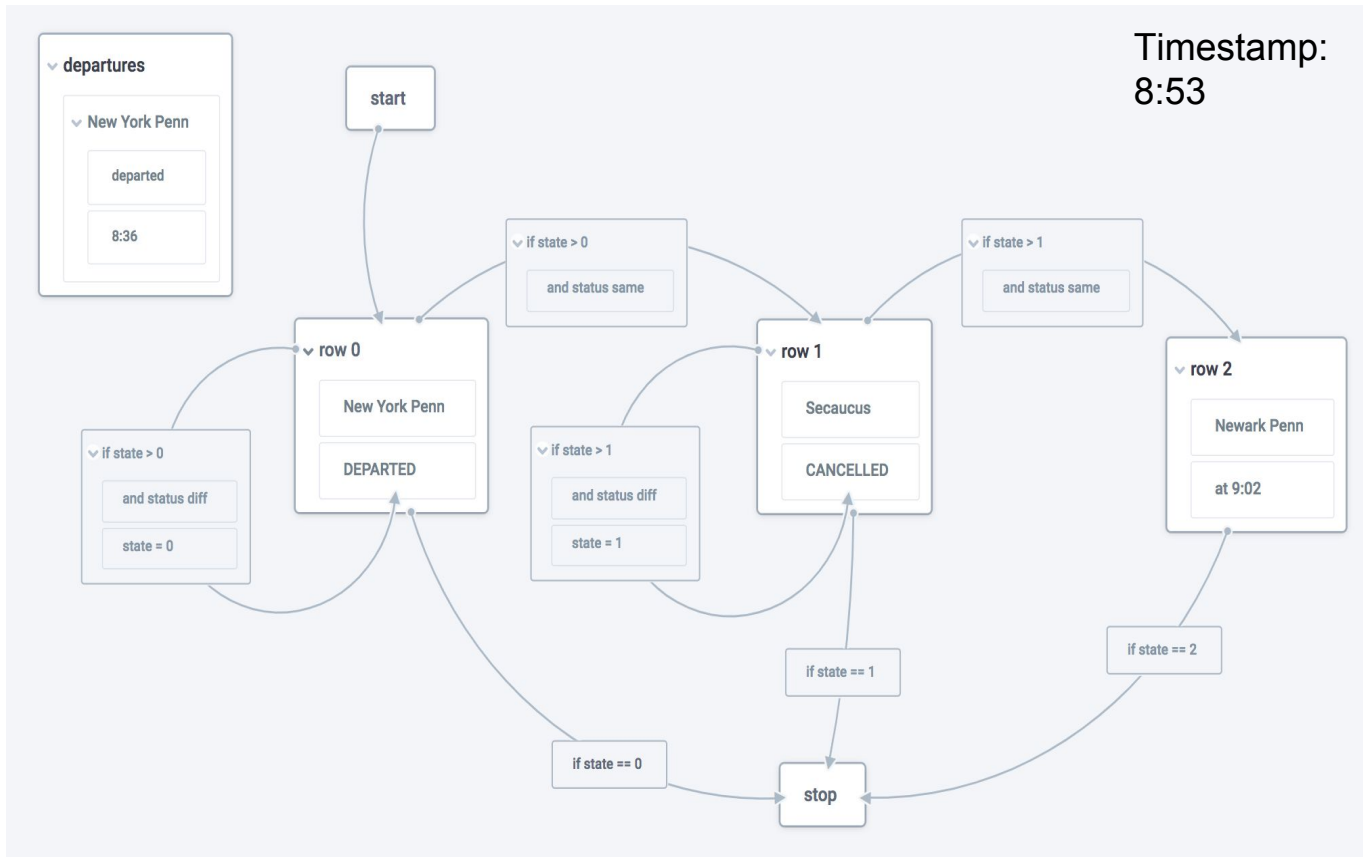


# Parser

update()

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

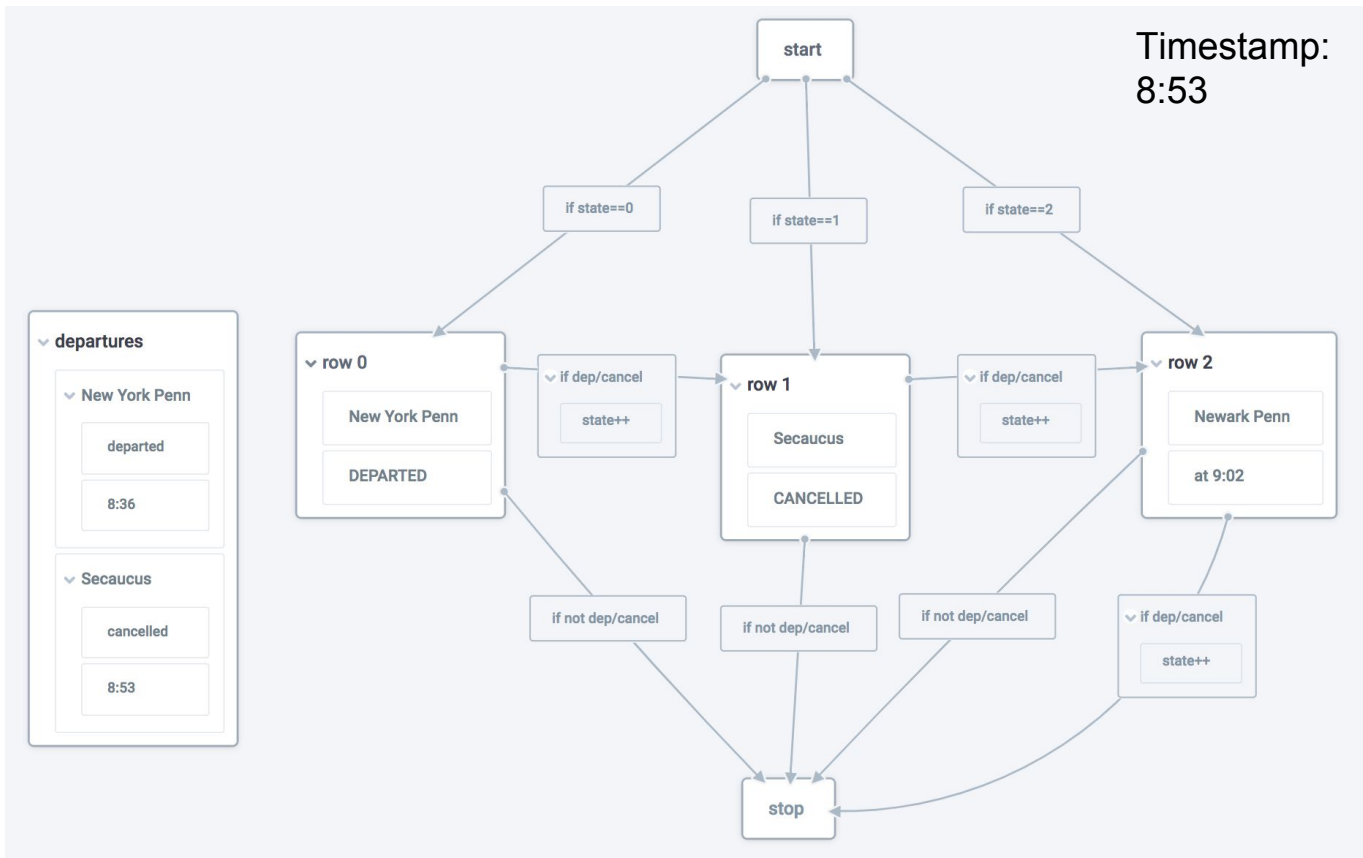
- resets state to before modified row



# Parser

check\_next()

- start from  
state==1 again





# 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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd
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 Corrd



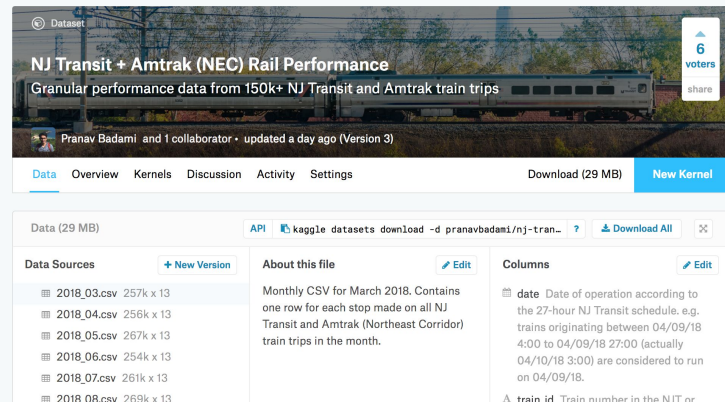
## Data quantity and quality

- 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)



---

**What have we done with this data?**

# Published on Medium

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

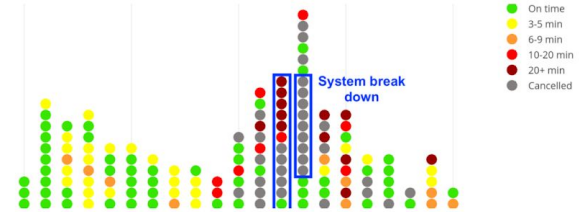
*What are the chances that NJ Transit will cause you to miss the Dinky? (Michael)*

*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

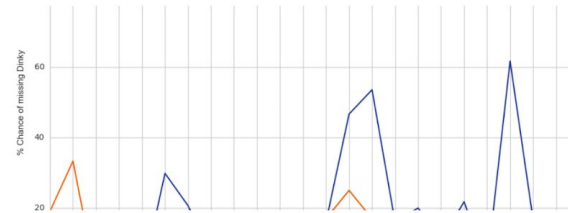
268

2 responses



Michael Zhang

Mar 18 · 7 min



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

254

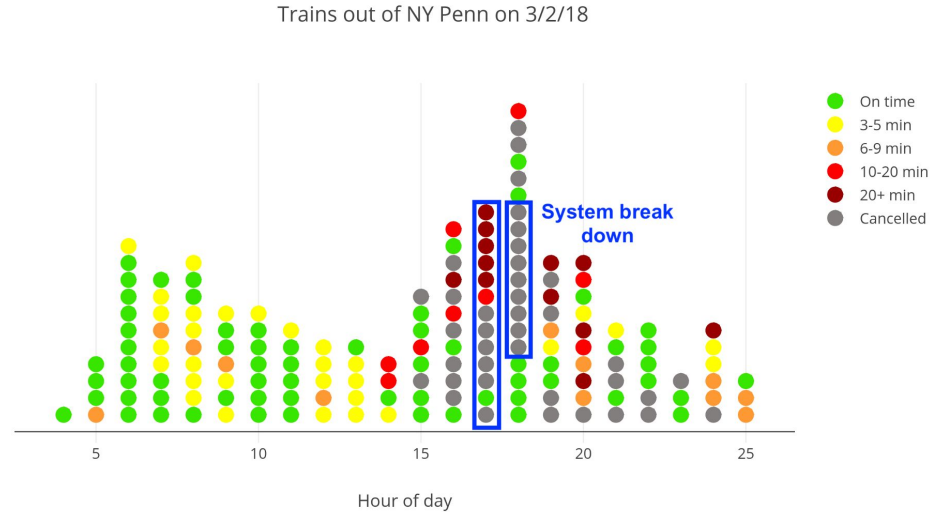
3 responses

# "The 5 Stages of a System Breakdown on NJ Transit"

## Step 1: Nor'easter

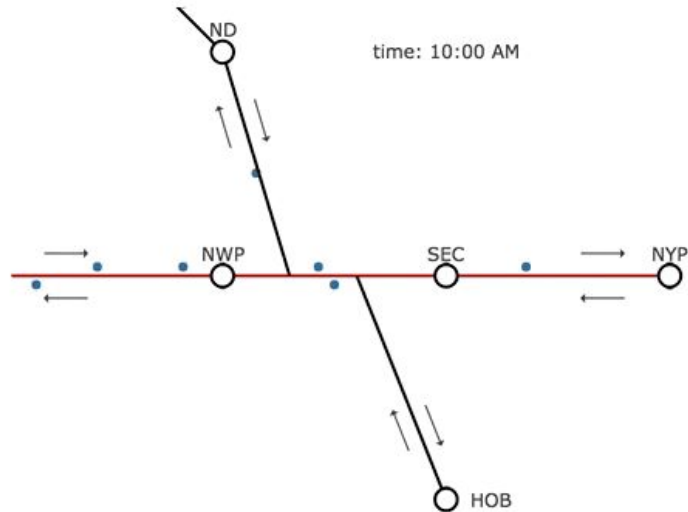


## Result?

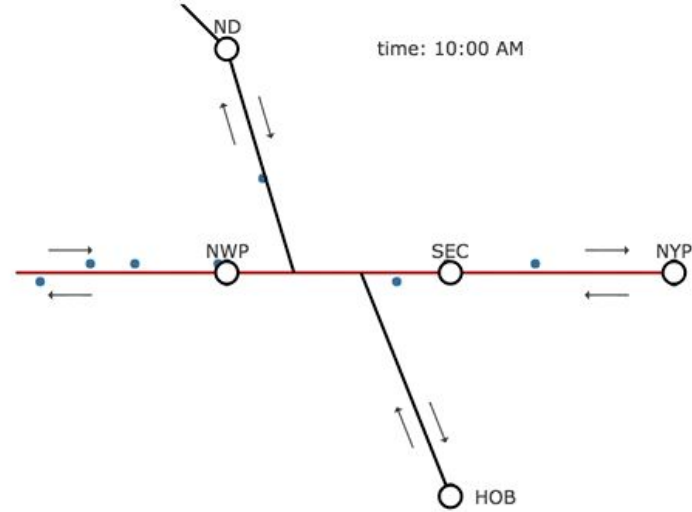


# The calm before (during) the storm

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

  See NJ TRANSIT - NEC's other Tweets

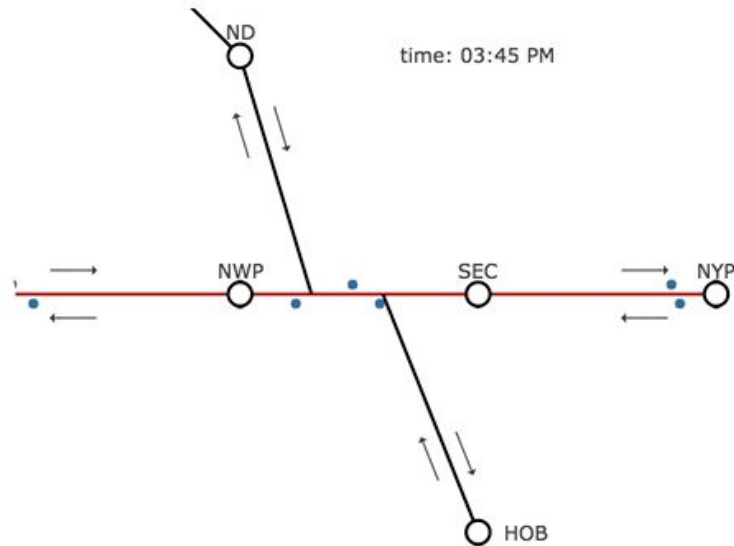




New York Penn Station Departures 4:39 PM Select a train to view station stops					
DEP	TO	TRK	LINE	TRAIN	STATUS
4:03	Harrisburg		KEYSTONE	A651	DELAYED
4:29	Trenton -SEC	8	Northeast Corrdr	3861	BOARDING
4:31	Jersey Ave		Northeast Corrdr	3723	DELAYED
4:33	Trenton -SEC		Northeast Corrdr	3947	DELAYED
4:36	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

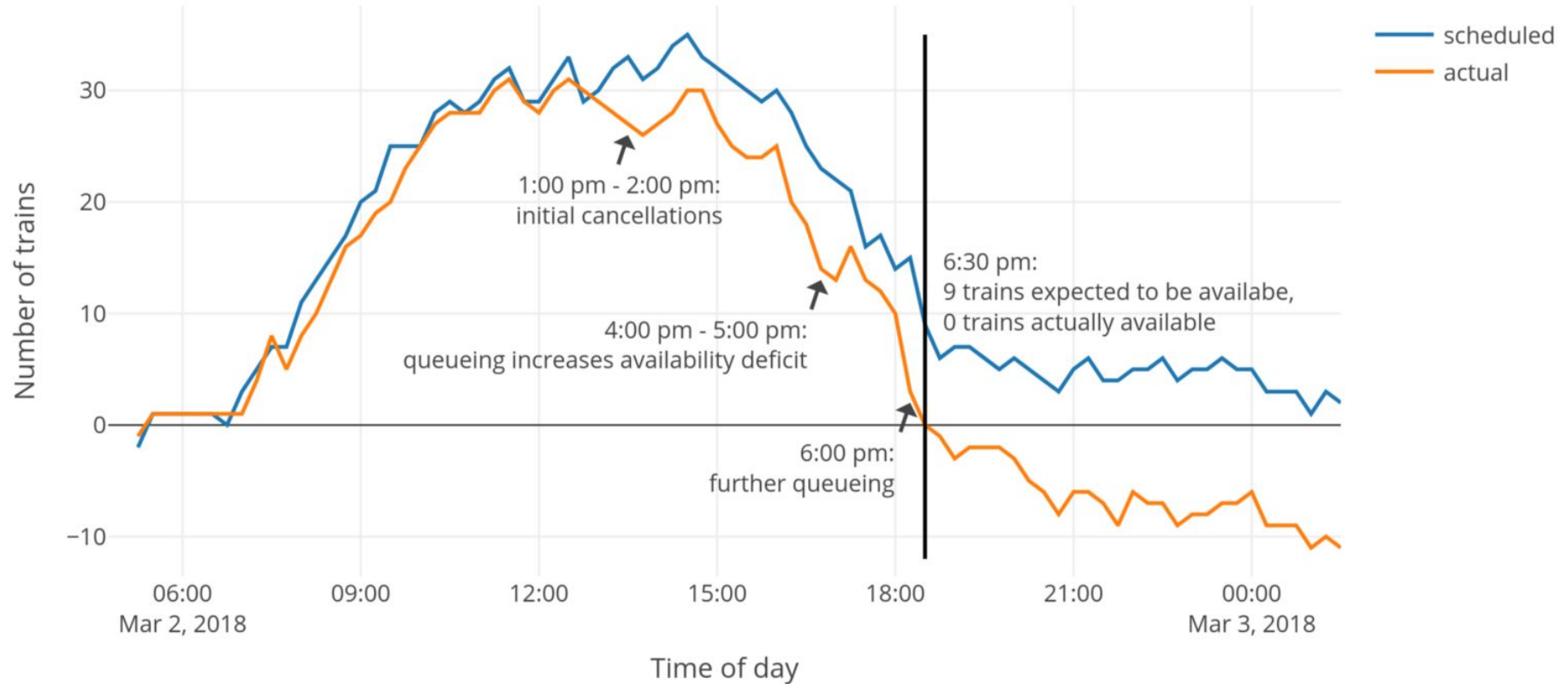
# Queueing begins...

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





## Number of Trains Available at New York Penn Station



**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/nitransit>

Medium:

<https://medium.com/@pranavbadami>  
<https://medium.com/@mzhang13>