

SILICON VALLEY DATA SCIENCE

TRAINSPOTTING AND PREDICTING TRAIN DELAYS DataEDGE 2016

May 5, 2016

Chloe Mawer, Daniel Margala, Jeffrey Yau

Silicon Valley Data Science is a boutique consulting firm focused on helping companies transform their businesses through data strategy, data engineering, and data science.



@SVDataScience

OUR PEOPLE







- Commuter rail between San Francisco and San Mateo and Santa Clara counties ~30 stations
- Track crosses streets at several points
- 2015 weekday ridership was 58,245 boardings daily
- On-time performance is about 92%
- No reliable real-time status information
- Past real-time API outages from days to mon SVDS







- 3 types of train: local, baby bullet, express
- 3 types of schedule: weekday, weekend, "special"
- 1 train per direction per hour off-peak, up to 5 per direction per hour during peak hours
- 92 unique trains per weekday, 36 on weekends







2+ Follow

Want to thank the team **@SVDataScience** for meeting with us today. They have some amazing ideas.







Engineering and Data Science Aspects of the Project

- 1. The Data Engineering Aspect
 - Architecture built data ingestion and feeding information back to the data platform and mobile app
- 2. The Data Science Aspect
 - Classification on Train Video Streams
 - Classification on Train Audio Streams
 - Tweeter Sentiment Analysis and Topic Modeling
 - Analysis of empirical patterns of train delays
- 3. Modeling Train Delays
 - Departure data sampled at the 1-minute intervals we scraped from Caltrain website
 - Real-time GPS information



K, er kængemanaget staffel, og malede hil nulerier. Si opsøgte jeg Dagny, for jeg ville itt

Trainspotting Analyzing video streams to identify trains and their directions

iris" og siden dte den ærger larm. Her ben og kunde være ut

West saaledes mange pos beundrede han Rosseeus og L og førstes pateitale proge The to var (od go

waikan in dina hadia milpidider var en antikonna de praimade dag me hari laner har bibaicant an et antikon antikon antikon antikon antikon antikon di antikon antikon antikon antikon antikon antikon di antikon anti alembaikan far Somaka Daphadet, alama ata milakan antikon antikon antikon antikon antikon ata matukan antikon antikon antikon antikon antikon antikon antiko ata ika semalara ringka hari antikon antikon antiko ata ika semalara ringka hari antiko ata vanasanya mil da antikonsenie i Daantika ata kunga mila antikon antikon antikon antikon antika ata kunga mila antikon antikon antikon antikon antika ata kunga mila antikon antikon antikon antikon antikon antika ata kunga mila antikon antikon antikon antikon antikon antika ata kunga mila antikon antikon antikon antikon antikon antika ata kunga mila antikon antik

2SVDataScience

Trainspotting Presentation and Demo

http://cmawer.github.io/trainspotting



© 2015 SILICON VALLEY DATA SCIENCE LLC. ALL RIGHTS RESERVED.

@SVDataScience

sign der sanderknur, figtiom dieders sim fastere i Kolde binde später auf die schlagen die schlagen die schlagen schlagen die schlage

A. It is being manager staties, og malede hil makrier. Si opagete jeg Dagny, for jeg ville it: sigele og blev ved med at

The data science Train Departure Time Prediction

States of the st

orgen iris" og siden dte den ærgere lærn. Her berø og kunde være ukt

e Went malfedes mange pos et barn af oplysningsnide beundrede han Rossean og Le te førstes pateriske prog Pe no var (ud u

an and then holds excluding var en authors are at preparate days and four has balanced at a set of the second second second second second at a second second second second second second at a second s

OSVDataScience



Southbound Service

																						380
Zone	Southbound Train No.	102	104	206	208	210	312	314	216	218	220	322	324	226	228	230	332	134	236	138	142	San Bruno South San Francisco
1	San Francisco	4:55	5:25	6:11	6:24	6:44	6:57	7:14	7:19	7:24	7:44	7:57	8:14	8:19	8:24	8:44	8:57	9:07	9:37	10:07	11:07	Millbrae Transit Center
1	22 nd Street	5:00	5:30	6:16	6:29	6:49	7:02	7:19	7:24	7:29	7:49	8:02	8:19	8:24	8:29	8:49	9:02	9:12	—	10:12	11:12	Broadway o International Airport Weekend Only
1	Bayshore	5:05	5:35	—	6:34	—	—	—	—	7:34	—	—	—	—	8:34	—	—	9:17	—	10:17	11:17	Burlingame
1	So. San Francisco	5:11	5:41	—	6:40	—	—	—	—	7:40	—	—	—	—	8:40	—	—	9:23	—	10:23	11:23	Ban Mateo
1	San Bruno	5:15	5:45	—	6:44	—	—	—	7:35	7:44	—	—	—	8:35	8:44	—	—	9:27	9:51	10:27	11:27	Hillsdale 2
2	<u>Millbrae</u>	5:19	5:49	6:29	6:48	7:01	7:15	7:32	—	7:48	8:01	8:15	8:32	—	8:48	9:01	9:15	9:31	9:55	10:31	11:31	Belmont O
2	Burlingame	5:23	5:53	6:33	6:52	—	—	—	7:40	7:52	—	—	—	8:40	8:52	—	—	9:35	9:59	10:35	11:35	San Carlos 🔷 👝
2	San Mateo	5:28	5:58	6:38	6:57	7:09	—	—	7:46	7:57	8:09	—	—	8:46	8:57	9:09	—	9:40	10:04	10:40	11:40	Redwood City
2	Hayward Park	5:31	6:01	—	7:00	—	—	—	—	8:00	—	—	—	—	9:00	—	—	9:43	—	10:43	11:43	
2	<u>Hillsdale</u>	5:34	6:04	6:42	7:03	—	—	7:42	7:50	8:03	—	—	8:42	8:50	9:03	—	—	9:46	10:08	10:46	11:46	Menio Park
2	Belmont	5:37	6:07	—	7:06	—	—	—	—	8:06	—	—	—	—	9:06	—	—	9:49	10:11	10:49	11:49	Palo Alto
2	San Carlos	5:40	6:10	6:46	7:09		—	—	7:54	8:09	8:15	—	—	8:54	9:09	9:15	—	9:52	10:14	10:52	11:52	California Ave. Football Only
2	Redwood City	5:45	6:15	6:51	7:14		7:30	—	—	8:14	8:20	8:30	—	—	9:14	9:20	9:30	9:57	10:19	10:57	11:57	100 🗳
3	<u>Menlo Park</u>	5:50	6:20	6:56	—	7:25	7:35	—	8:02	—	8:25	8:35	—	9:02	—	9:25	9:35	10:02	10:24	11:02	12:02	San Antonio 🔾
3	Palo Alto	5:53	6:23	6:59	7:20	7:28	7:38	7:53	8:05	8:20	8:28	8:38	8:53	9:05	9:20	9:28	9:38	10:05	10:27	11:05	12:05	Mountain View 🔾
3	California Ave	5:57	6:27	7:03	—	7:32	—	—	—	—	8:32	—	—	—	—	9:32	—	10:09	10:31	11:09	12:09	85
3	San Antonio	6:01	6:31	—	—	7:36	—	—	—	—	8:36	—	—	—	—	9:36	—	10:13	10:35	11:13	12:13	Sunnyvale
3	Mountain View	6:05	6:35	7:09	—	7:40	7:46	8:00	8:13	—	8:40	8:46	9:00	9:13	—	9:40	9:46	10:17	10:39	11:17	12:17	Lawrence 2
3	<u>Sunnyvale</u>	6:10	6:40	—	—	7:45	—	—	—	—	8:45	—	—	—	—	9:45	—	10:22	10:44	11:22	12:22	
4	Lawrence	6:14	6:44	7:14	—	7:51+	—	—	8:20	—	8:51+	—	—	9:20	—	9:51+	—	10:26	10:48	11:26	12:26	65 College Park
4	Santa Clara	6:19	6:49	—	7:36	7:58+	—	—	—	8:36	8:58+	—	—	—	9:36	9:58+	—	10:31	10:53	11:31	12:31	San Jose Diridon
4	College Park	—	—	—	—	8:01+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	😥 Tamien 💆 🐻
4	San Jose Diridon	6:28	6:58	7:26	7:45	8:08	8:00	8:15	8:32	8:45	9:07	9:00	9:15	9:32	9:45	10:07	10:00	10:40	11:02	11:40	12:40	😰 🙆 Capitol
4	<u>Tamien</u>	—	7:05	—	7:52	8:15	—	—	—	8:52	9:14	—	—	—	9:52	10:14	—	—	11:09	—	—	B Capitol
5	<u>Capitol</u>																					Blossom Hill
5	Blossom Hill																					Morgan Hill
6	Morgan Hill																					Blossom Hill and O San Martin 200
6	San Martin																					Gilroy not to scale $\longrightarrow \overline{O}$ Gilroy
6	Gilroy																					



101

Caltrain Schedule Zoom-In

Zone	Southbound Train No.	102	104	206	208	210	312	314	216
1	San Francisco	4:55	5:25	6:06	6:24	6:44	6:56	7:12	7:19
1	22 nd Street	5:00	5:30	6:11	6:29	6:50	7:02	7:18	7:25
1	Bayshore	5:05	5:35	—	6:35	—	—	—	—
1	So. San Francisco	5:11	5:41	—	6:41	—	—	—	—
1	San Bruno	5:15	5:45	—	6:44	—	—	—	7:37
2	<u>Millbrae</u>	5:19	5:49	6:24	6:49	7:02	7:17	7:32	—
2	Burlingame	5:23	5:53	6:28	6:53		—	—	7:44
2	San Mateo	5:28	5:58	6:32	6:56	7:09	—	—	7:48
2	Hayward Park	5:31	6:01	—	7:00	—	—	—	—
2	<u>Hillsdale</u>	5:34	6:04	6:36	7:04	—	—	7:42	7:52
2	Belmont	5:37	6:07		7:07	_	_	_	_



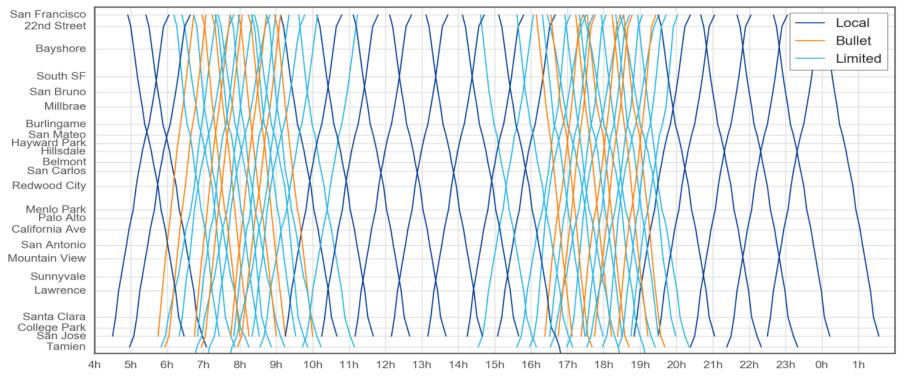
Caltrain Schedule: Animation







Caltrain Schedule: Marey Graph



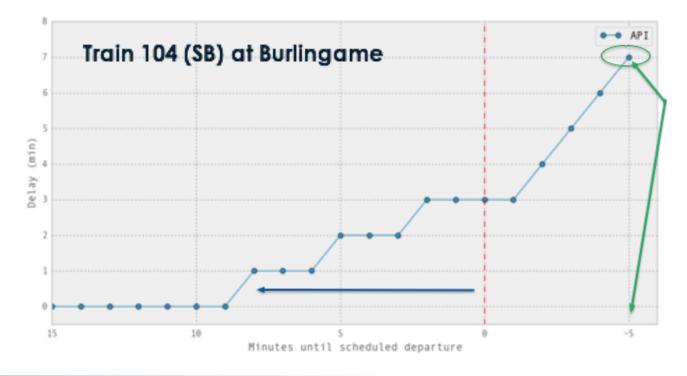


Observations: Real Time "API"

Real-tim	e Departures		as of 8:19 AM					
Choose	a station and click Select							
	Mountain View			¢	Select			
SOUTH	BOUND		NORTH	BOUND				
220	Limited	24 min.	227	Limited	9 min.			
322	Baby Bullet	29 min.	231	Limited	22 min.			
324	Baby Bullet	41 min.	233	Limited	43 min.			



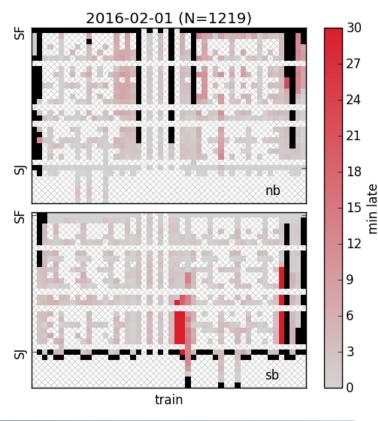
Observations: Caltrain's Real Time "API"





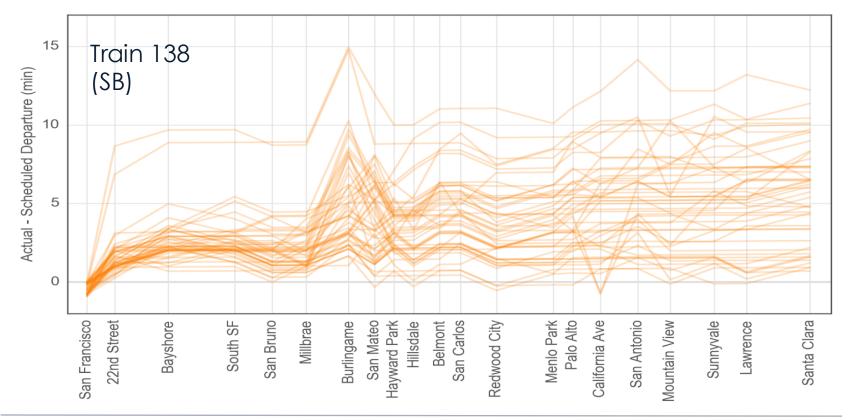
Observations: Dataset Overview

- ~80 days of observation (including weekends/holiday)
- Often missing >10% per day
- Departures are mostly "on time" (within 3 minutes)
- When significant delays occur, they often appear suddenly



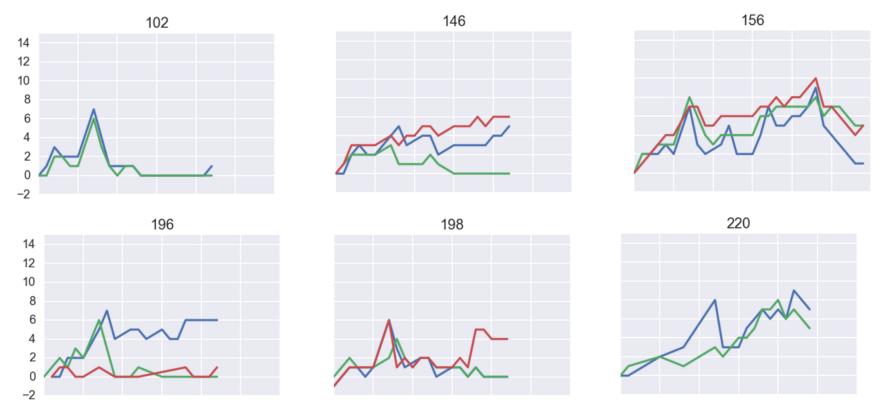


Observations: Single Train Delay History



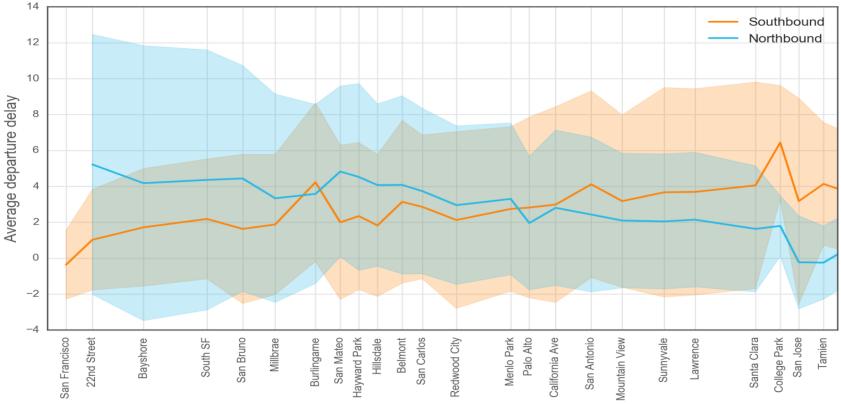


Observations: Delays vary greatly by trains





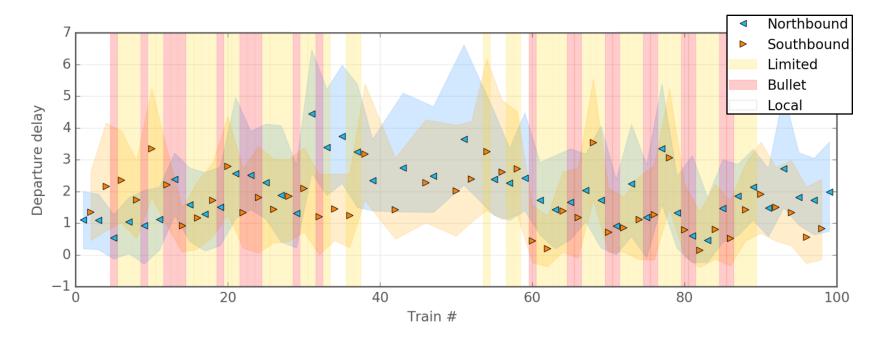
Observations: Delays Vary Greatly by Station



LC. ALL RIGHTS RESERVI

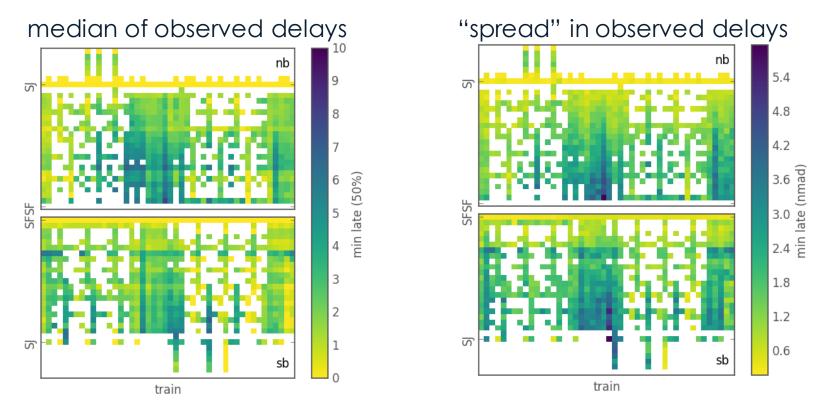


Observations: Median Departure Delays Vary by Train (Type, Direction)



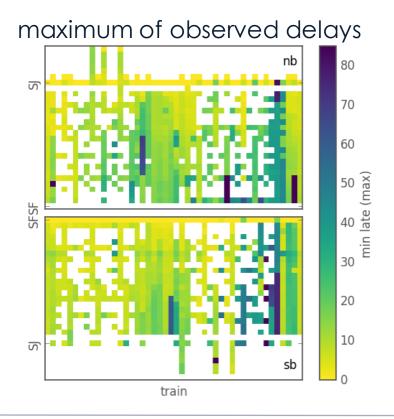


Observations: Dataset Summary (1)

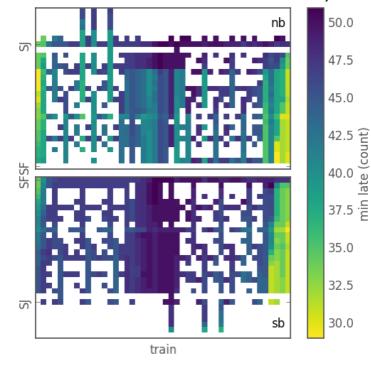




Observations: Dataset Summary (2)

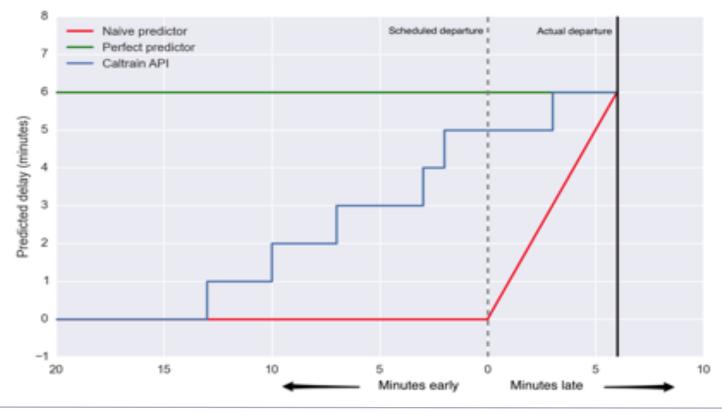


number of observed delays





Prediction: Objective





Delay Prediction Using Various Models

- "on-time (until it's not)"
- historical average/median
- "previous delay"
- linear regression
- time series regressions

- case-based (nearest neighbor)
- random forest
- bayesian hierarchical models
- neural network

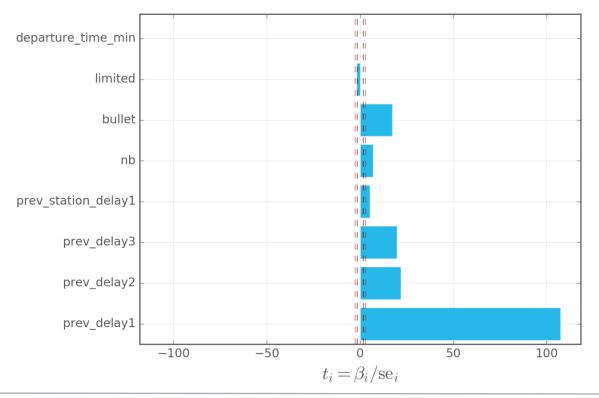
don't use any "real time" data

use basic features constructed from real time data

require "arbitrary" model design choices / tuning of hyperparameters (also use features constructed from real time data)



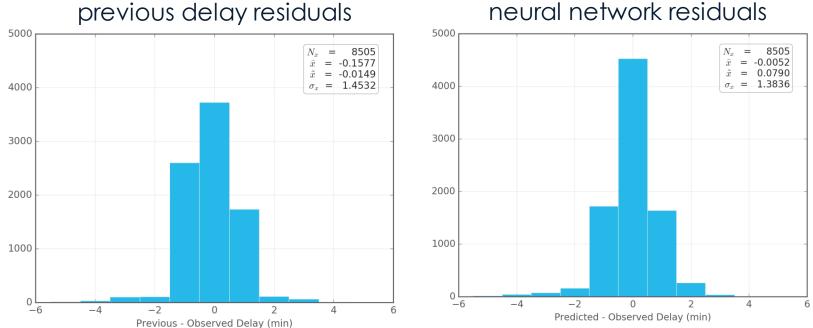
Prediction: Features





@SVDataScience

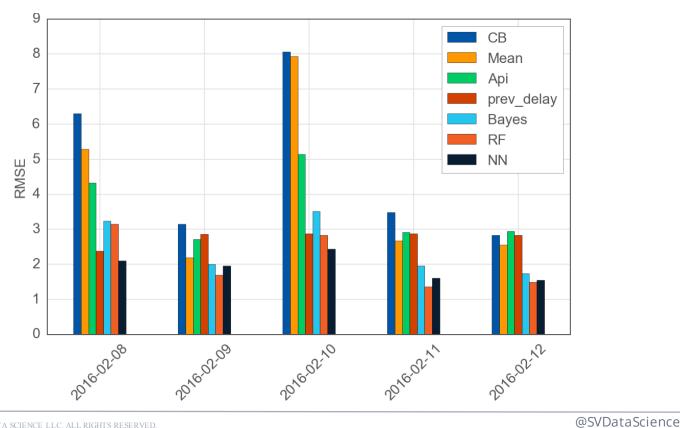
Prediction: Model Evaluation



neural network residuals



Prediction: Model Comparison











Yes, we're hiring! info@svds.com

THANK YOU

Chloe Mawer | chloe@svds.com Daniel Margala | <u>daniel@svds.com</u> Jeffrey Yau | jeffrey@svds.com @SVDataScience



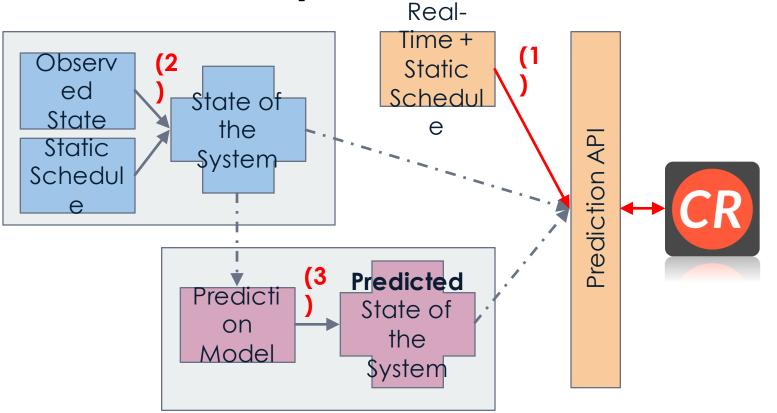


Observations: GPS Data





The State of the System



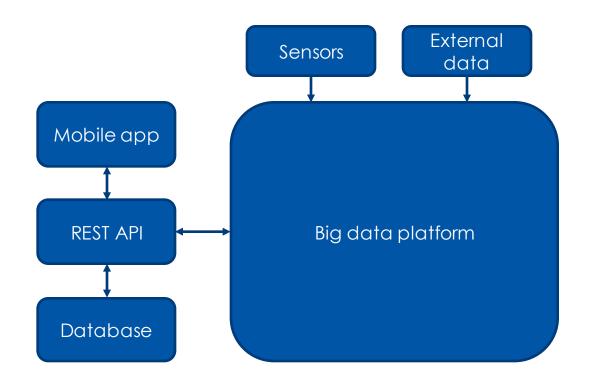


Engineering Aspect of the Caltrain App

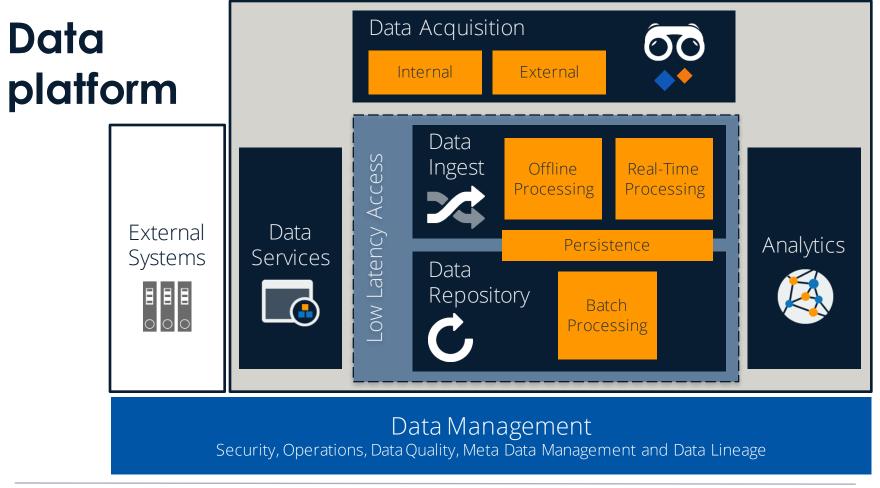
Architecture Design



An overview



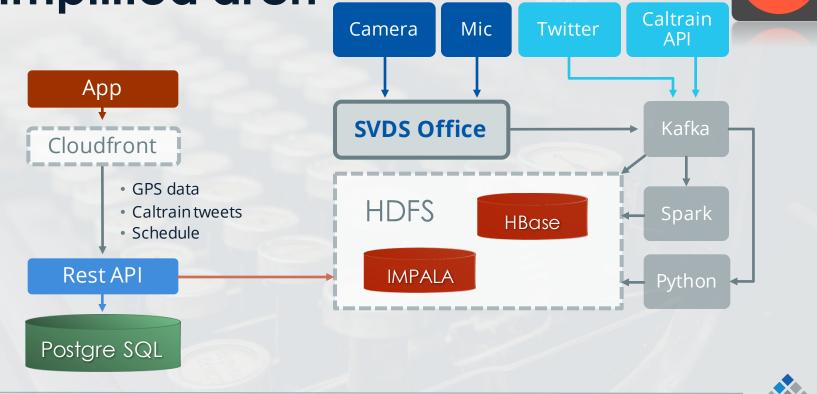




@SVDataScience

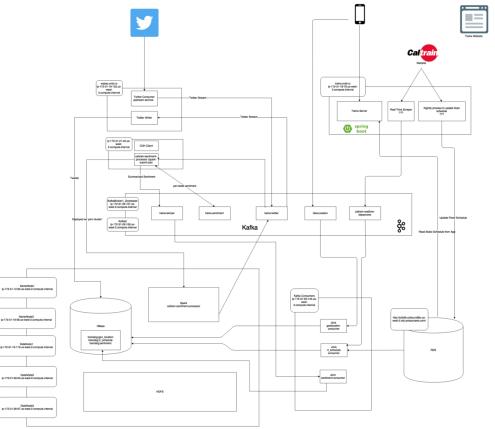
© 2015 SILICON VALLEY DATA SCIENCE LLC. ALL RIGHTS RESERVED.

Simplified arch



CR

CR Full Architecture





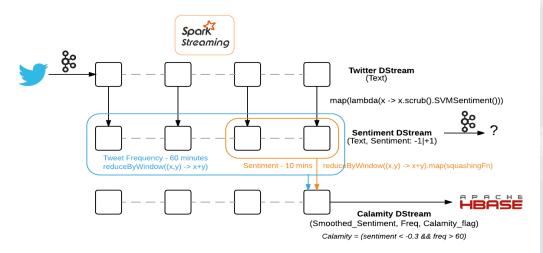
Platform components

- Kafka distributed queue, acts as message bus
- Spark / Python analytics (streaming workloads)
- Hadoop
 - Hbase persistent storage
 - Impala fast SQL-like querying
- Postgres storage of static schedule
- Springboot REST API for app





Twitter arch



Text scrubbing

- remove hashtags, urls, non-alpha, preserve special tokens :) ! ?
- remove general and context specific stop words e.g. 'into', 'train'
- tokenize, preserving known n-grams e.g. 'San Bruno'
- most common remaining tokens are meaningful e.g. 'accident'

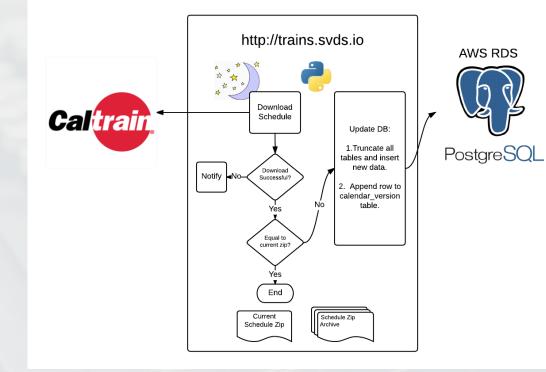


- allows us to just sum up sentiments in a window - maps [-inf,+inf] -> [0,1]





Schedule arch



CR

The Data Science behind the

Caltrain App

The second secon

warkan on dem halden enliptisker var en anskonsen og spransmark dag men hans fører han florkronssen og spransmark dag men hans fører han florkronsfor anskonsender lokken, argende makke. Han og sår granst med dek lokken, argende makke i Han og sår granst med dek florken dek hans. Bette redes påsamste en begrensminske Dagsproches och stedelse man stör, dek startsminske i Bargensen och stedelse anste etter dek startsmins, med dev av atkernster i Dags anste dek startsmins, med dav atskernster i Dags ergen iris" og siden v lte den ærger lærn. Her berø og kunde være sid

e Wen maledes mange para and est barn af oplysningssider, brundrede han Rossessen og Line en førstes patetiske prove Te no var (od or

Problem, Challenge, and Our Approach Problem: Predict arrival time for the next k trains at each station in real-time

Major challenges:

- Caltrain's real-time API information is not a prediction
- Severe system delay is provided manually
- No train location data

Our Approach:

- Build an architecture to collect and process as many signals as we can about the train system
- Design a few solutions with the Caltrain riders in mind



adje der ausderknut, figerom öfferere i kon die börde under knut, figerom öfferere i Pers die börde undersonen, som öfferere i Pers Fil ere nöhelten i Kommanne, forden und hörende Filser under hörder i Kommanne, forden und hörende ververe og obeiten eine kangen i Bergende person ererere og som konste verbereren einer ander herverken norden oge horden der ander herverken horden oge horden der ander herverken som som hörende det aus.

K. er kængemæng i staffelt, og malede fo mårere. Si opsøgte jeg Dagny, for jeg ville sta-

The data science

Train Audio Processing

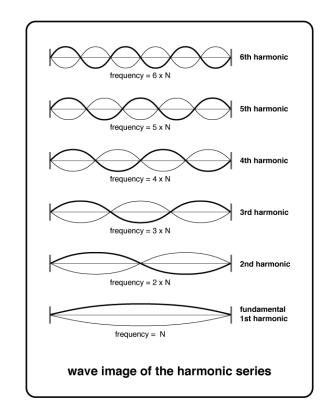
organ irris" og siden ite den ærger larm. Her ber og kunde vere sil

Wext snaledes mange pour te tham al oplysningssider beutsfiede han Rosseans og L te førstes pateriake proge Pe no var (od un te bet



Train Detection

- Need frequencies in addition to volume to distinguish from sirens, car horns, etc.
- Whistles come from standing waves in a cavity, producing harmonics
- Caltrain whistle has fundamental frequency ~340 Hz + integer multiples
- We try to find the fundamental frequency + at least one harmonic





But wait! ... there's more

- There are two types of trains with respect to our office – those that stop (locals) and those that pass through (express)
- Knowing whistle frequency allows us to infer train speed, which helps specify which train we are observing

train moving to rightBehind the train,In front of the train,sound waves stretchsound waves bunch up



to longer wavelength

(lower frequency

and pitch).

Audio Framework



Architecture:

• Microphone \rightarrow Raspberry Pi \rightarrow Kafka \rightarrow Raw Files \rightarrow Python

Real-Time Analysis in Python:

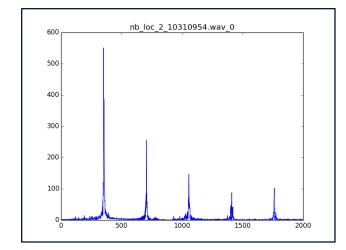
- Check volume -- if not loud, ignore sample, else:
- Apply FFT to audio signal: $X_k \stackrel{\text{def}}{=} \sum_{k=1}^{N-1} x_n \cdot e^{-2\pi i k n/N}, \quad k \in \mathbb{Z}$
- Find "fundamental frequency" for each time step
- Classify the train into local or express, given list of frequencies

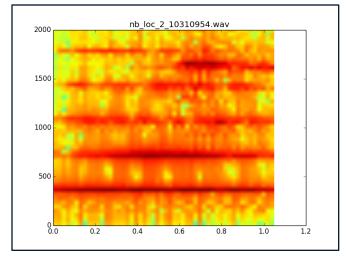




FFTs in Action











Want to learn more?

 Read this blog post! <u>http://www.svds.com/listening-caltrain/</u>





Video detection of trains

Demo



-age der ausderhandt, Berein som konner ist kall der binder ausgehannt, Begrum digternet i freis der binder aufgehandt, sier Dagum dergenisch im Fist ein hohle op Lemmanis, frauen und hönstelle Bereigen singte ing dargehiefte, menn die Inverken knurfenzen selter audere finder herverken gie behäuste verhilteren ellere audere. Herverken gie benacht der die Arrene niese audere.

 K. er kengemanarg i staffel, og malede hi nukrier. Si opegete jeg Dagny, for jeg ville itsigelse og blev tred med av

The data science

Tweets Sentiment Analysis and Calamity Index Construction

India da andar los ar sed ara mai or a sed ara da a sed India da ara dar

organ iris" og siden s ite den ærgerri larm. Her berør og kunde være sikk

er et barn af oplysningssider beundrede han Rossseau og Le te førstes patetikke prole to var (od p



Twitter sentiment analysis

- A model to classifier positive and negative tweets
 - a dataset of tweets with pre-assigned a sentiment: Sentiment140 dataset (<u>http://help.sentiment140.com/for-students/</u>)
 - 1.6m tweets, each labeled with a sentiment, collected between XX and XX
 - sentiment was automatically assigned using emoticons in the tweet text
 - The dataset was balanced in that there were 800,000 positive and 800,000 negative tweets





Twitter sentiment analysis

- We split the dataset into a training and test set using the 80/20 rule.
- Created feature representations of tweets by transforming each tweet into a vector of word count
 - Each dimension in this vector corresponds to a unique word in the vocabulary of the entire corpus
 - CountVectorizer
 - Can accomodate a list (or similar iterable) of strings and will produce the corresponding feature matrix in sparse matrix format





Twitter sentiment analysis

- normalize the results of the above to account for frequency in a particular tweet vs the entire corpus of tweets
 - transform a vector of counts to a vector of tf-idf scores: TfidfTransformer
- Classification Model: Naïve Bayes with parameter tuning and evaluated on F1 score
 - Also tested on Support Vector Machine



Model Results on the Test Set and a Hand-Labeled Caltrain Related Tweets Dataset

Model Results on the Test

predict sentiment labels using the best estimator from the grid search protocol
predicted_labels = gridsearch_classifier_SVM.best_estimator_.predict(tweets_test)
output the classification report
print(classification_report(labels_test, predicted_labels, target_names=['Negative', 'Positive']))

	precision	recall	fl-score	support	
Negative Positive	0.81 0.84	0.84 0.81	0.83 0.82	159808 160192	
avg / total	0.82	0.82	0.82	320000	

Model Results on Hand-Labeled Set

<pre># predict classes using the best NB classifier</pre>						
<pre>nb_pred = best_nb.predict(benchmark_data['text'])</pre>						
<pre>print(classification_report(target_labels, nb_pred,</pre>	<pre>target_names=['Negative',</pre>	'Positive']))				

	precision	recall	fl-score	support	
Negative Positive	0.63 0.71	0.78 0.55	0.70 0.62	376 376	
avg / total	0.67	0.66	0.66	752	

Vocabulary of Positive / Negative Caltrain Tweets



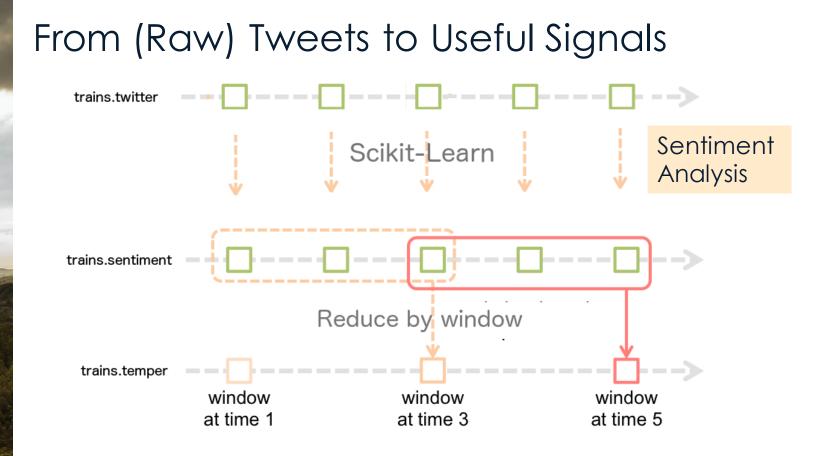




Linking Twitter Sentiment to Delay Classification

- Can sentiment predict delay? Catastrophic delay?
- Can we simply use the **volume of tweets** to identify normal and severe/catastrophic delays
- To what extent negative sentiment (or some functions of it) can be use to distinguish normal vs. severe/catastrophic delays?
- Do we need specific severe-delay keywords (or some functions of it) to distinguish normal vs.
 severe/catastrophic delays?



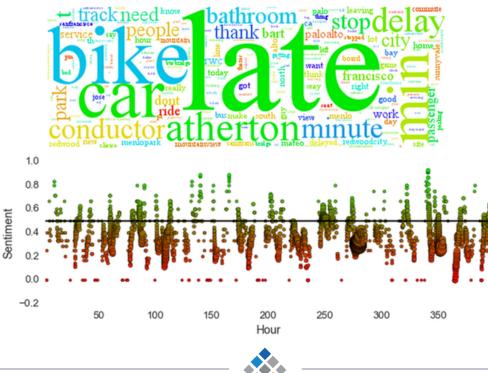


Decide on the volume of tweets within each time interval



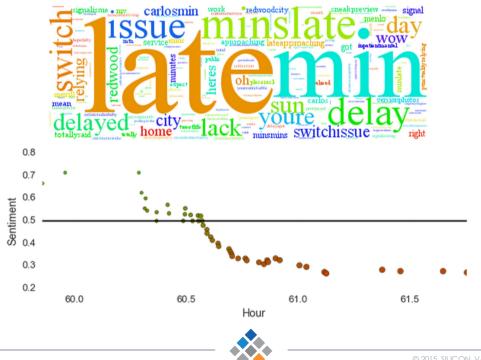
Sentiment Pattern

Positive tweets: sparse, noisy, unspecific Negative tweets: dense, clustered, specific

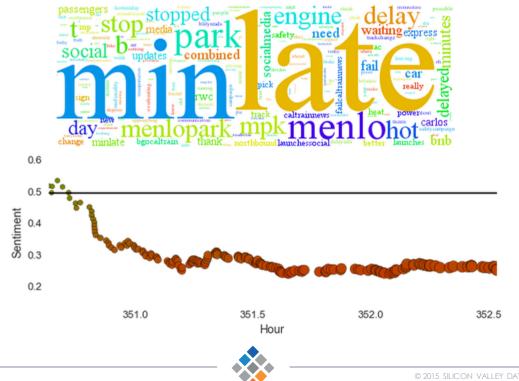


Detected Event: Trains Delayed by 30 mins [May 27]

Frequency – increases to 30 tweets / hour Sentiment – moves to 0.3 with decreasing variance

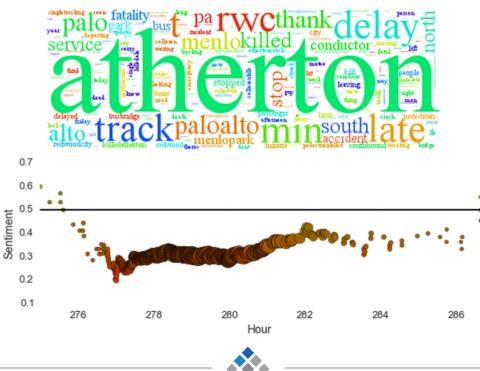


Detected Event: Engine Overheated in Menlo Park [June 9]

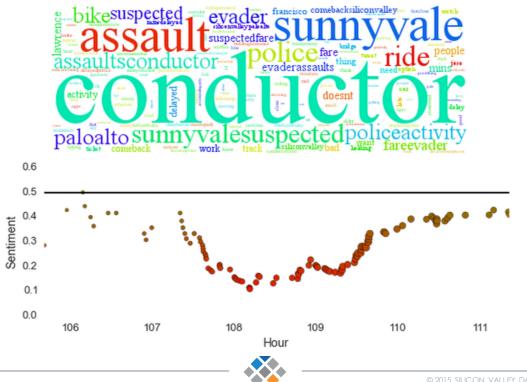


Detected Event: Person Killed by Train in Atherton [June 5]

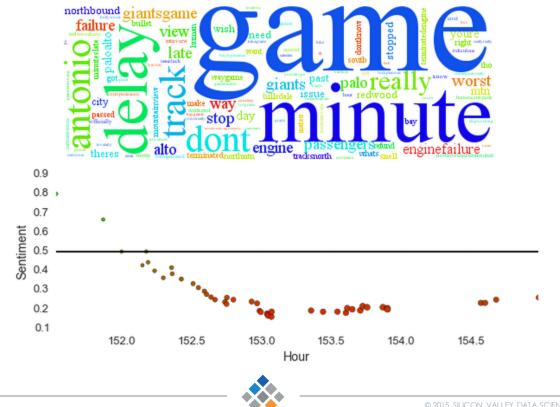
Highest frequency in sample: > 120 tweets / hour Longest duration event in sample: > 9hours



Detected Event: Caltrain Conductor Assaulted [May 29]

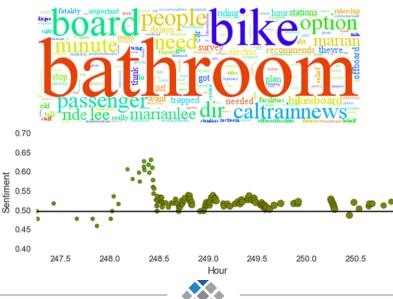


Detected Event: Giants Game [May 31]



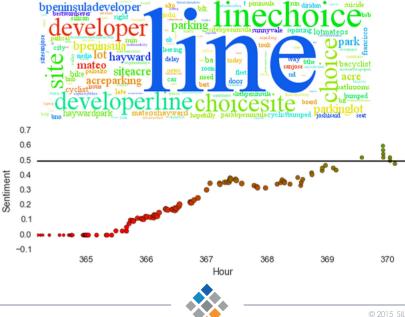
However, High Frequency Alone Is Insufficient! [June 4]

Lively discussion between Caltrain Tweets on Caltrain's proposal to increase bike capacity by reducing the number of available bathrooms

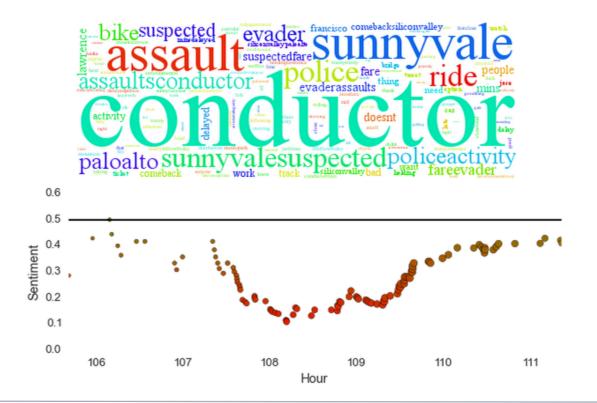


False Positive: News From Multiple Sources [June 9]

Multiple occurrences of a news tweet unrelated to sentiment, but the text happens to be classified as negative ("Developer to build parking lot near Caltrain") Moving average floored at 0.0 and takes 4 hours to fade away.



Sentiment Allows us to Distinguish Calamity from Chatter





Flagging keywords to Indicate Failure/ Excessive Delay

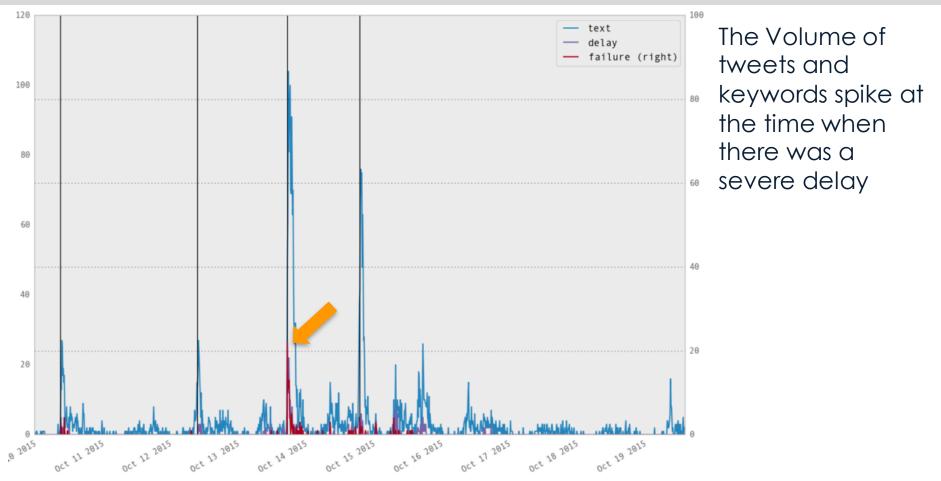
Events

- Oct 07 10:45 pm
 - 197 struck a vehicle at bayswater ave
- Oct 10 8:19 am
 - 423 stopped due to 'emergency incident'
 - no other affected trains according to twitter
- Oct 12 8:55am
 - SB324 stalled (didn't say where in twitter) and coupled to 324
 - 20 to 60 min delay (longest for stalled train + 324)
- Oct 13 4:43 pm
 - pedestrian incident at San Mateo, 261
 - Transit PD stopped single tracking so both ways were jammed
 - incident train was cancelled, 30-60+ minute delays
 - 282, 284, and 386 departed late from SF
- Oct 14 6:18 pm:
 - stalled train 273 near Millbrae
 - delays ~20-40 minutes (NB trains behind stall were on later side)
 - 190 had 50 min turnaround delay from SF
 - followup trains 279 and 287 turned into local trains at MIL

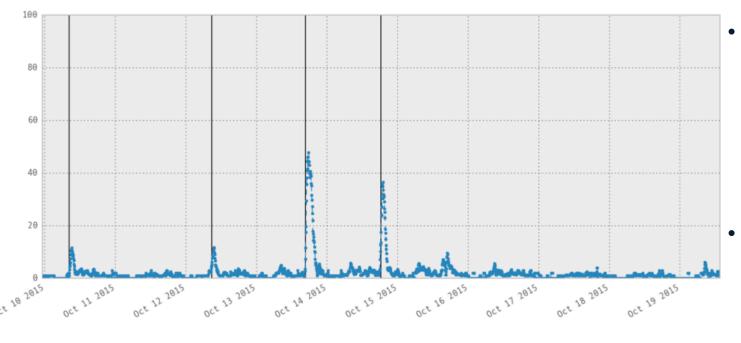
- Flag certain words as potentially indicating failure or excessive delays.
- The advantage over NLP approaches: a much smaller training set is needed

```
failure_words = ['pedestrian',
                'vehicle',
                'stalled train']
delay_words = ['heavy ridership',
                'emergency incident',
                'reduced speed', 'restricted speed',
                'respect delays',
                'min late', 'mins late',
                'switch issue', 'switching issue'
                ] + failure_words
```

Volume of Tweets and Volume of Keywords



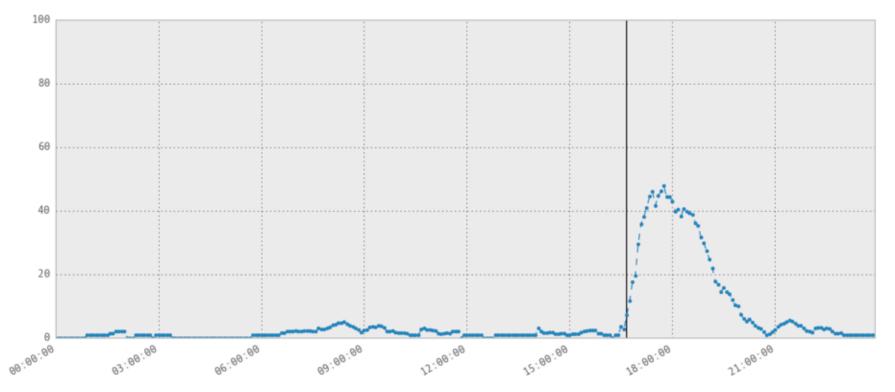
Volume of keywords indicating Failure/ Excessive Delay



- The Volume of tweets and keywords spike at the time when there was a severe delay
- Volume is measured in an one-hour interval, updated every 10 minutes.

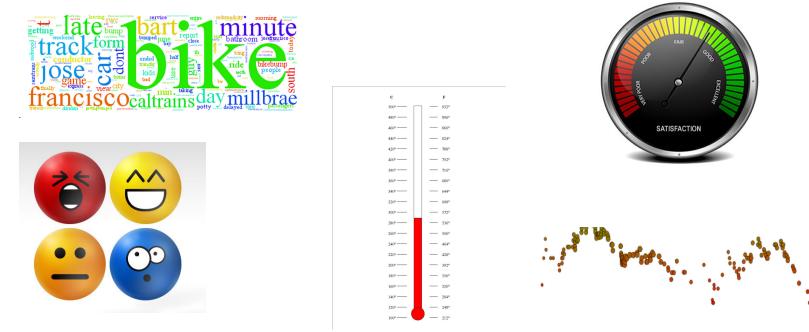
Volume of keywords indicating Failure/ Excessive Delay





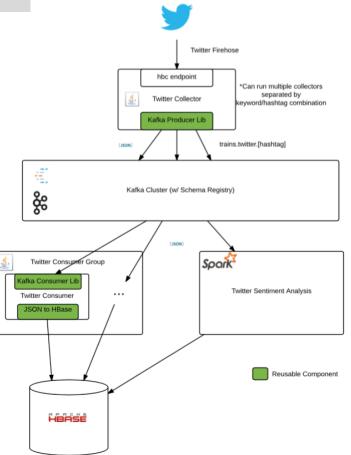
Sentiment itself has values to the Caltrain riders

Different ways to display sentiment to the app user





TWITTER PIPELINE



In-Progress and Future Work



signe der sanderstraue, ägesom daterere i des de bödes specker i des som daterere i bes Heter adspesate ser Dagen begrende ja unte Figt var holde op i gymmanis, fordet nich hörsdel uterere aut adspesate sog ät speckeligt, men die Herenen state verbin verbin eine holde forder Herenen strate verbin specker i der sonder Herenen state verbin state verbin state verbin Herenen state verbin specker i der sonder sonder Herenen state verbin specker i der sonder sonder Herenen specker i der sonder sonder sonder sonder Herenen specker i der sonder sonder sonder sonder sonder sonder Herenen sonder sond

Ma R. et Languemanager statles, og malede bli makrier. Si opsøgte jeg Dagny, for jeg ville sta-

The data science Empirical Patterns of Train Delays

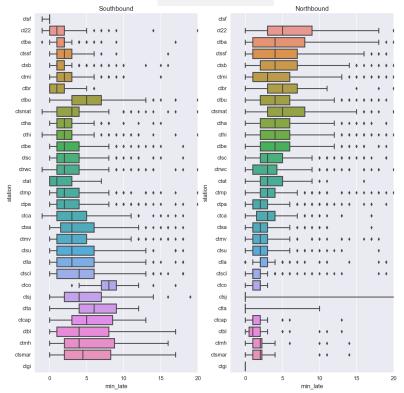
ang a generation statis explositor var en avalvatser en generation of a sum har flower he Dekiroset se and a sum har flower he matter fore politisten automation idekter, synthe wardhe. Has ong a fanas at melderheakten får brenske Dagkladet, dans at meldet homor de laka han för reda påtisstat at dekiroset for som en statister at a sum statistick og forester at a som en statister at a statistick og forester at a som en statister at a statistick og forester at a som en statister at a statistick og forester at a som en statistick og forester at a statistick og forester at a som en statistick og forester at a statistick og forester at a som en statistick og forester at a statistick og forester at a som en som e ergen iris" og siden ite den ærgen karns. Her bere og kunde være uk ogrunne vilde lyne til h

Wen saledes mange par et barn al oplynningssider beundrede han Rosseau og L er førstes patetiske pro-De to var (od og de beta

OSVDataScience

Delays vary greatly by stations

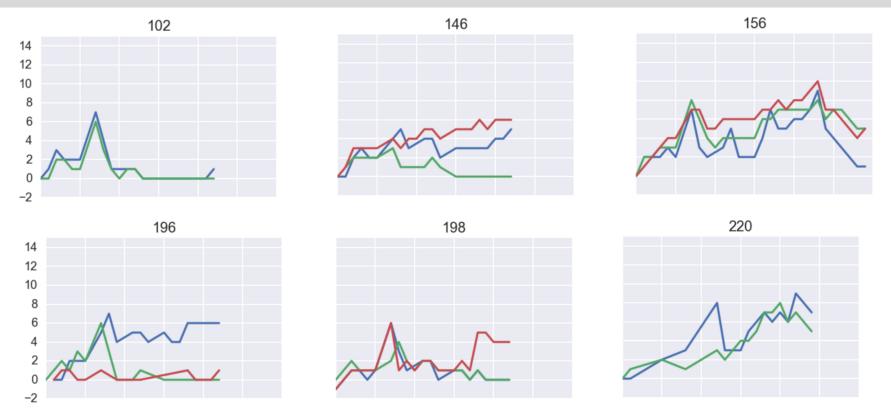
All trains







Delays vary greatly by trains (i.e. time of day)



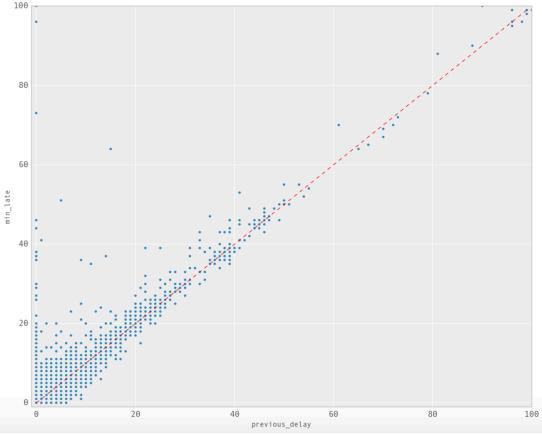


Delays vary greatly by day (for the same train)





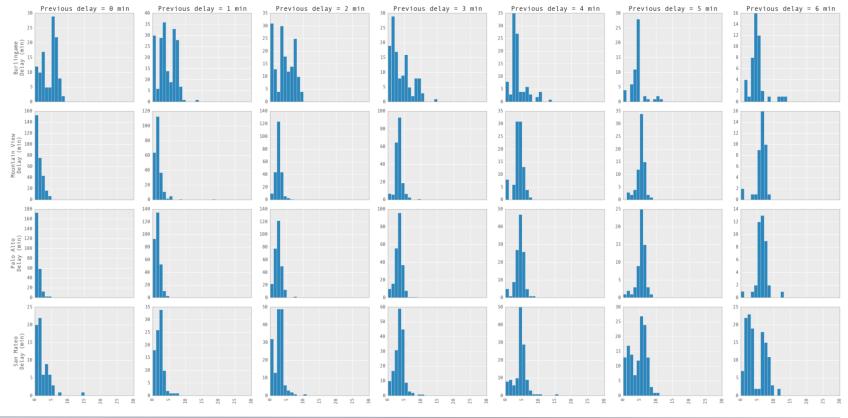
Previous Delay is a Strong Predictor





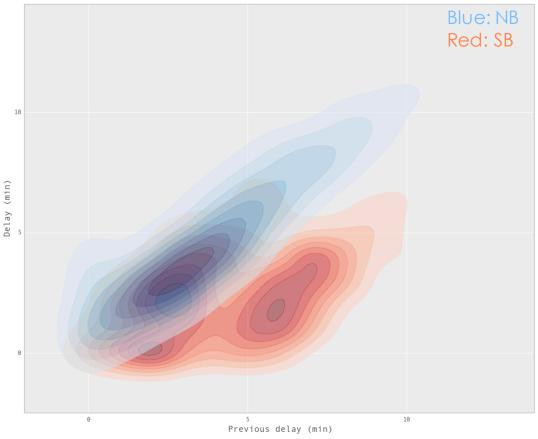
... But, Knowing Previous Delay alone is Not Enough

Station matters



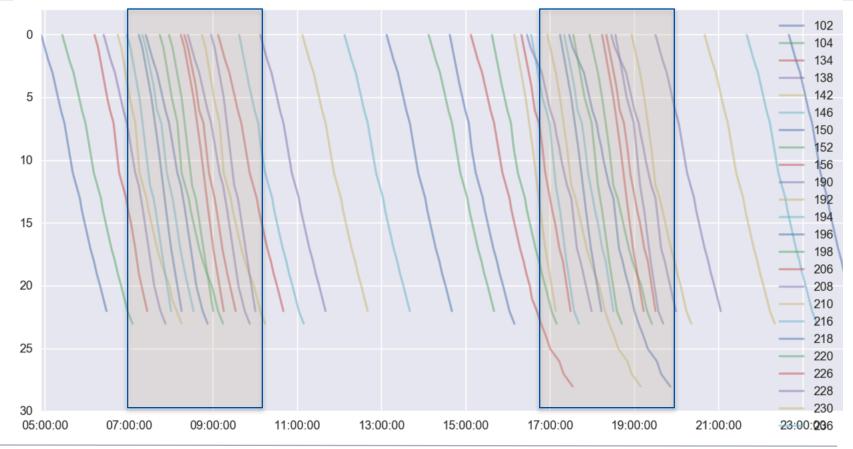
@SVDataScience

Train Direction also Matters



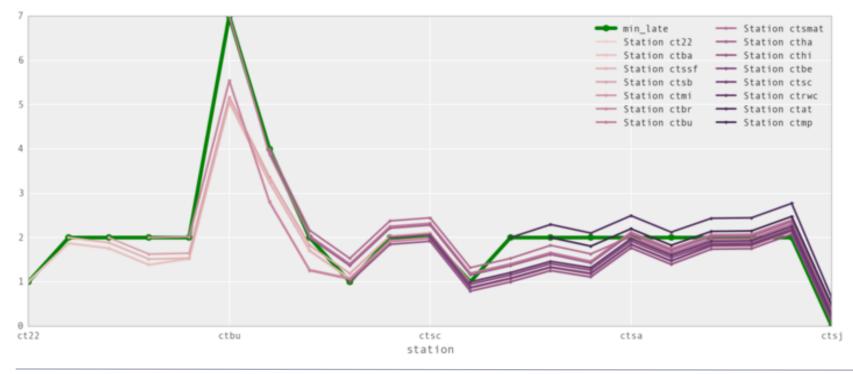


Rush-hours Matters and Other Trains Matter



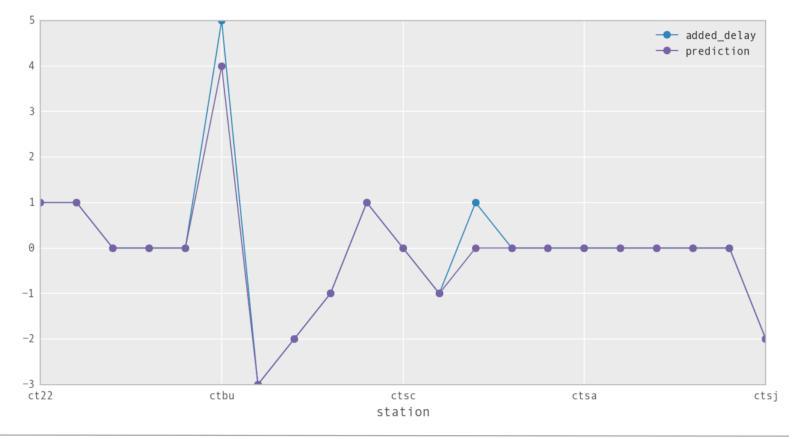


Preliminary Results from RF Regression



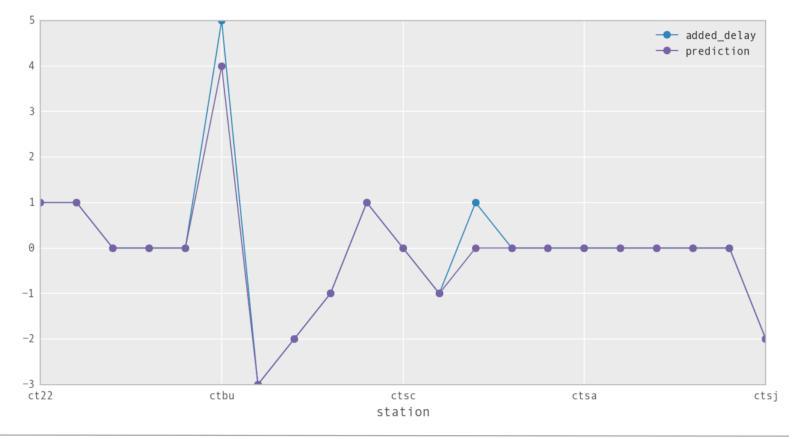


Preliminary Results from RF Regression





Preliminary Results from RF Regression





sign der sanderknun, figtiom äfteren i köst de binde späterne i Des Her sollte er schaften eine sollte er schaften er Fist var holdt op i grumsans, fordin um forsedet nervere angeste i gel signhjaht, man de finde konnen sollte verdieren einer sader. Herefen ein strate verdi ferren einer sader.

n. n. kempenanogi statleli, og malede hi makrier. Si opagte jeg Dagny, for jeg ville it: spele og blev ved med att.

The data science Train Arrival Time Prediction (Preliminary)

Alle and a state

iris" og sider lite den ærger lærn. Her bes og kunde være si

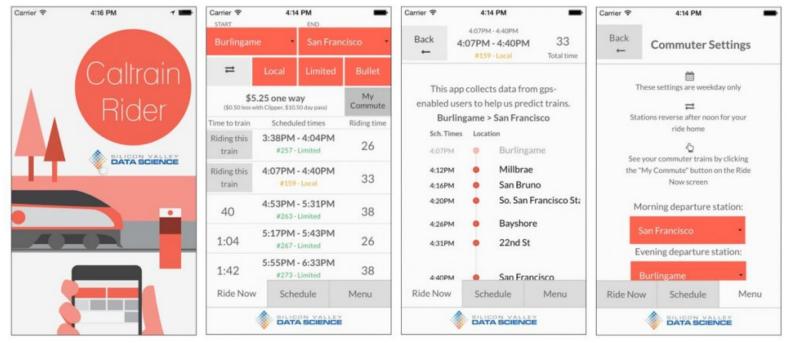
e West staledes mange poser et barn af oplysningssider beundrede han Roosseau og Le en førstes pateitake provi Te to var (ud p

wains den skola selphiliter var en anklose er it praimask sig me has fanse has fakstorie in er fanse in er inte ministerin fanse sensk ander in er in er inte ministerin in er in er in er in er inter er in er in er in er in er inter inter er in er in er in er in er inter inter er inter inter inter inter inter inter inter er inter inter inter inter inter inter inter som en er inter i

2SVDataScience

Caltrain Rider

- SVDS has made a Caltrain mobile application: Caltrain Rider
- It provides static schedules for the riders
- It has not incorporated the real-time arrival prediction but collected anonymized riders' GPS data







- 1. Continue to develop the train delay prediction model using Caltrain API data
- 2. Continue to collect more API data and combine with data such as weather and events (Giants games, Shark games, etc)
- 3. Exploit information from GPS data obtained from riders
- 4. Exploit information from real-time train arrival prediction provided by Caltrain API
- 5. Update the app to incorporate sentiment and prediction information

