

Alpitour



Outline

1. Introduction
2. Data
3. Models
4. Deliverables

Team



Stephen Slater

A.B./S.M. CS/CSE



Yuting (Koko) Kou

S.M. Data Science



Tommaso Scarlatti

M.S. Computer Eng.



Eleonora Cappuccio

M.S. Comm. Design

Collaboration Infrastructure

- Communication: Slack
- Version control: GitHub
- Conference calls: FB Messenger



Client

- Alpitour is Italy's leading integrated tourism company
- Founded in 1947 and now achieves annual revenue larger than €1B
- Businesses:
 - Tour operating
 - Aviation
 - Hotel management
 - Travel agencies



Problem Statement

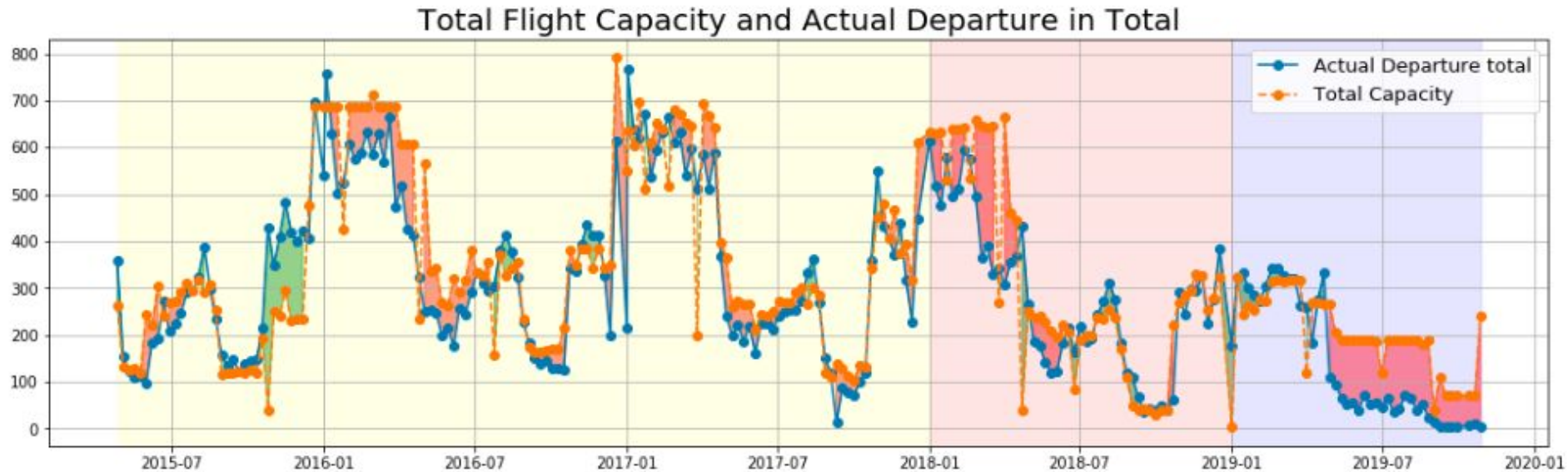
- **Predict** the demand (orders/departures) for travel to Cuba in the next year
- Challenging time series problem, with noise due to human behavior
- Current prediction system (linear)

Preclosing - NOVEMBRE 2018



			Inverno 2017/2018				Inverno 2018/2019				Δ vs. PY		Budget						
			Consuntivo		Portafoglio al 25/3		Portafoglio al 24/3/2019			18/19 vs. 17/18		2018/2019							
Week	dal	al	Gar	Pax	NEOS	% Occ.	Gar	Pax	NEOS	% Occ.	Gar	Pax	% Occ.	NEOS	Δ Gar	Δ Pax	Gar	Pax	NEOS
1	1/11	1/11	0	0	0	-	0	0	0	-	0	0	-	0	0	0			
2	2/11	8/11	14	13	0	93%	14	13	0	93%	57	50	88%	0	43	37			
3	9/11	15/11	35	29	0	83%	35	29	0	83%	49	49	100%	0	14	20			
4	16/11	22/11	20	20	0	100%	20	20	0	100%	37	33	89%	0	17	13			
5	23/11	29/11	20	18	0	90%	20	18	0	90%	25	25	100%	0	5	7			
6	30/11	30/11	0	0	0	-	0	0	0	-	0	0	-	0	0	0			
Casablanca			89	80	0	90%	89	80	0	90%	168	157	93%	0	79	77	93	86	0

Capacity data

- Capacity of planned flights and real departures for those flights



Business Value

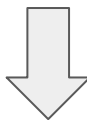
- Alpitour books flights up to 1 year in advance, so they need predictions of departures up to 1 year in advance
-  If Alpitour doesn't fill its capacity, it may **lose money**
-  If Alpitour fills its capacity, maybe it could have **sold more trips**

This is where our project comes in!

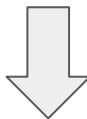
Let's go back in time and look at our trajectory from the beginning...

Trajectory: Data and Features

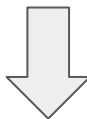
- Internal data from Alpitour



- Features such as moving averages, news sentiment, Google Trends



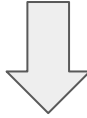
- Received more data including Alpitour web sessions, Google Ads



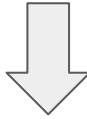
- Scraped Google Search “Cuba” query history for proportion of travel news

Trajectory: Models

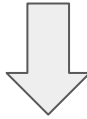
- Classification of orders for the next 1-3 months (up/down)



- Regression of average departures in the next 1-3 months



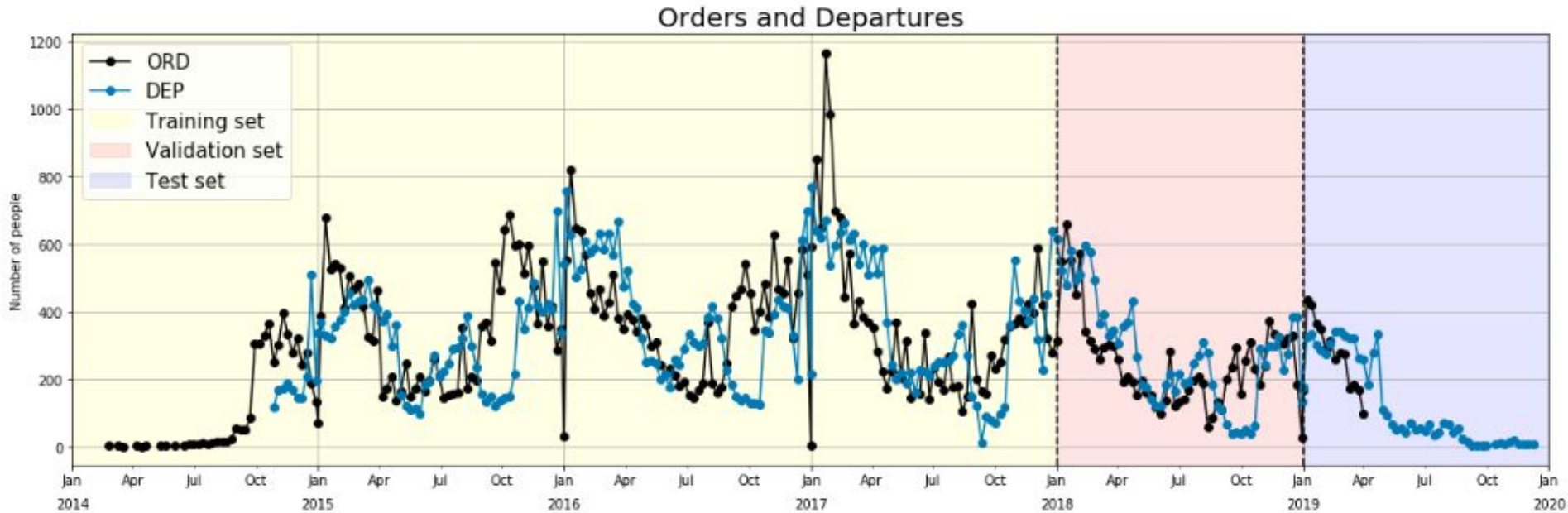
- Regression of departures for each of the next 12 weeks



- Regression of departures for each week in the next 1 year

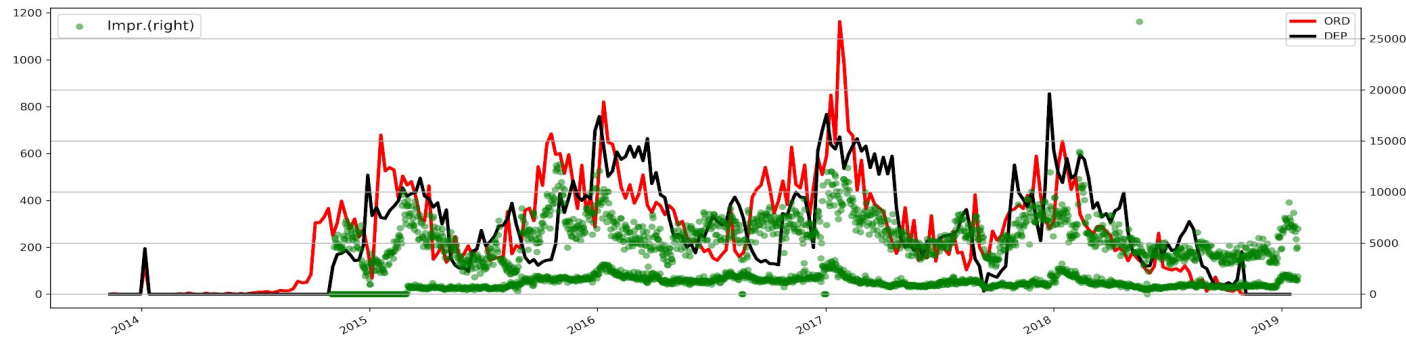
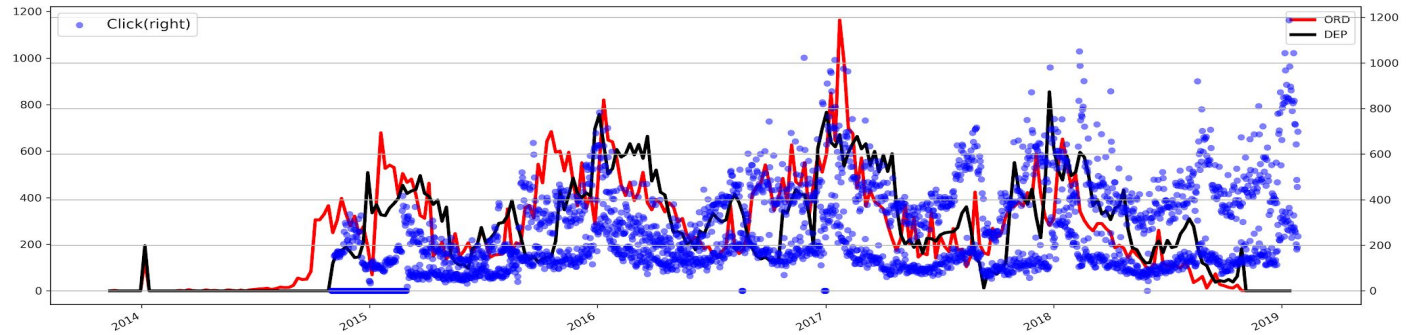
Internal Data

1. Booking data: 11/2013 to 4/2019 (33,900 rows): location, passengers, price...



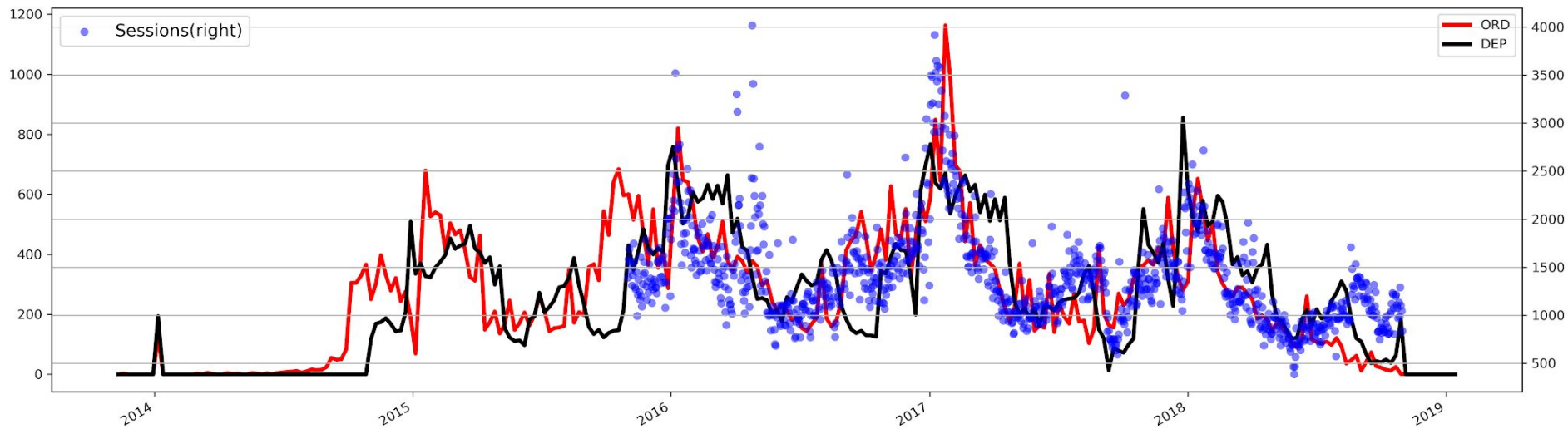
Internal Data

2. Google Ads for Cuba and Hotel Caraibi (49,100 rows): clicks and impressions



Internal Data


3. Search session history on Alpitour's website (5,000 rows)



External Data: Media

- Gathered articles from Italian news outlets
 - Consumer-related: Repubblica, ANSA
 - Trade-related: Travelquotidiano, Agenzia di viaggi, Giornale del turismo
 - News aggregator: European Media Monitor (EMM)

Viaggio a Cuba: 12 luoghi da non lasciarsi scappare Easyviaggio [🔗](#)

 msn--it Wednesday, March 6, 2019 4:39:00 PM CET | [info](#) [en](#) [en] [other]


La capitale di Cuba ha una bellezza e un'anima antica che vi faranno innamorare. Il nucleo coloniale spagnolo del centro è semplicemente affascinante, con le sue chiese affrescate e i palazzi civici restaurati che oggi sono stati riconvertiti in musei e ristoranti....

Lega Assisi si oppone a presenza in città di Aleida, figlia di Che Guevara [🔗](#)

 247libero Wednesday, March 6, 2019 4:13:00 PM CET | [info](#) [en](#) [en] [other]

. Al di là della simpatica paternale che la Francesca Vignoli ha redatto, del cui testo, tra l'altro, sono rimasto piacevolmente sorpreso, in quanto ho scoperto una satira ed uno humor che non avrei creduto; mi sono soffermato sul contenuto del suo articolo (che cito per intero) che risulta povero.....

KappaViaggi, tutto è pronto per un'estate tailor made [🔗](#)

 247libero Wednesday, March 6, 2019 1:10:00 PM CET [en](#) [en] [other]

Denso di novità il 2019 di KappaViaggi , già pronta ad affrontare la stagione estiva con ben dieci strutture commercializzate in Italia , di cui sei Coralia Club e quattro strutture a marchio Kappa Club . La ricca selezione di soggiorni e tour targati KappaViaggi è dedicata sia a chi in vacanza.....

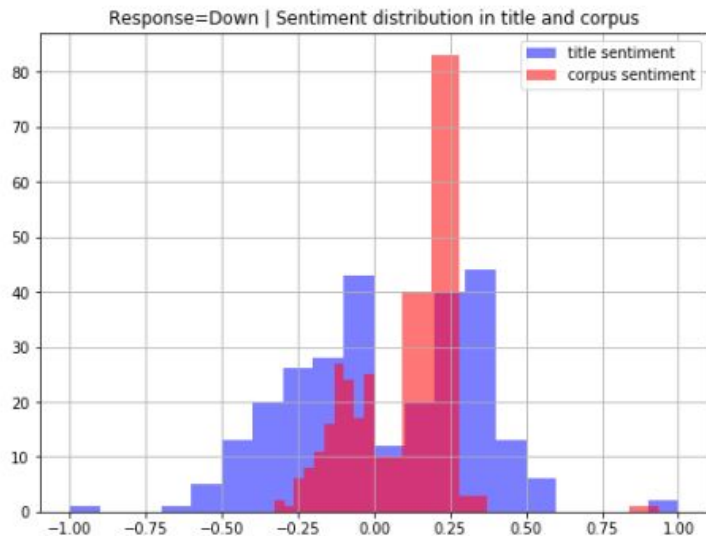
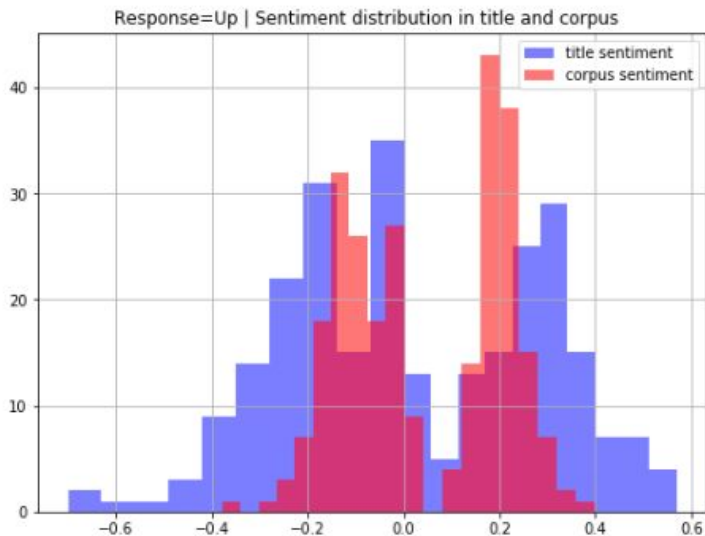
Date, Title (it), Corpus (it)

Google Translate API,
Group by week

Date, Title (en), Corpus (en)

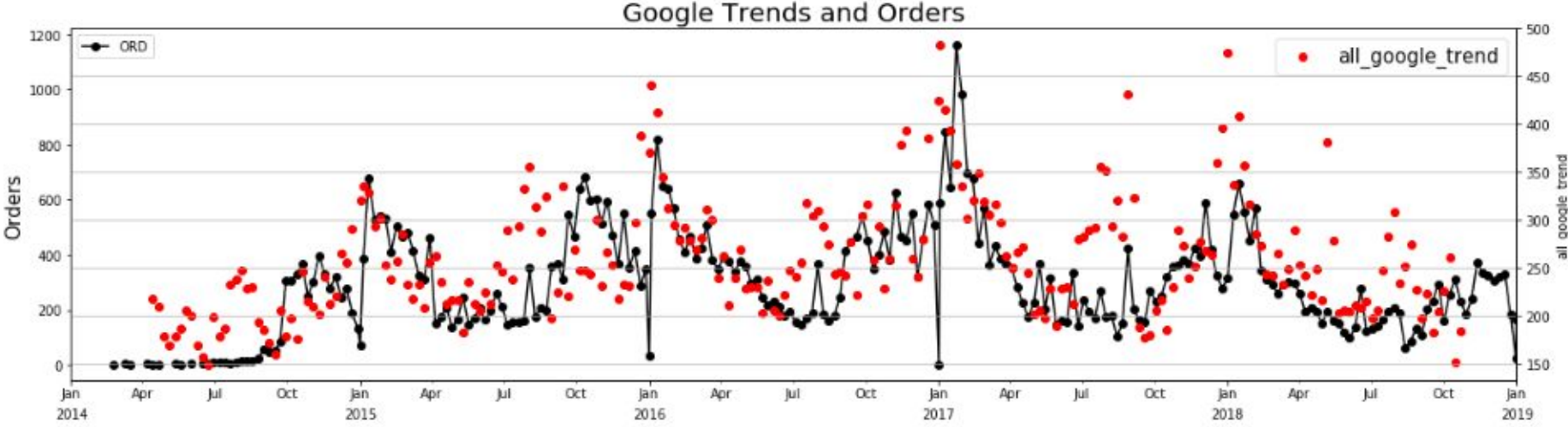
External Data: Media

- Each article (title and corpus) has a sentiment between $[-1, 1]$
- We expected this to have predictive power for our classification models:



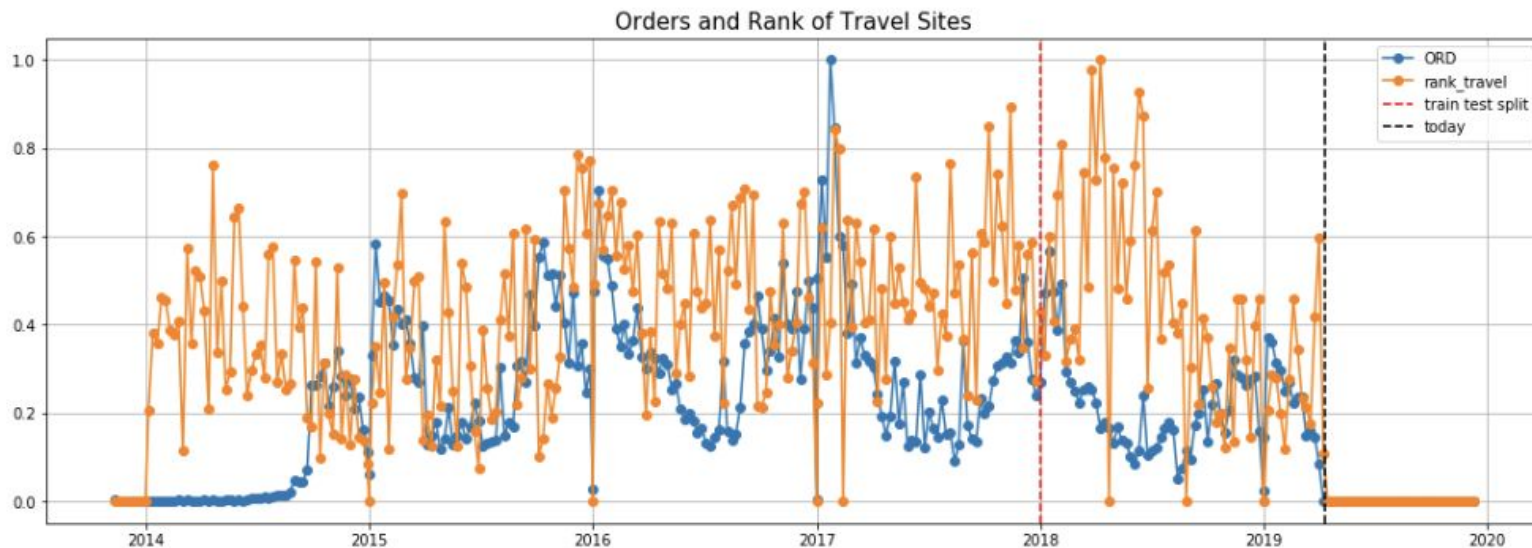
External Data: Google Trends

- Google Trends searches for Cuba vacation, Havana, etc.



External Data: Google Search

- Retrieved the top 100 results for the query “Cuba” for each week since 2014
- Assigned a travel ranking score inversely proportional to the ranks of the pages



Literature Review: Topics and Justification

- Time series data:
 - ARIMA
 - RNN (sequence modeling for time series prediction)
 - Additive Model

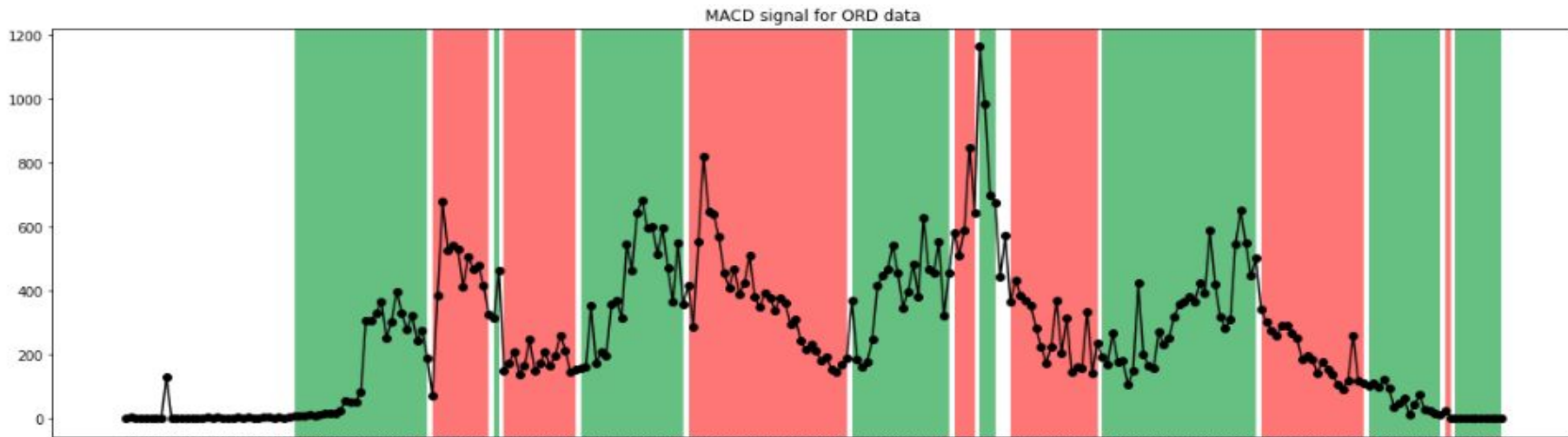
- NLP feature generation (make use of small amount of news information):
 - Bag of words (BOW)
 - Term frequency–inverse document frequency (TF-IDF)
 - Latent Dirichlet Allocation (LDA)

Literature Review: Topics and Justification

- Classification:
 - Random Forest Classifier
 - Multi-layer Perceptron
 - Logistic Regression
- Regression:
 - ARIMA
 - RNN (sequence modeling for time series prediction)
 - Additive Model
 - Gaussian Process
 - Other: MLP, Random Forest Regressor

Feature Generation: Booking data

- MACD: Moving average convergence divergence
- Captures momentum in demand for target Y at each time step (in weeks)
 - MACD line: $MA(Y, 12 \text{ weeks}) - MA(Y, 26 \text{ weeks})$
 - Signal line: $MA(\text{MACD line}, 9 \text{ weeks})$
 - “Up” feature: $\text{Indicator}\{\text{MACD line} > \text{Signal line}\}$

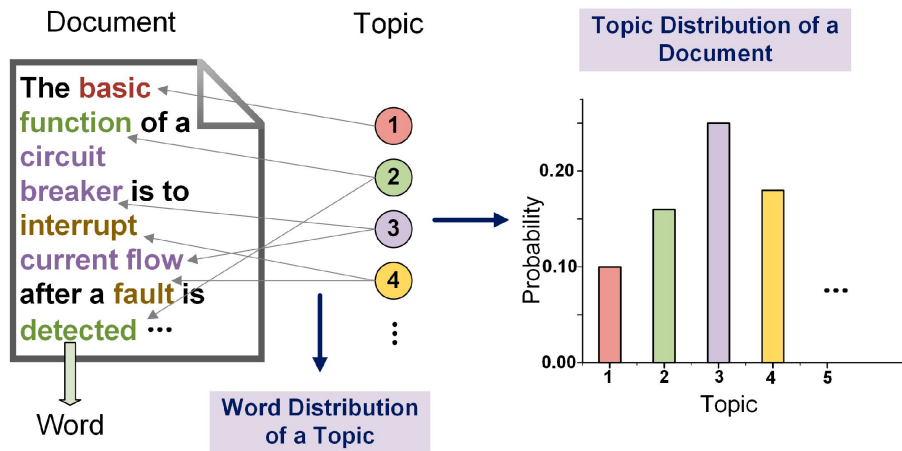


Feature Generation: NLP on media news

- BOW, TF-IDF
 - Each week, for titles and corpuses joined together, we built:
 - **BOW (Bag of Words)**: a sparse matrix with frequencies of each word
 - **TF-IDF**: sparse matrix with values in range [0, 1]
 - TF (Term Frequency): weight of a term proportional to its frequency
 - IDF (Inverse Document Frequency): specificity as an inverse function of the number of documents in which it occurs
 - Used as a weighting factor

Feature Generation: NLP on media news

- LDA (Latent Dirichlet Allocation)
 - Tried to extract topics from news and correlate them to sentiment/trends
 - “Fuzzy” clustering
 - No meaningful topic extracted, high overlap



Feature Generation

- We also created polynomial features from our original features
- i.e. we considered how the target variable (orders or departures) varied with the squares or cross-terms of our features

Classification Models

- Given the features of the current week, will the average number of orders per week in the next 1 month increase or decrease? 2 months? 3 months?
- Classification models we used:
 - Logistic regression, random forest, multi-layer perceptron, decision trees, etc.

Benchmark: Logistic Regression, 1 month

- Using D features, predict the label (1/0) with probability p

$$p = \sigma(\theta^\top x)$$

$$x, \theta \in R^D$$

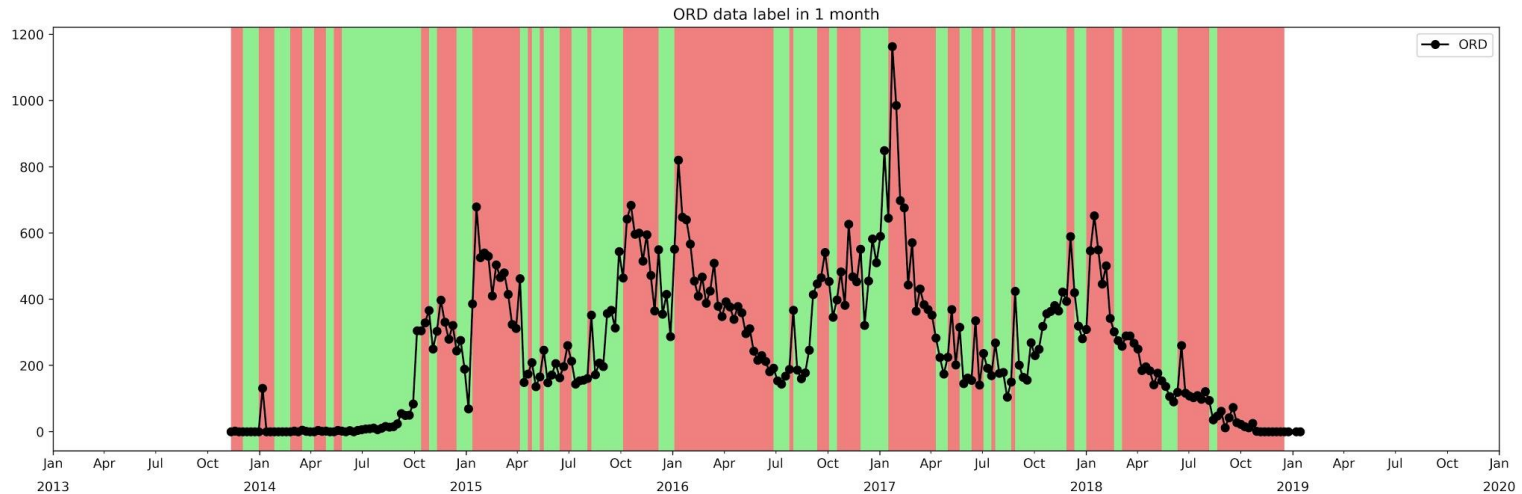
- Results:
 - **Mean AUC:** 0.566
 - **Mean Accuracy:** 0.575
- Next, we attempted to use our generated features

Classification Models Results

- Media News data:
 - Adding sentiment features did not significantly improve performance
 - In general, news sentiment was not a strong predictor of future demand
- Added MACD feature: buy/sell signal
 - **AUC:** 0.600
 - **Accuracy:** 0.624
- At the end of March, we received new data, which corrected some empty values in Alpitour's booking data from December 2018

Classification of future avg. order demand (up/down)

AUC of order direction	1 mo.	2 mo.	3 mo.
Multi-layer Perceptron	0.75	0.78	0.78
Logistic Regression	0.73	0.76	0.79



Regression Models

- Benchmark Model:
 - Direct modeling using ARIMA, Random Forest Regressor, MLP, etc.
 - Best one: Random Forest Regressor

- Time Series Models:
 - Additive seasonal model with various regression models on trend
 - Gaussian Process, MLP, RF
 - Gaussian Process regression with seasonality kernel
 - LSTM

Regression Models

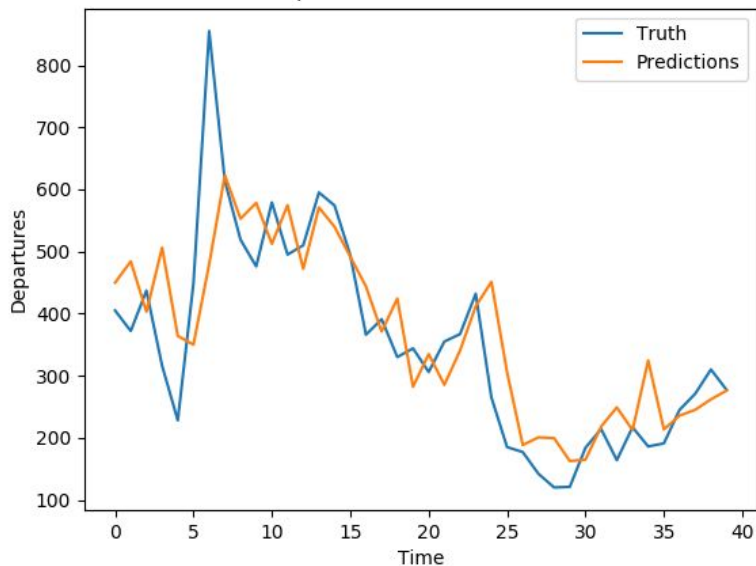
- Benchmark Model:
 - Direct modeling using ARIMA, Random Forest Regressor, MLP, etc.
 - Best one: Random Forest Regressor

- Time Series Models:
 - Additive seasonal model with various regression models on trend
 - Gaussian Process, MLP, RF
 - Gaussian Process regression with seasonality kernel
 - LSTM
 - Best one: Additive seasonal model with Gaussian Process

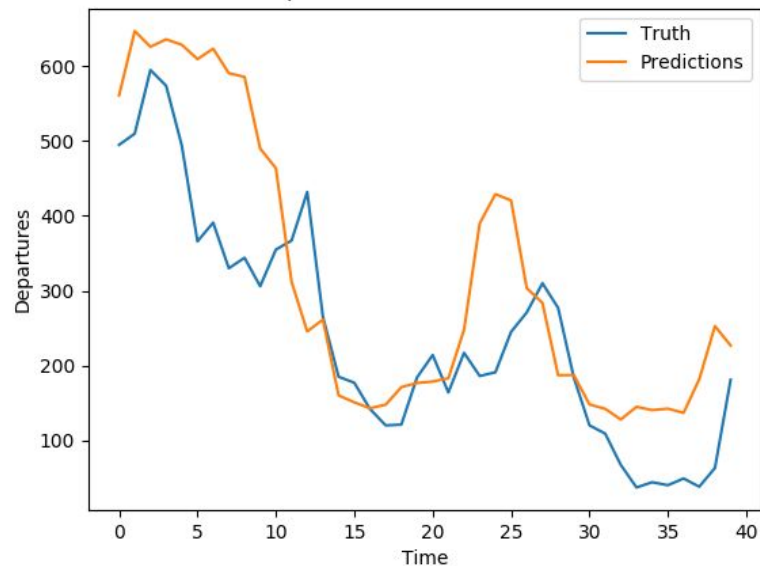
Benchmark Regression of future departures (RF)

Week	1	2	3	4	5	6	7	8	9	10	11	12
R ²	0.65	0.15	0.24	0.19	0.50	0.35	0.43	0.32	0.11	0.13	0.42	0.34

Prediction of departures in the next 1 week. R^2 : 0.65.



Prediction of departures in the next 12 weeks. R^2 : 0.34.

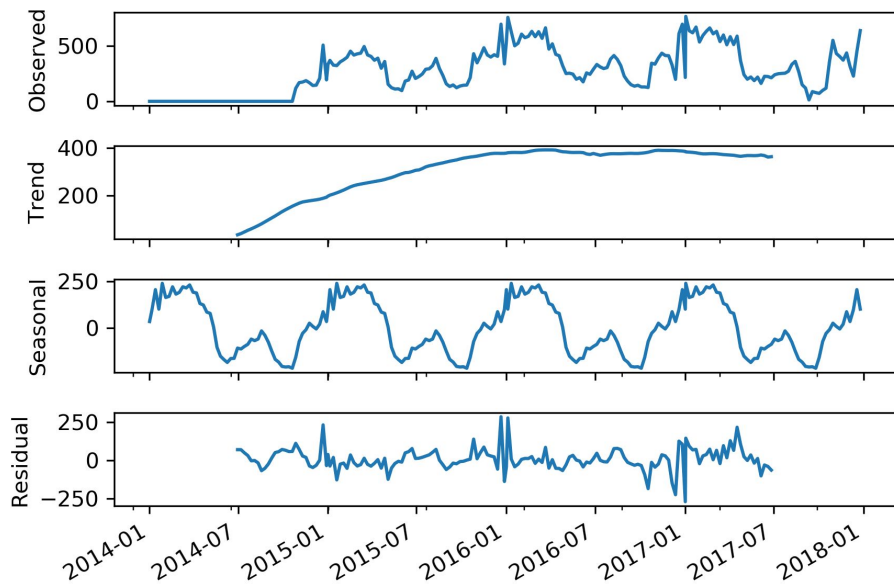


Final Models: Motivation

- Previous models do not consider:
 - Time series
 - Our current models consider each time step to be an independent snapshot of features
 - Seasonal behavior
 - Uncertainty bounds
 - Prediction for cities

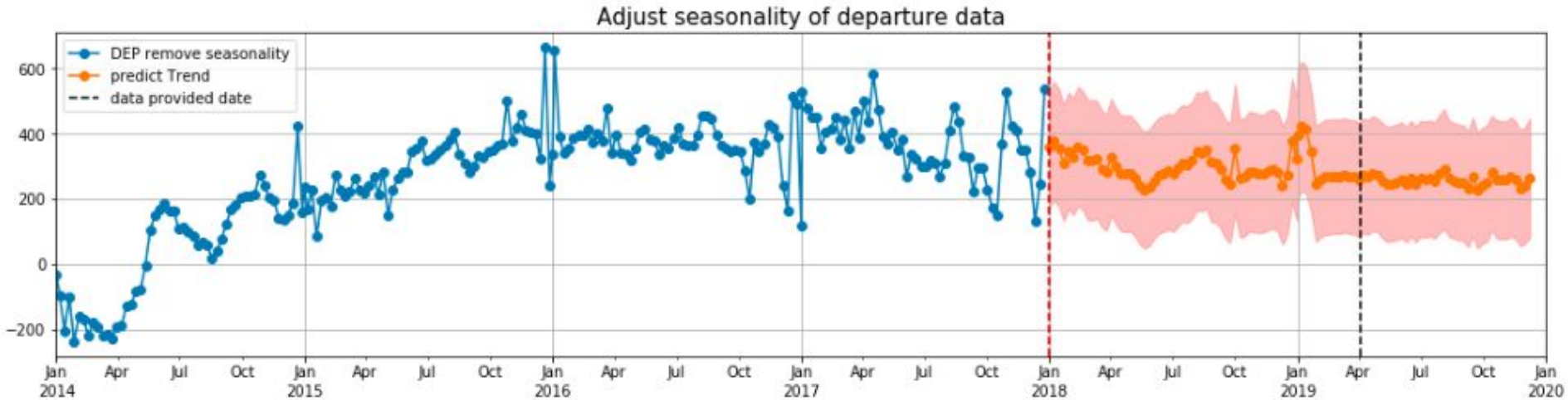
Additive model with Gaussian Process regressor

- Additive model:
 - $Y[t] = S[t] + T[t] + E[t]$
 - Periodicity: 52

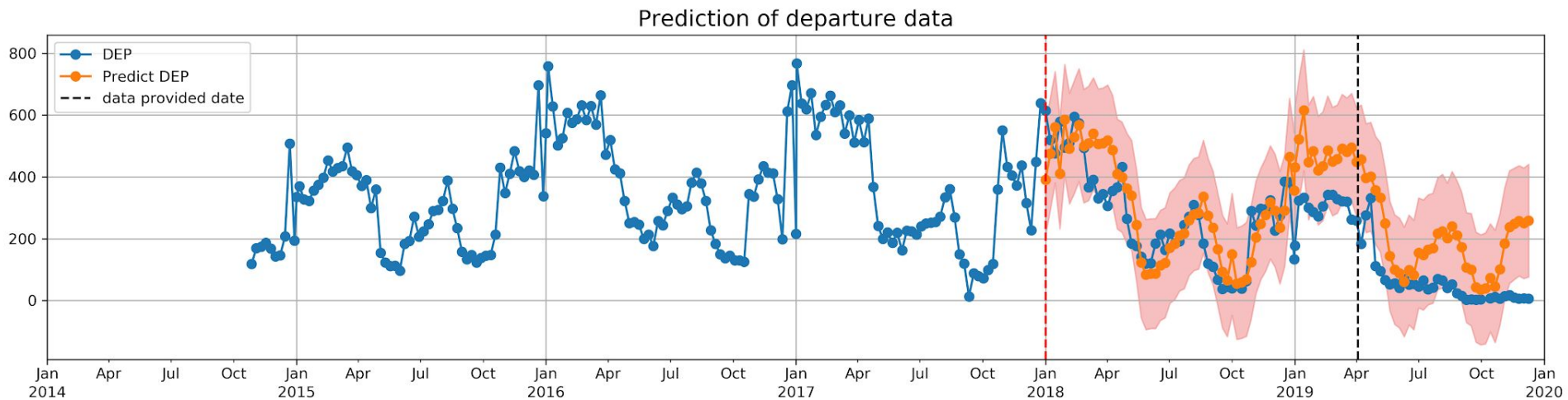


Additive model with Gaussian Process regressor

- Additive model: $Y[t] = S[t] + T[t] + E[t]$
- Then, Gaussian Process regression to predict the trend
 - Use all features such as Google Trends (next step: feature selection)
 - Kernel: Dot Product + White Noise Kernel

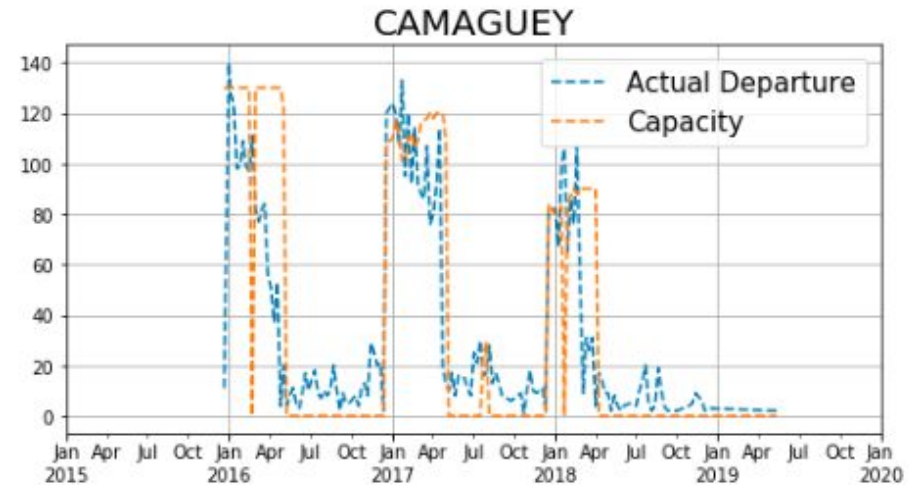
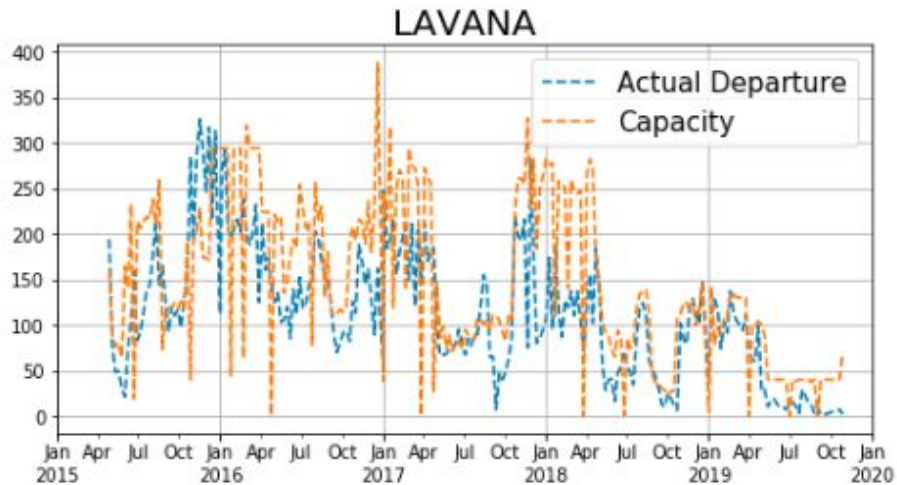


Additive model with Gaussian Process regressor



Application: Predictions for each city

- Unequal number of data points for each city
- Capacity plot for the most and least popular cities: L'Avana, Camagüey



Predictions of departures to L'Avana

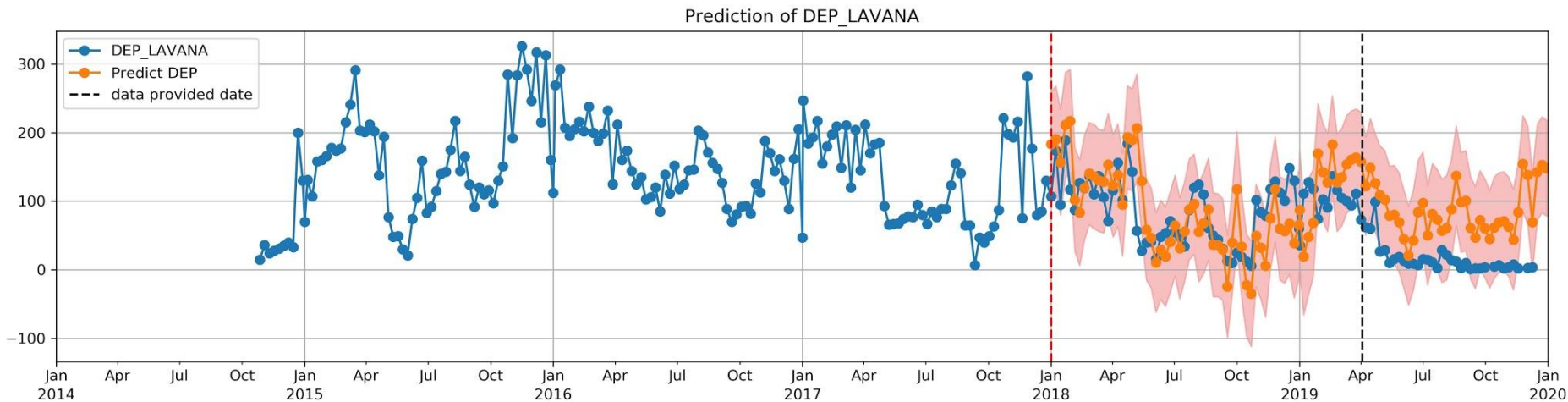
- Example: L'Avana
- Benchmark: Predict the year 2018 departures with mean of training set
- Metric:

$$\text{Ratio of MSE} = \frac{\text{MSE}_{\text{benchmark}}}{\text{MSE}_{\text{model}}}$$

Model	Ratio of MSE: (larger number is better)
Additive + Gaussian Process regression	2.37
Additive + MLP regression	1.50
Additive + RF regression	1.40



Prediction of departures to L'Avana

- Using Gaussian Process regression on the trend:



Now, can we quantify the
business value?

Business Value

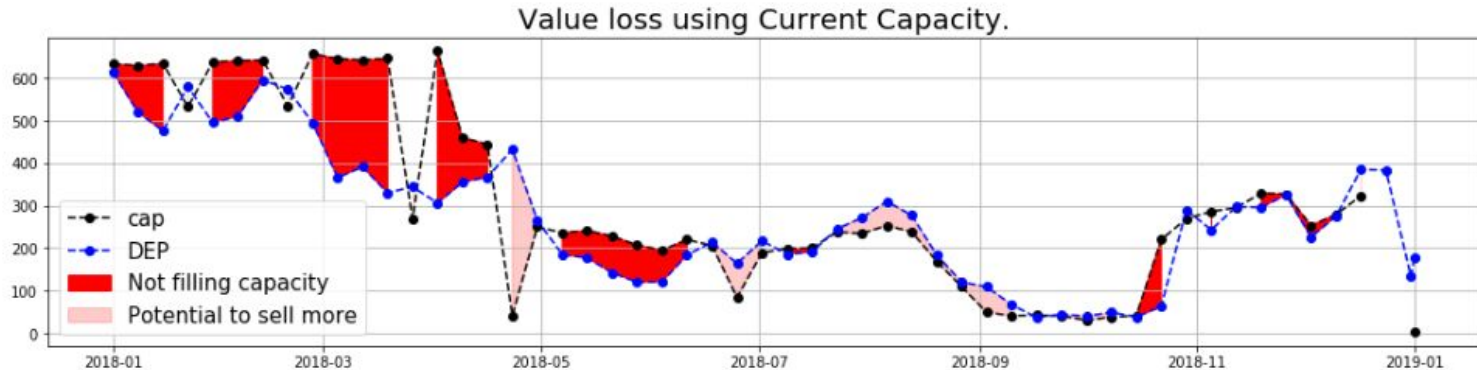
- Alpitour books flights up to 1 year in advance, so they need predictions of departures up to 1 year in advance
-  If Alpitour doesn't fill its capacity, it may **lose money**
-  If Alpitour fills its capacity, maybe it could have **sold more trips**

Estimated Business Value (Revenue)

- We want to quantify the value Alpitour can gain by using our model
- Use the predicted departures for each week in the validation set 2018 as the updated capacity
- Comparing the value lost by Alpitour using its current planned capacity vs. the capacity from our predicted departures

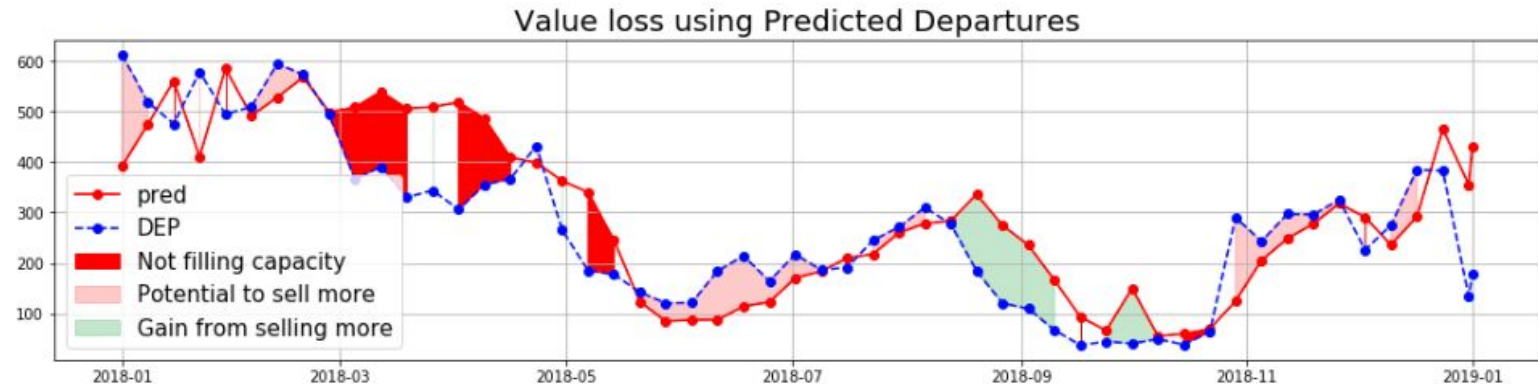
Estimated Business Value

- In week t , Y_t is our target variable (departures). P_t is Alpitour's price/person for a flight for a booked trip in week t , estimated at 10% of the avg. price/person for a booked trip at week t
- Value lost using current planned capacity:
 - If $C_t > Y_t$ then Alpitour does not fill its capacity and loses value: $P_t^* (C_t - Y_t)$
 - If $Y_t > C_t$ then Alpitour could book more flights: $P_t^* (Y_t - C_t)$



Estimated Business Value

- Value lost with predicted departures Y'_t as capacity instead of planned capacity:
 - If $Y'_t > Y_t$ then:
 - if $Y_t < C_t$: $P_t * (Y'_t - Y_t)$ Booked more flights than demand
 - if $Y_t > C_t$: 0 Earn back the potential demand
 - If $Y_t > Y'_t$ then Alpitour could book more flights and increase revenue:
 - $P_t * (Y_t - Y'_t)$ (opportunity cost)



Estimated Business Value

- Unbooked value lost using current capacity:
 - € 505,413.52
- Unbooked value lost using predicted departures as the planned capacity:
 - € 440,070.02
- We save **€ 65,343.50** of value for the Cuba case in 2018
- This is revenue, not profit, so the true profit will be smaller
- This depends on Alpitour's operating profit margin

Deliverables and Impact

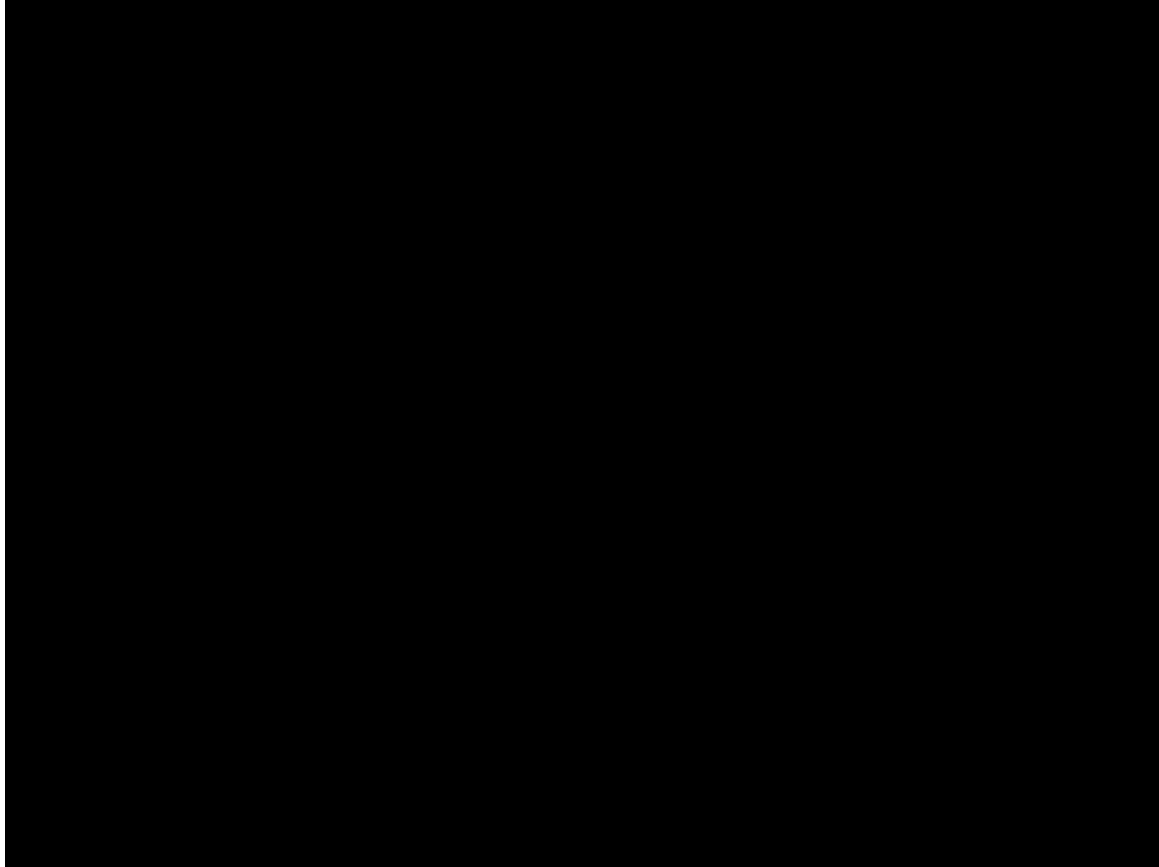
1. Front-end dashboard: learn how to use our code for future case studies

- <https://datashack2019.github.io/alpitour-datashack2019/>

2. Back-end pipeline:

- Enable engineers and data analysts to replicate our methodology for other locations
- Provide supply managers with predictions to inform capacity decisions
- <https://github.com/Harvard-IACS/2019-AC297rs-alpitour/blob/master/notebooks/DEMO.ipynb>

Frontend Dashboard



Cuba

Havana-HVA

[Back to the map](#)

From January 2019



To January 2020



How to read it



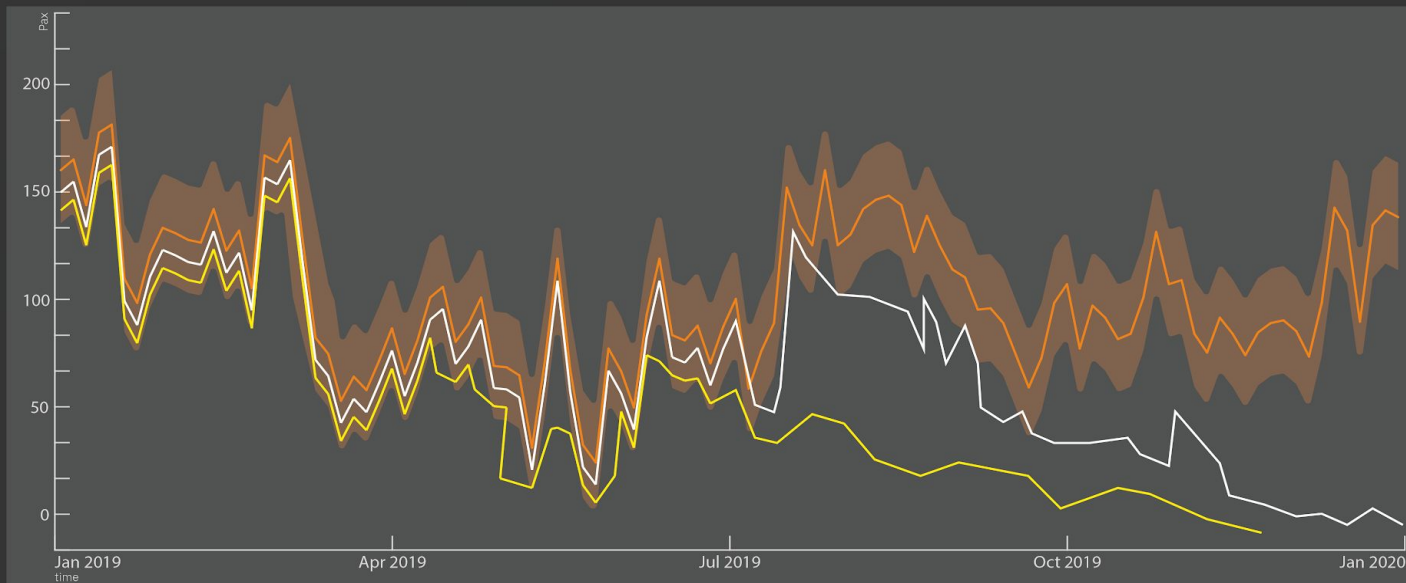
Predicted



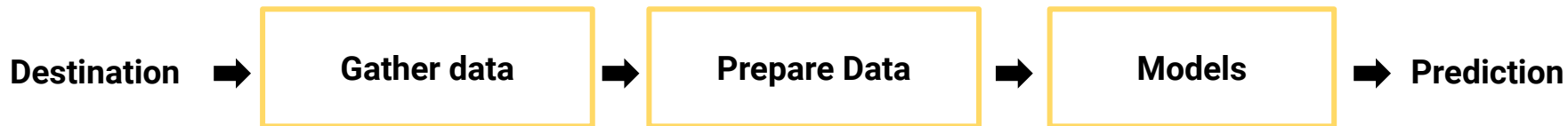
Capacity (number of orders purchased by Alpitour)



Number of orders purchased by the customers



Back-end pipeline



Example Input

- Engineers update internal datasets for desired location
- Use the gather data module to update external data
- One click to process all data

Example Output

- Prediction of departures in future periods with uncertainty estimation

Future Work for Alpitour

- Model improvement:
 - Update data of useful features
 - Improve the method of handling missing feature values in future weeks:
 - Use the mean for the corresponding week, rather than mean overall
- Combine predictions of departures with:
 - Ability to change capacity of flights per-seat vs. booking a full new plane
 - Hotel availability

Literature Review: Citations

- **ARIMA**
 - Hillmer, Steven Craig, and George C. Tiao. "An ARIMA-model-based approach to seasonal adjustment." *Journal of the American Statistical Association* 77.377 (1982): 63-70.
- **Recurrent neural networks (RNN/LSTM)**
 - Connor, Jerome T., R. Douglas Martin, and Les E. Atlas. "Recurrent neural networks and robust time series prediction." *IEEE transactions on neural networks* 5.2 (1994): 240-254.
- **Generalized additive models (GAM)**
 - Hastie, Trevor J. "Generalized additive models." *Statistical models* in S. Routledge, 2017. 249-307.
- **Gaussian process**
 - Rasmussen, Carl Edward. "Gaussian processes in machine learning." *Summer School on Machine Learning*. Springer, Berlin, Heidelberg, 2003.

Thank you