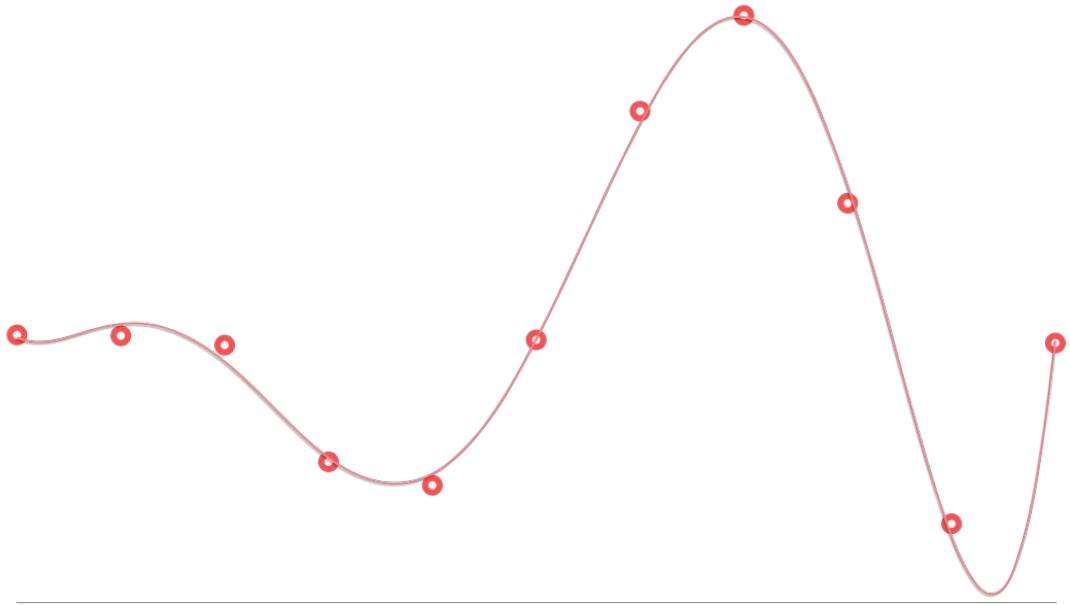
Kai Brusch / April 18th, 2019 / Data Council SF

Delphi: a hybrid approach to forecasting a global marketplace

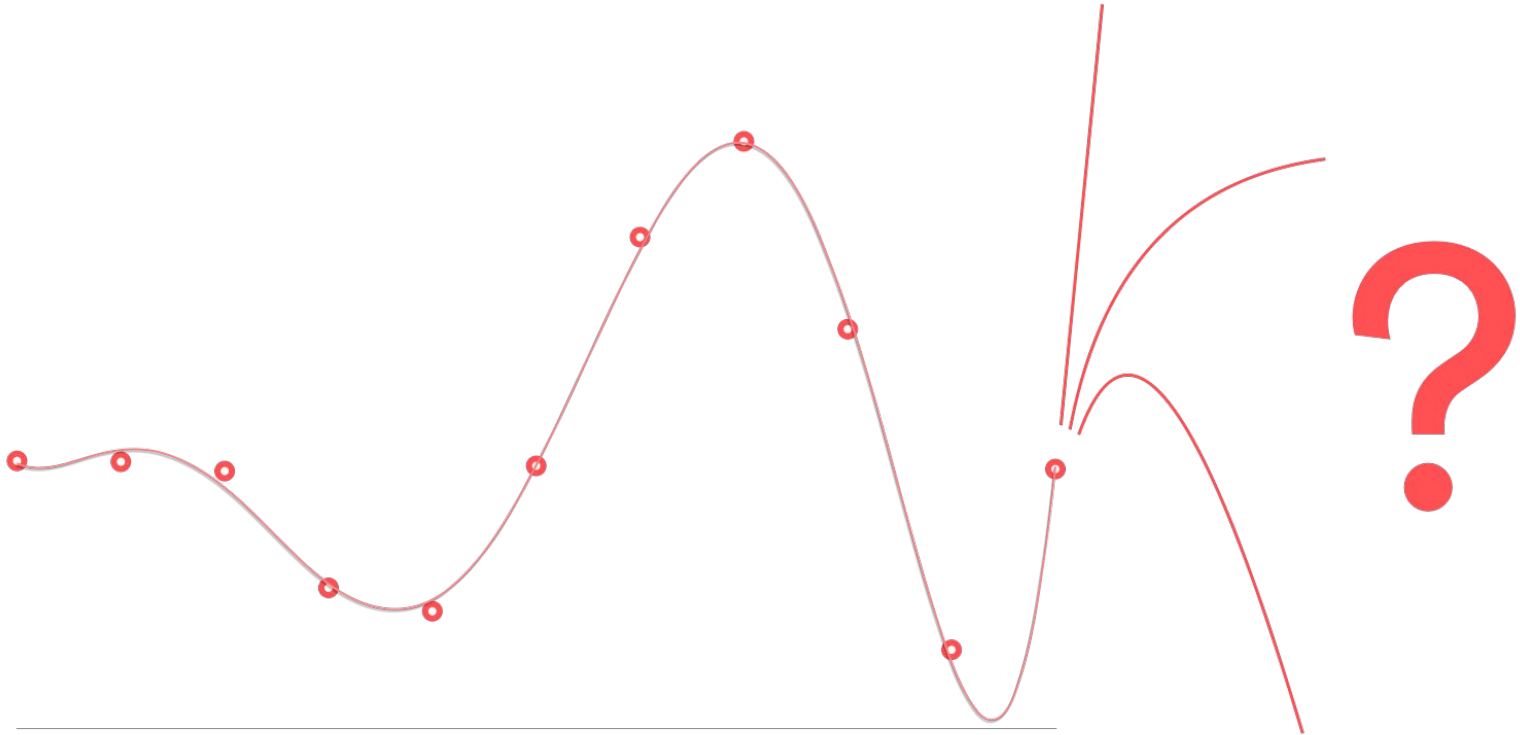
Machine Learning is very good at interpolation

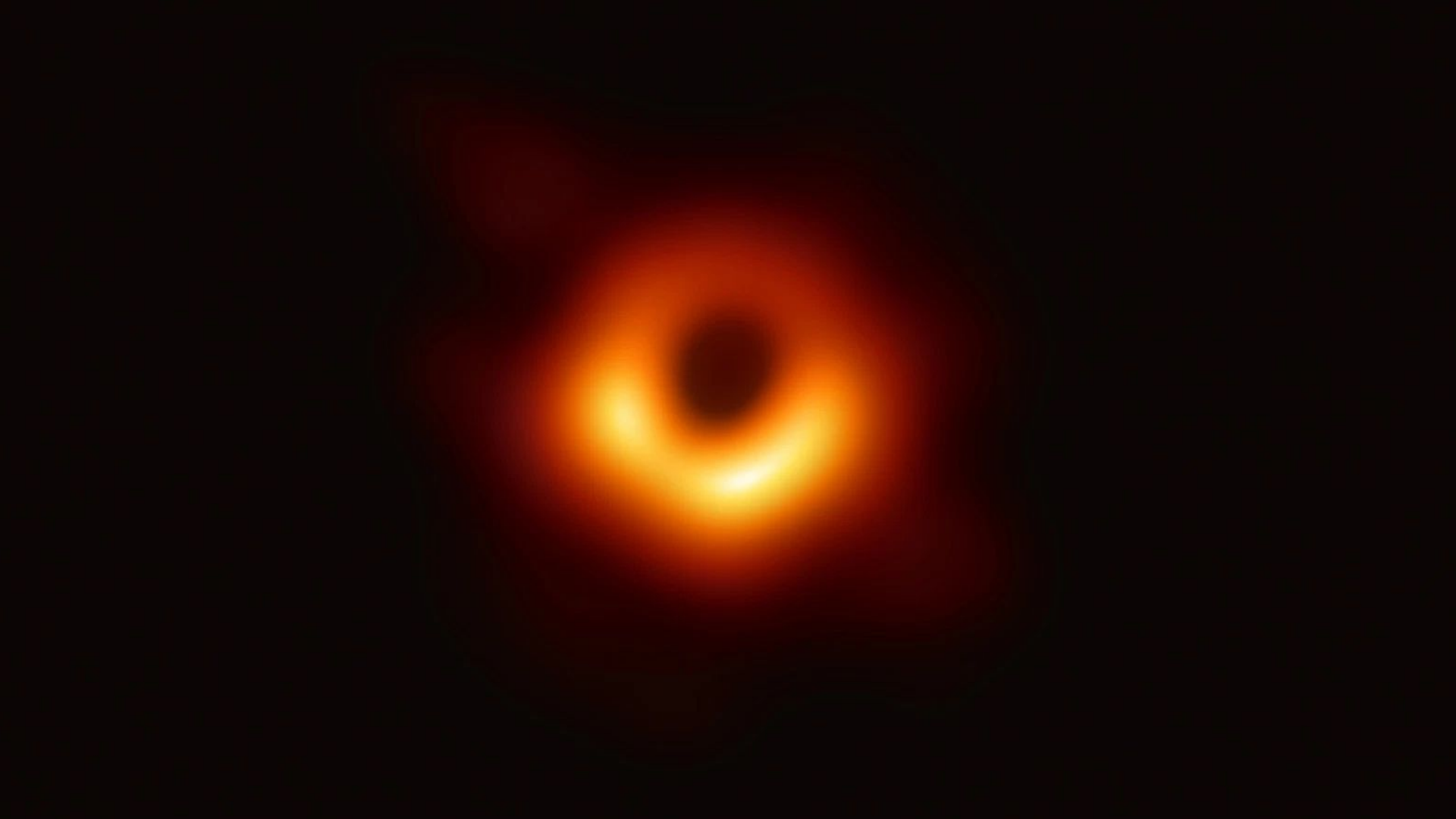
Purely optimizing the error function with an arbitrary number degree of freedom will always be able to perfectly fit



But pure Machine Learning struggles with extrapolation

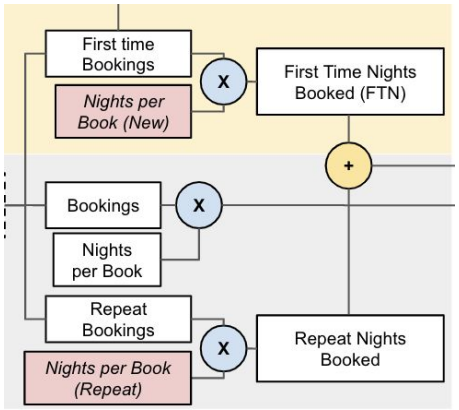
Predictions on out of training samples are a notoriously hard problem

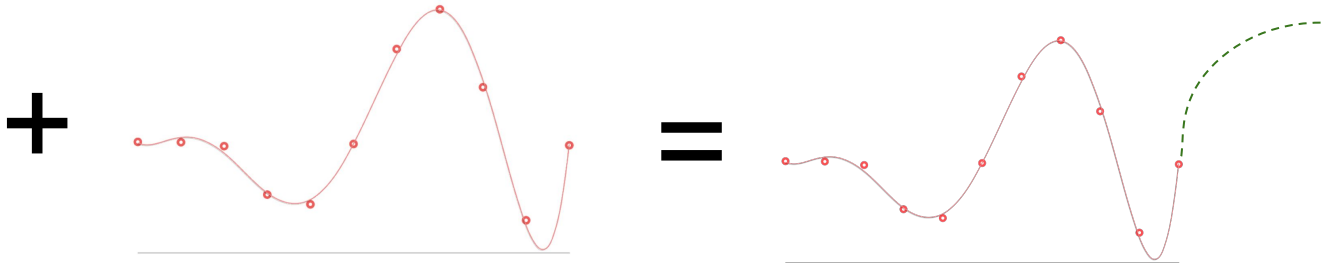




A hybrid between statistical and causal extrapolation

A strong theoretical framework allows to reliably forecast a global marketplace





Agenda

Intro

What is our approach to forecasting and how do we think about metrics?

Statistical Forecasting

How do we estimate the seasonality of supply and demand?

Metric Graph

How do we define the underlying theoretical framework?

Delphi

How does Delphi realize this hybrid approach?

Agenda

Intro

What is our approach to forecasting and how do we think about metrics?

Statistical Forecasting

How do we estimate the seasonality of supply and demand?

Metric Graph

How do we define the underlying theoretical framework?

Delphi

How does Delphi realize this hybrid approach?

Agenda

Intro

What is our approach to forecasting and how do we think about metrics?

Statistical Forecasting

How do we estimate the seasonality of supply and demand?

Metric Graph

How do we define the underlying theoretical framework?

Delphi

How does Delphi realize this hybrid approach?

Agenda

Intro

What is our approach to forecasting and how do we think about metrics?

Statistical Forecasting

How do we estimate the seasonality of supply and demand?

Metric Graph

How do we define the underlying theoretical framework?

Delphi

How does Delphi realize this hybrid approach?

Agenda

Intro

What is our approach to forecasting and how do we think about metrics?

Statistical Forecasting

How do we estimate the seasonality of supply and demand?

Metric Graph

How do we define the underlying theoretical framework?

Delphi

How does Delphi realize this hybrid approach?

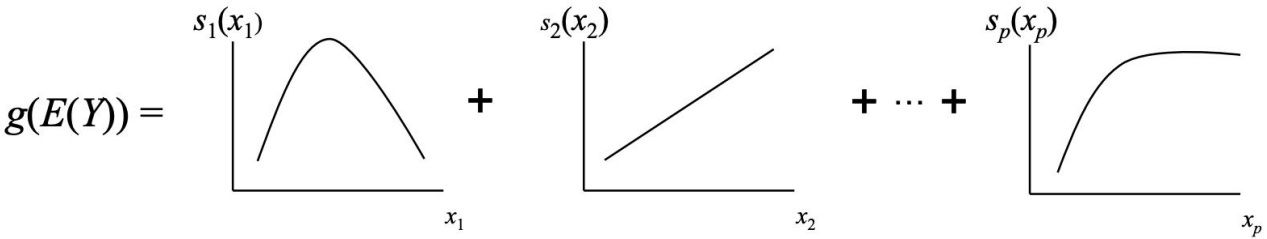
Regression + extensions are the answer to interpretability

Our hybrid approach dictates the model selections to interpretable models

- Interpretable models > black box
  - Main assumption for connection to metric graph
  - Only way to derive business value is interpretability
- Generalized Linear Model (GLM) is the statistical foundation
- Expected: seasonality + events
  - GLM + seasonality = Generalized Additive Model (GAM)
- Unexpected events
  - GLM + random effects = Generalized Linear Mixed Models (GLMM)

Seasonal estimation with Generalized Additive Models

GAM extend the GLM framework with seasonality estimation



- Models the expectation of link function as sum of unknown smoothing functions
- Represent smoothing functions as B-Splines (mgcv)
- Example: Estimate bookings with a nights booked model

[1,2]

Every booking happens from a date

20.3

nights booked: date: date_x: delta:

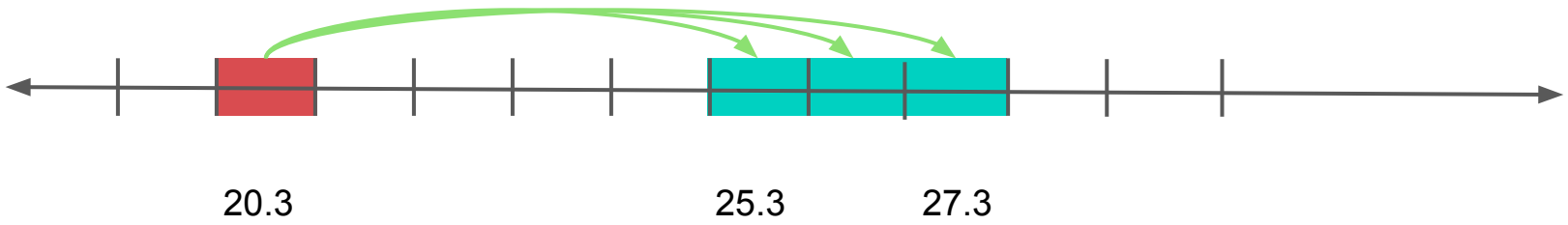
For several future nights on date_x

(20.3 ; 25.3 ; 1)
(20.3 ; 26.3 ; 1)
(20.3 ; 27.3 ; 1)

nights booked: date: date_x: delta:

Add the delta between date and date_x

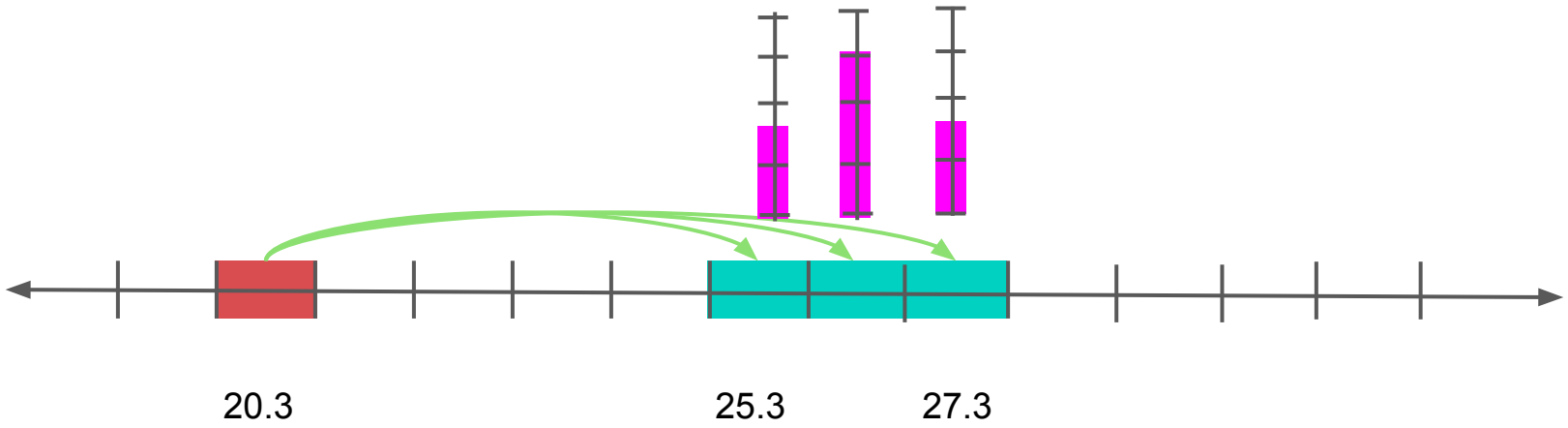
(20.3 ; 25.3 ; 1 ; 5)
(20.3 ; 26.3 ; 1 ; 6)
(20.3 ; 27.3 ; 1 ; 7)



nights booked: date: date_x: delta:

Those future dates already have some bookings

(20.3 ; 25.3 ; 1 ; 5 ; 0,2)
(20.3 ; 26.3 ; 1 ; 6 ; 0,9)
(20.3 ; 27.3 ; 1 ; 7 ; 0,3)



nights booked: date: date_x: delta: occupancy:

20.3 ; 25.3 ; 27 ; 5 ; 0,2
20.3 ; 26.3 ; 30 ; 6 ; 0,9
20.3 ; 27.3 ; 11 ; 7 ; 0,3
20.3 ; 25.3 ; 3 ; 5 ; 0,2
21.3 ; 26.3 ; 2 ; 5 ; 0,9
...
12.3 ; 27.3 ; 9 ; 7 ; 0,3
11.3 ; 25.3 ; 4 ; 5 ; 0,2
19.3 ; 26.3 ; 1 ; 6 ; 0,9
30.12 ; 31.12 ; 21 ; 7 ; 0,99

```
model_gam = bam(
    value
    ~ 0
    + weekday
    + early_growth + last_12_months
    + last_24_months + last_36_months
    + last_48_months + last_60_months
    + event_index:event
    + weekday:event
    + s(share_of_year, k=length(knotsYear), bs="cc")
    + s(delta, k=length(knots_delta), by = weekday)
    + s(share_of_year_x, k=length(knotsYear), bs="cc")
    + s(share_of_year_x, k=length(knotsYear), by=weekday_offset, bs='cc')
    + weekday_x
    + event_index_x:event_x
    + event_x:weekday_offset
    + growth_x:weekday_offset
    + offset(-occupancy_index)
      , family=quasipoisson()
)
```

nights booked:

date:

delta:

date_x:

Occupancy index:

```
model_gam = bam(
    value
    ~ 0
    + weekday
    + early_growth + last_12_months
    + last_24_months + last_36_months
    + last_48_months + last_60_months
    + event_index:event
    + weekday:event
    + s(share_of_year, k=length(knotsYear), bs="cc")
    + s(delta, k=length(knots_delta), by = weekday)
    + s(share_of_year_x, k=length(knotsYear), bs="cc")
    + s(share_of_year_x, k=length(knotsYear), by=weekday_offset, bs='cc')
    + weekday_x
    + event_index_x:event_x
    + event_x:weekday_offset
    + growth_x:weekday_offset
    + offset(-occupancy_index)
      , family=quasipoisson()
)
```

Event detection with Generalized Linear Mixed Model

GLMM extend the GLM framework with random effects

$$\mathbf{y} = \mathbf{X}\beta + \qquad \epsilon.$$

- Observations come from groups which may have varying slopes and intercepts
- GLMM uses random and fixed effects hence the name mixed models (lme4)
- Example: We have several observations of each date in the future

[3]

Event detection with Generalized Linear Mixed Model

GLMM extend the GLM framework with random effects

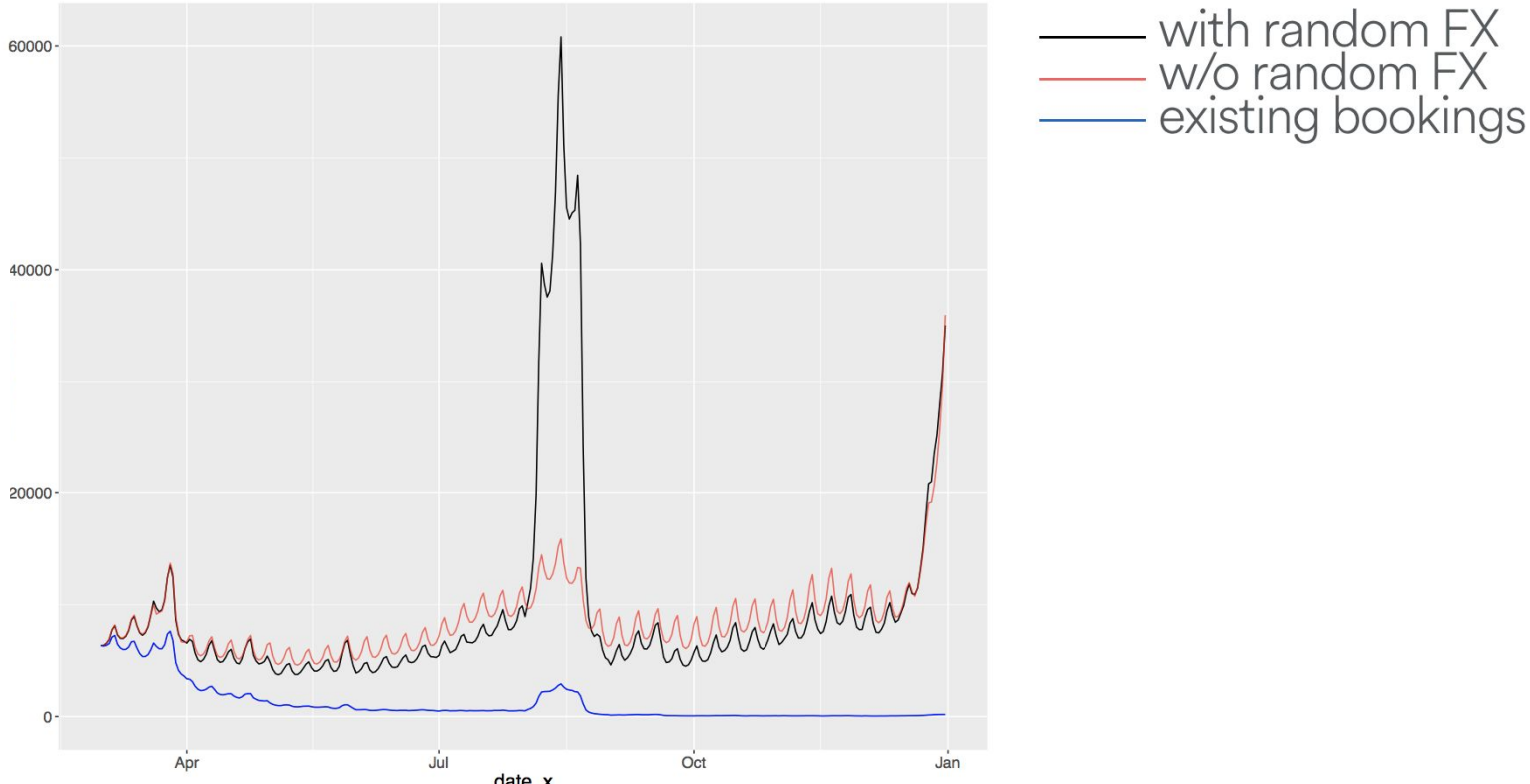
$$\eta \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}),$$
$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\eta + \epsilon.$$

- Observations come from groups which may have varying slopes and intercepts
- GLMM uses random and fixed effects hence the name mixed models (lme4)
- Example: We have several observations of each date in the future

[3]

Leveraging pre-existing information to detect events

Successfully detected events we didn't expect



Agenda

Intro

What is our approach to forecasting and how do we think about metrics?

Demand and Supply

How do we estimate the seasonality of supply and demand?

Metric Graph

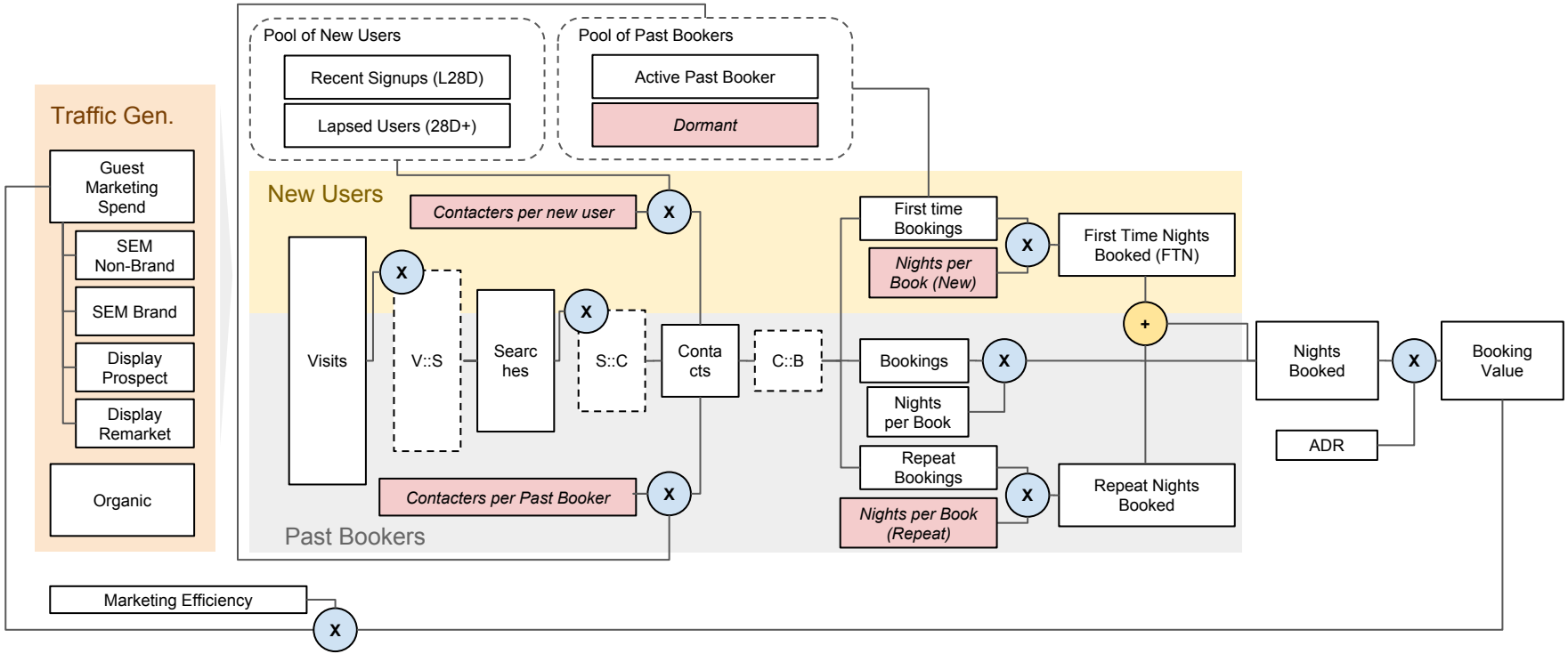
How do we define the underlying theoretical framework?

Delphi

How does Delphi realize this hybrid approach?

Human input: underlying causal framework

Causal relationships between metrics expressed as a graph



Agenda

Intro

What is our approach to forecasting and how do we think about metrics?

Demand and Supply

How do we estimate the seasonality of supply and demand?

Metric Graph

How do we define the underlying theoretical framework?

Delphi

How does Delphi realize this hybrid approach?

Delphi provides a singular interface for a hybrid approach

A DAG to generate DAGs

- Implements a singular interface for statistical models and causal graph
- Produces
  - An Airflow DAG for scalable estimation of statistical models (language independent)
  - Computational engine (Cython) to fuses estimates together
- And a GUI to allow investigation and access to computational engine
- Computational engine facilitates the scenario building:
  - Forward: If I pull now what outcome will I achieve
  - Backward: What levers do I need to pull to get to a goal

with metric()

with facet()

timeshiftOccupancyModel()

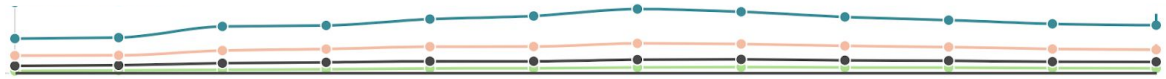
```
with metric('nights booked', date_x='date night', shifted_name='trips in progress before cancellations'):
    with facet(
            [destination_dim, 'guest type l2'], ['date'],
    ):
        DataWarehouse()

    with facet(
            [destination_dim, ('listing stage', 'rookie')], ['delta', 'date'],
    ):
        # If we use the default batch then the query returns too much data and times out. Therefore don't batch
        # any dimension
        DataWarehouse(batch=[])
        Timeshift()
        RookieNightsModel()

    with facet(
            [destination_dim], ['delta', 'date'],
            export='destination', export_shifted=True,
    ):
        # If we use the default batch then the query returns too much data and times out. Therefore don't batch
        # any dimension
        DataWarehouse(batch=[])
        TimeshiftOccupancyModel(
            supply=('active listings', {'listing stage': Select('veteran')}),
        )

    with facet(
            [origin_dim, 'guest type l2'], ['date'],
    ):
        DataWarehouse()
        Pullback(origin_dim, destination_dim, 'propensity nights contacted')
```

History

Charts | Summary | About

total | booking | | from | | Total | Total

2012 2013 2014 2015 2016 2017 2018 2019



Share Metric

Monthly

Linear

Adjust Trend

Plan Config Show Info

Summary

2018 YoY	2018 Exit Rate YoY	2019 YoY	2019 Exit Rate YoY

Markus Schmaus
(Creator)

Jerry Chu, Didi Shi, Chris Lindsey (Engineering)
Jackson Wang, Jiwoo Song, you? (FP&A)

[1] https://multithreaded.stitchfix.com/assets/files/gam.pdf

[2] Simon Wood. Generalized Additive Models : an introduction with R . CRC Press/Taylor & Francis Group, Boca Raton, 2017

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin.
Bayesian Data Analysis Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004