SMART POWER

Can Data Science Change the Way We Invest in Infrastructure?

By A. Issaoui, M. Parent, N. Fazeli, S. Wei, P. Lerchi

Data-X Spring 2018: End of Semester Presentation

OVERVIEW

Investors in Energy Infrastructure are using outdated tools to bet on future electricity markets.

By current market standards, commodity prices are treated as over-simplified features that are incapable of generating reliable long-term predictions. Our group wants to change this practice by predicting retail energy prices to a five year horizon. This project utilizes statistical feature selection methods, time series modeling, and time-variant feature segmentation to develop a flexible tool that can yield insight into financial and policy-based decisions.

APPROACH

PIPELINE DIRECTION

- FEATURE SELECTION
- TIME SERIES PREDICTION
- BASELINE REGRESSION
- VALIDATION
- TIME VARIANCE

FEATURE SELECTION

Check correlation with outcome variable

- 74 original features from last year
- 11 variables with Pearson > 80%
- Easy to swap in result of FS team

Test for autocorrelation, reduce redundancy

- Variance inflation factor (VIF)
- Confirm result with industry experts

	Feature	Pearson
28	Biofuels.Consump.TrillBTU.	0.923888
29	Total.Biomass.Consump.TrillBTU.	0.918654
19	Bio.Prod.Trillion.BTU.	0.914748
20	BioMass.Prod.Trill.BTU.	0.910235
59	Natural.Gas.Consumed.by.the.Transportation.Sec...	0.859504
14	GDP	0.849559
23	GeoConsump.TrillBtu.	0.843676
10	GenCalifornia...other.thousand.megawatthours	0.826934
30	Total.Renewable.Consump.TrillBTU.	0.816141
21	Total.RenProd..TrillBtu.	0.812076
27	WasteConsump.TrillBTU.	0.803539

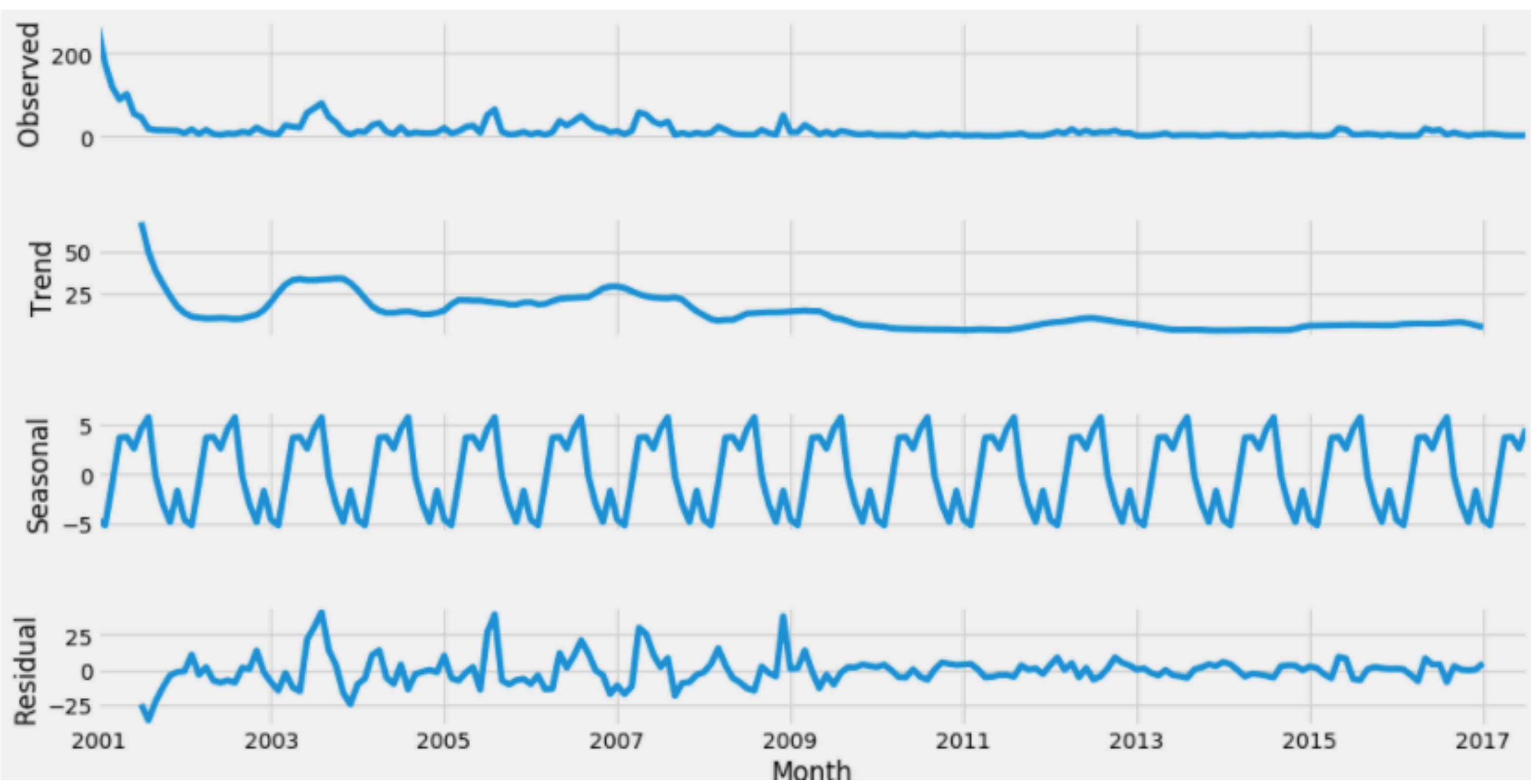
TIMESERIES PREDICTION

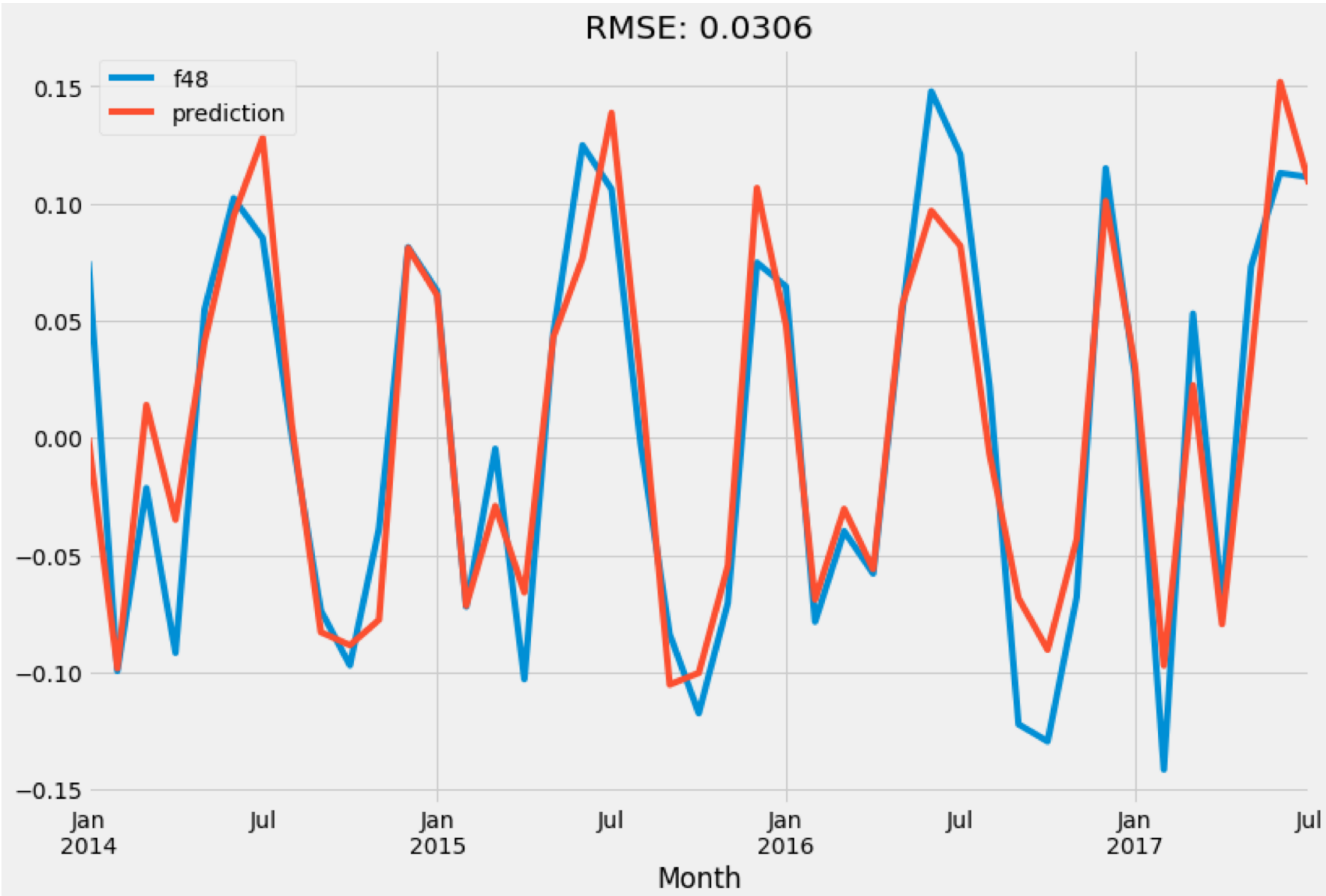
Model Type: S-ARIMA

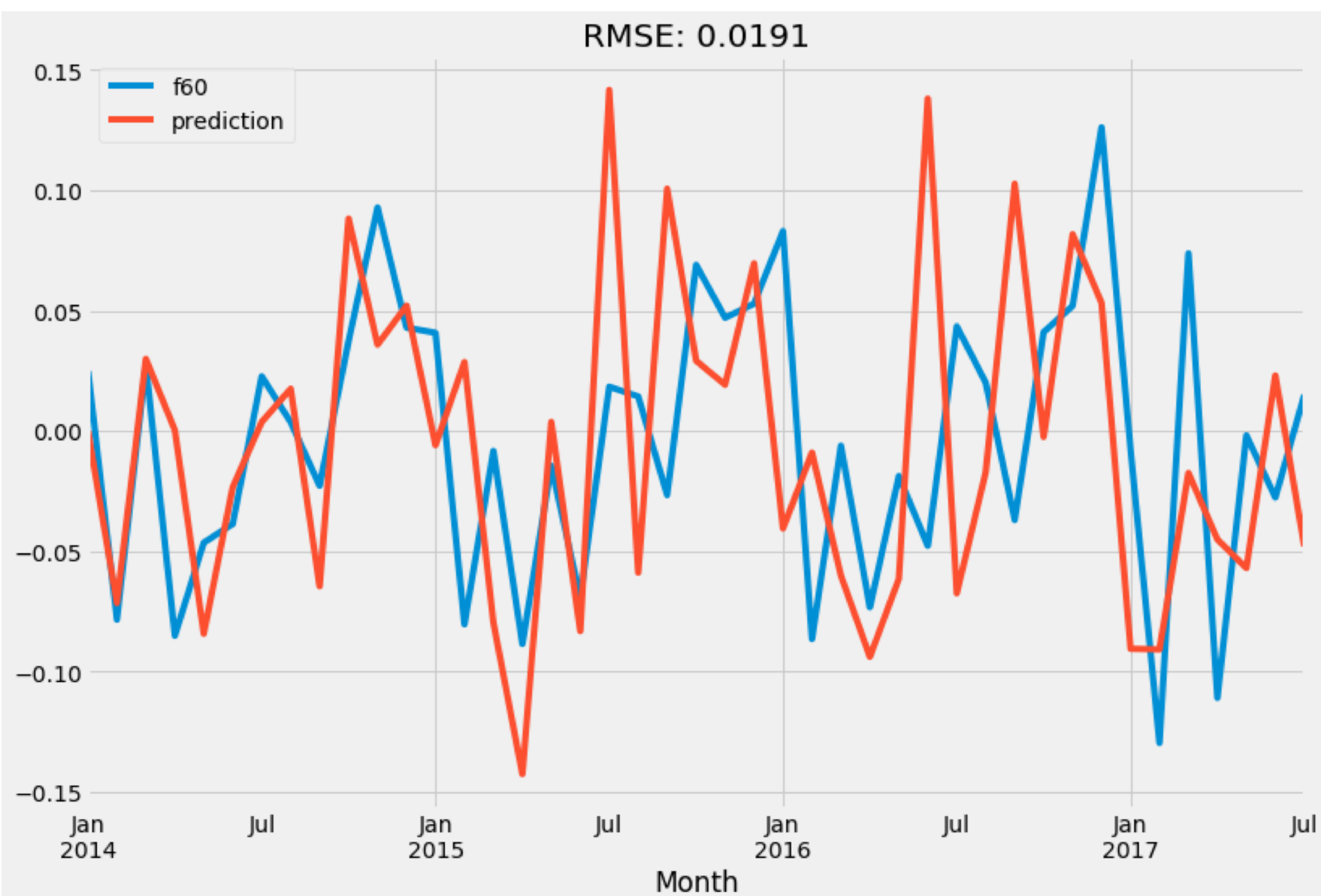
Seasonal Autoregressive Integrated Moving Average

- Done on a per feature basis
- Dickey Fuller test
- First differencing yields stationary series
- Decompose seasonality from moving average
- Parameter selection

Also considered: AR, MA, Dynamic models

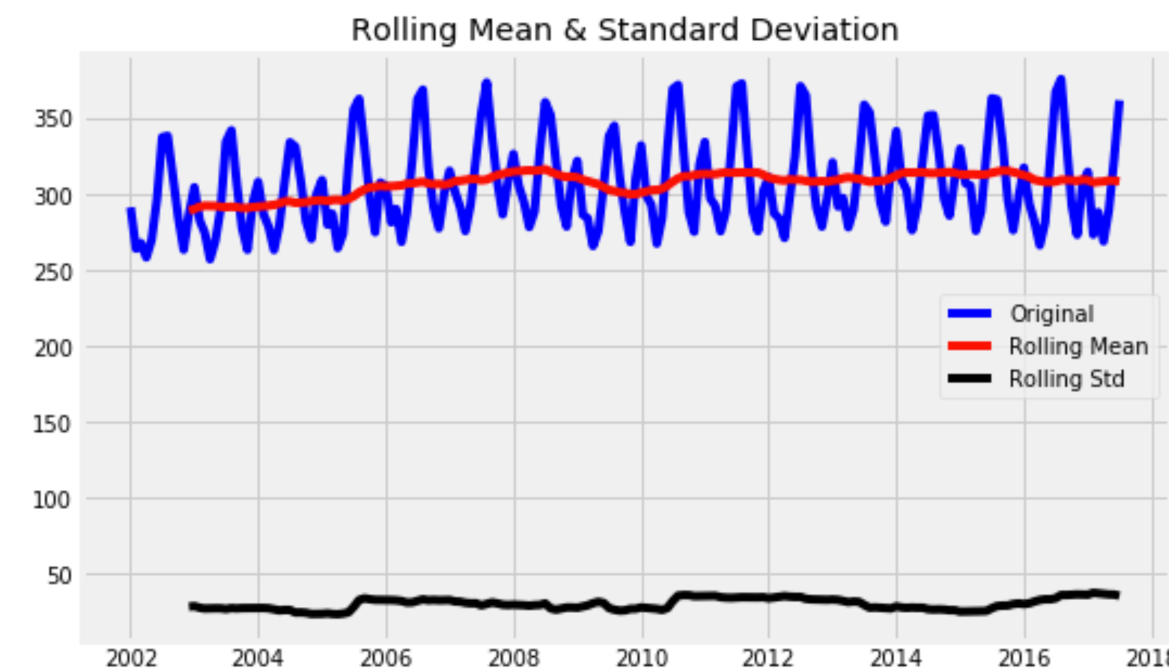


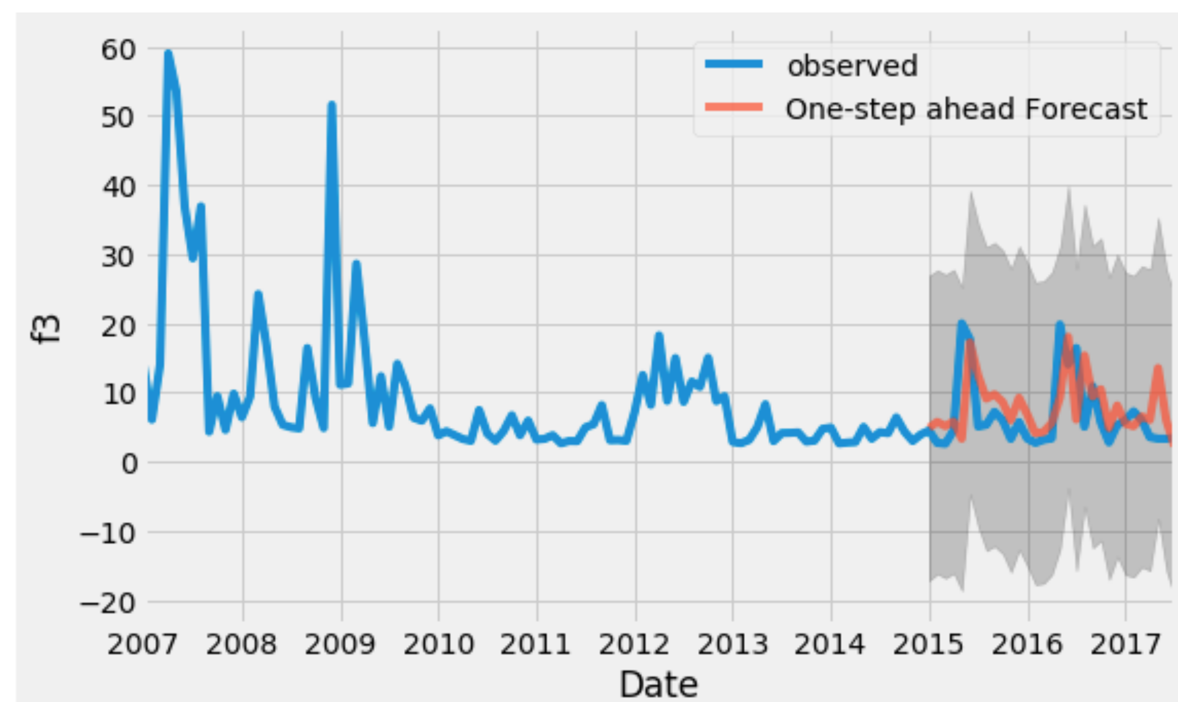




RESULT

- All features were 1st order stationary
- Tuned parameters for each feature
- Reasonable accuracy (most cases)
- Potential time shift?





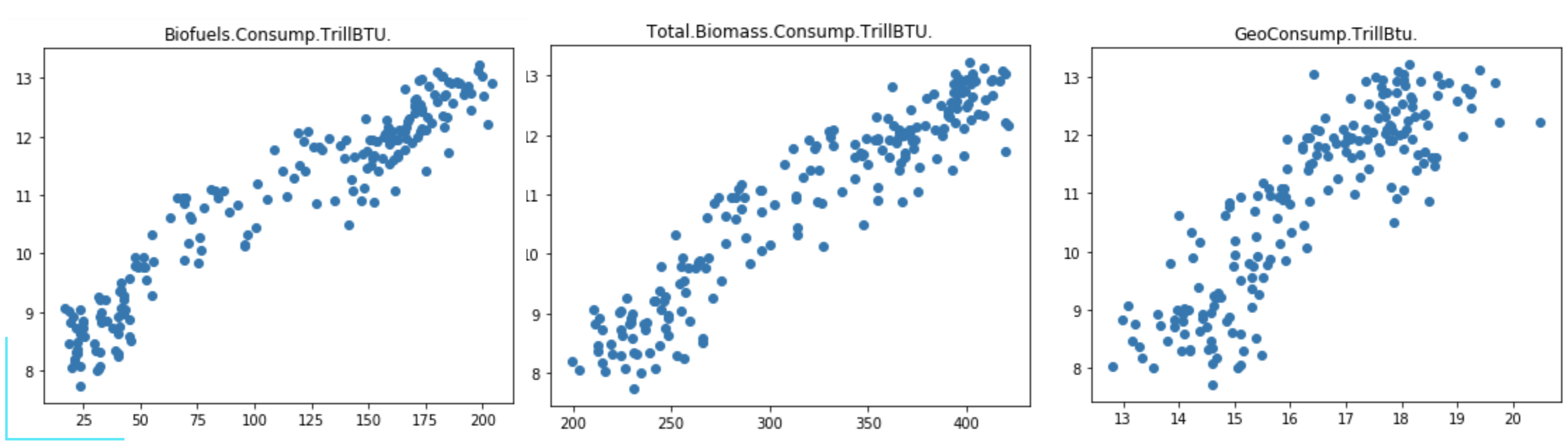
BASELINE REGRESSION

Quantifying each feature's effect on price:

- Train on historical data set
- Fit each feature with a significance coefficient
- Result: linear combination of features & weights
- Validate accuracy on separate historical set
- Good accuracy for linear features

$$y(t) = \varphi^T(t)\theta \qquad \hat{\theta}(N) = \arg\min_{\theta} \sum_{t=1}^{N} \|y(t) - \varphi^T(t)\theta\|^2$$

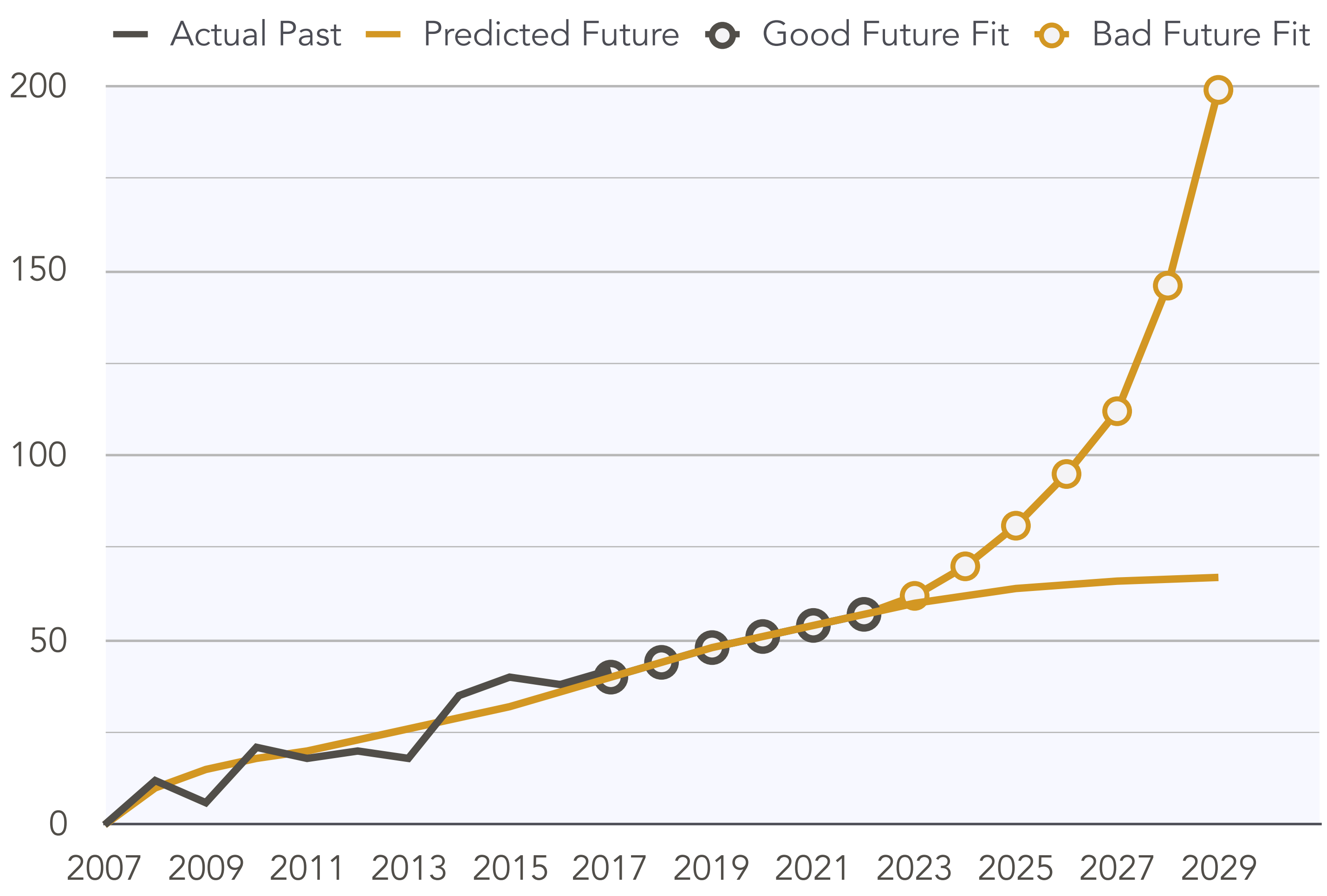
```
Best Linear Predictors:
 ['Biofuels.Consump.TrillBTU.', 'Total.Biomass.Consump.TrillBTU.', 'GeoConsump.TrillBtu.']

  R-Squared:               Coefficients:                            Intercept:
   0.41662115151058926      [ 0.02414756   0.00441986 -0.19547787]    9.955992839935849
```



MODELING DISRUPTIONS

WHY DOES THE BASELINE DEVIATE?

- *Changing feature influence (coal vs. solar)*
- *Unpredictable market shocks (drought)*
- *Introduction of new technology (EVs)*



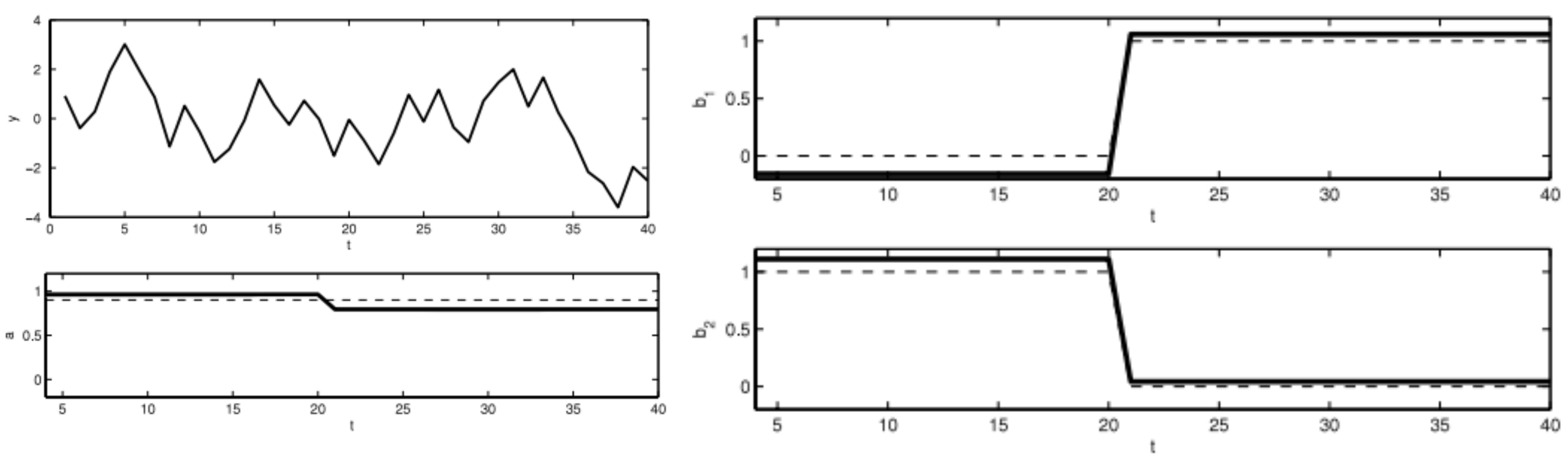
TIME-VARIANT SEGMENTATION

$$\min_{\theta(t)} \sum_{t=1}^{N} \|y(t) - \varphi^T(t)\theta(t)\|^2 + \lambda \sum_{t=2}^{N} \|\theta(t) - \theta(t-1)\|_{\text{reg}},$$

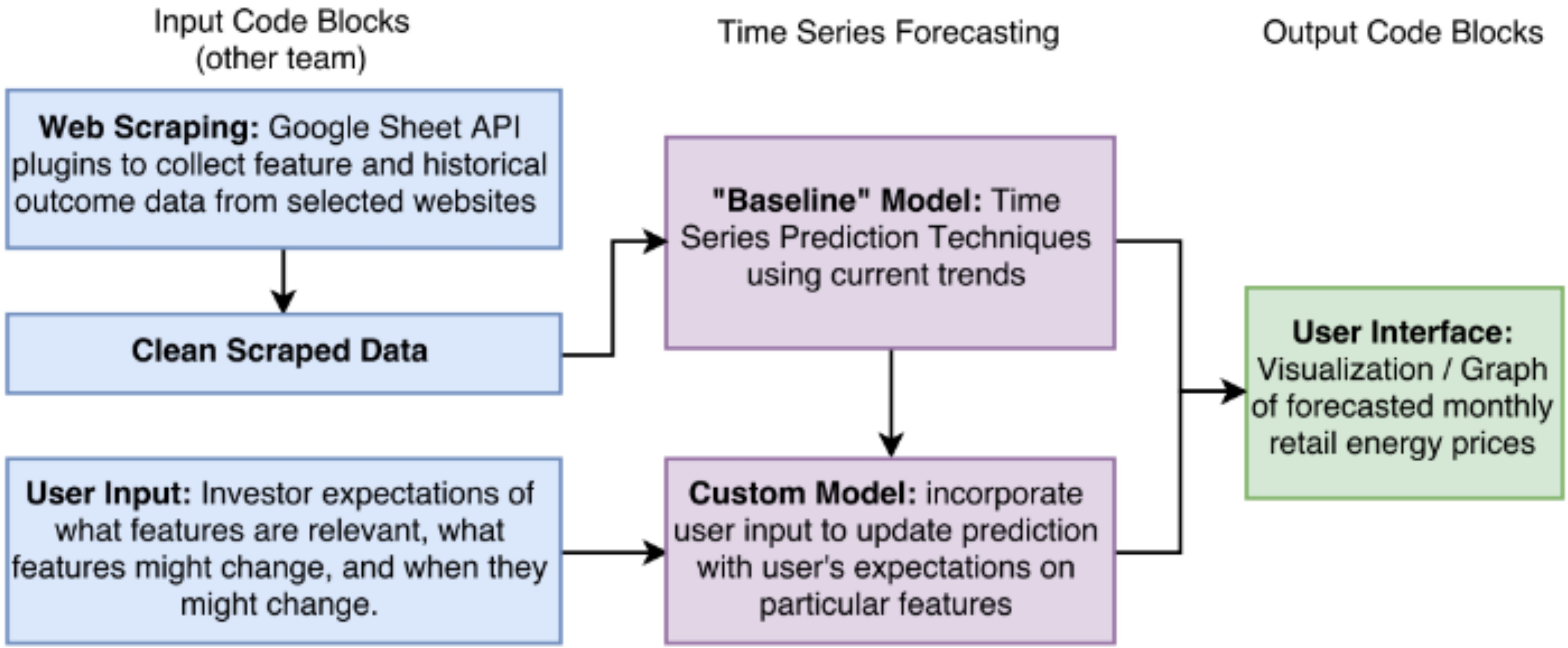
- Feature influence changes
- $\Delta\theta$ is penalized to avoid overfit
- Tune λ for desired segmentation

HOW IS THIS RESULT USEFUL?

- Capture the effect of market disruptors
- Reflect a realistic, changing economy
- Simulate a range of future outcomes



ARCHITECTURE



USER INTERFACE

Baseline Features

	Feature	Pearson
28	Biofuels.Consump.TrillBTU.	0.923888
29	Total.Biomass.Consump.TrillBTU.	0.918654
19	Bio.Prod.Trillion.BTU.	0.914748
20	BioMass.Prod.Trill.BTU.	0.910235
59	Natural.Gas.Consumed.by.the.Transportation.Sec...	0.859504
14	GDP	0.849559
23	GeoConsump.TrillBtu.	0.843676
10	GenCalifornia...other.thousand.megawatthours	0.826934
30	Total.Renewable.Consump.TrillBTU.	0.816141
21	Total.RenProd..TrillBtu.	0.812076
27	WasteConsump.TrillBTU.	0.803539

ADD

REMOVE

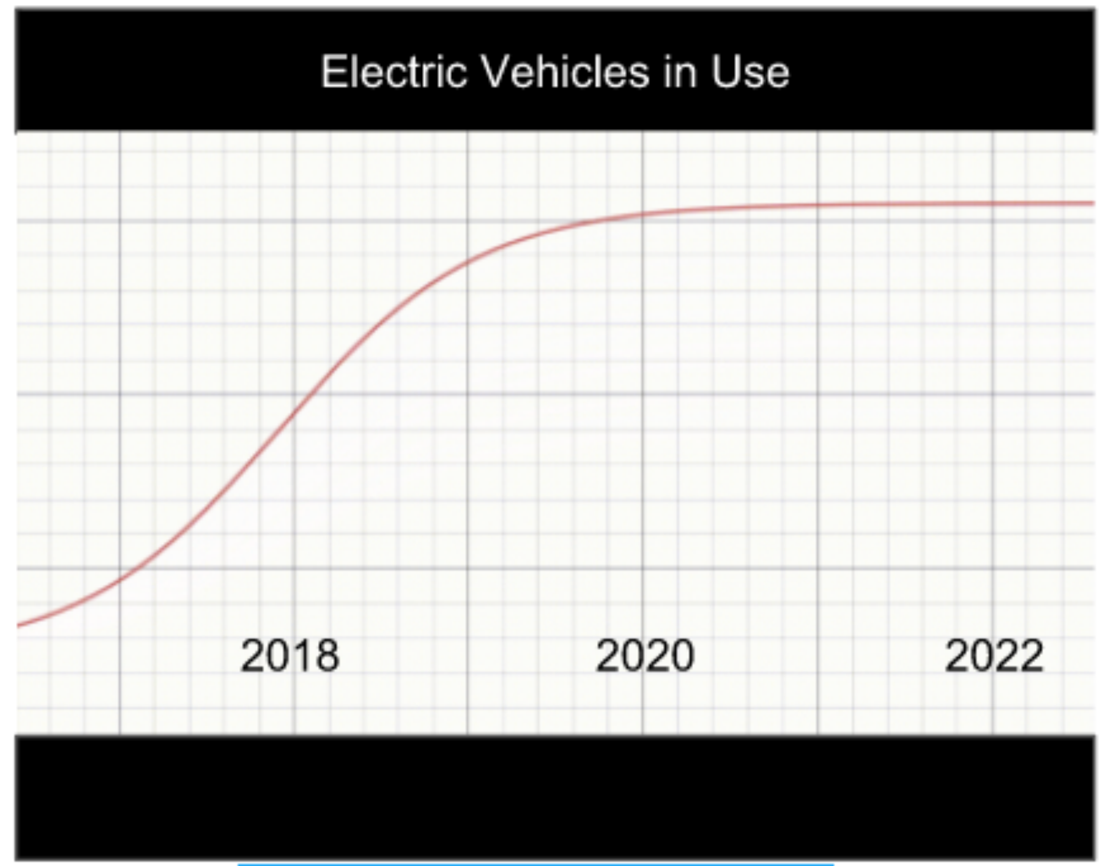
TRAIN MODEL

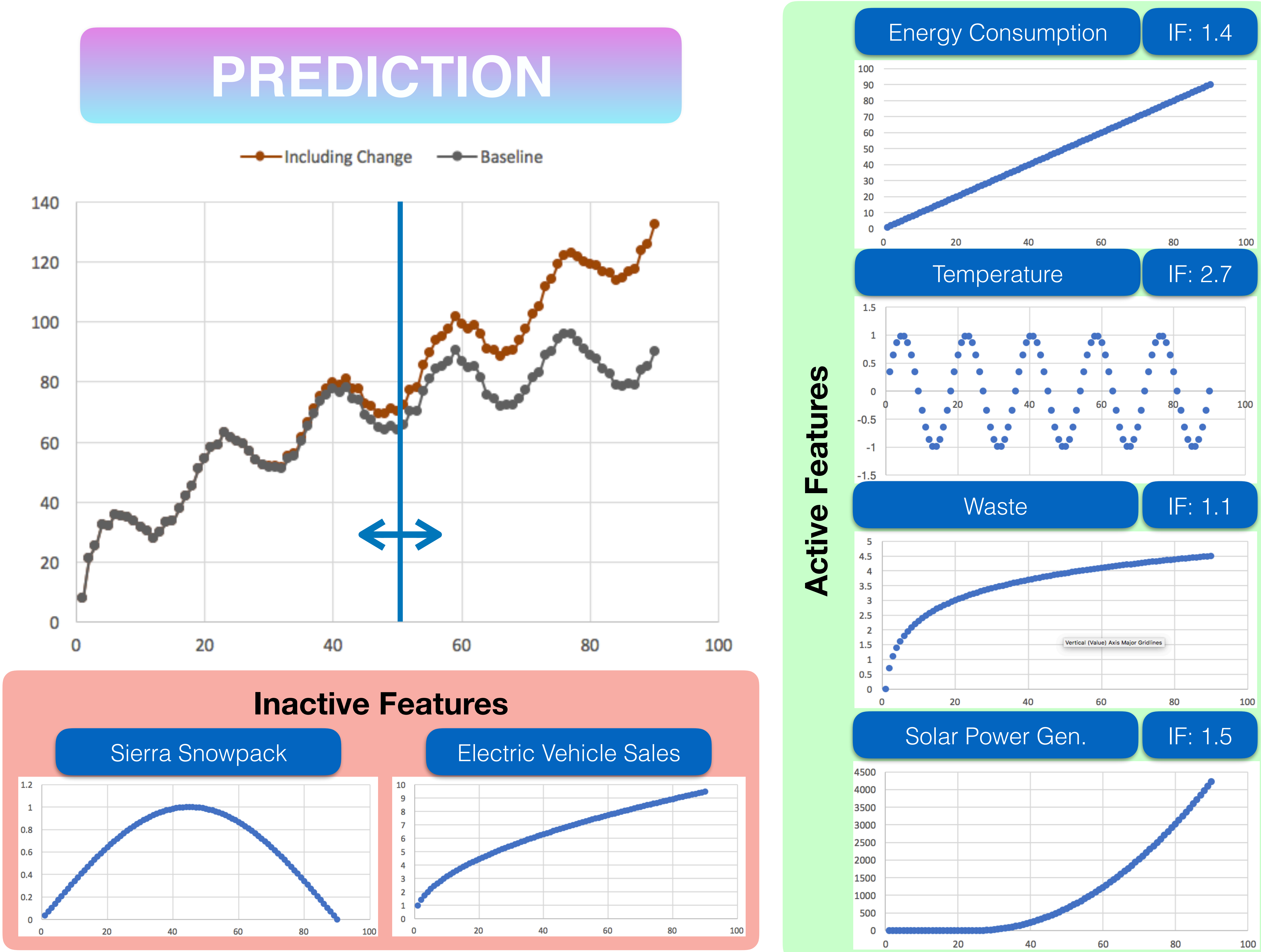
Include Dynamic Impact: Please select...

Electric Vehicles

Sierra Nevada Snowpack

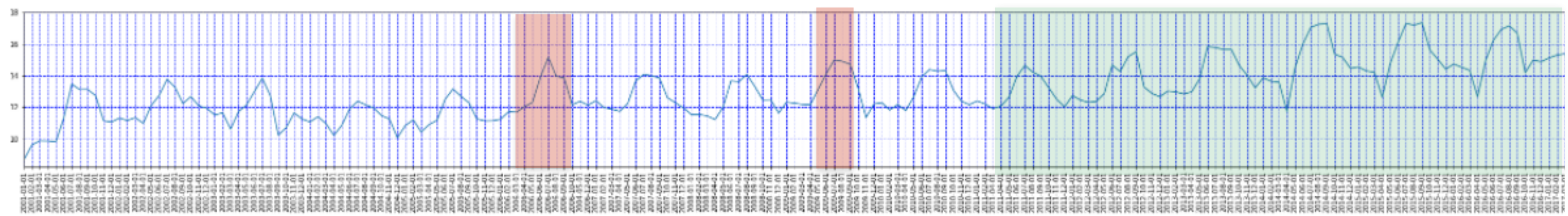
Solar Production



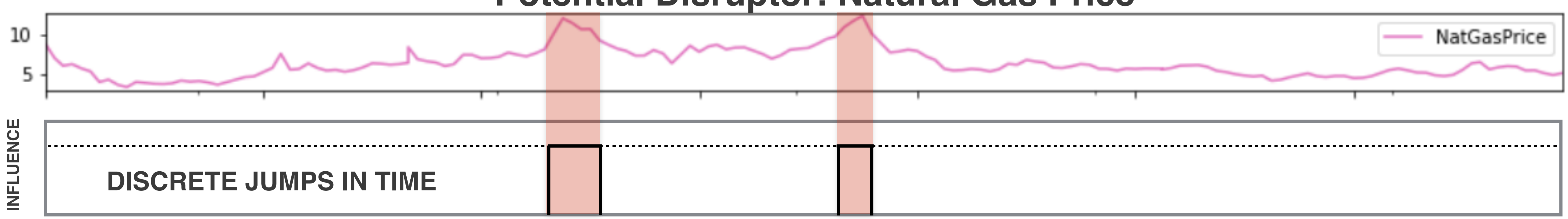

PREDICTION
Including Change
Baseline
0 20 40 60 80 100 120 140
0 20 40 60 80 100
Inactive Features
Sierra Snowpack
Electric Vehicle Sales
Active Features
Energy Consumption
IF: 1.4
Temperature
IF: 2.7
Waste
IF: 1.1
Vertical (Value) Axis Major Gridlines
Solar Power Gen.
IF: 1.5

FEATURE INFLUENCE

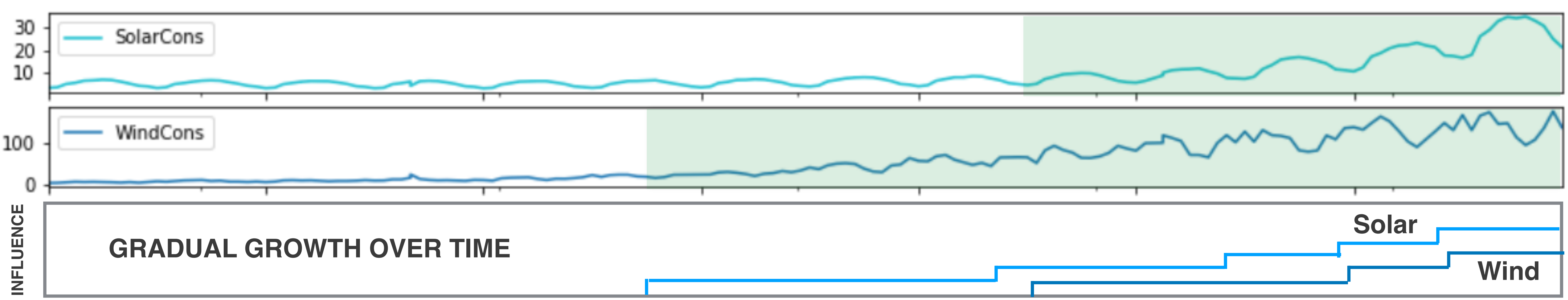
OUTPUT: Retail Electricity Price



Potential Disruptor: Natural Gas Price



Potential Disruptor: Renewable Energy



LEARNING PATH

JANUARY	FEBRUARY	MARCH	APRIL

Phase 1 — UNDERSTAND THE PROBLEM

- **1.1** Meet with senior advisors, read academic papers, identify key features
- **1.2** Understand the limits of predictive modeling

Phase 2 — DEVELOP TIMESERIES

- **2.1** Implement various feature selection methods, check for stationarity
- **2.2** Build timeseries + linear regression notebooks

Phase 3 — ADD FLEXIBILITY, TIME VARIANCE

- **3.1** Incorporate the Stanford Paper into time-varying weight adjustment
- **3.2** Recruit feedback

FUTURE STEPS

1. ADD NONLINEAR REGRESSION TO TIMESERIES POST-PROCESSING
2. FEATURE ENGINEERING: TIME DELAYED VARIABLES, FUTURE DATA
3. VALIDATE ON MORE TEST DATA, GENERATE UI FOR INTERACTIVE USE
4. OUTCOME VARIABLE: CONSIDER SUPPLY SIDE ENERGY COST OVER RETAIL PRICE (DEMAND SIDE)

OUR TEAM

Ahmed Issaouai
Master of
Industrial Eng.

Negi Fazeli
Haas School
of Business

Patrick Lerchi
Civil & Environ.
Engineering

Marie Parent
Master of
Public Policy

Shaojin Wei
Master of
Industrial Eng.

SENIOR ADVISORS:

Shikhar Verma
Master of Engineering
Project Founder

Sina Khoshfetrat
Senior Data Scientist
C3 IoT

INDUSTRY MENTORS:

Antonio Vitti
Merchant Atlas CFO
Haas Alumni

Steven Gustafson
MAANA Data Scientist
Head of GE Discovery Lab

THANKS FOR LISTENING!

Your feedback is greatly appreciated

Code can be found <u>HERE</u>

Ahmed Issaoui
Negi Fazeli
Marie Parent
Shaojin Wei
Patrick Lerchi