Automated Time Series Analysis

Dr. Clifton Phua
Senior Director
clifton@datarobot.com
https://www.linkedin.com/in/cliftonphua/

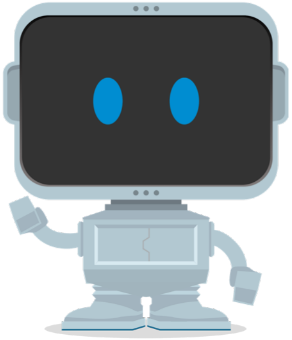
Agenda

- ✓ Introduction
- ✓ Automated time series analysis
  - ✓ Feature engineering
  - ✓ Target transforms
  - ✓ Model backtesting
  - ✓ Time series modeling
- ✓ Sales forecasting use case and demo
- ✓ Advanced topics in time series analysis

INTRODUCTION

About Me

- Customer-Facing Data scientist @ DataRobot
- ~12 years of real-world business experience in machine learning
- In various industries, on many use cases



Previously…

A*STAR Institute for Infocomm Research

sas

ncs making IT happen

DataRobot

BANKING **INSURANCE** **TELCO** **RETAIL** **AND MORE**

2012
Founded
HQ in Boston, MA

250+
Data scientists &
Engineers (of 500+)

$225M+
In funding

800,000,000+
Models built on
DataRobot cloud

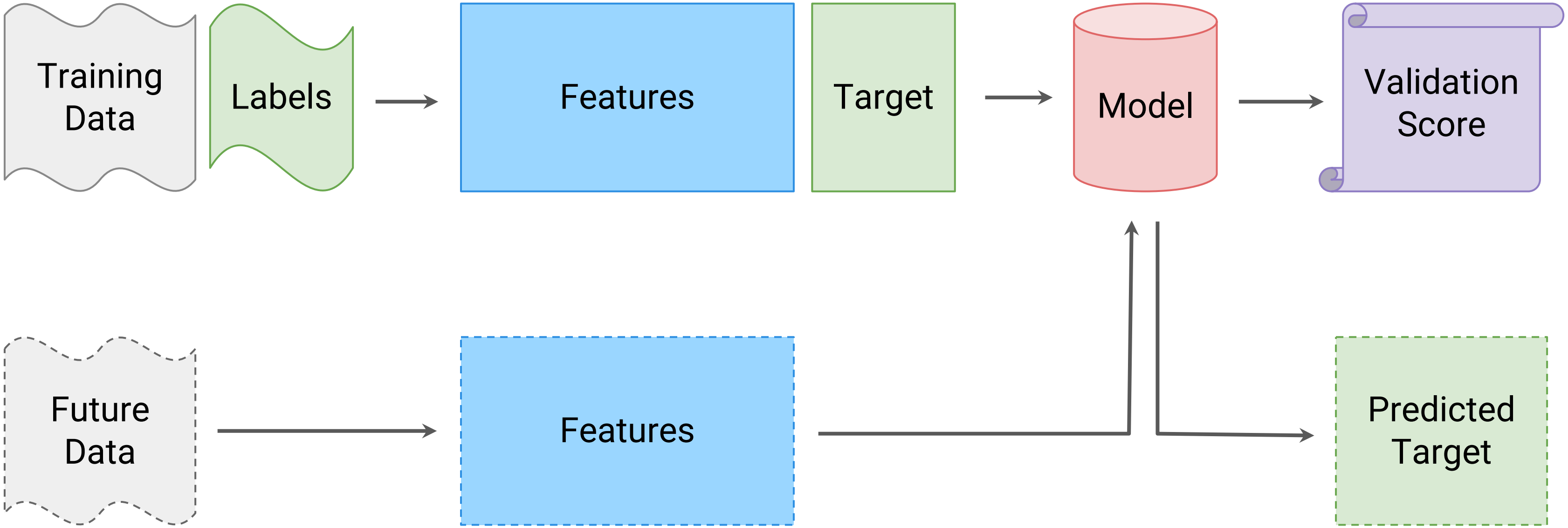
4
kaggle #1 ranked
Data scientists

50+
kaggle Top 3 finishes

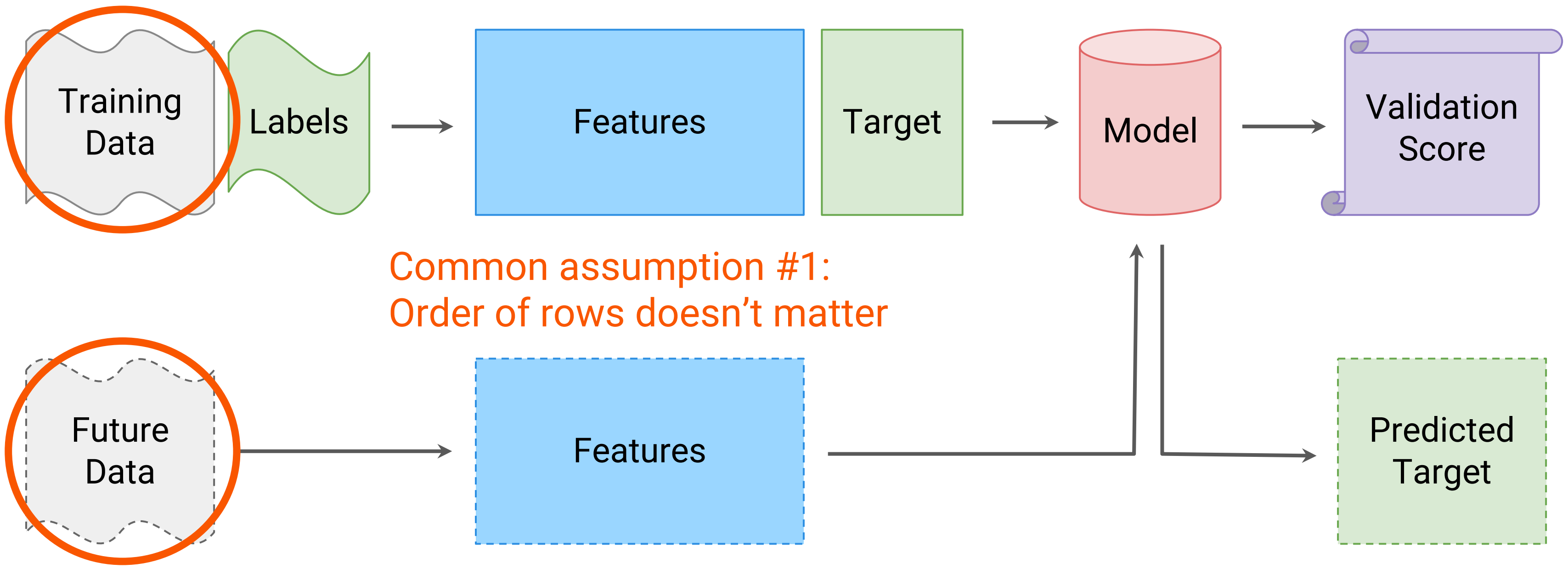
H2O

AUTOMATED TIME SERIES ANALYSIS

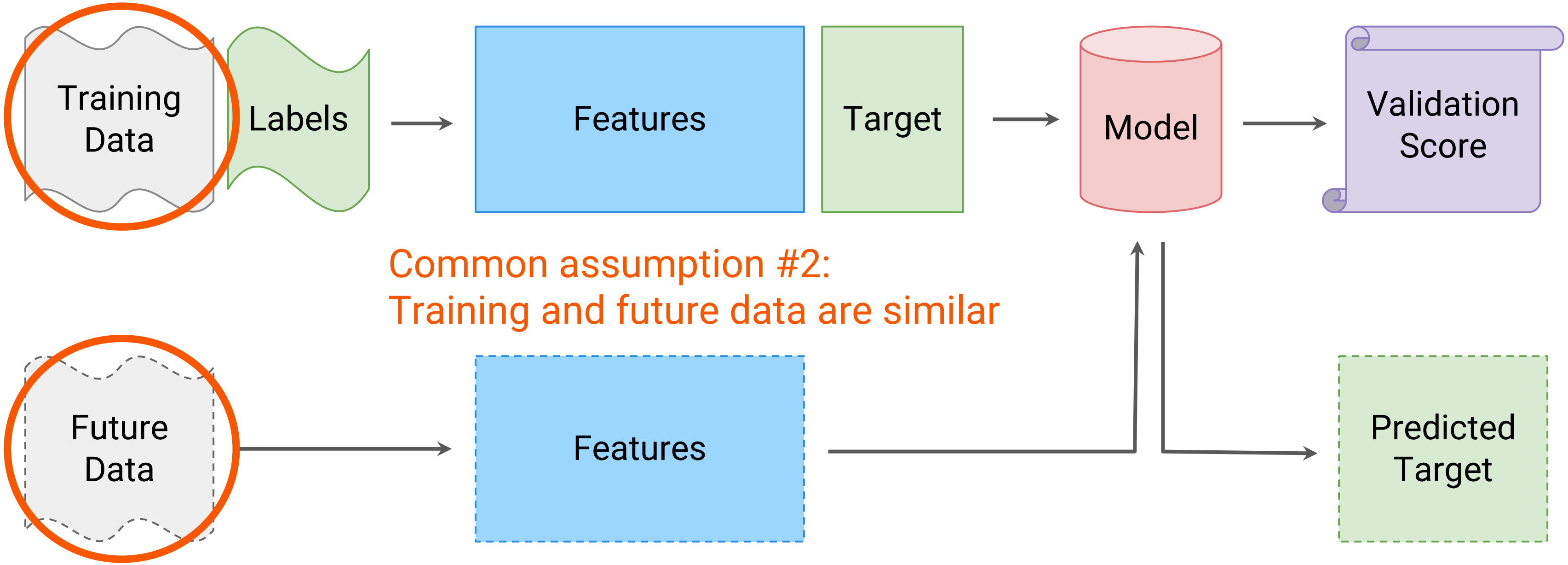
Typical Machine Learning Flow



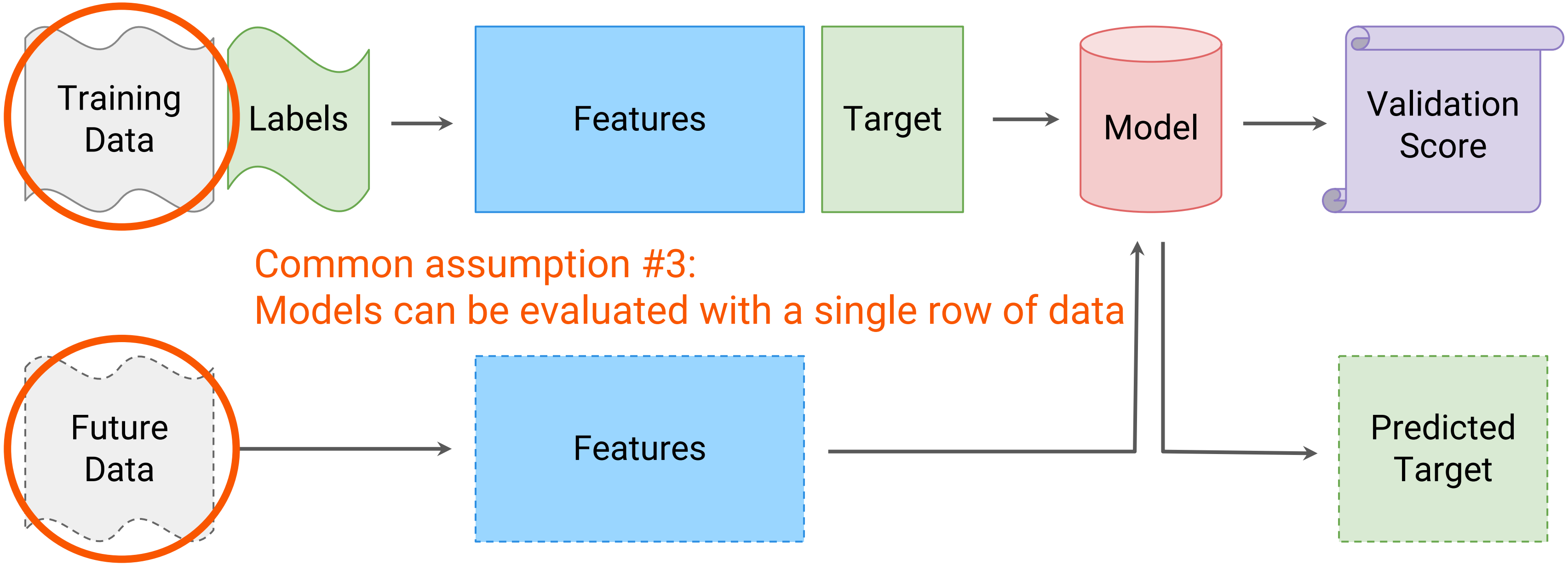
Typical Machine Learning Flow



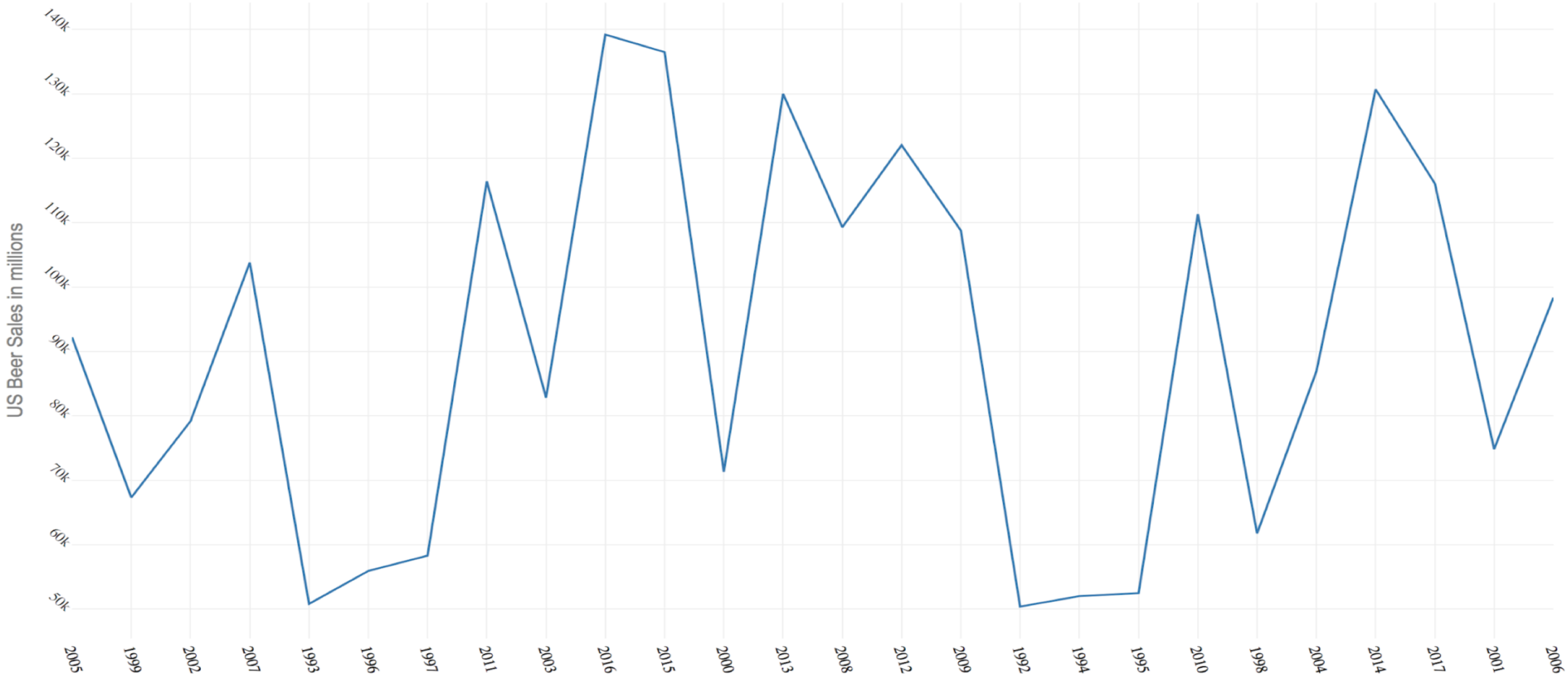
Typical Machine Learning Flow



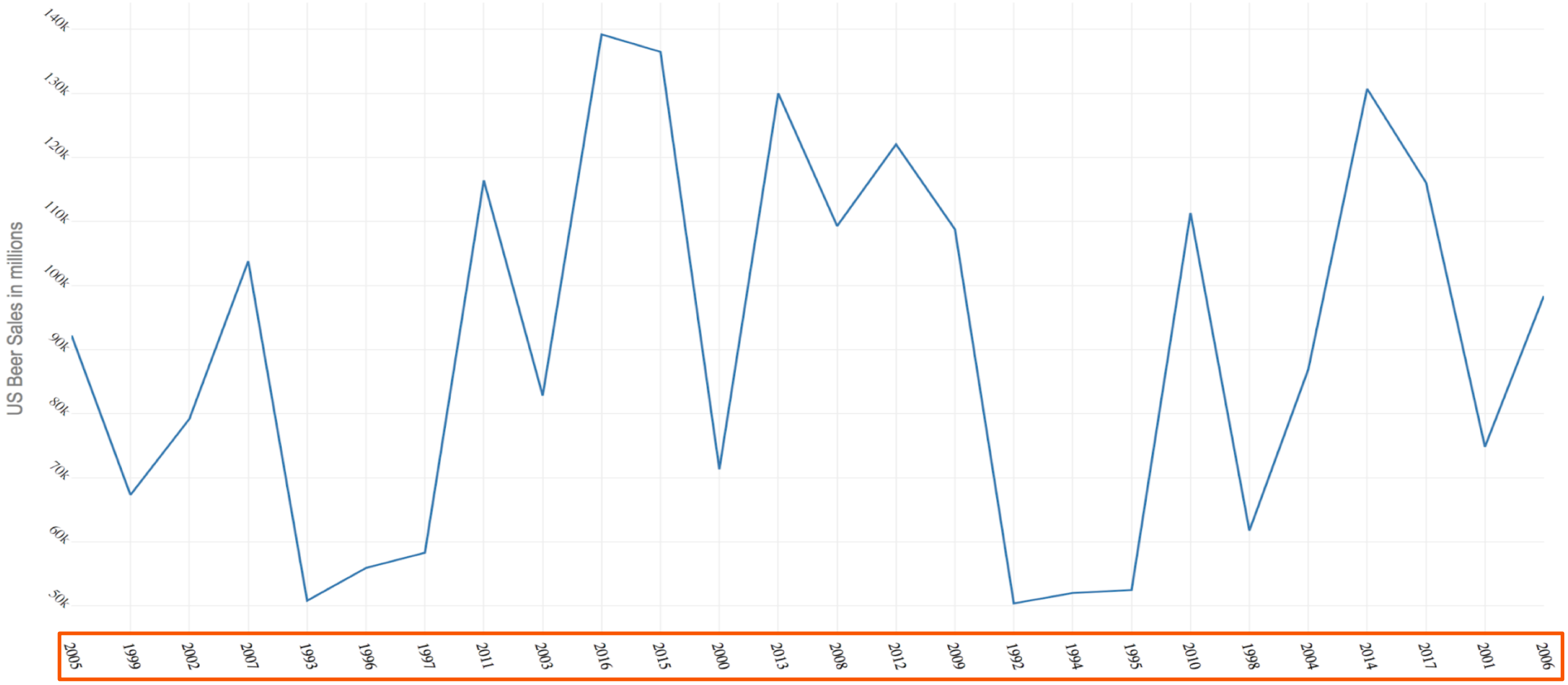
Typical Machine Learning Flow



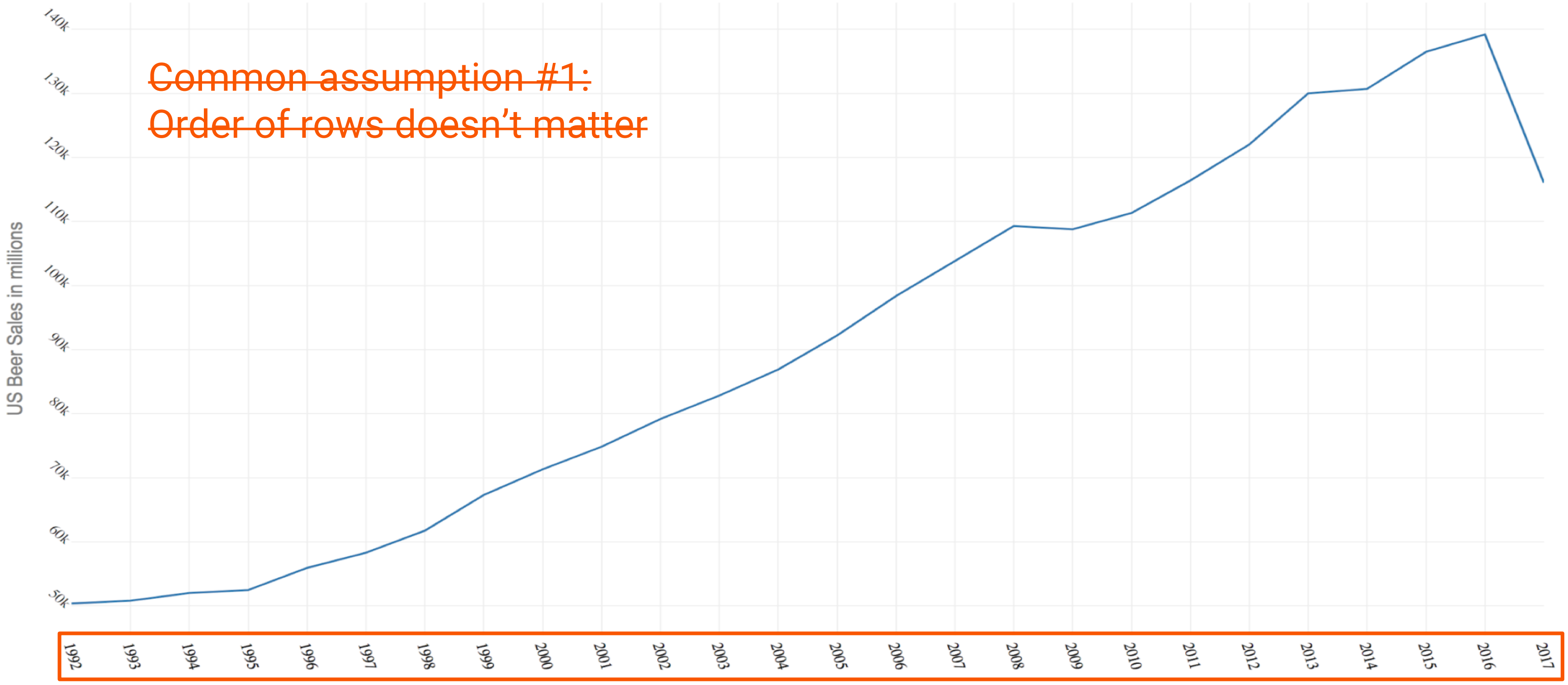
U.S. Beer Sales



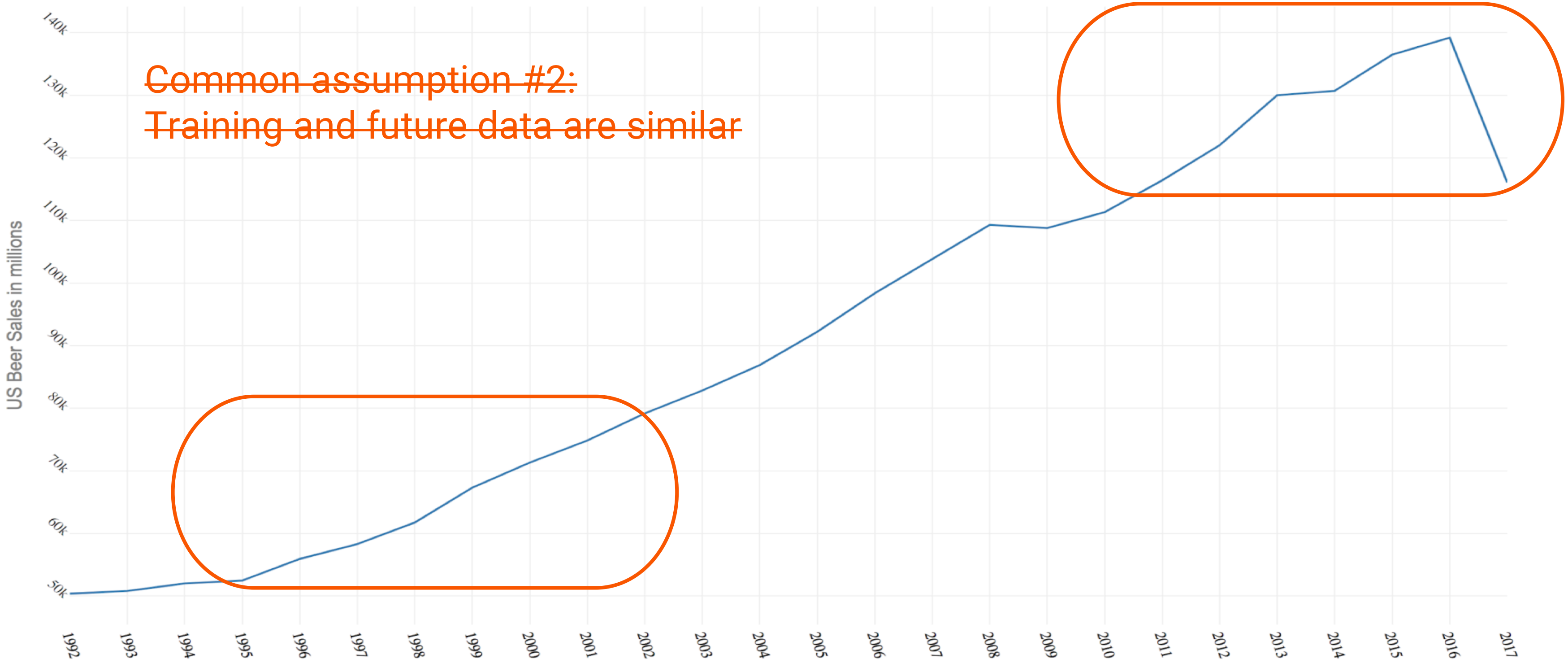
U.S. Beer Sales



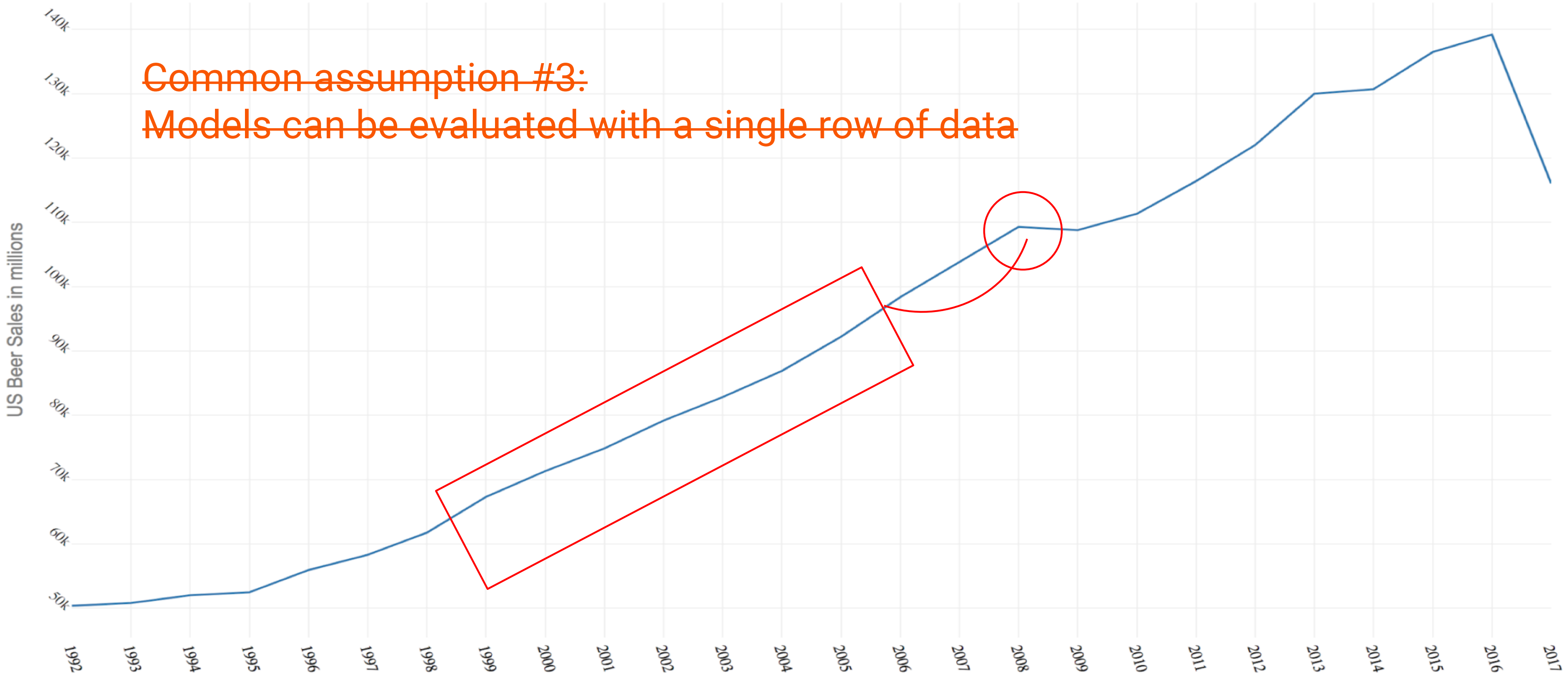
U.S. Beer Sales



U.S. Beer Sales



U.S. Beer Sales



Forecasting

	Date	Sales	Employees Present	Ad Spending	Inventory Rate	Weather
	11/09/17	$432,897	16	$3,000	47%	Rain
	11/10/17	$474,306	19	$2,100	46%	Heavy Wind
	11/11/17	$415,434	17	$1,200	41%	Cloudy
	11/12/17	$289,290	20	$1,800	31%	Cloudy
	11/13/17	$355,786	15	$1,300	36%	Cloudy
today	11/14/17	$375,284	13	$600	39%	Sunny

Forecasting

	Date	Sales	Employees Present	Ad Spending	Inventory Rate	Weather
	11/09/17	$432,897	16	$3,000	47%	Rain
	11/10/17	$474,306	19	$2,100	46%	Heavy Wind
	11/11/17	$415,434	17	$1,200	41%	Cloudy
	11/12/17	$289,290	20	$1,800	31%	Cloudy
	11/13/17	$355,786	15	$1,300	36%	Cloudy
today	11/14/17	$375,284	13	$600	39%	Sunny
	...	...	...	...	...	...
tomorrow	11/15/17	?	10	$1100	43%	Sunny

Forecasting

	Date	Sales	Employees Present	Ad Spending	Inventory Rate	Weather	
	11/09/17	$432,897	16	$3,000	47%	Rain	
	11/10/17	$474,306	19	$2,100	46%	Heavy Wind	
	11/11/17	$415,434	17	$1,200	41%	Cloudy	
	11/12/17	$289,290	20	$1,800	31%	Cloudy	
	11/13/17	$355,786	15	$1,300	36%	Cloudy	
today	11/14/17	$375,284	13	$600	39%	Sunny	
	...	...	...	...	...	...	
tomorrow	11/15/17	?	10	$1100	43%	Sunny	leakage!

We Can Introduce Lags

	Date	Sales	Employees Present	Ad Spending	Inventory Rate	Weather
	11/09/17	$432,897				
	11/10/17	$474,306	16	$3,000	47%	Rain
	11/11/17	$415,434	19	$2,100	46%	Heavy Wind
	11/12/17	$289,290	17	$1,200	41%	Cloudy
	11/13/17	$355,786	20	$1,800	31%	Cloudy
today	11/14/17	$375,284	15	$1,300	36%	Cloudy
	...	...	...	...	...	...
tomorrow	11/15/17	?	13	$600	39%	Sunny

Lag = 2

	Date	Sales	Employees Present	Ad Spending	Inventory Rate	Weather
	11/09/17	$432,897				
	11/10/17	$474,306				
	11/11/17	$415,434	16	$3,000	47%	Rain
	11/12/17	$289,290	19	$2,100	46%	Heavy Wind
	11/13/17	$355,786	17	$1,200	41%	Cloudy
today	11/14/17	$375,284	20	$1,800	31%	Cloudy
	...	...	...	...	...	...
tomorrow	11/15/17	?	15	$1,300	36%	Cloudy

Lag = 3

	Date	Sales	Employees Present	Ad Spending	Inventory Rate	Weather
	11/09/17	$432,897				
	11/10/17	$474,306				
	11/11/17	$415,434				
	11/12/17	$289,290	16	$3,000	47%	Rain
	11/13/17	$355,786	19	$2,100	46%	Heavy Wind
today	11/14/17	$375,284	17	$1,200	41%	Cloudy
	...	...	...	...	...	...
tomorrow	11/15/17	?	20	$1,800	31%	Cloudy

We Can Also Lag the Target

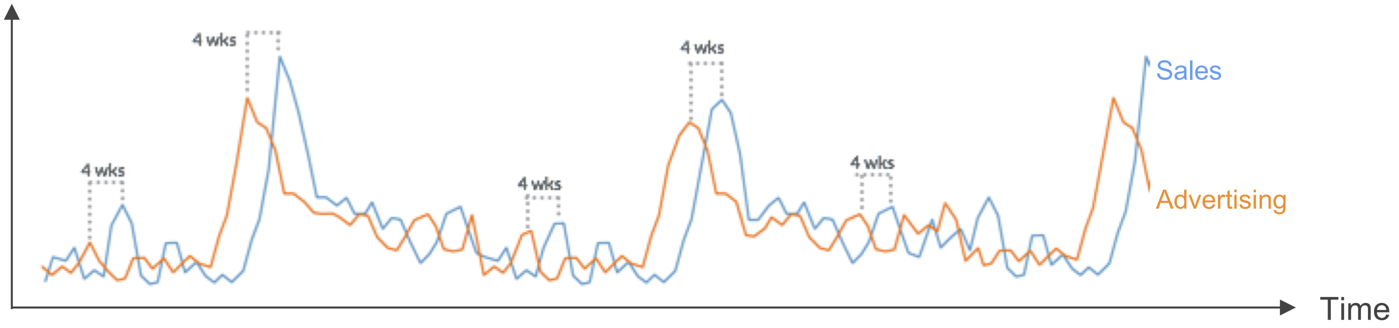
	Date	Sales	Sales 3 days ago	Employees Present	Ad Spending	Inventory Rate	Weather
	11/09/17	$432,897					
	11/10/17	$474,306					
	11/11/17	$415,434					
	11/12/17	$289,290	$432,897	16	$3,000	47%	Rain
	11/13/17	$355,786	$474,306	19	$2,100	46%	Heavy Wind
today	11/14/17	$375,284	$415,434	17	$1,200	41%	Cloudy
	...	...		...	...	...	...
tomorrow	11/15/17	?	$289,290	20	$1,800	31%	Cloudy

We Can Also Mix Lags

	Date	Sales	Sales 3 days ago	Employees Present	Ad Spending	Inventory Rate	Weather
	11/09/17	$432,897					Rain
	11/10/17	$474,306			$3,000		Heavy Wind
	11/11/17	$415,434		16	$2,100		Cloudy
	11/12/17	$289,290	$432,897	19	$1,200	47%	Cloudy
	11/13/17	$355,786	$474,306	17	$1,800	46%	Cloudy
today	11/14/17	$375,284	$415,434	20	$1,300	41%	Sunny
	...	...		...	...	...	...
tomorrow	11/15/17	?	$289,290	15	$600	31%	Cloudy

Why Does Lagging Work?

Because real world data has delays



Rolling Statistics (Numeric)

Date	Sales	Employees Present	Employees Present (7 day mean)	Employees Present (14 day mean)	Employees Present (21 day mean)	Employees Present (28 day mean)
....	....	....	....	....	....	....
12/09/17	$430,327	16	16.66	17.19	17.51	17.05
12/10/17	$572,309	19	17.33	15.75	18.39	18.40
12/11/17	$399,494	17	16.81	17.08	18.02	17.88
12/12/17	$250,290	20	17.74	17.98	18.76	18.44
12/13/17	$389,786	15	16.18	14.42	17.01	17.12
12/14/17	$366,284	13	15.60	17.35	16.33	16.92

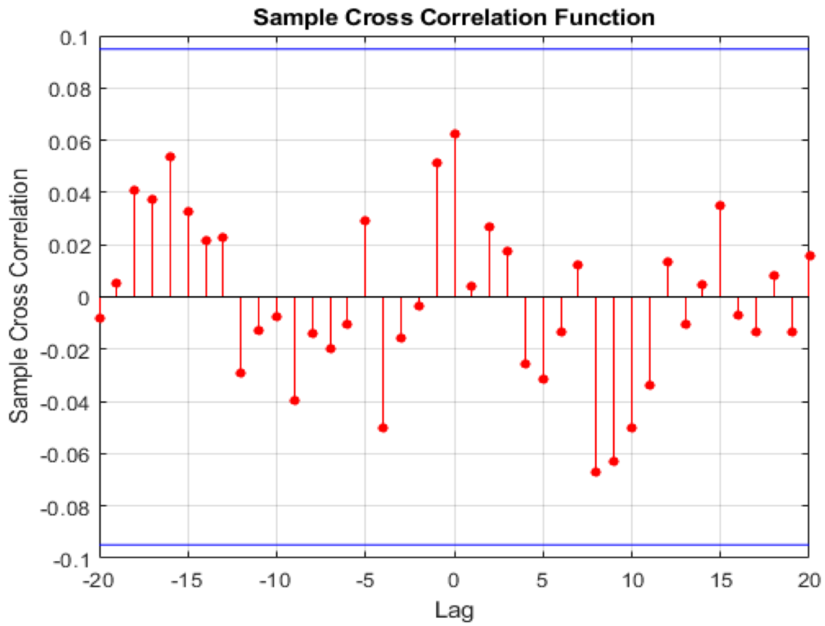
Rolling Statistics (Categorical)

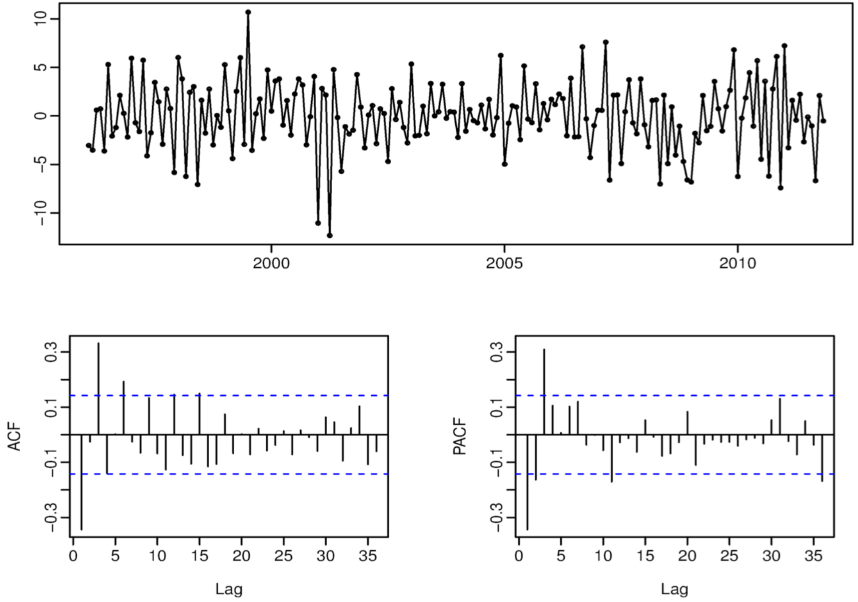
Date	Sales	Weather	Weather (7 day n_unique)	Weather (14 day n_unique)	Weather (21 day n_unique)	Weather (28 day n_unique)
....	....	....	....	....	....	....
12/09/17	$430,327	Rain	2	2	4	5
12/10/17	$572,309	Heavy Wind	2	3	7	8
12/11/17	$399,494	Cloudy	3	4	6	9
12/12/17	$250,290	Cloudy	1	4	5	6
12/13/17	$389,786	Cloudy	1	2	6	7
12/14/17	$366,284	Sunny	4	5	8	10

Rolling Statistics (Text)

Date	Sales	Product Offers	Product Offers Length (2nd lag)	Product Offers Length (7 day mean)	Product Offers Length (14 day std)
….	….	….	….	….	….
12/09/17	$430,327	Home, Kitchen	14	15.44	4.49
12/10/17	$572,309	Diary, Grocery, Personal Care	26	16.50	7.18
12/11/17	$399,494	Frozen Food, Magazines	13	19.36	6.19
12/12/17	$250,290	Beverages, Seafood	29	17.28	3.22
12/13/17	$389,786	Wine, Home Appliances, Cleaning Supplies	22	23.33	5.20
12/14/17	$366,284	Home, Wine, Frozen Food, Candy	18	20.85	2.12

Which Features Should I Use?





Wait, is it just automated feature engineering?

Automated Feature Engineering

- Automated lag selection
- Automated window statistics
- Rolling numeric, categorical, text information

Automated Target Transforms

- Stationarity, exponential, periodicity detection
- Automated differencing, offsets, log transformations
- Additive or multiplicative models

Automated Model Backtesting

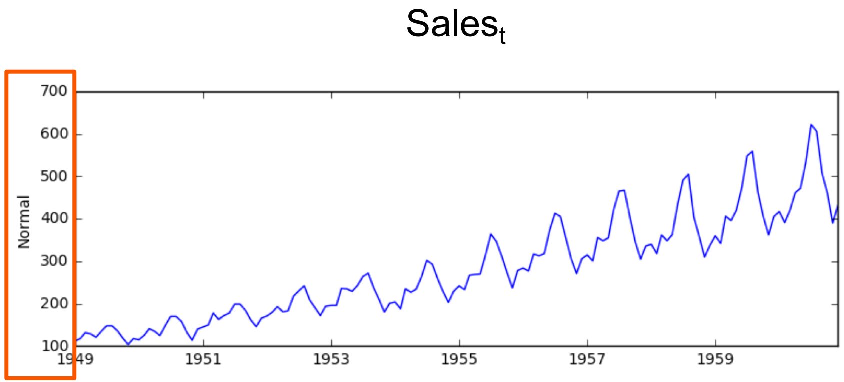
- Time-aware data partitioning and validation
- Automated or configurable backtesting strategies
- Refit on most recent data

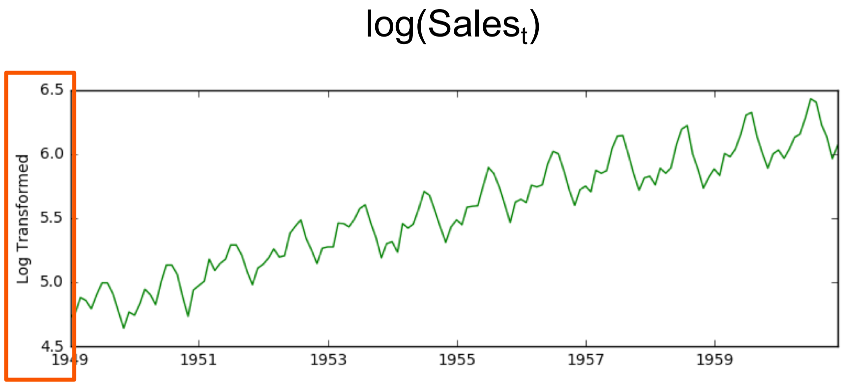
Automated Time Series Modeling

- Classical time series models (ARIMA, ETS, etc.)
- Time-aware xgboost, distance modeling, etc.
- Deploy to dedicated prediction service

Automated Target Transformation

Log Transform

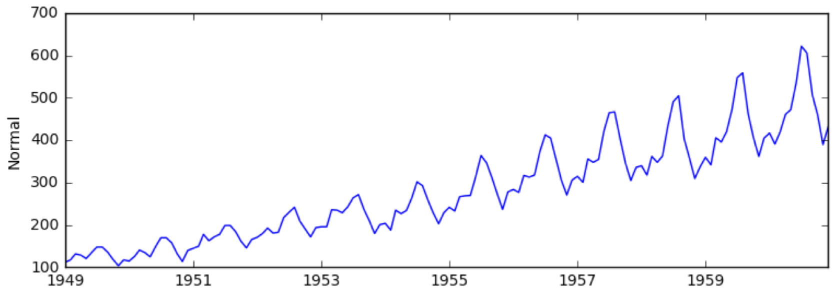




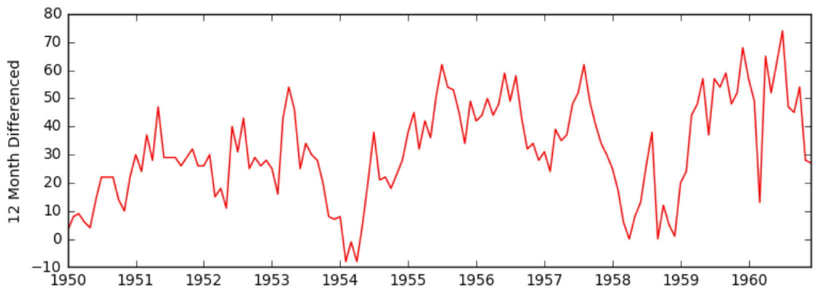
Automated Target Transformation

Differencing

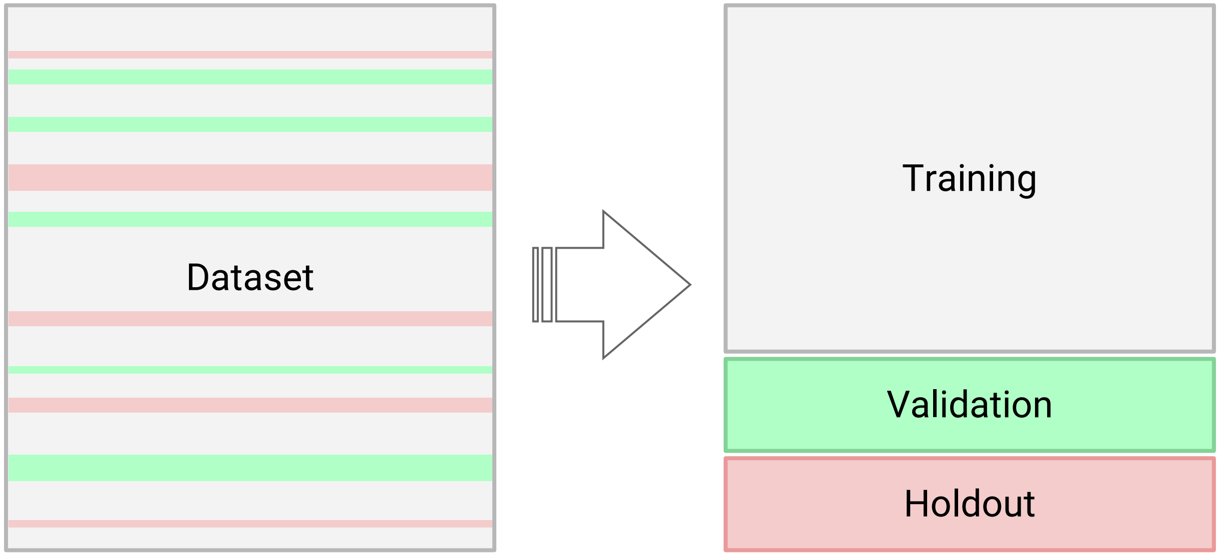
$Sales_t$



$Sales_t - Sales_{t-1}$



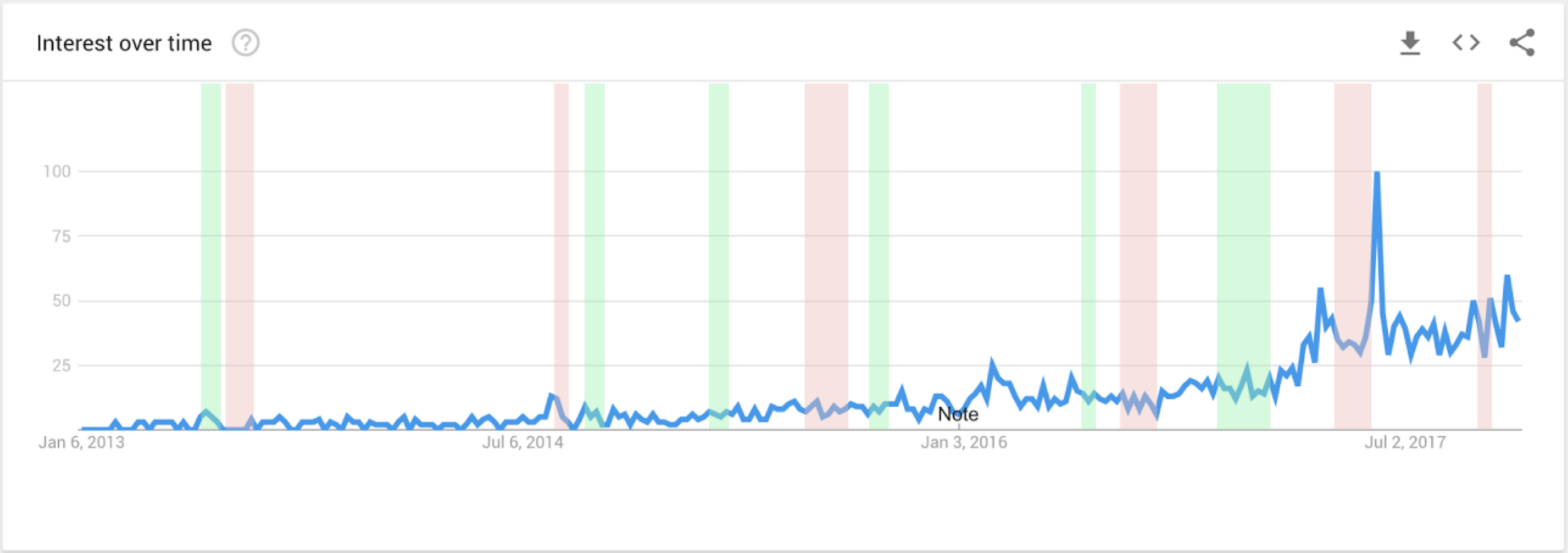
Automated Model Backtesting



Conventional machine learning approach?

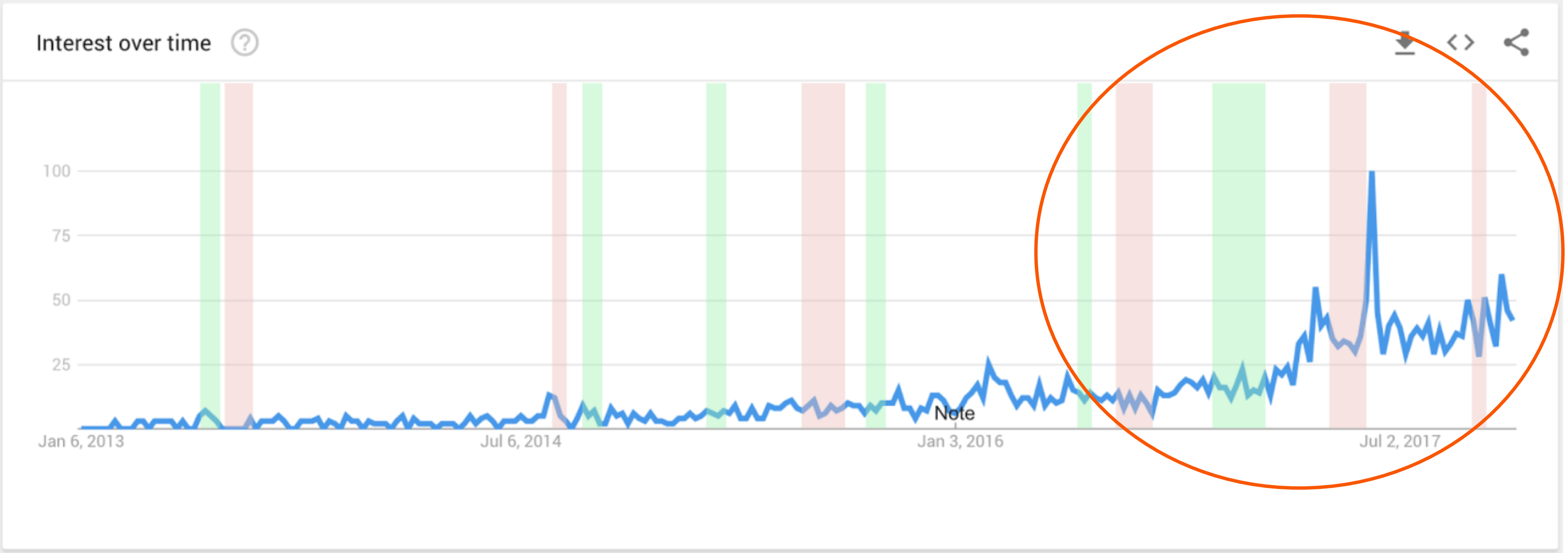
Automated Model Backtesting

Problem!

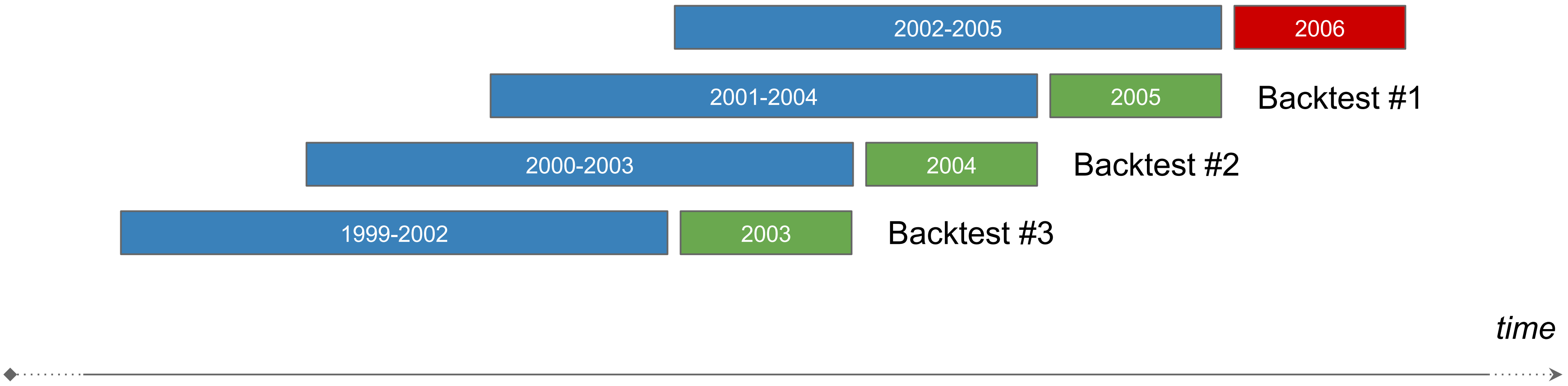


Automated Model Backtesting

Using future data to predict the past

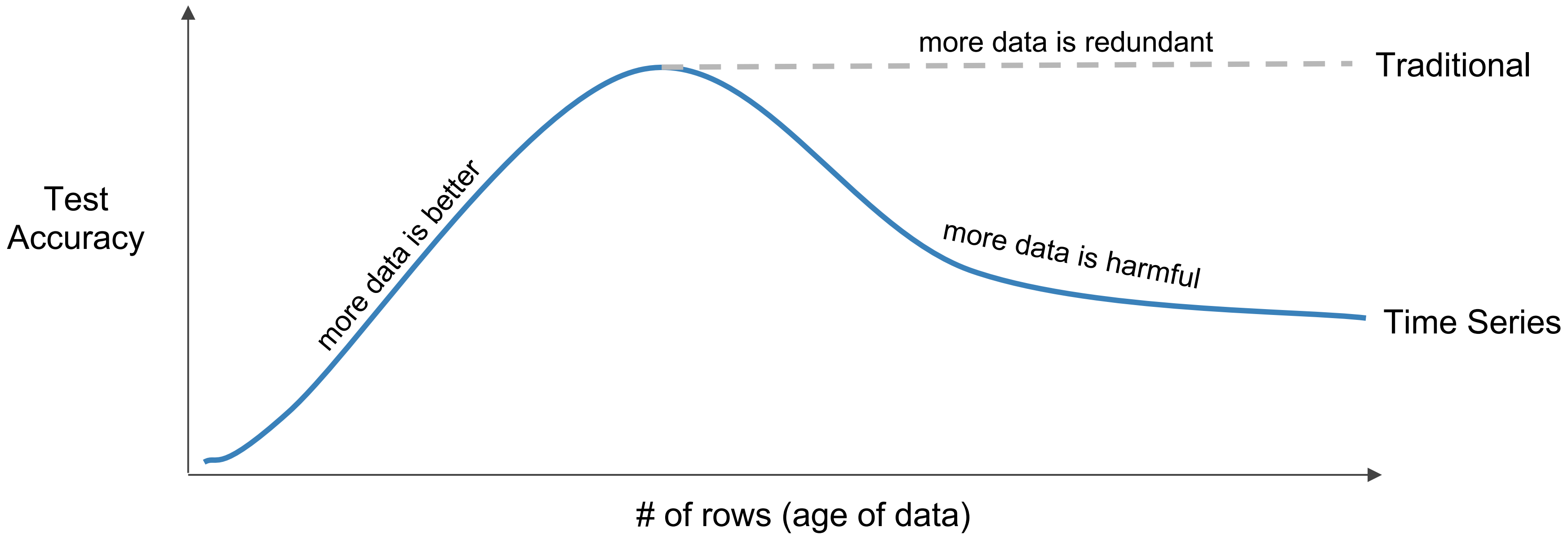


Automated Model Backtesting



Backtesting score = avg(validation score #1 + validation score #2 + validation score #3)

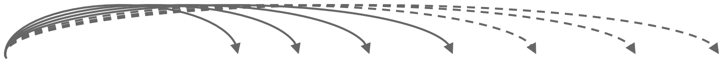
Time Series Learning Curves



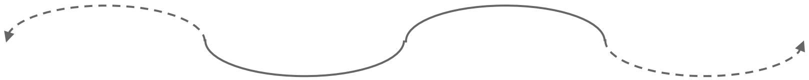
Automated Time Series Models

Integrated Models

Forecast Distance Models



Trends and Decomposition Models



Some Time Series Use Cases

Industry	Use Case(s)
Aviation	Predict frequent flyer points balance, demand for flights
Energy	Predicting electricity demand
Entertainment	Predict visitor number
Finance / Oil&Gas	Predict commodity values
Healthcare	Predict the outbreak of a disease (e.g., Zika), patient health trajectory (5 years out), emergency / patient visits
Insurance	Predicting the number of contracts and claims over time
Investments	Predicting unemployment rates, market indices, stock or market volatility
Manufacturing	Predicting production output, sensor values outside limits, machine failure
Marketing	Predicting Google AdWords bid prices, marketing attribution
Restaurants	Sales prediction per restaurant
Retail	**Sales prediction per product/store**
Telco	Predict cell (or service) usage, capacity planning
Utility	Demand prediction (gas/electric)

Use Case:
Sales Forecasting

Project Statement

Our 10 retail stores do around $78k in sales per day on average. Average daily sales can be as high as $273k during Black Friday or a low as $0 when the store is closed. You would like to be able to accurately forecast sales over the next week as well as understand what factors impact sales and to what degree. If successful, this information will used by executives as a baseline for judging store performance and evaluating overall business operations.

DEMO

Advanced Topics

What level of aggregation should I model?

What is the business question?

Should we predict the mean or the total?

Granularity

- Minutely
- Hourly
- Daily
- Weekly
- Monthly
- Quarterly

Attributes:

- Less stable target
- More data to train models
- Higher likelihood time series models will capture interesting dynamics
- Potentially too many zeros to model data well

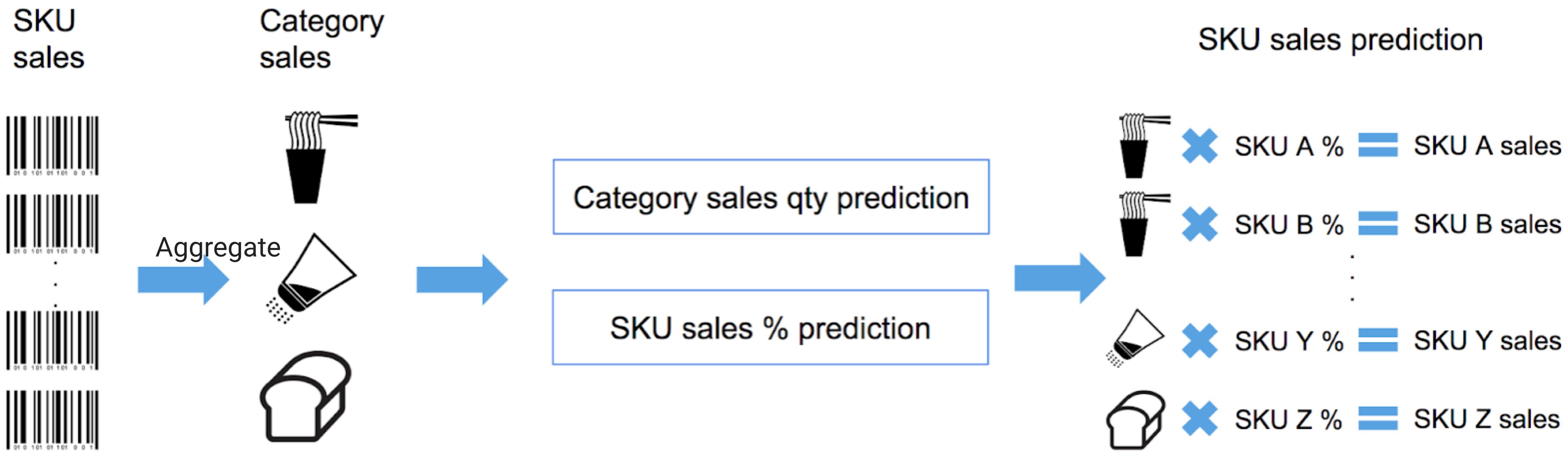
Attributes:

- More stable target
- Less data to train models
- Dynamics are often damped
- Tend to be worse ML problems

Explore adding features such as mean, total, min, max, slope while aggregating

Can hierarchical strategies help?

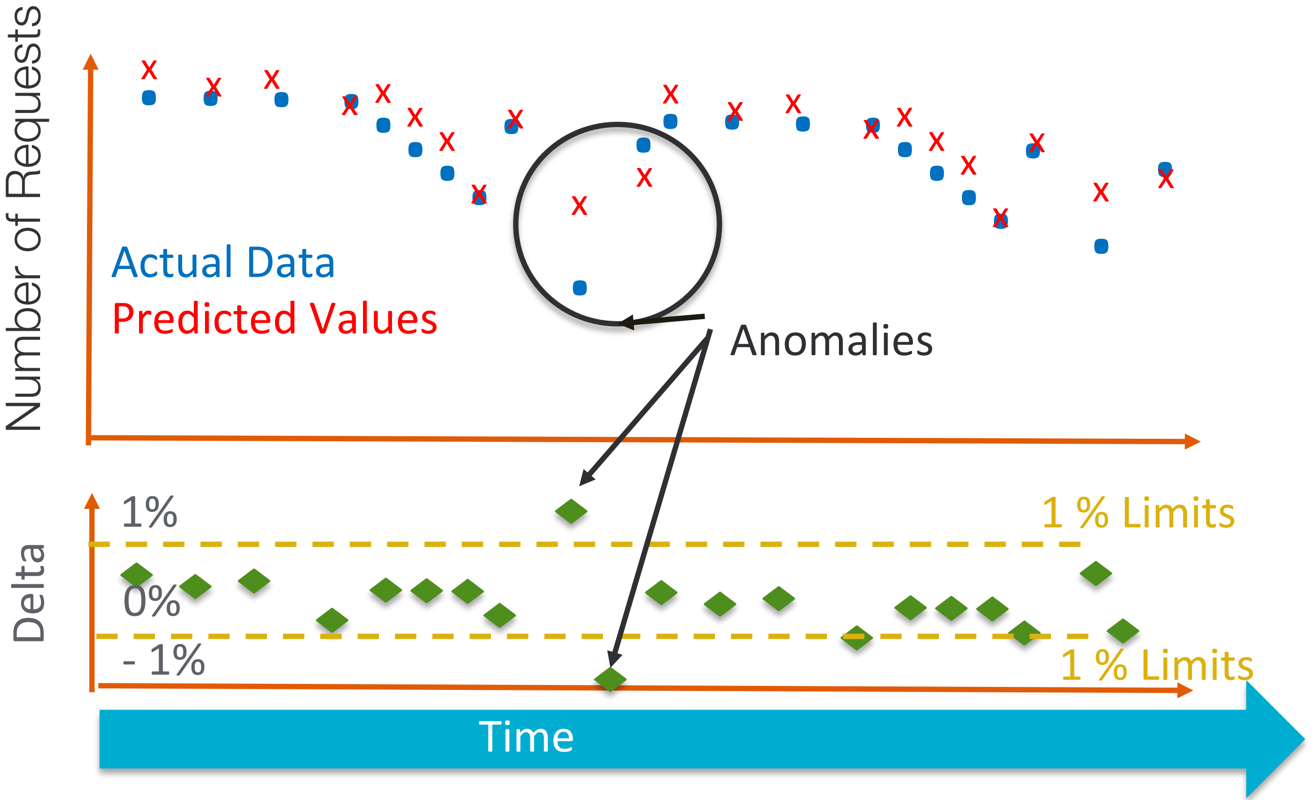
YES - Hierarchical modeling can improve performance for multi-series data where many of the series are very low-volume.



How does dynamic thresholding work?

Time series models can be used to monitor a process health

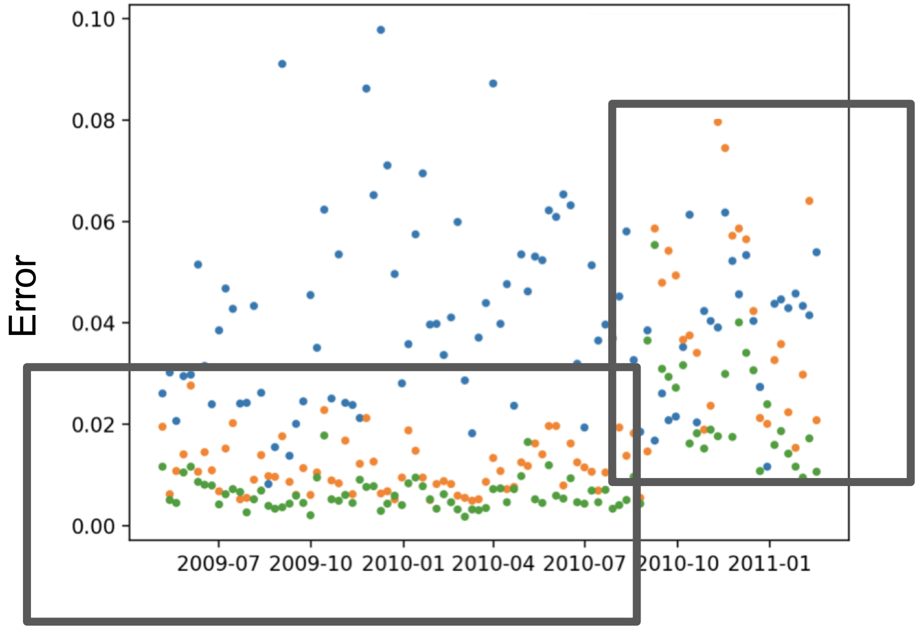
Difference between Actual - Prediction



What goes wrong?

Sudden system change

- New product launch
- Competitor opens a new store
- Regulatory changes



Good Performance Suddenly Falls Apart

We Are Hiring in Seoul!

- Customer-Facing Data scientist @ DataRobot
- Do you have?
  - 4-5+ years of real-world business experience in a Data science role
  - Hands-on experience building and implementing predictive models using machine learning algorithms
  - Strong customer interaction and project management skills
  - Excellent organizational, communication, writing, interpersonal skills
  - Familiarity with variety of technical tools for manipulation of datasets
  - Fluency with scripting (Python / R)
- Contact me!

Customer Facing Data Scientist

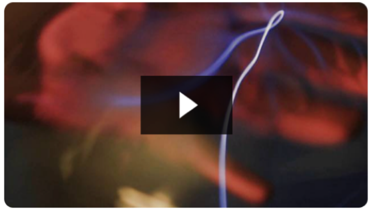
Seoul, South Korea

Customer Facing Data Scientists (CFDSs) are critical to making our customers successful. An ideal CFDS candidate should have strong fundamentals of applied data science in business setting, and should enjoy communicating and evangelizing data science solutions to business stakeholders.

Roles and responsibilities :

- Product
    - Representing the DataRobot product from a technical standpoint to customers – including demonstrations, conducting proof-of-concept trials, helping clients evaluate success criteria, and training users
    - Providing the customer's point of view to DataRobot's Product team, informing the direction of future product feature development

Data Science



Our Data Science team is composed of people from around the globe united by their passion for

DataRobot

clifton@datarobot.com
https://www.linkedin.com/in/cliftonphua/