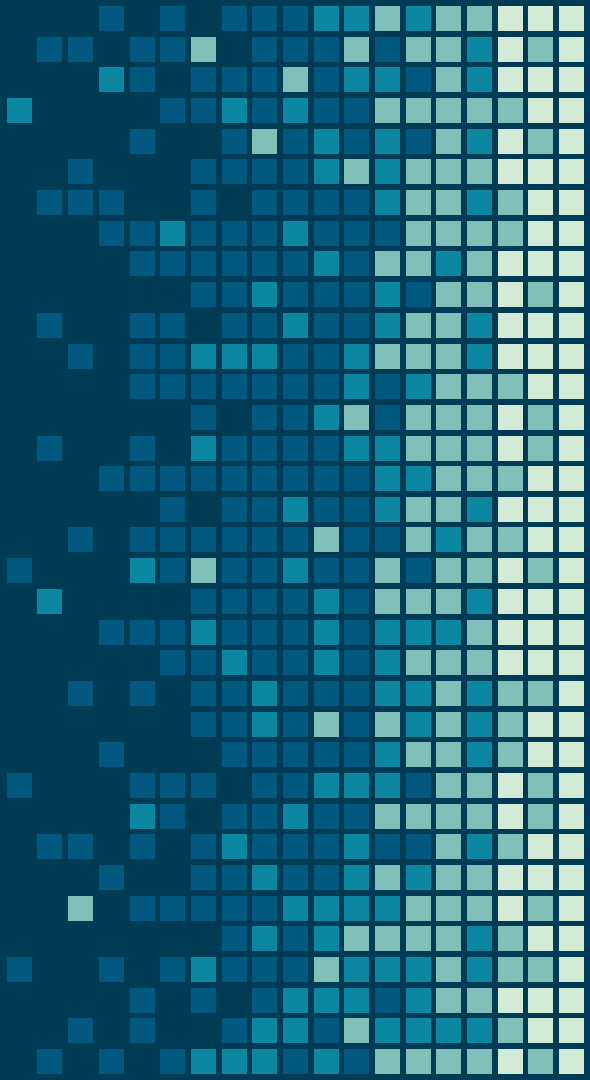
Anomaly Detection

A to Z



Why Listen?

Intro Anomaly Detection

- What & Why

Anomaly Detection Methods

- Basic
- More Complex
- Advanced

John Hogue

Lead Data Scientist

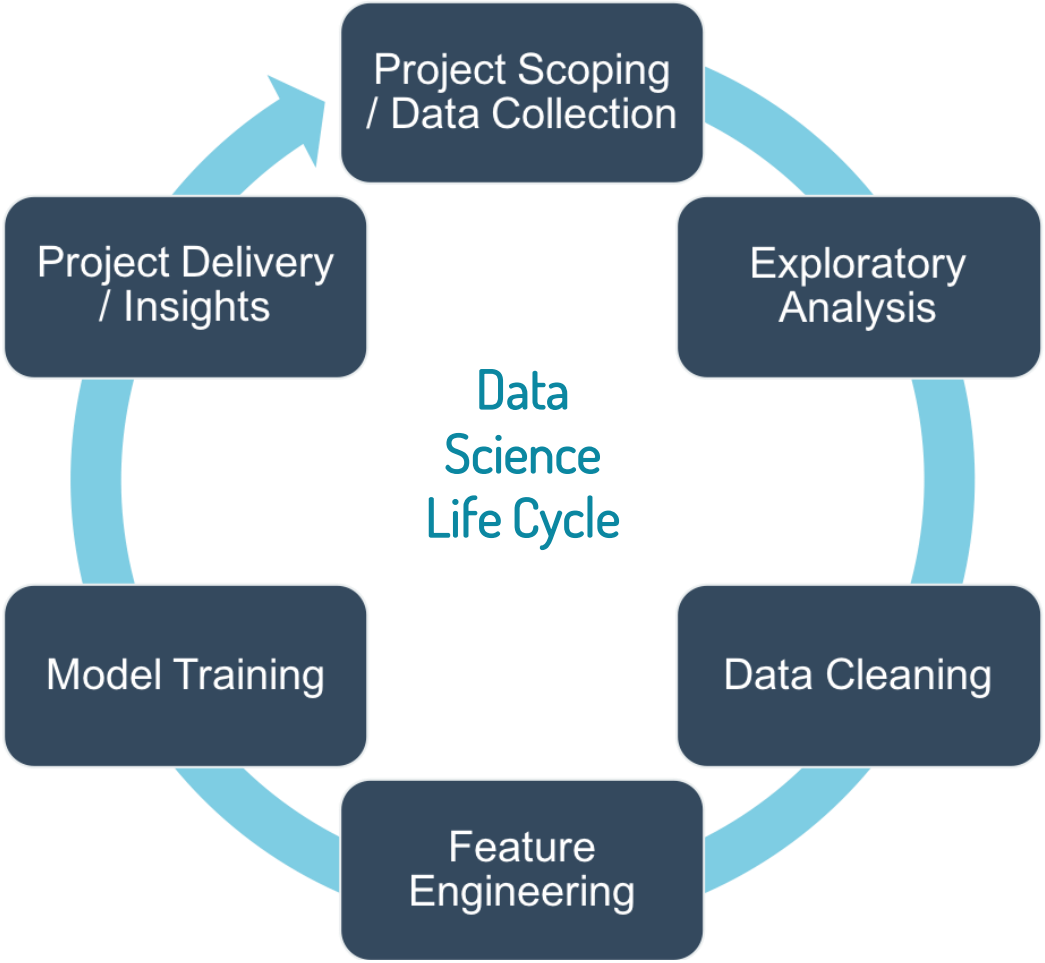
- General Mills

CoFounder

- Social Data Science

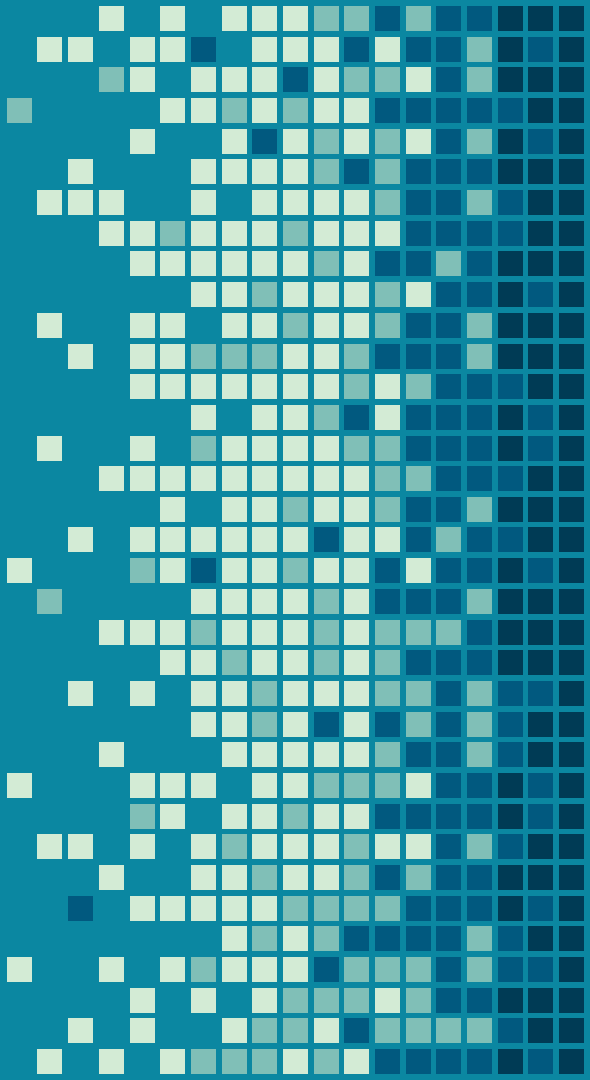
Executive Advisor

- MinneAnalytics


Project Scoping / Data Collection
Exploratory Analysis
Data Cleaning
Feature Engineering
Model Training
Project Delivery / Insights
Data Science Life Cycle

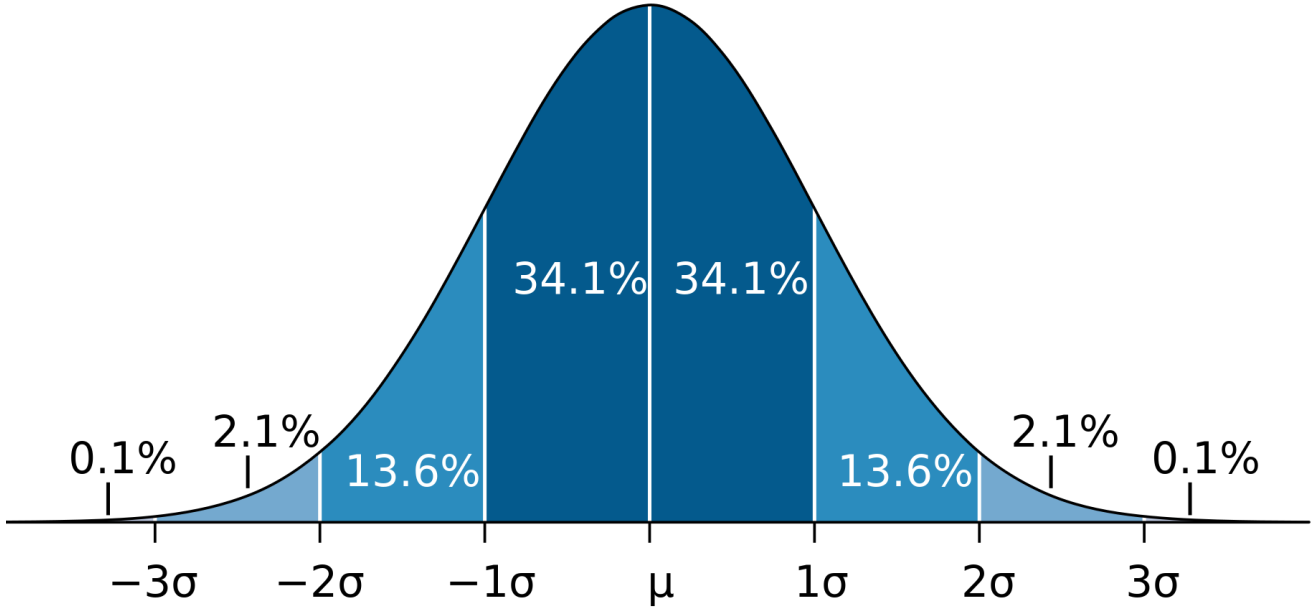
“Anomaly

Something that deviates from what is standard, normal, or expected.



What is ‘standard’, ‘normal’ or ‘expected’?

Mathematically…

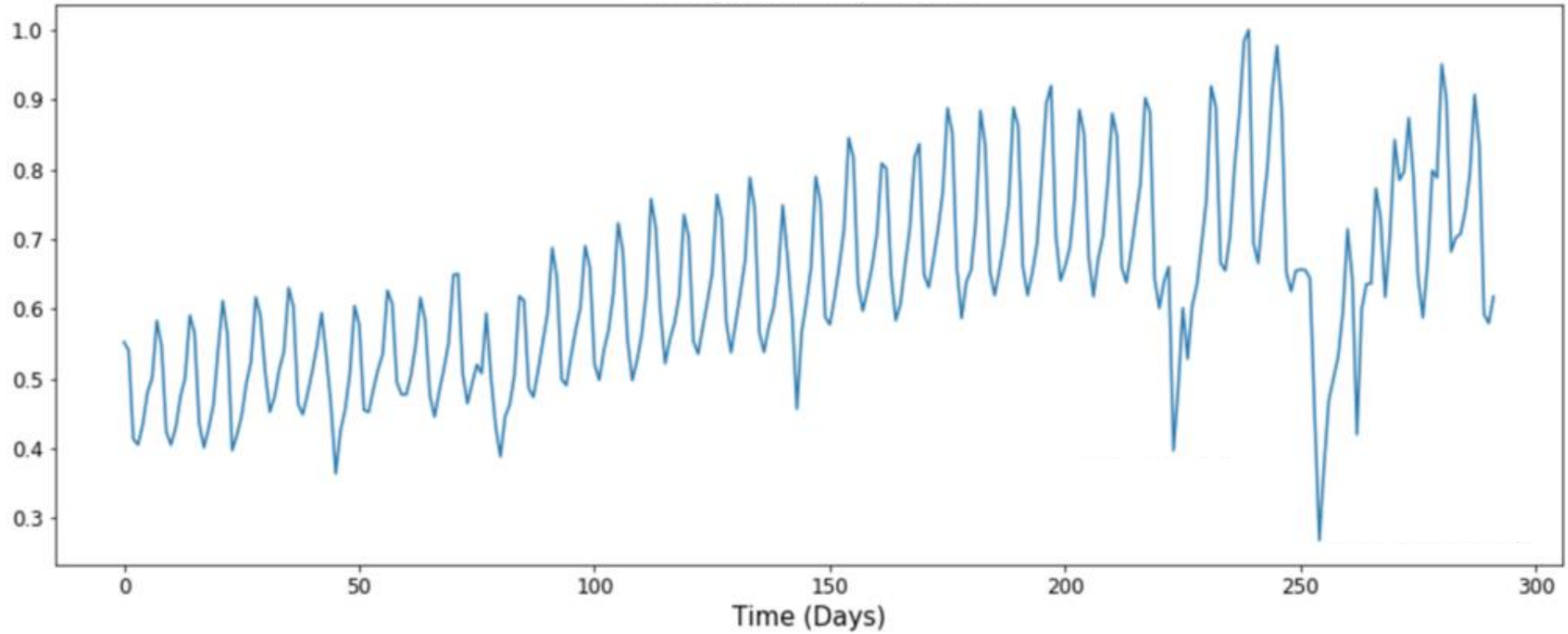


What is ‘standard’, ‘normal’ or ‘expected’?

Contextually…

- One’s person’s anomaly is another’s common sense.
- **Business Acumen** decides what is actionable.

But what about context?

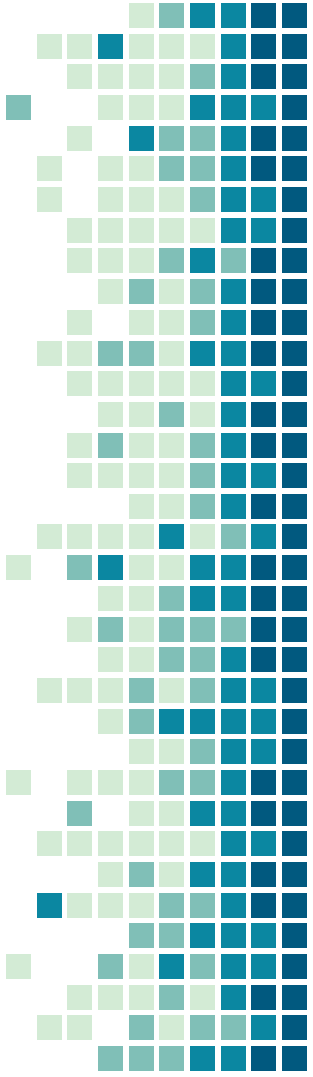


Why bother detecting?

- Accurately compare period over period
- Identify changepoints

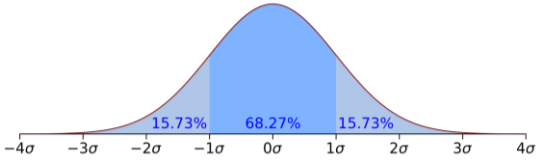
Domains

- Health & Life Sciences
- Fraud Detection
- Network Intrusion/Abuse
- Marketing Effectiveness
- Manufacturing

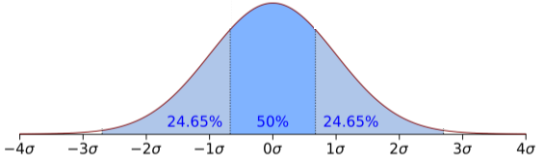


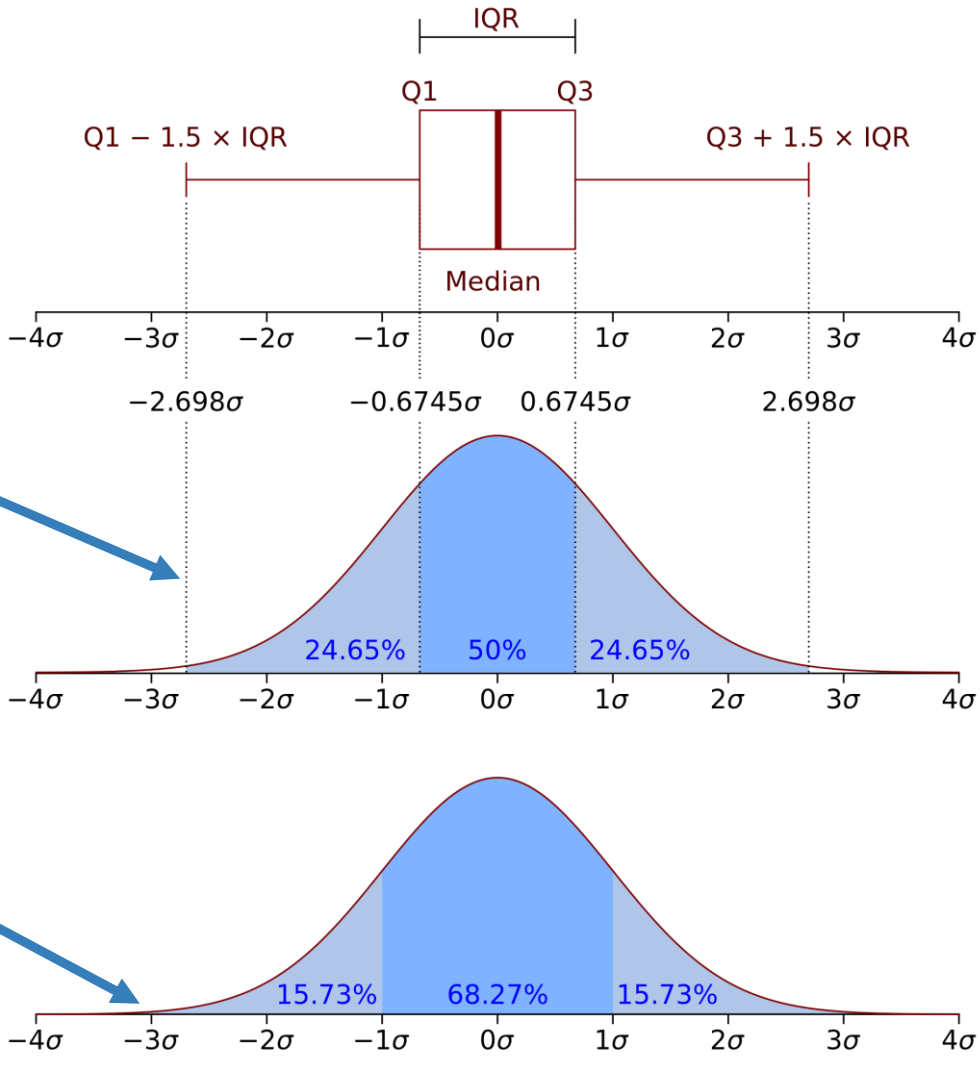
Basic outlier detection

- *Z-Scores*
  - $z = (x - \mu) / \sigma$
  - $\mu \mp 3\ \sigma$ is a default
    - 0.3% of data outlier



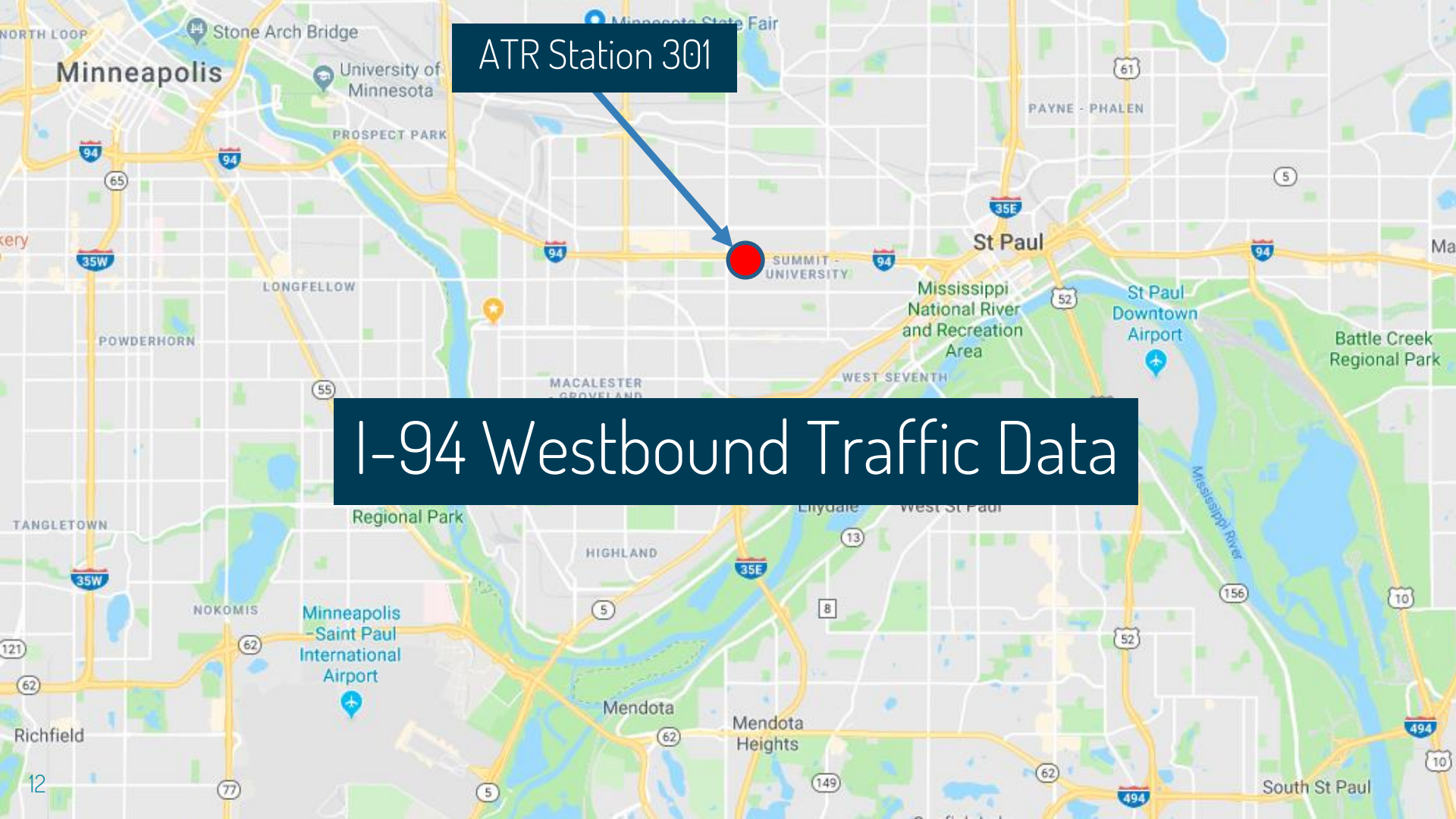
- *Tukey's Inter Quartile Range*
  - ±1.5IQR
    - 0.7% of data outlier

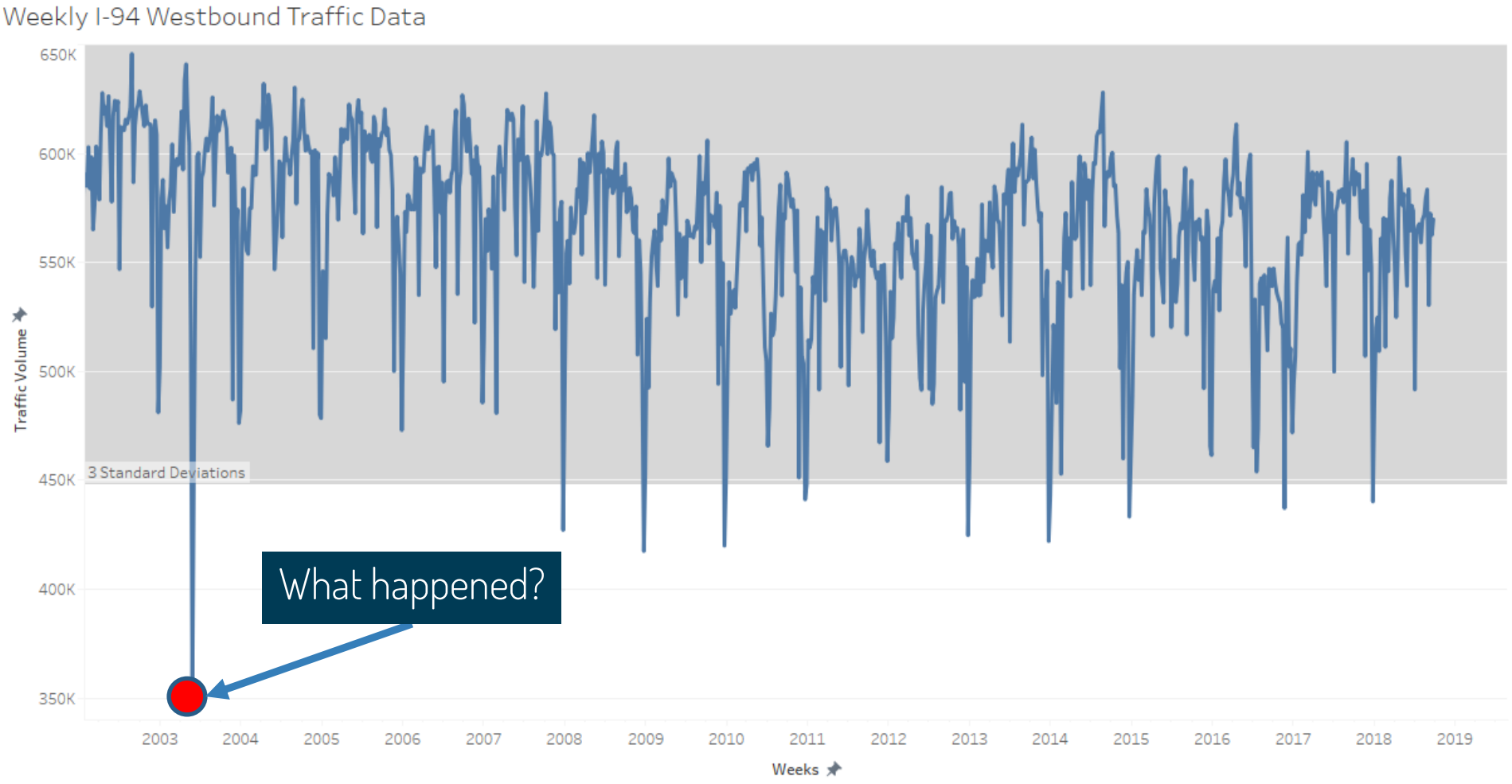




Q1 -1.5*IQR

Z Score of 3


ATR Station 301
I-94 Westbound Traffic Data
Minneapolis
Stone Arch Bridge
University of Minnesota
PROSPECT PARK
LONGFELLOW
POWDERHORN
MACALESTER
WEST SEVENTH
St Paul
SUMMIT - UNIVERSITY
Mississippi National River and Recreation Area
St Paul Downtown Airport
Battle Creek Regional Park
PAYNE - PHALEN
TANGLETOWN
Regional Park
HIGHLAND
NOKOMIS
Minneapolis -Saint Paul International Airport
Mendota
Mendota Heights
Richfield
South St Paul
Mississippi River


Weekly I-94 Westbound Traffic Data
650K
600K
550K
500K
450K
400K
350K
Traffic Volume
3 Standard Deviations
What happened?
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Weeks

The art of creating a mediocre model

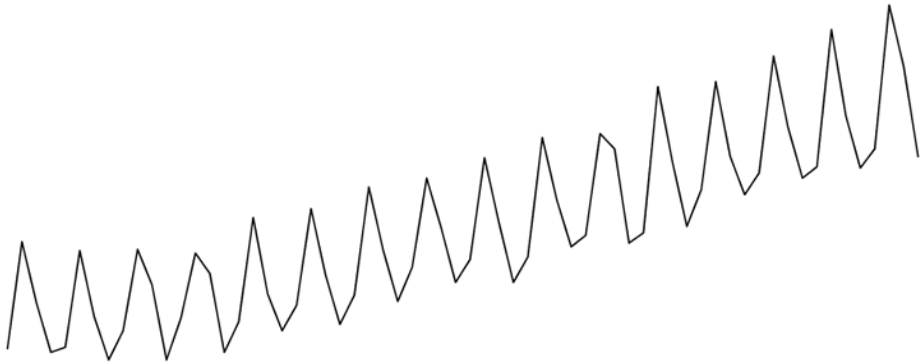
- *Generalizing the pattern* is the most critical

- **Too good of a model**
  - Too few anomalies
- **Too crappy of a model**
  - Everything is an anomaly

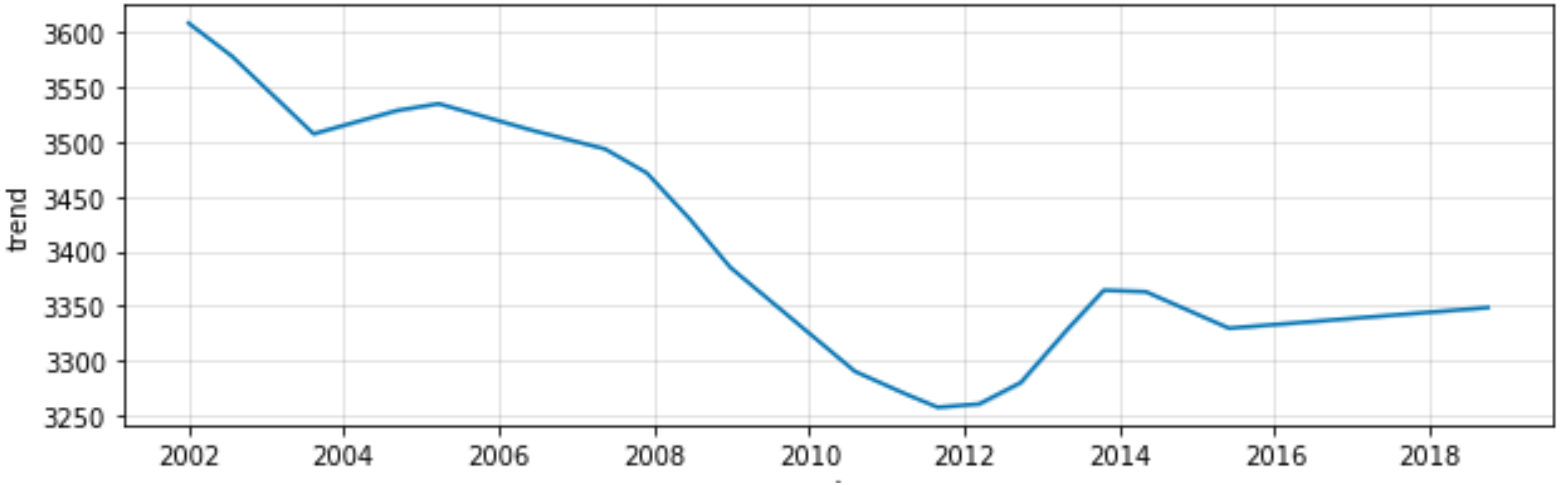
Seasonality & Trend

- *Seasonality*
  - Yearly
  - Weekly
  - Daily
  - Holiday Related
    - Moving Holidays**
- *Trend*
  - Change in behavior
  - Change in population

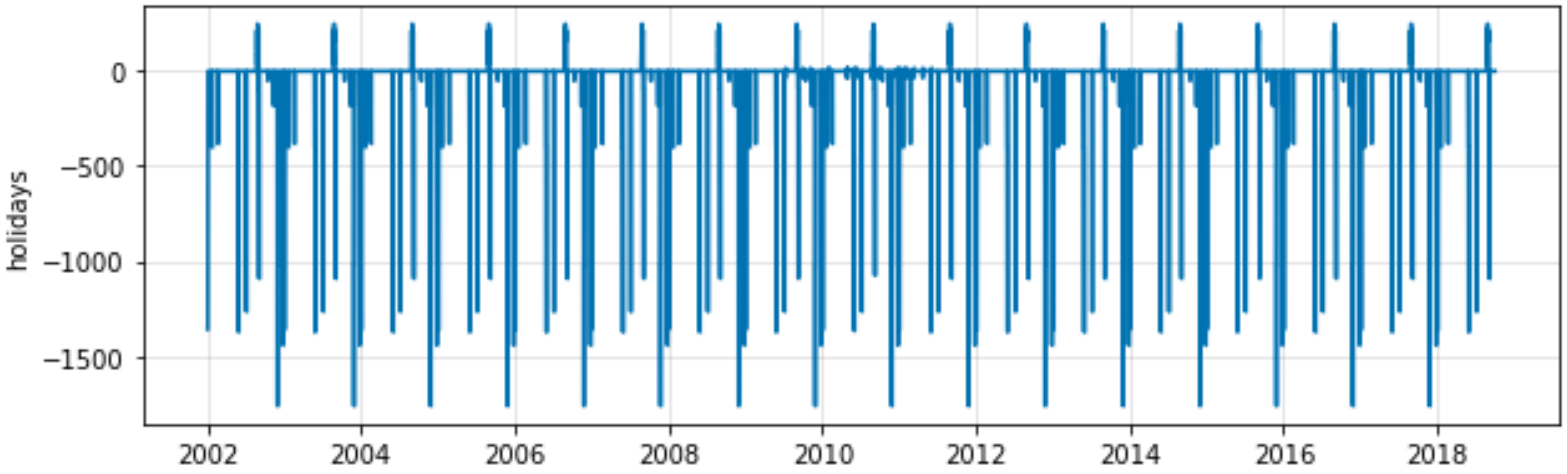
Original



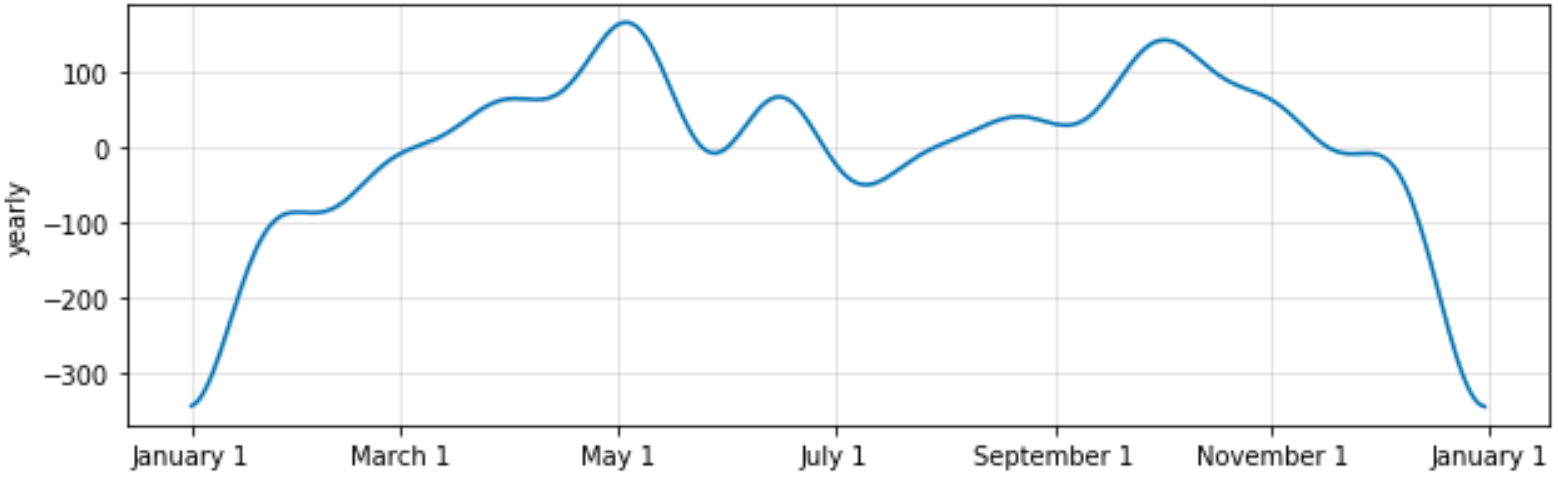
I-94 Westbound Trend



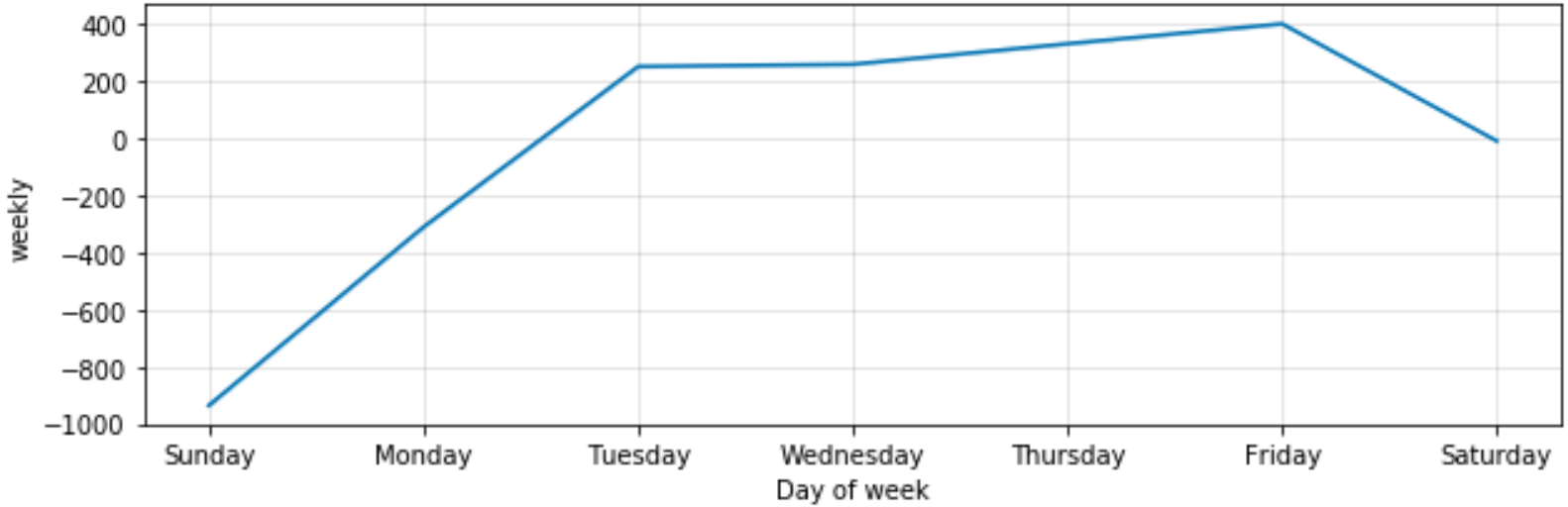
I-94 Westbound Holidays



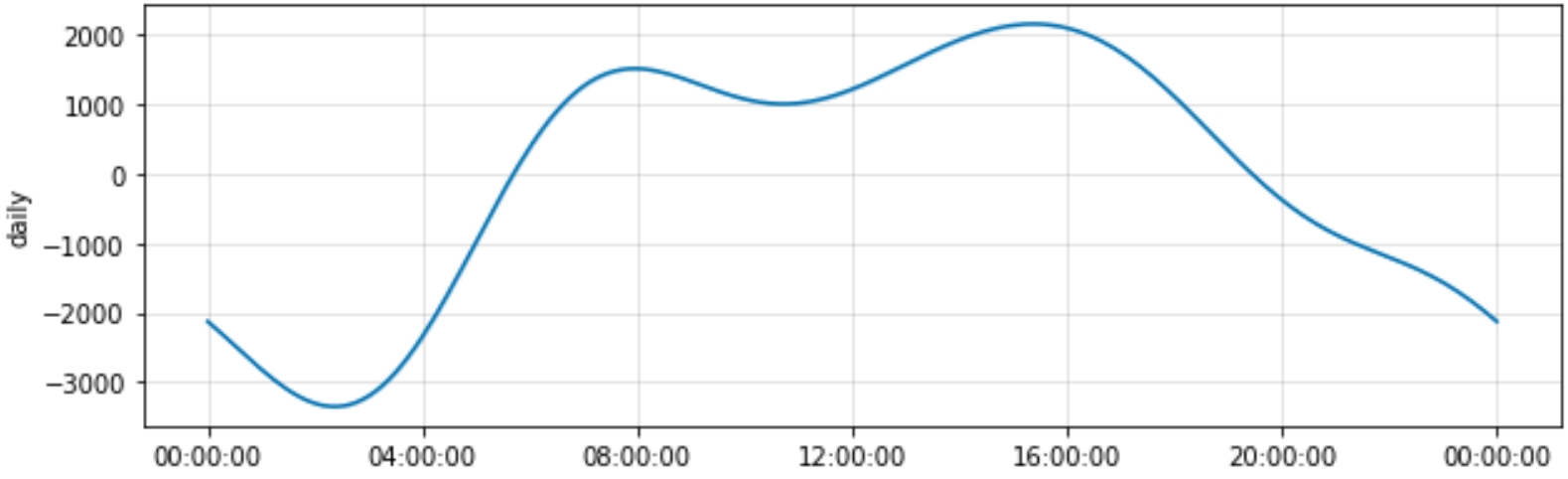
I-94 Westbound Yearly



I-94 Westbound Weekly**

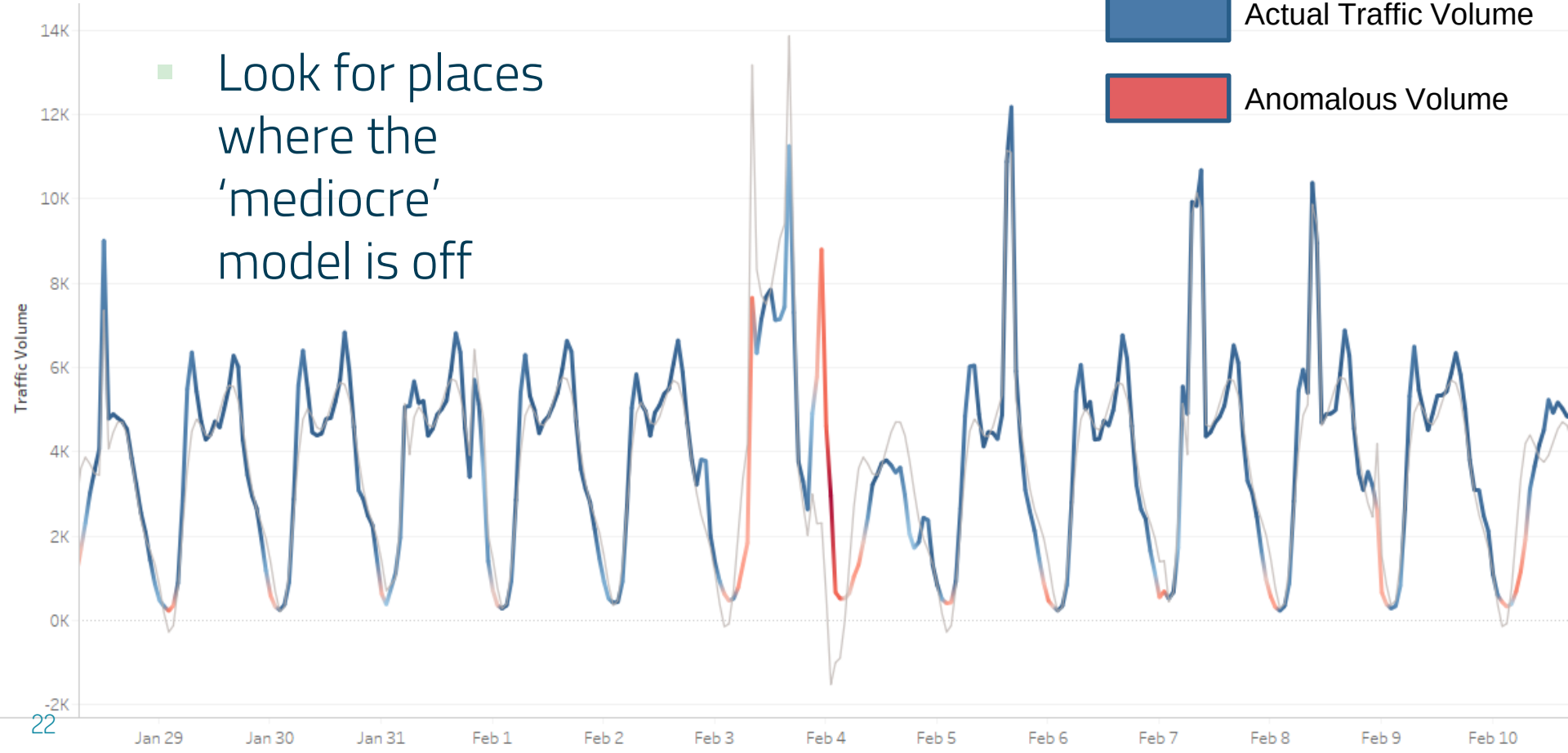


I-94 Westbound Daily

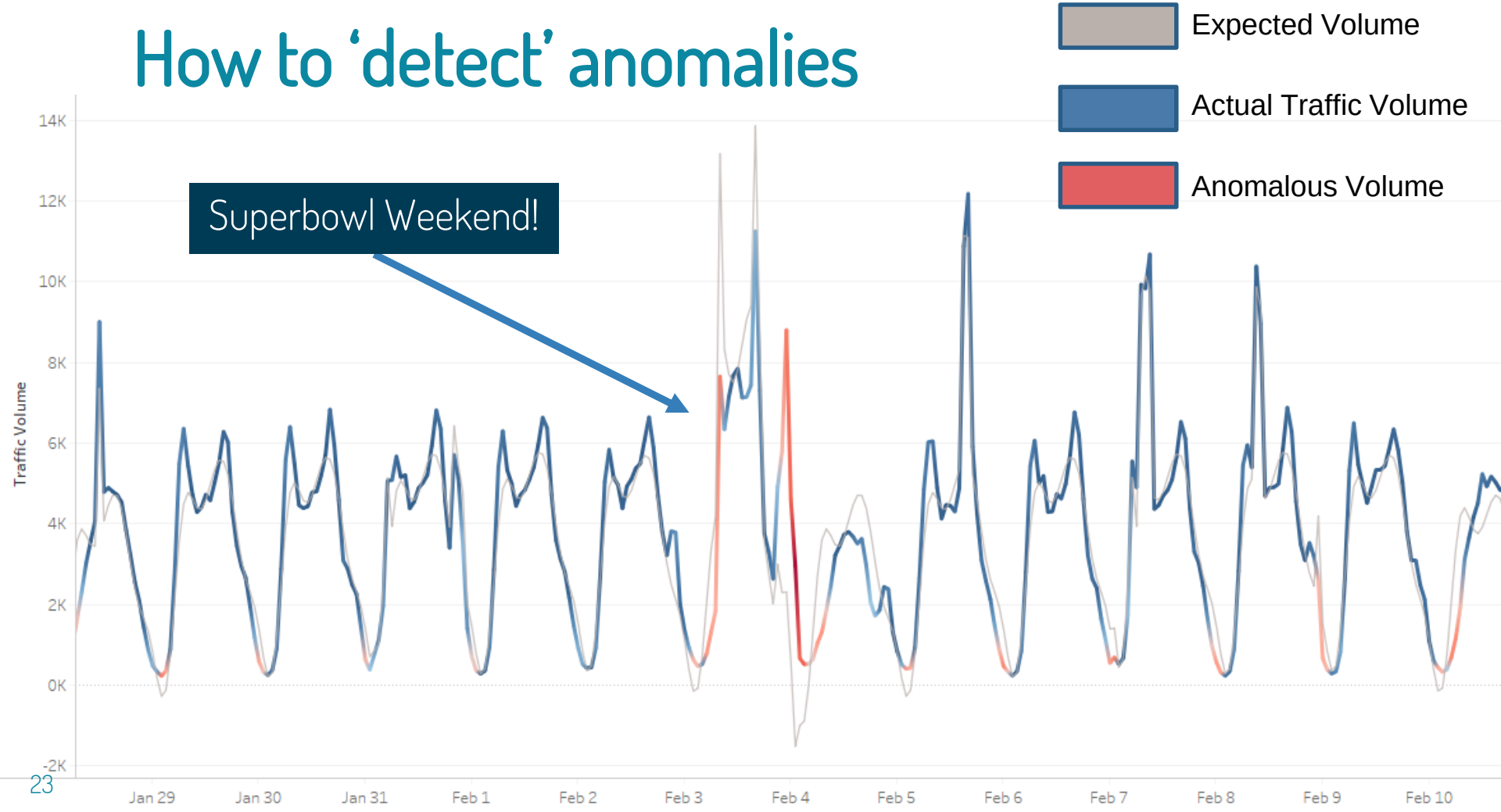


How to ‘detect’ anomalies

- Look for places where the ‘mediocre’ model is off

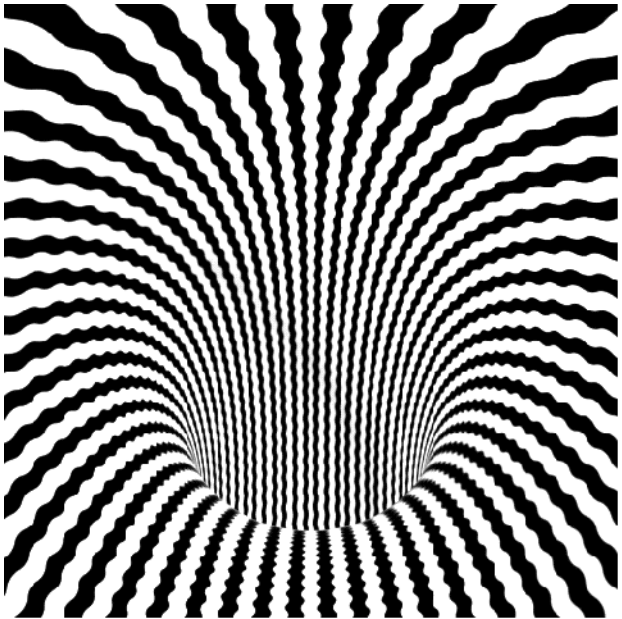


How to ‘detect’ anomalies

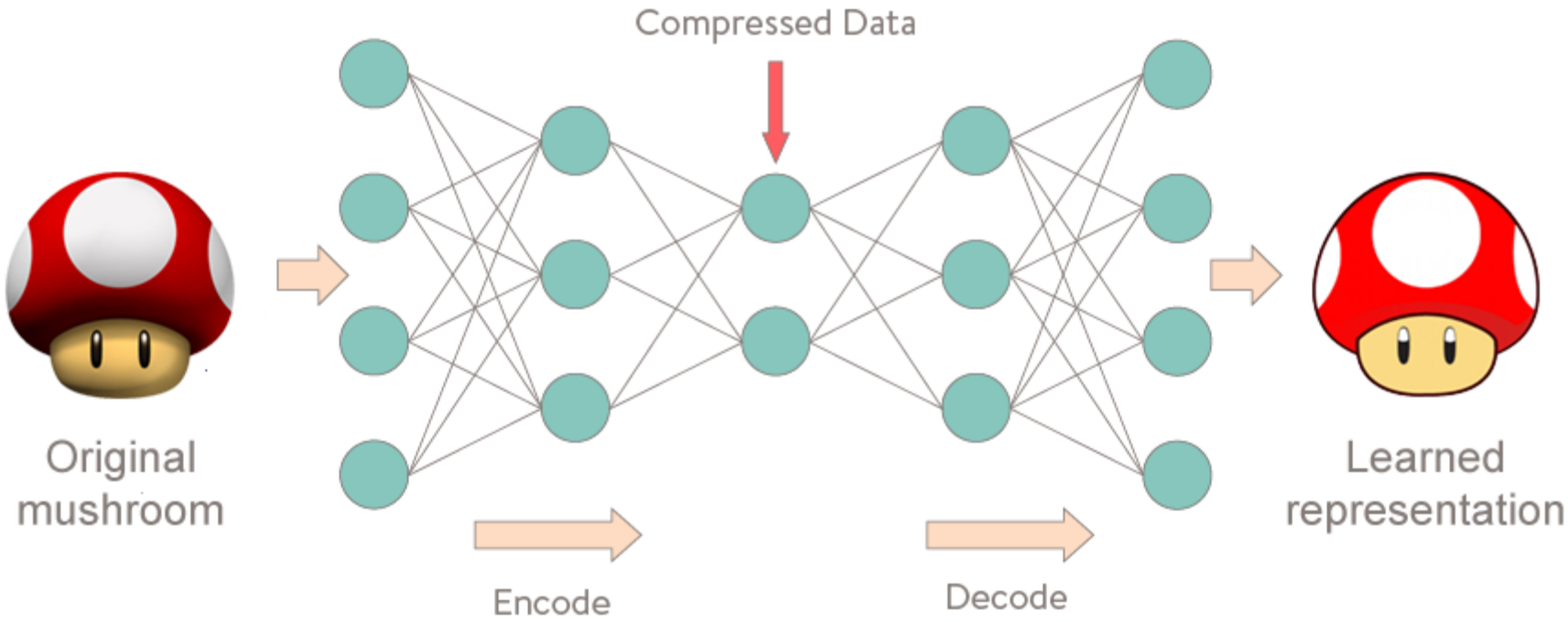


What if we had more things to model?

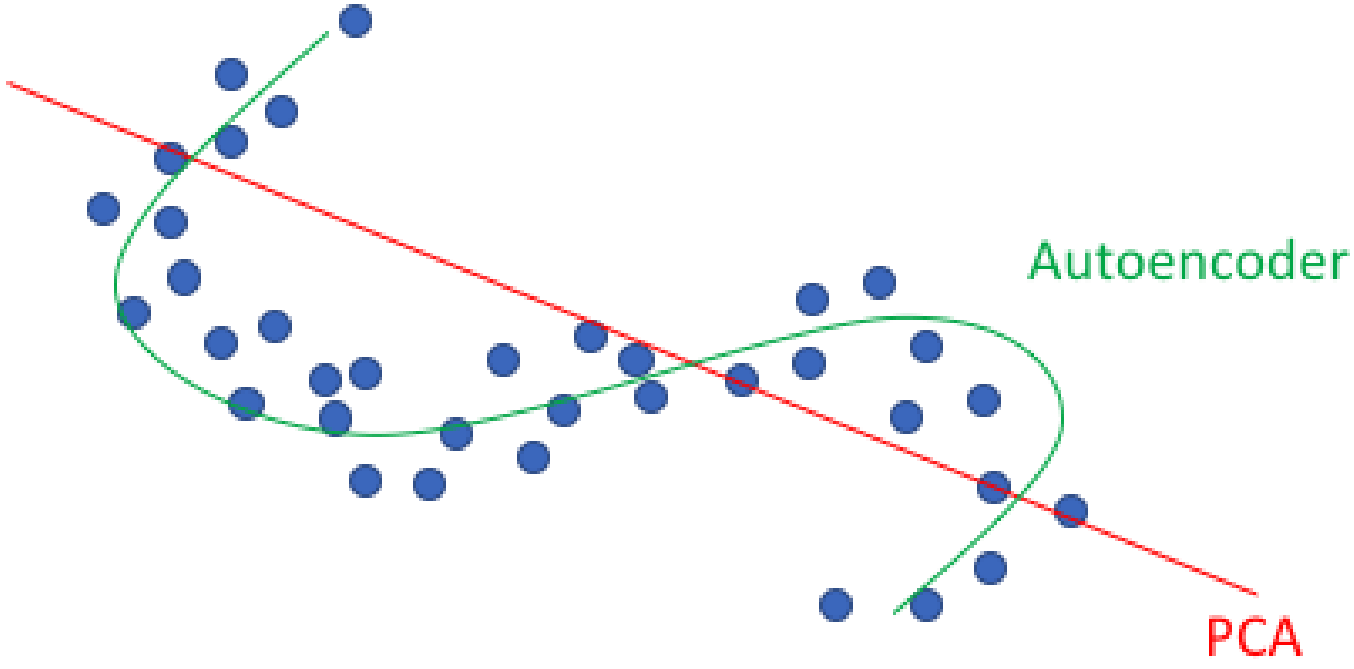
- The case for Multivariate
  - Rain?
  - Snow?
  - Accidents?
  - Lane Closures?
  - It never ends!



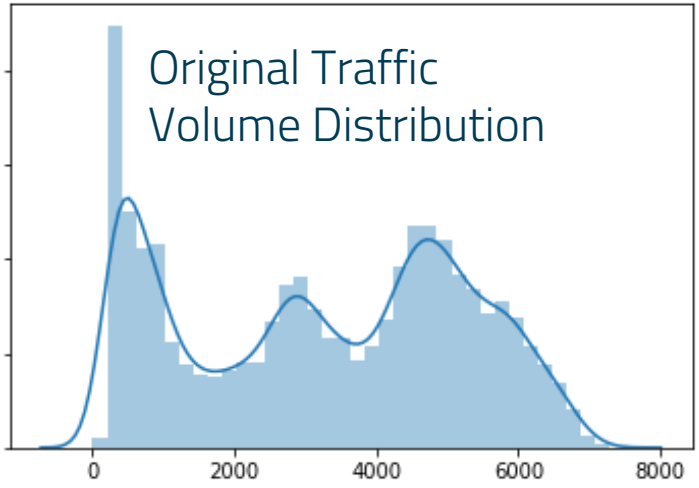
Autoencoders

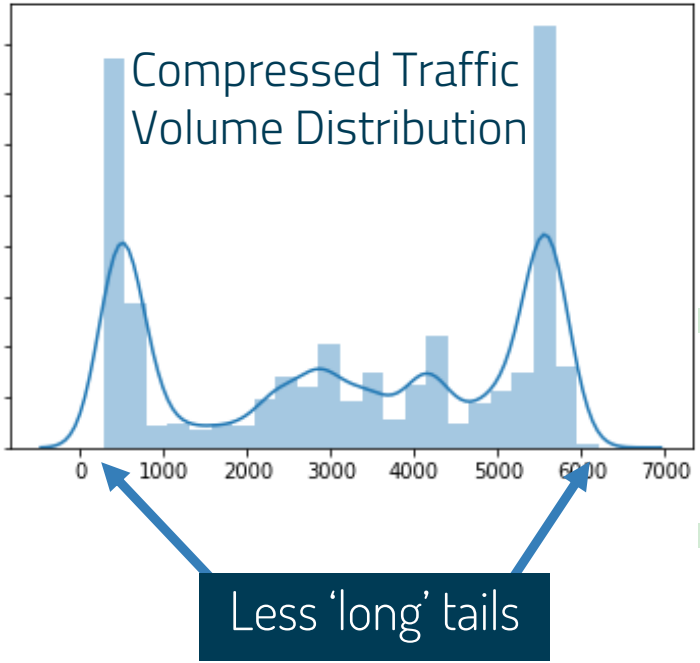


Capturing Non Linear Trends

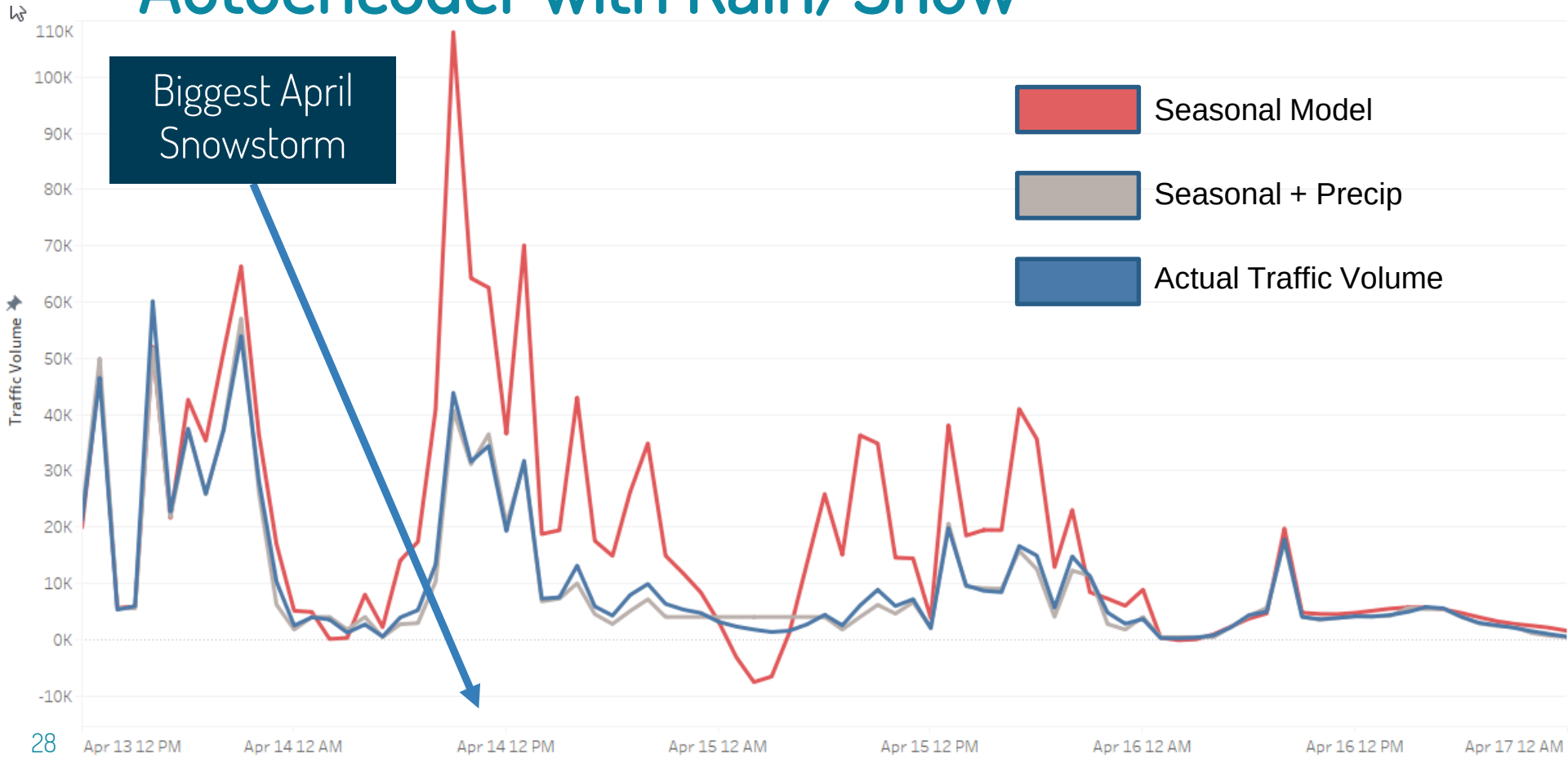


Back to basics: Traffic Volume





Autoencoder with Rain/Snow



General Mills Hiring!

- **Senior Data Scientist**
  - careers.generalmills.com

- **GitHub**
  - dreyco676/Anomaly_Detection_A_to_Z