Time Series

Forecasting the Future

Charles Parker
VP ML Algorithms, BigML, Inc

What is Time Series?

- Traditional machine learning data is assumed to be Independent & Identically Distributed
    - Independent (points have no information about each other's class)
    - Identically distributed (come from the same distribution)

Independent Data

Color	Mass	PPAP
red	11	pen
green	45	apple
red	53	apple
yellow	0	pen
blue	2	pen
green	422	pineapple
yellow	555	pineapple
blue	7	pen

Discovering patterns:

- Color = “red” ⇒ Mass < 100
- PPAP = “pineapple” ⇒ Color ≠ “blue”
- Color = “blue” ⇒ PPAP = “pen”

Independent Data

Color	Mass	PPAP
green	45	apple
blue	2	pen
green	422	pineapple
blue	7	pen
yellow	0	pen
yellow	9	pineapple
red	555	apple
red	11	pen

Patterns still hold when rows re-arranged:

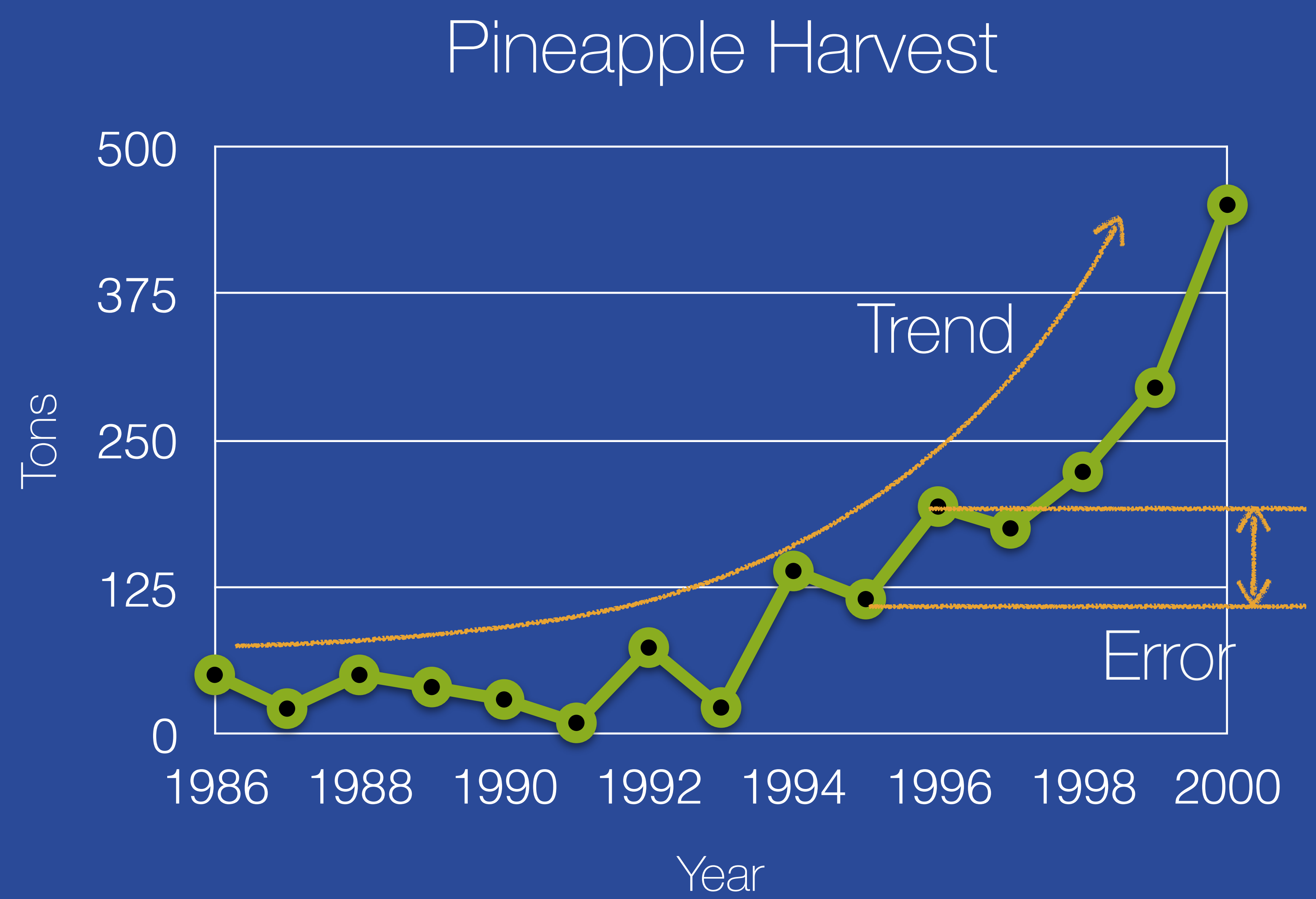
- Color = “red” ⇒ Mass < 100
- PPAP = “pineapple” ⇒ Color ≠ “blue”
- Color = “blue” ⇒ PPAP = “pen”

What is Time Series?

- Traditional machine learning data is assumed to be Independent & Identically Distributed
    - Independent (points have no information about each other's class)
    - Identically distributed (come from the same distribution)
- But what if you want to predict just the next value in a sequence?
    - `1, 2, 3, 2, ???`
    - No longer independent, not identically distributed!

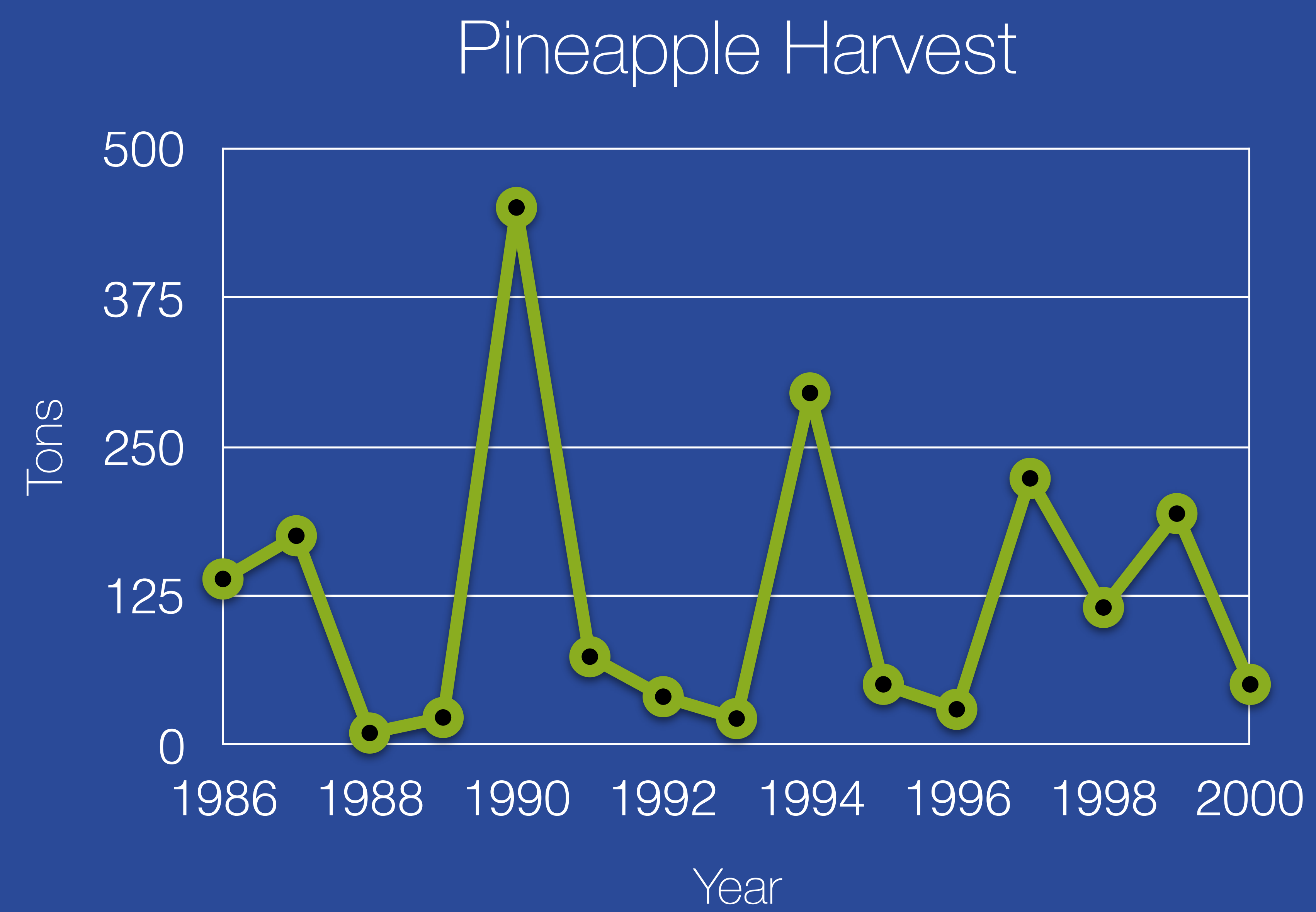
Dependent Data

Year	Pineapple
1986	50.74
1987	22.03
1988	50.69
1989	40.38
1990	29.80
1991	9.90
1992	73.93
1993	22.95
1994	139.09
1995	115.17
1996	193.88
1997	175.31
1998	223.41
1999	295.03
2000	450.53



Dependent Data

Year	Pineapple
1986	139.09
1987	175.31
1988	9.91
1989	22.95
1990	450.53
1991	73.93
1992	40.38
1993	22.03
1994	295.03
1995	50.74
1996	29.8
1997	223.41
1998	115.17
1999	193.88
2000	50.69



Rearranging Disrupts Patterns

What is Time Series?

- Traditional machine learning data is assumed to be Independent & Identically Distributed
    - Independent (points have no information about each other's class)
    - Identically distributed (come from the same distribution)
- But what if you want to predict just the next value in a sequence?
    - `1, 2, 3, 2, ???`
    - No longer independent, not identically distributed!
    - This also changes how we evaluate!

Random Train / Test Split

Train

Test

plasma glucose	bmi	diabetes pedigree	age	[illegible]
148	33.6	0.627	50	TRUE
85	26.6	0.351	[illegible]	FALSE
183	23.3	0.672	[illegible]	TRUE
89	28.1	0.167	[illegible]	FALSE
137	43.1	2.288	[illegible]	[illegible]
116	25.6	0.201	30	[illegible]
78	31	0.248	26	TRUE
115	35.3	0.134	29	FALSE
197	30.5	0.158	53	TRUE

Linear Train / Test Split

Train **Test**

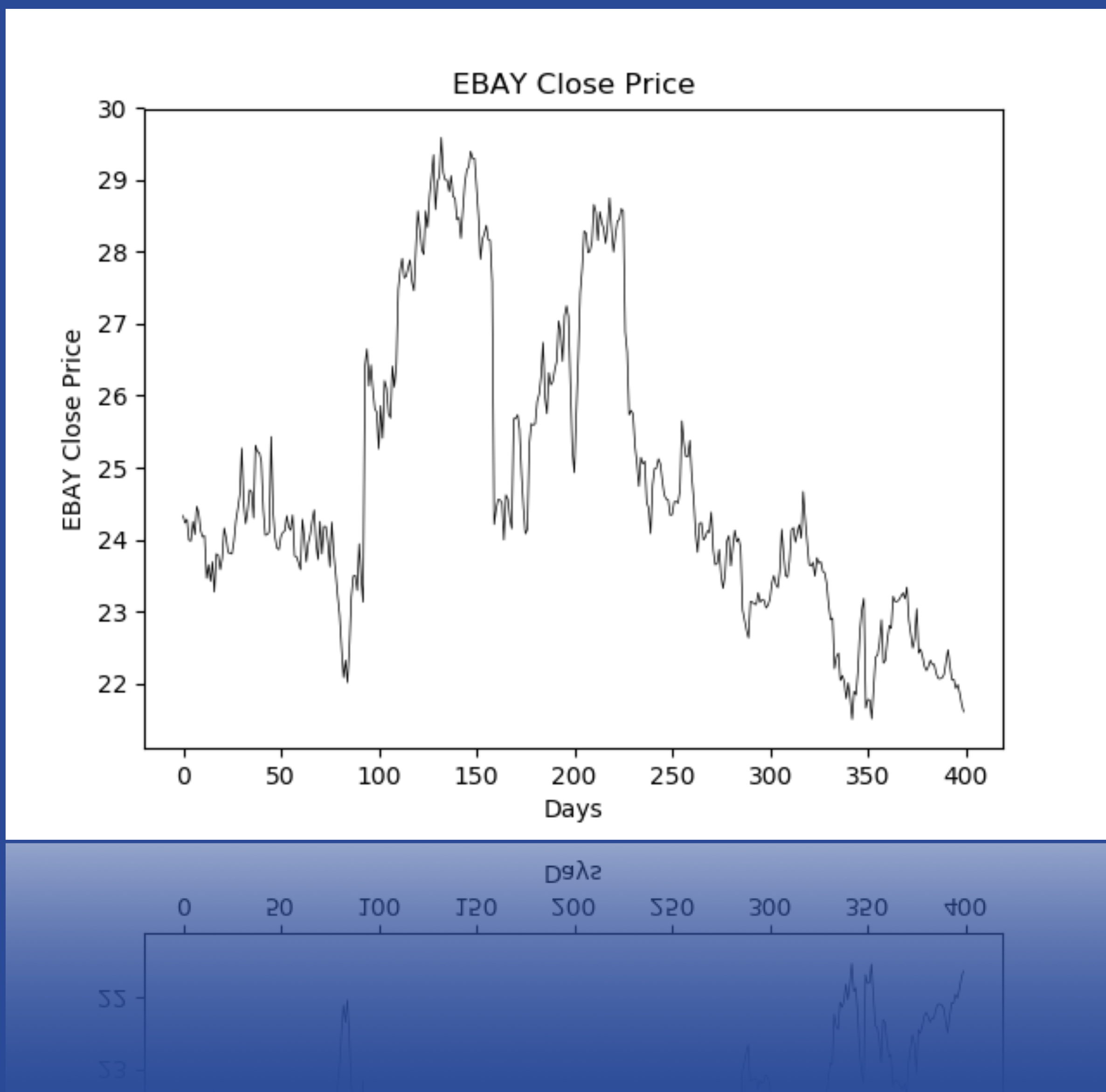
Year	Pineapple
1986	50.74
1987	22.03
1988	50.69
1989	40.38
1990	29.80
1991	9.90
1992	73.93
1993	22.95
1994	139.09
1995	115.17
1996	193.88
1997	175.31
1998	223.41
1999	295.03
2000	450.53

Forecast

COMPARE

Exponential Smoothing

Idea: Each new value in the series depends on all previous values with a decaying weight

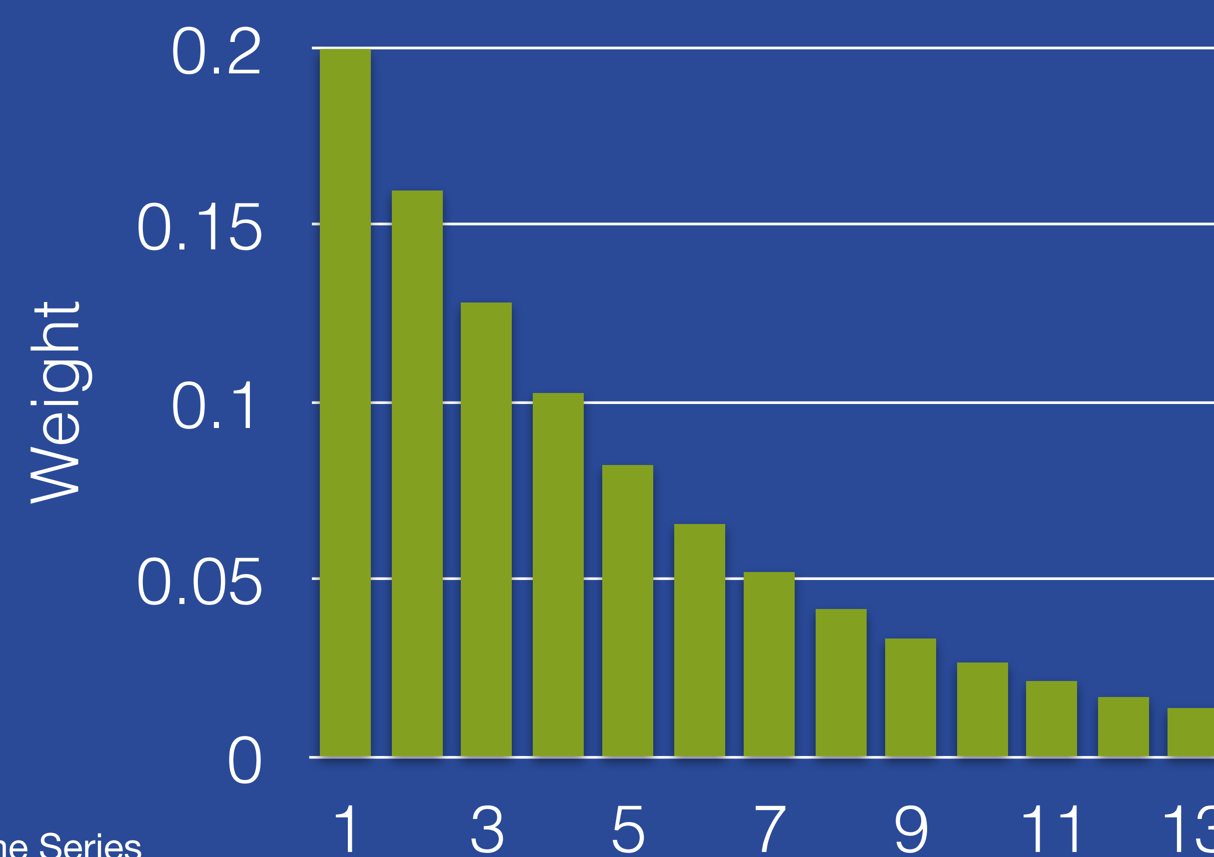


For training values x_t

Smoothing Factor $0 < \alpha < 1$

Predicted values s_t

$$s_t = \alpha \cdot x_t + (1-\alpha) \cdot s_{t-1}$$



Smoothing Factor

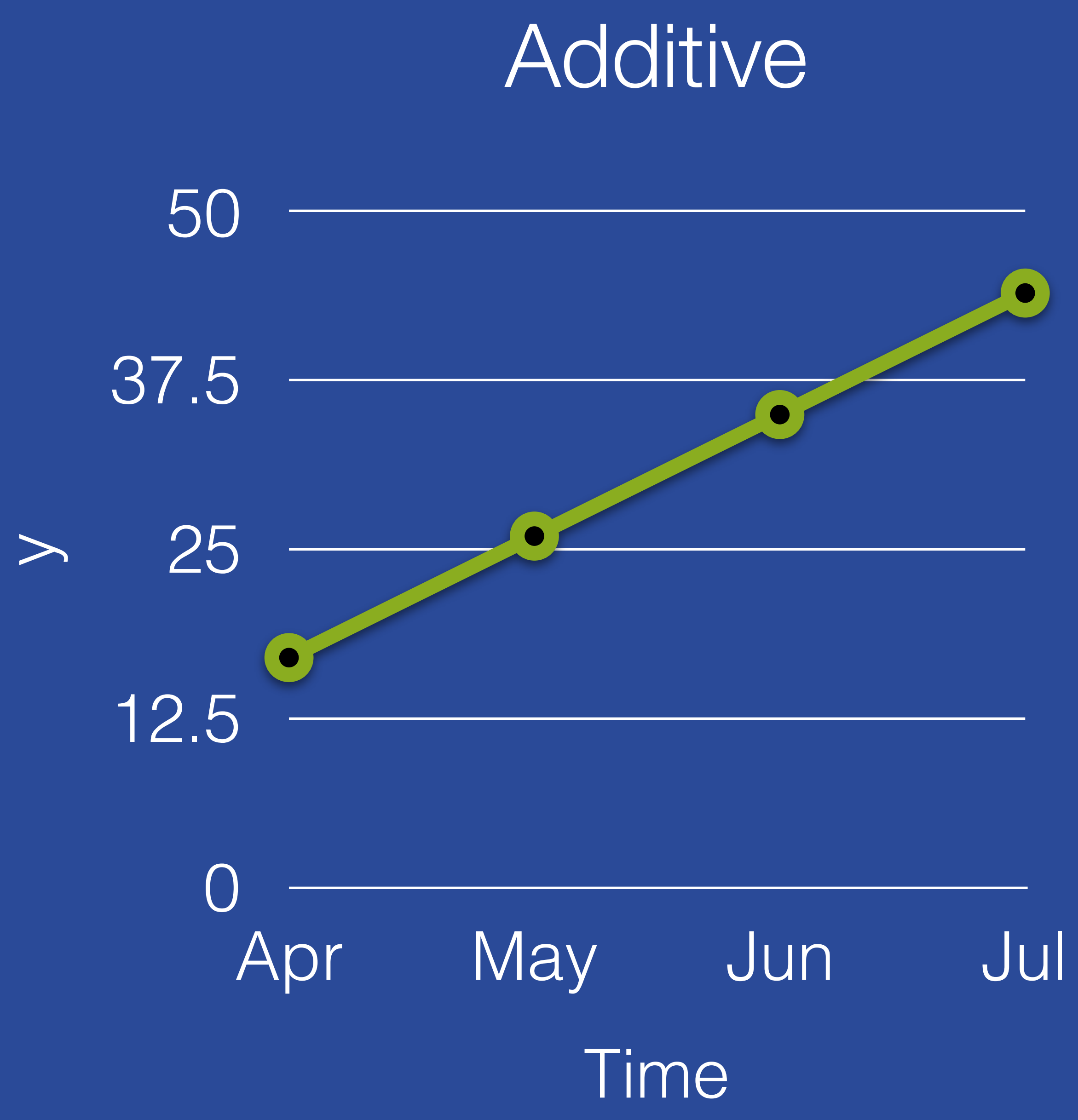
$$s_t = \alpha \cdot x_t + (1-\alpha) \cdot s_{t-1}$$

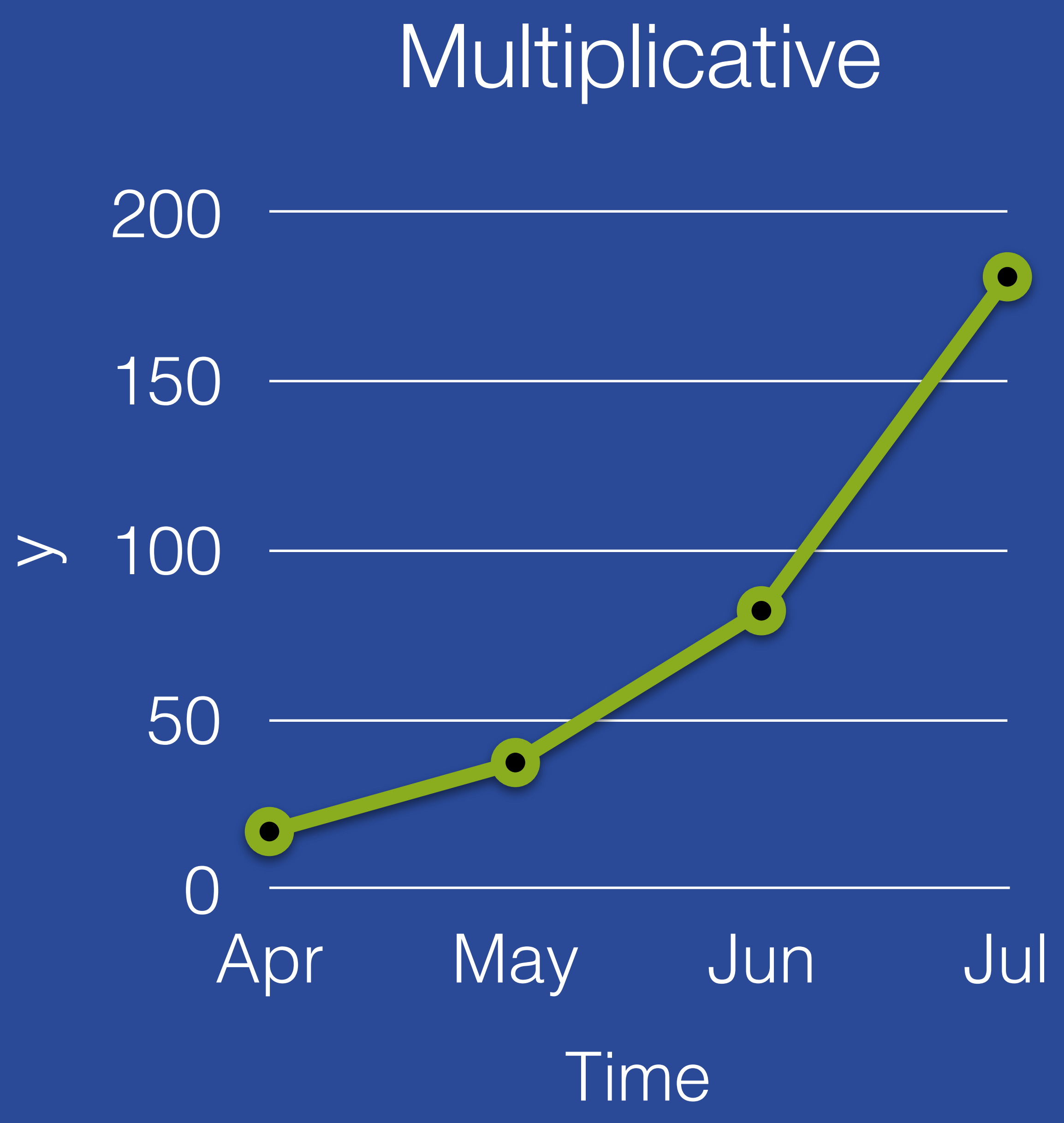
- $\alpha \rightarrow 0$
  - Series relies more heavily on past values
- $\alpha \rightarrow 1$
  - Series relies more heavily on current value
- $\alpha = 1$
  - Series is the current value

Problem: Real-world data is more complicated than this…

Trend

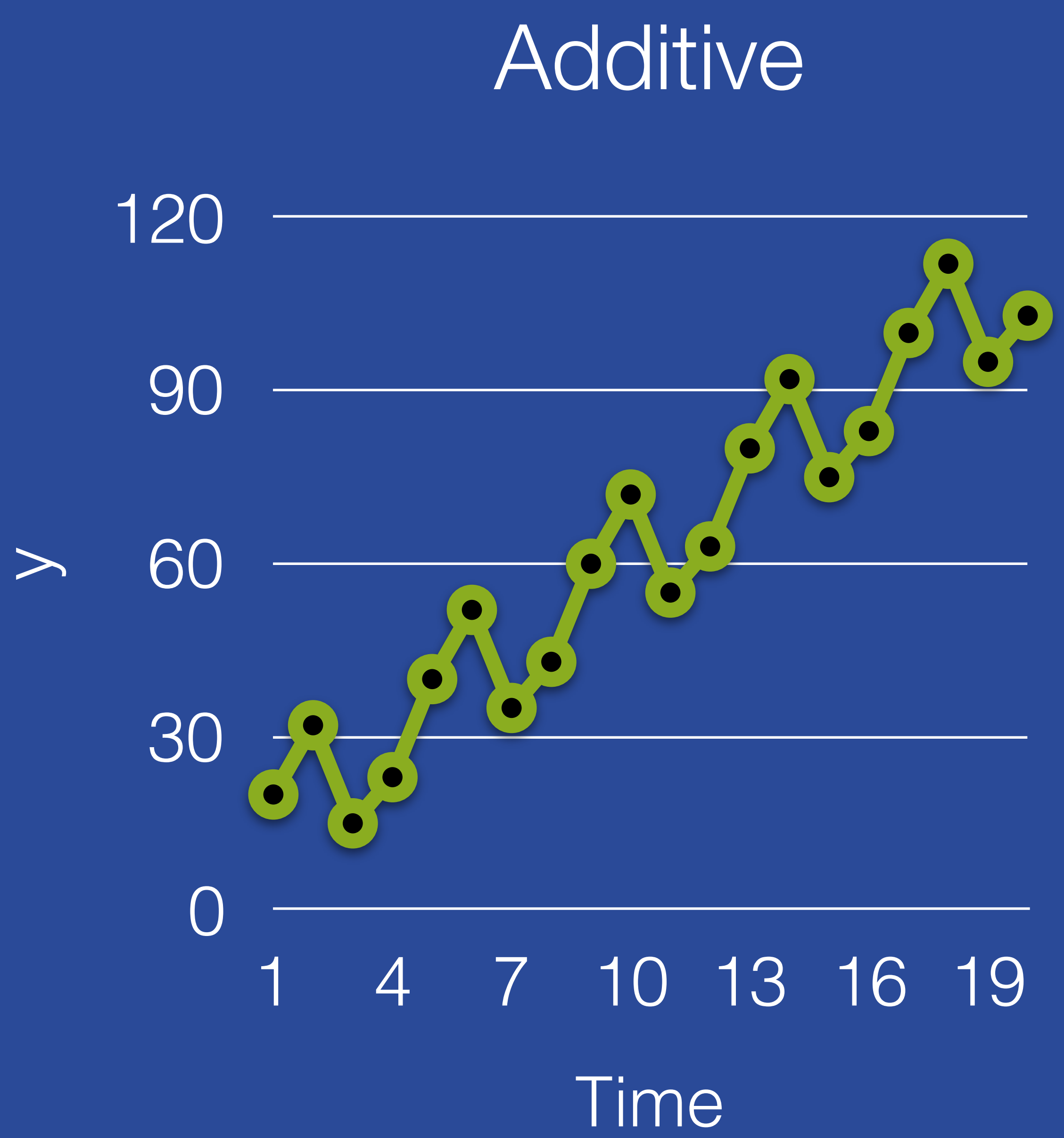
Trend: A persistent long-term pattern

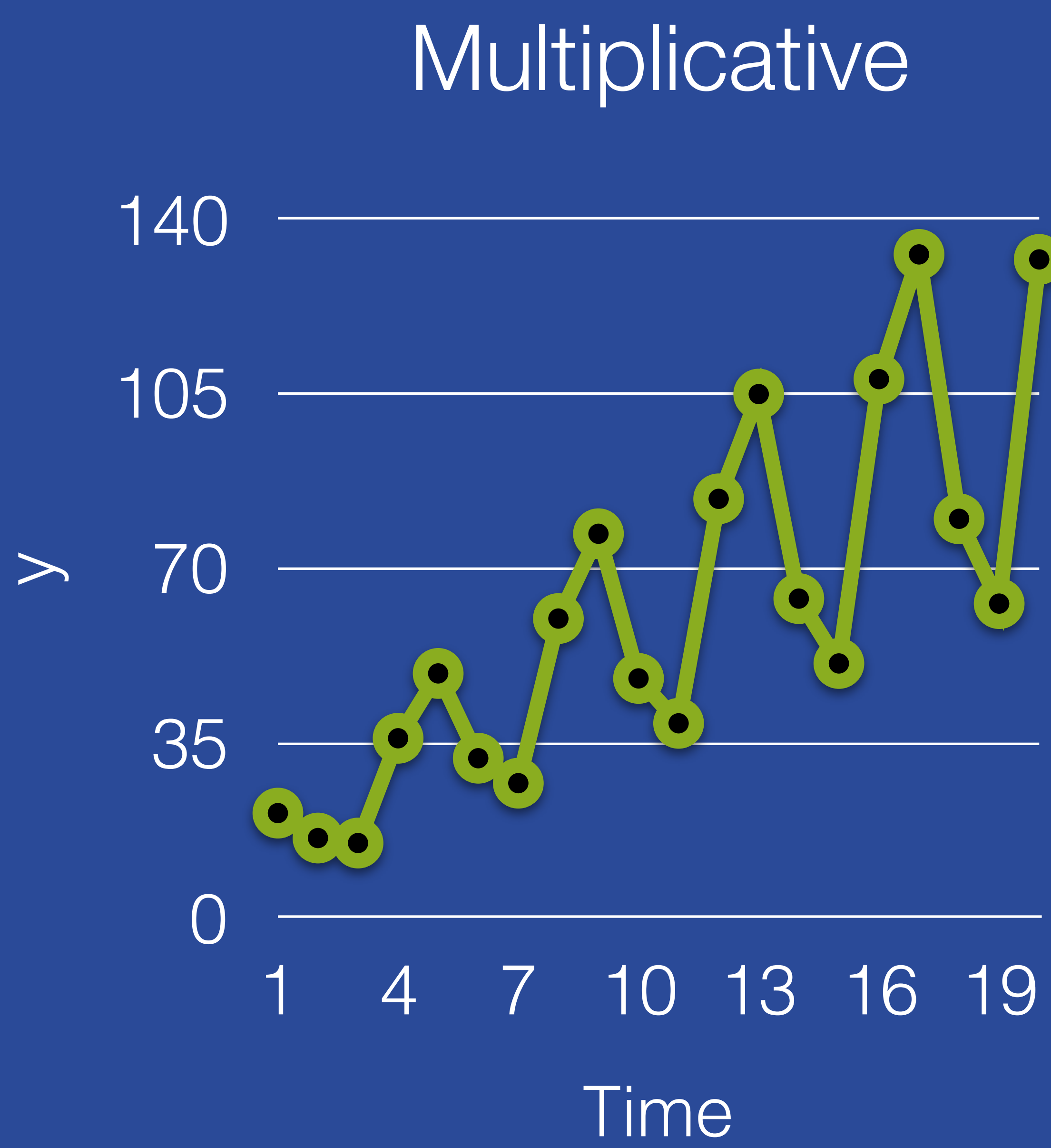




Seasonality

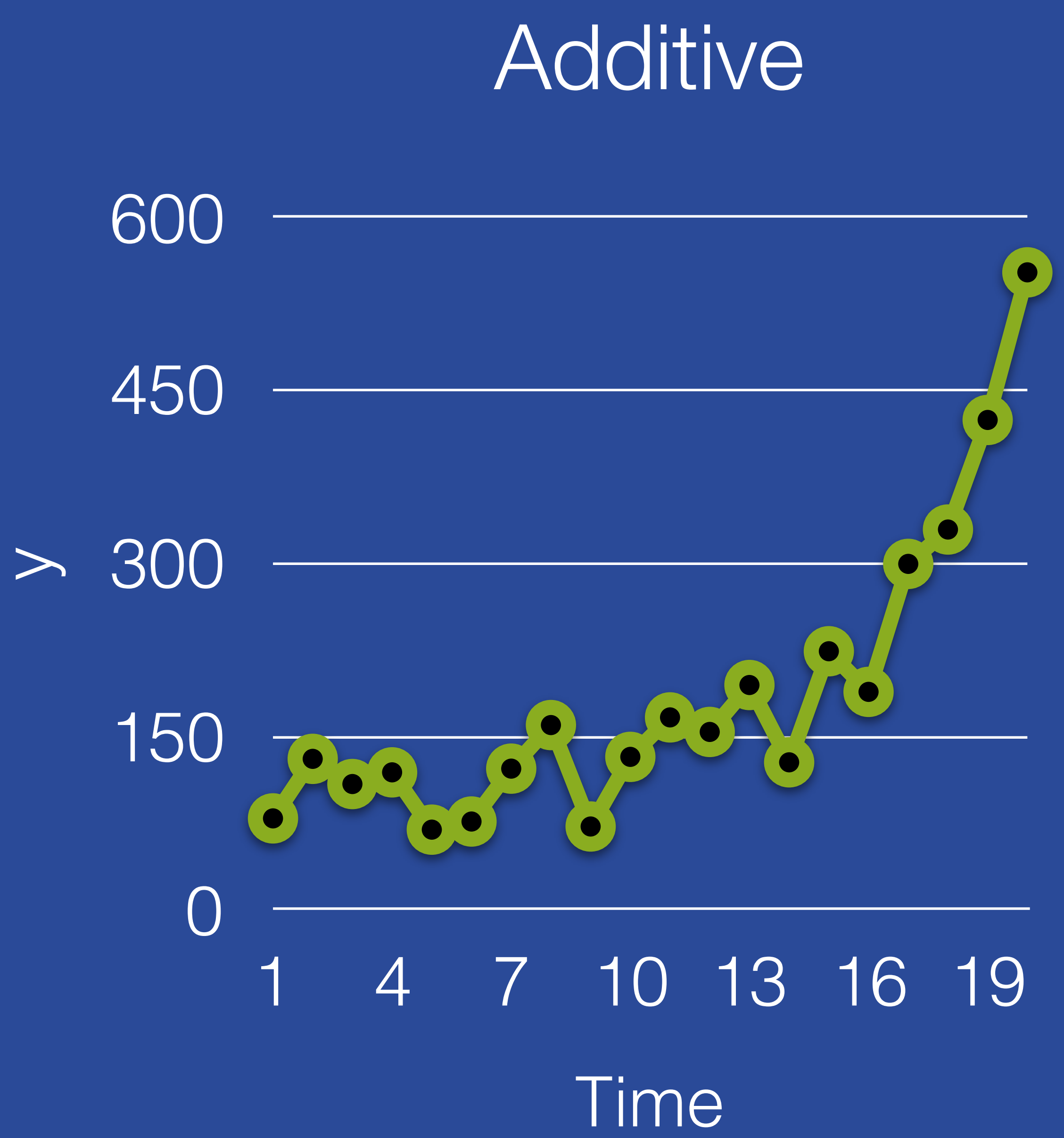
Seasonality: A recurring shorter-term pattern

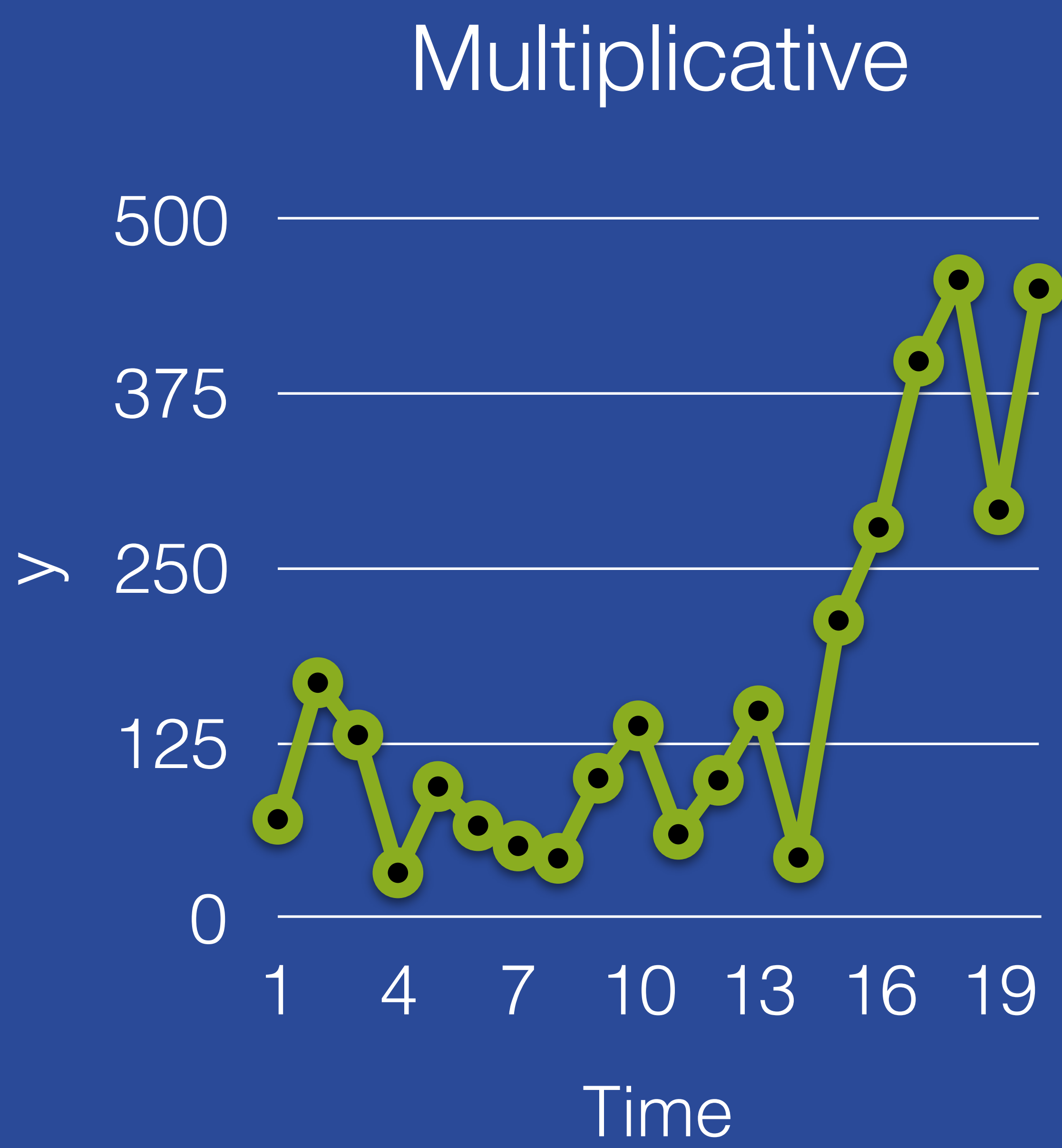




Error

Error: Cumulative error from the smoothing





Time Series Model Matrix

These can all be modeled with time series as well!

M,N,A

Multiplicative Error

No Trend

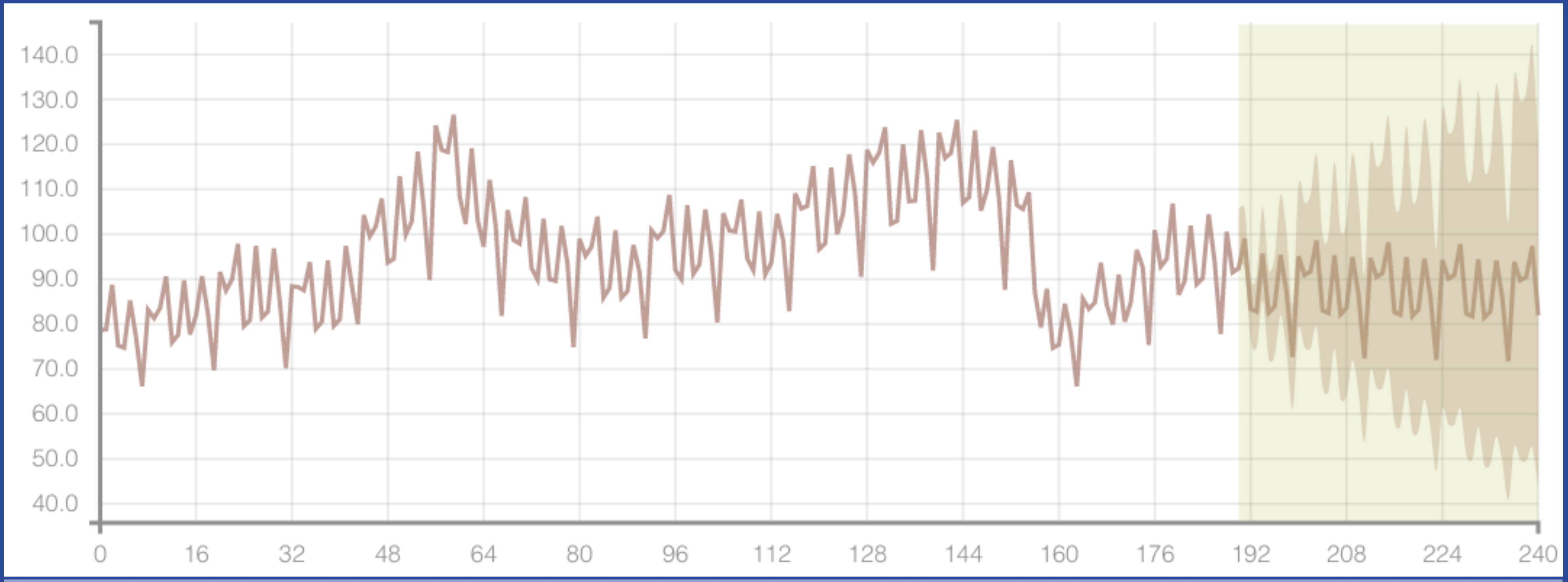
Additive Seasonality

	None		Additive		Multiplicative	
None	A,N,N	M,N,N	A,N,A	M,N,A	A,N,M	M,N,M
Additive	A,A,N	M,A,N	A,A,A	M,A,A	A,A,M	M,A,M
Additive + Damped	A,Ad,N	M,Ad,N	A,Ad,A	M,Ad,A	A,Ad,M	M,Ad,M
Multiplicative	A,M,N	M,M,N	A,M,A	M,M,A	A,M,M	M,M,M
Multiplicative + Damped	A,Md,N	M,Md,N	A,Md,A	M,Md,A	A,Md,M	M,Md,M

- Question: Which one works best?

Time Series Forecasts

Use the data from the past to predict the future

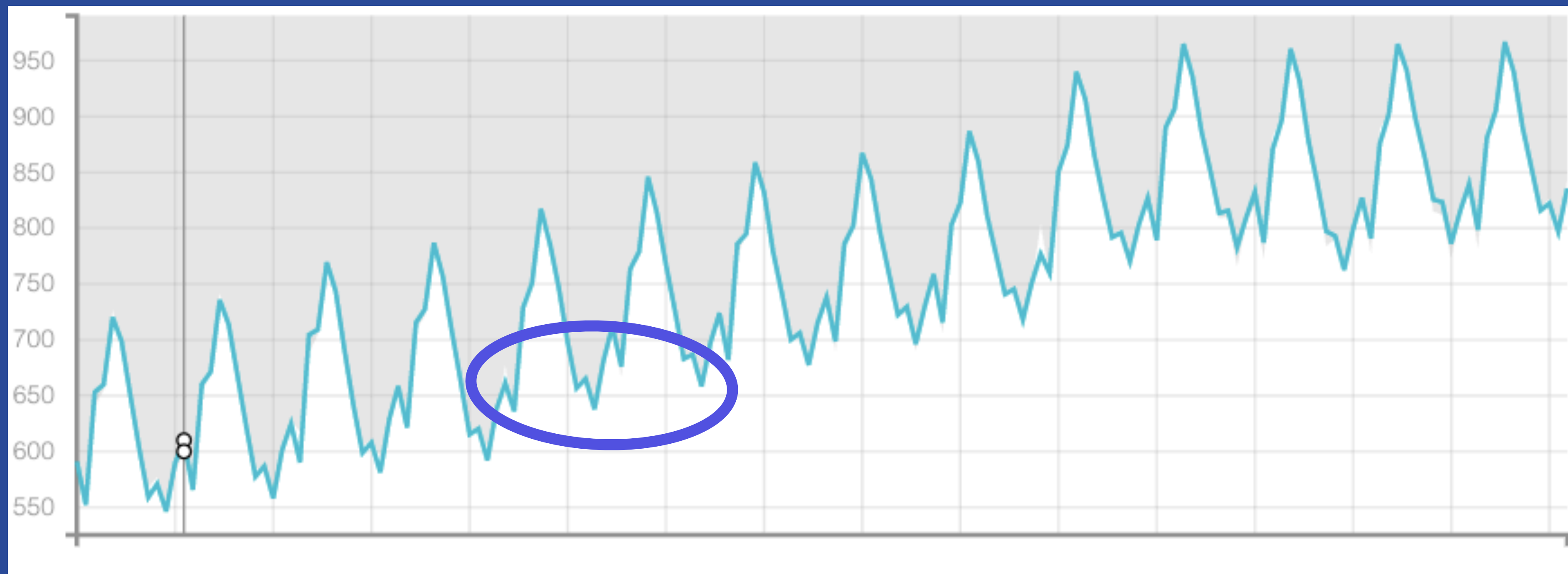


Time Series Demo #1

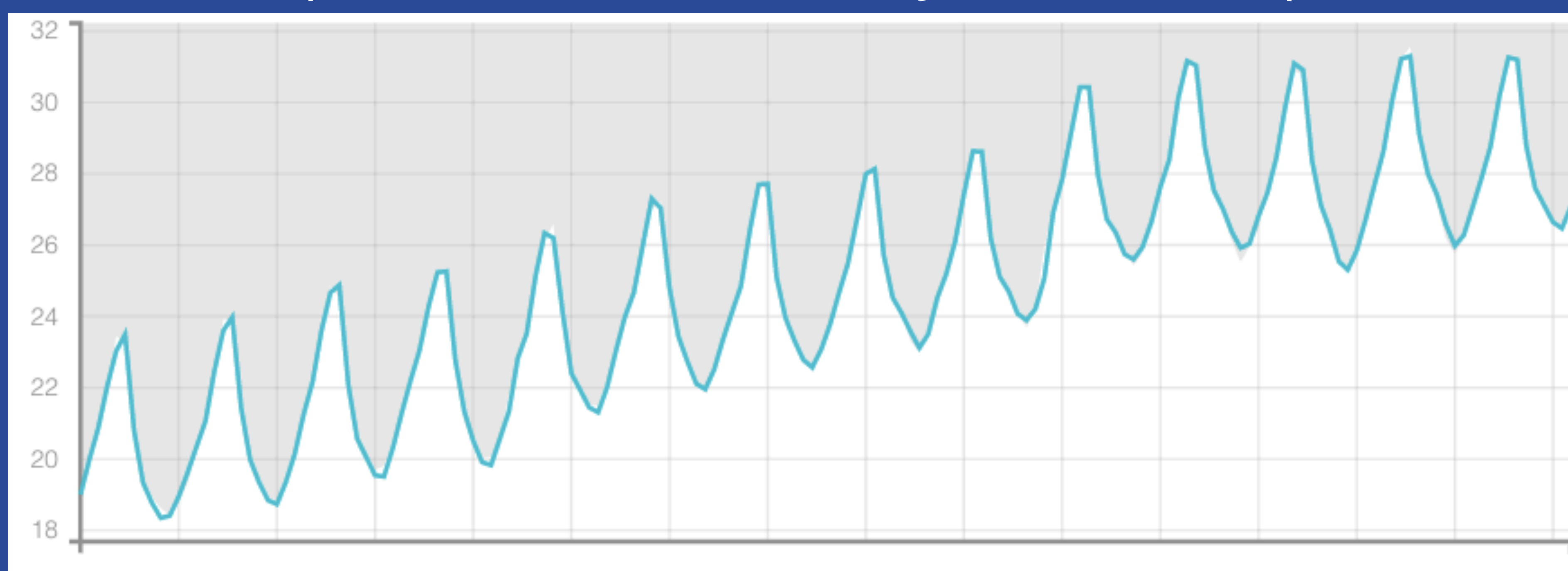
Your Turn!

- Upload the Milk Production Source
- Create a Dataset and a 1-click Time Series
- Forecast the monthly milk production for 50 months

Calendar Correction



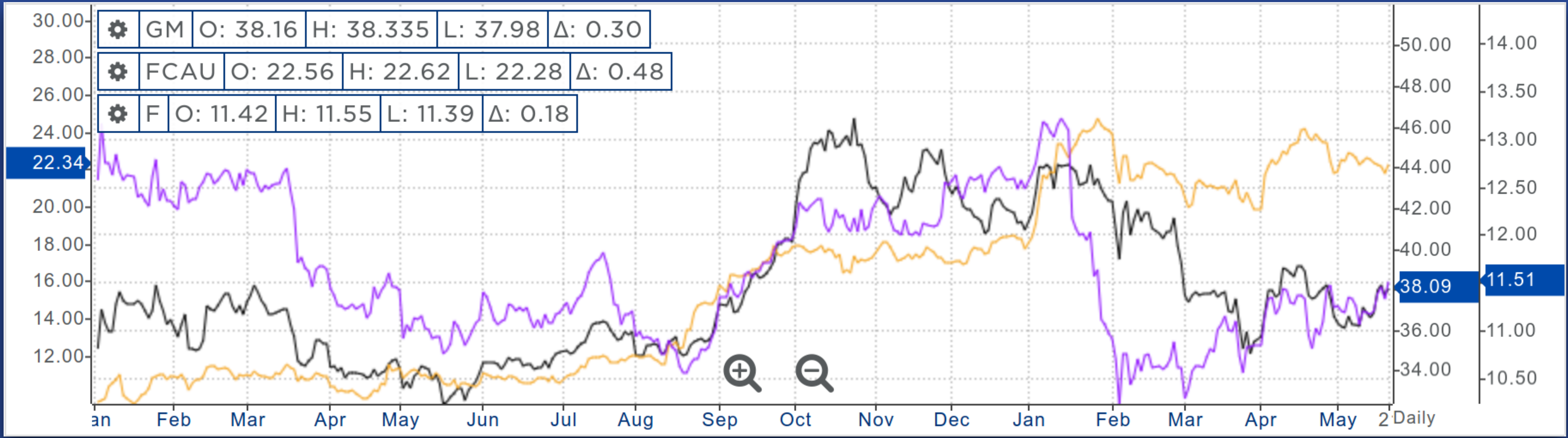
- Time Series data can show variations due to aggregation
- For example: “pounds/month” produced
- Transform: pounds/month ÷ days/month = pounds/day



Calendar Correction

Multi-Variate TimeSeries

- We built a Time Series that predicted two objectives - but this is not multi-variate time series.
- A "convenience" feature. Results are identical to fitting separate individual time series models
- Planned for a future release!



bigml®