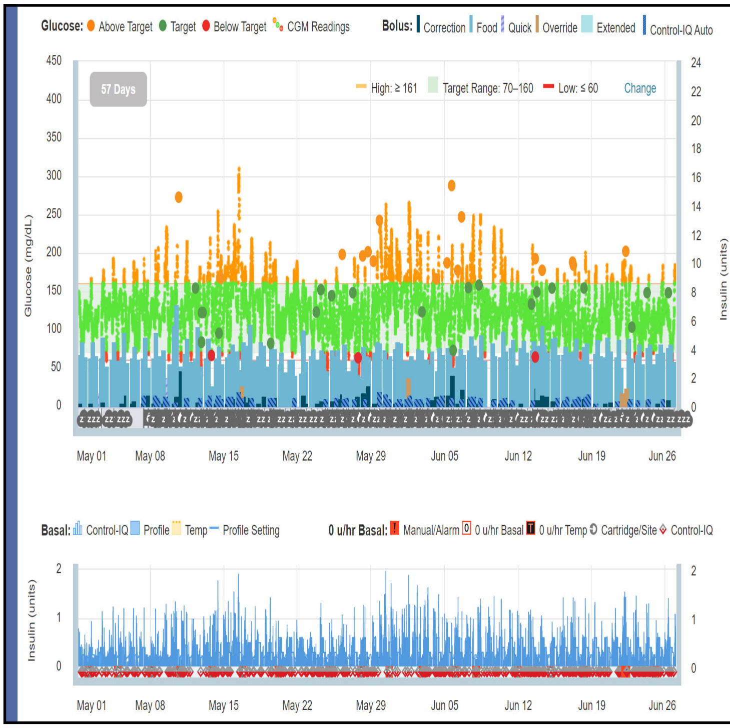
Title

CGM and insulin pump data to introduce classical and machine learning time series analysis concepts to students

Author
Sanchez, Juana

Publication Date
2021-08-10

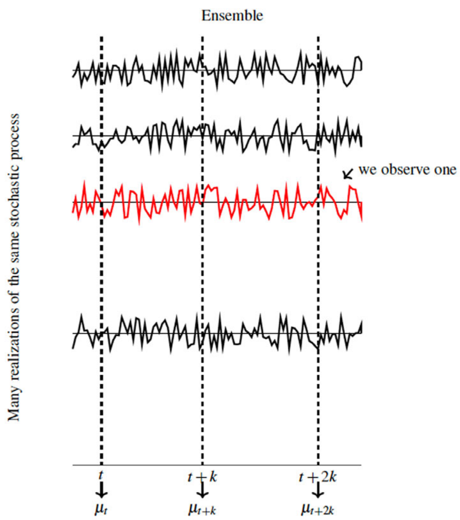


CGM and insulin pump data to introduce classical and machine learning time series analysis concepts to students

Juana Sanchez
UCLA Department of Statistics
Joint Statistical Meetings, August 2021
github.com/juanasanchez
jsanchez@stat.ucla.edu

- **Time series data analysis: rarely taught to undergraduates in college.**
- **Many students graduate with:**
  - **the i.i.d mindset,**
  - **without ever having corrected for autocorrelation in regression,**
  - **without ever using their regression models to predict out-of-sample.**
- **In the rare cases where an upper division elective time series course is offered to juniors and seniors,**
  - **the iid mindset interferes with the learning of basic time series concepts,**
  - **students have to invest a large amount of time learning concepts that could have easily been taught at the intro stats level.**
- **The volume, velocity and variety of timestamped data (smart cities, medical devices, finance, economy, climate, water quality, energy…) is making it increasingly necessary to include at least some basic time series education in the intro stats course.**

2



For example, the red one could be the blood glucose sequence observed for DDD (Diabetic Data Donor) between May 1st and June 26th 2021.

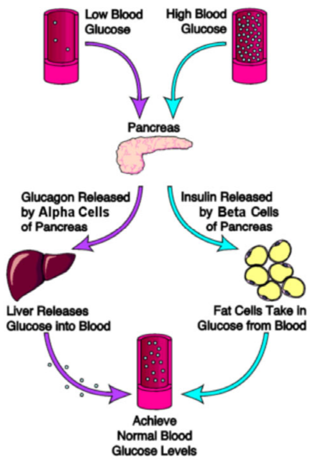
A statistical time series is a random realization of one of the many possible sequences of values of a variable generated by a stochastic model.

This talk is a case study involving data produced by closed-loop technogy for the management of T1D of the DDD. Follows GAISE recommendations.

1. The context: Blood glucose regulation for a T1D (Type I Diabetic) person
2. The multivariate and real timestamped sensor data donated by a DDD is produced by the technology with a purpose: to help non-statistician health care providers routinely guide DDD's health care management.
3. Our objective is to use this context and data to engage students in the discovery of time series features by means of basic graphs and summary statistics that they learn in an introductory statistics course, by investigations and critical thinking.

4

1. The context: blood glucose regulation

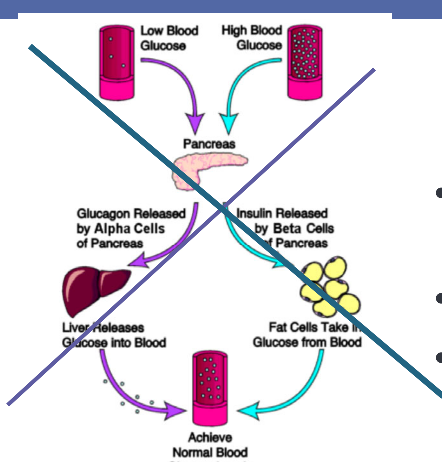


Low and high blood sugar are self-regulated by the pancreas. When the pancreas functions well:

- If high blood sugar level: Insulin released by the Beta Cells of Pancreas makes sugar go to cells to produce energy. This helps maintain normal sugar levels in the blood.
- If low blood sugar level: Glucagon released by Alpha Cells of Pancreas makes the liver release glucose to the blood. This helps maintain normal sugar levels in the blood. Not desirable even in normal persons. Better eat that cookie.

5

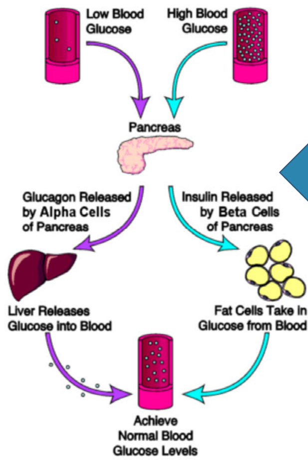
For a T1D (for 1.6 million people in the US of every age, race, shape, and size)



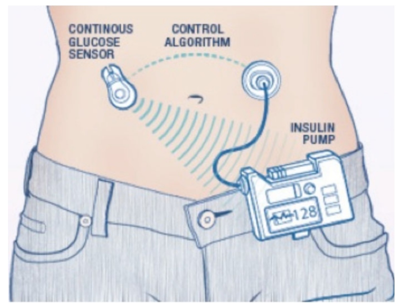
- **Normal blood glucose self regulation does not work.**
- **All the blood sugar stays in the blood.**
- **Artificial insulin is needed.**

6

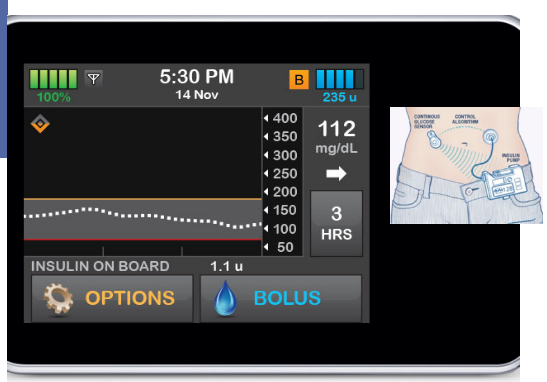
Since Jan 2020 close-loop system technology adjusts insulin levels and provides readings of interstitial blood glucose. The DDD in this case study uses the technology.



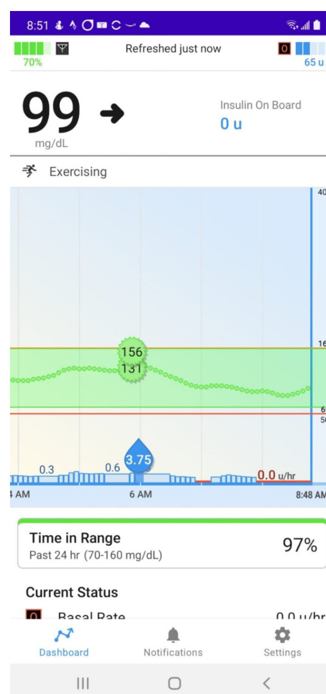
Objective: mimic normal blood glucose control



DDD's view of the data produced by the technology



DDD can view sensor glucose readings in the insulin pump display up to 24 hours

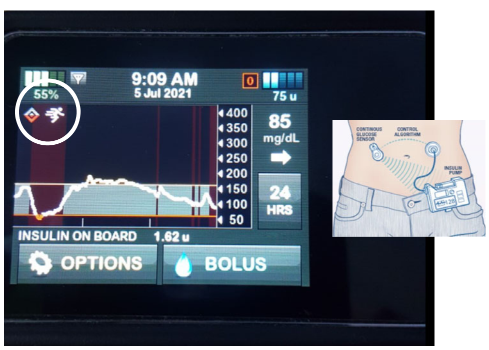


DDD can also view sensor glucose readings in the smart phone

DDD can see more than glucose readings

DDD also validates cgm readings by periodically checking "true" blood sugar manually and periodically intervenes.

The algorithm is unknown



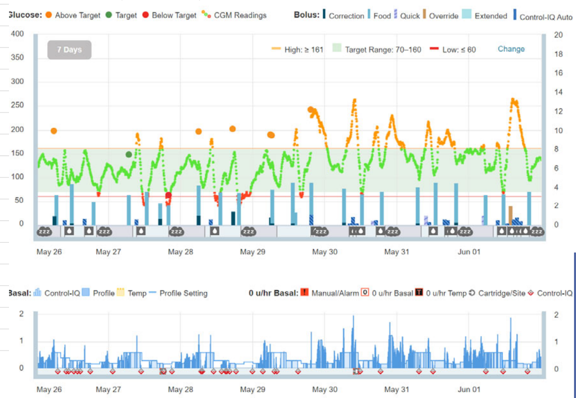
The pump seems to act based on predictions of blood glucose for the next half hour. When sleeping, that works very well. When awake, with unexpected events, body reactions, food, etc. it gets complicated. DDD would like to understand the secret algorithm inside the technology better.

2

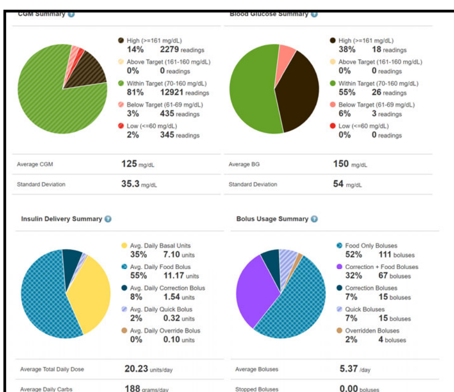
EventDateTime	Readings (CGM / BGM)
2021-05-01T00:02:00	122
2021-05-01T00:07:00	125
2021-05-01T00:12:00	127
2021-05-01T00:17:00	129
2021-05-01T00:22:00	132
2021-05-01T00:27:00	134
2021-05-01T00:32:00	133
2021-05-01T00:37:00	132
2021-05-01T00:42:00	133
2021-05-01T00:47:00	133
2021-05-01T00:52:00	133
2021-05-01T00:57:00	
2021-05-01T01:02:00	
2021-05-01T01:07:00	
2021-05-01T01:12:00	
2021-05-01T01:17:00	

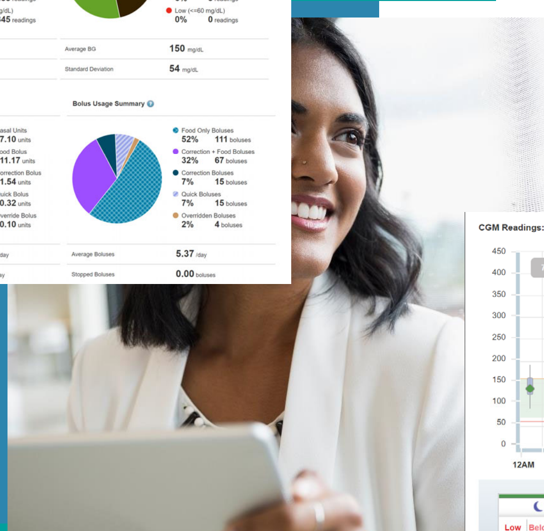
The technology stores DDD's timestamped data and plots.

IOB	81	2021-05-01T05:40:17	0.28
IOB	81	2021-05-01T05:50:18	0.25
IOB	81	2021-05-01T06:00:17	0.21
IOB	9	2021-05-01T06:05:40	0.1
IOB	9	2021-05-01T06:05:49	0.1
IOB	81	2021-05-01T06:10:17	0.19
IOB	9	2021-05-01T06:15:22	0.1
IOB	16	2021-05-01T06:15:23	0.19
IOB	64	2021-05-01T06:15:41	0.19
IOB	55	2021-05-01T06:15:56	0.19
IOB	20	2021-05-01T06:17:12	3.8
IOB	81	2021-05-01T06:17:12	0.19
IOB	81	2021-05-01T06:20:17	3.82
IOB	81	2021-05-01T06:30:17	3.75
IOB	81	2021-05-01T06:40:17	3.64
IOB	81	2021-05-01T06:50:17	3.43
IOB	81	2021-05-01T07:00:17	3.22
IOB	81	2021-05-01T07:10:18	2.97
IOB	81	2021-05-01T07:20:17	2.68

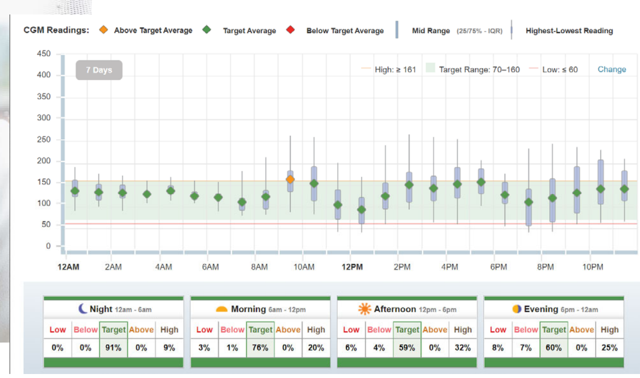


11

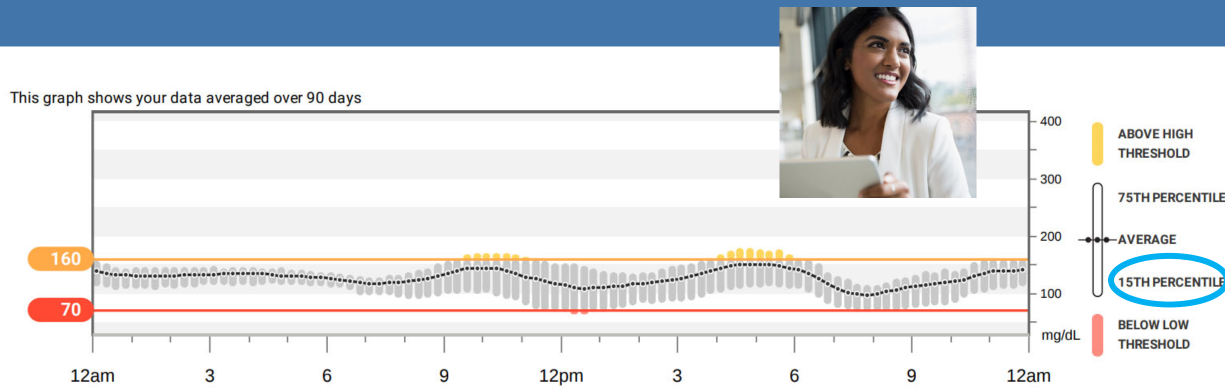




The purpose of the data storage: to allow non-statistician nurses and doctors inspect all the plots and summaries every three months to guide DDD's adjustments to therapy.

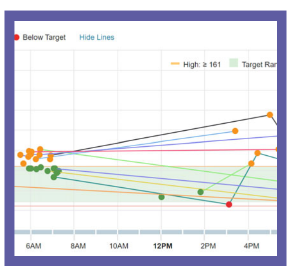


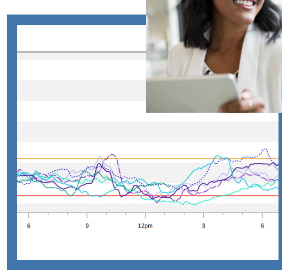
12

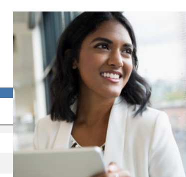


Graphs and summary statistics are not quite those taught at the intro stats level.

Lots of time series graphs and data stored in the technology's server







Hourly average glucose readings summarized by hour of the day.

Something close to seasonal box plots

Some kind of trends

Something close to trends during the day

Spaghetti plots

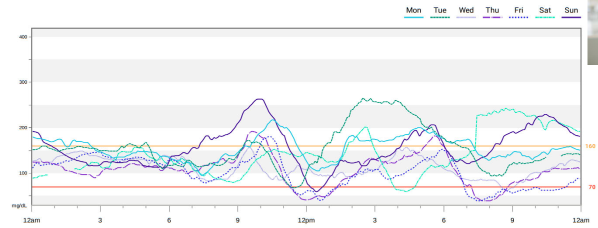
Something close to spaghetti plots

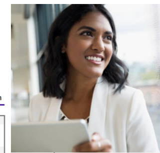
Streaming data.

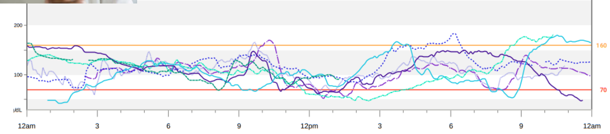
Something like multivariate time plots, customized, but not line plots.

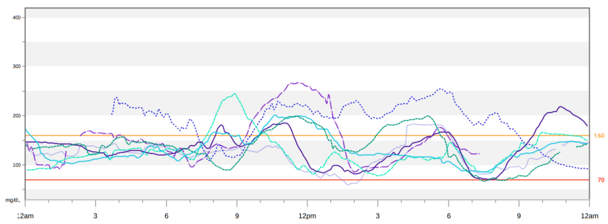
Not all weeks are the same, not all days are the same

In some weeks, all days are roller-coasters



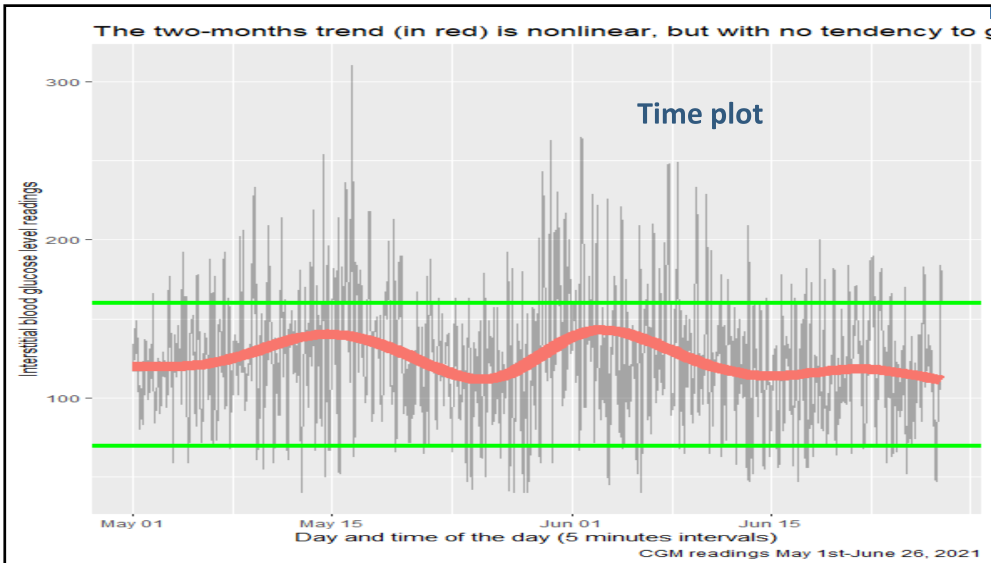






There is a rich story to tell students with just blood glucose built-in graphs. We analyze 16017 observations of blood glucose of May 1st-June 26, 2021 and 9055 observations of insulin interventions during the same period.

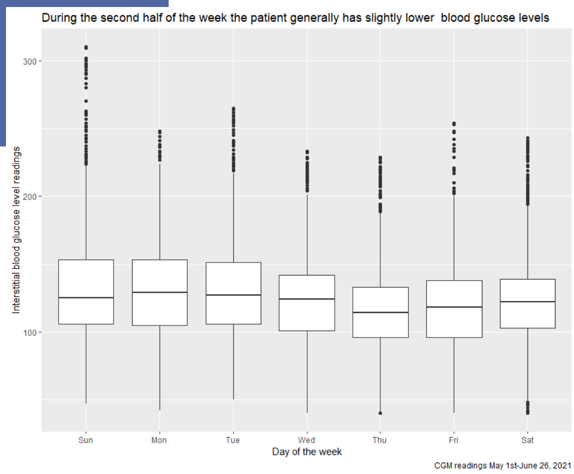
15



3. Teaching basic time series concepts in the intro stats class with R

Non-linear trend according to a moving average smoother

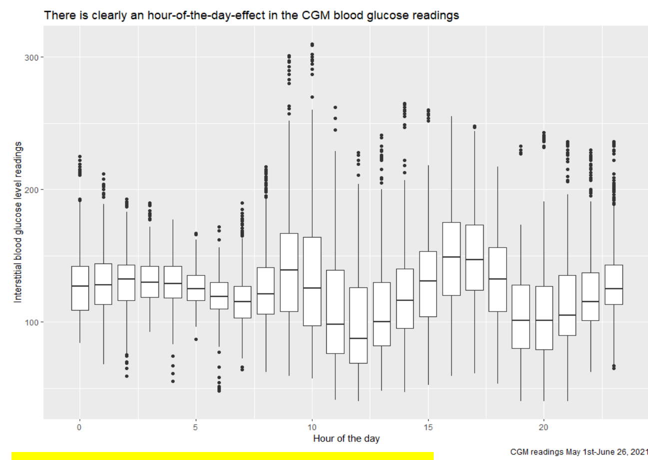
Q. What does this plot reveal about the technology? About DDD? Would you expect this trend in a time series of household electricity use recorded at 5 minute intervals?



Weekly seasonality. Median lower on thursday-Sat seen with a **seasonal box plot**

Q. How does it differ from weekly seasonality of household electricity use?

Daily seasonality seen with seasonal box plot of DDD's blood glucose.



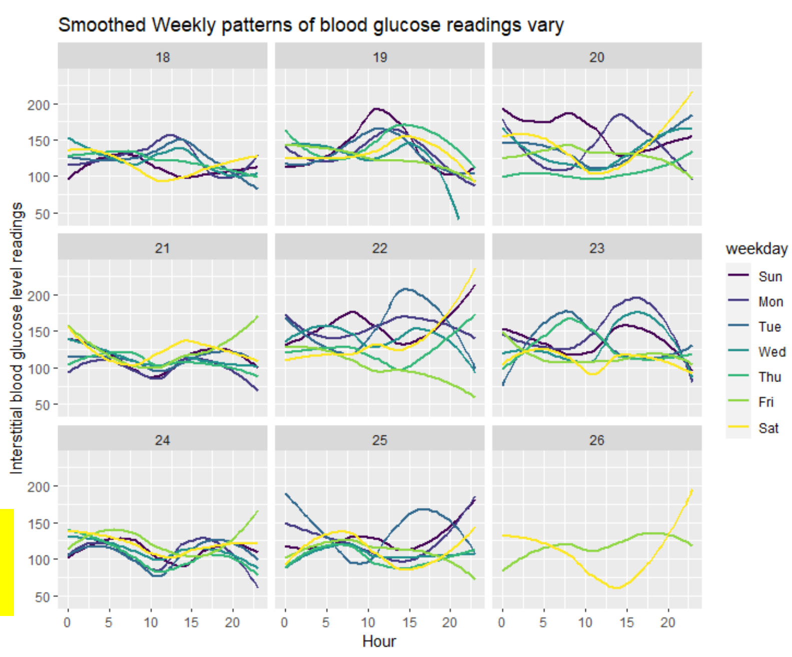
Q. How does it differ from daily seasonality of household electricity use?

17

Daily seasonality different each week

- The daily seasonality pattern that we noticed in the seasonal box plot, when viewed week by week and day of the week by day of the week, does not appear so regular.
- Some weeks show more variability than others.

Q. What variables could help explain the differences in the weeks.

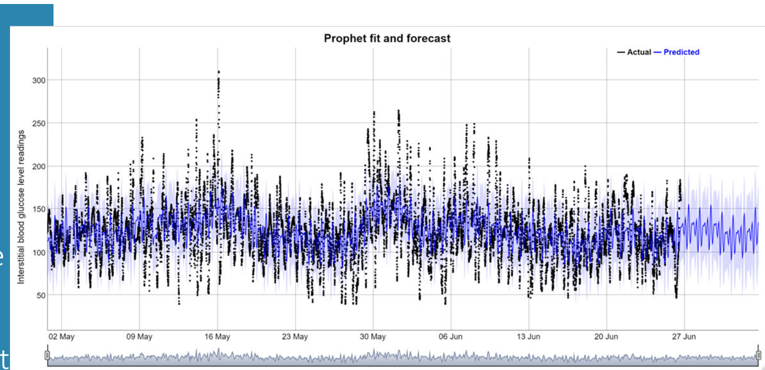


18

Forecasting like analysts in the loop with Prophet

- Cannot expect intro stats students to forecast with sophisticated time series models.
- But students without any time series training work sometimes as analysts in the loo (use a proprietary automated forecasting routine). Facebooks' prophet is an example.

Q. What do you observe about the forecasts out-of-sample as compared with the training data?



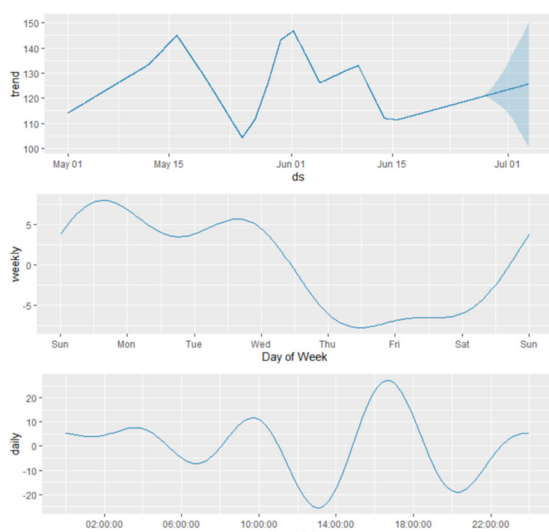
Glucose = trend + hour +day of week+holidays

Forecasting like analysts in the loop with Prophet

- Prophet models the components of the time series that we analyzed with our plots.
- Glucose = trend + hour +day of week+holidays

Q. What is happening to our uncertainty when we forecast farther ahead?

Another way to do time series decomposition.



Machine learning: features engineering and clustering to be able to combine insulin and glucose data

- We involve now the insulin interventions by the pump and DDD, but because the recording time interval is different than for glucose we use features engineering :

- **Use summaries familiar to the intro stats student: interquartile range, min, max, median, frequencies of different types of boluses.**
- **By the hour.**

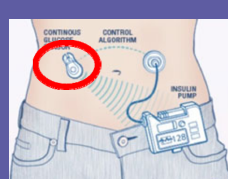
```
# A tibble: 1,367 x 20
# Groups:   day [57]
     day  hour iob_med_hour iob_sd_hour iob_IQR_hour iob_min_hour iob_max_hour iobmonth iobday  n_81   n_9  n_16  n_20  n_55  n_64 cgm_med_hour
   <dbl> <int>        <dbl>       <dbl>        <dbl>        <dbl>        <dbl>    <dbl>  <int> <int> <int> <int> <int> <int> <int>        <dbl>
 1   121     0        0.1       0.0742       0.128          0.01         0.18        5      1     6     0     0     0     0     0         132
 2   121     1        0.15      0.0186       0.0225         0.12         0.17        5      1     6     0     0     0     0     0         128.
 3   121     2        0.26      0.0235       0              0.21         0.28        5      1     6     0     0     0     0     0         138
 4   121     3        0.325     0.00816      0.0100         0.32         0.34        5      1     6     0     0     0     0     0         143
 5   121     4        0.385     0.0387       0.0600         0.32         0.41        5      1     6     0     0     0     0     0         148.
 6   121     5        0.33      0.0520       0.0775         0.25         0.38        5      1     6     0     0     0     0     0         138.
 7   121     6        0.19      1.76         3.40           0.1          3.82        5      1     7     3     1     1     1     1         135
 8   121     7        2.54      0.505        0.697          1.89         3.22        5      1     6     0     0     0     0     0         122.
 9   121     8        1.01      0.420        0.595          0.53         1.62        5      1     6     0     0     0     0     0         106
10   121     9        0.19      0.136        0.18           0.03         0.37        5      1     5     0     0     0     0     0         105
# ... with 1,357 more rows, and 4 more variables: cgm_sd_hour <dbl>, cgm_IQR_hour <dbl>, cgm_min_hour <dbl>, cgm_max_hour <dbl>
```

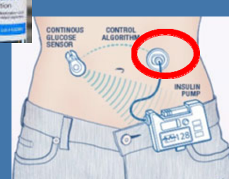

Clustering of features

Q. What explains the differences in the clusters of the hours? Investigate using data not used in clustering.

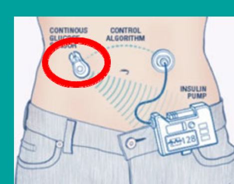

```
iob_med_hour
   0.8557059
   0.6131099
   1.1281836
```


```
cgm_med_hour
    88.41529
   127.66491
   177.18164
```



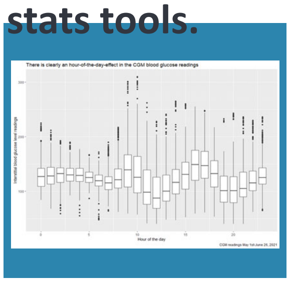



```
iob_IQR_hour
   0.6495647
   0.4184902
   0.4602930
```

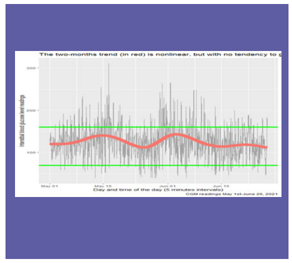



```
cgm_IQR_hour
   10.304118
    9.724774
   13.486328
```

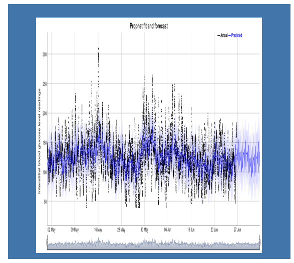
Conclusion: the case study engages students and makes them use the tools they know to investigate a complex process. At the same time, they learn basic time series concepts using only their intro stats tools.



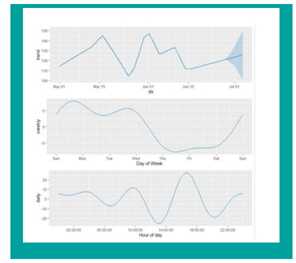
The concept of seasonality



The concept of long term trend



The concept of forecast



The concept of decomposition of a time series into its components

And with all of the above, an intuition for the concept of autocorrelation is gained before introducing the ACF and PACF.

Of course, having time series for more than one DDD would make the study more interesting. After all, the technology designers are inspired by time series like ours of thousands of individuals and design to target the average individual.

Thank you for your attention

The paper (with references), simple R programs and data for this talk can be found at github/juanasanchez shortly after the JSM meetings.
jsanchez@stat.ucla.edu

I thank the DDD for the data and for all the information provided that helped me understand the data. Without the DDD's help I would not have been able to complete this presentation.