The data you were promised... and the data that you got

A story about Reykjavík's Thermal Pools

Hlynur Hallgrímsson
The City of Reykjavík

November 9th, 2021

Who am I?

- **Hlynur Hallgrímsson**
  Senior Data Scientist
  Office of Data Services
  City of Reykjavík
- **Economist in a past life**
  Background in economics and econometrics, but I've been working as a data scientist since 2016 (I was a data analyst before that).
- **The data science team is new.**
  Created at the end of 2019 (I joined in March of 2020).
  We're a small team. There's only 6 of us.
- **What we do**
  We create data products for other departments. Often times this means doing the data work from start to finish, but as there are other analysts, statisticians and data scientists working throughout the City of Reykjavík, we're sometimes brought in to help them. For instance, taking their analysis to the next level by either reimplementing at a larger scale, or putting them into production.

What is this talk about?

It's about swimming pools!

The population of Iceland is 375.000 people. The City of Reykjavík is home to 135.000 of them, or about 36%[1]. A large share of those people is **crazy** about swimming pools. So the city runs seven public geo-thermal swimming pools, six of which are open 363 days a year[2].

But mostly it's about how data science is, like... really hard... 😭

Ideally this talk would be titled *"The good life: How data science becomes super easy when you have the right tools!"*. But, as you might know, **the hardest thing in data science isn't really any specific data science thing**. In my experience the hardest things are stuff like group intercommunication and project logistics (like dealing with vendors 🙃).

[1] The greater Reykjavík area (a.k.a. the capital region) actually accounts for 64% of all Icelanders (240.000).

[2] The pools are closed on Christmas Day and New Year's Day.

But what's the deal with these pools?

- Iceland has abundant geothermal resources. Water doesn't need to be heated, we just pump hot water out of the ground at 80°C/176°F. So, hot water in Iceland is incredibly cheap.
- Iceland is a fishing nation. When every **1 out of 5 people works in fishing**, as was the case on average for the whole of the 20th century, it's really important that those people know how to swim. So every able-bodied person that goes through the Icelandic school system knows how to swim - as we have to take mandatory swimming lessons once a week through 1st to 10th grade (that's roughly **400 hours of *mandatory* swimming**).
- **It's cold and it's dark.**
  Iceland is cold, and since sunset is just 4 hours after sunrise during the darkest winter days in Reykjavík (as little as 2 hours in more northerly towns) we don't get enough Vitamin D from sunlight. But we can fight that by laying in our bathing suits in pools and hot tubs, with hot water protecting us from the freezing wind - our skin grasping on to what little sun light there is to go around.
- So, people go there to swim and to enjoy some sunlight in a 42°C/108°F hot tub... but there's a third reason. They also go there to socialize, meet people, talk politics and gossip.

A thermal pool data science project

The idea: to show which swimming pools are crowded and which ones are not

- Since the swimming pools are popular, citizens have contacted the city offices to ask if there's any way to see which pools are crowded and which ones have few patrons at the moment.
- Welfare workers have also expressed an interest in such a solution (for clients that are susceptible to sensory overload)
- A neighbouring town, Hafnarfjörður, a suburb of Reykjavík, now offers a website dashboard which shows you how crowded their two swimming pools currently are. They have their staff manually count patrons.
- During Covid, swimming pools have had to limit availability to 50% of normal capacity for certain periods (this was the final catalyst).

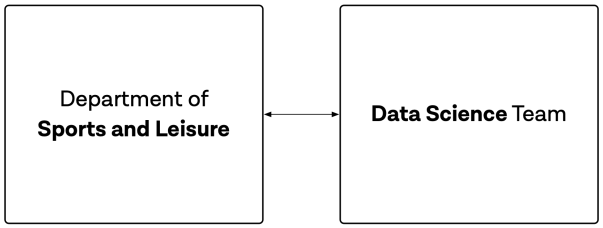
The pitch (*rephrased*)

"We have updated the gates at our swimming pools to modern electronic turnstiles. With these modern gates we can now get information on how many people are at each swimming pool through an API (application programming interface).

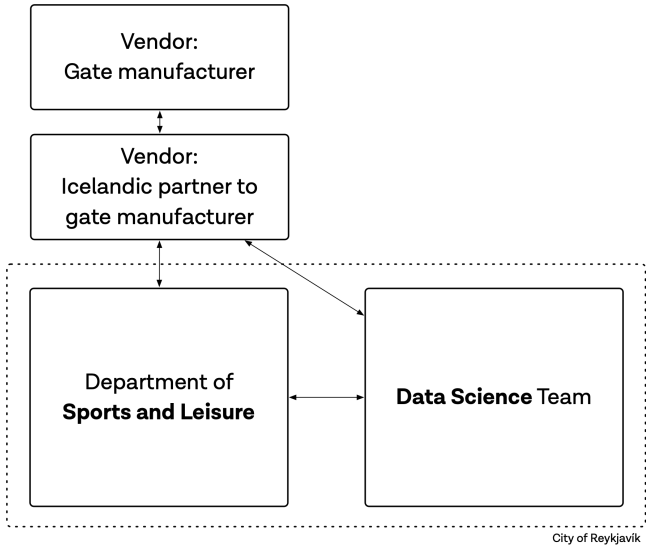
Can you help us present this data - showing how crowded each pool is?"

- Department of Sports and Leisure (*not a direct quote*)

So we're in contact with Sports and Leisure



But it's actually more complicated than that



"But still, should be pretty straight forward once we get the data, right?" 🤡

How I envision the project:

1. Read data from API using *{httr}*
2. (Surely) Do some data cleaning with the *{tidyverse}* and *{purrr}*
3. Visualize the data with *{ggplot2}*
4. Make that visualization interactive with *{ggiraph}* or *{plotly}*
5. Make that into a reactive *{shiny}* app and deploy it to *shinyapps.io* or *RStudio Connect*

This is of course a tragic miscalculation on my part because:

a) I'm assuming that getting the data will take a somewhat reasonable time

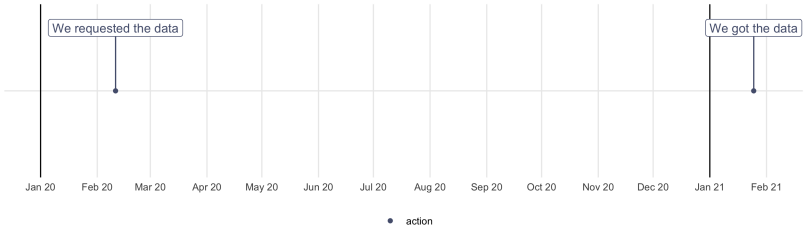
b) I'm assuming lots of things about the data and the API (see five item list above)

A little bit about why simple things can take a long time

- Priorities
- Group intercommunication
- Forced communication protocols
- Red tape (the worst of which is "lawyer stuff")

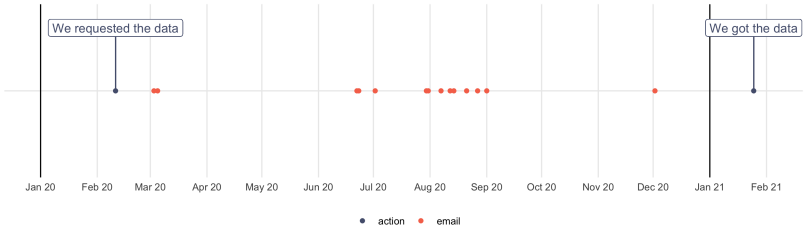
A little bit about why simple things can take a long time

The timeline for getting API access to swimming pool gate data



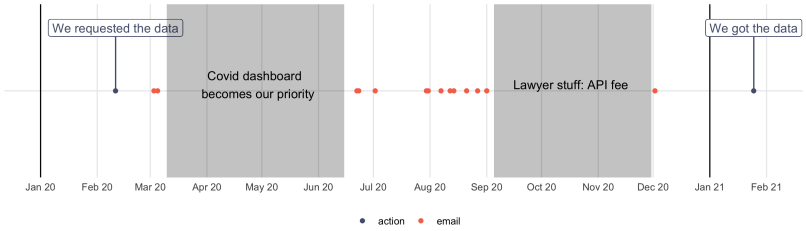
A little bit about why simple things can take a long time

The timeline for getting API access to swimming pool gate data

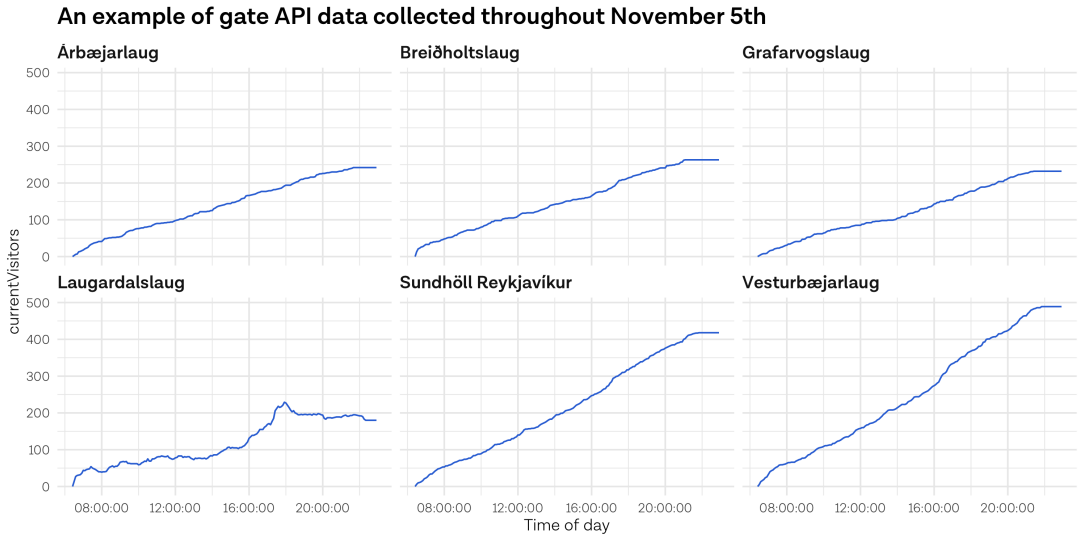


A little bit about why simple things can take a long time

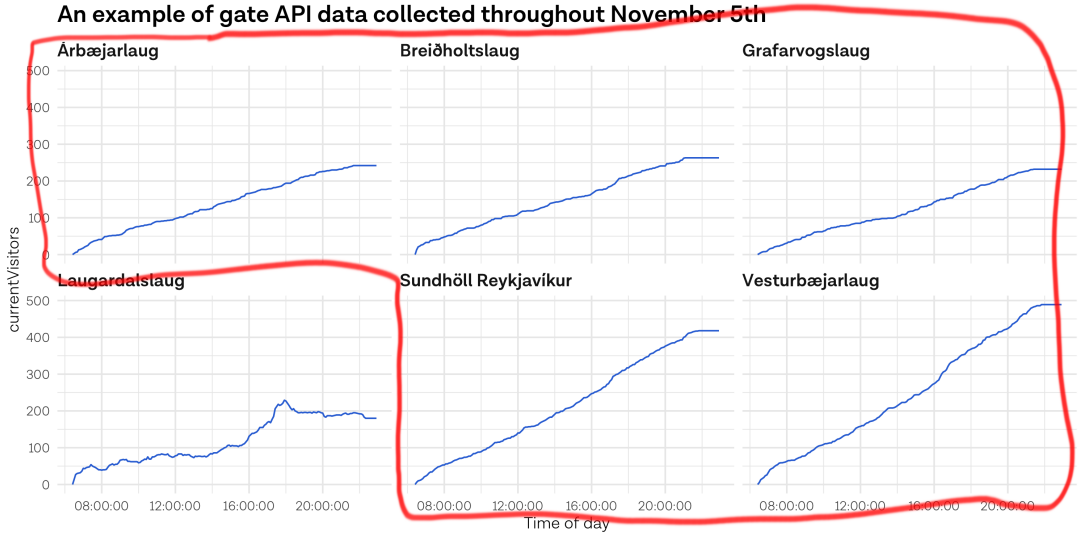
The timeline for getting API access to swimming pool gate data



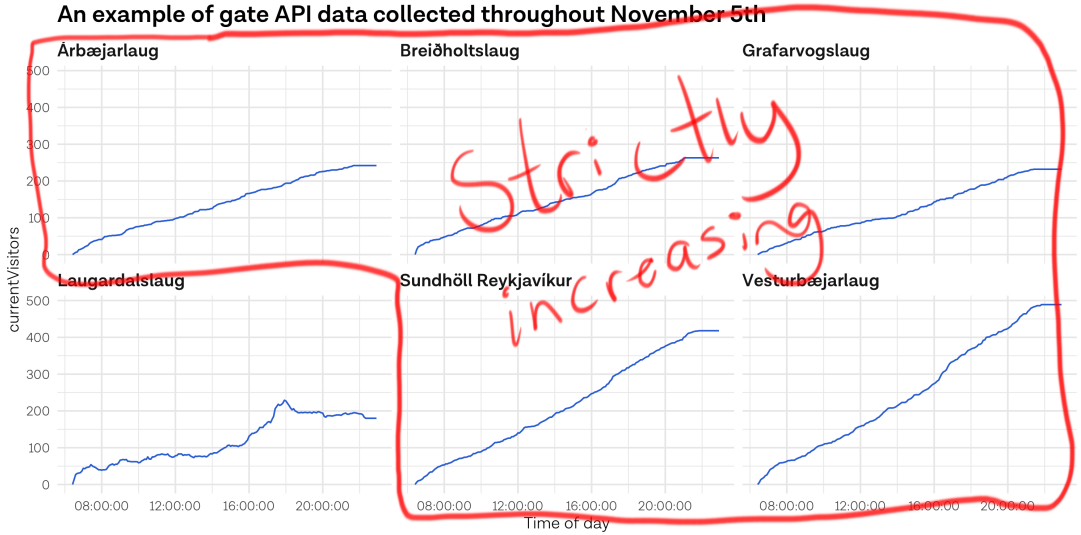
But then you get the data and clean it



But then you get the data and clean it



But then you get the data and clean it



A realization

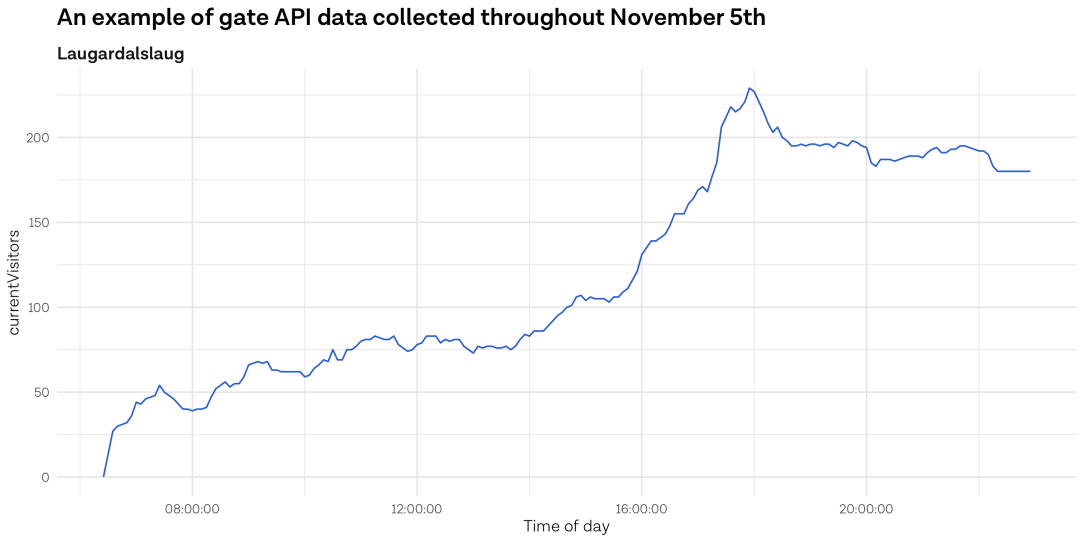
So it turns out

- The gate hardware for five of the six pools we want to show in our app only counts visitors **into** the pool. So *currentVisitors* isn't actually the number of current visitors, but the cumulative sum of total visitors up to the point of the GET request.

But also

- The API can't show us anything except each pool's *currentVisitors* value at the time of the GET request. (So what you saw in the earlier slide is actually the result of sending GET requests at five minute intervals throughout the day).

But wait! There's more! 🎉



But wait! There's more! 🎉

What next?

From the API we don't know when people leave

so we ask:

What if we model the duration?

The idea 💡

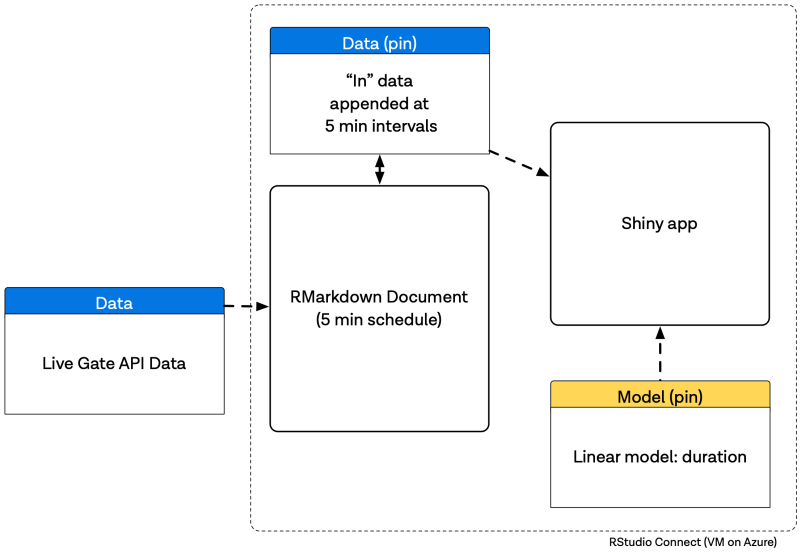
- We request *historical* data through the Icelandic vendor (reminder: Laugardalslaug has *in* and *out* data)
- Assumption 1: Peoples' stays are not inherently different in duration between Laugardalslaug and other pools
- Assumption 2: Peoples' stays are not inherently different in duration between people who are counted out and people who are not counted out

If we get the historical data + assumptions 1 and 2 hold:

- We can train a model on historical Laugardalslaug stay durations of people that are counted out

- And then predict on live API data for other pools

So that's what we did



So that's what we did

On a local machine

1. We read historical data of in/out duration from Google Cloud into R using *{bigrquery}*[1].
2. Fit a few different linear models and evaluate them using *{tidymodels}*; *{yardstick}*[2].
    - **Predictors**:
    - *weekend* - is it the weekend (or a weekday)? 0/1
    - Natural spline term of *hours* (which is hour of the day)[3]
    - *from_solstice* - absolute number of days from the summer solstice
3. We then publish that model to RStudio Connect through the *{pins}* package.

[1] Later we'd turn to RStudio Professional Drivers and *{DBI}*

[2] We ended up with a super simple linear splines regression model... and I LOVE it!

[3] We use 7 knots selected based on staff knowledge but also evaluating different combinations on the training set

So that's what we did

On RStudio Connect

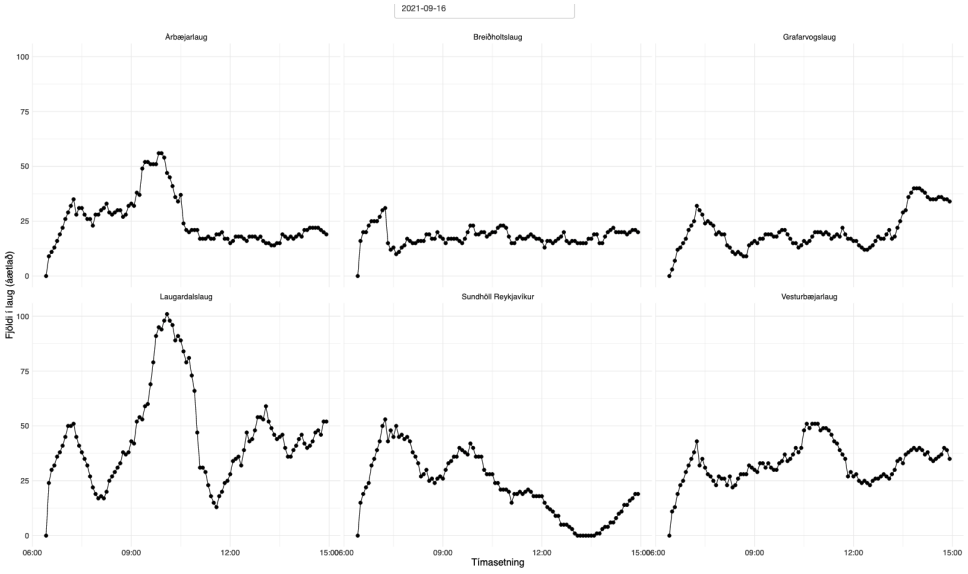
An RMarkdown document is scheduled to do the following every 5 minutes:

- Read data from the API
- Clean that API data and wrangle it into a table
- Read data from the "In" dataset pin
- Append cleaned API data table to the "In dataset"
- Pin (save) the updated "In" dataset to the RStudio board

A Shiny application

- Reads the "In" dataset pin
- Reads the lm model pin
- Runs *predict* for the "In" data to create an estimate of *currentVisitors*

An app for the managers of the swimming pools to evaluate:



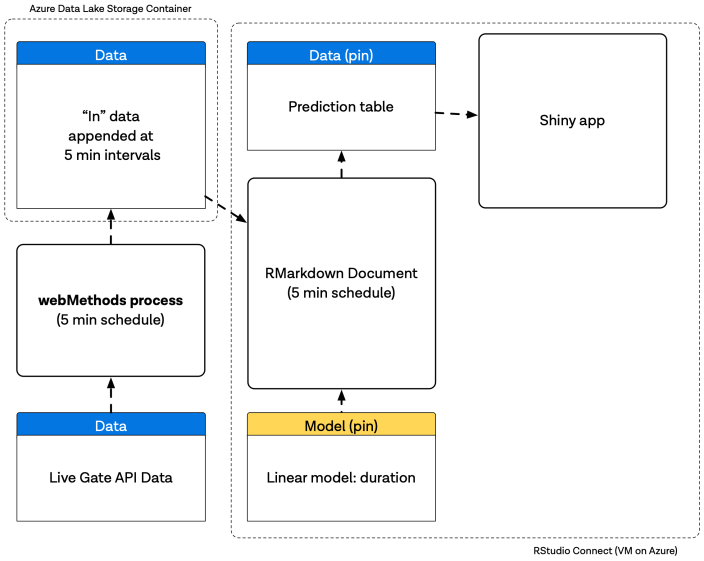
Why is this super cool?

It's all R and RStudio Connect.

So we can move at our own pace - for this part of the project there are no outside constraints or inefficiencies in how we communicate with other groups.

We also get to run into problems ourselves and then we know what to take into account if parts of the process are *outsourced* beyond the data team.

The current setup



But we have to talk about Laugardalslaug

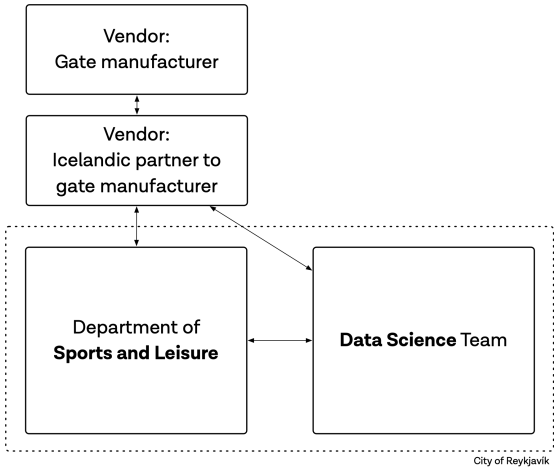
- The data for Laugardalslaug is **absolutely whack!** (technical term)
- And it has everything to do with the kindness of the staff working there!

DAMN YOU KINDNESS! YOU RUINED MY DATA! 😠

Things we're trying

- **Primary**: Ask the vendor to make changes to the API to separate (or show only) **"In"** data for Laugardalslaug
  - We're waiting on this
- **Secondary**: Use a different counter located by the actual pools in Laugardalslaug (not by the entrance)
  - This has proven unfruitful (bidirectional)
- **Secondary**: Use count data from the smart lockers in Laugardalslaug
  - Unclear if the firmware can be updated to the most recent version to support *live* data

BTW. this was a gross oversimplification



Still to-do:

Account for gym patrons that use the swimming pools at:

- Laugardalslaug
- Breiðholtslaug

Suuuuurely, something else!

Thank you

On Linkedin: in/hlynurh (Hlynur Hallgrímsson)

On Twitter: @hlynur