Digital Twin

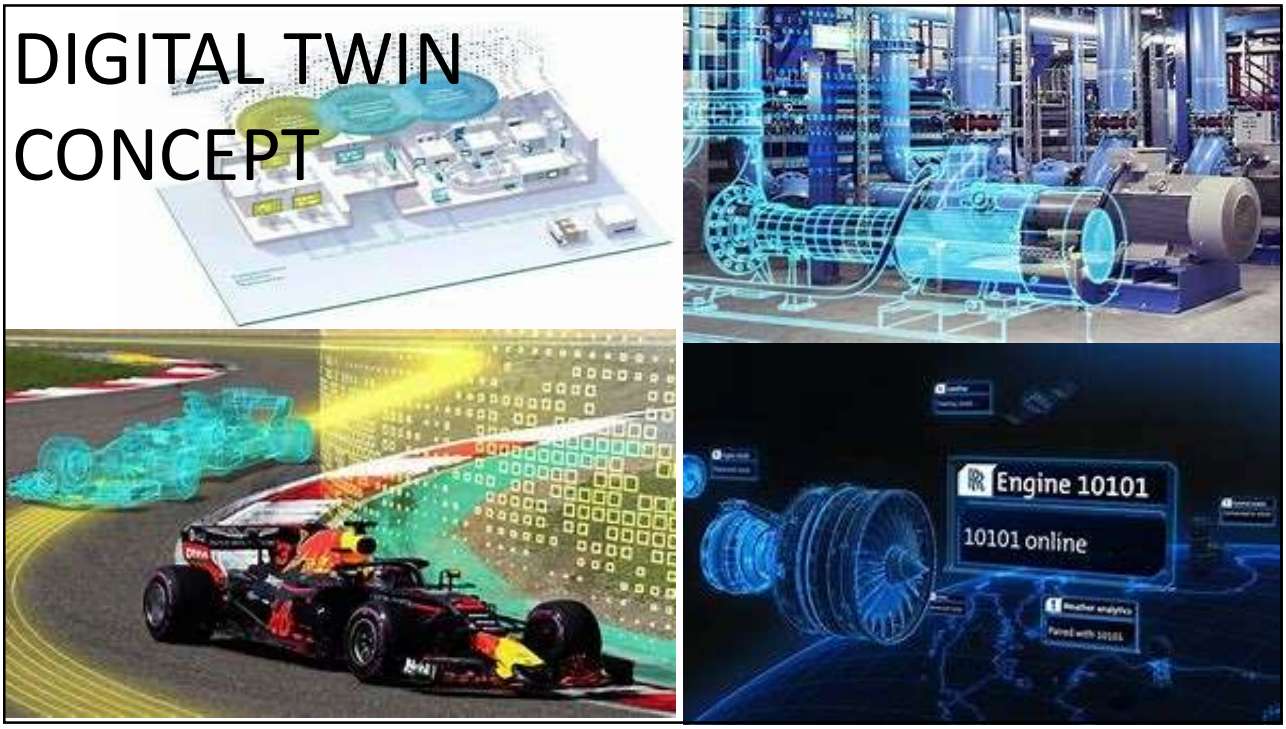
+ An Integrated Pest Management case study

Richard Hinton - Head of Technology Solutions + Enterprise Architecture Planning
Eleanor Koller - Data Scientist
Armando Mendez – IPM Coordinator - Conservation

NATURAL HISTORY MUSEUM

1

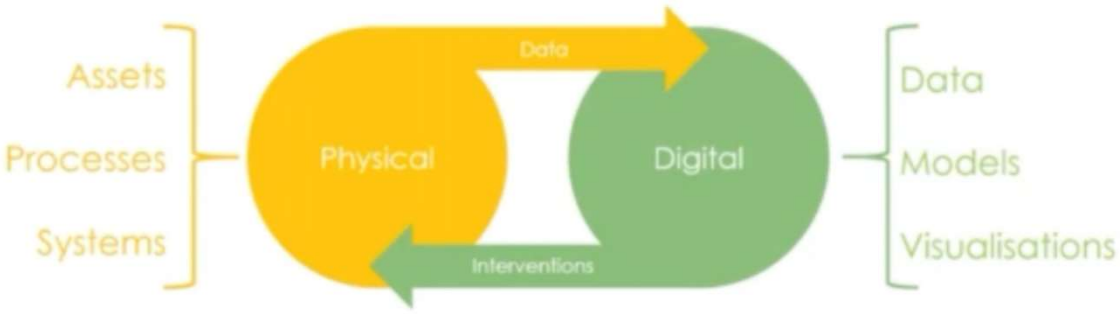
DIGITAL TWIN CONCEPT



2

Digital Twin

Delivering value: "better decisions faster"



Driven by purpose

3

A COMMON THEME ACROSS ALL OUR STRATEGIC PRIORTIES

Our **vision is of a future where both people and the planet thrive**.
To achieve this, we will harness the powerful combination of our three key assets: our collection, our scientific research and our reach to a worldwide audience, in our mission **to create advocates for the planet**.

Our five interlinked strategic priorities will drive our activities over the next 12 years:

Secure the future of our collection
Ensuring our collection is safe, accessible and digitally available - for future innovations and generations.

Develop our gardens and galleries
Creating new spaces, inside and out, combining heritage and experience to connect to nature.

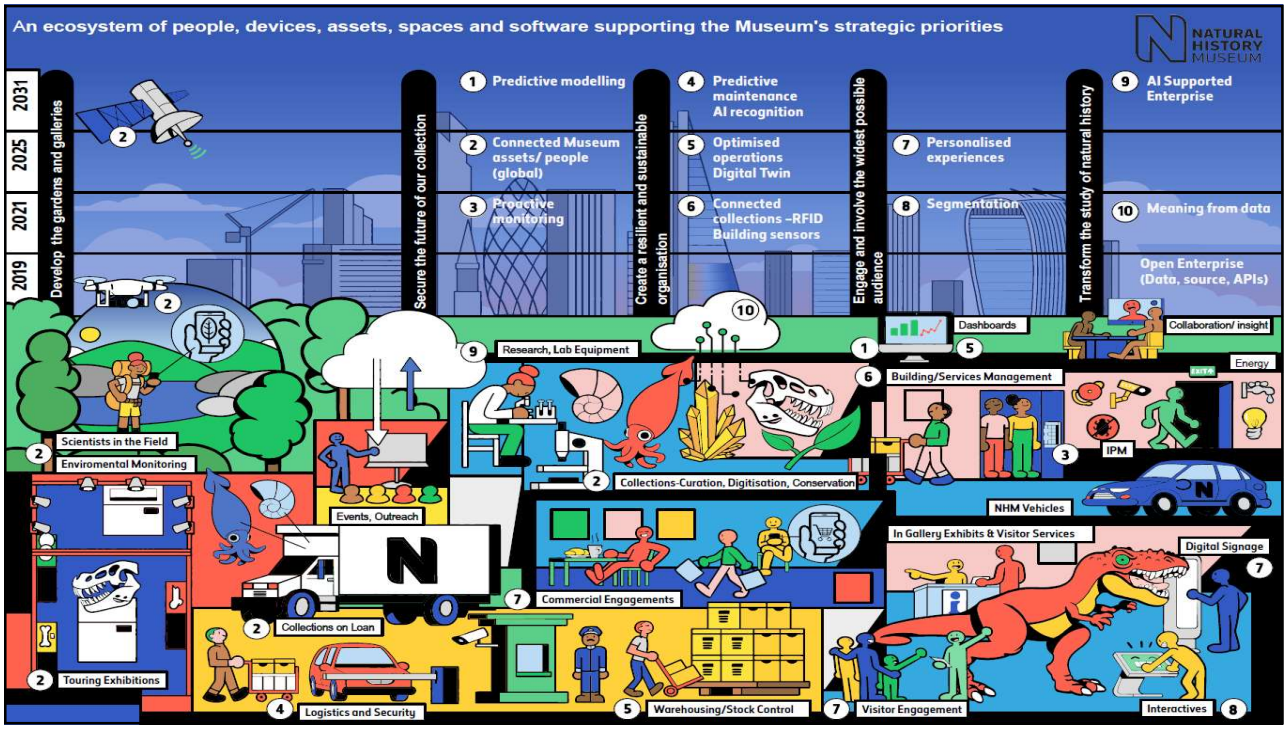
Engage and involve the widest possible audience
Reaching out nationally and globally, onsite and online to create advocates for the planet.

Create a resilient and sustainable organisation
Investing in people, technology and our estate, striving towards financial and environmental sustainability.

Transform the study of natural history
Applying technological innovations to our collection, collections and science, ensuring benefit to people and planet.
Training future generations of scientists.

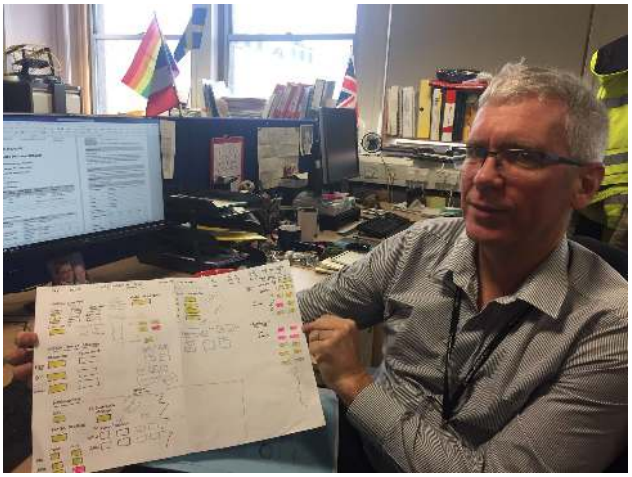
- WHAT KEY QUESTIONS ARE WE TRYING TO ANSWER ?
- WHAT DATA IS REQUIRED?
- ACTIONABLE INSIGHT
- DATA SKILLS/ APPROACH

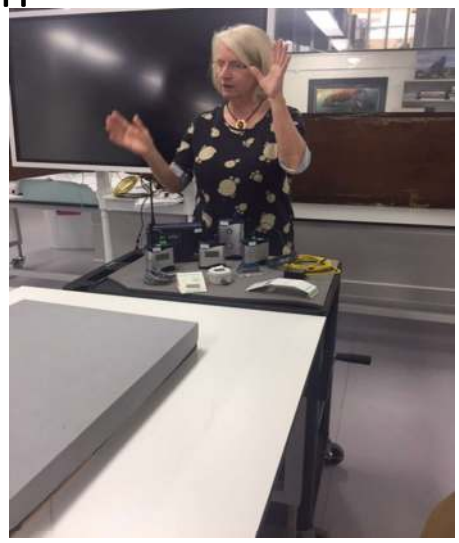
4



5

Building a shared vision

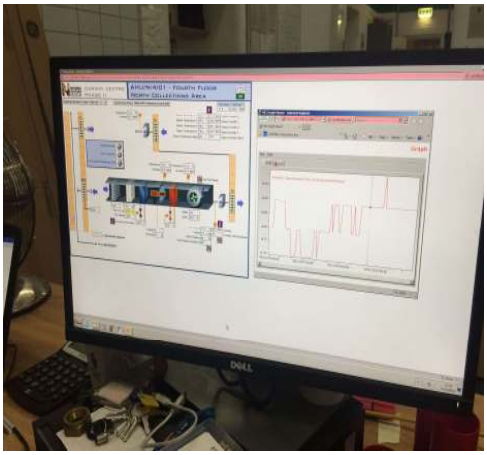




- **Real time dashboard** – bringing together different datasets to inform operations
  (e.g. utilities, heat, humidity, vibration, occupancy, IPM, etc)
- **Unlock insight** – NHM control of data; ability to find and bring together the right data to answer the question
  (e.g. trend analysis, troubleshooting, predictive maintenance)
- **Efficiency** – improve responsiveness

6

Looking for the digital twin



Taking the digital pulse of the Museum

7

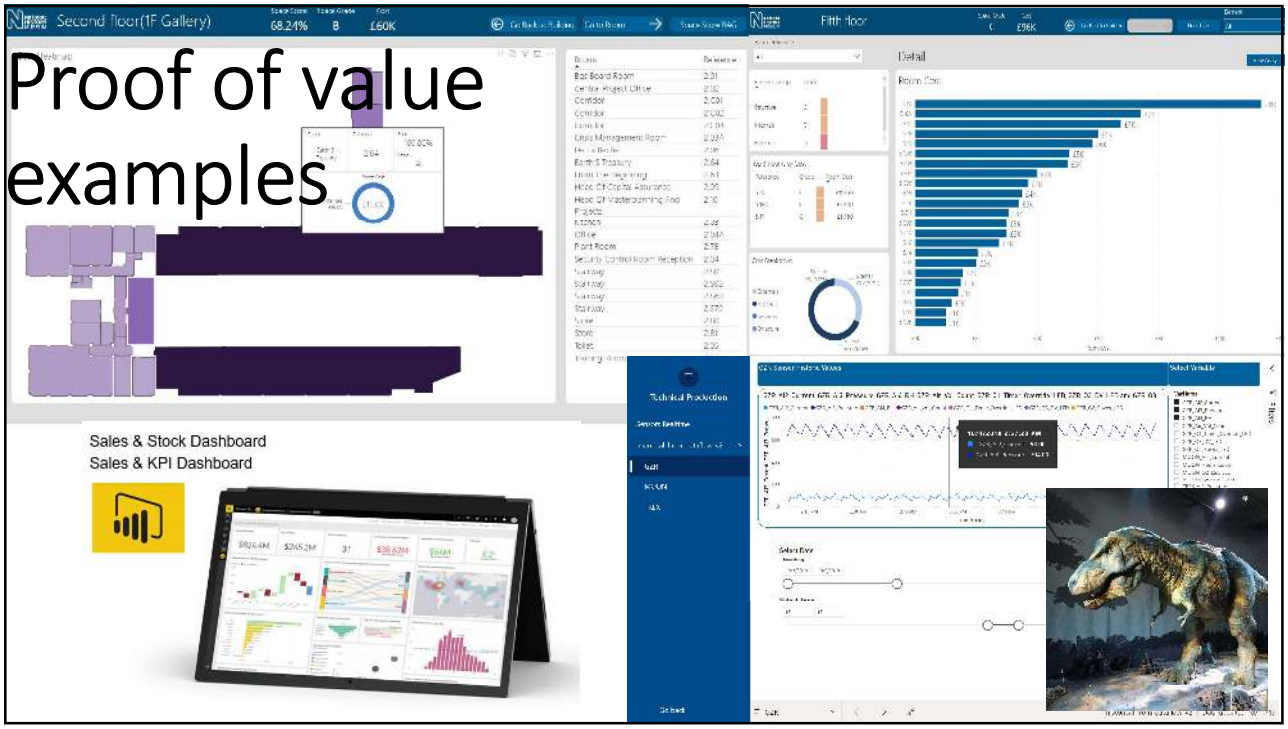
MANTRA

"IF YOU DON'T MEASURE IT, YOU CAN'T MANAGE IT"

"EVERYTHING HAPPENS SOMEWHERE + SOMETIME"

8

Proof of value examples



9

DIGITAL TWIN
Integrated Pest Management case study
AI proof of value

Eleanor Koller - Data Scientist

NATURAL HISTORY MUSEUM

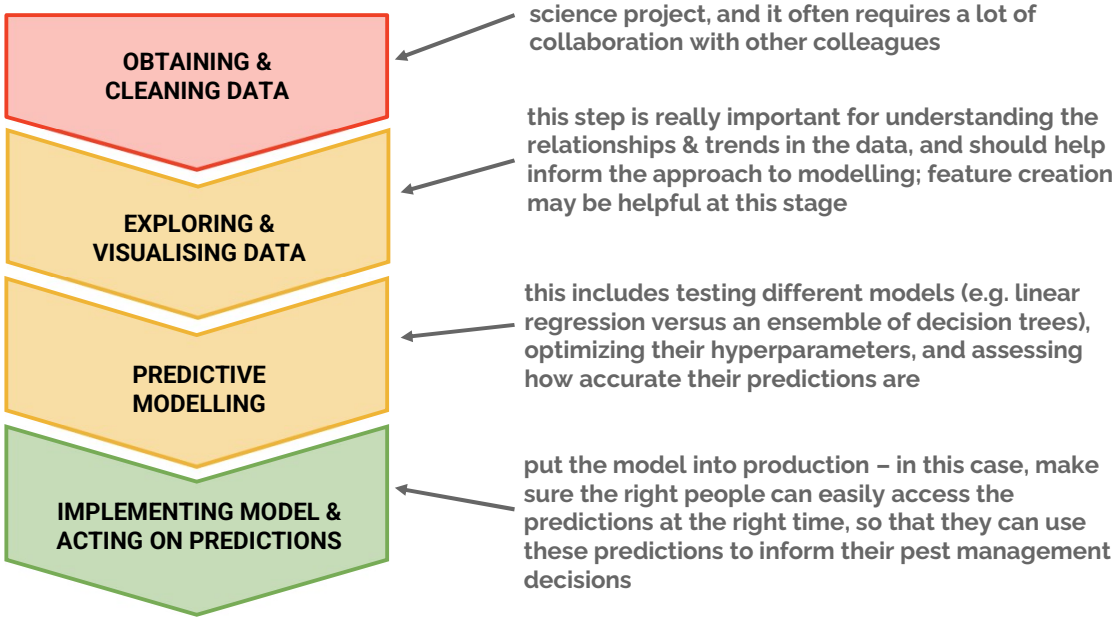
10

Settling on the Problem: Predicting Pests

- Pests are a huge threat to our collection
- If we could predict where and when pests will turn up, we could target our pest management solutions more efficiently and deal more rapidly with infestations. For example, we could:
  - Ask scientists to use stricter protocols with specimens for the next month
  - Keep environmental conditions within a narrower acceptable band
- This seemed like a good problem for us to look at to test out the potential benefits of developing our AI capabilities through a proof of concept:
  - We had a clear problem statement - we knew what we were trying to predict, and what variables would help us make those predictions (e.g. environmental data from sensors in the building)
  - We had good quality data collected for more than a decade
  - We had the support of subject matter experts – our IPM colleagues were really keen to help us understand the context

11

Approach



12

Step 1: Obtaining & Cleaning Data

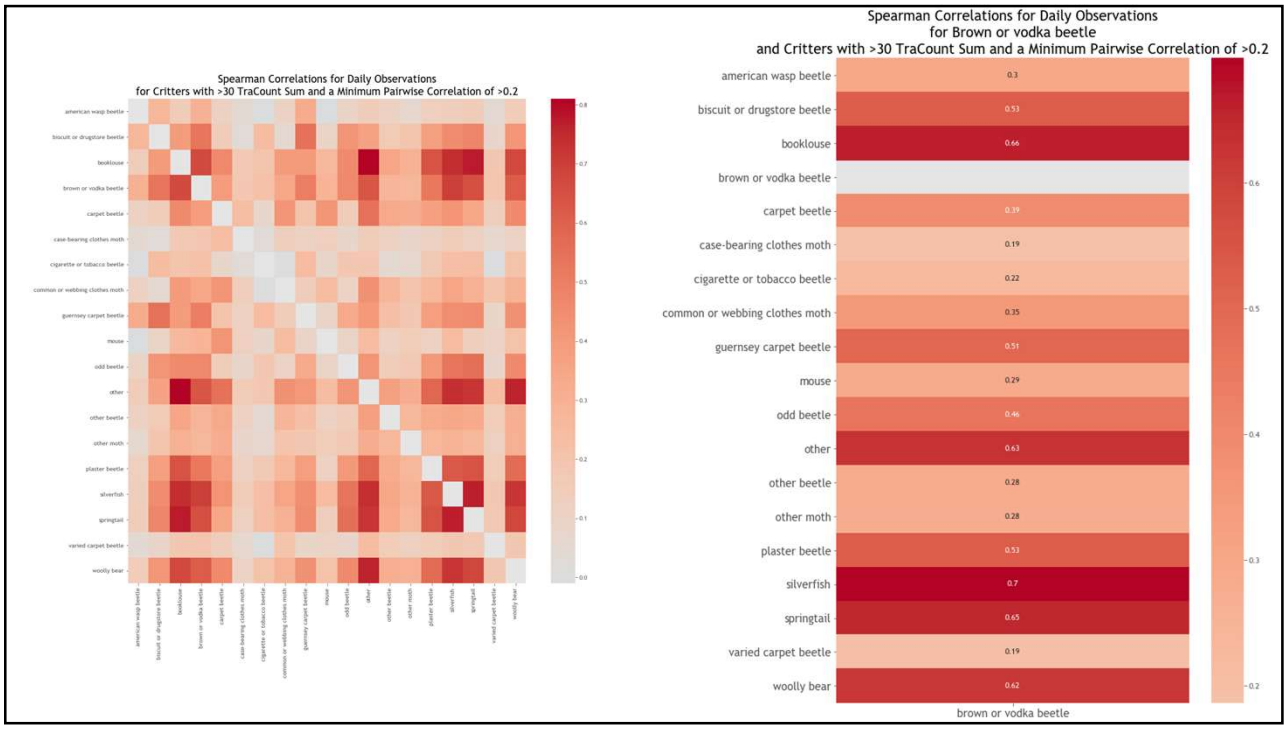
- Collaborating with huge amounts of data required us to work out some data storage & sharing solutions (huge thanks to Jerry Smallwood for his work on this!)
- The locations in our different datasets didn't always match up
- All of our datasets required a bit of cleaning before they were usable – for example:
  - Some inconsistencies with the way pest data was entered – e.g. location data missing, but sometimes present in other fields
  - Humidity readings of 866% and temperature readings of 1660°C
  - In the official historical visitor data broken down by entrance, there's a period of time in 2013 when the DC entrance has negative values recorded, which shouldn't be possible

13

Step 2: Exploratory Data Analysis

- This step is really broad, and entails getting to really understand the data through collaboration with subject matter specialists. For example, before we can even begin any meaningful EDA, we need to answer to questions like:
  - What does 'Indicator – RH' mean in the field 'TraEcoType'?
  - What's the difference between 'TraCount' and 'TraTotalPest'?
  - Were there any notable changes to our pest management regime over the last 15 years that might undermine the continuity in our dataset?
  - How often are traps checked? Is that the same across different spaces?
- This step is where we begin to understand & visualise trends over time, seasonal patterns, and correlations between different variables in our data.
- We also might need to create features at this stage: for example, from the Eltek data, we calculated daily means, maximums, and minimums, as well as features like, "What percentage of time over the last month did the humidity exceed 73%?"

14



15

Can we group some rooms together as rooms that often get pests around the same time?

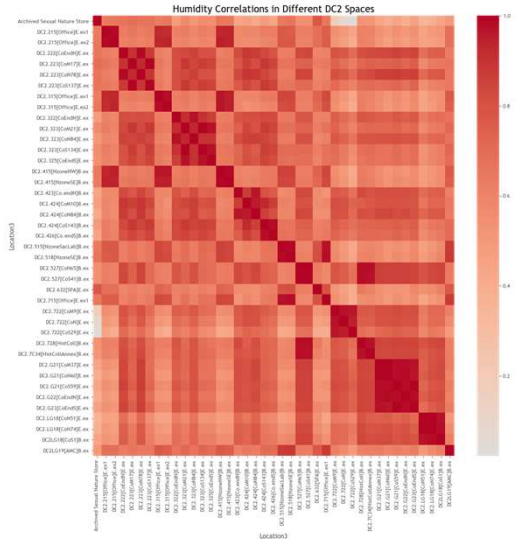
I used correlation matrices to look at that, and also tried a bit of clustering.

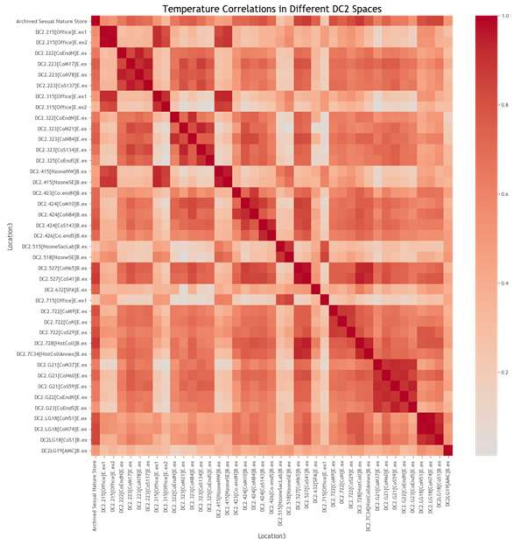
The screenshot at the right shows part of the result of an Agglomerative Clustering Model.

```
Group 3
Index(['B1', 'B14', 'B15', 'B45', 'B46', 'B52', 'B57', 'B6', 'B9'], dtype='object')
--------------------------------------------------
Group 19
Index(['B10'], dtype='object')
--------------------------------------------------
Group 12
Index(['B11', 'B12', 'B16', 'B17', 'B24', 'B29', 'B3', 'B30', 'B31', 'B32',
       'B33', 'B34', 'B36', 'B41', 'B42', 'B43', 'B44', 'B47', 'B48', 'B49',
       'B50', 'B51', 'B53', 'B54', 'B55', 'B56', 'B58', 'B59', 'B60', 'B61',
       'B62', 'B63', 'B64', 'B68', 'B69', 'B7', 'B70', 'B8', 'BB1', 'BB2',
       'BB3', 'ZP1', 'ZP2', 'ZP3', 'ZP4', 'ZP5'],
      dtype='object')
--------------------------------------------------
Group 20
Index(['B13'], dtype='object')
--------------------------------------------------
Group 6
Index(['B18', 'B19', 'B20', 'B21', 'B22', 'B23', 'B25', 'B26', 'B27', 'B28'], dtype='object')
--------------------------------------------------
Group 4
Index(['B2', 'B5'], dtype='object')
--------------------------------------------------
Group 11
Index(['B35', 'B37', 'B38', 'B39', 'B4', 'B40'], dtype='object')
--------------------------------------------------
Group 0
Index(['B65', 'B66', 'B71', 'B72', 'B74, WE.2.S06A', 'B75', 'B76', 'B77',
       'B78', 'B79',
       ...
       'ZZS19', 'ZZS20', 'ZZS21', 'ZZS24', 'ZZS8', 'behind case VI',
       'behind case XIII', 'behind case XIX', 'behind case XV', 'ore room'],
      dtype='object', length=357)
--------------------------------------------------
```

16

This shows that we can probably mostly use humidity and temperature data across the whole of DC2 – but we're more certain about that for humidity than we are for temperature, and maybe it'd be better to group some of the rooms together that are highly correlated than take the whole of DC2





17

Step 3: Modelling

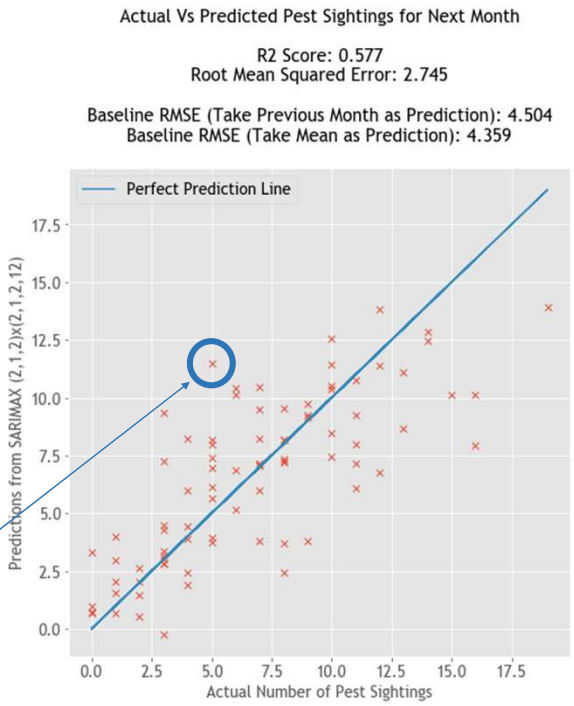
- Once we've explored the data and consulted the subject matter experts, we hopefully have a good sense for what variables might help a model to predict pest infestations. For example, we want to tell the model to consider:
  - Pests over the previous three months, this time last year, and this time two years ago
  - Non-pests over the previous three months
  - Humidity min, max, mean, std, and proportion of previous period spent over certain thresholds (e.g. 73%)
  - Temp min, max, mean, std, and proportion of previous period spent over certain thresholds (e.g. 16.9C)
- Data input constraints:
  - In some spaces, traps are only checked every other month, or the cadence is inconsistent from year to year – this is a challenge for modelling!
  - We also don't have Eltek sensors in every space that we have pest traps – this is another challenge for modelling!
- Pest infestations are not very common events (thanks to all of the efforts of the IPM team!), so we've had some challenges creating a model with high levels of granularity in terms of location, timing, and pest species, even in spaces with decent data collection & Eltek data availability
- When we look at a big enough space with regular pest collection and we group some pests together, though, we can generate some good predictions!

18

Step 3: Modelling

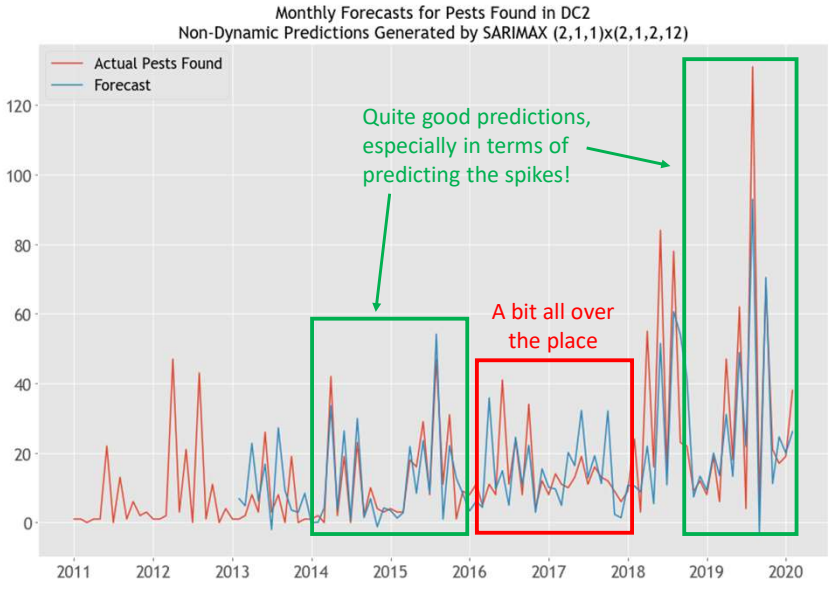
- For a model predicting the number of pest sightings for the next month in DC2, we get an R2 score of 0.58
  - This is a measure of how much better our model is at predicting pest sightings than simply guessing the mean every time, where 1 would be perfect predictions, and 0 would be 'you might as well just guess the mean'.
- The typical error for our model's predictions is 2.7, compared with a typical error of 4.4 or 4.5 if you only ever guessed the mean or the previous month's number.

This is a month where there were 5 pest sightings in DC2, but the model thought there would be 11 or 12



19

Step 3: Modelling

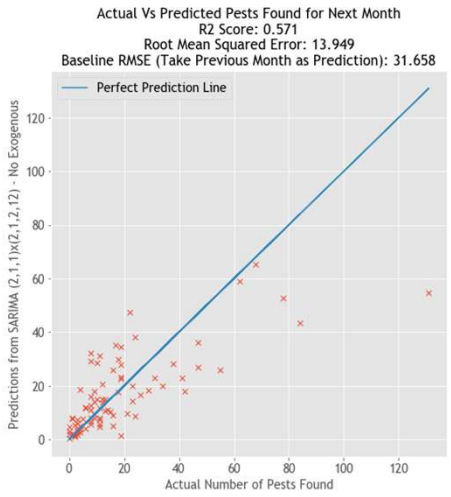


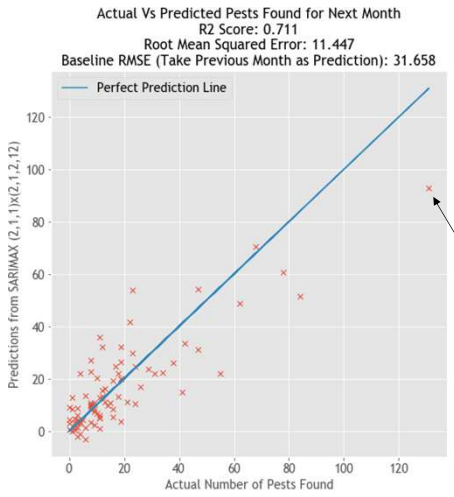
- We can do even better if we try to predict the total number of pests found (rather than the number of pest sightings).
- This shows the forecasts generated by a SARIMAX (2,1,1)x(2,1,2,12) model, which looks at:
  - The number of pests in each of the last two months, and the error in the last month's prediction
  - The number of pests this time last year and the year before, and the error for both of those predictions
  - Exogenous variables:
    - Number of non-pests
    - Humidity
    - Temperature

20

Step 3: Modelling

add exogenous variables, and the model performs better!





EXOGENOUS VARIABLES:
- number of non-pests over the last three months
- proportion of last month hotter than 22.9C
- proportion of last month and 3 months ago more humid than 73%
- min monthly temp 3 months ago
- mean monthly temp last month

In particular, adding exogenous variables has improved the model's prediction of spikes – especially the big outlier of over 120!

21

Step 3: Modelling – Next Steps

- We've had some useful feedback from Armando that we're going to incorporate into our next iteration of models:
  - Add weather data and visitor data
  - Rather than predicting pests overall, look at beetles, moths, and woolly bears separately
- More extensive optimisation:
  - We've looked at a few different types of models (Extra Trees, XGBoost, SARIMAX), but it wasn't comprehensive – there's more to try
  - Could try more extensive hyperparameter optimisation (e.g. for SARIMAX models, how many previous data points should the model consider? 1? 2? 3?)
  - Which features should we add to the model that add predictive value rather than bogging it down with extra dimensions?
- We've only modelled for DC2 so far – there is the small matter of the whole rest of the Museum (although some spaces don't have the data for it, for example due to infrequent collection)
- Step 4 - How would we set the models up to provide predictions on an ongoing basis?

22

23