Introducing a data science perspective on predictive modelling within a large introductory statistics course: Connecting research with practice

DR ANNA FERGUSSON

Te Kura Tatauranga | Department of Statistics

Waipapa Taumata Rau | University of Auckland

a.fergusson@auckland.ac.nz

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
DEPARTMENT OF STATISTICS

Introducing a data science perspective on predictive modelling within a large introductory statistics course: Connecting research with practice

DR ANNA FERGUSSON

Te Kura Tatauranga | Department of Statistics

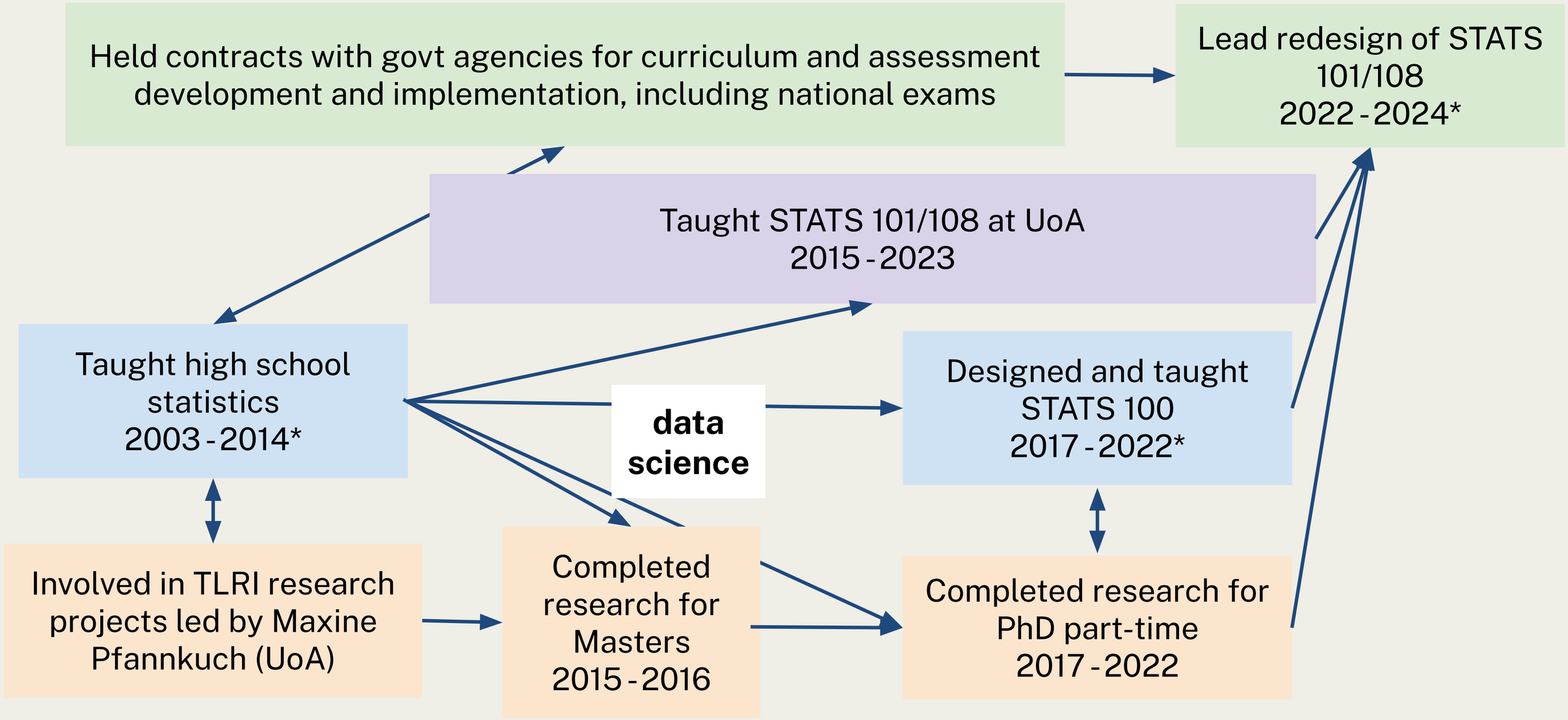
Waipapa Taumata Rau | University of Auckland

a.fergusson@auckland.ac.nz

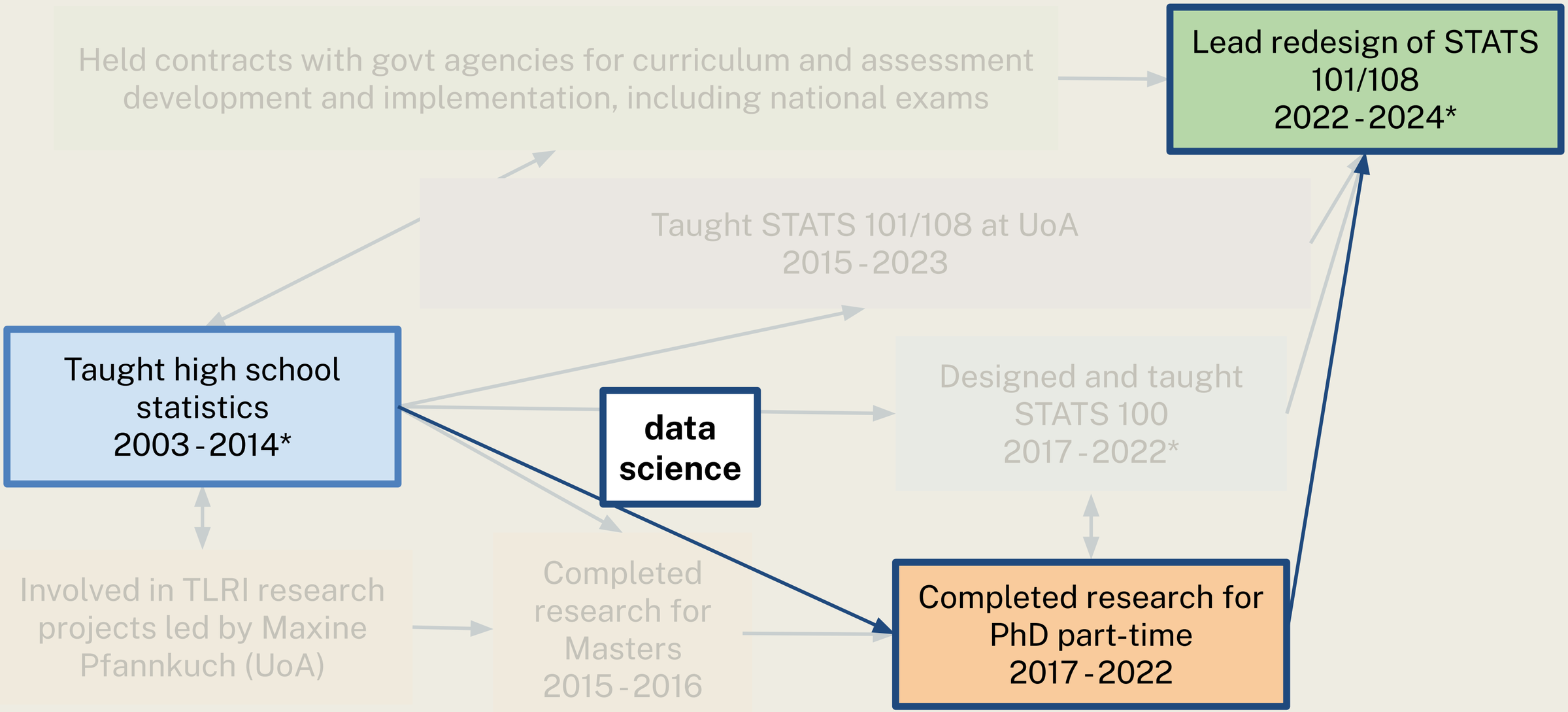
THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
DEPARTMENT OF STATISTICS

BACKGROUND: HOW DOES MY RESEARCH CONNECT WITH PRACTICE?



BACKGROUND: HOW DOES MY RESEARCH CONNECT WITH PRACTICE?



WHY A FOCUS ON PREDICTIVE MODELLING?

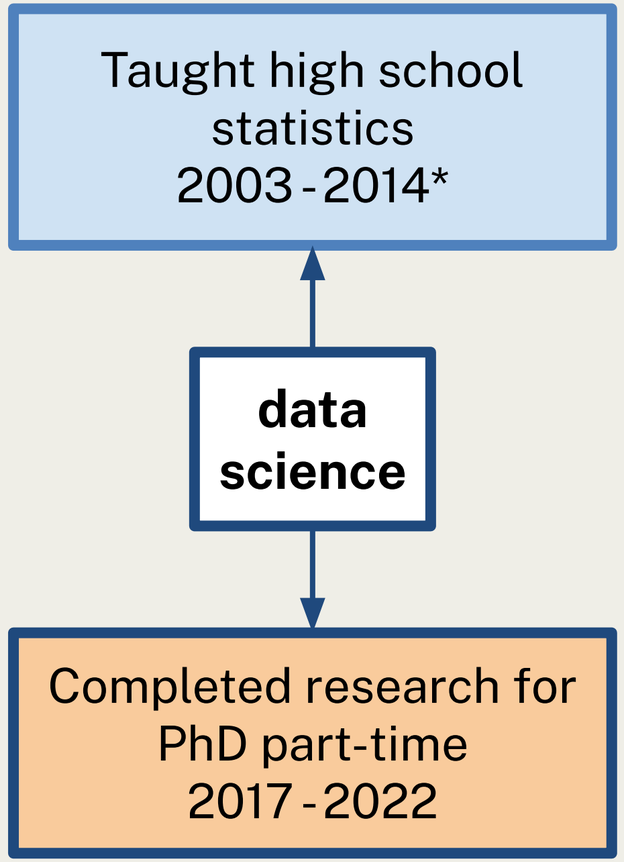
Data + data technologies

What it means to learn from data is different because of advances in computing (Baumer et al, 2017)

Learning experiences need to reveal more of the world of data faster (Wild, 2015) and use dynamic data contexts that engage students (e.g., Fergusson & Bolton, 2018)

Predictive modelling has many modern applications (e.g., using past customer transactions to predict future purchasing behaviours)

Teaching about algorithmic bias and data ethics should highlight the subjectivity of analytical and modelling processes



WHY A FOCUS ON PREDICTIVE MODELLING?

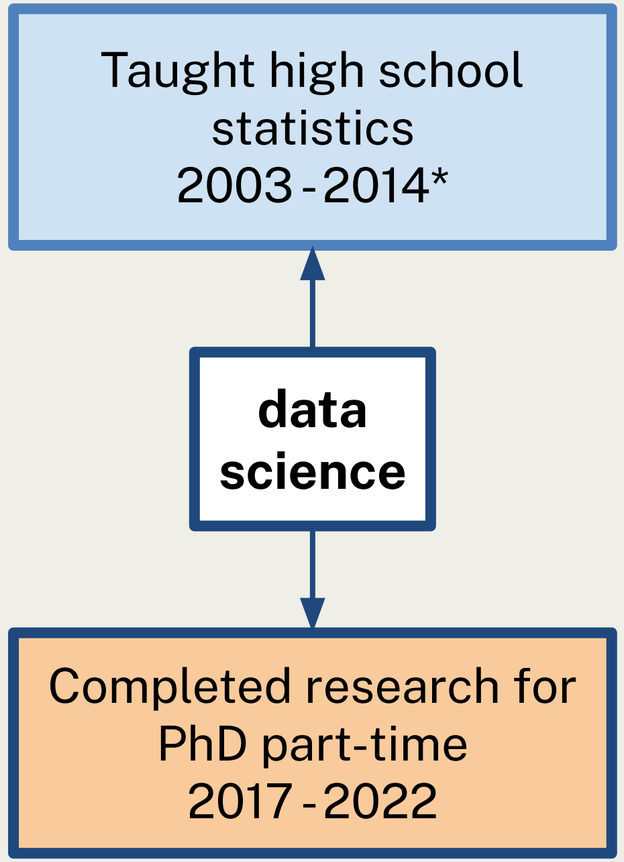
Modelling perspective

Central to statistical thinking is the use of statistical models (Wild & Pfannkuch, 1999)

A *modelling* perspective is needed for learning statistics (Garfield et al., 2012)

Place greater emphasis on predictive modelling (e.g., Biehler & Schulte, 2017)

Algorithmic models could offer a more accessible and conceptually simpler mechanism to introduce students to data science than inferential methods (Gould, 2017)



WHY A FOCUS ON PREDICTIVE MODELLING?

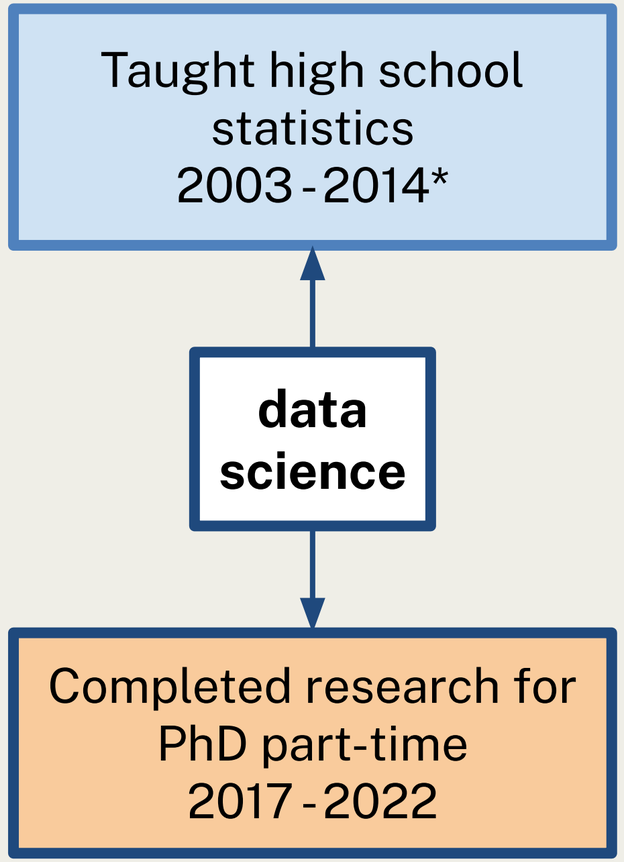
Different purposes for data

Description: Describe a dataset or a population of interest: Involves calculating estimates based on groups and identifying clusters

Prediction: Predict what will happen in a new instance or at a future time: Involves making predictions and forecasts at the individual level.

Explanation: Explain why things have happened, often so we change how they happen in the future: Involves discovering and investigating causes.

(cf. Carlin & Moreno-Betancur, 2023)



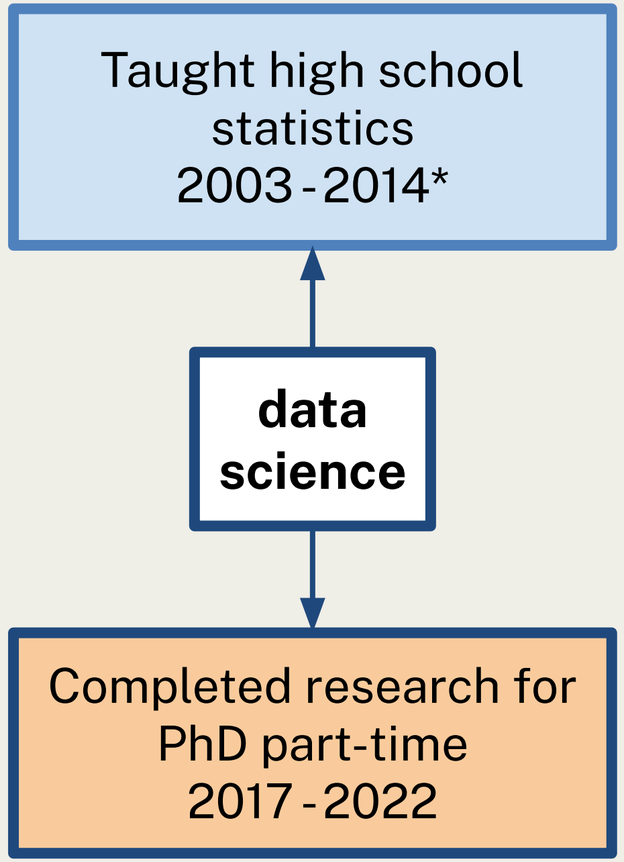
WHY A FOCUS ON PREDICTIVE MODELLING?

The New Zealand teaching context

Senior high school statistics students are expected to use linear regression models to make point predictions

Students are not currently expected to:

- use a prediction model developed with one set of data to generate predictions for cases within a different set of data
- generate prediction intervals from a model
- evaluate a model in terms of predictive accuracy
- discuss precision versus accuracy
- access APIs as a source of data
- use computer programming as part of the predictive modelling process



CONSIDERATIONS WHEN DESIGNING PhD RESEARCH TASK

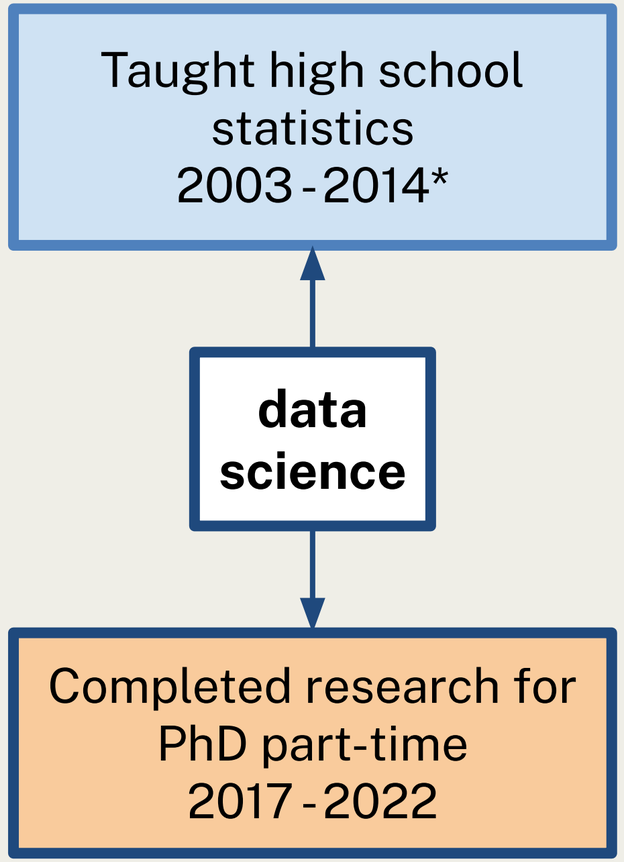
Recommendations

Immerse learners in data-rich contexts by sourcing dynamic (“live”) data from the internet (e.g., Hardin, 2018)

Use APIs and custom-built tools to support student interactions with data and modelling (Fergusson & Wild, 2021)

Build from familiar understandings of linear regression but include more emphasis on validation through residual analysis and predictive accuracy (Biehler & Schulte, 2017)

Draw on the success of informal inference research (e.g., Makar & Rubin, 2018) by using an informal approach



KEY FEATURES OF RESEARCH APPROACH (Fergusson, 2022)

Participants

A design-based research approach was used (e.g., Reeves, 2007)

Participants were six experienced Grade 12 statistics teachers, teachers were positioned as learners

Task implemented during second day of four full-day professional development workshops

Teachers worked in pairs, were given access to one laptop computer to assist with completing the task, and asked to “think aloud” as they completed the task (Van Someren et al., 1994).

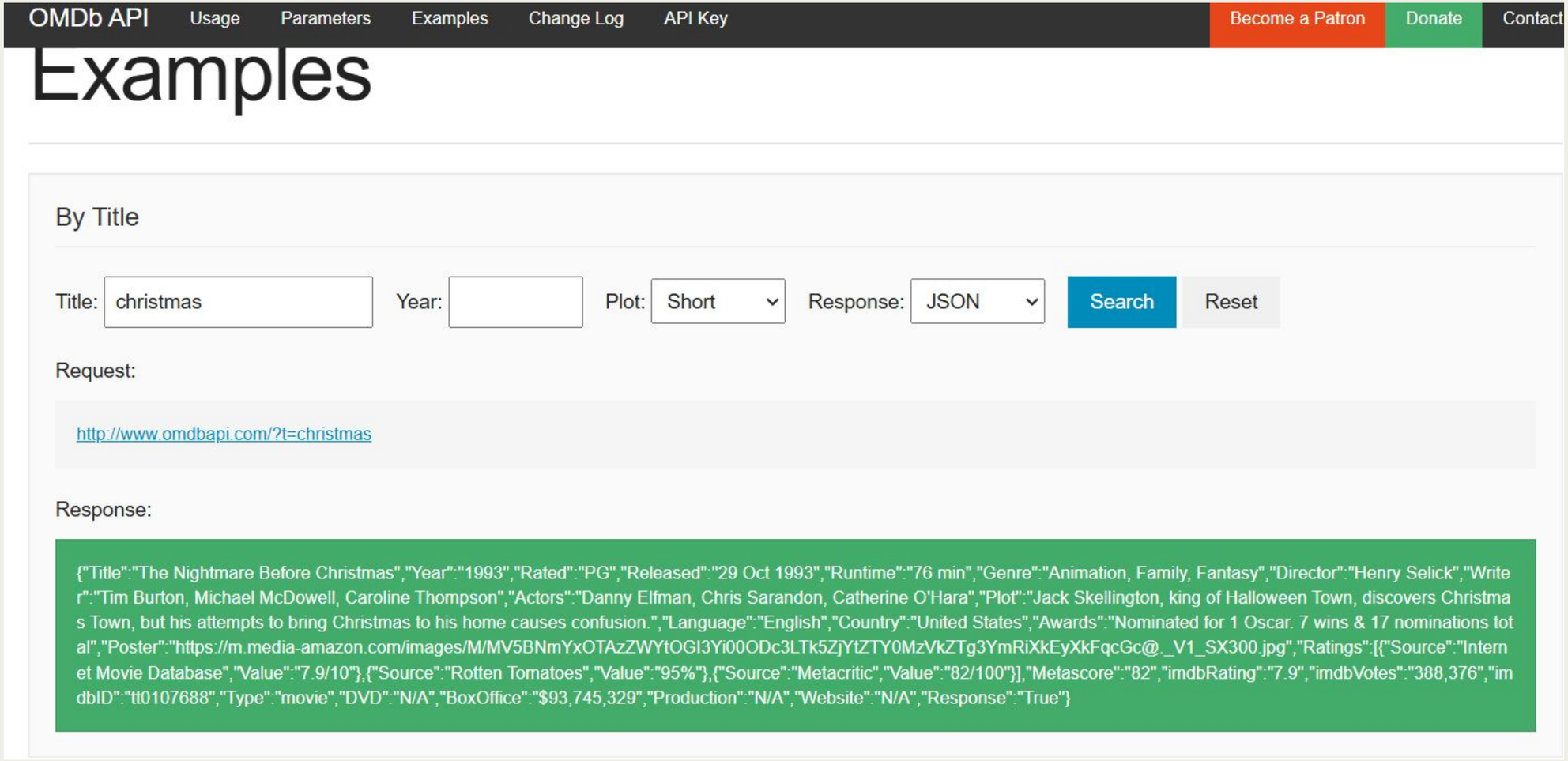
The statistical modelling approach was an **informal method** that relied on **teachers' reasoning with features of visualisations** to construct a model that **generated prediction intervals.**

The form of the prediction model was predicted $y = a + bx \pm \text{error}$, where a and b are the y-intercept and slope of a linear model respectively, and error is a numeric value visually estimated by the teachers to model prediction error.

KEY FEATURES OF TASK

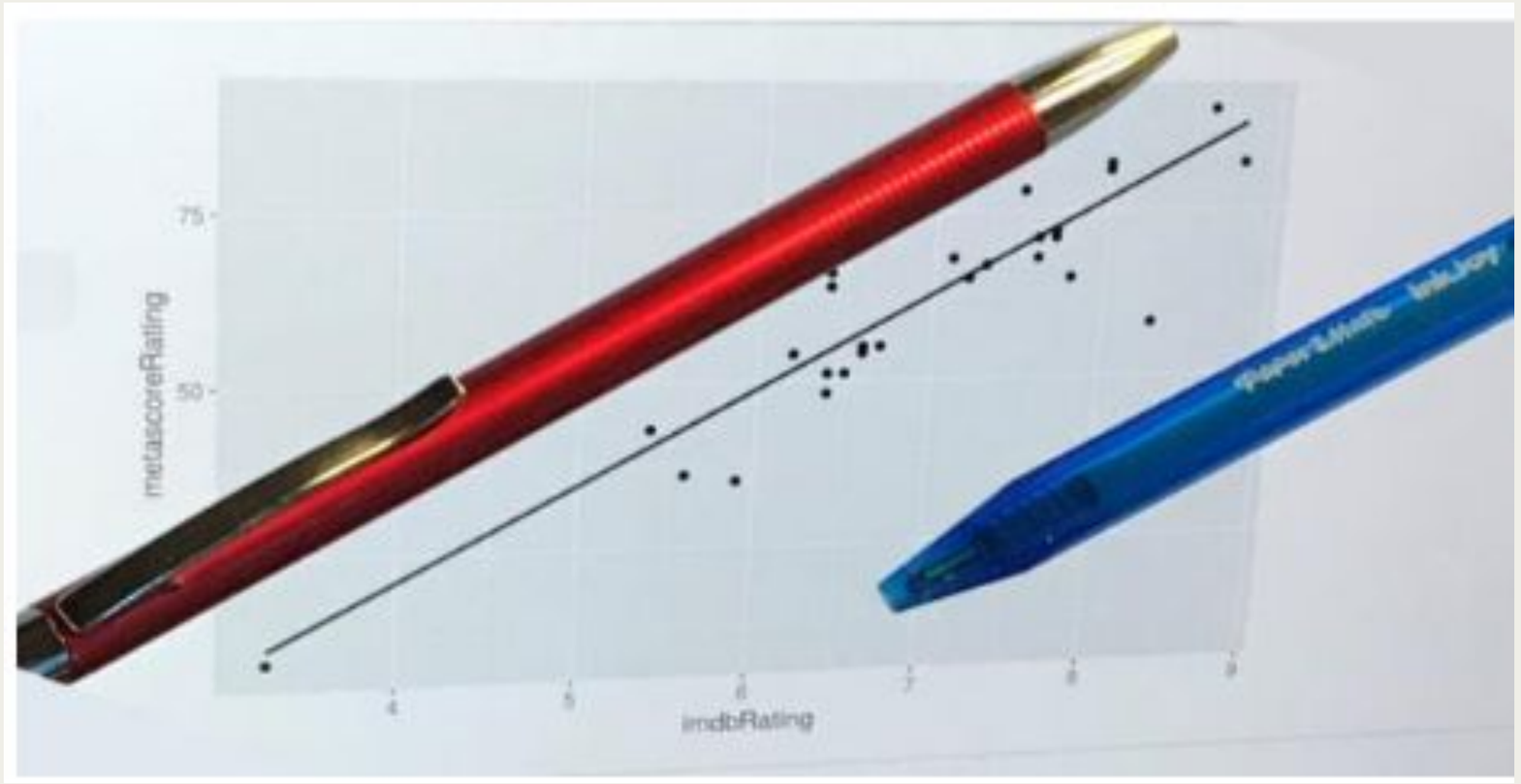
Teachers were able to generate movie ratings data dynamically using an API

Teachers could create different data sets to develop and test model using different search terms



KEY FEATURES OF TASK

Teachers used features of the scatter plot to decide on where to position “parallel lines” to the fitted linear model

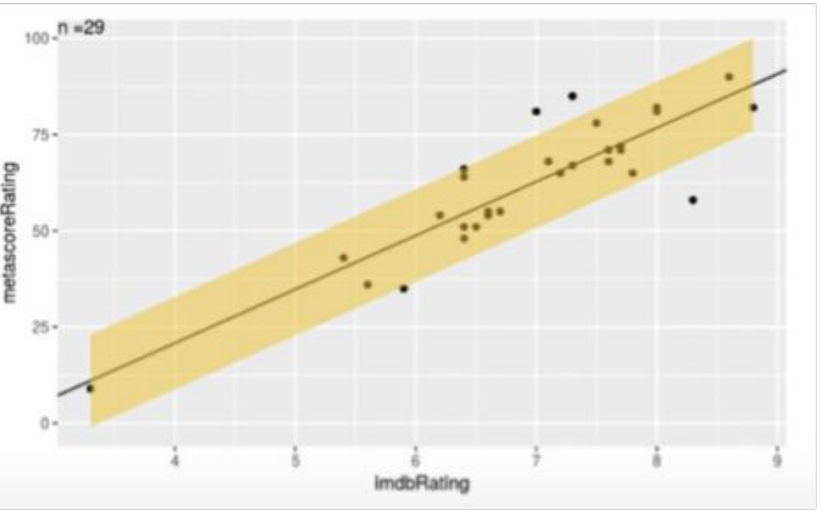


KEY FEATURES OF TASK

Teachers were able to create visualisations of the prediction intervals generated from their model by using R code based on their values for the y-intercept, slope, and error

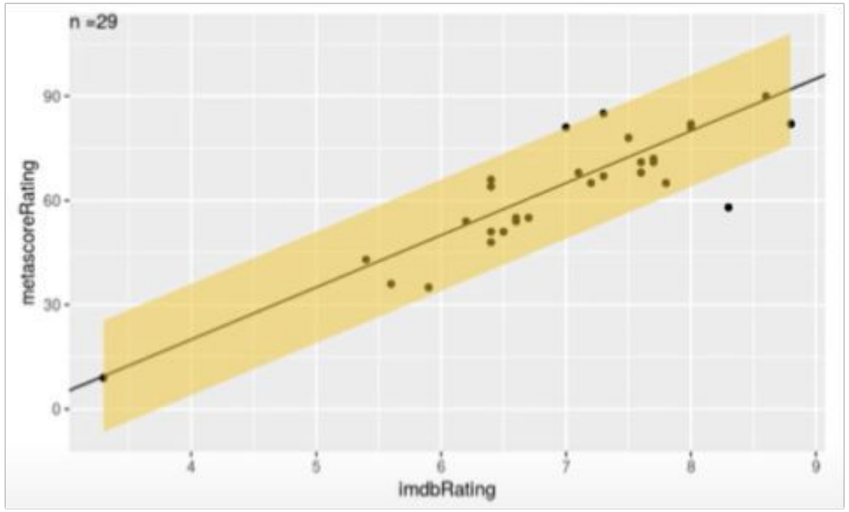
(a) HN

```
5 a <- -35.05967 # y-intercept
6 b <- 13.97598 # slope
7 error <- 12 # error
```



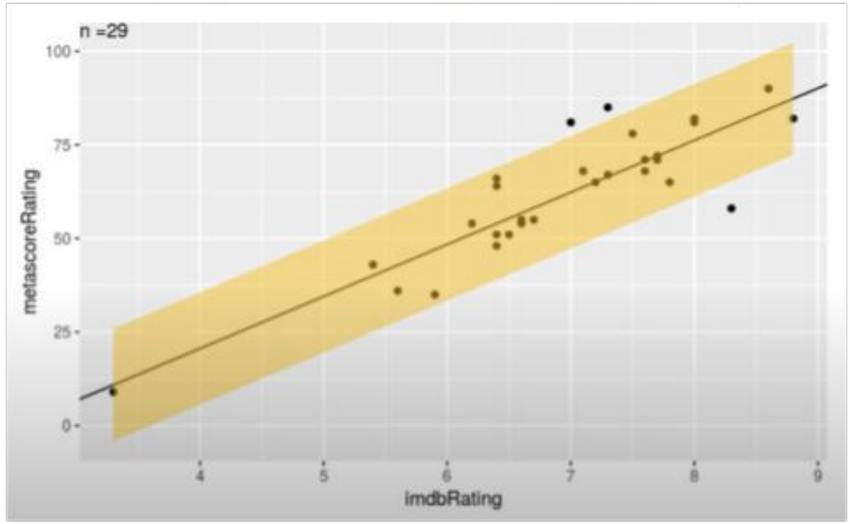
(b) AN

```
5 a <- -40 # y-intercept
6 b <- 15 # slope
7 error <- 16 # error
```



(c) AI

```
5 a <- -35 # y-intercept
6 b <- 13.9 # slope
7 error <- 15 # error
```



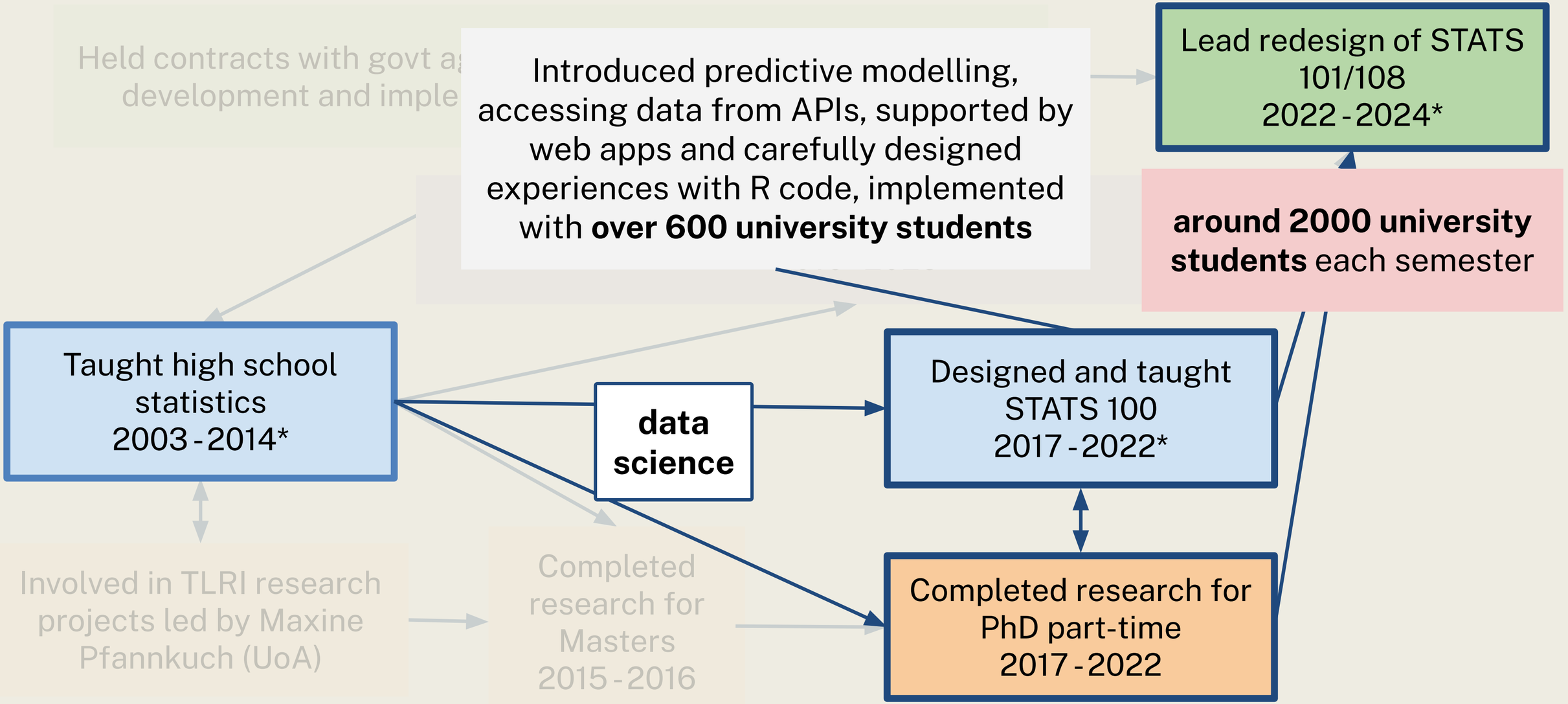
RELEVANT FINDINGS

The informal approach to developing a prediction model, primarily the visualisation of error, provided an opportunity for teachers to develop their own reasoning for what made a good prediction model e.g. *"Take out that bottom one! We're not using all the dots, are we?"*

The visualisation of prediction intervals appeared to support teachers to assess their models in terms of both predictive precision and accuracy e.g., *“Well, you want it [the prediction interval] to be narrow but you also want it to be realistically narrow. It's no good saying you want it to be narrow if it doesn't actually predict very well.”*

Use of an API to access and use more than one data set as part of the modelling process appeared to help teachers appreciate the predictive goal of the modelling task e.g., *“Actually one of the things that impressed me was that we didn't change the gradient or the intercept and yet that line fitted everything we tried pretty well. It pretty much went through the points no matter which thing we tried it on which was really good.”*

OTHER CONSIDERATIONS FOR CONNECTING RESEARCH WITH PRACTICE




FISHER & PAYKEL
APPLIANCES
AUDITORIUM

WHAT IS THE STATS 101/108 TEACHING CONTEXT?

The students

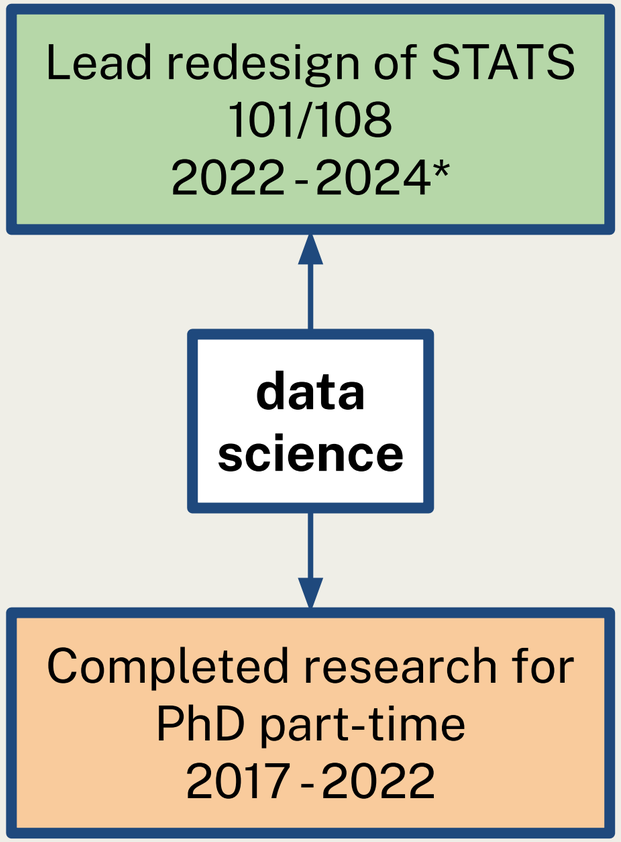
- Predominantly mix of business students and science students
- Range of specialisations e.g. Psychology, Biological sciences, Environmental science, Accounting, Marketing, etc.
- Open entry - range of student prior knowledge in Statistics
- Large student cohort ~2200 students per semester

The teaching

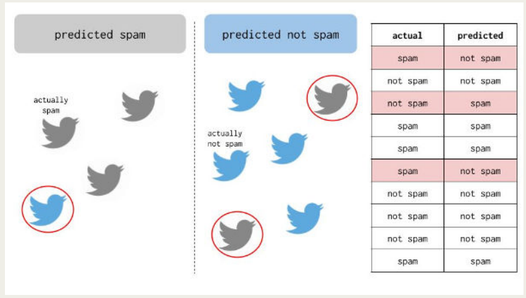
- Taught in four/five streams
- Large teaching team
- Three x 50 minute lectures each week
- Drop-in sessions for student support

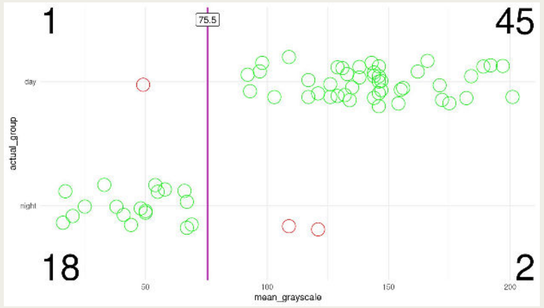
The changes

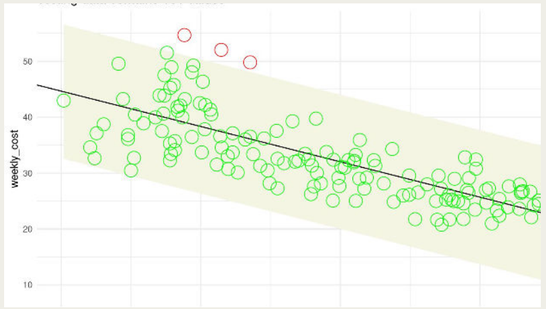
- **New module with a focus on data technologies and prediction**
- New online interactive course book
- New assessment structure (weekly quizzes/chapter tasks)
- New lectures with a focus on student interaction

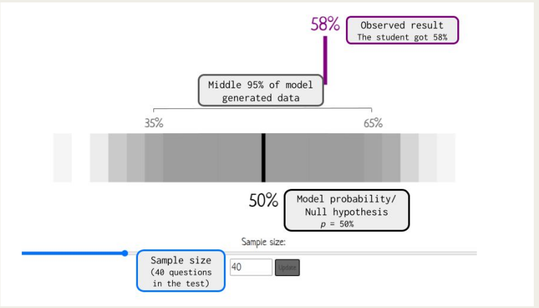


WHEN IS PREDICTIVE MODELLING INTRODUCED?









Chapter 1 **Datafication**

Informal approaches to binary classification models, where the input is a categorical variable.

Chapter 2 **Classification**

Informal approaches to binary classification models, where the input is a numeric variable.

Chapter 3 **Prediction**

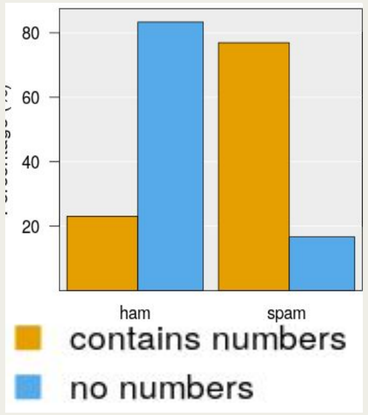
Informal approaches to prediction models, where output is a prediction interval for a numeric value.

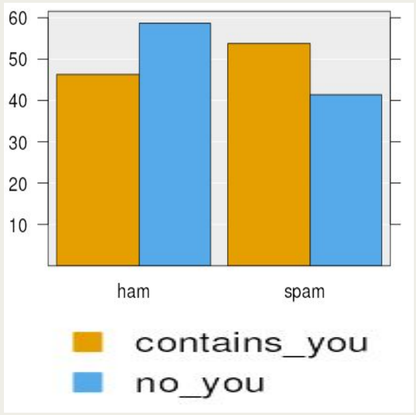
Chapter 4 **Randomisation**

Simulation-based inference approaches, where observed data is compared to model-generated data.

CONCEPTUAL PATHWAY

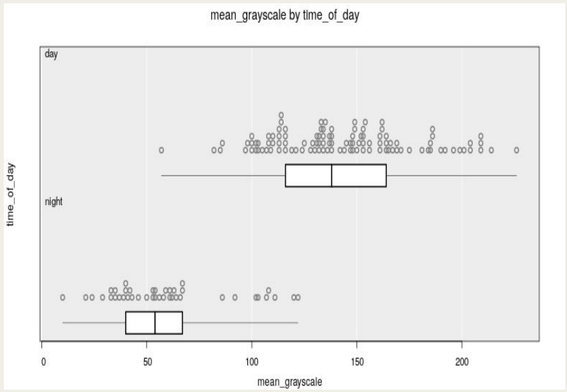
CONCEPTUAL PATHWAY: LAYING FOUNDATIONS (Datafication + Classification)

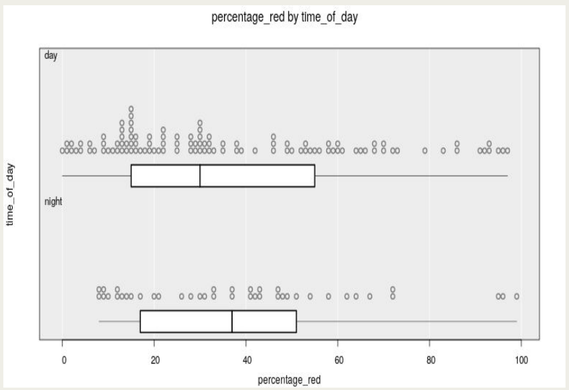




Chapter 1
Datafication

- **Datification** - familiarisation with data and tech - Google Sheets, importing data into analysis software (iNZight), generating plots and summary statistics
- Considering which variables will be useful for building a classification model and why





Chapter 2
Classification

- Introduction to **training/testing** - use one set of data to build a model, use another set of data to test the model
- Considering which variables will be useful for building a classification model and why
- Determining cut-off values and writing decision rules

- **Accuracy** - calculate the PCC and compare to a **baseline model**

CONCEPTUAL PATHWAY: BUILDING ON FOUNDATIONS (Prediction)

Baseline model

Compare and contrast baseline models for categorical (binary) and numeric target variables

Concept(s): The “baseline model” for a prediction model is a simple “no information” prediction model that always predicts the average of the target variable. As the baseline model provides just one value for the prediction, you only have “one shot” at getting the prediction correct.

Prediction errors

Use informal and estimation-based graphical approaches with a focus on predictions and prediction errors

Concept(s): We need to account for how “far off” our predictions are from “the truth”. Smaller prediction errors means are predictions are more precise and the size of the prediction errors is a combination of the model used (which you can change) and the distributional features of the data (which you can’t change).

Prediction intervals

Develop prediction intervals for single numeric variables, numeric + categorical variables, then pairs of numeric variables

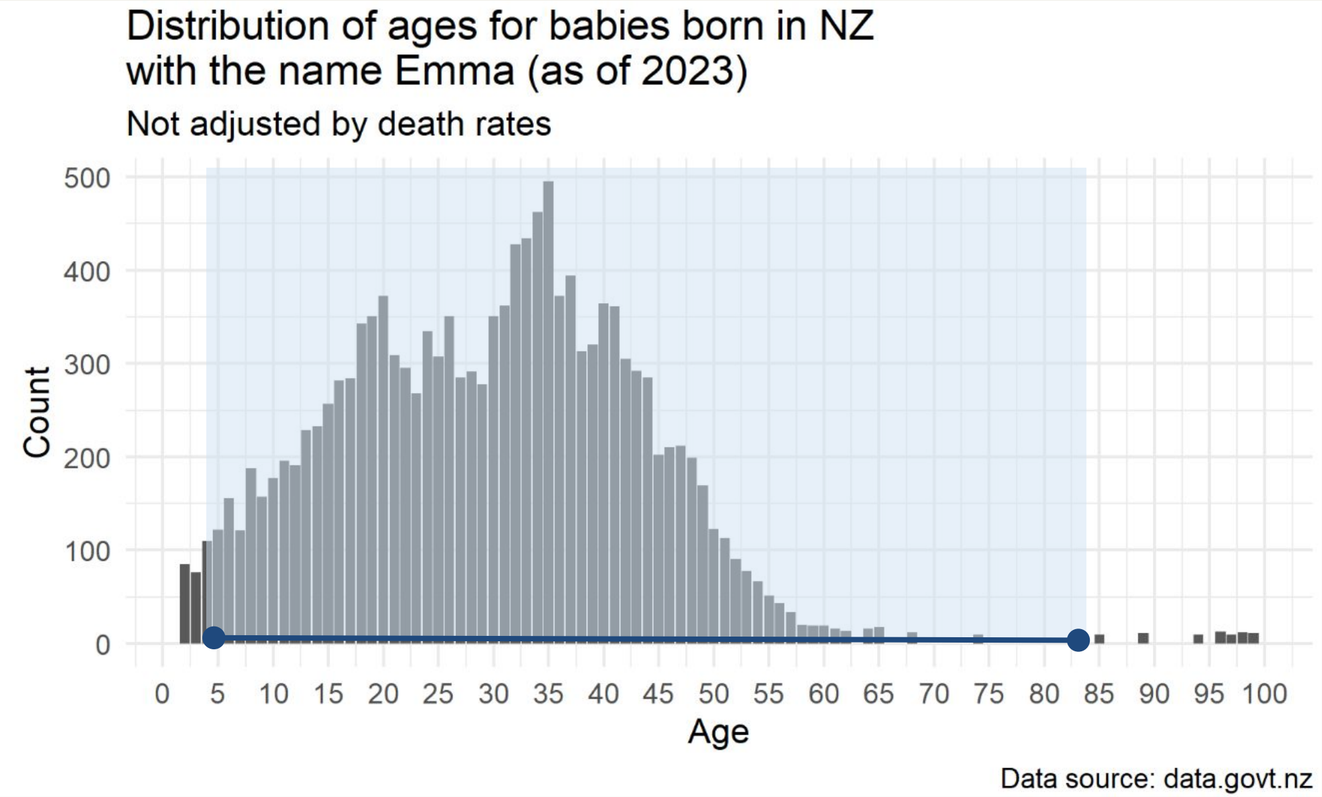
Concept(s): A prediction interval gives a range of values for each prediction, between a lower limit and an upper limit. Making our prediction interval as wide as the middle 95% of the data we should cover most of the variation seen in our data. We might be able to get narrow prediction intervals if we take into account other variables in our prediction model.

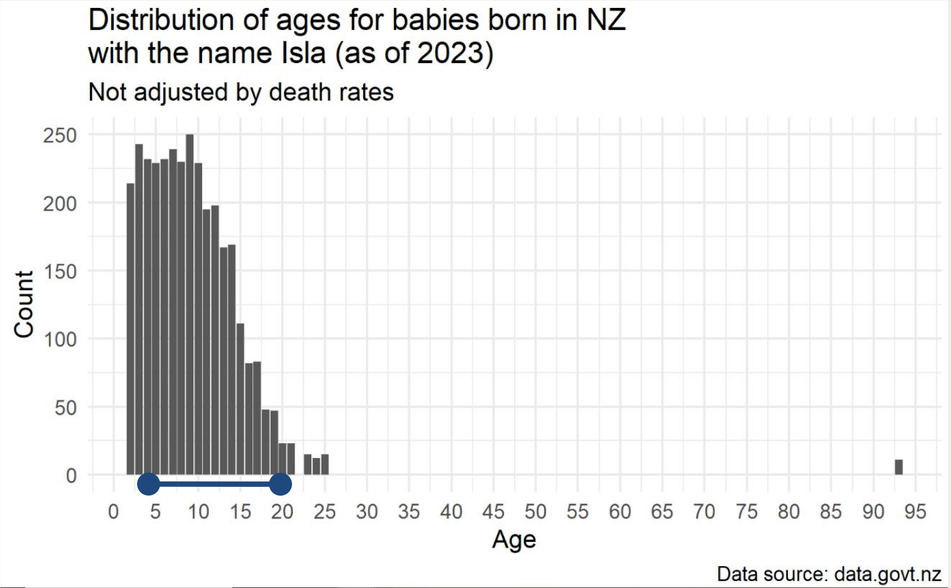
Training/testing data

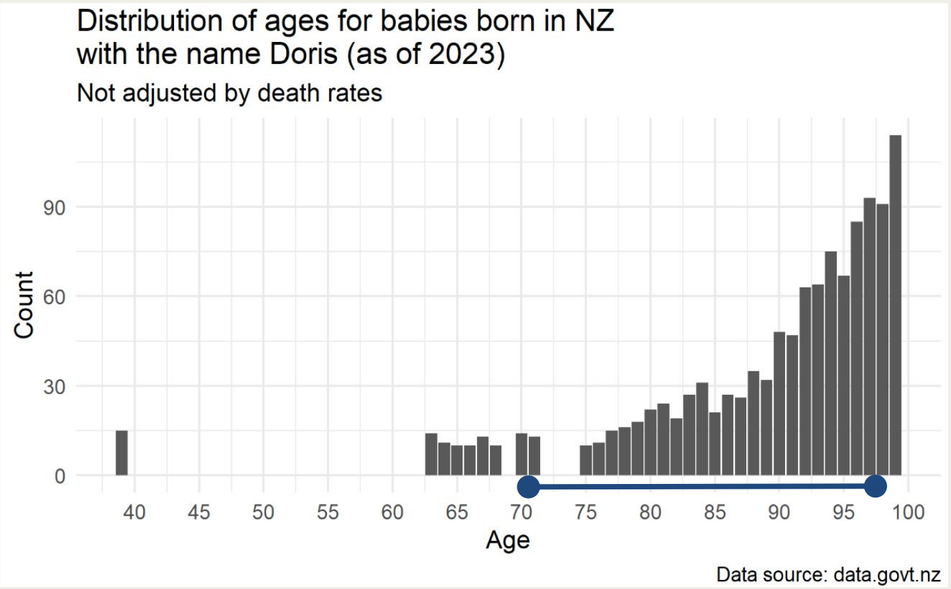
Provide data and interactive tools where students can select and use different subsets for training and testing models

Concept(s): We use a training and testing approach to building models to stop ourselves from being overly optimistic about how great our model is, which can lead to “overfitting”. Even if we train a model that gets the most amazing test results, if we try to use it to predict for situations where the “same rules don’t apply” in terms of the relationship being modelled, we’re not necessarily going to get accurate predictions.

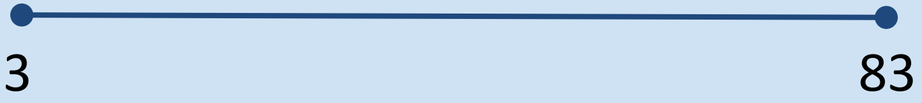
EXAMPLE 1: PREDICTING AGES LECTURE ACTIVITY



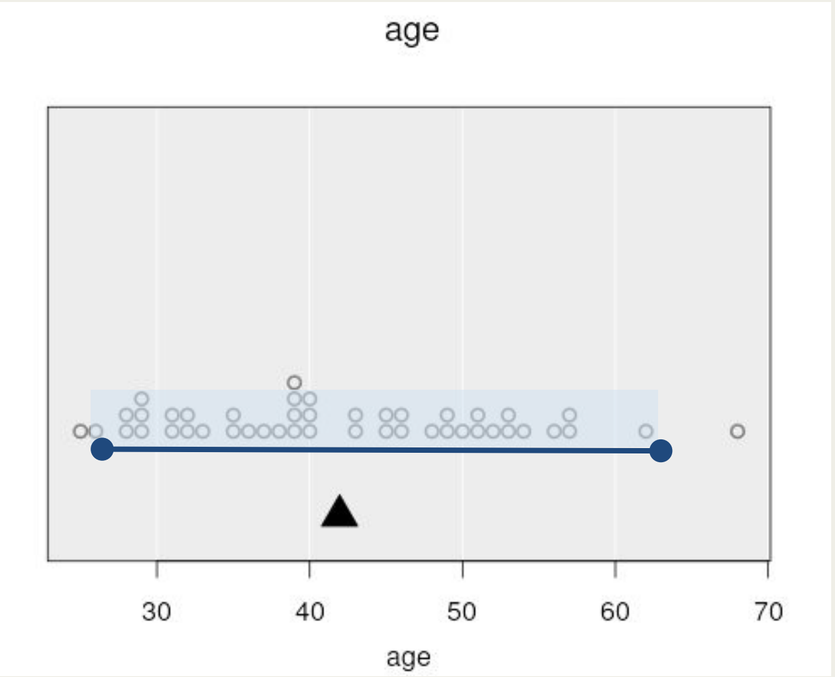


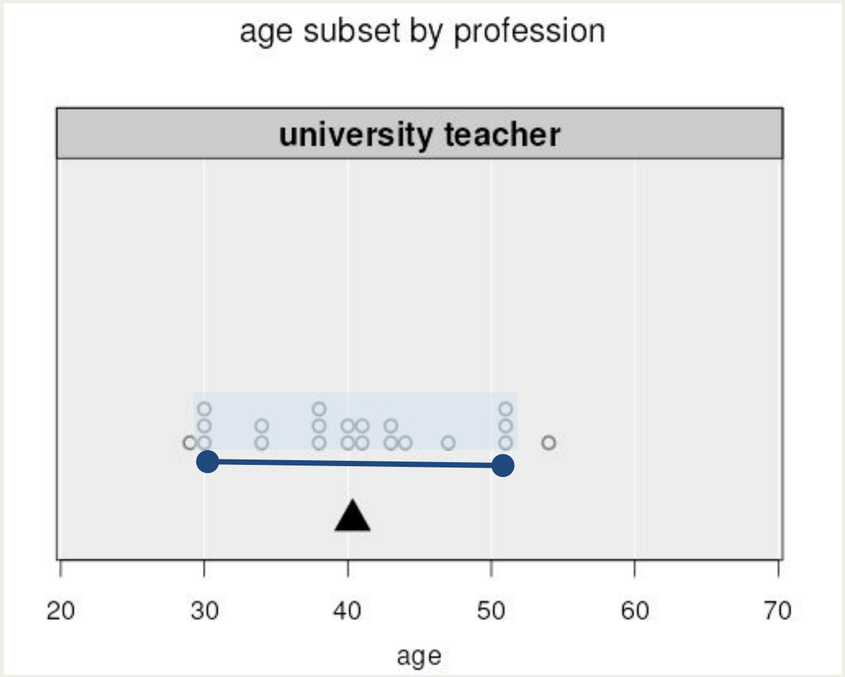


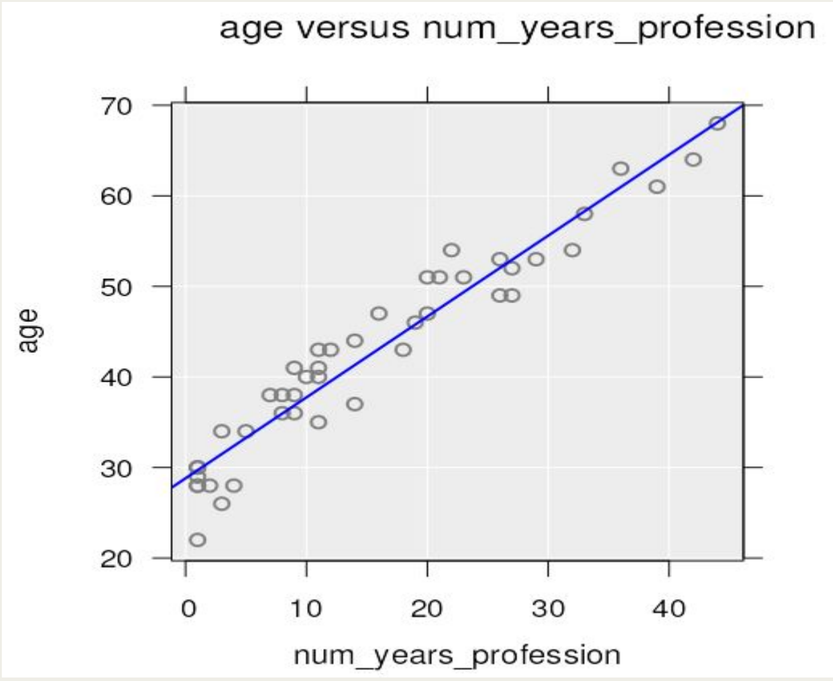
“Someone named Emma could be between 3 and 83 years old”



EXAMPLE 1: PREDICTING AGES LECTURE ACTIVITY







Age only

predicted age: (25, 63)

Age by profession

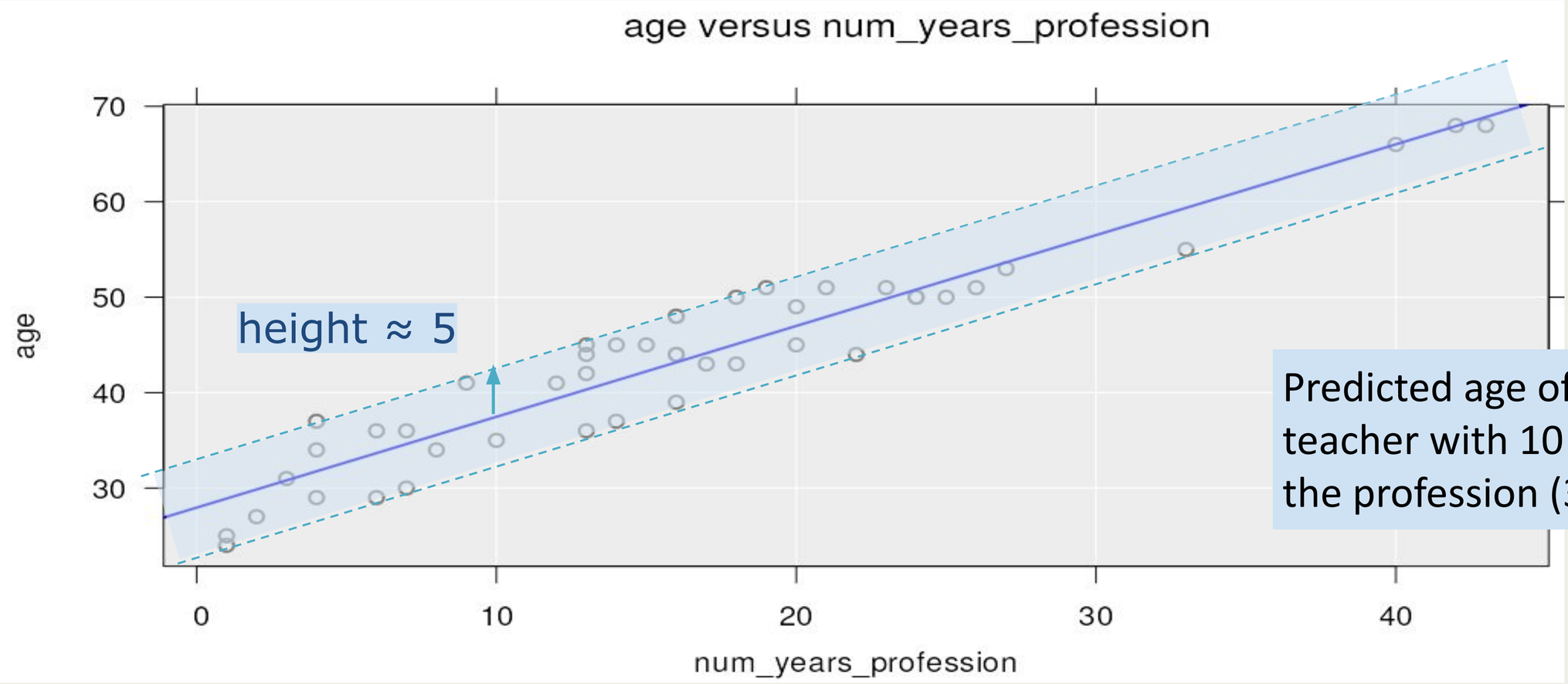
predicted age: (30, 51)

Age by number of years in profession

predicted age = 28.81+ 0.8934 * num_years_profession

EXAMPLE 1: PREDICTING AGES LECTURE ACTIVITY

predicted age = 28.81+ 0.8934 * num_years_profession ± 5

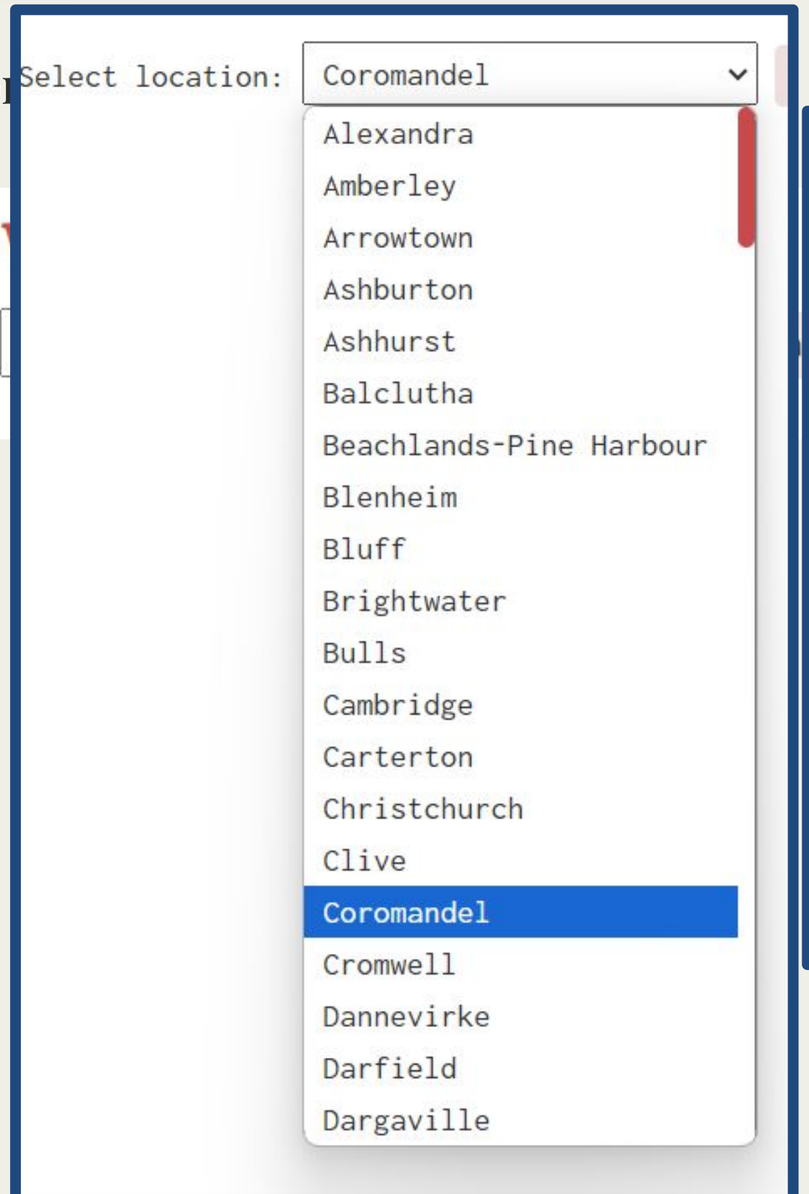


EXAMPLE 2: PREDICTING WEATHER TASK

Choose locatio

Predicting

Select location:



DECEMBER 1, 2023 BY VISUAL CROSSING

Timeline Weather API

The Timeline Weather API is the simplest and most powerful way to retrieve weather data. You can request data over any time window including windows that span the past, present, and future. The API will take care of the combining historical observations, current 15-day forecasts, and statistical weather forecasts to create a single, consolidated dataset via a single API call.

Looking for the /forecast and /history end point docs?

Our existing Weather API endpoints for /forecast and /history queries are still fully supported. The documentation for these endpoints can be found here.

The Timeline API offers complete, global weather data coverage both geographically and chronologically. It always picks the best available data sources to answer any weather API query. These sources include:

- Current weather conditions
- Daily historical, forecast and statistical forecast data (depending on dates requested)
- Hourly historical observations and 15-day forecast
- Weather alerts
- Astronomical observations including sunrise, sunset and moon phase.

EXAMPLE 2: PREDICTING WEATHER TASK

Develop model

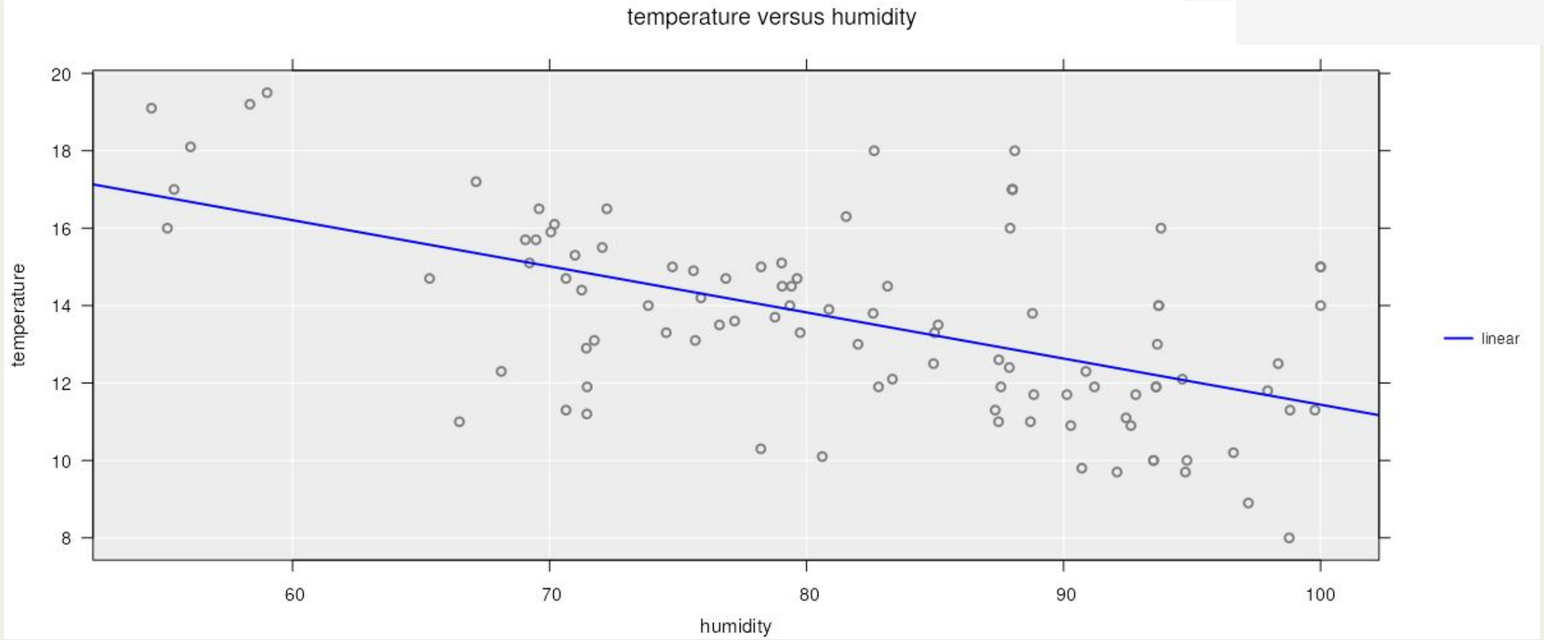
Select location: Coromandel

Get **training data**

Link to training data CSV:

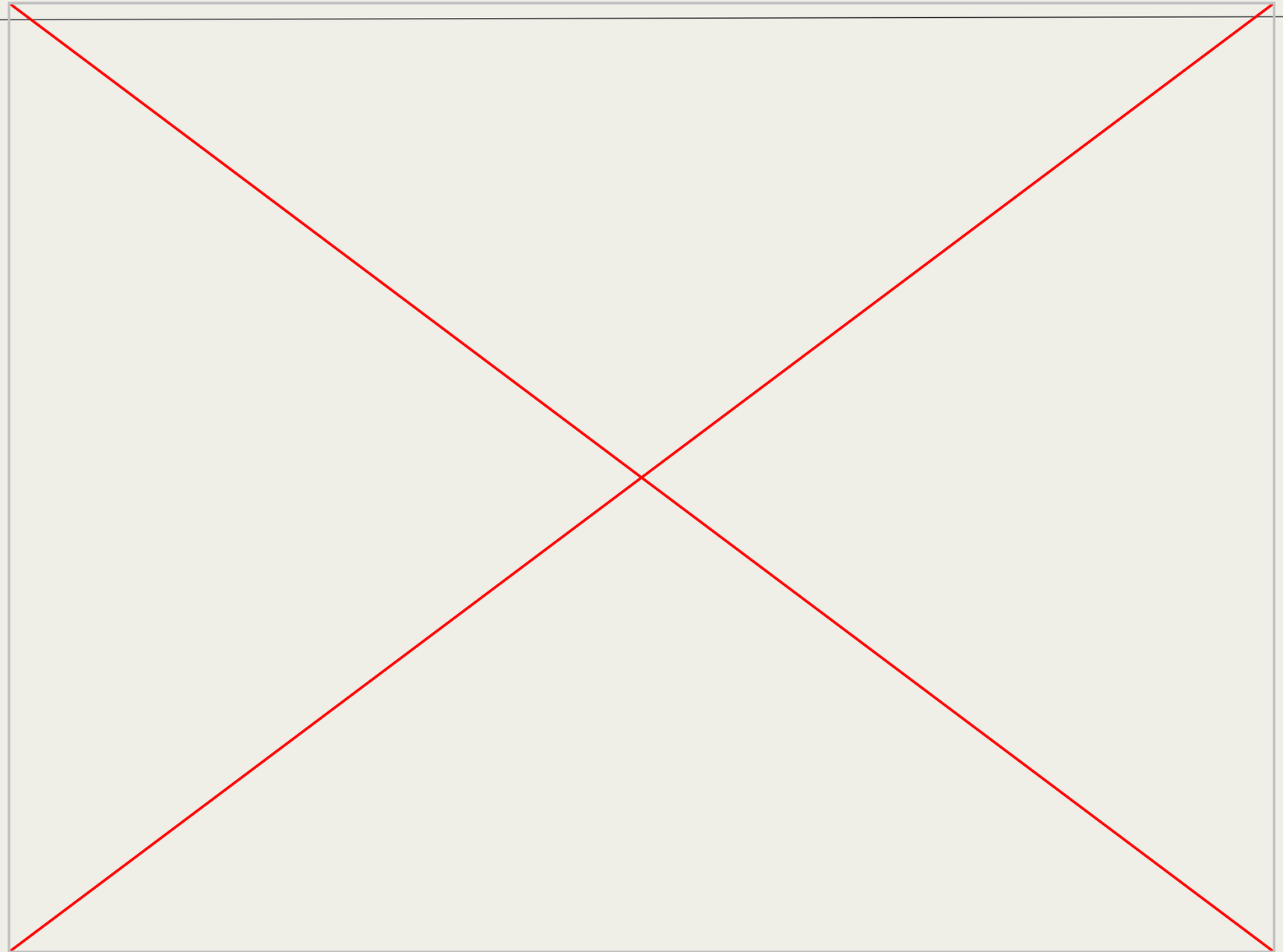
https://dataexplorations.online/csv/tosr675953c4eebde.csv

```
Linear trend:

    temperature = 23.35 - 0.1191 * humidity
    Linear correlation: -0.56

Rank correlation: -0.50  (using Spearman's Rank Correlation)
```



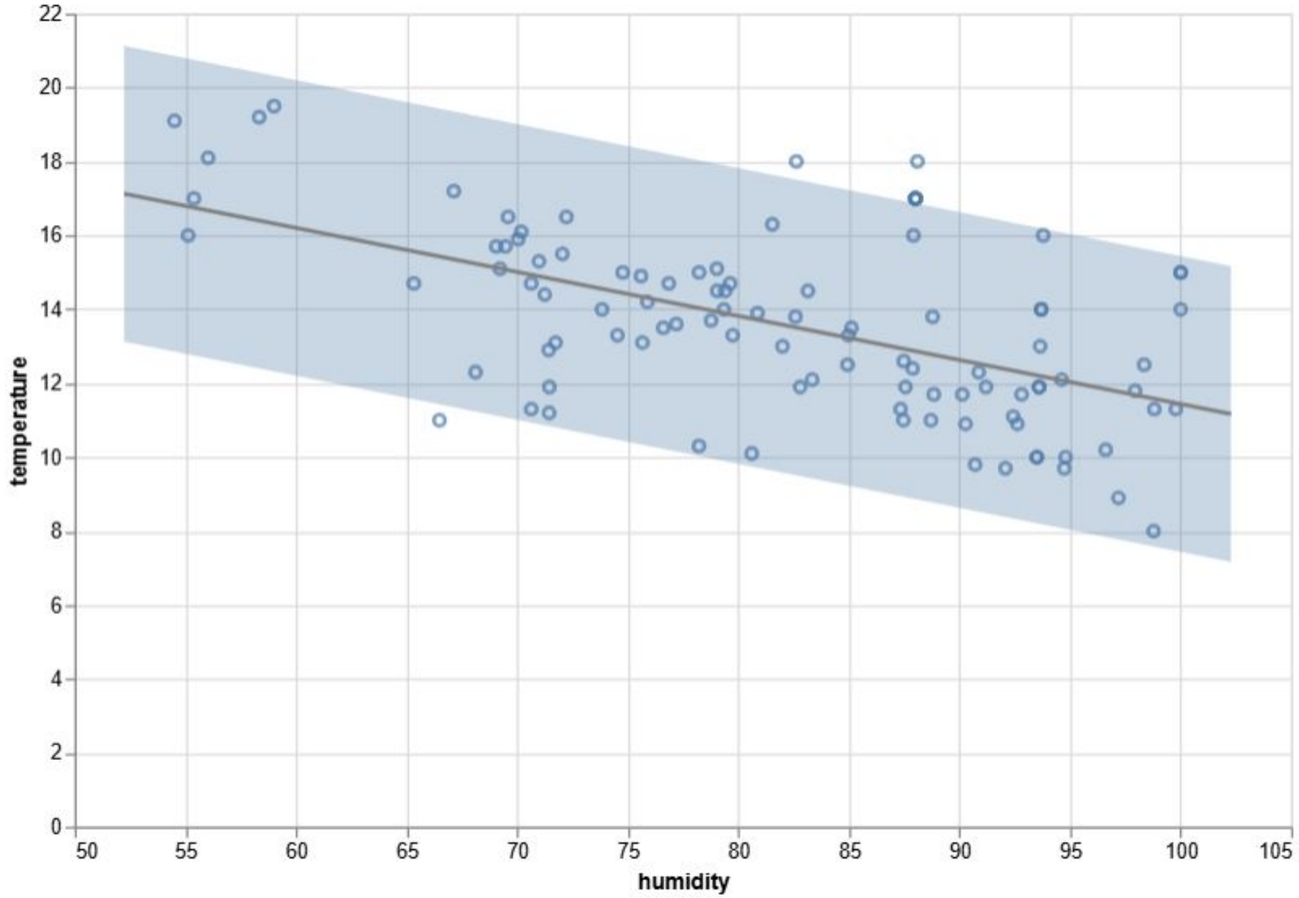
EXAMPLE 2: PREDICTING WEATHER TASK

Develop model



EXAMPLE 2: PREDICTING WEATHER TASK

Develop model



Response variable: temperature

Explanatory variable: humidity

Linear model components

y-intercept (a): 23.35

slope (b): -0.1191

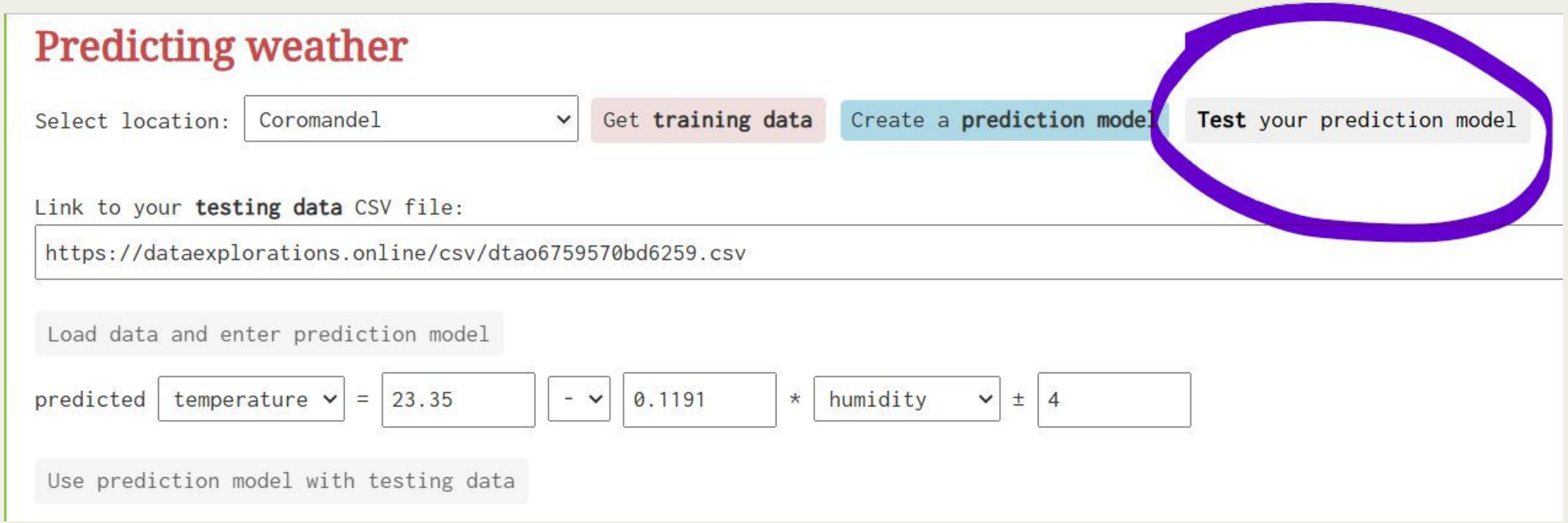
Error component

Error: 4

Update graph

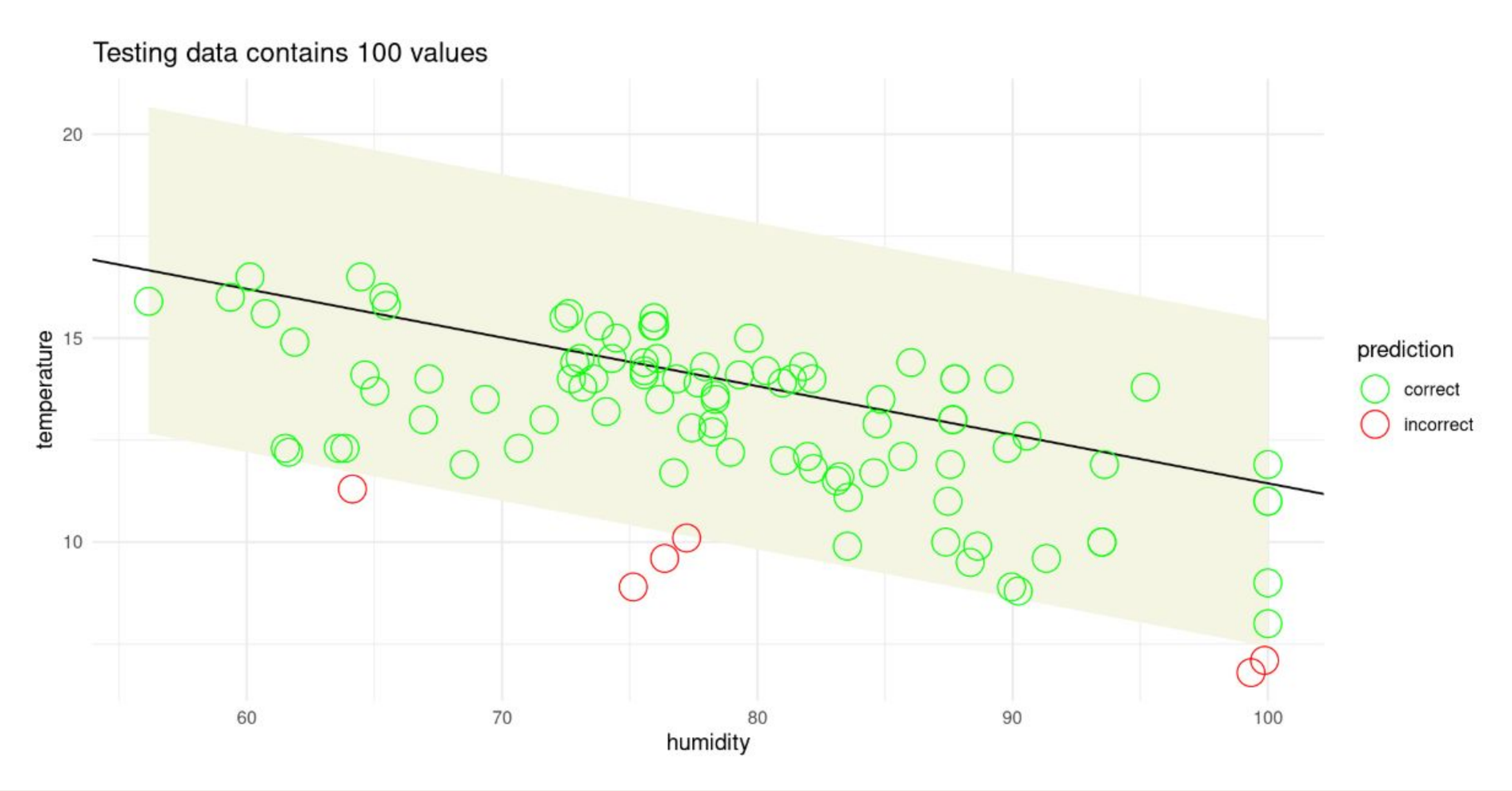
EXAMPLE 2: PREDICTING WEATHER TASK

Test model



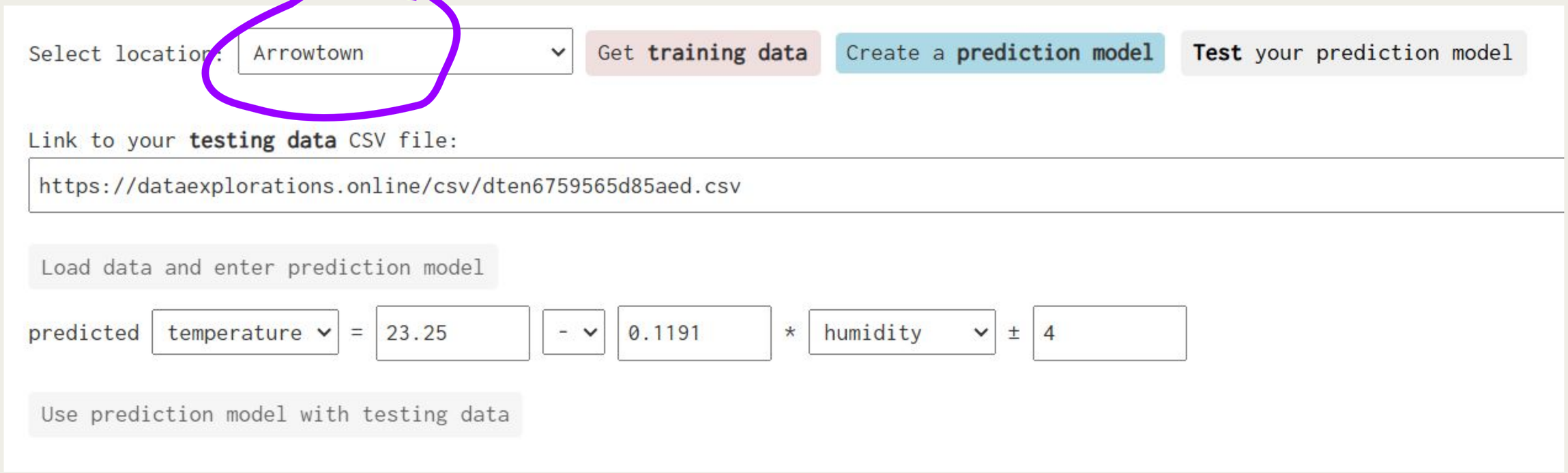
EXAMPLE 2: PREDICTING WEATHER TASK

Test model



EXAMPLE 2: PREDICTING WEATHER TASK

Test model



Arrowtown

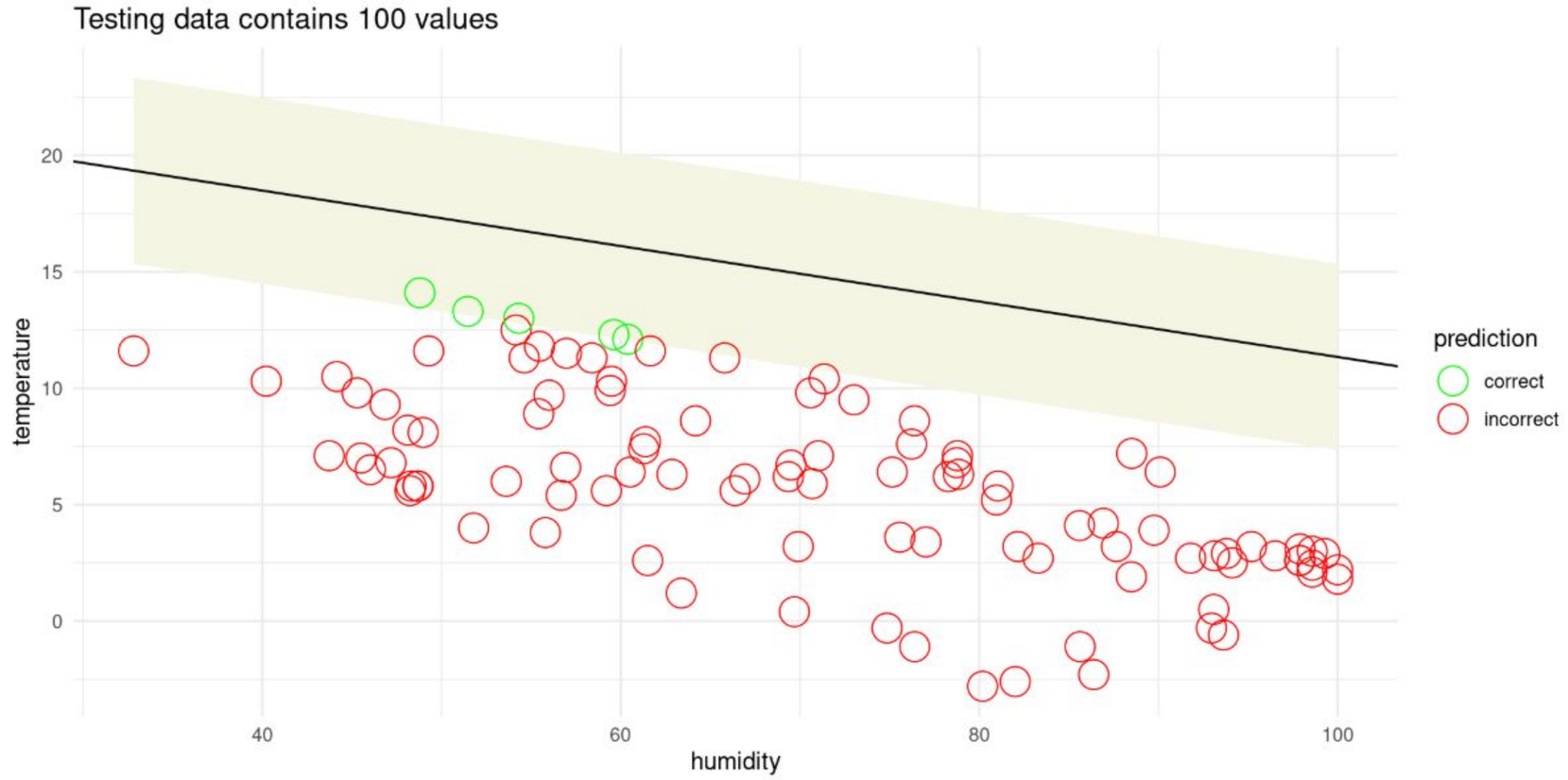
Coromandel

EXAMPLE 2: PREDICTING WEATHER TASK

Test model

predicted temperature = 23.25 - 0.1191 * humidity ± 4

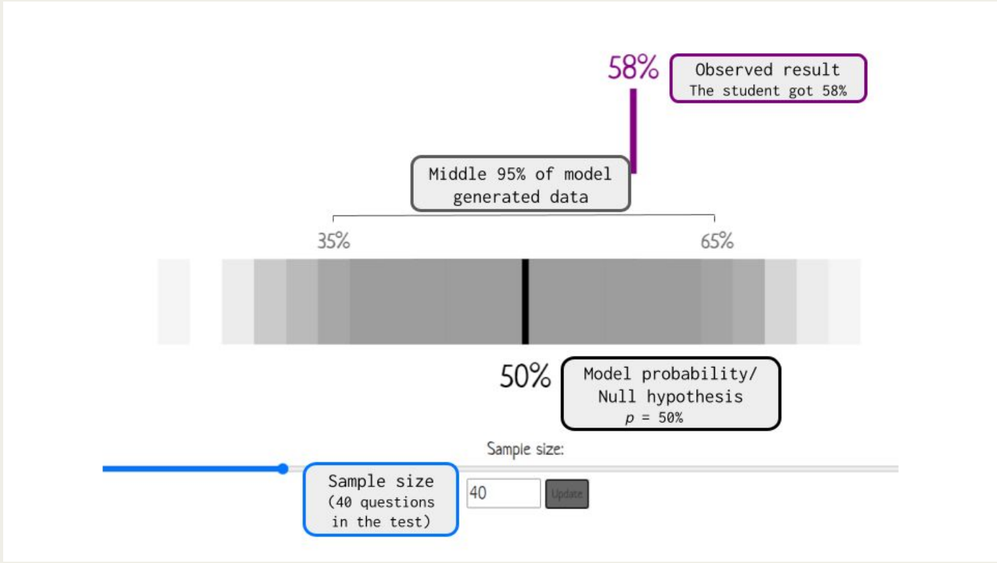
Use prediction model with testing data



CONCEPTUAL PATHWAY: EXTENDING MODELLING IDEAS (Randomisation)

Generating data from a model using simulation

Using middle 95% for “usual”

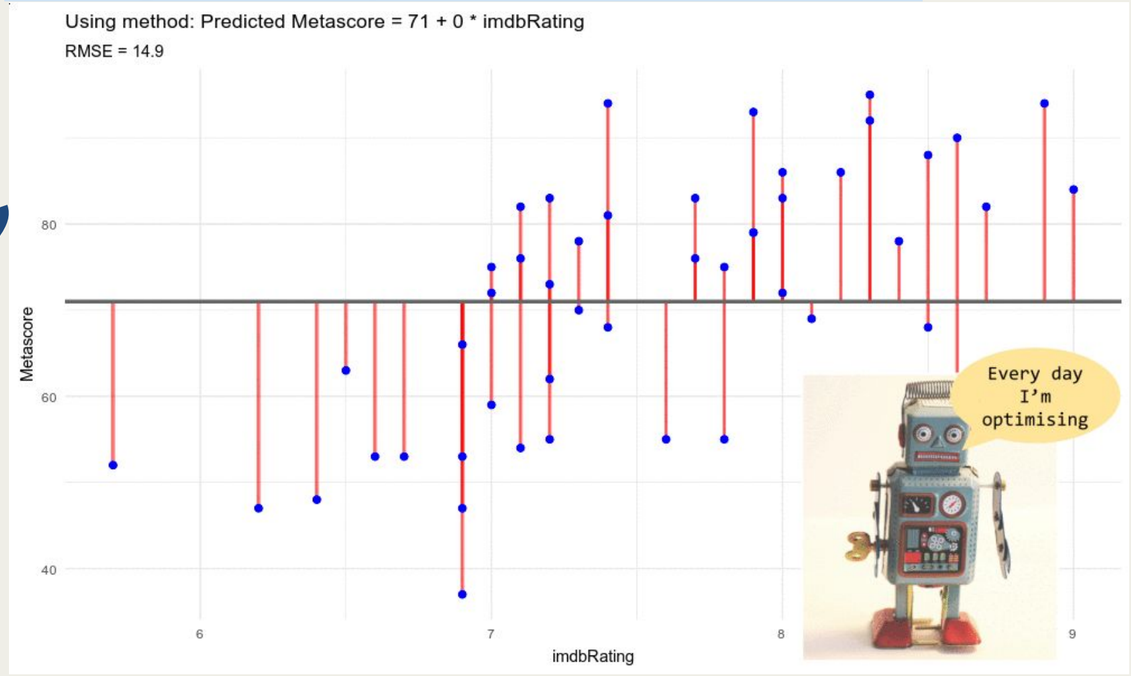


Tool: learning.statistics-is-awesome.org/threethings

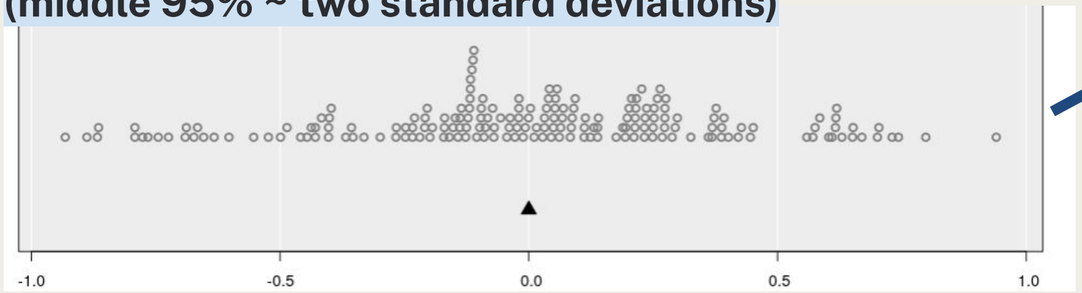
- Build on earlier ideas of models that generate data
- Continue to use visualisations and tools that support students to engage with modelling
- Emphasise that all model-generated data is conditional
- Use statistical models to help us reason with what we observe in the “real world” (non-model world)

CONCEPTUAL PATHWAY: EXTENDING MODELLING IDEAS (Regression)

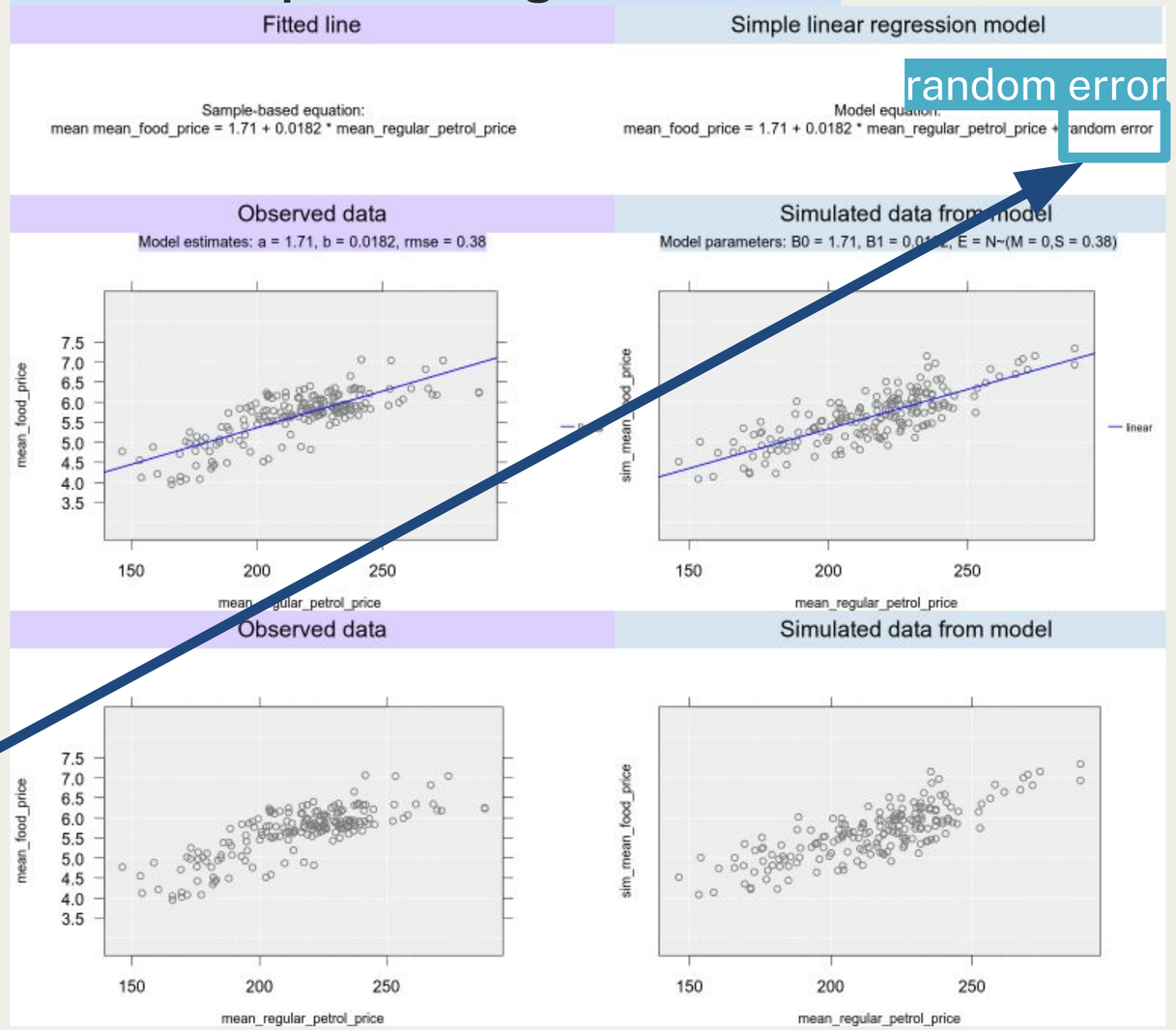
Linking RMSE to prediction errors (for SLR)



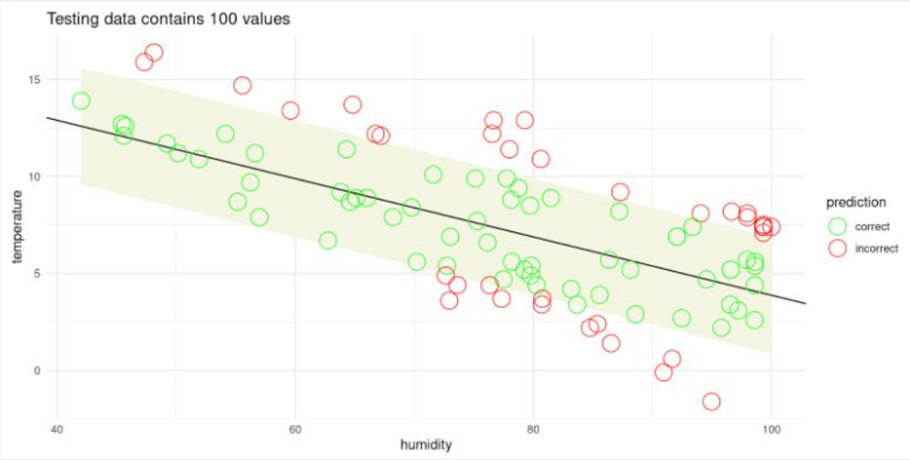
Modelling prediction errors using Normal distribution (middle 95% ~ two standard deviations)



Visualise simple linear regression model



OUR INITIAL REFLECTIONS ON STUDENT LEARNING



Potential benefits

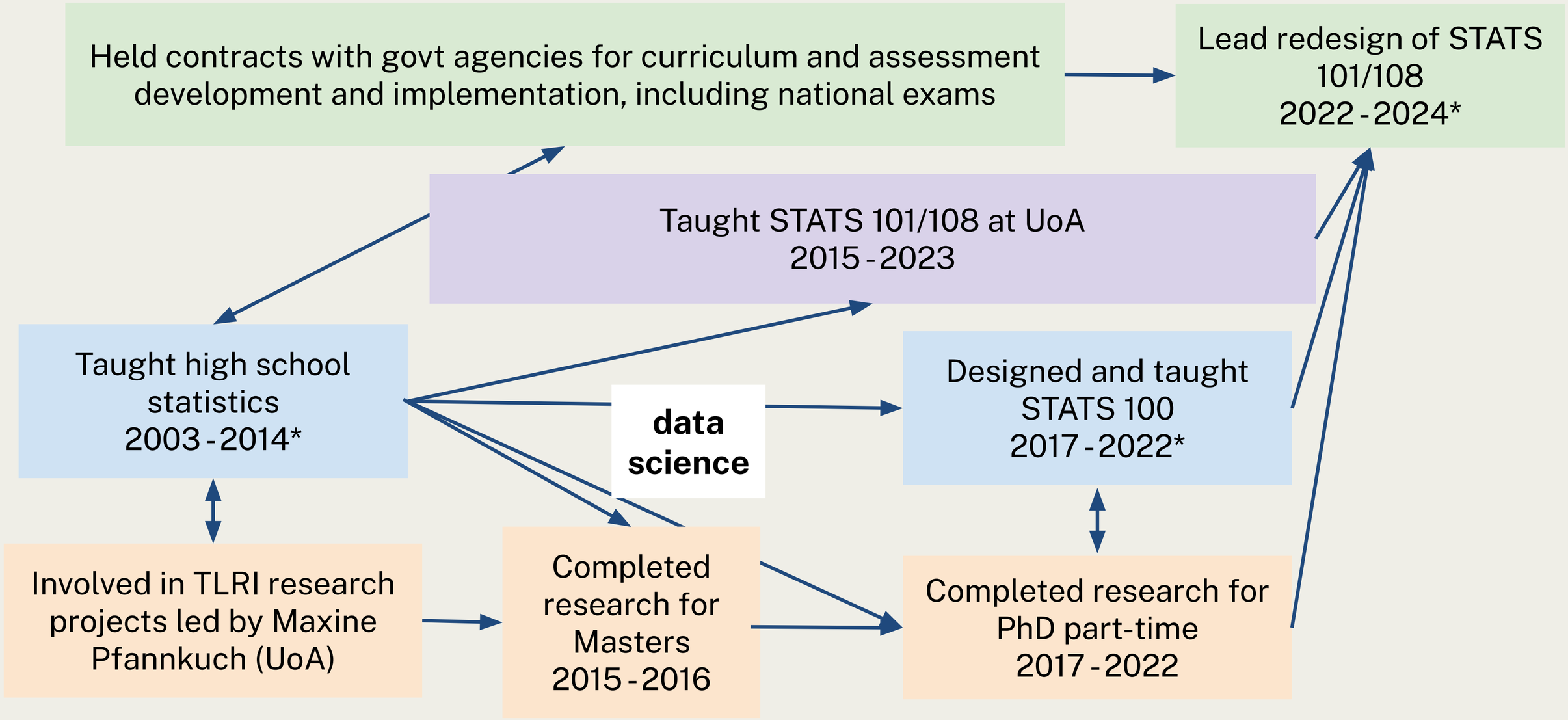
- More “concrete” to focus on individual cases and modelling situations where you can check if you got things “right” (prediction)
- Use of “data landscapes” supports students to personalise their learning and creates a variety in task responses (large scale)
- Simple modelling approach, but the learning experiences provide foundations for key ideas related to predictive modelling (delayed formalisation)

Potential challenges

- Students see lines fitted to data all the time at high school and think they already know it all (engagement)
- Students struggled more with estimating error components than we anticipated (graphicacy)
- Training/testing ideas are not intuitive for students when used to “one and done” approach and non-genuine reasons for prediction models (purpose)

✅

ESTABLISHING & SUSTAINING CONNECTIONS BETWEEN RESEARCH & PRACTICE



Introducing a data science perspective on predictive modelling within a large introductory statistics course: Connecting research with practice

Thanks to Maxine Pfannkuch, Chris Wild & Stephanie Budgett (my PhD supervisors), my research participants, the SERJ Data Science Special Issue editors and reviewers, and the STATS 101/108 re-development team (Emma Lehrke, Lars Thomsen, Anne Patel), for helping develop many of the ideas presented.

DR ANNA FERGUSSON

Te Kura Tatauranga | Department of Statistics

Waipapa Taumata Rau | University of Auckland

a.fergusson@auckland.ac.nz | e.lehrke@auckland.ac.nz

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
DEPARTMENT OF STATISTICS

References

Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2021). Modern data science with R. Chapman & Hall. https://doi.org/10.1201/9780429200717

Biehler, R., & Schulte, C. (2017). Perspectives for an interdisciplinary data science curriculum at German secondary schools. In R. Biehler, L. Budde, D. Frischemeier, B. Heinemann, S. Podworny, C. Schulte, & T. Wassong (Eds.), *Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts* (pp. 2–14). Universitätsbibliothek Paderborn.

Carlin, J. B., & Moreno-Betancur, M. (2023). On the uses and abuses of regression models: a call for reform of statistical practice and teaching. *arXiv preprint arXiv:2309.06668.* https://arxiv.org/abs/2309.06668

Fergusson, A., & Bolton, E. L. (2018). Exploring modern data in a large introductory statistics course. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* International Statistical Institute. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_3C1.pdf?1532045286

Fergusson, A., & Pfannkuch, M. (2022). Introducing high school statistics teachers to predictive modelling and APIs using code-driven tools. *Statistics Education Research Journal, 21*(2). https://doi.org/10.52041/serj.v21i2.49

Fergusson, A., & Wild, C. J. (2021). On traversing the data landscape: Introducing APIs to data-science students. *Teaching Statistics, 43*, S71-S83. https://doi.org/10.1111/test.12266

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM, 44*(7), 883–898. https://doi.org/10.1007/s11858-012-0447-5

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal, 16*(1), 22–25. https://doi.org/10.52041/serj.v16i1.209

Hardin, J. (2018). Dynamic data in the statistics classroom. *Technology Innovations in Statistics Education, 11*(1). https://doi.org/10.5070/T5111031079

Makar, K., & Rubin, A. (2018). Learning about statistical inference. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 261–294). Springer. https://doi.org/10.1007/978-3-319-66195-7_8

Wild, C. J. (2015). Further, faster, wider. Online discussion of Cobb, G.W. (2015), “Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up”. The American Statistician, 69, 266–282. https://doi.org/10.1080/00031305.2015.1093029

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x