Quantitative and Qualitative Analysis

What's out there and what you need to know.

Devan Becker
Public Health Agency of Canada; National Microbiology Laboratory

February 7th, 2022
For Western Postdoctoral Scholars

Introduction

Land Acknowledgement

UWO exists on the traditional territories of:

- Anishinaabek
- Haudenosaunee
- Lūnaapéewak
- Chonnonton

These lands are connected with the London Township and Sombra Treaties of 1796 and the Dish with One Spoon Covenant Wampum.

About Me

- Education
  - B.Sc. Math (Laurier)
  - M.Sc. Stats/Biostats
  - Ph.D. Stats
- Work
  - Postdoc - SARS-CoV-2 in wastewater
  - Public Health Agency of Canada
- Life
  - Music, reading, outdoorsy stuff
  - Crying about the housing market

Foreshadowing

Outline

- **Quantitative:** Dealing with numbers
  - Any number in a range
  - Only 0's and 1's (maybe a 2)
  - Things we can turn into numbers
- **Qualitative:** Dealing with descriptions
  - Using your brain
  - Using your computer
- **Meditative:** Dealing with everything
  - How to get started
  - Accessing resources
  - Not being afraid of coding

Before we begin

- Interrupt at any time
- All notes/links/resources/R code are on GitHub
- Ask future questions in the PAW Slack chat
- I have allowed myself **ONE** equation.

```
The GitHub version also has my (approximate) script inside
:::notes::: tags, which show up as text in the pdf.
```

What to watch for

Keep an eye out for the following concepts:

1. Garbage In, Garbage Out (GIGO)
2. Numerical summaries lie - you need plots!
3. Models are models.
4. Models are wrong.

Regression

Terminology

- The **Target** could be any number in a range.
  - A.k.a. dependent variable or response.
- The Features could be any data type
  - A.k.a. explantory or independent variables (IVs)

The Data

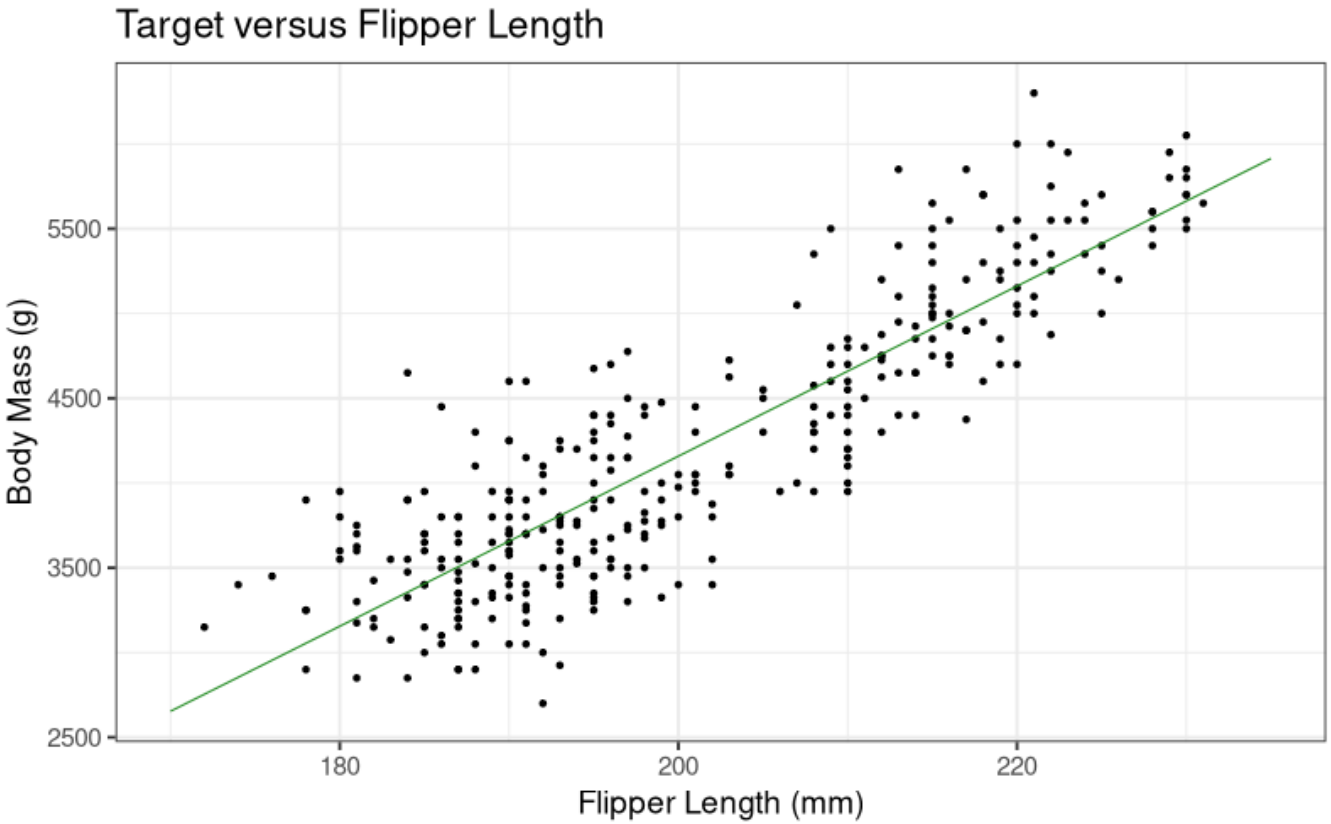
mass	bill_len	flipper_len	species	island	sex
3750	39.1	181	Adelie	Torgersen	male
3800	35.3	187	Adelie	Biscoe	female
4150	42.0	210	Gentoo	Biscoe	female
5350	48.7	222	Gentoo	Biscoe	male
3725	52.7	197	Chinstrap	Dream	male
3750	51.3	197	Chinstrap	Dream	male
3400	50.1	190	Chinstrap	Dream	female

Goal

How does body mass change with flipper length?

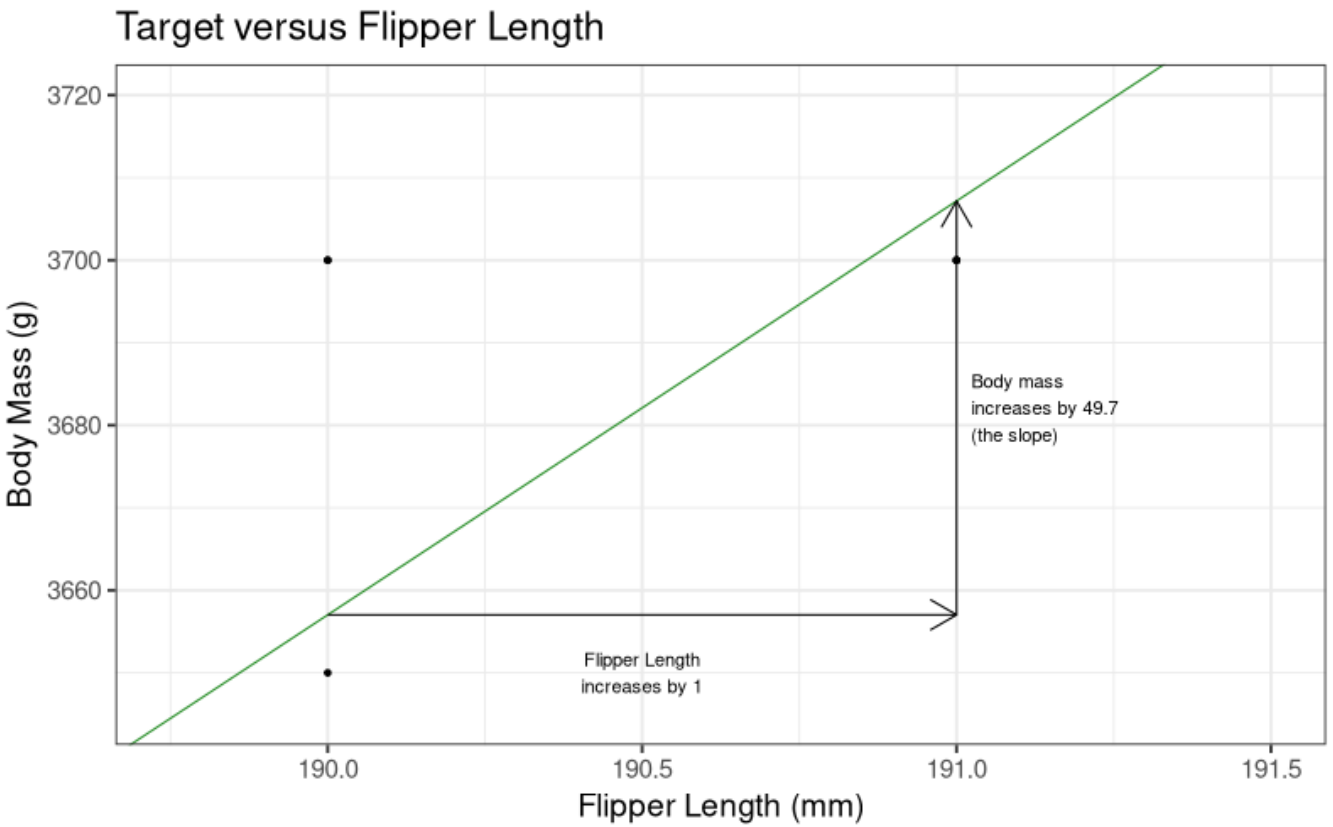
- The **slope** quantifies this change
- If the data are good, estimated slope is similar to *population* slope

Intro to linear models



- Find the slope and intercept to best fit the cloud of points.
  - Slope: rise over run.
  - Intercept: the body mass when flipper length is 0.

Linear models: slopes



Binary Features

Suppose we have a variable that is labelled either 0 or 1.

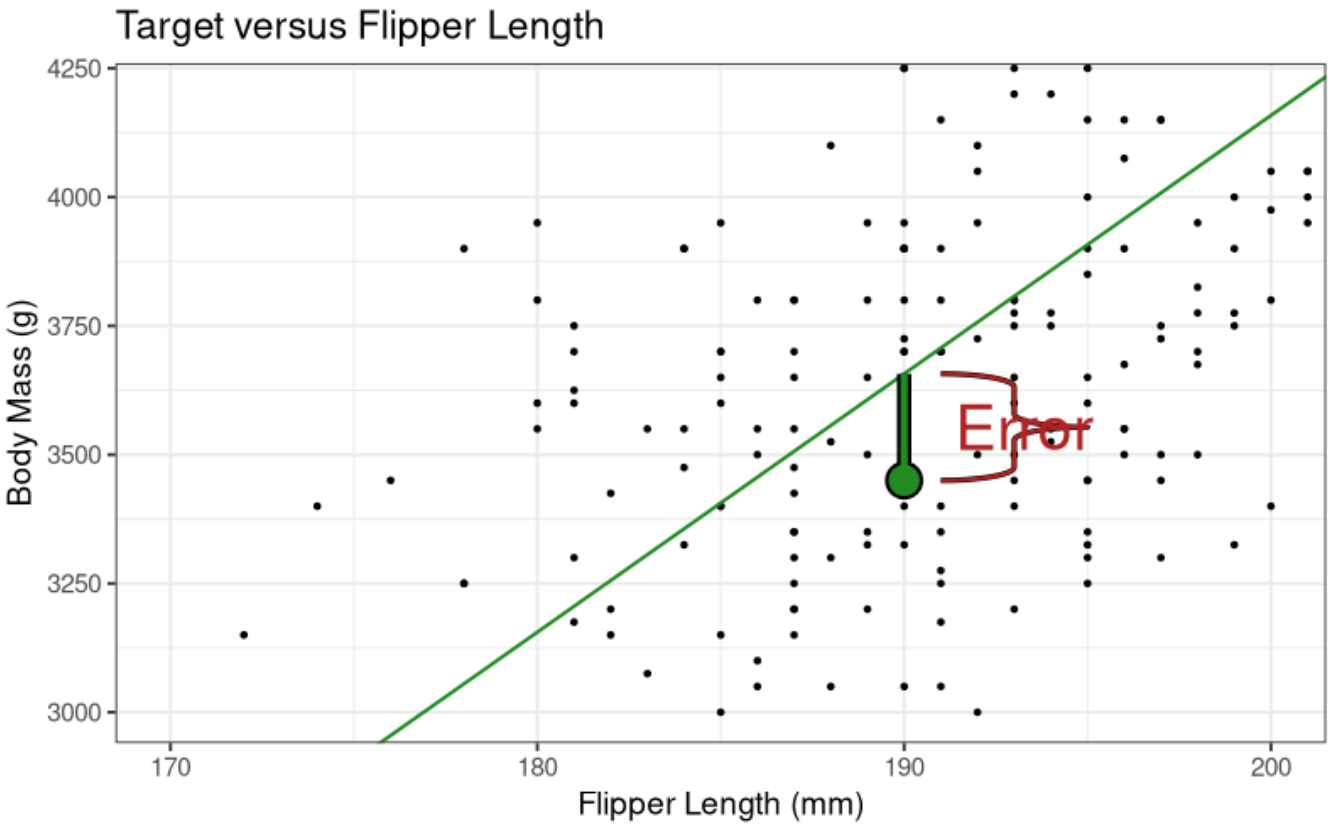
What does the slope represent?

The story so far

- The Intercept is a mathematical necessity
- The Slope answers our questions

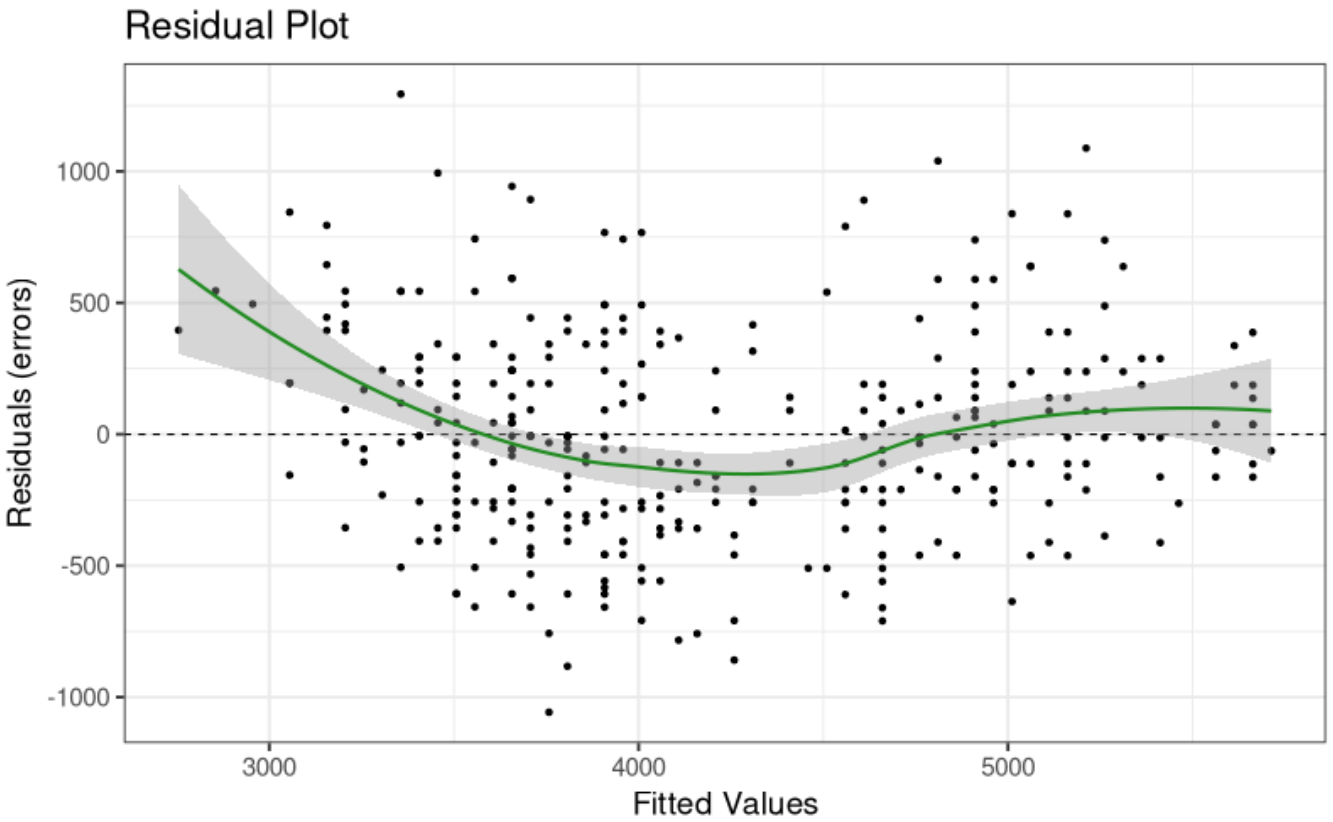
But how good is our model?

The most important part!



- The line will never go through every point perfectly!
- Know where the model fails can tell you everything!

Residual plots: residuals versus predicted



- A perfect residual plot should show no pattern.
- This plot looks like there's a slight pattern. . .

The pattern

Each species has a slightly different relationship!!!

Putting it all together

1. Get data
   - Data cleaning is the hardest part.
2. Check data
   - If you haven't plotted it, you're doing it wrong.
3. Fit model
4. Check model
   - If you haven't plotted it, you're doing it wrong.

Other

- Mixed Models

Machine Learning

What is Machine Learning?

- Statistics, but done by a computer scientist...

–OR–

- Anything algorithm that tries to get information from data!

This includes linear regression!

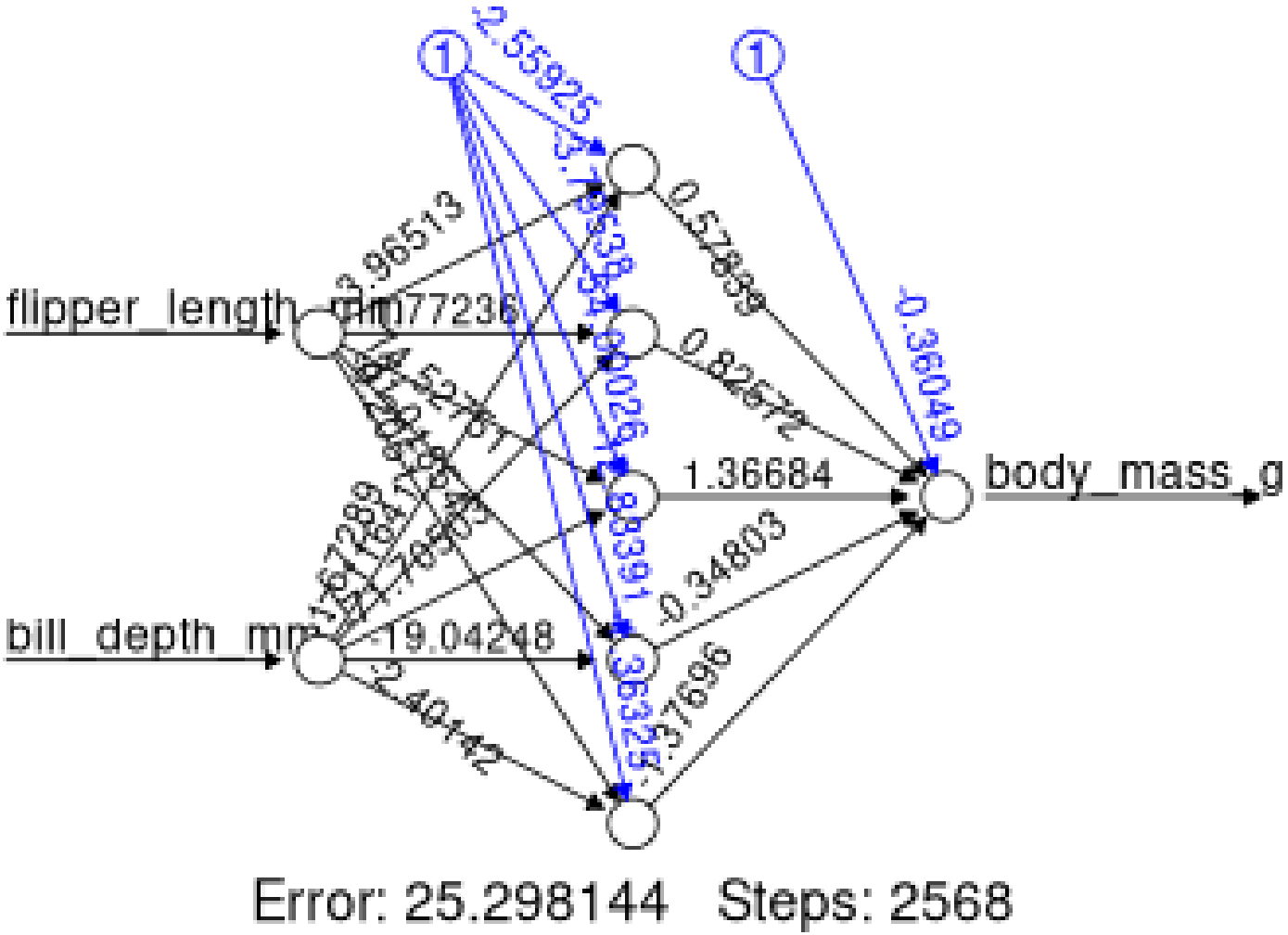
Regression in Machine Learning

- **Lasso Regression**
  - It's like linear regression, but it automatically removes features.
  - Related: Ridge regression, ElasticNet
- **xgBoost**
  - Remember the residual plots? What if we fit a regression to those residuals?
- **Neural Nets**
  - ...

Neural Nets

- What most people think of as ML.
  - Deep Learning: fancy neural nets.
- Loosely based on the way synapses work.
- Just a bunch of linear regressions

Neural Net Setup

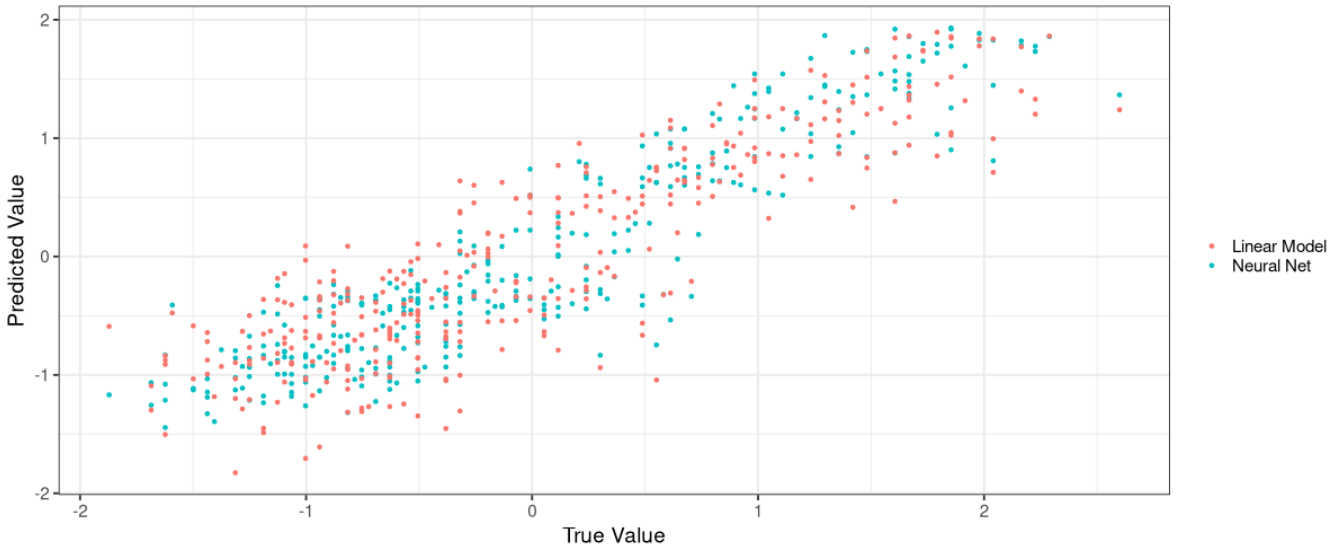


Linear Models Versus Neural Nets

- LM gives interpretable slopes
- NN accounts for complex interactions

- LM is better for **inference**
- NN is better for **prediction**

Is NN always better than LM?



No.

ML and Ethics

- ML finds patterns that exist
  - It perpetuates existing patterns, e.g. black recidivism
- ML is hard to audit
  - Is it just looking at peoples' race? Hard to say!
  - Explainable AI (XAI)
- ML doesn't answer email
  - Why did it make a certain decision?
  - Ca't plead your case.

https://delphi.allenai.org/?a1=Using+AI+to+determine+ethics

Classification

Binary Target

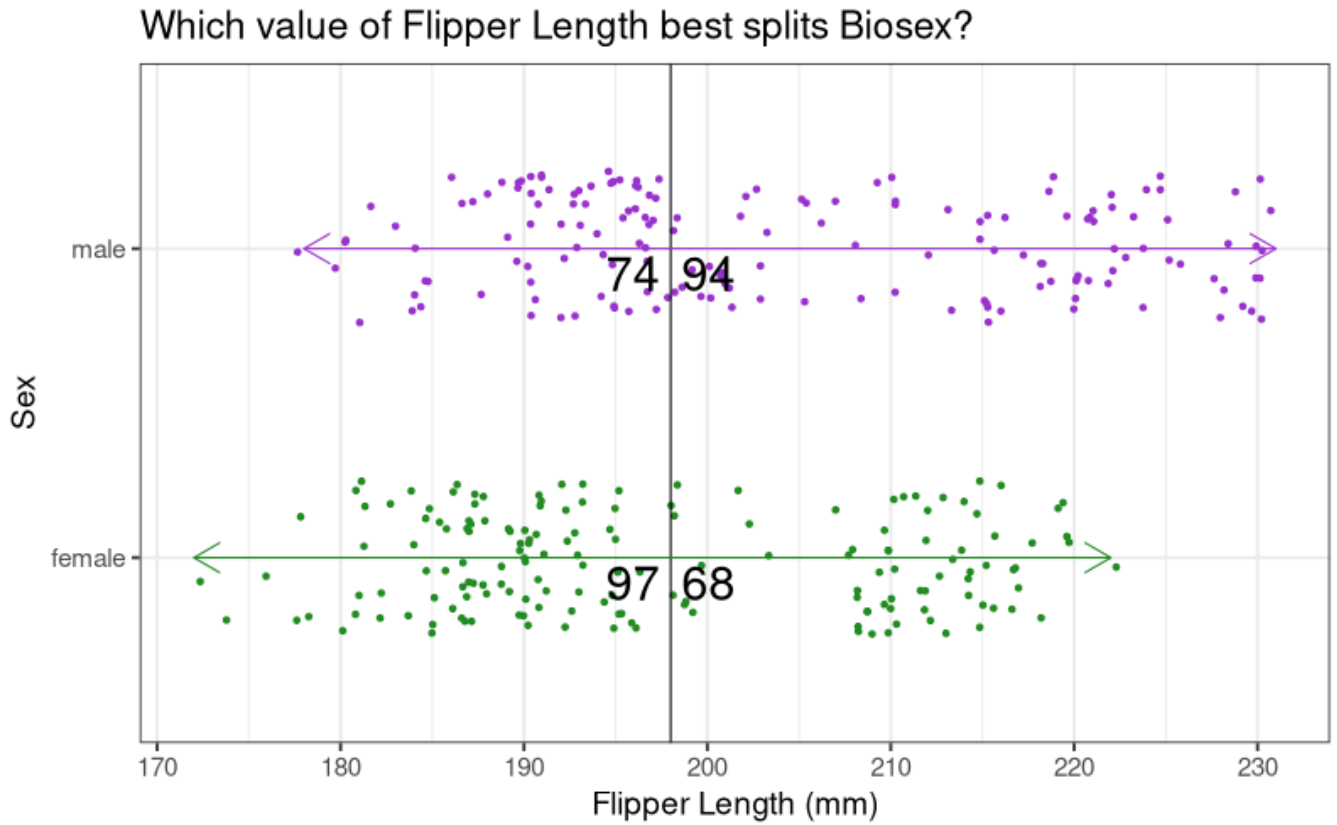
mass	bill_len	flipper_len	species	island	**sex**
4500	46.1	211	Gentoo	Biscoe	female
5150	46.8	215	Gentoo	Biscoe	male
4600	48.2	210	Gentoo	Biscoe	female
5400	48.4	220	Gentoo	Biscoe	male
4200	45.5	210	Gentoo	Biscoe	female
5550	50.4	224	Gentoo	Biscoe	male

Ethics: biosex versus gender

- Chinstrap penguins have a higher-than-average occurence of homosexual behaviour.
  - Tufts University, Feb 2021: "What's With All the Gay Penguins?"
- Gentoo penguins have less rigid gender roles.
  - NBC news, Sept 2019: "Gay penguins at London aquarium are raising 'genderless' chick"
- Adelie penguins of any gender all want to be like Adele
  - <Citation Needed>

"Biosex" is a fundamentally imperfect measurement of gender roles.

Choosing between two options



- When Flipper Length is below 198, most are female.
- This is called SVM, or Support Vector Machines

But how were we wrong?

If we label any penguin with Flipper < 198 as female:

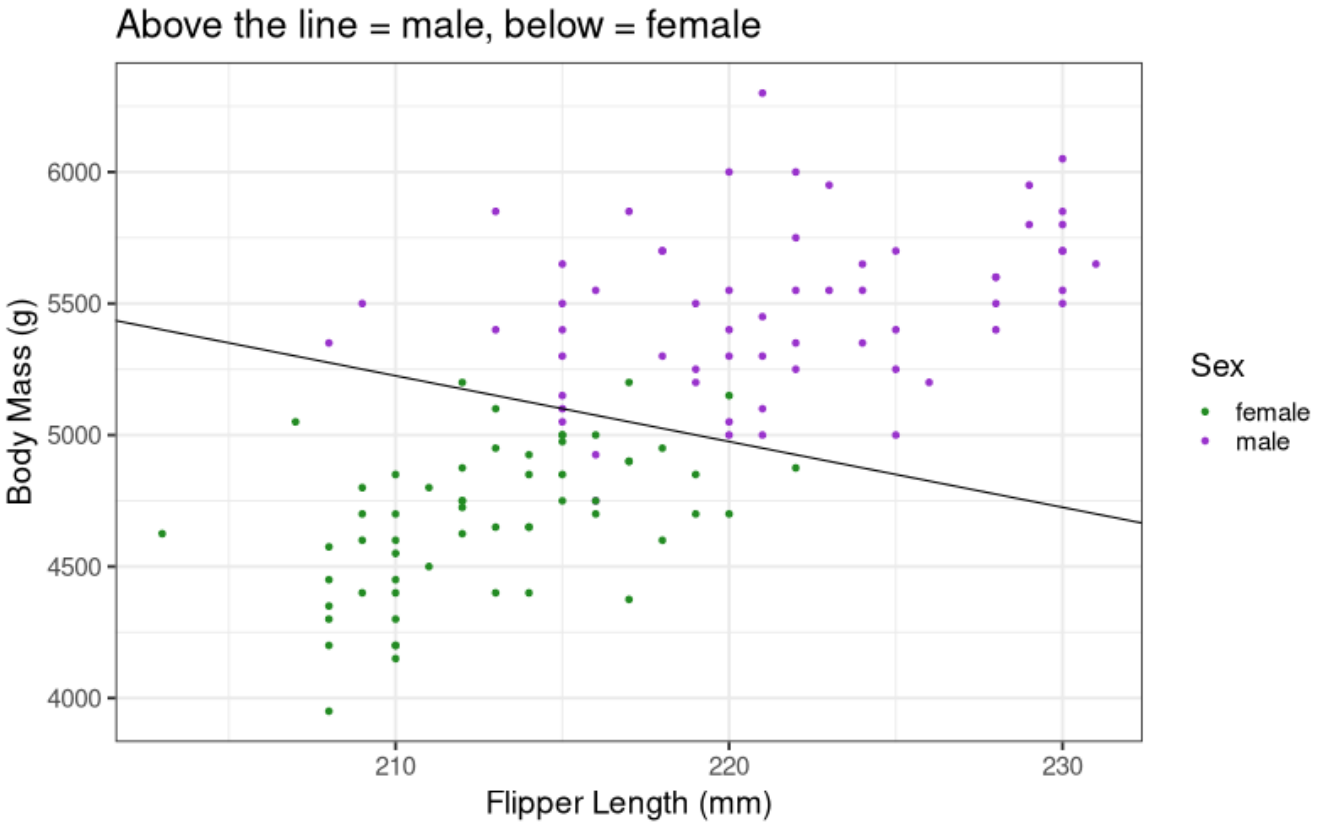
	Flipper < 198	Flipper ≥ 198
Male	74	94
Female	97	68

- When we label them female, they're actually female 97/(74+97)=56% of the time.
- When they're actually female, we label them female 97/(68+97)=58% of the time.

This is a **Confusion Matrix**.

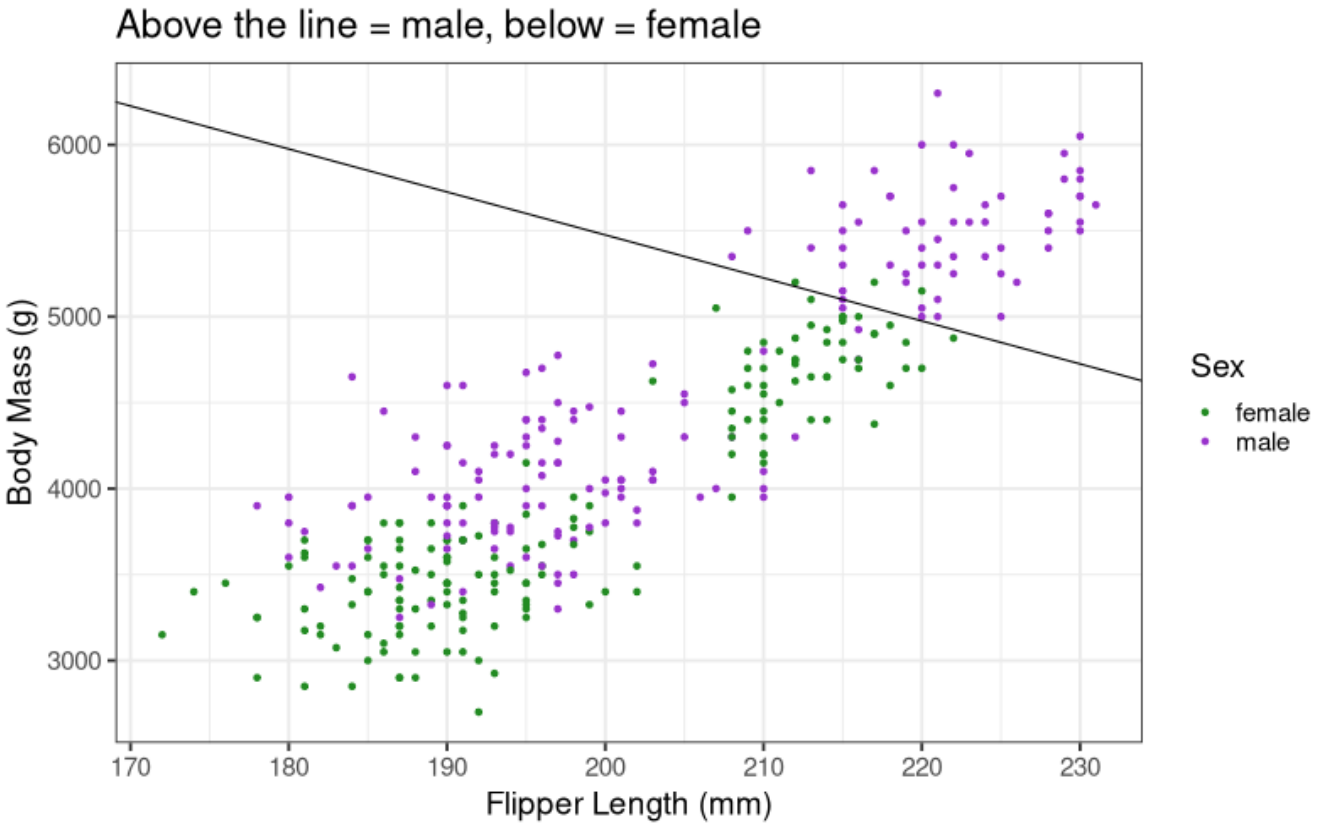
See also: sensitivity, specificity, precision, recall, F1 score, ROC/AUC curves.

More dimensions!

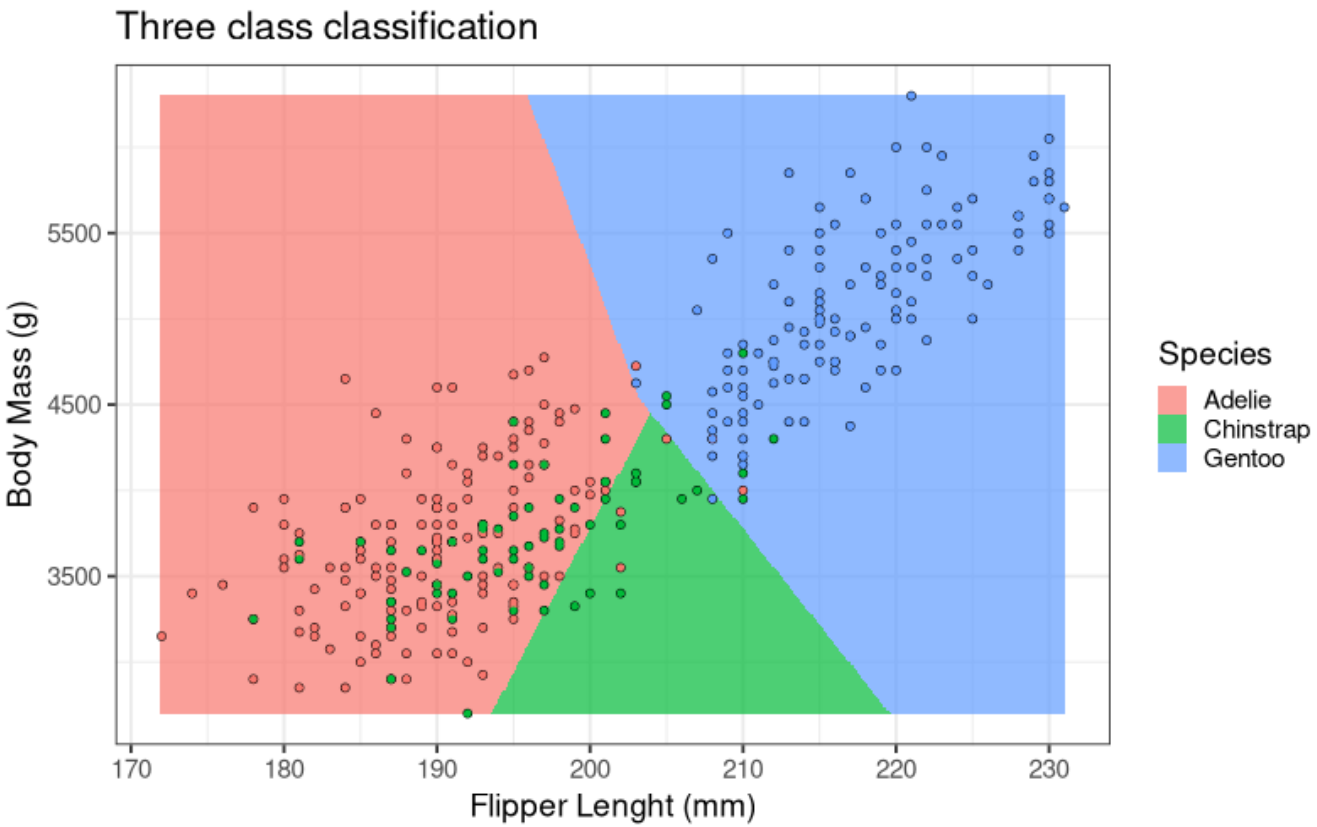


- With more information we can fit a better model!
- ...

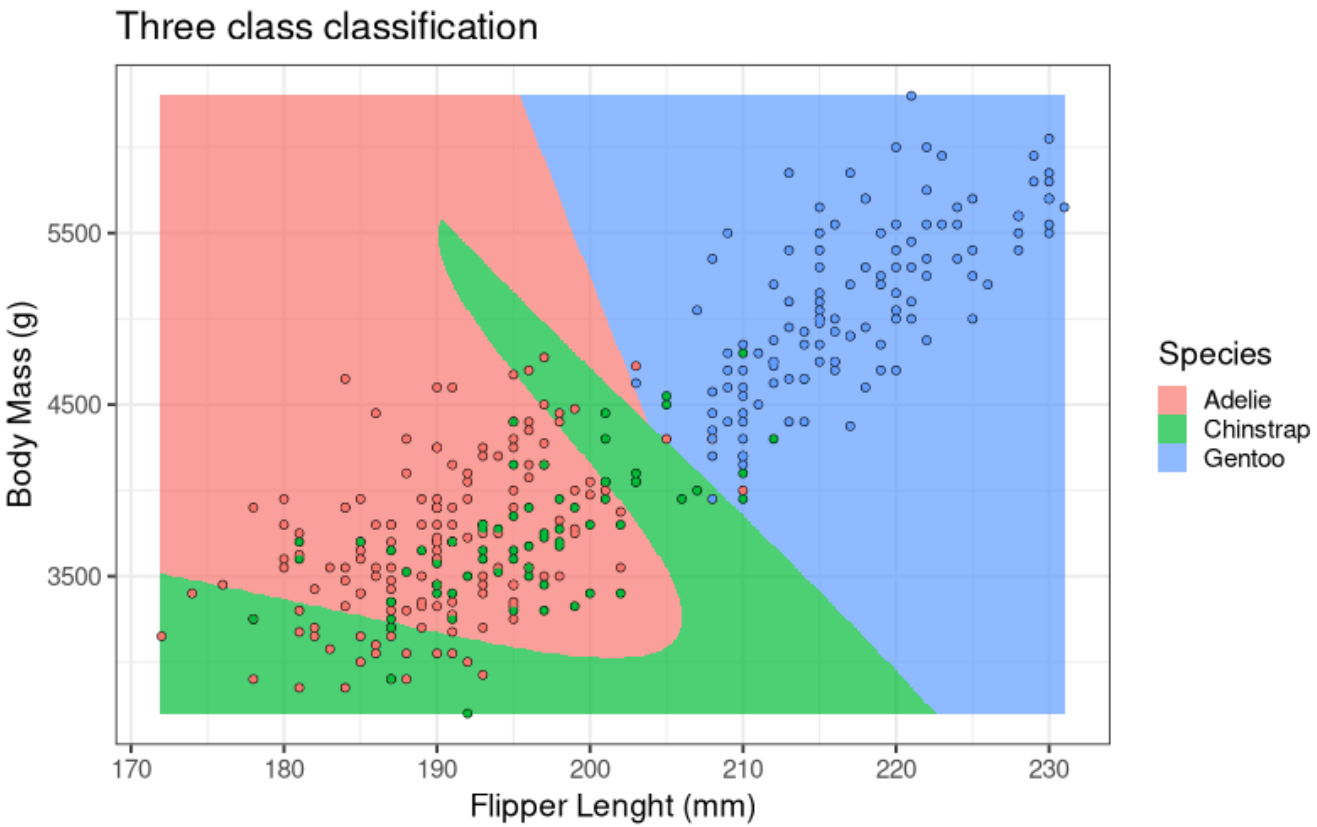
. . . but there's a reason I only used Gentoo



Three categories: Species



It doesn't need to be linear!



Other Classification Models

- **Logistic Regression**
  - Basically, LM for probabilities
    - Interpretable results!
  - For multiclass, Multinomal Regression
- **Decision Trees** and **Random Forests**
  - Very important models that I'm not covering
  - Still a linear model at heart
- Naive Bayes Classifiers
- K-Nearest Neighbours (KNN)
- Neural Nets!

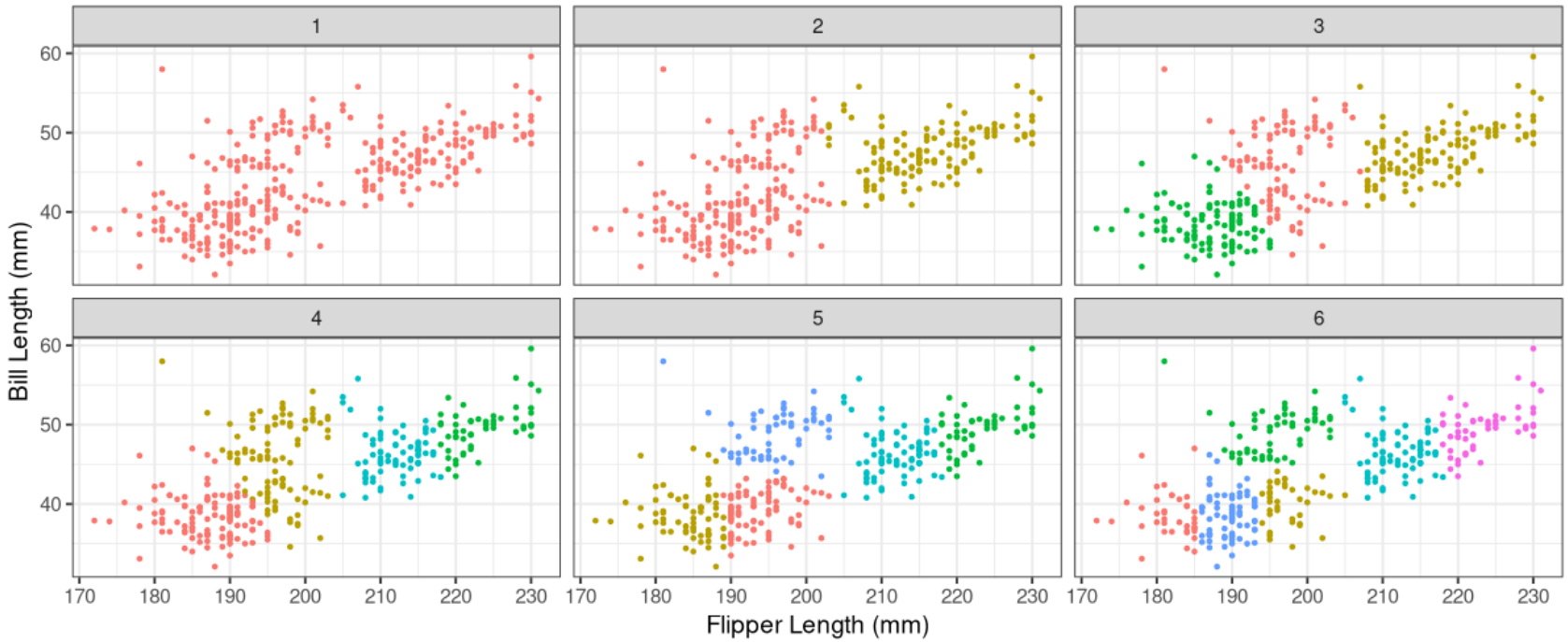
Unsupervised Learning

Definition

In classification, we're predicting labels *and checking if we're right.*

Unsupervised learning means *we don't know the labels*.

K-means Clustering



- Pretend that Species info is *NOT* available.
- How many clusters are there?

Dimension Reduction

Motivation

Why use many features when few features do trick?

By combining features, we might:

- Find out which features have similar effects on the target.
- Find clusters

Principal Components Analysis (PCA)

A *Principle Component* is a combination of the features (NOT target).

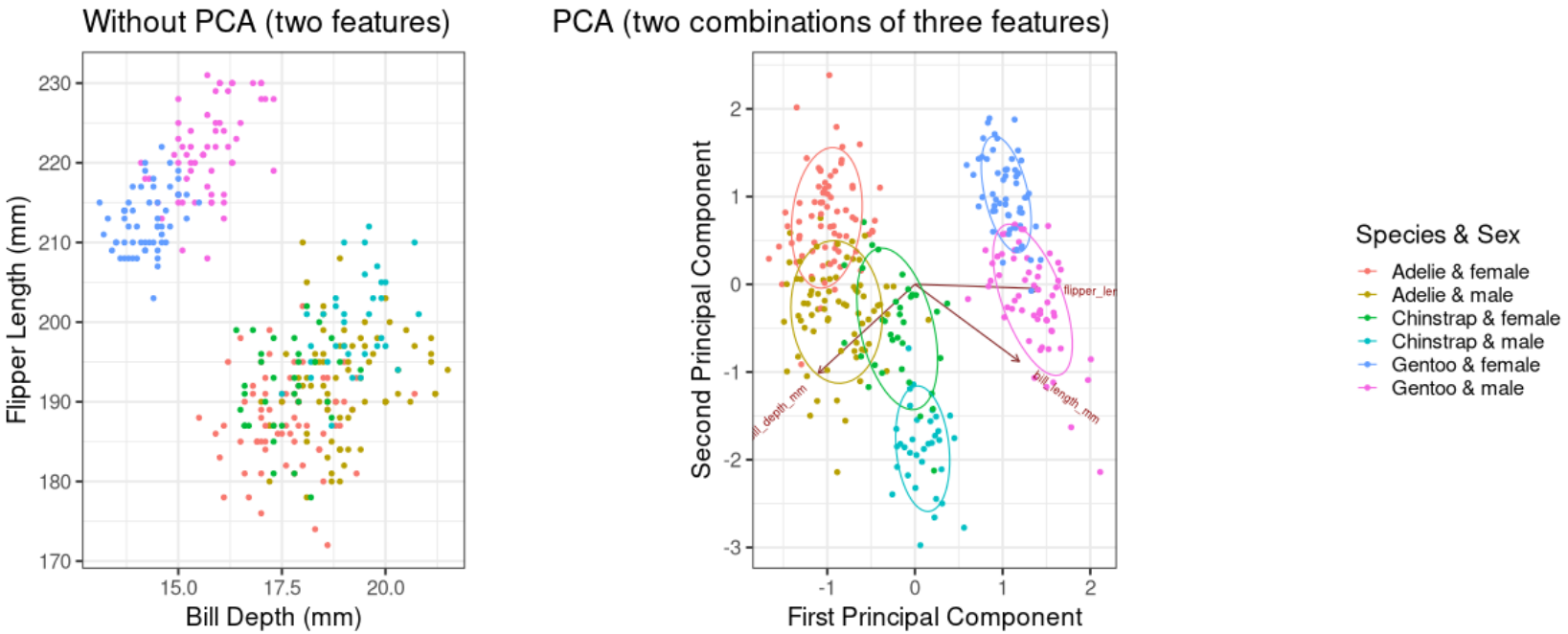
Each component is unrelated to the others.

$$PC1 = 0.55 * bill\_length - 0.51 * bill\_depth + 0.65 * flipper\_length \quad (1)$$

$$PC2 = -0.65 * bill\_length - 0.75 * bill\_depth - 0.03 * flipper\_length \quad (2)$$

... (3)

Principle Components - clustering



Qualitative Analysis

Qualitative Data

Quality: the properties/characteristics of a thing (not numbers)

- Survey responses
  - "A lot of people seem to talk about painful things . . . "
- Categories
  - "Registered democrats tend to have these qualities . . . "
- Texts
  - "The grammar in this act is different from Shakespeare's usual style . . . "
- Concepts
  - "These documents could be categorized by their use of . . . "

Qualitative Data **Analysis**

- Fully manual: read everything, pay attention, take notes, compare.
  - I can't help you with this.
- Some computer: search within documents, word clouds, etc.
  - Audio/image/video transcription via neural networks
  - Semantic analysis
- Much computer: **Natural Language Processing**
  - It's machine learning, but for words!

Natural Language Processing

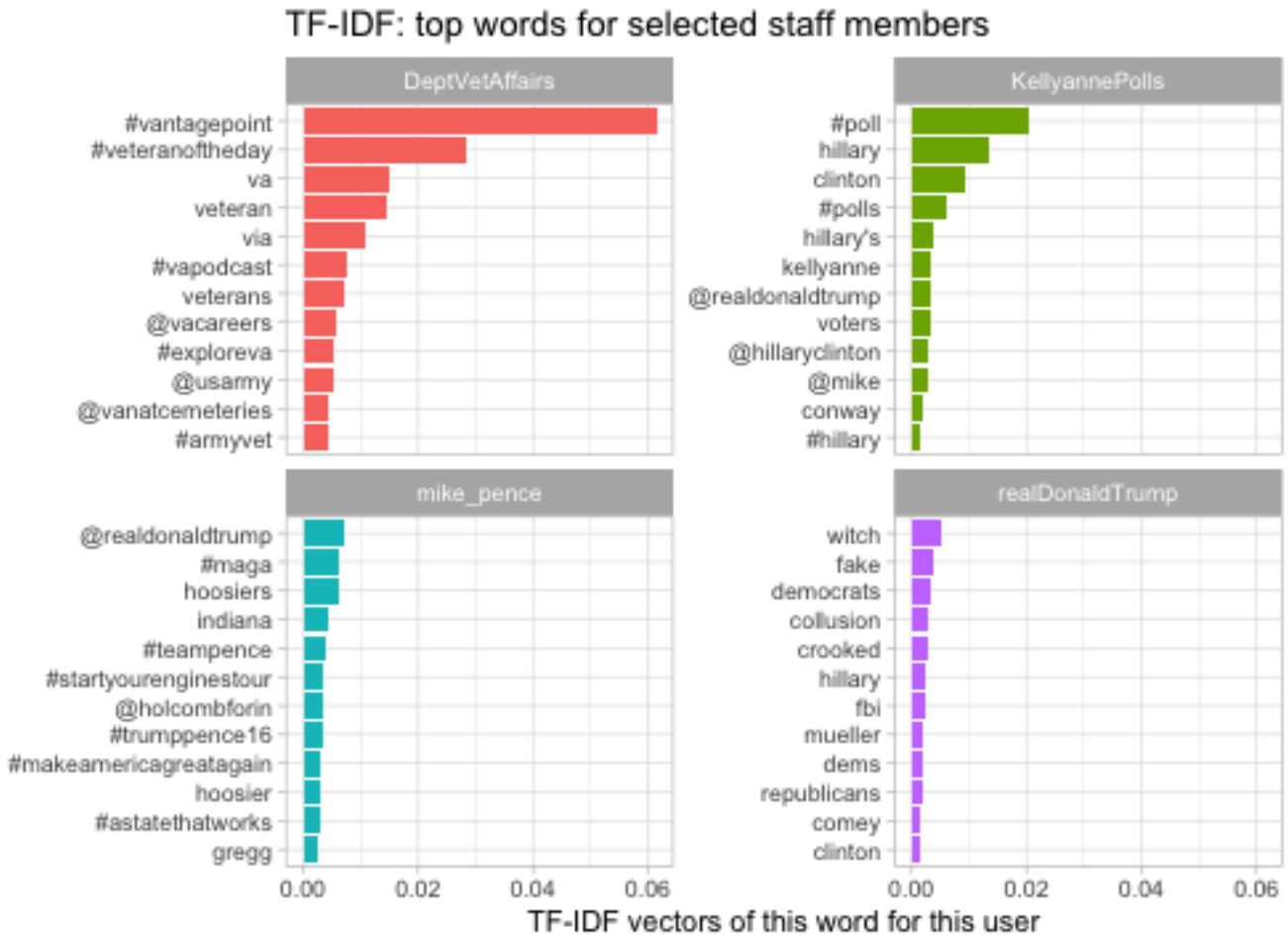
Code is perfect and English is awful

- It's cold outside, yes?
- It's cold outside, no?

Sometimes, yes and no mean the same thing.

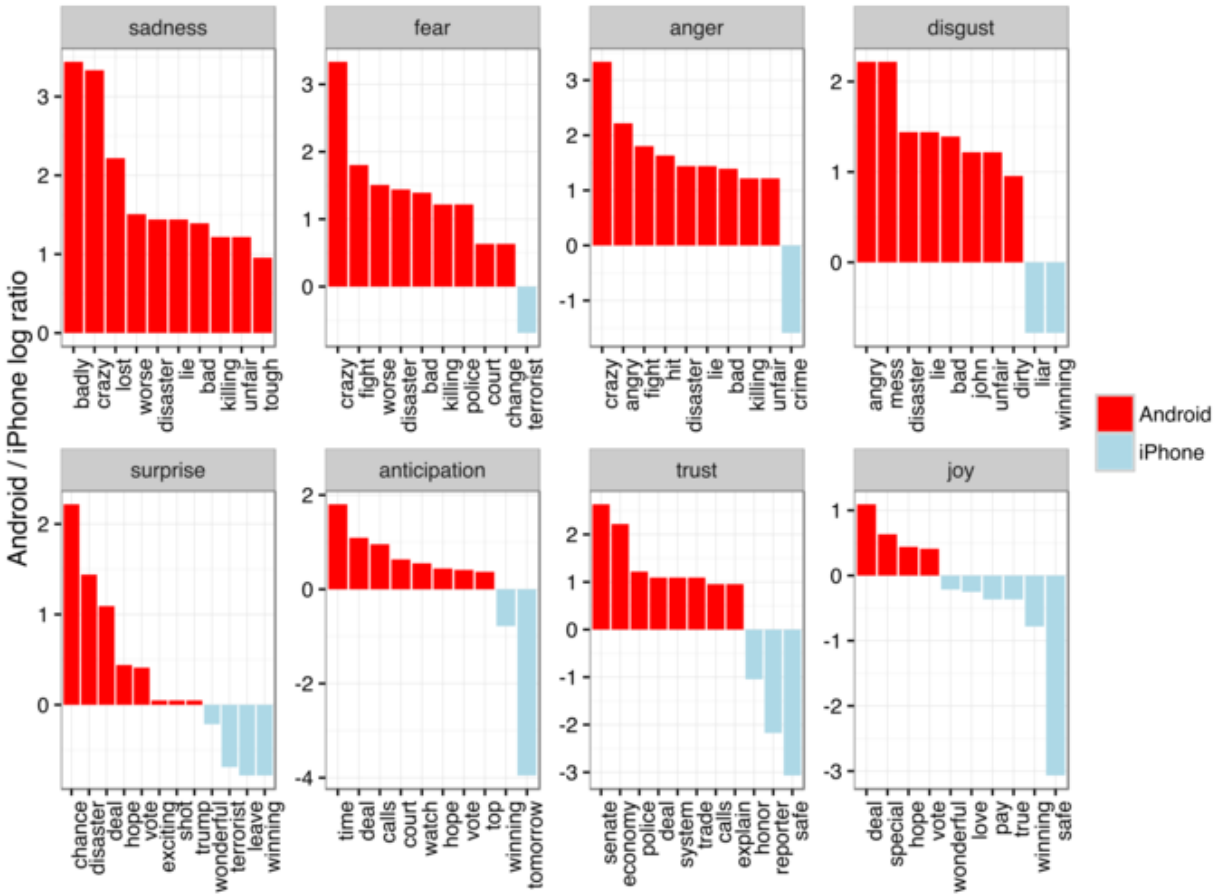
How the heck could a computer have a chance?!?

TF-IDF: Who wrote the Op-Ed?



Source: http://varianceexplained.org/r/op-ed-text-analysis/

Sentiment Analysis: Trump Uses an Android



Source: http://varianceexplained.org/r/trump-tweets/

More Advanced Natural Language Processing

- Parts of Speech
  - Nouns, verbs, etc.
- Topic modelling
  - Words that show up in similar sentences prob. have similar topics
- Bag of Words (Word2Vec)
  - How often are words used together?

All of the above can be based on Nueral Nets!

Meditative

Summary

Same ideas throughout:

1. GIGO
2. Plot everything
3. Learn to code
4. Plot everything

Learning Path

- Take notes on a basic coding tutorial
- Work through an *easy* passion project
  - Visualize olympic medals (Kaggle)
  - Basic linear model for bitcoin values
- Backpropagate your new knowledge
- Write a tutorial for yourself, share it on GitHub.
- Search Twitter, follow relevant topics/people

Important things we didn't cover

- **Data Cleaning** (don't use Excel)
- Inference versus Prediction
- **Cross Validation**
- Version control and best practices (GitHub!)
- Scrutinizing data

R versus Python versus Other

- R is stats focused
  - Python has cutting edge machine learning and general purpose
- R has `dplyr` and `ggplot2`
  - Python teaches/requires better coding skills
- RMarkdown is astounding
  - Black holes were imaged in Jupyter
- Both will work for any analysis
  - Basicaly, use what your colleagues use.

FWIW, I used R for this workshop and code is available.

Thank You!

See you on the slack chat!

Bonus Topic: Bayesian Statistics

I flip a coin and get heads.

What's the probability of heads?

- Frequentist (the usual way): 100%
- Bayesian: I though it was 50%, now I think it's closer to 60% maybe?

Probabilities

- Usual way: long term frequency (e.g., after 100 coin flips, 50 are expected to be heads)
- Bayesian: Uncertainty (e.g., I think the next flip is heads with 50% certainty)

In linear models

- I think the slope is probably 10, on average.
  - There's variance in my belief
- With a small data set, I move my belief closer to what the data says
  - Larger data means estimate is closer to data AND smaller variance.

The *posterior* distribution comes from updating the *prior* with the data (likelihood).

Why Bayesian?

- The posterior is a *distribution*, not a point estimate
- A 95% *credible interval* contains the true mean 95% of the time!
- Much, much, much more flexible models
  - Especially mixed moe