Using Big Data to Find Average Cat Weights
Using Machine Learning to Analyze Big Data

Adam Campigotto
DVM, MSc Candidate

Advisors: Dr. Theresa Bernardo, Dr. Elizabeth Stone,
Dr. Zvonimir Poljak, Dr. Deborah Stacey

Overview

Introduction

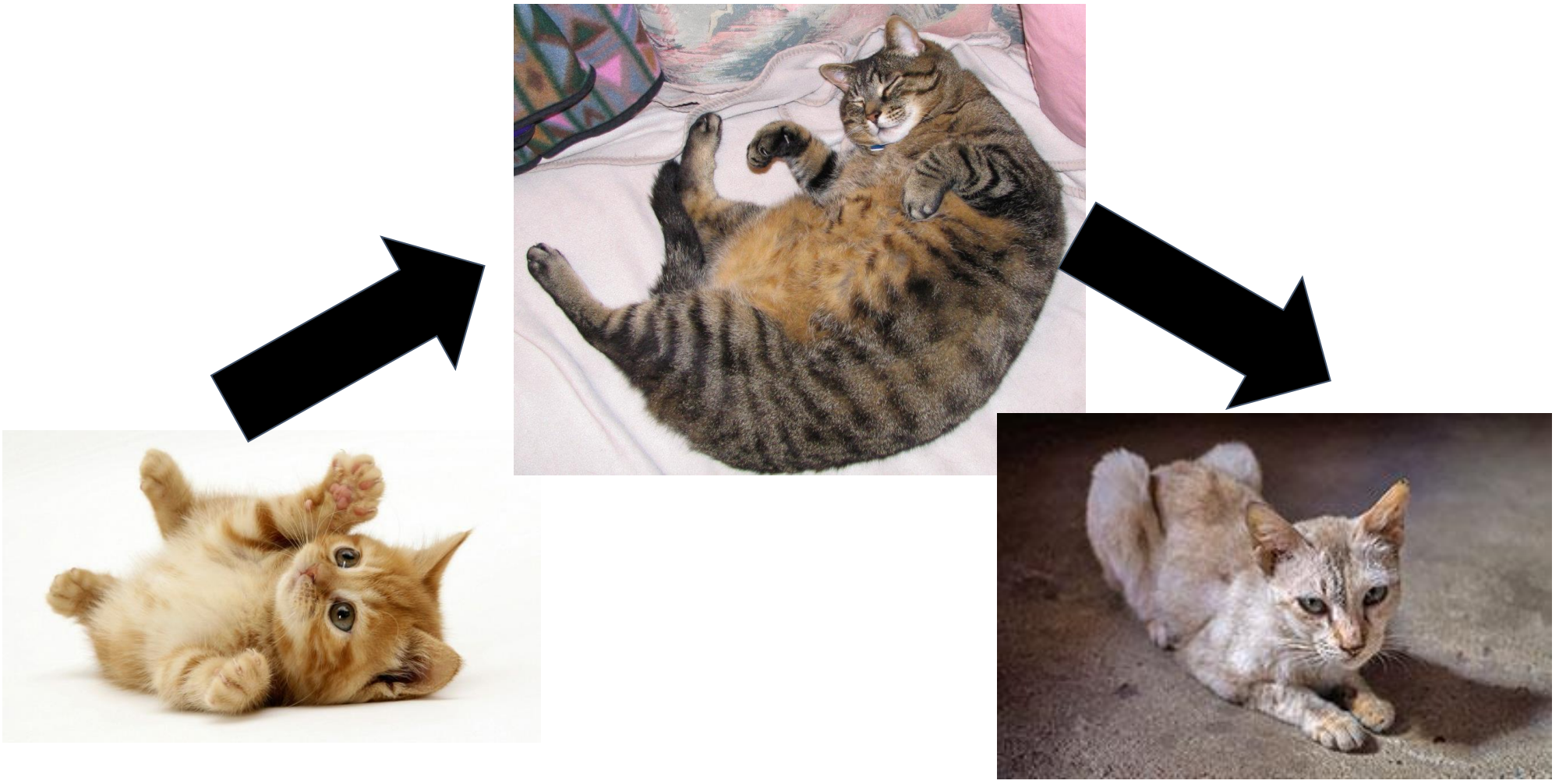
- Cat Weight
- Big Data

Data Collection and Cleaning

Methods & Results

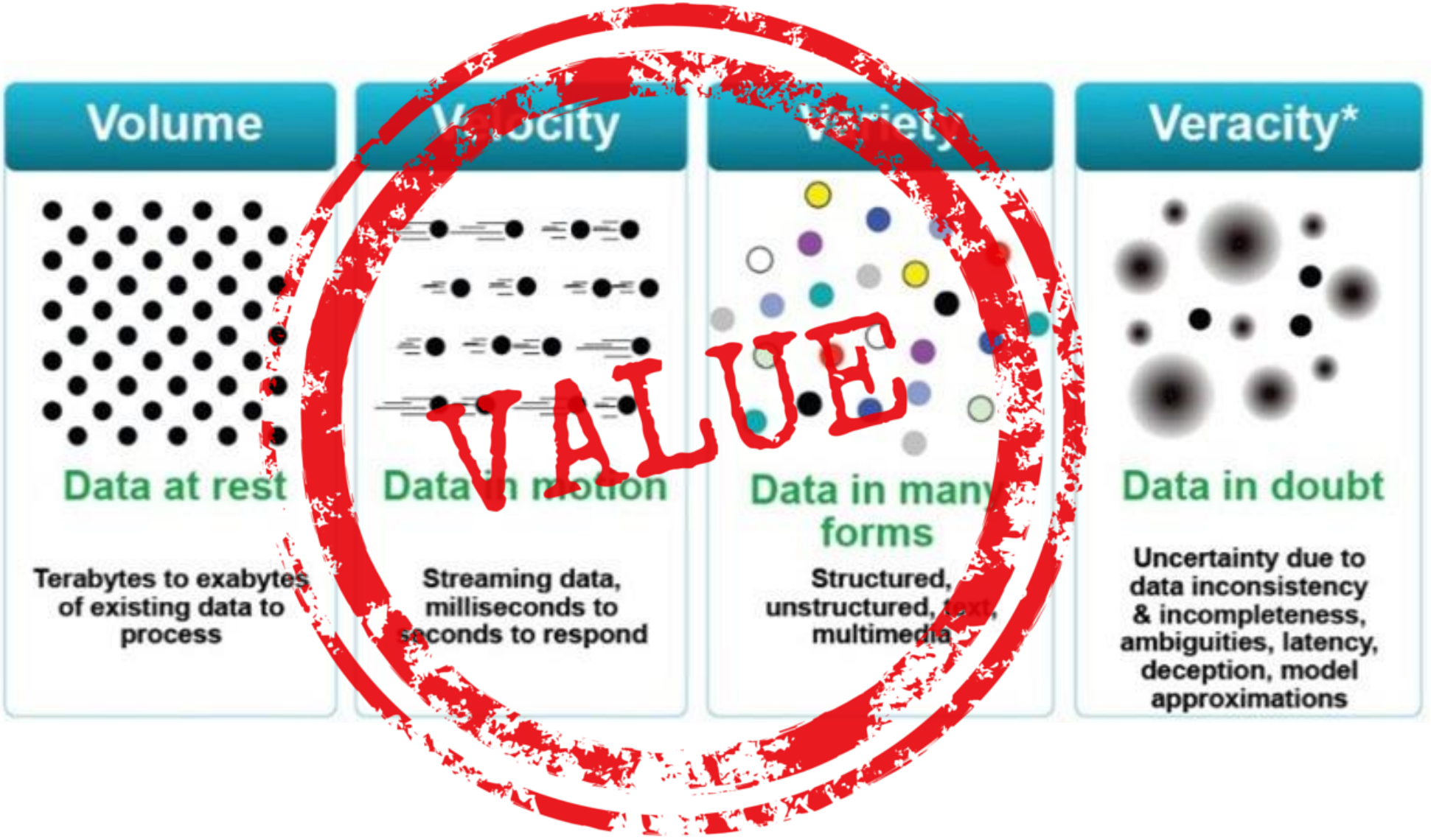
- Descriptive Analysis
- Regression

Conclusion



There is little information on weight of different breeds of cats.

What is Big Data?



https://image.slidesharecdn.com/hadoop-141023223426-conversion-gate01/95/hadoop-20141024-12-638.jpg?cb=1414121124

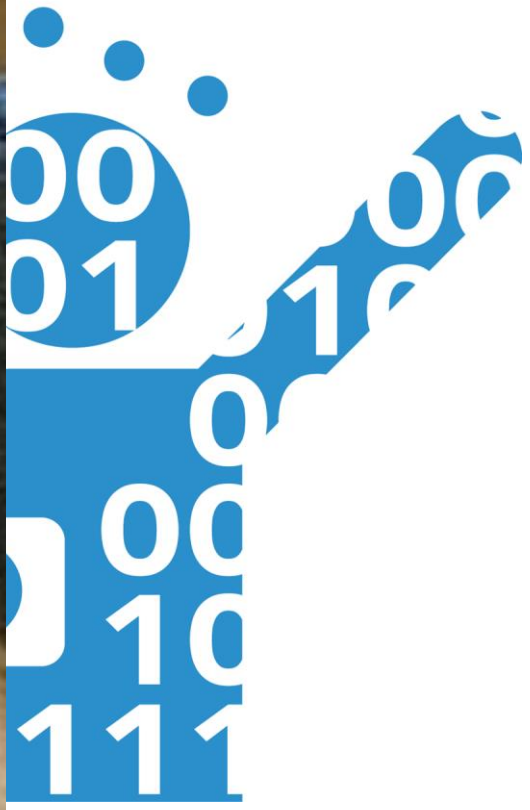
Source of Big Data

- Data from 19,416,753 'feline' records from Intravet, Impromed, Avimark, Cornerstone from 1981-2016
- Data included: Animal ID, Clinic ID, birthdate, gender, breed, weight, units of weight, timestamp of weight
- Provided by IDEXX Laboratories
- Provided as JSON line-delimited format
- Cleaned and analyzed using Python 3.5.2

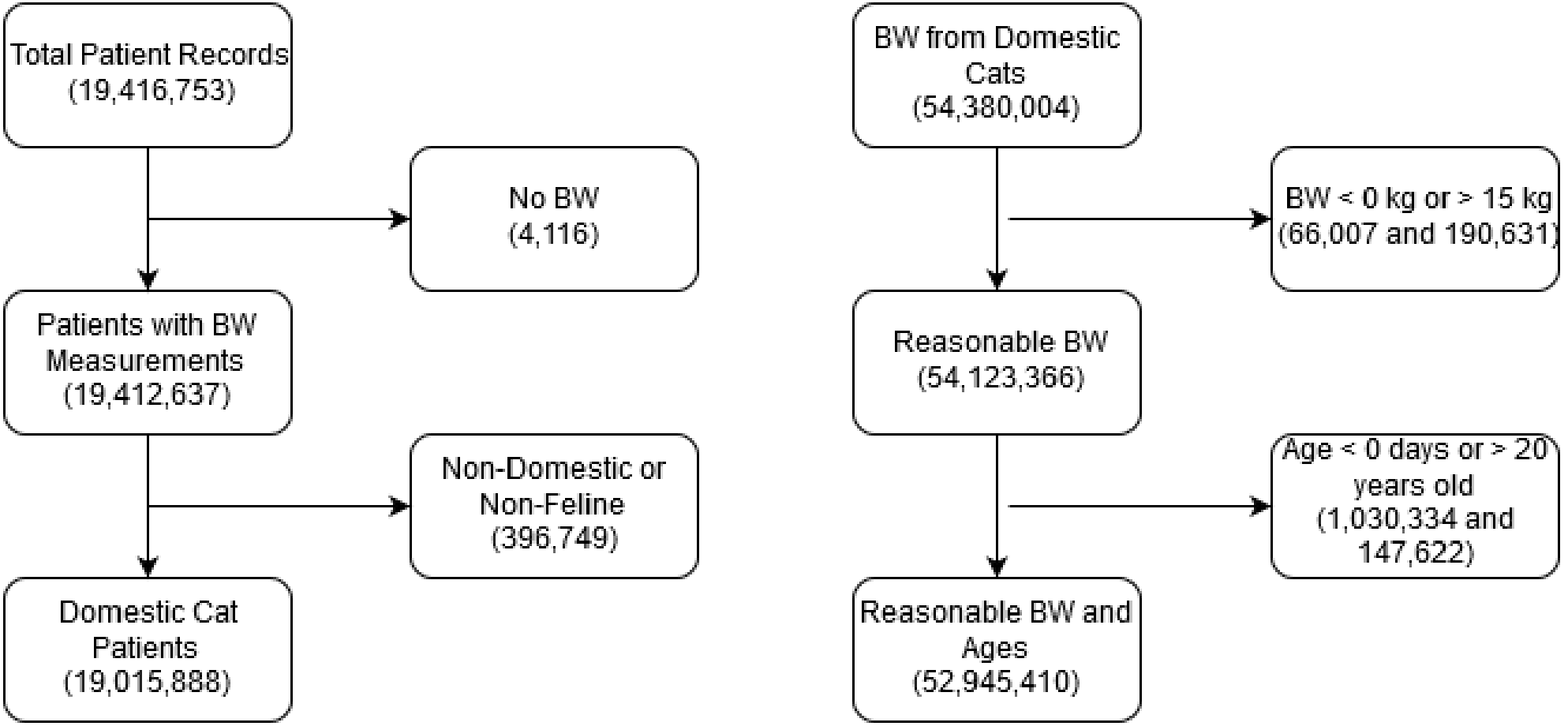
python™

Data Cleaning

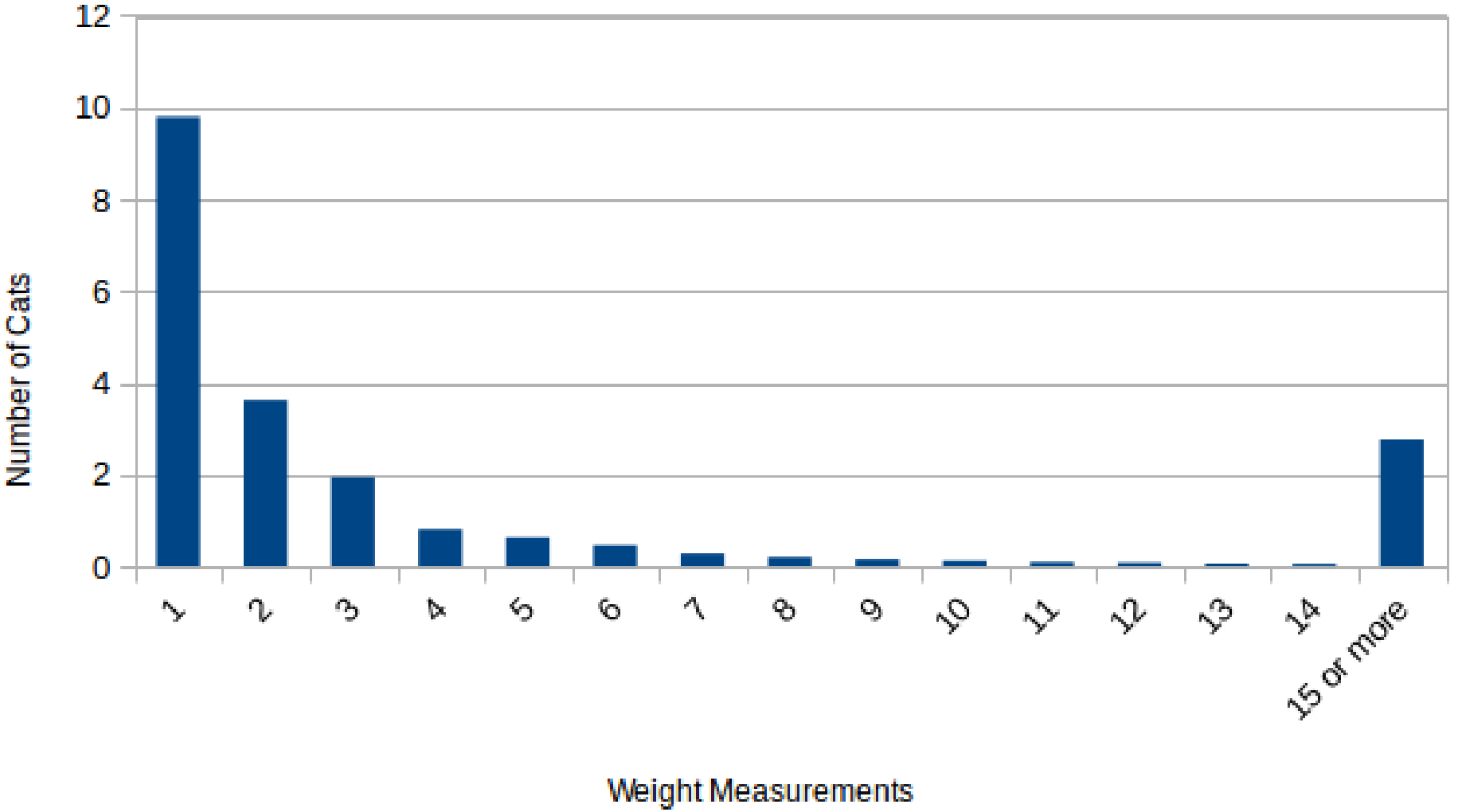
- Continuous
  and weight
- 4 discrete va
- 4 discrete va
  units
- 5057 discret
  breed
  - Fuzzy M
    Manua



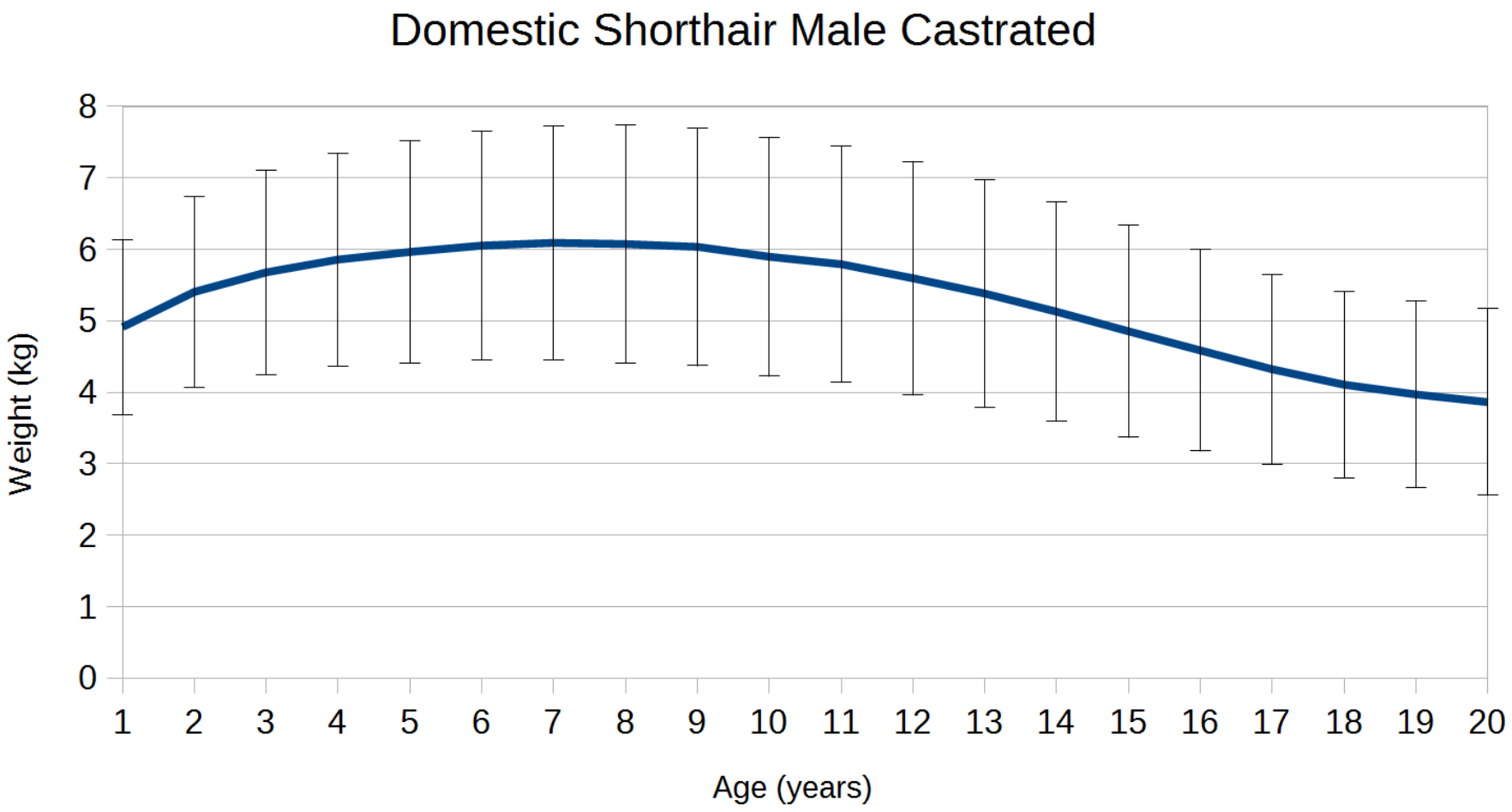
Exclusion Criteria



Descriptive Results



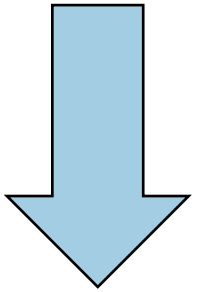
Descriptive Results



OLS Regression

Problem:

Performing Ordinary Least Squares (OLS) regression on the full data set couldn't run



Solution 1:

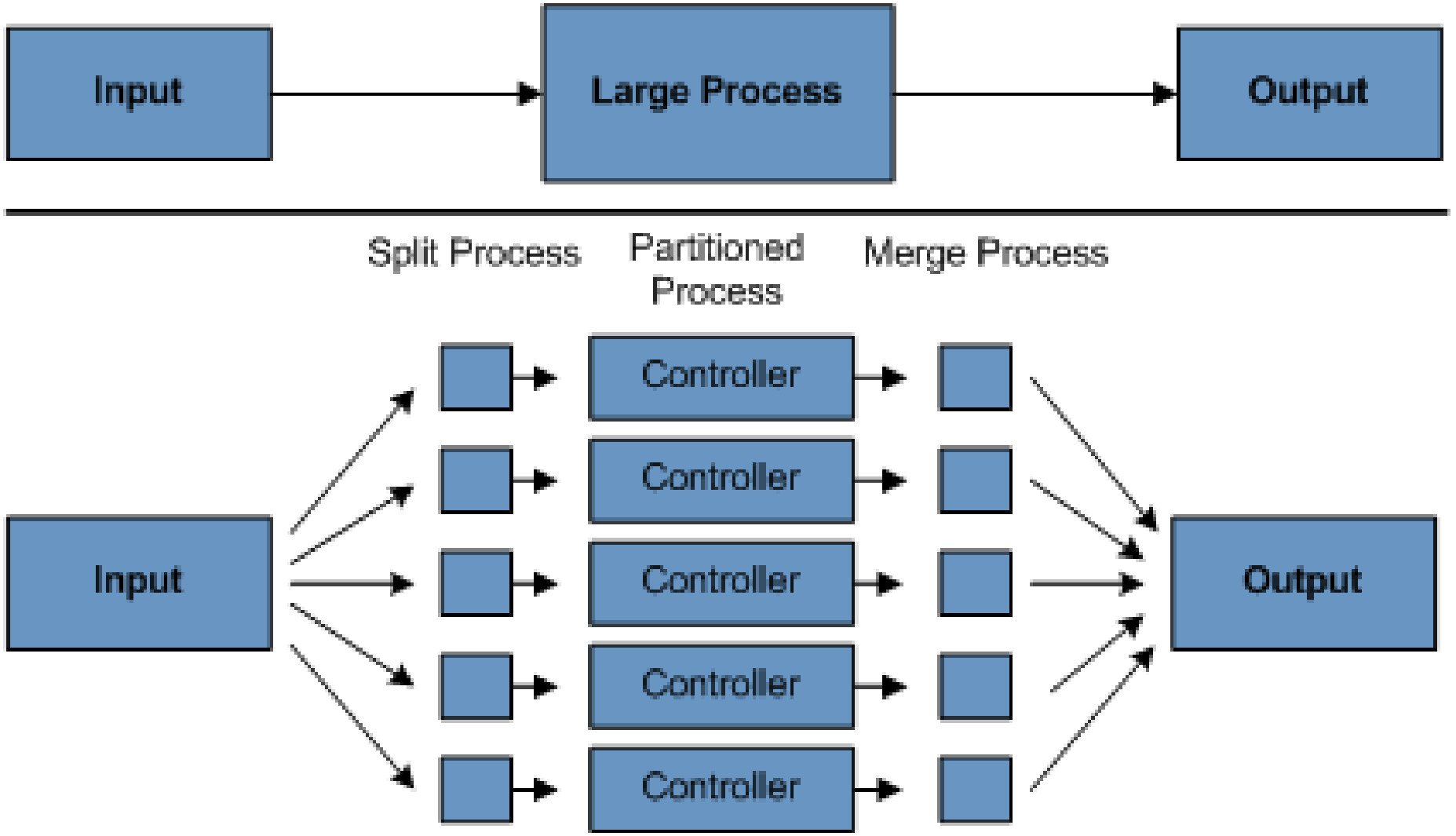
Took random sub-sample of 5,000,000 data points

Found variations in coefficients depending on the subsample chosen

Regression Results

Variable	p-Value
Intercept	<0.001
Age	<0.001
Age^2	<0.001
Breed (categorical)	<0.001
Gender (categorical)	<0.001

Working with Big Data



http://technologist-work.com/wp-content/uploads/2014/09/SQL-Data1.png

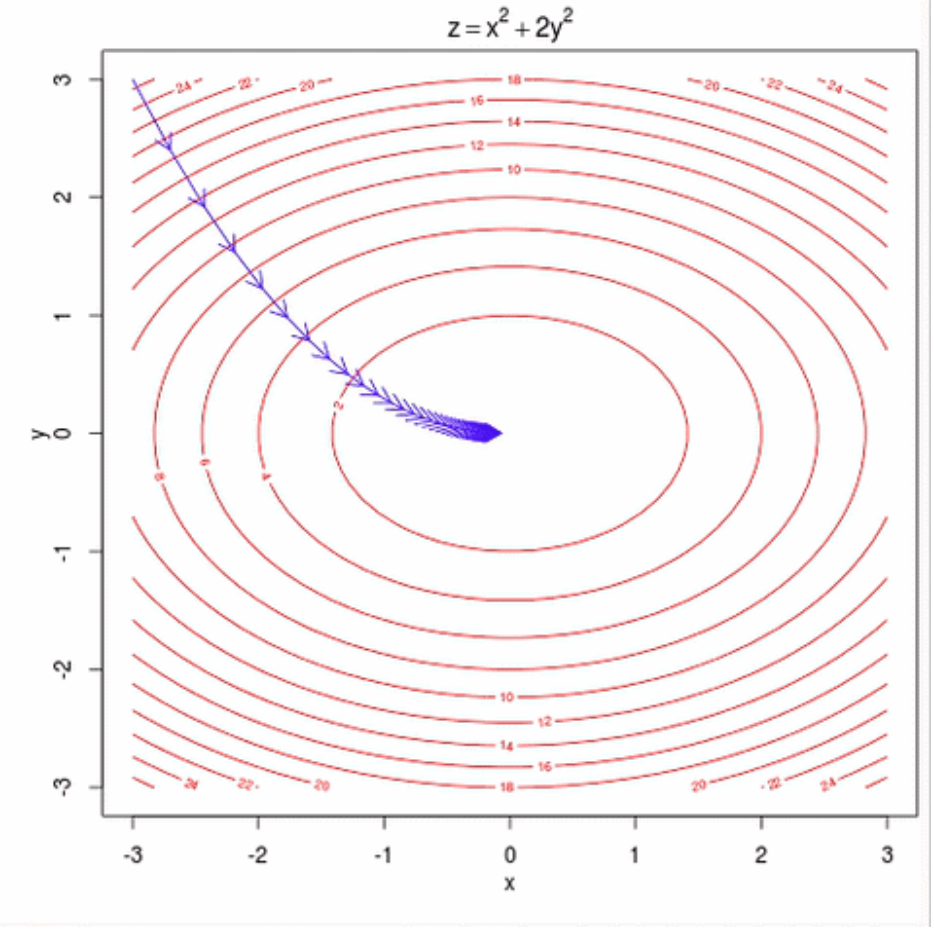
SGD Linear Regression

Problem:

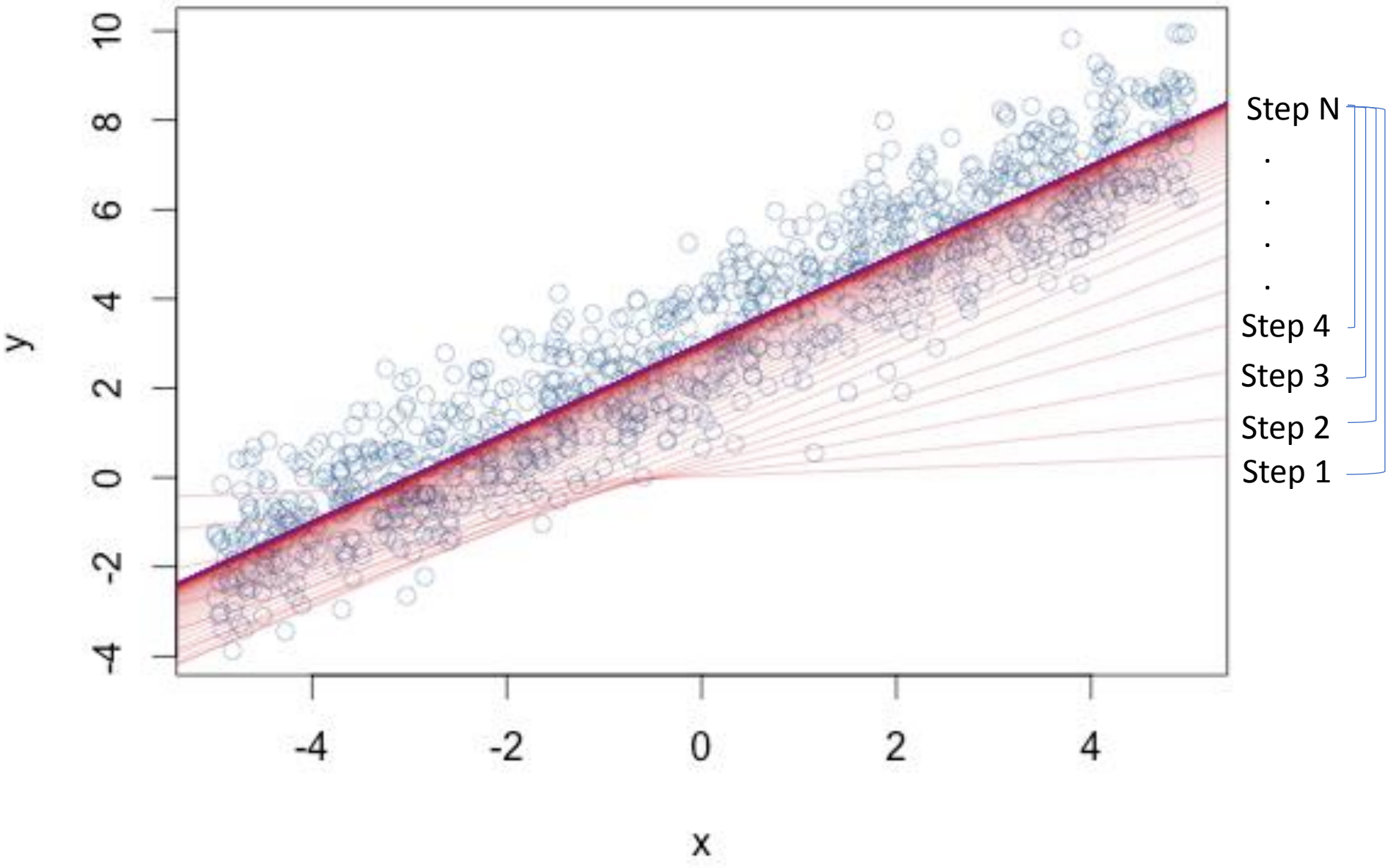
Performing Ordinary Least Squares (OLS) regression on the full data set couldn't run

Solution 2:

Linear regression equation optimized by stochastic gradient descent (SGD)



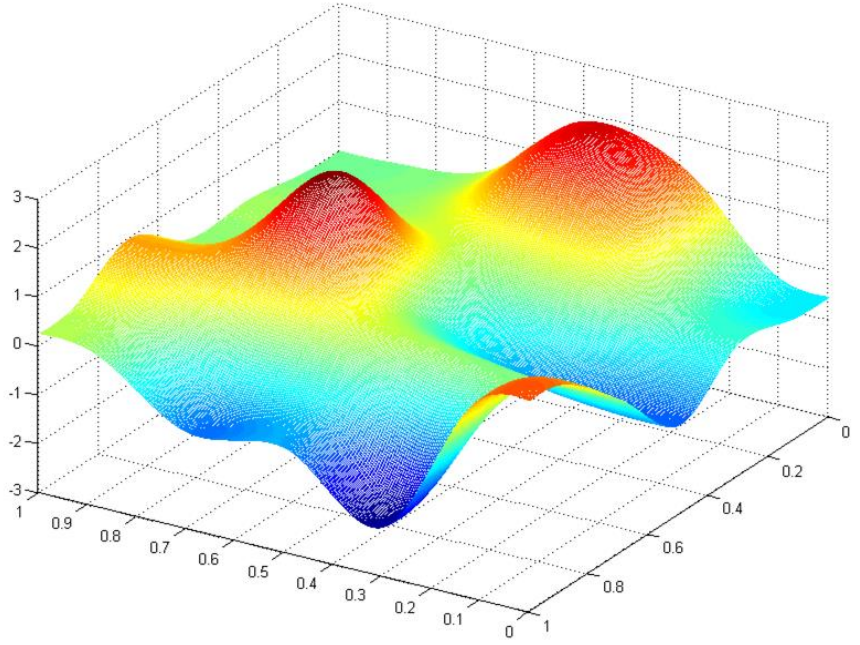
Linear Regression by Gradient Descent



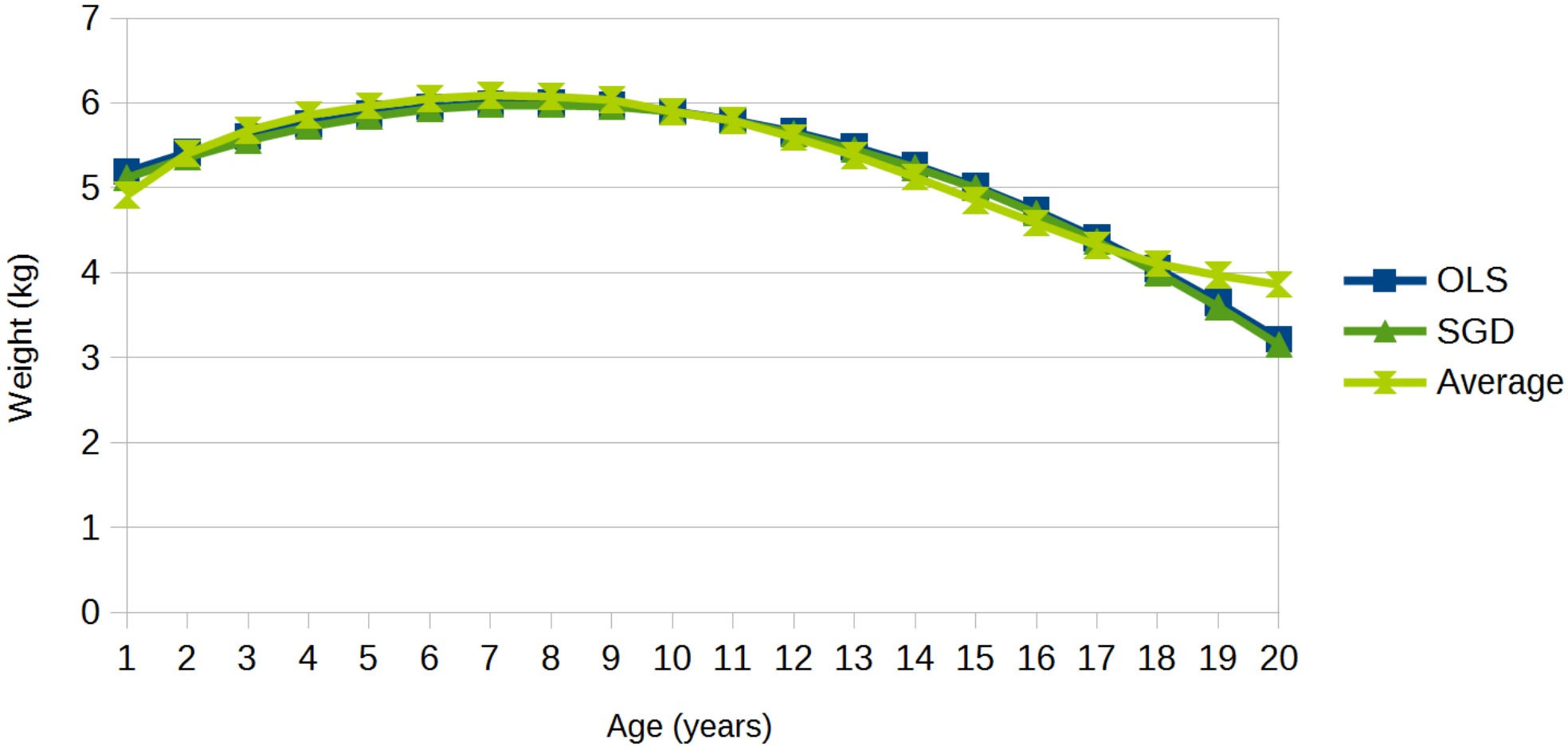
http://i1.wp.com/blog.hackerearth.com/wp-content/uploads/2016/11/GD_R_plot-1.png

Potential Concerns

- No guarantee that will find the absolute minimum error!
- Results can depend on learning rate, iterations, sequence...

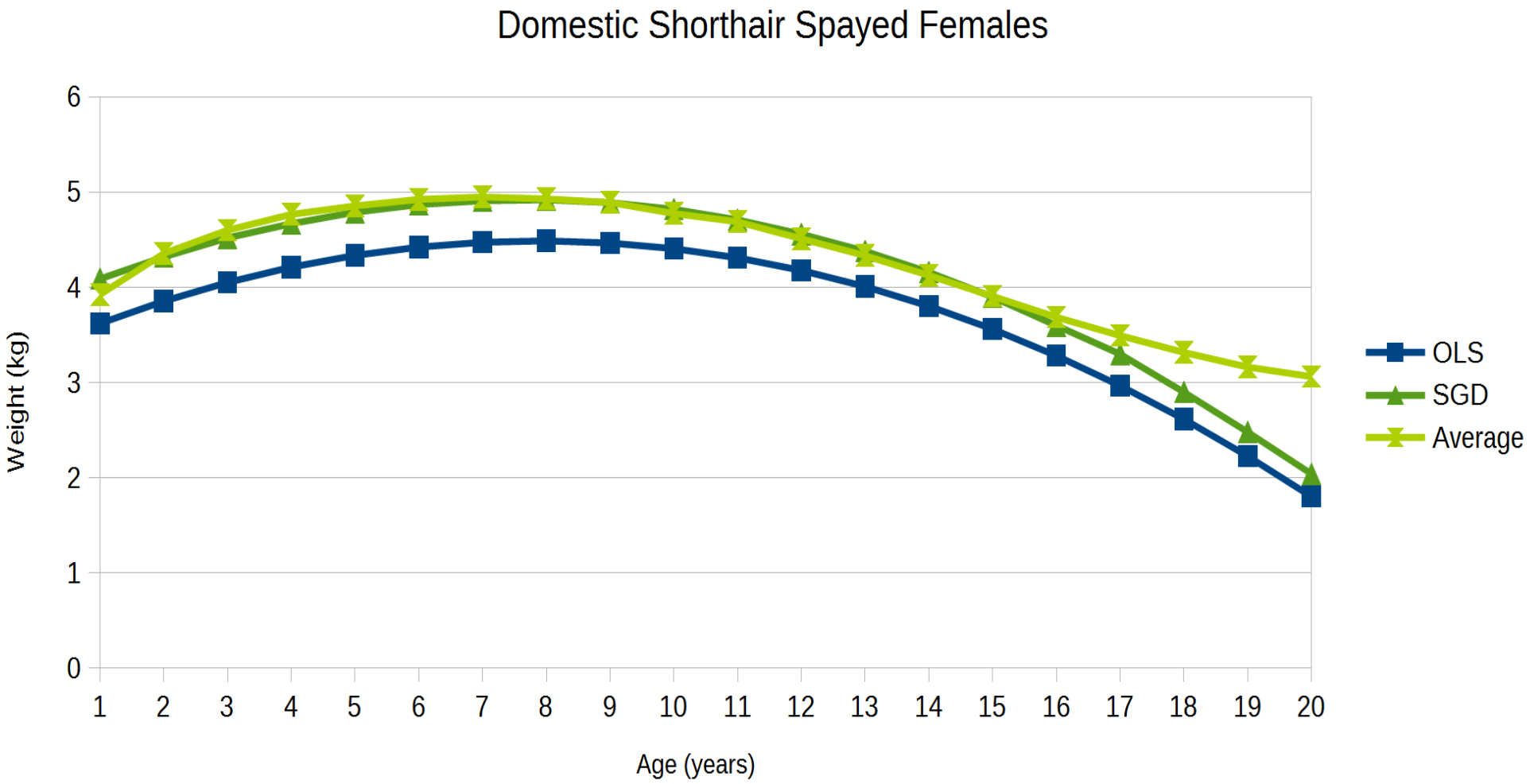


Domestic Shorthair Castrated Male



Adjusted $R^2 = 0.21$
RMSE = 1.4

Domestic Shorthair Spayed Female

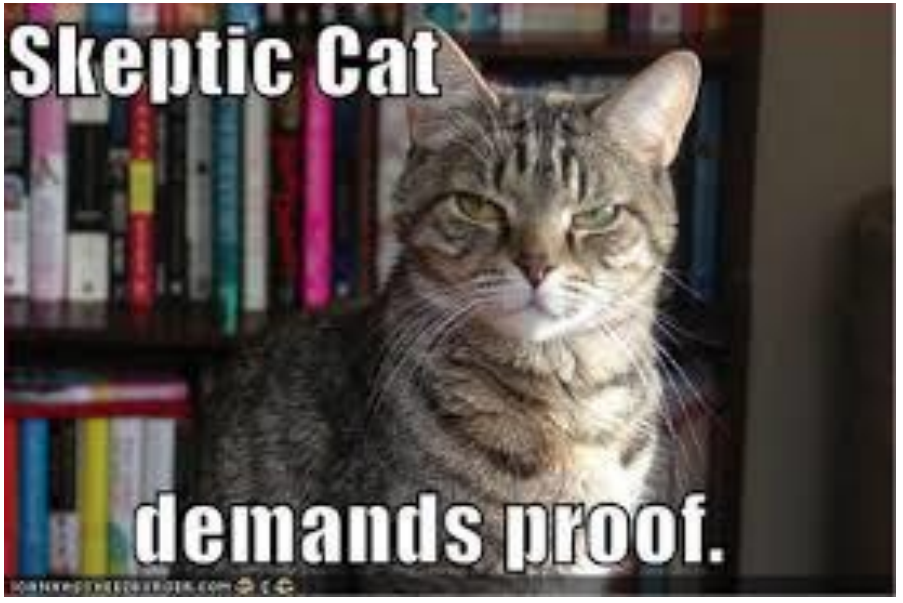


Trends

- Increase weight until 6-10 years.
  - Larger breeds tend to reach maximum weight earlier in life.
- Decline shortly after reaching maximum weight.
- Intact animals reached a lower maximum weight compared to neutered animals.

Advantages of SGD

- SGD regression provides similar predictive ability as OLS
- Ability to update as more data enters
- Ability to work with large data sets

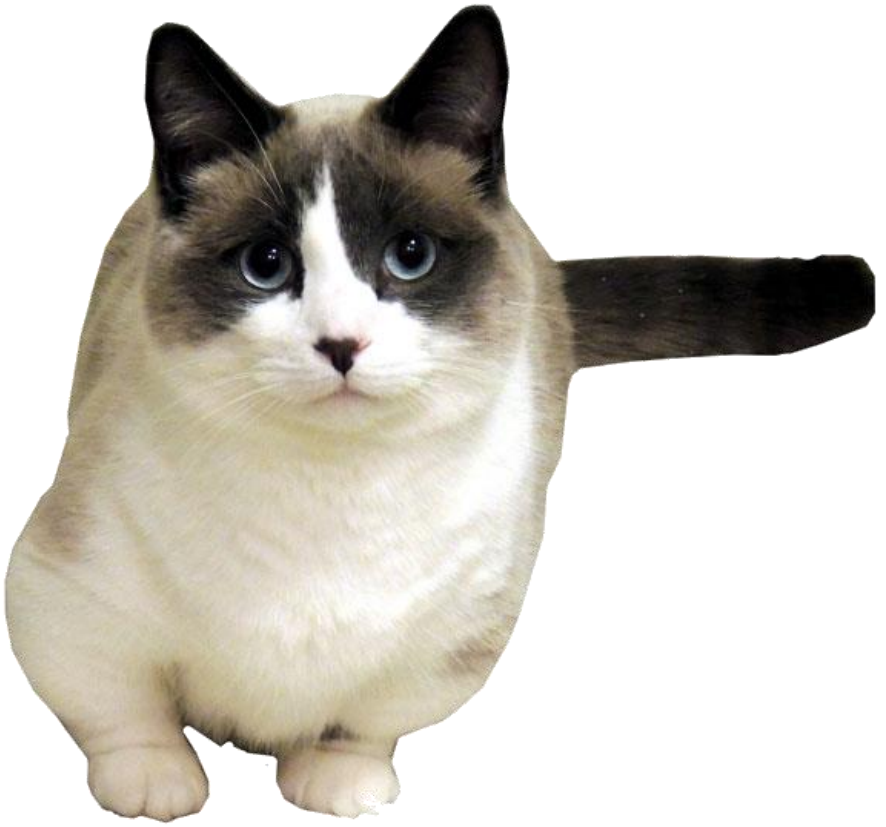


Disadvantages of SGD

- No confidence intervals
- No p-values
- Coefficients may not correspond to OLS equivalent

Future Considerations

- Use same techniques for health parameters
- Future research can include development of methods and use of predictive methods to health data and longevity



Acknowledgements

Advisors

- Dr. Theresa Bernardo
- Dr. Zvonimir Poljak
- Dr. Deborah Stacey
- Dr. Elizabeth Stone