# Visualization for Data Science

Leland Wilkinson
Chief Scientist
H2O
Adjunct Professor of Computer Science, UIC
https://www.cs.uic.edu/~wilkinson/

NISS National Institute of Statistical Sciences

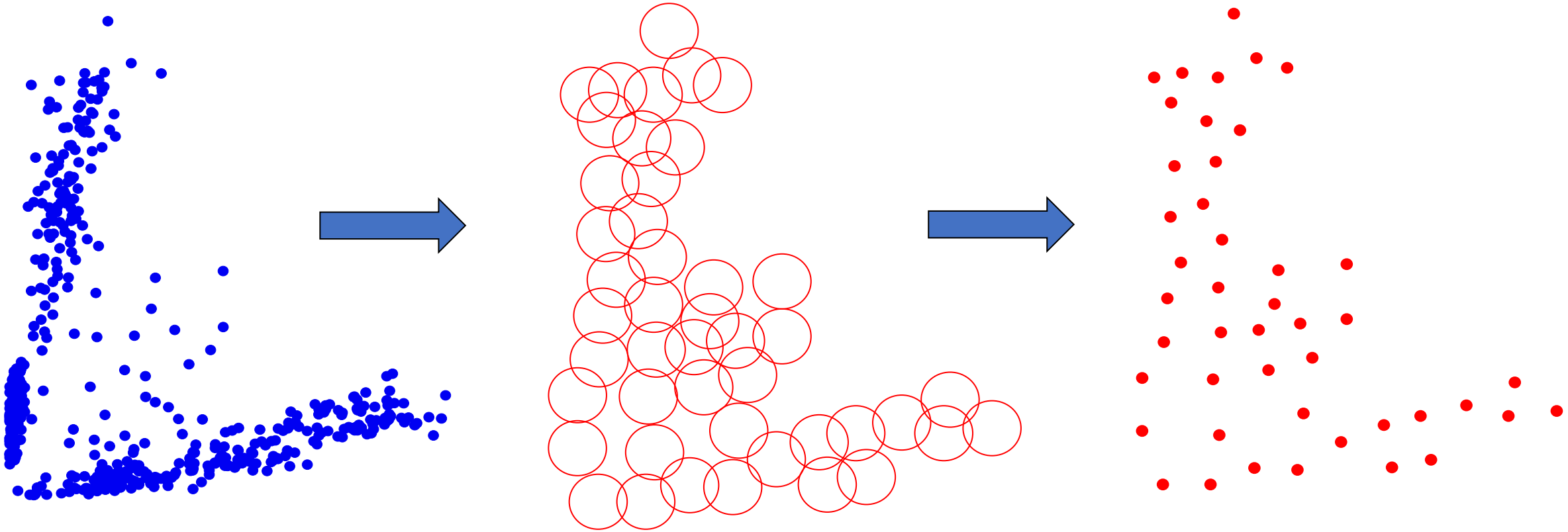
# This talk is about visualization needed for AI

- It concerns how visualization can help identify anomalies.
  - Outliers
  - Distributional anomalies
  - Logical anomalies
  - Other diagnostics

# Visualizing Big Data

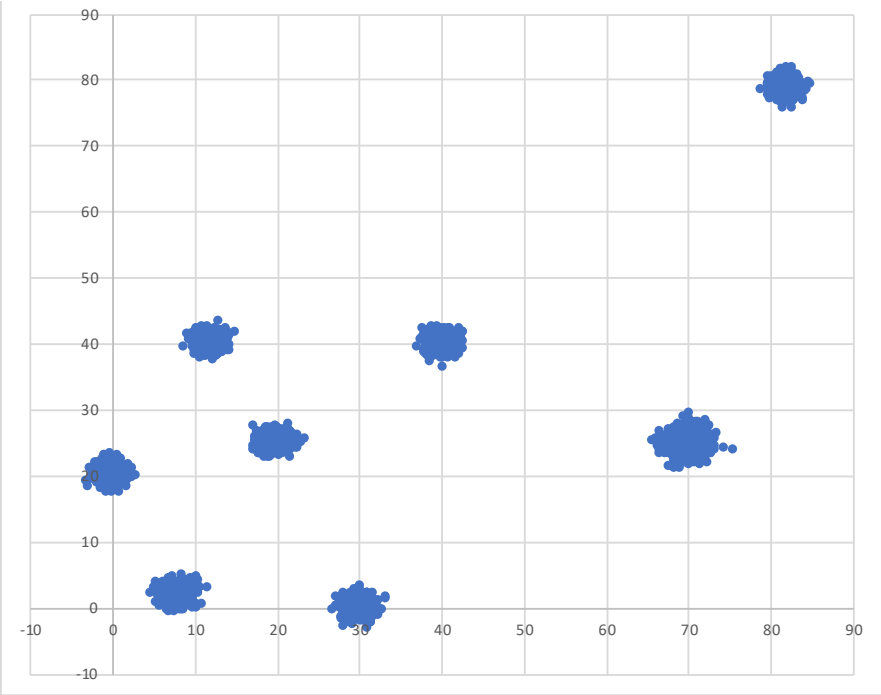
- Complexity: Many visualization functions are polynomial or exponential
- Curse of Dimensionality: distances tend toward constant as $d \to \infty$
- Chokepoint: Cannot send big data over the wire
- Real Estate: Cannot plot big data on the client

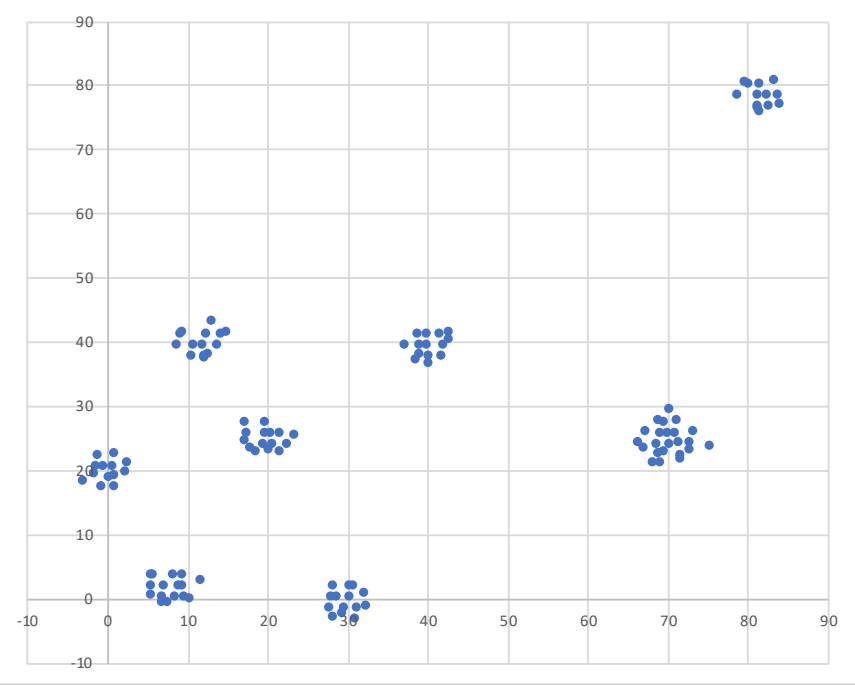
# Big Data



set cover (core sets)

# Big Data


90
80
70
60
50
40
30
20
10
0
-10
-10
0
10
20
30
40
50
60
70
80
90


90
80
70
60
50
40
30
20
10
0
-10
-10
0
10
20
30
40
50
60
70
80
90

# High-Dimensional Aggregation

1. Map categorical variables to continuous values (SVD).
2. If $p$ large, use random projections to reduce dimensionality.
3. Normalize columns on [0, 1]
4. Choose radius of balls to make aggregated file n ~ 1,000
5. Aggregate in one pass through file.
6. We call data points on which balls are centered *exemplars*
7. We call data points falling inside balls *members*

# Outliers

1. Aggregate
2. Compute nearest-neighbor Euclidean distances between exemplars.
3. Fit exponential distribution to largest distances.
4. Reject points in upper tail of this distribution.

Visualizing Big Data Outliers Through Distributed Aggregation. IEEE Trans. Vis. Comput. Graph. 24(1): 256-266 (2018)
https://cran.r-project.org/web/packages/HDoutliers/index.html

# Anomalies

- An anomaly is an observation inconsistent with a set of beliefs.
  - An outlier anomaly is an observation inconsistent with a set of points.
    - The points are presumed generated by a probabilistic process in a vector space.
    - All outliers are anomalies but not all anomalies are outliers
  - A distributional anomaly is a distribution of data that doesn't fit conventional expectations.
    - Multimodality
    - Skewness, …
  - A logical anomaly is an observation inconsistent with the the axioms of a system
    - A report that someone has been in a relationship with another longer than he/she has been alive
    - A finding that a certain car gets negative miles per gallon
  - A model anomaly is an observation inconsistent with a model
    - Observations less than zero fed to a Poisson model
    - Survival times less than zero fed to a Cox proportional-hazards model
    - Residuals that violate assumptions for fitting a model

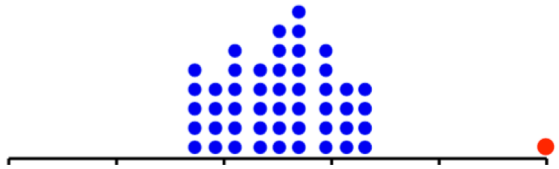
# Outliers

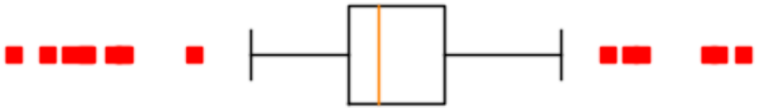
Outlier detection has more than a 200 year history.

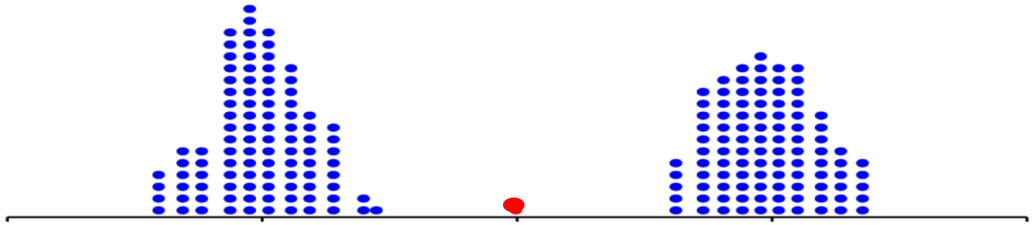
- The goal was to reduce bias in models
- The goal today is to learn interesting stuff from examining outliers
- Statisticians no longer delete outliers. They use robust methods.
- The best references on outliers were written by statisticians.
  - Barnett & Lewis (1994), *Outliers in Statistical Data*.
  - Rousseeuw & Leroy (1987). *Robust Regression & Outlier Detection*.
  - Hartigan (1975) *Clustering Algorithms.*

# Outliers

- Univariate outliers
  - Distance from Center Rule
  - Gaps Rule

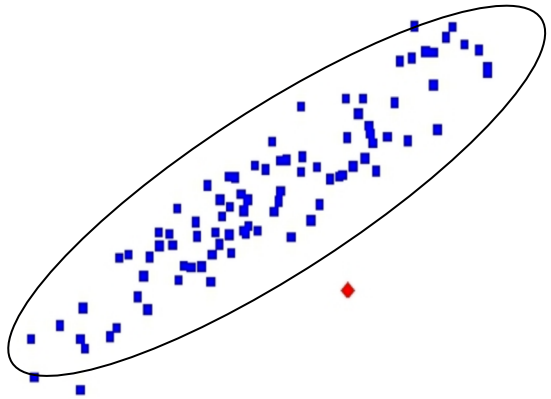




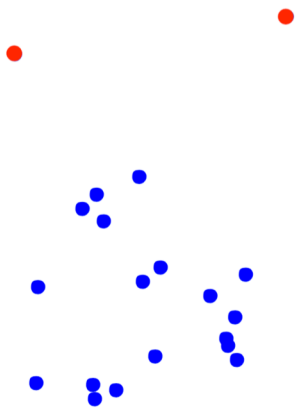


# Outliers

- Multivariate outliers
  - Distance from Center Rule

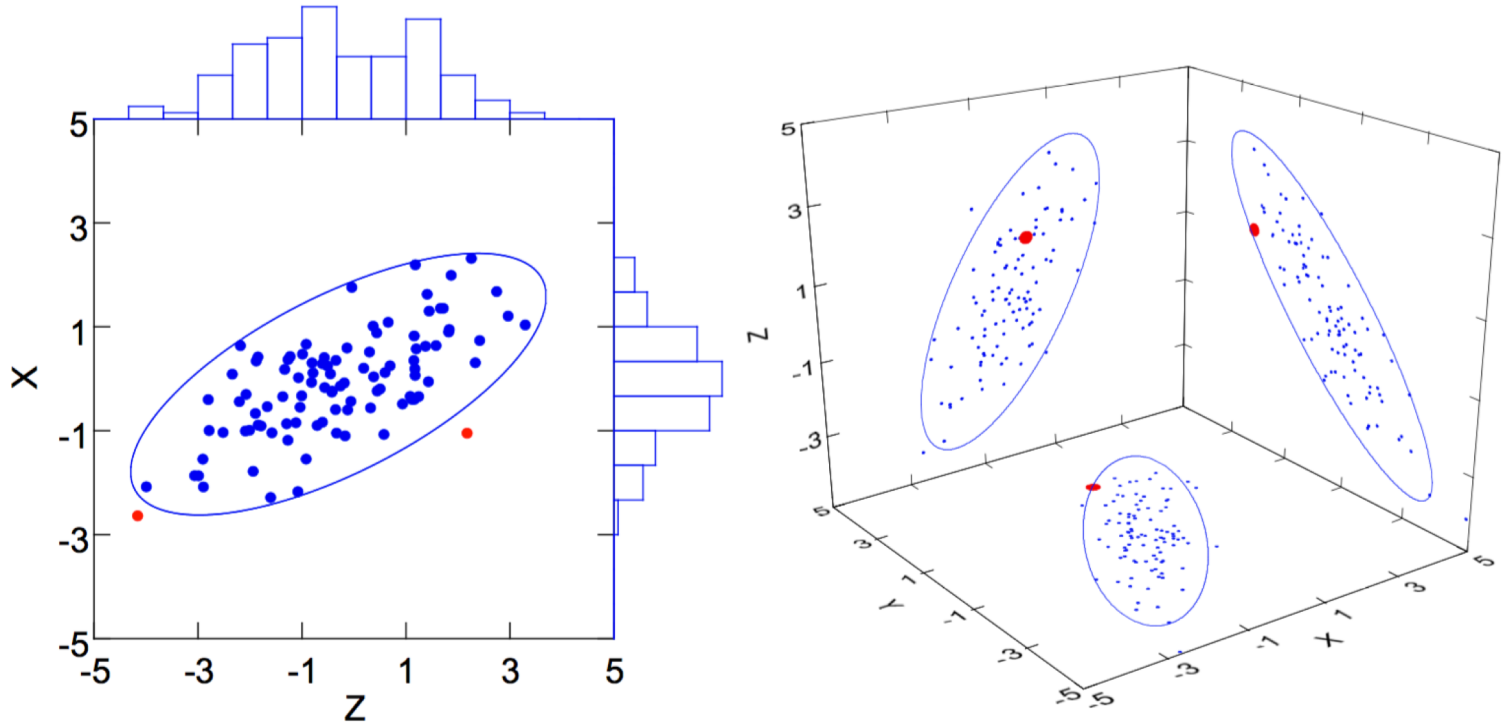


  - Gaps Rule



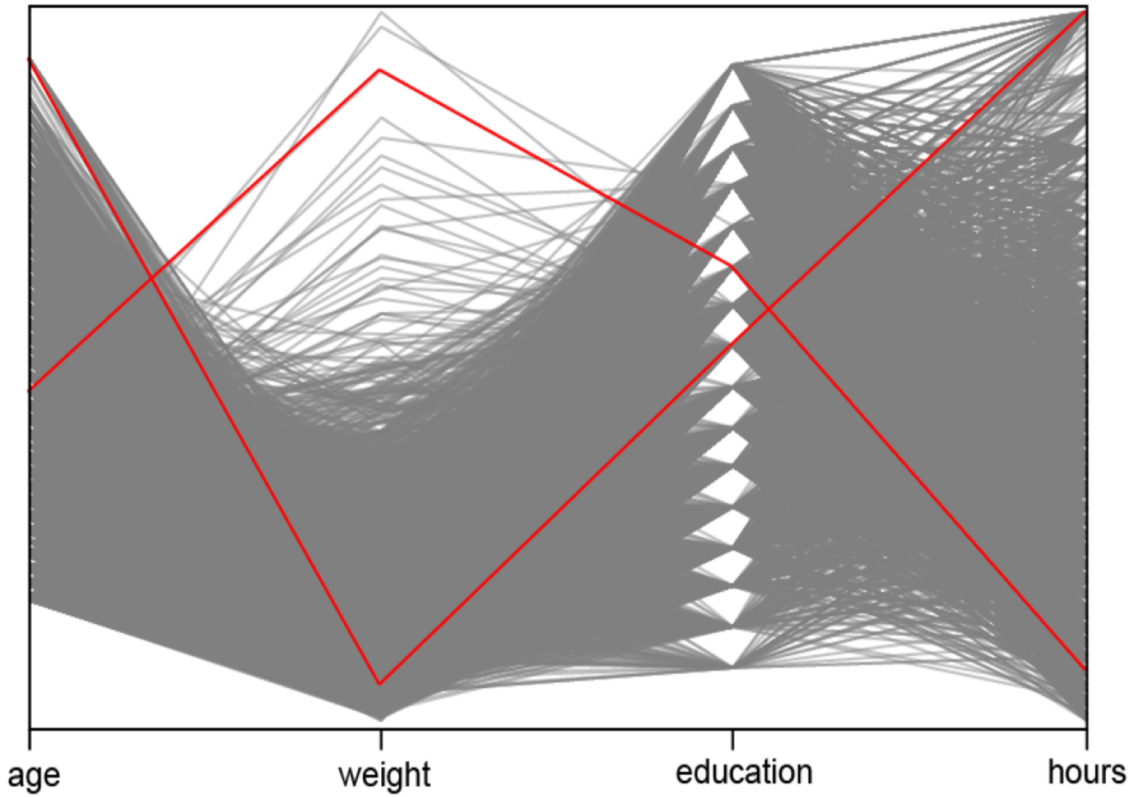
# Outliers

- Low-dimensional projections are not reliable ways to discover high-dimensional outliers.



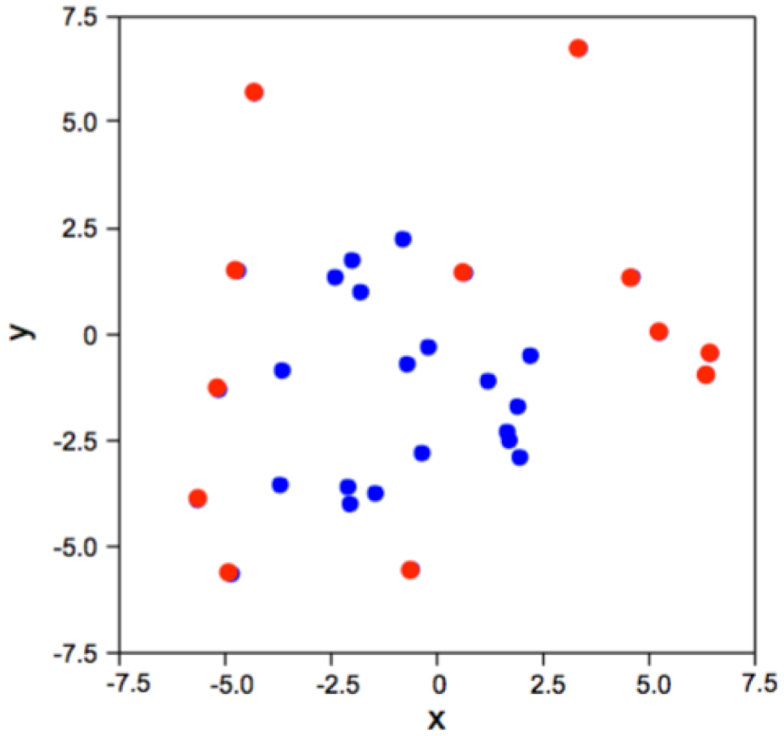
# Outliers

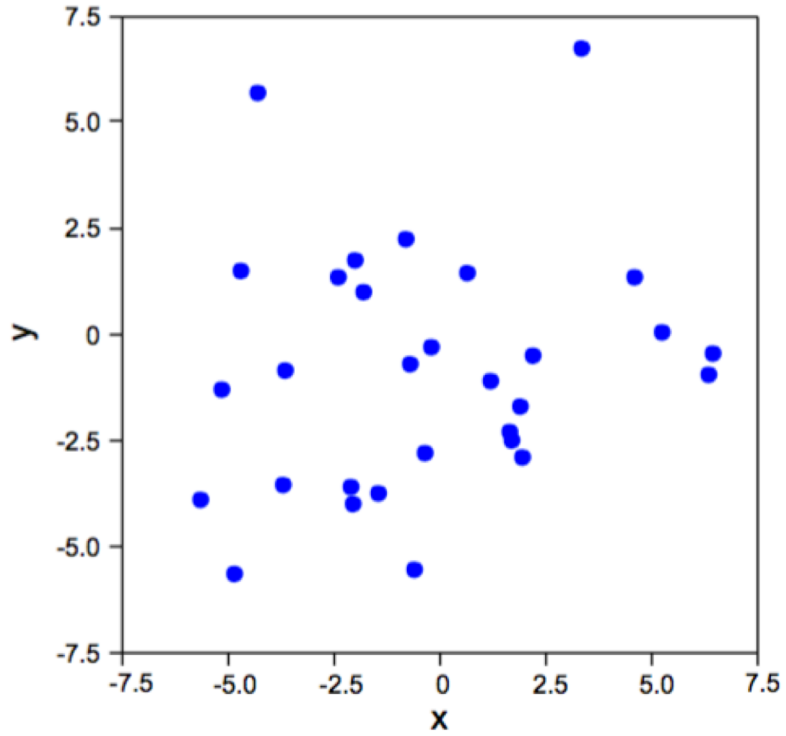
- Parallel coordinates and other multivariate visualizations are not reliable ways to discover high-dimensional outliers.



# Outliers

- Popular ML algorithms are not reliable ways to identify outliers.



# Outliers Summary

- Regression methods (OLS, GLM, …) are susceptible to outliers.
  - Workarounds:
    - Delete outliers – a bad idea unless you KNOW they are mistakes.
    - Use robust regression (best of all, use median regression).
- Tree-based methods are susceptible to outliers.
  - People who claim tree methods are robust think only of univariate predictor outliers.
  - Regression trees on data having dependent variable univariate outliers are susceptible.
  - Trees on data with univariate predictor outliers are usually not susceptible.
  - Trees on data with multivariate outliers are usually susceptible.
  - Random forests and GBM trees are not exceptions to these rules.
  - Workarounds:
    - Delete outlier cases –again, usually a bad idea.
    - Use robust loss functions.

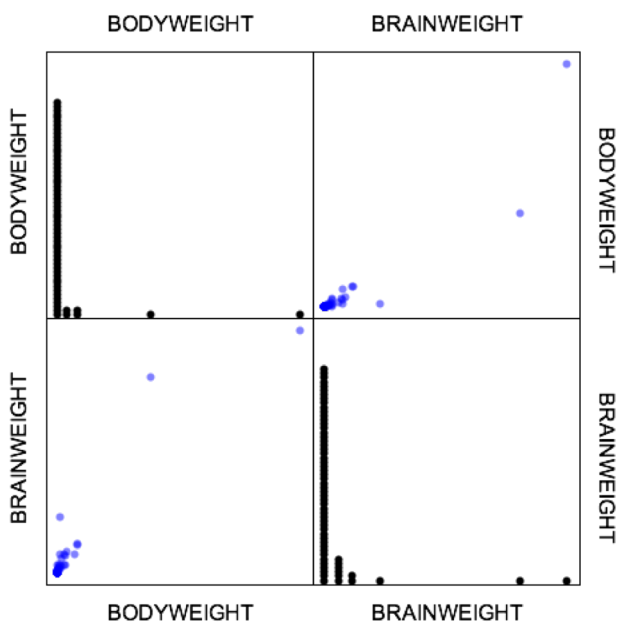
# Distributional Anomalies

Anomalous distributions demand special attention.

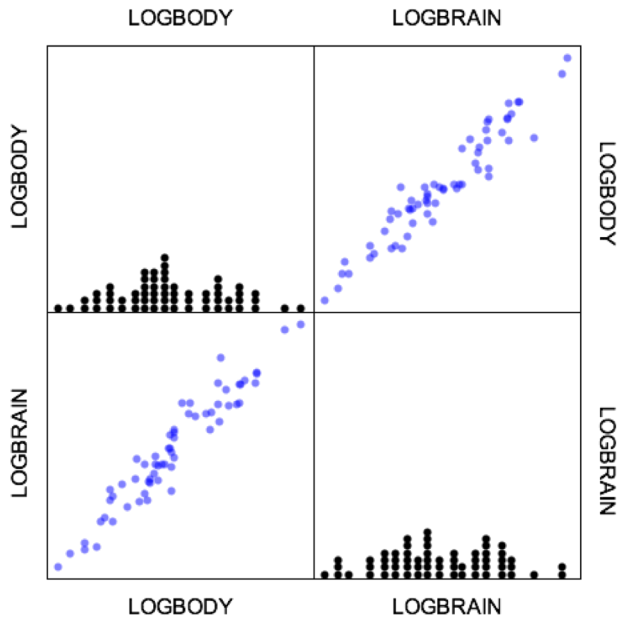
- Most classical statistical prediction methods assume normal distributions
- Other statistical prediction methods were designed for alternative distributions
  - Logistic regression
  - Poisson regression
  - Negative binomial regression
  - Zero inflated regression …
- Skewed distributions can sometimes be transformed to look more closely Normal.
- Judicious use of transformations can improved even tree models.

# Distributional Anomalies

Raw Data



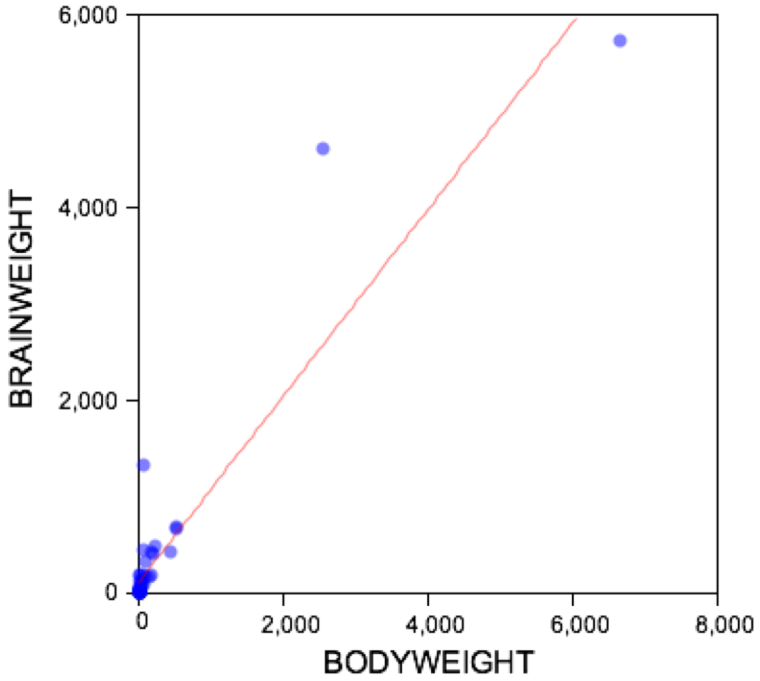

Log Data



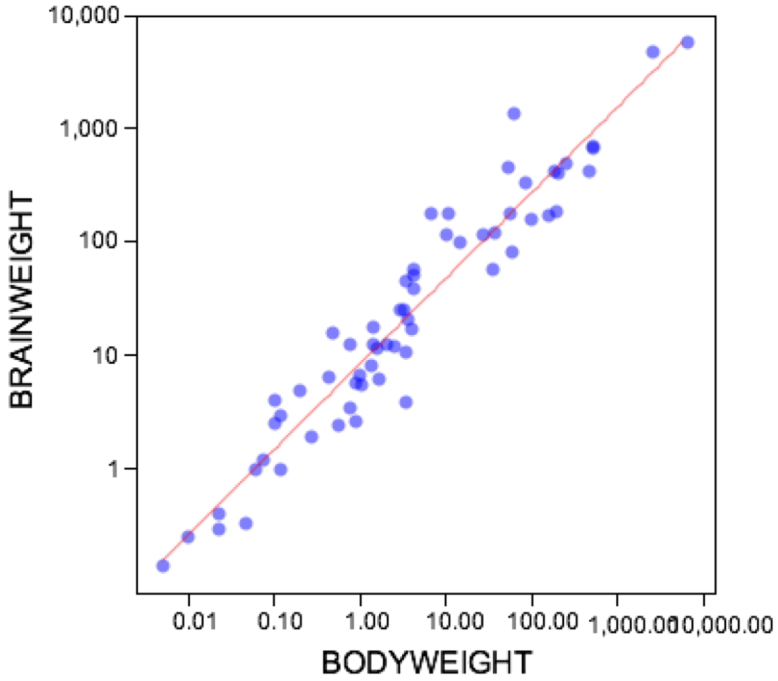

Allison, T. and Cicchetti, D. (1976). Sleep in mammals: Ecological and constitutional correlates. *Science*, *194*, 732–734.

# Distributional Anomalies

Raw Data
r-squared = .873



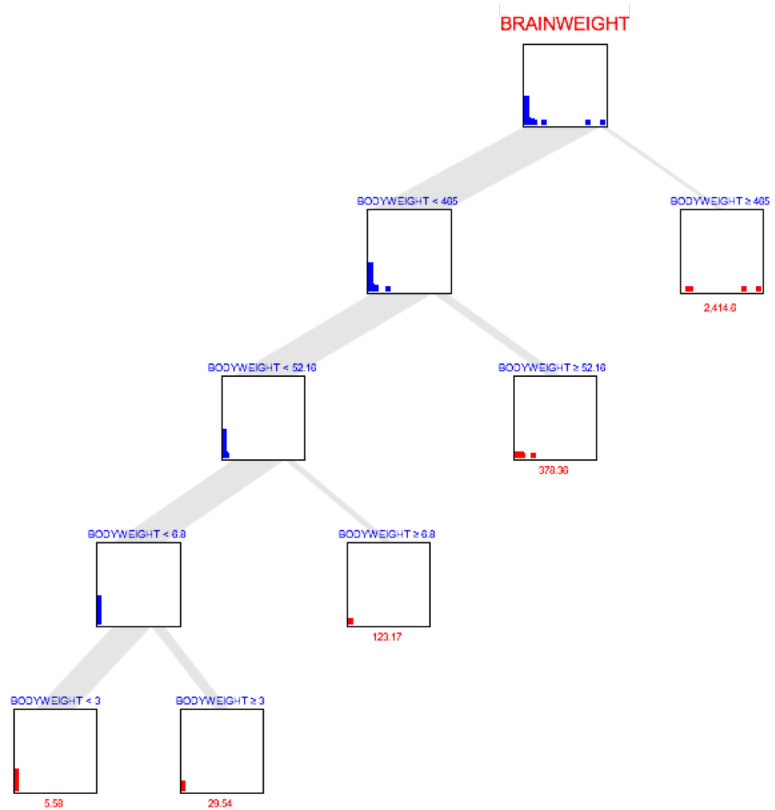

Log Data
r-squared = .921



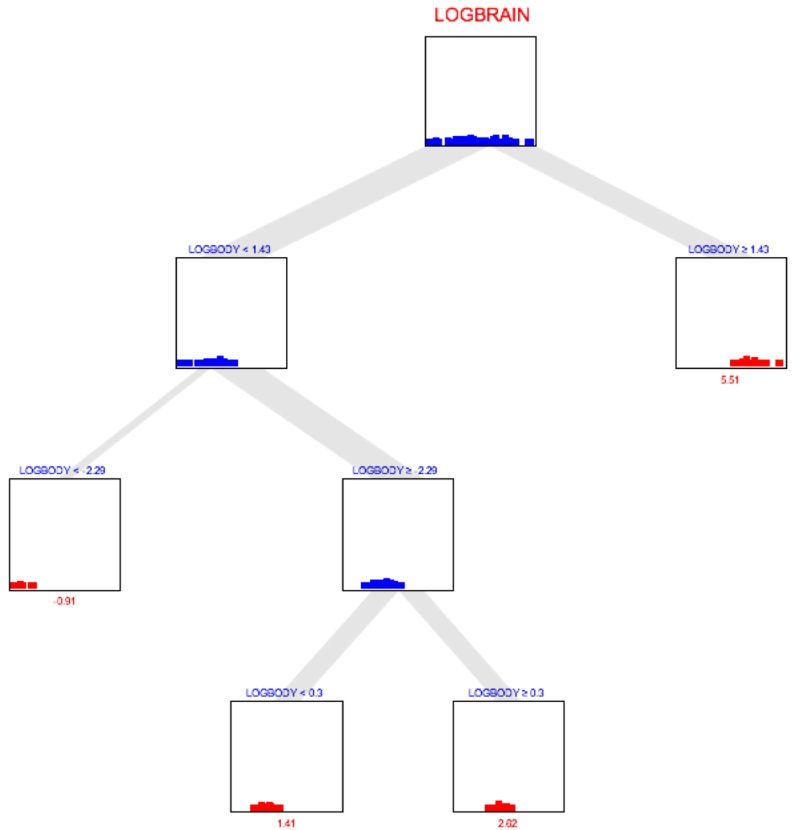

NISS National Institute of Statistical Sciences

# Distributional Anomalies

Raw Data
PRE = .490




Log Data
PRE = .849




NISS National Institute of Statistical Sciences

# Distributional Anomalies Summary

## Tukey Ladder of Powers (re-expressions)

Assume data are positive, or use $X + 1$ if non-negative

Tukey formula

$X \longmapsto X^p$

Box & Cox formula (derived from Tukey's idea)

$X \longmapsto (X^p - 1) / p$

Values of p

$p = 2$ yields $X^2$

$p = 1$ yields $X$

$p = .5$ yields sqrt($X$)

$p = 0$ yields log($X$)

$p = -1$ yields $1 / X$

For Box & Cox formula

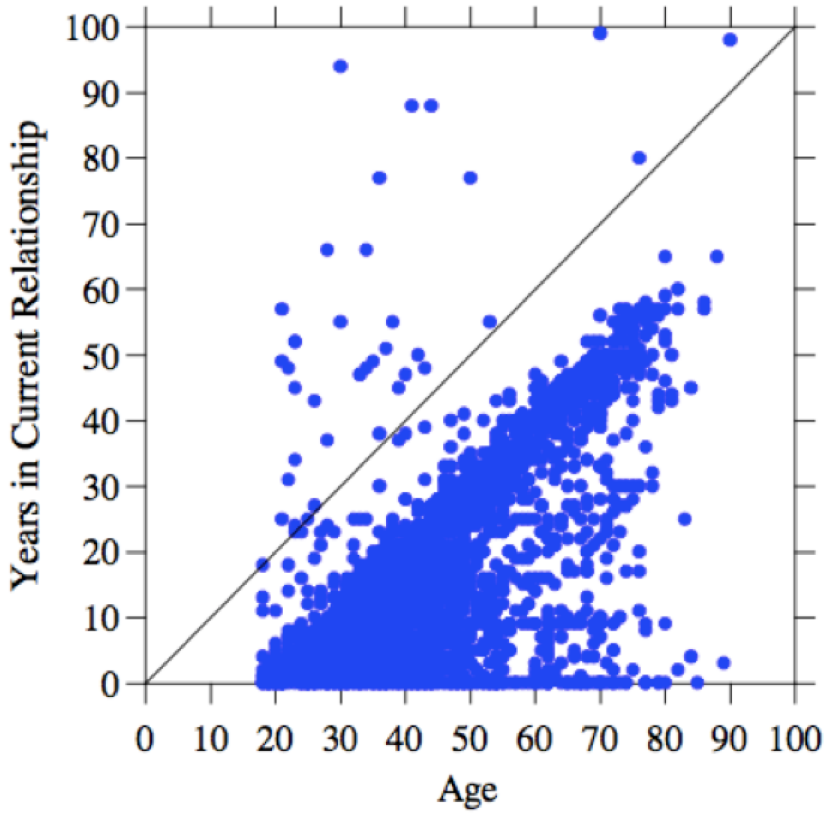
$p = 0$ yields log($X$) because $\lim_{p \to 0} (X^p - 1) / p = \log(X)$

Also, dividing by $p$ in Box & Cox formula preserves polarity of X

Ascending the ladder ($p > 1$) spreads out large values and compresses small values.
Descending the ladder ($p < 1$) compresses large values and spreads out small values.

# Logical Anomalies

Some people report they have been in a relationship longer than they have been alive.



# Logical Anomalies Summary

- How do we deal with logical anomalies?
  - Look for implicational vs distributional scatterplots
  - Apply Berkson's **intraocular traumatic test**
  - Apply logical filters (negative age, negative counts, percents > 100, etc.)

# Model Anomalies

These anomalies become apparent after fitting a model.

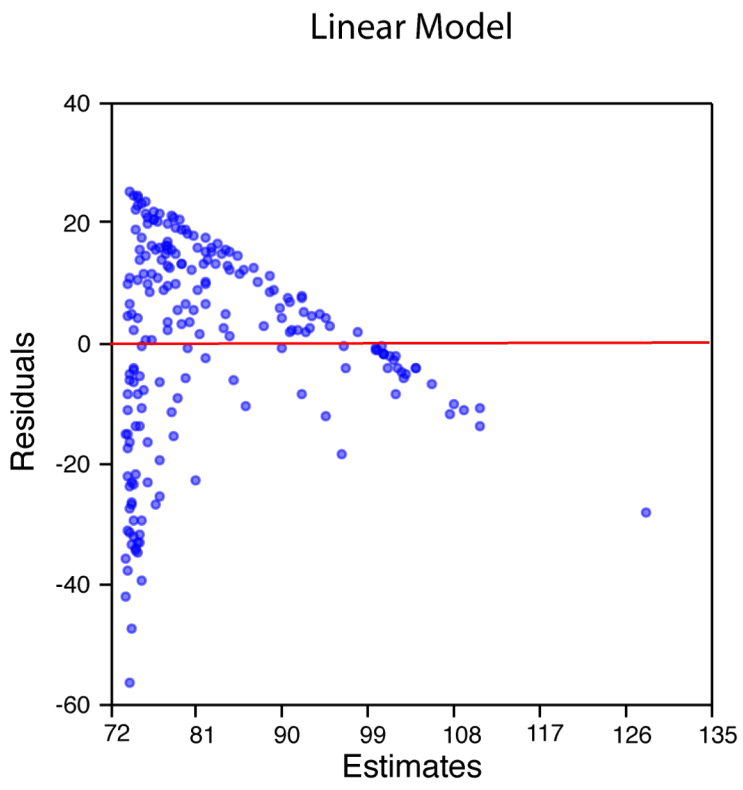
We discern them through diagnostic methods.

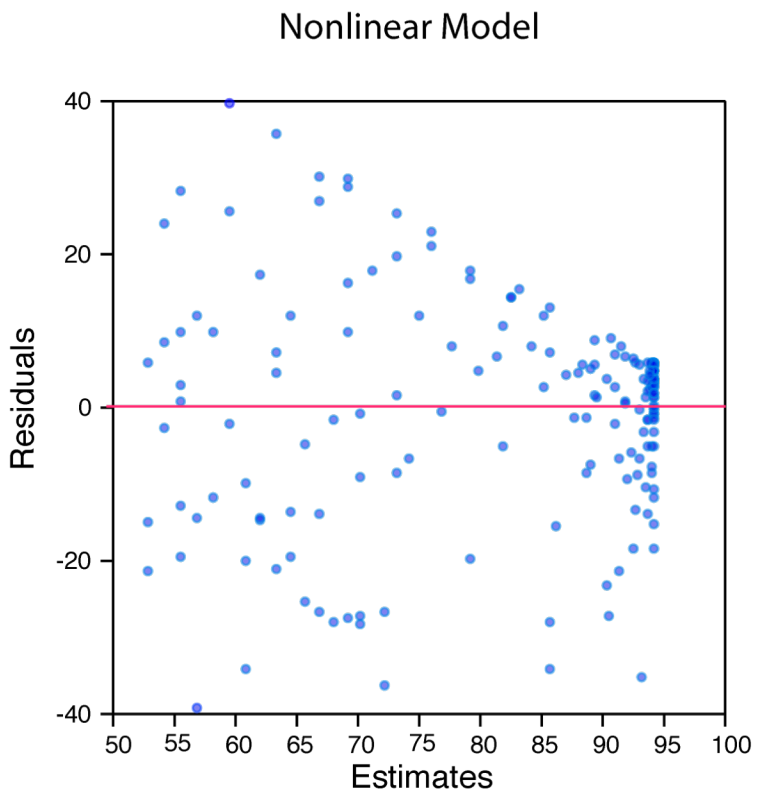
Each type of model has different diagnostics.

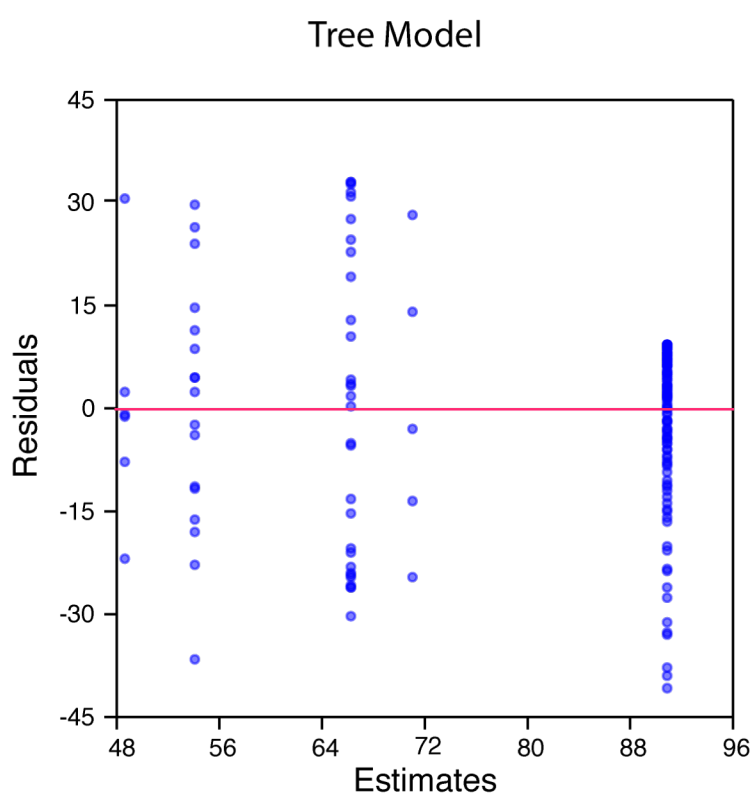
Most prediction models diagnostics involve examining residuals.

# Diagnostics

## Residual plots.



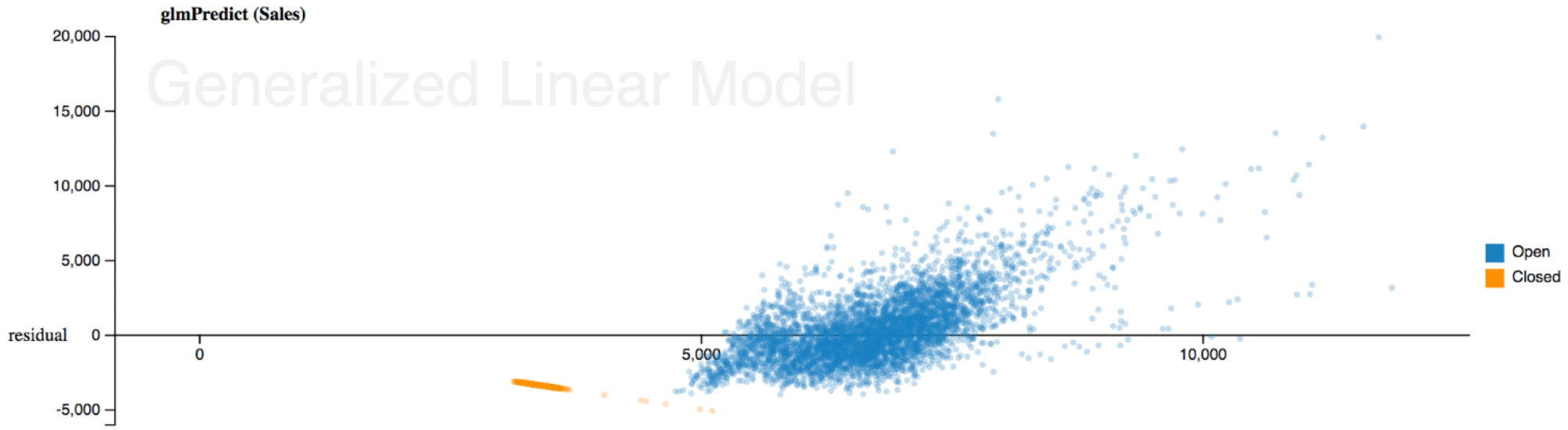



# Diagnostics

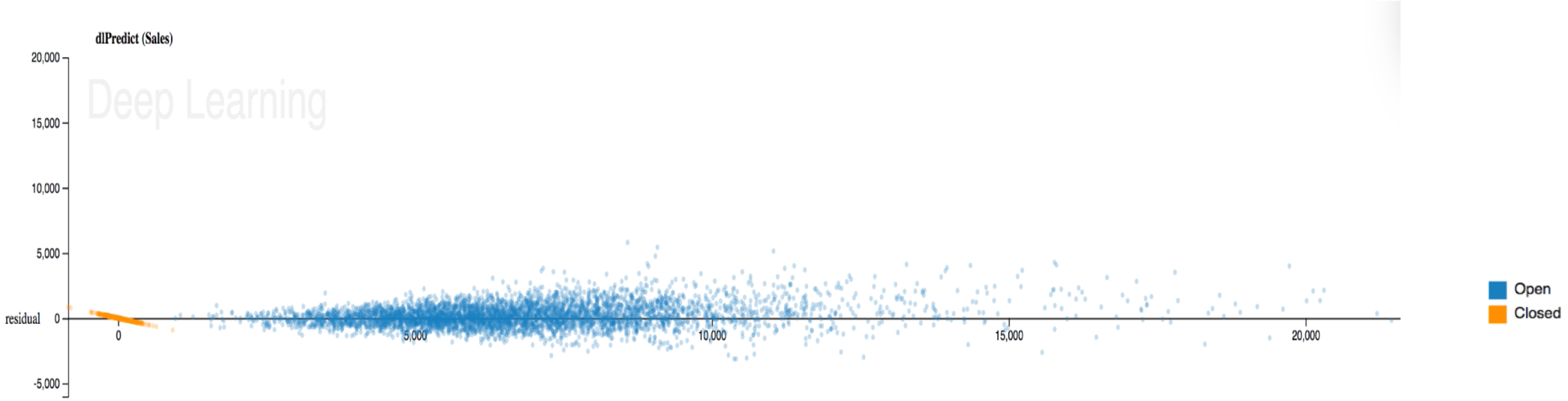
## Residual plots.

Rossmann Stores Kaggle dataset (https://www.kaggle.com/c/rossmann-store-sales)



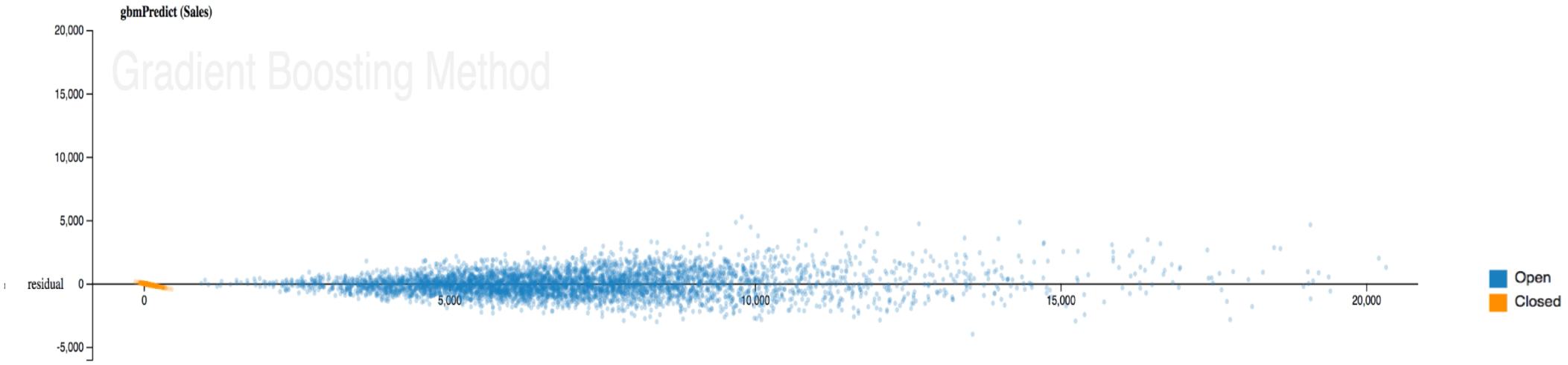
# Diagnostics

Residual plots.



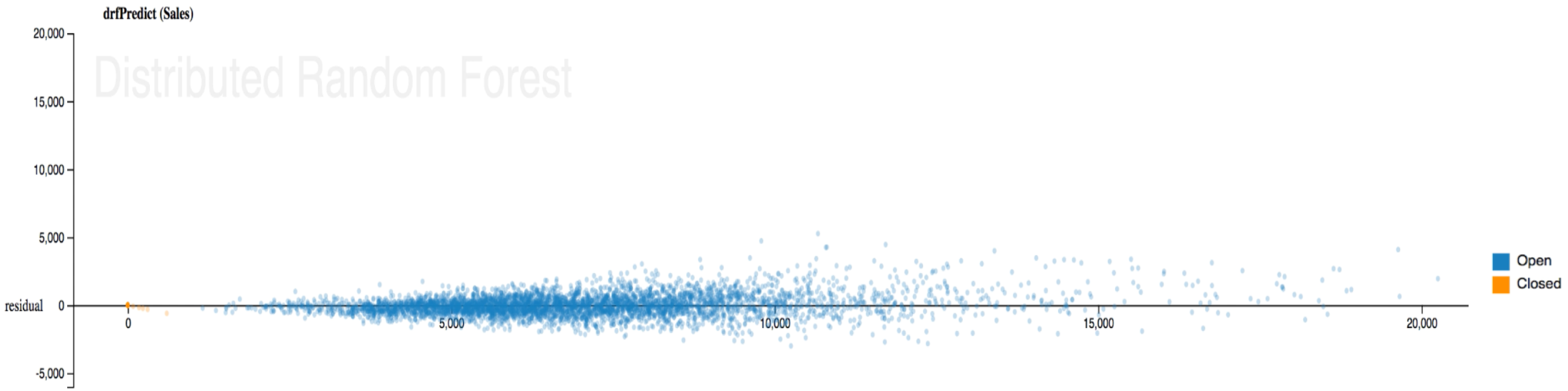
# Diagnostics

Residual plots.



# Diagnostics

Residual plots.



# Diagnostics Summary

- ALWAYS look at residual plots after fitting a model.
  - Whether the fit was automatic or manual, this is essential.
  - A statistic (e.g., RMSE) is a complement to a graphic, not a substitute.
- Features to look for in residual plot:
  - Band of residuals symmetrically distributed around horizontal line at zero
    - This implies there is no trend across fitted (estimated) values
  - Uniform spread of residual values in all local regions of fitted values
    - If you use classic statistics (*t, F*, …) residuals in local regions must be normally distributed.
- If residuals don't seem to meet these conditions, re-model.
  - Add new predictors and/or try transformations of variables.
  - Then look at residuals for new model.
  - Don't be afraid to iterate this process.