Philosophical Detour

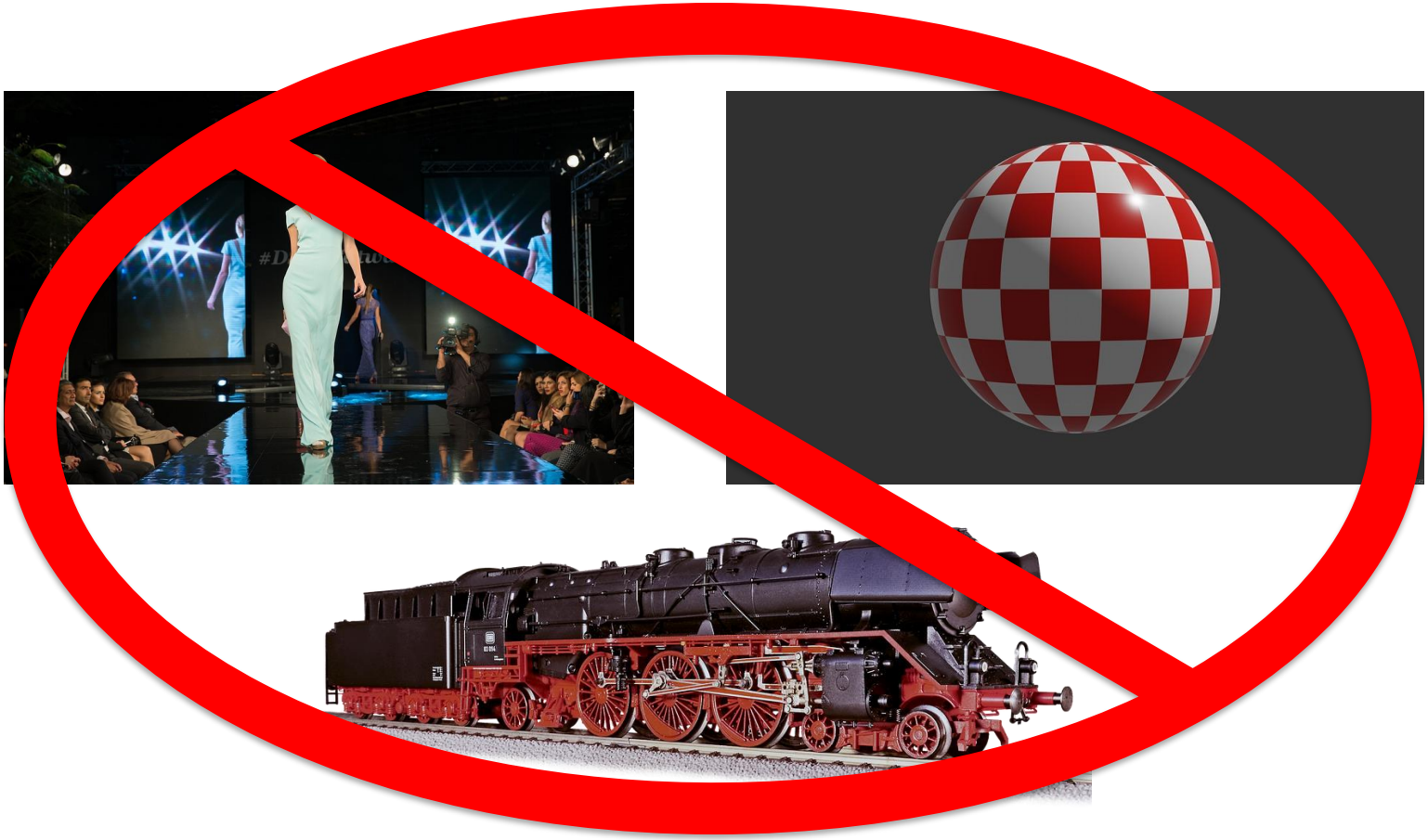
Modeling

Real-world Research

- Nature of Phenomena
  - "Complex"
  - Quantification (measurement) difficult IF possible
- To circumvent these issues we use models
  - Simplifies the "Reality"
  - Allows for unexplained ("error")

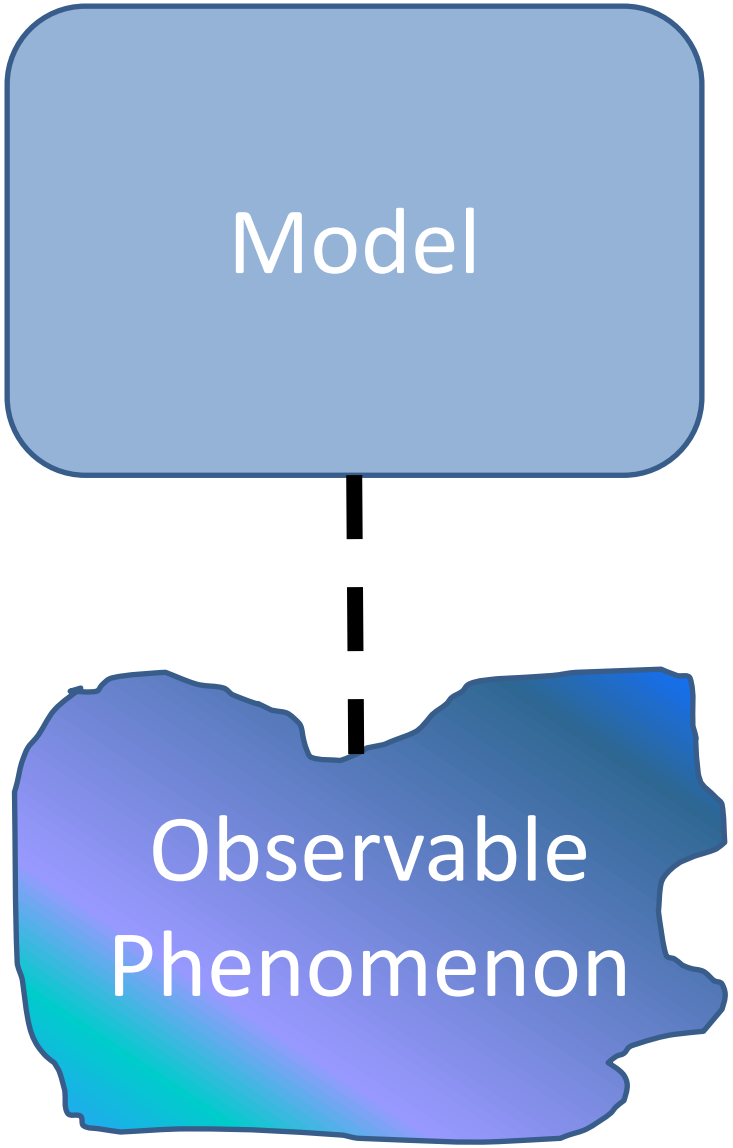
WHAT'S A MODEL

This is not the type of model I'm talking about



Model

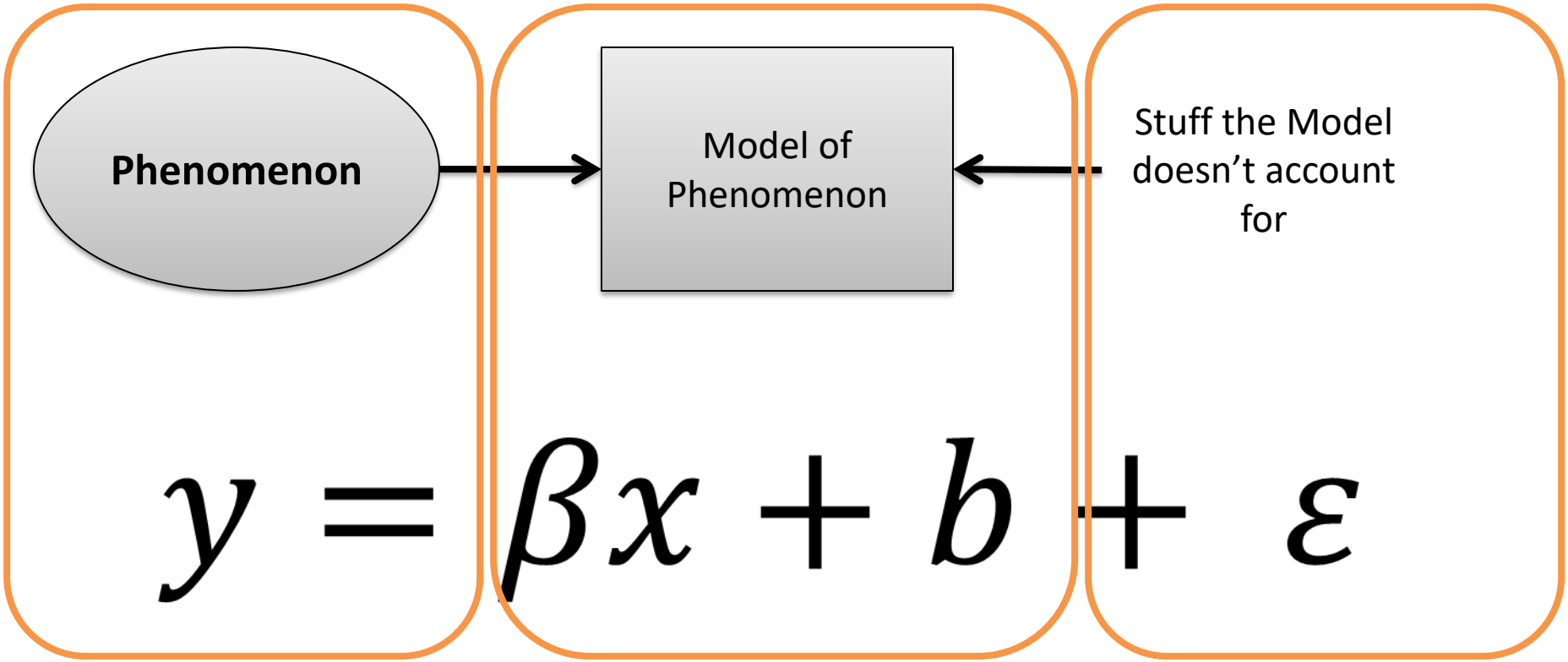
- Model
  - Simplified representation of something observable
    - Describe, Predict, and potentially Change
  - Ex:
    - $A = \pi * r^2$
    - $a^2 + b^2 = c^2$
    - $F = m * a$



- Approximation
- Preserves ‘key elements’
  - Rectangular shape
  - Bluish/grey color

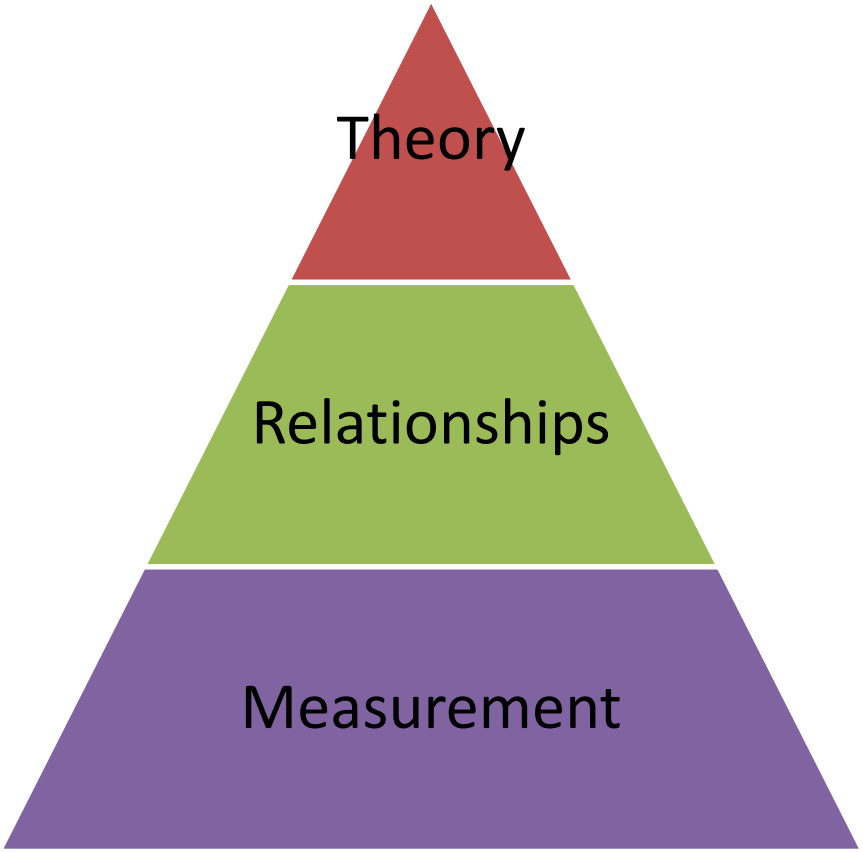
- Dotted line is the **fit** between the model and the phenomenon

Models


Phenomenon
Model of Phenomenon
Stuff the Model doesn't account for
$y = \beta x + b + \varepsilon$

How are we modeling?

- Measurement
  - Variable type
  - Scale of measurement
- Relationships
  - Statistical models
- Theory
  - Inferences we draw from results
  - (Not as much of a concern for this class)



8180 Review:

VARIABLES AND SCALES OF MEASUREMENT

Variable Types

- Qualitative
  - *Categorical*- Names a category
- Quantitative
  - *Discrete*- Comes in nice little packages (can't have half units)
    - Ex: People (can't have half a person)
  - *Continuous*- Varies infinitely between two values (fractionally)
    - Ex: Height (can be 5' 10.234817485712354351235436"

Scales of measurement

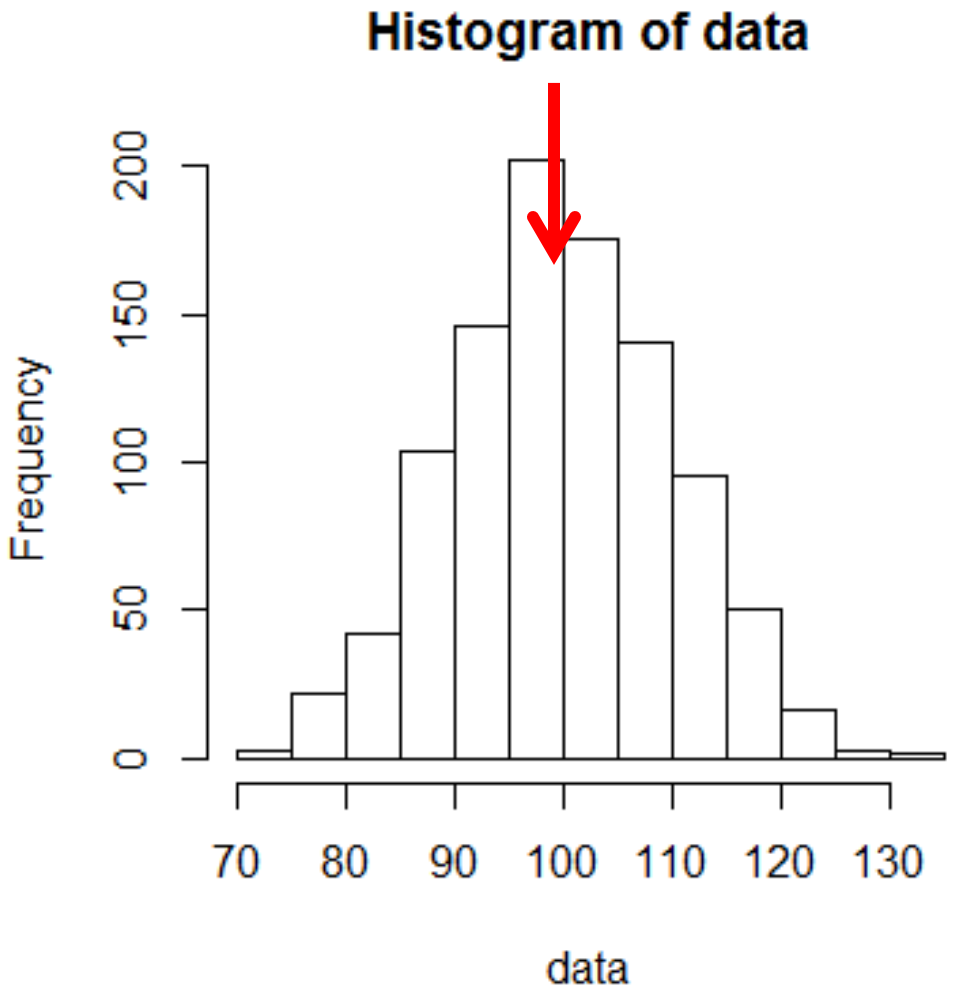
- **Nominal-** (Naming scales) Categorical
  - Ex: Group A, Blues, Person C, #1023
- **Ordinal-** (Ordering scales) Ranks individuals in sequences; distances between scores not same size
  - Ex: 1st place, 2nd place, 3rd place,…
- **Interval-** Intervals between scores of the same size and comparable
  - Ex: 20°C change is always the same; 10-pt increase in IQ is always the same
- **Ratio-** Same as interval scales, but with an absolute zero (a gold standard)
  - Ex: 20°K (0°K = no heat)

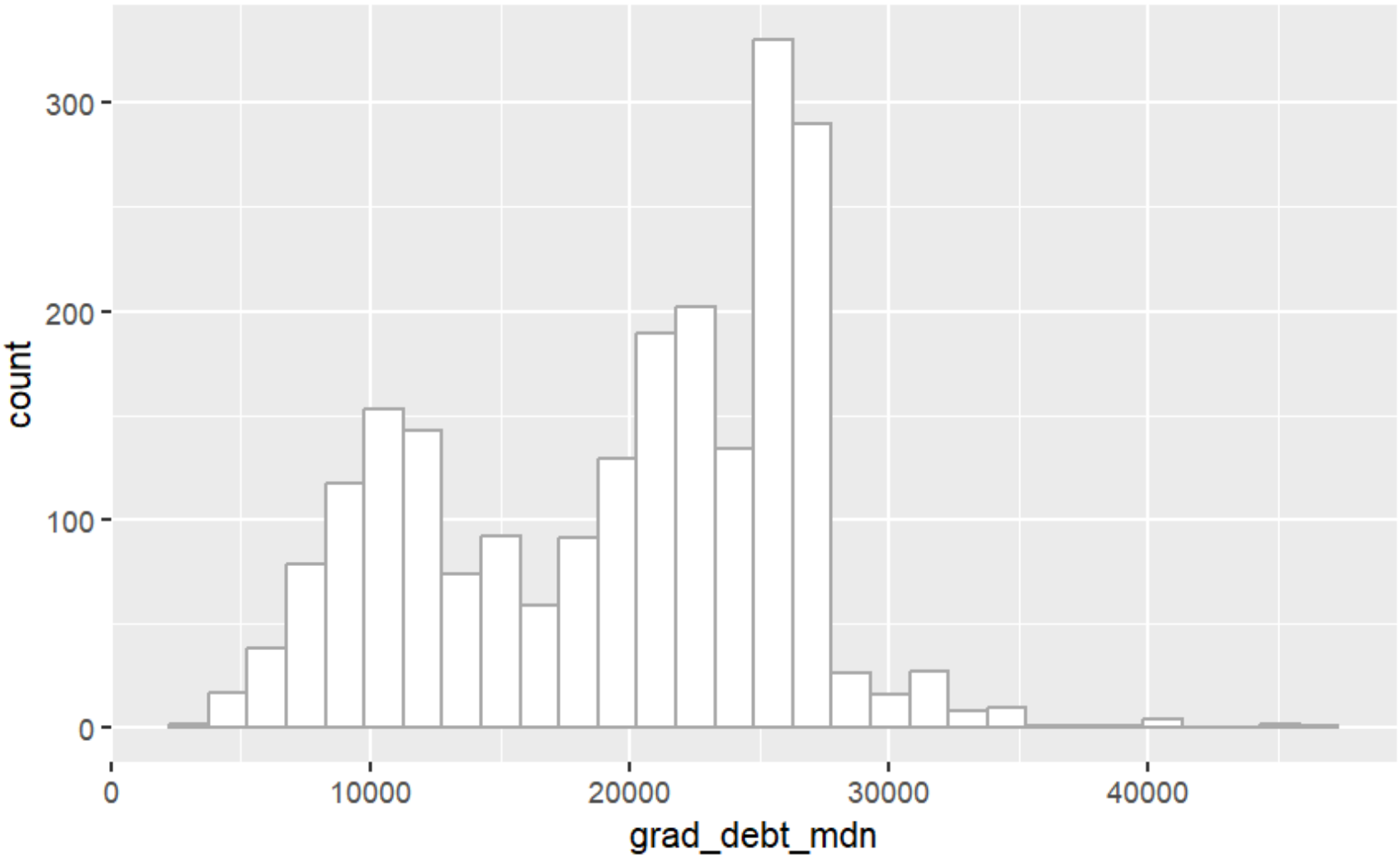
WHY IS THIS IMPORTANT

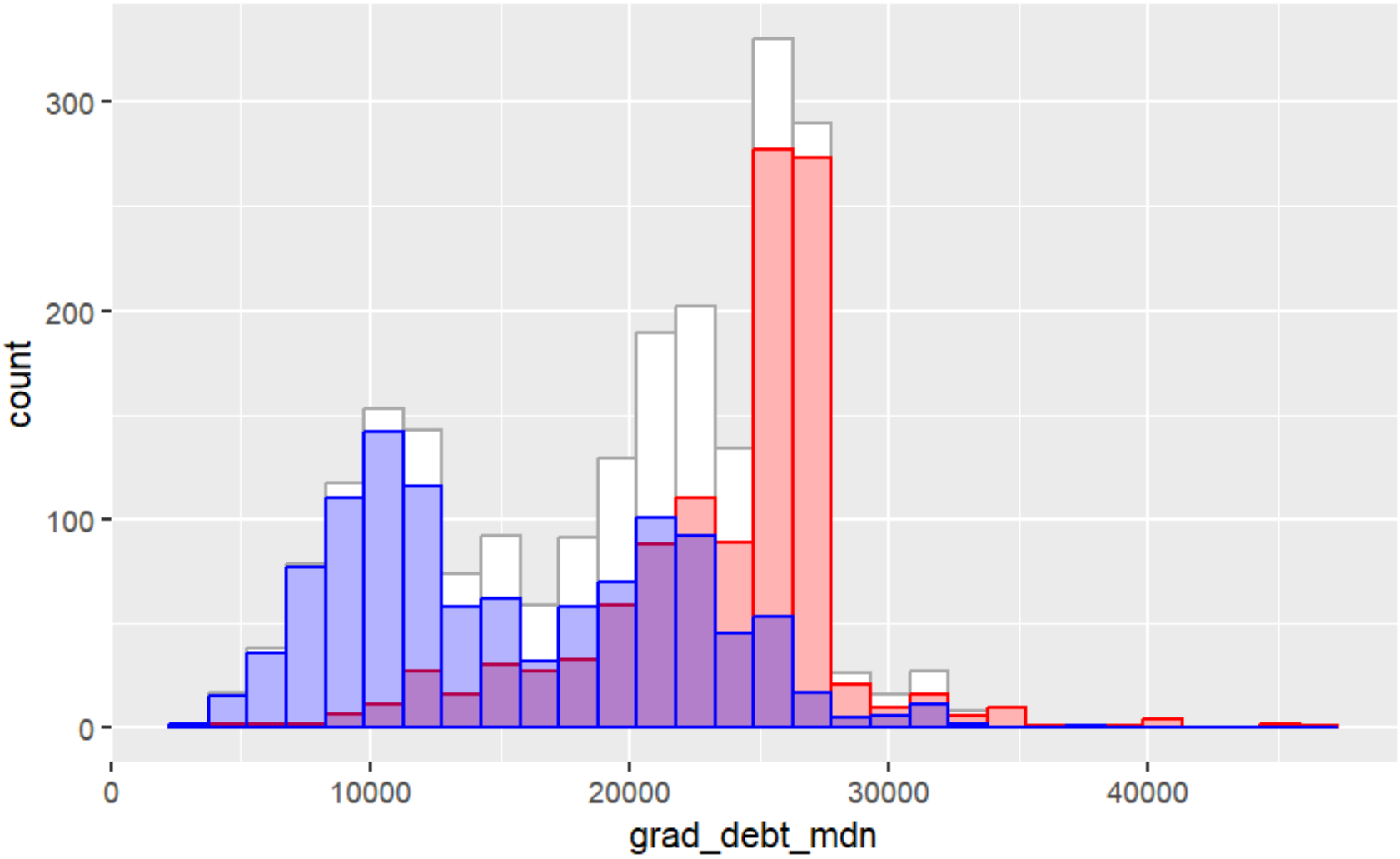
Central tendency

- Because the data ‘centers’ around them
- These are the simplest statistical models that we have
  - **Median-** the score that falls in the middle of an ordered list
  - **Mode-** the most frequently occurring score
  - **Mean-** the arithmetic average score

Median	Mean
99.60	99.80




300
200
100
0
count
0
10000
20000
30000
40000
grad_debt_mdn


300
200
100
0
count
0
10000
20000
30000
40000
grad_debt_mdn

Discretizing (Binning)

- Equal-frequency binning
- Equal-width binning
- K-means clustering

Discretizing (Binning)

- Equal-frequency binning
  - n-tiles
    - Medians, quartiles, quintiles, deciles, etc.
  - Equal representation across range
  - Parallels the original distribution
    - Good for model input
- Equal-width binning

- K-means clustering

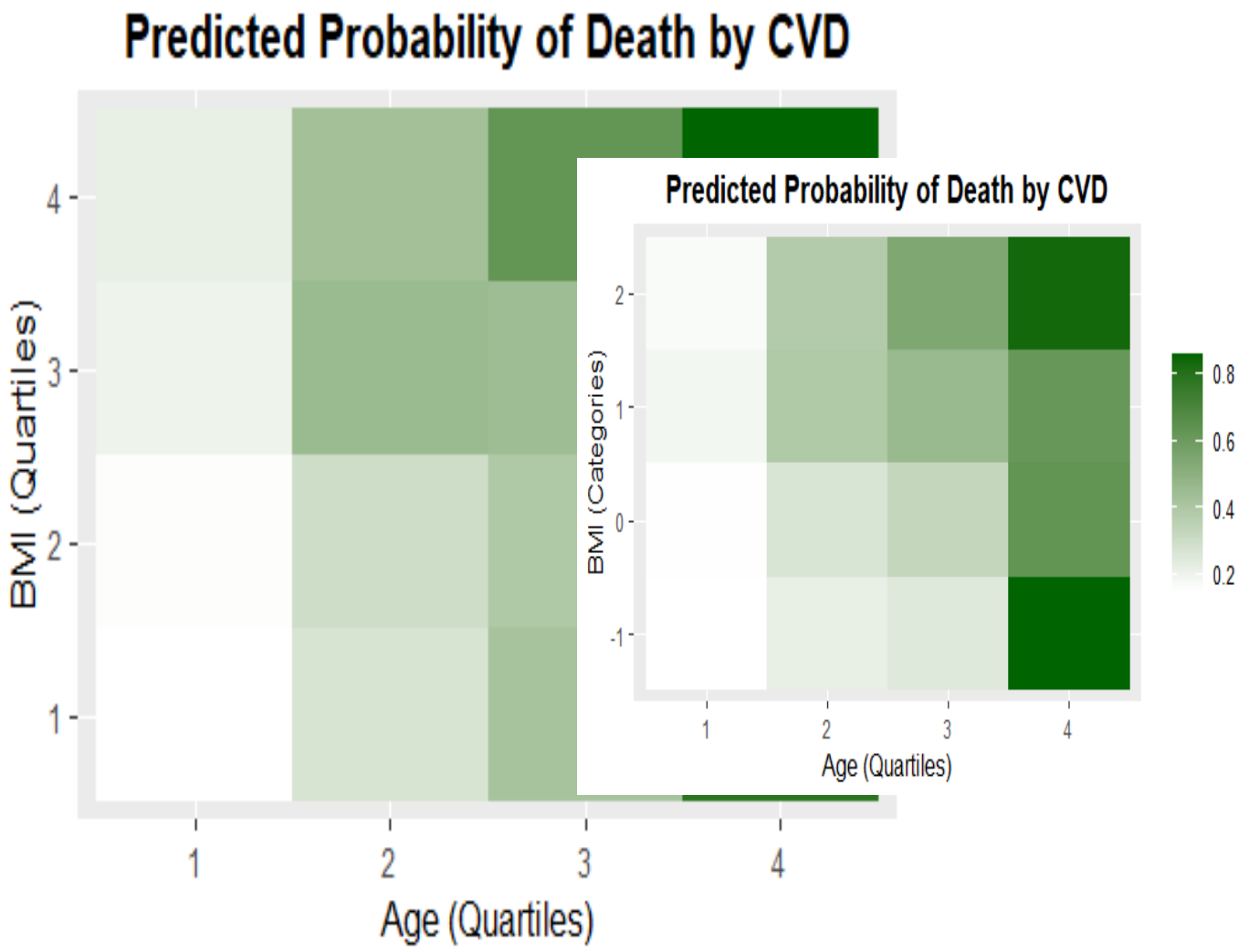
Discretizing (Binning)

- Equal-frequency binning

- Equal-width binning
  - Each bin is the same size of the range (width)
    - Age, GPA, etc.
  - Convenient for interpretation
  - Must take care when determining the width
- K-means clustering

Discretizing (Binning)

- Equal-frequency binning
- Equal-width binning
- K-means clustering
  - Each bin is determined Maximum Likelihood Optimization
    - Cases belong to the “closest mean”
  - Can identify useful profiles/typologies
  - Category labels must be interpreted *post hoc* and can be multidimensional


Predicted Probability of Death by CVD
BMI (Quartiles)
Age (Quartiles)
1
2
3
4
Predicted Probability of Death by CVD
BMI (Categories)
Age (Quartiles)
-1
0
1
2
0.8
0.6
0.4
0.2