MAYO
CLINIC

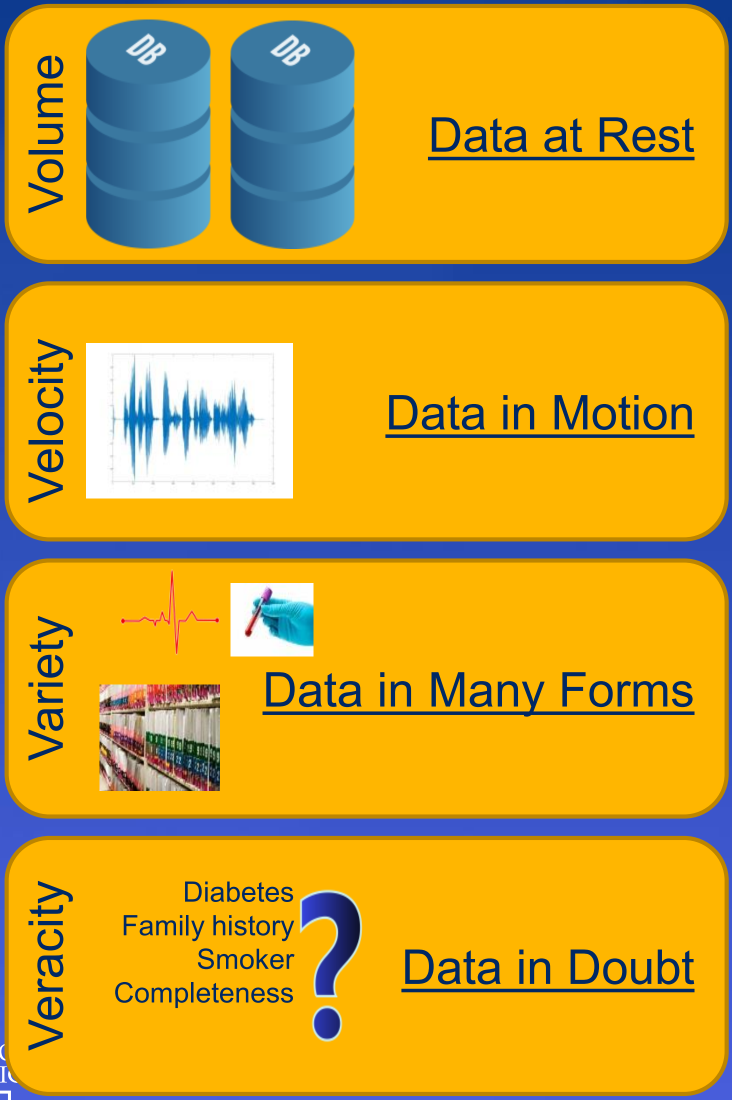
Session 7: Big Data

Rickey Carter, PhD
Professor of Biostatistics
Mayo Clinic
Jacksonville, FL

Conflicts of Interest

No conflicts of interest reported.

What is Big Data?



Volume

Data at Rest

Velocity

Data in Motion

Variety

Data in Many Forms

Veracity

Diabetes
Family history
Smoker
Completeness

Data in Doubt

Focus -- Veracity

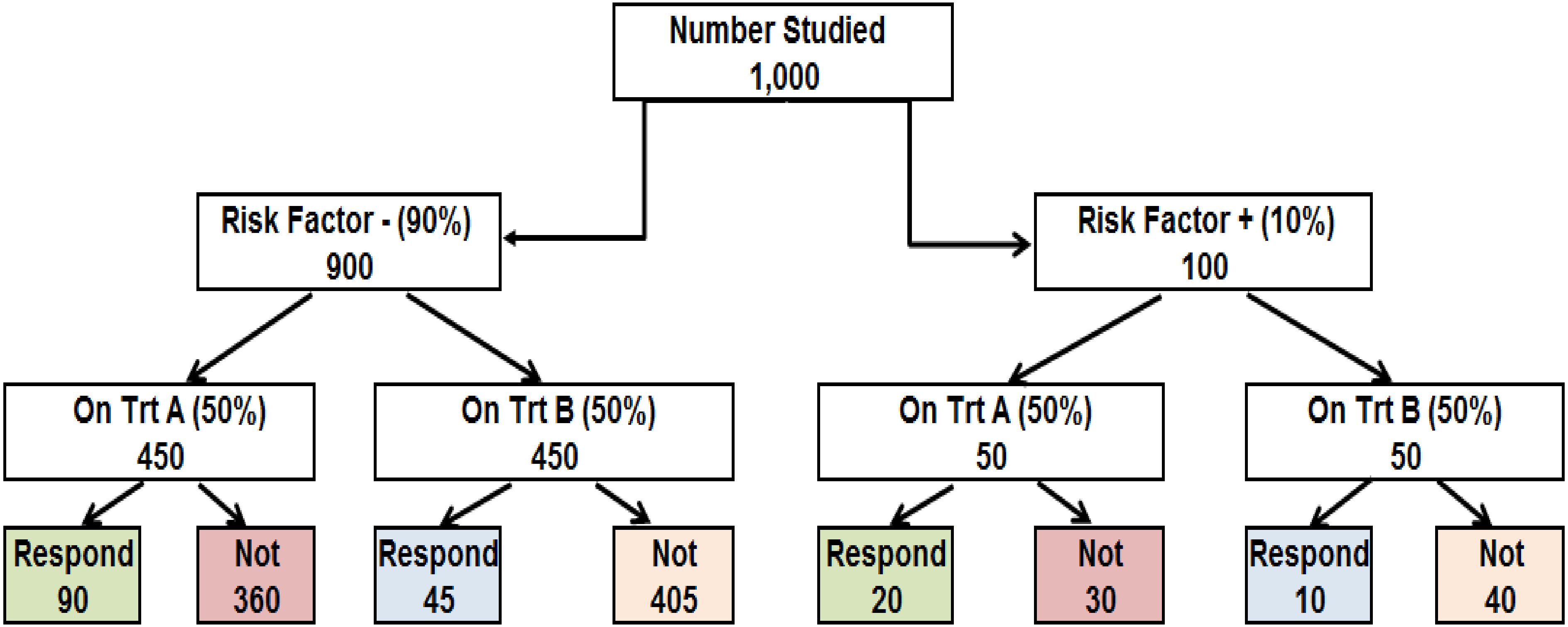
- Clinical Trials & Experimental Studies
  - “Rigor and Reproducibility”
- Is EHR data sound?

Big Data Alone Does Not Fix Confounding

- Example: 2 existing treatments (CER)
  - Suppose Risk Factor +/-
  - Different treatment response rates but same relative benefit
    - RF+
      - Treatment A 40%, Treatment B 20% (Relative Risk = 2)
    - RF-
      - Treatment A 20%, Treatment B 10% (Relative Risk = 2)

MAYO CLINIC

Example 1 – No Confounding (True RR = 2.0)


Number Studied
1,000

Risk Factor - (90%)
900

Risk Factor + (10%)
100

On Trt A (50%)
450

On Trt B (50%)
450

On Trt A (50%)
50

On Trt B (50%)
50

Respond
90

Not
360

Respond
45

Not
405

Respond
20

Not
30

Respond
10

Not
40

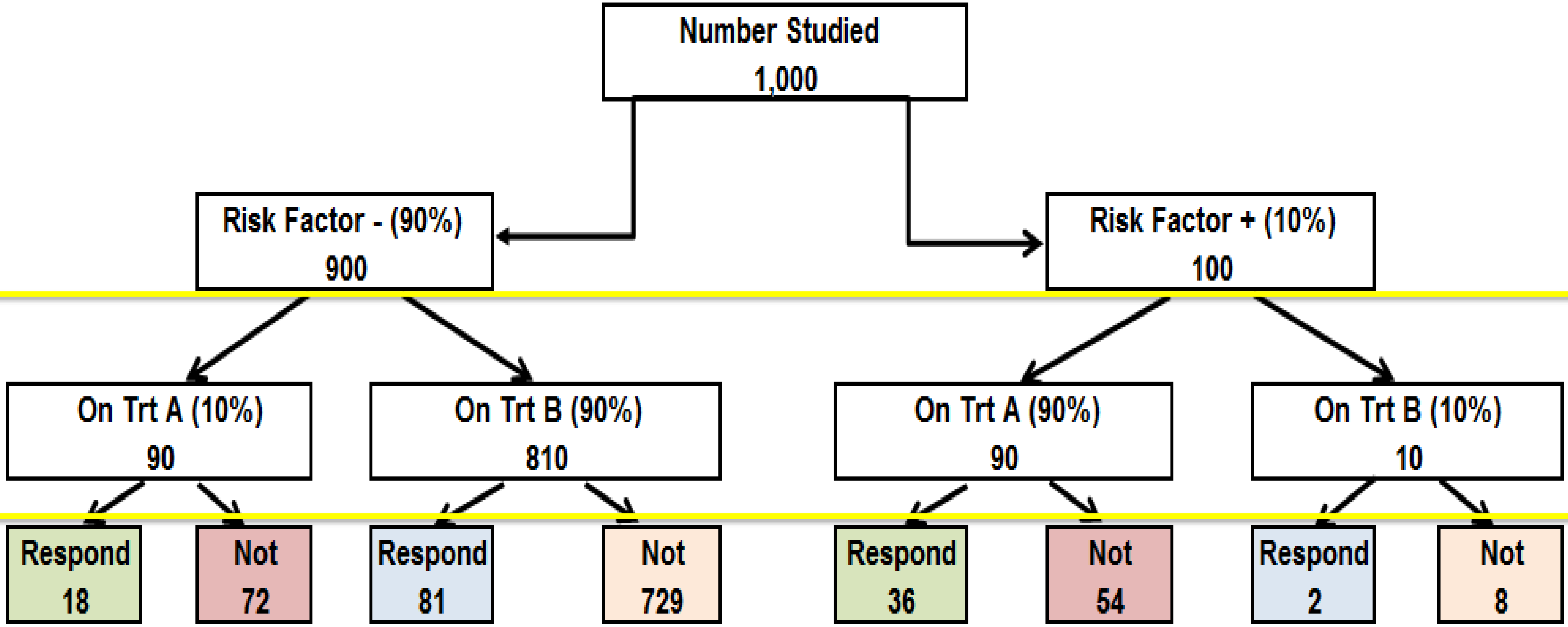
Unadjusted Results

	Respond	Not
Treatment A	110	390
Treatment B	55	445

Relative Risk: 2 (95% CI: 1.48 to 2.7)

% Error 0%

Example 2: Confounded Data



Number Studied
1,000

Risk Factor - (90%)
900

Risk Factor + (10%)
100

On Trt A (10%)
90

On Trt B (90%)
810

On Trt A (90%)
90

On Trt B (10%)
10

Respond
18

Not
72

Respond
81

Not
729

Respond
36

Not
54

Respond
2

Not
8

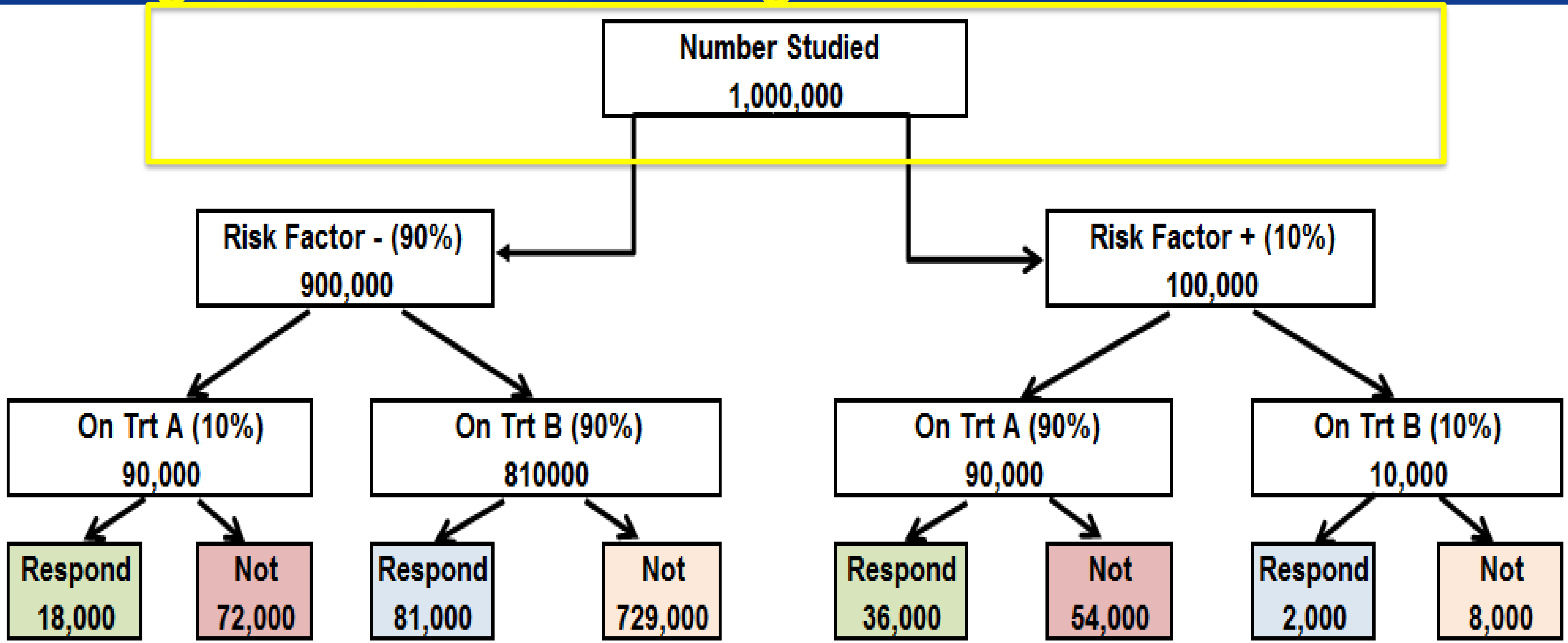
Unadjusted Results

	Respond	Not
Treatment A	54	126
Treatment B	83	737

Relative Risk: 2.96 (95% CI: 2.19 to 4.01)

% Error: 48%

Big Data is Still Wrong


Number Studied
1,000,000

Risk Factor - (90%)
900,000

Risk Factor + (10%)
100,000

On Trt A (10%)
90,000

On Trt B (90%)
810000

On Trt A (90%)
90,000

On Trt B (10%)
10,000

Respond
18,000

Not
72,000

Respond
81,000

Not
729,000

Respond
36,000

Not
54,000

Respond
2,000

Not
8,000

Unadjusted Results

	Respond	Not
Treatment A	54,000	126,000
Treatment B	83,000	737,000

Relative Risk: 2.96 (95% CI: 2.94 to 2.99)

% Error: 48%

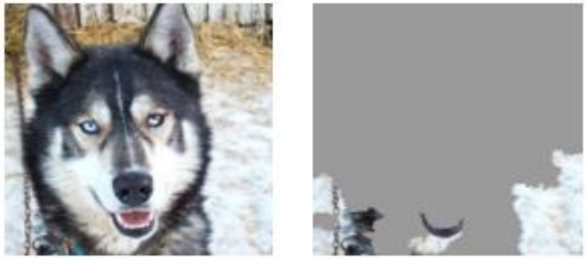
Algorithms – AI! Deep Learning!

Must challenge to look closer

- Evaluate drivers of predications and associations
- Overall model performance is inadequate

Local interpretable model-agnostic explanations (LIME)

Wolf
or
Dog?



(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

https://arxiv.org/pdf/1602.04938.pdf