Exploratory Data Analysis

What is EDA?

- Investigation
- Description
- Explanation

Data

EDA

Understanding

Challenges to Successful EDA?

- Data Quality
- Process / Method
  - Eg. H-test
- Organizational challenges
- Business / Strategic Challges.

Data Quality

- Missing → NULL
- Unknown Values
- Representative / Non R. Sample.
- Inconsistent Values (Data entry)

"Explanation" Issues

- Cause → Effect Analysis
  - E → C ; C → E
  - H → E ; H → C
  - E assoc. C

Distinguish?

- Hypothesis tests (on Bad Data?)

Organizational challenges

- Bureaucracy → Access, Speed
  → Change Process
  Esp. Data collection
- Ownership → Gate Keeping, Hoarding
- "Politics" → Status, Hierarchy, | Duplication, unclear Responsibility
- Governance → Legal
  → Internal Ethics

NB. HIPPO - "Senior" Opinions [Refute]

Business challenges

- define question
- ROI
- Time available
- Staff & "human" resources
- Strategic Value?

EDA Workflow

→ assumig most challenges solved

Universal

- Data Structure
- Distributions
- Correlations

} All Cols

- Filters + Distributions + Correlation

} i.e. "important" Rayes

- Graphs
  - → i.e. Factor Analysis

Cont.

Domain-Specific

- Formula (Eg. $BMI = \frac{w^2}{h}$)
- Predictive Models

  Eg. $y_{Expec.} = x_{BMI} + \cdots$
- Specialized Visuals

  Eg. Maps

Aim: Understand the data.

Example: Education

School Year 1 ... 2 ... 5

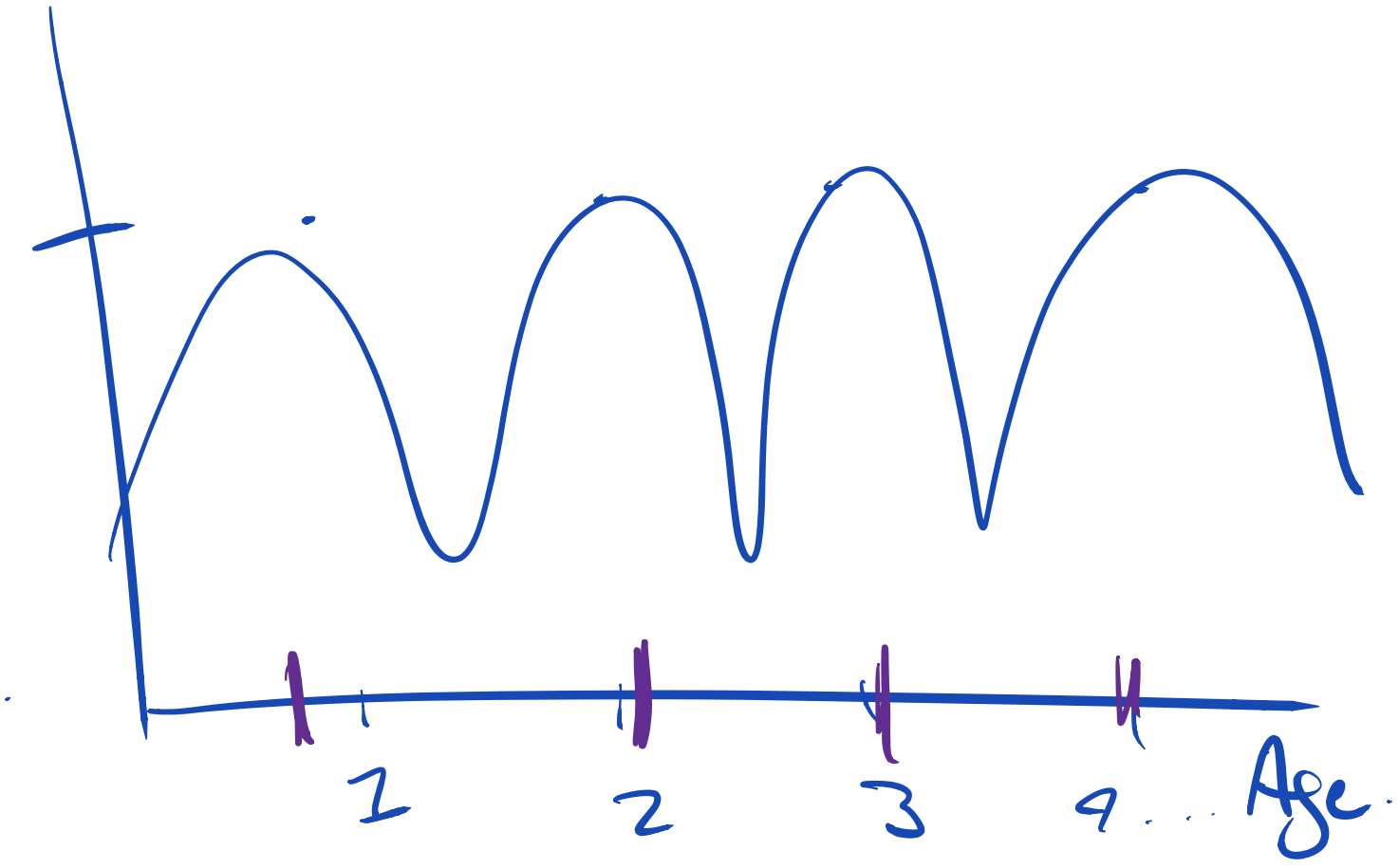
Data → ① Structure

Student ID

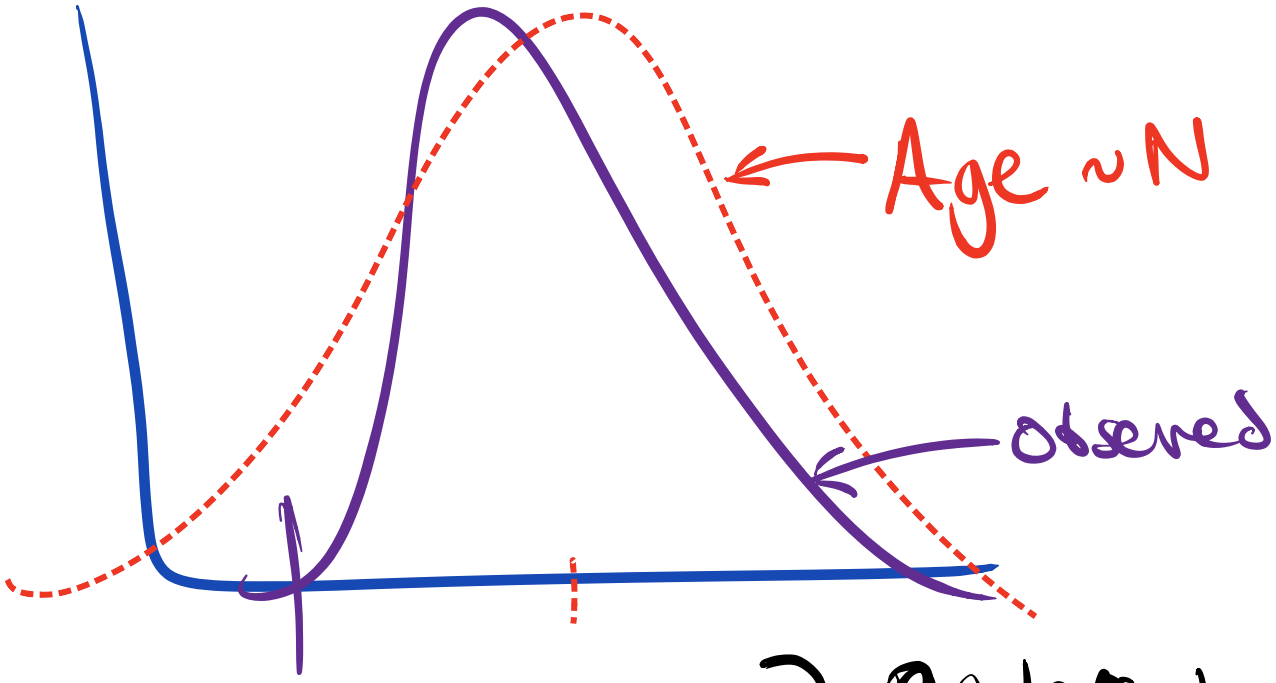
Year

Mean SAT

② Distribution



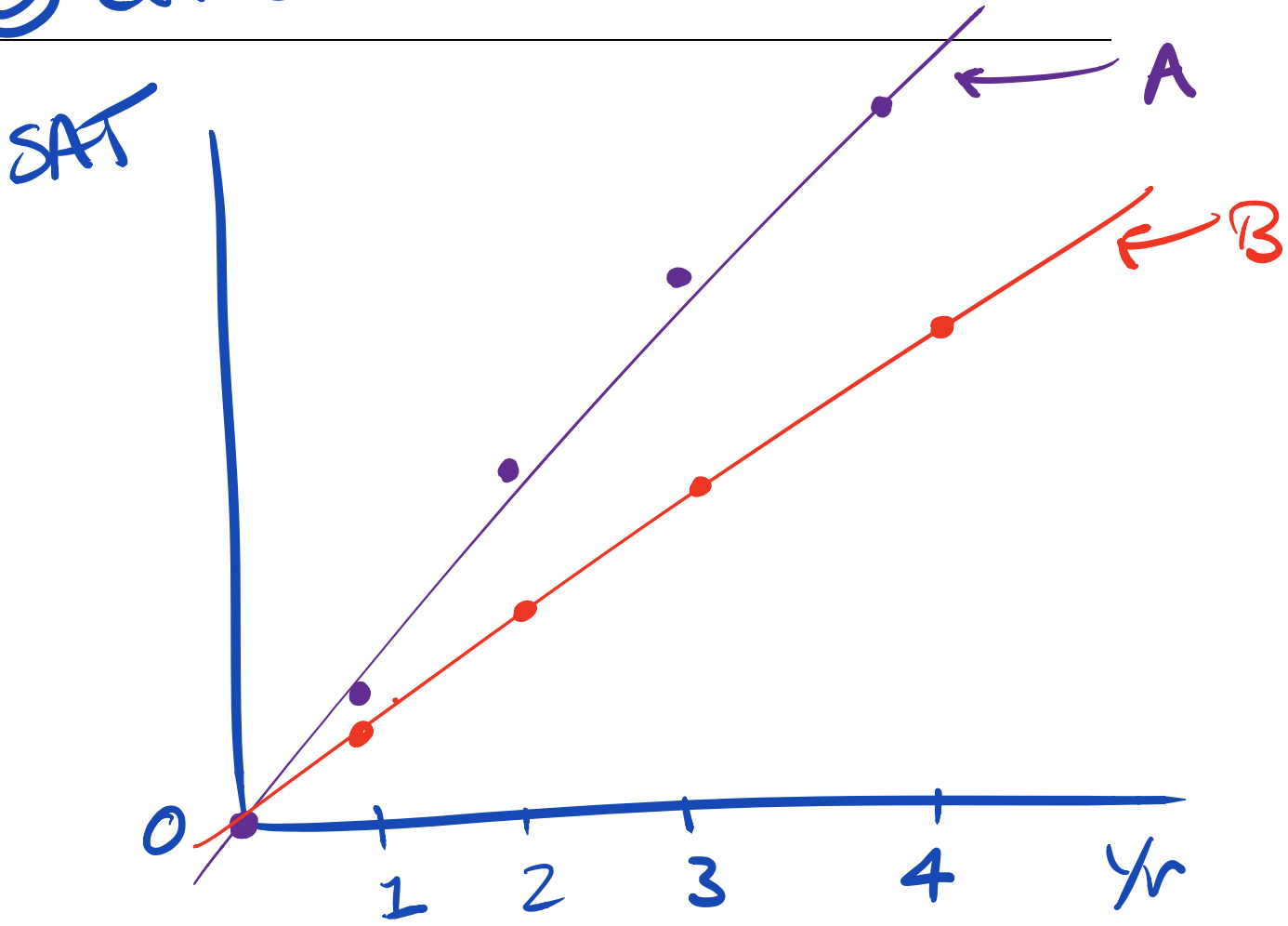
② Distn.



. . .

} Problems happen if you skip

③ Correlation

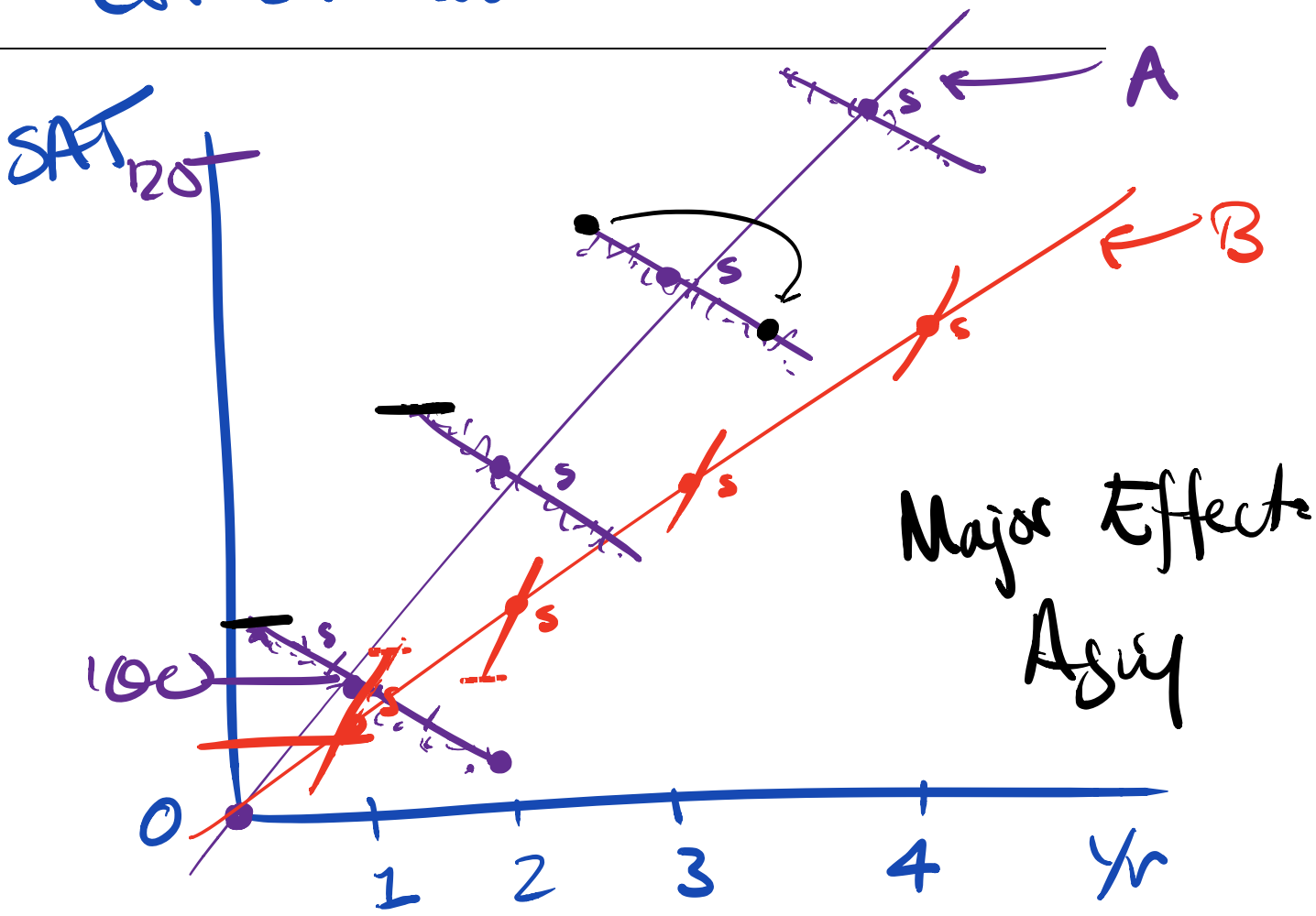


... Explain?

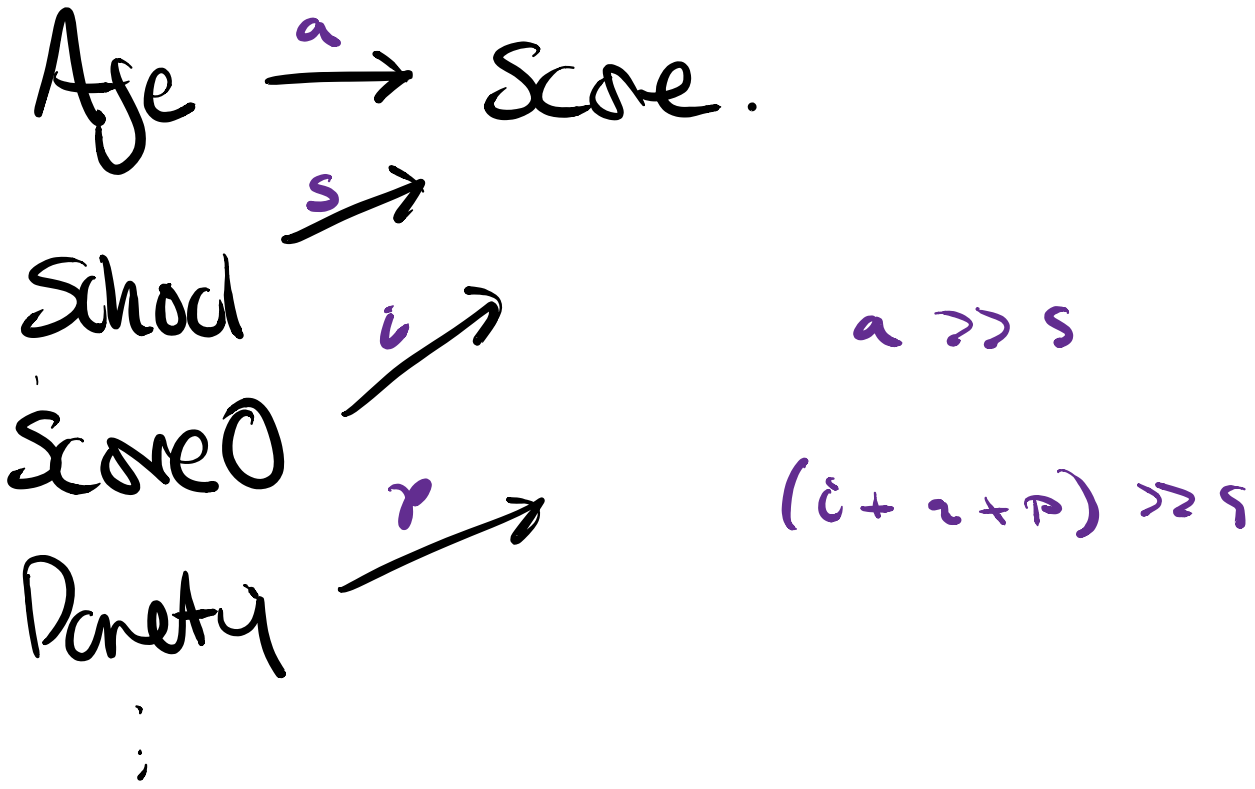
Is School A ?

School B ?

Correlation



Problem Caused Analysis



Simpsons Paradox

~~in/ex~~ $A > B$

though every subgroup A

$<$

every subgroup B.

Q&A