Integrative (Big?) Data Analysis

Patrick J. Curran
Andrea M. Hussong
Daniel J. Bauer

University of North Carolina at Chapel Hill

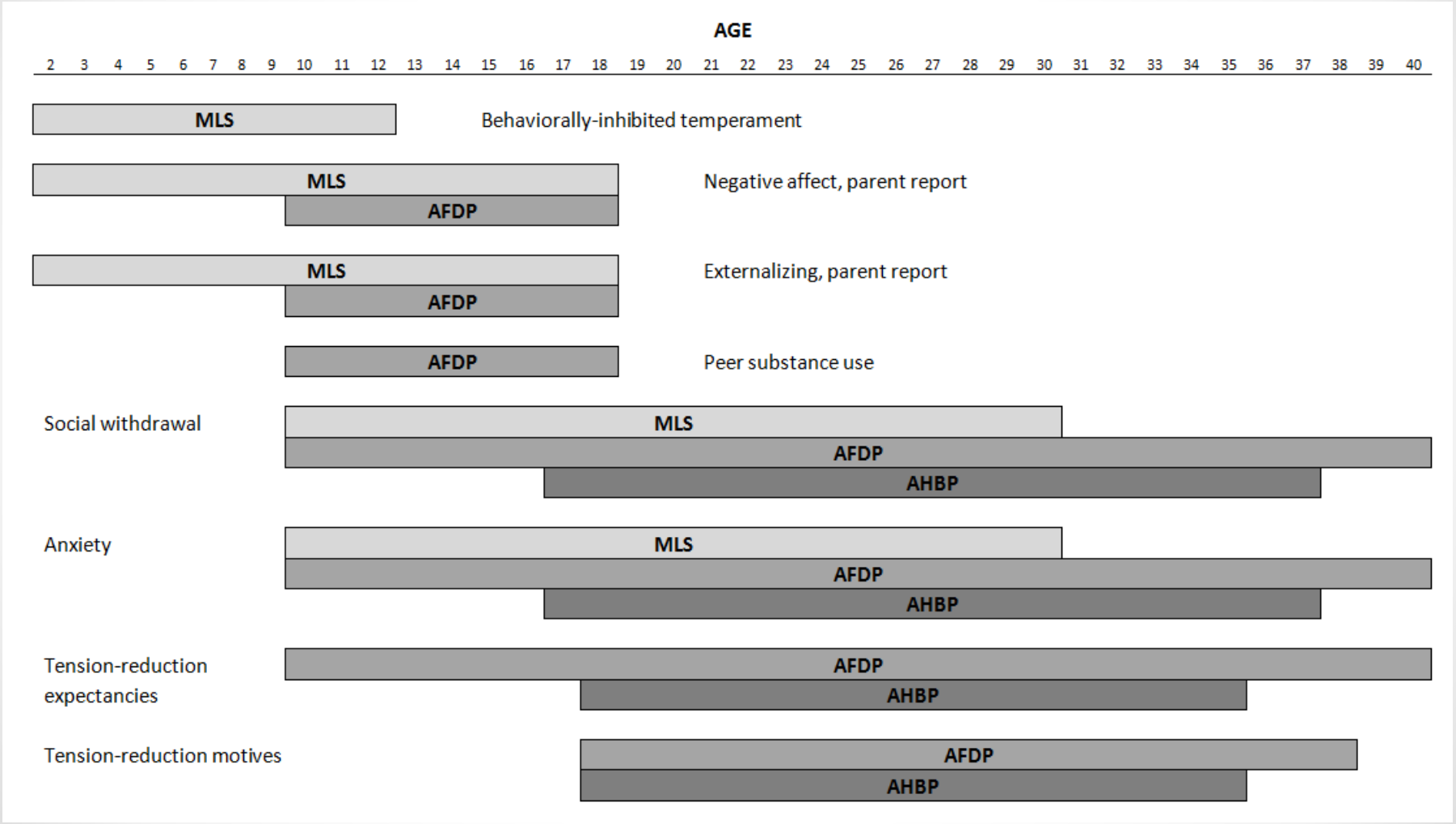
Integrative Data Analysis

- The fitting of statistical models to raw data that have been pooled across multiple independent samples
- More of a methodological framework than a specific set of analyses
- *Does the combination of multiple data sources provide a unique insight that is not possible in any individual data set when taken alone?*

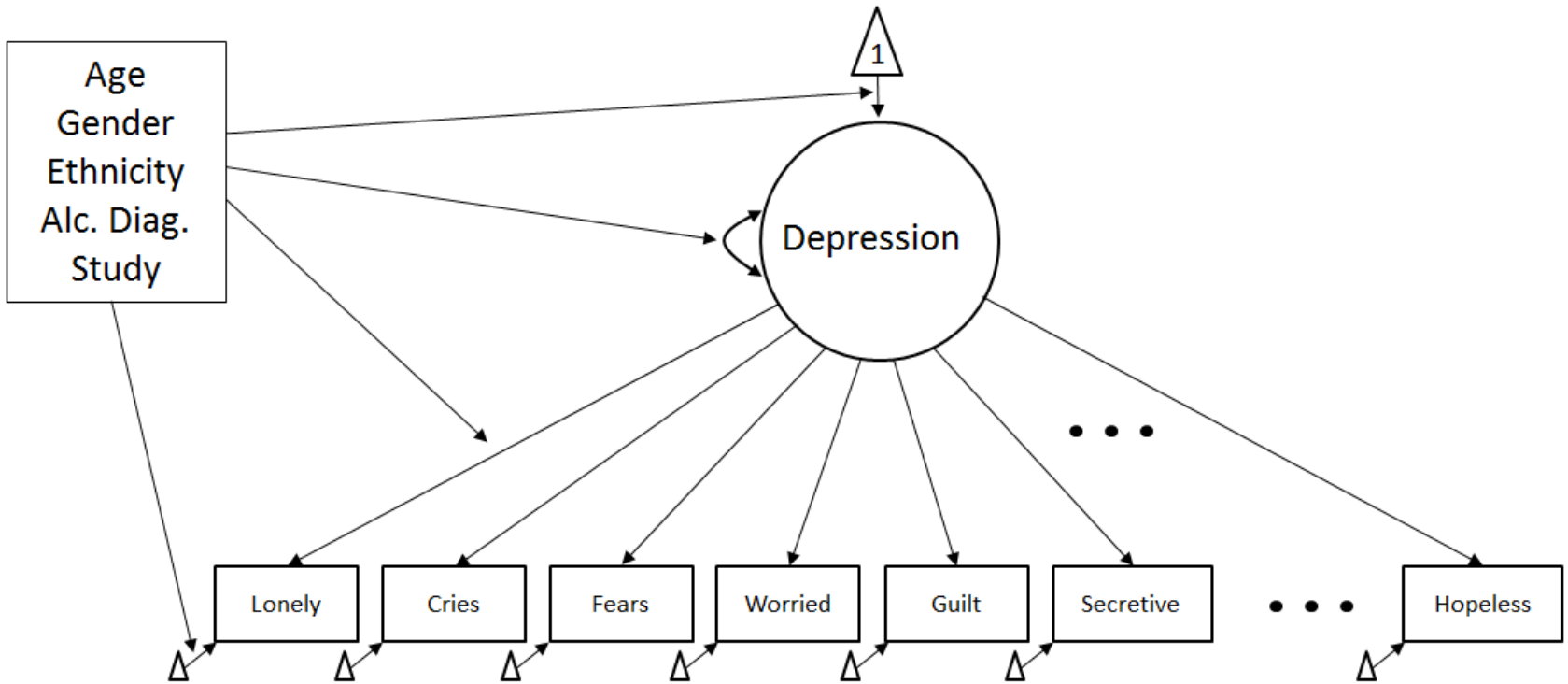
Integrative Data Analysis

- Potential advantages
  - o efficient use of resources
  - o greater age coverage
  - o greater power
  - o increased coverage of rare behaviors
  - o greater sample heterogeneity
  - o within- and across-study replication
- Potential disadvantages
  - o massive data management; complex methods
  - o critical: establish commensurate measures of all constructs across all contributing studies

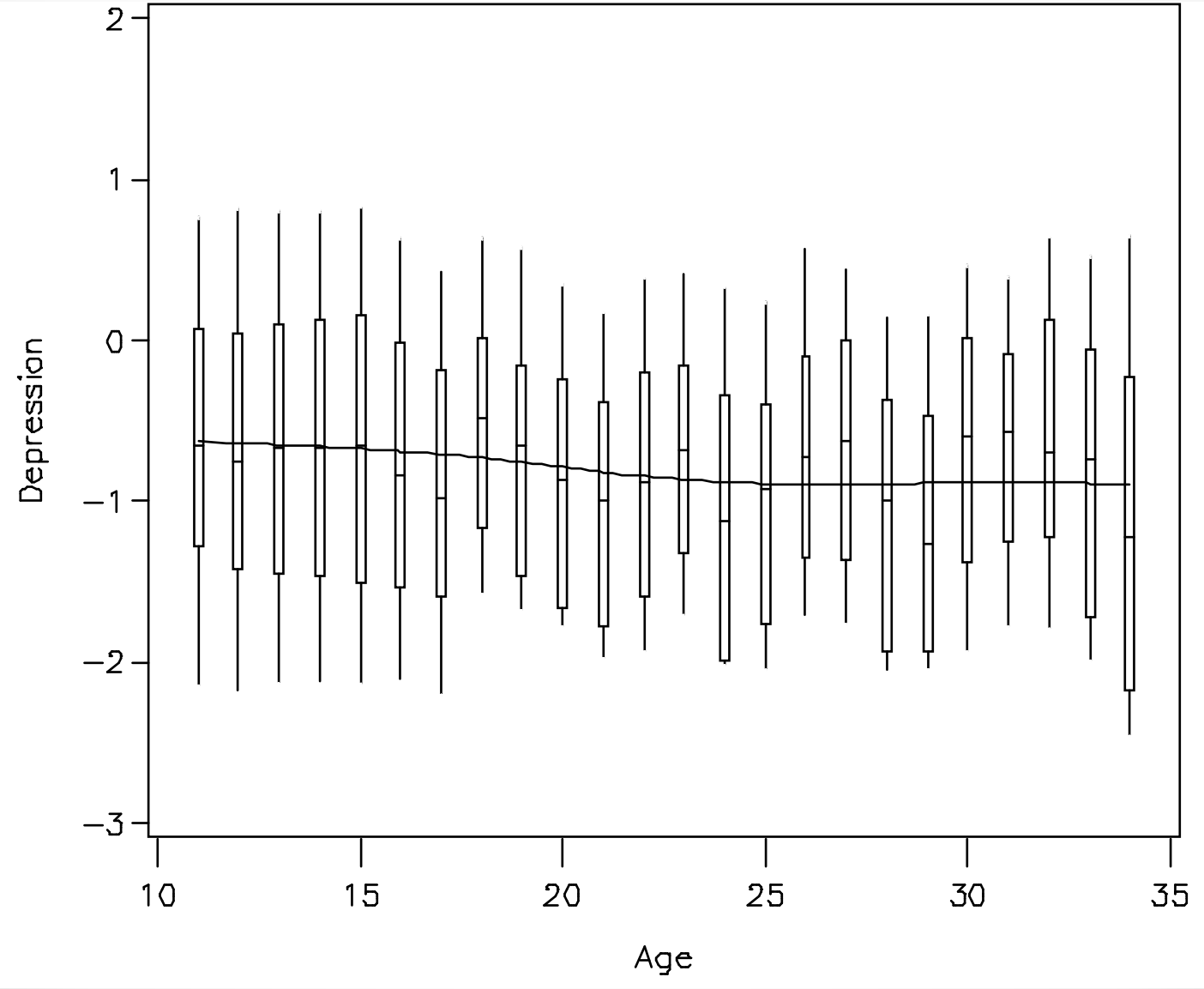
Very Brief Example



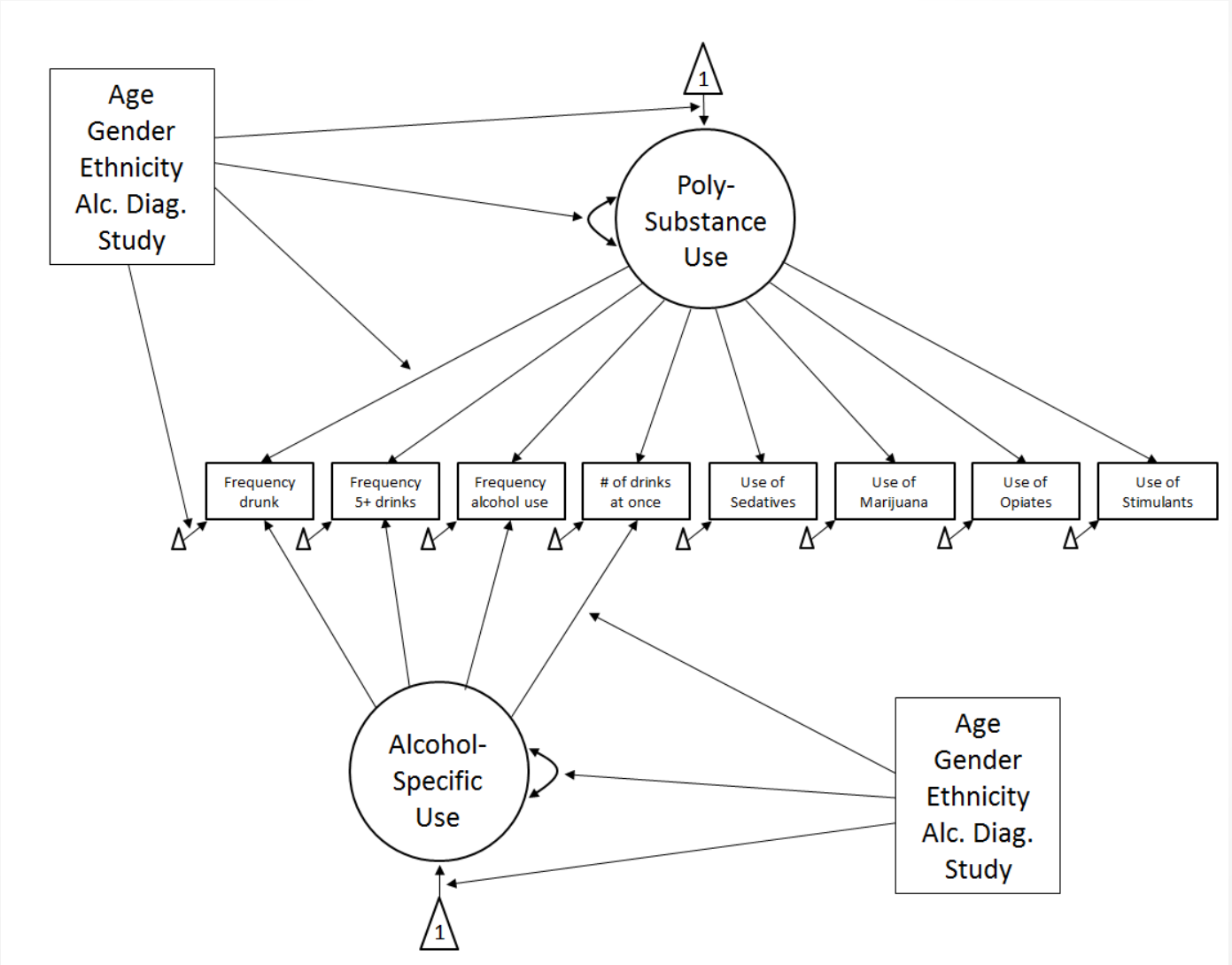
Measurement: Depression



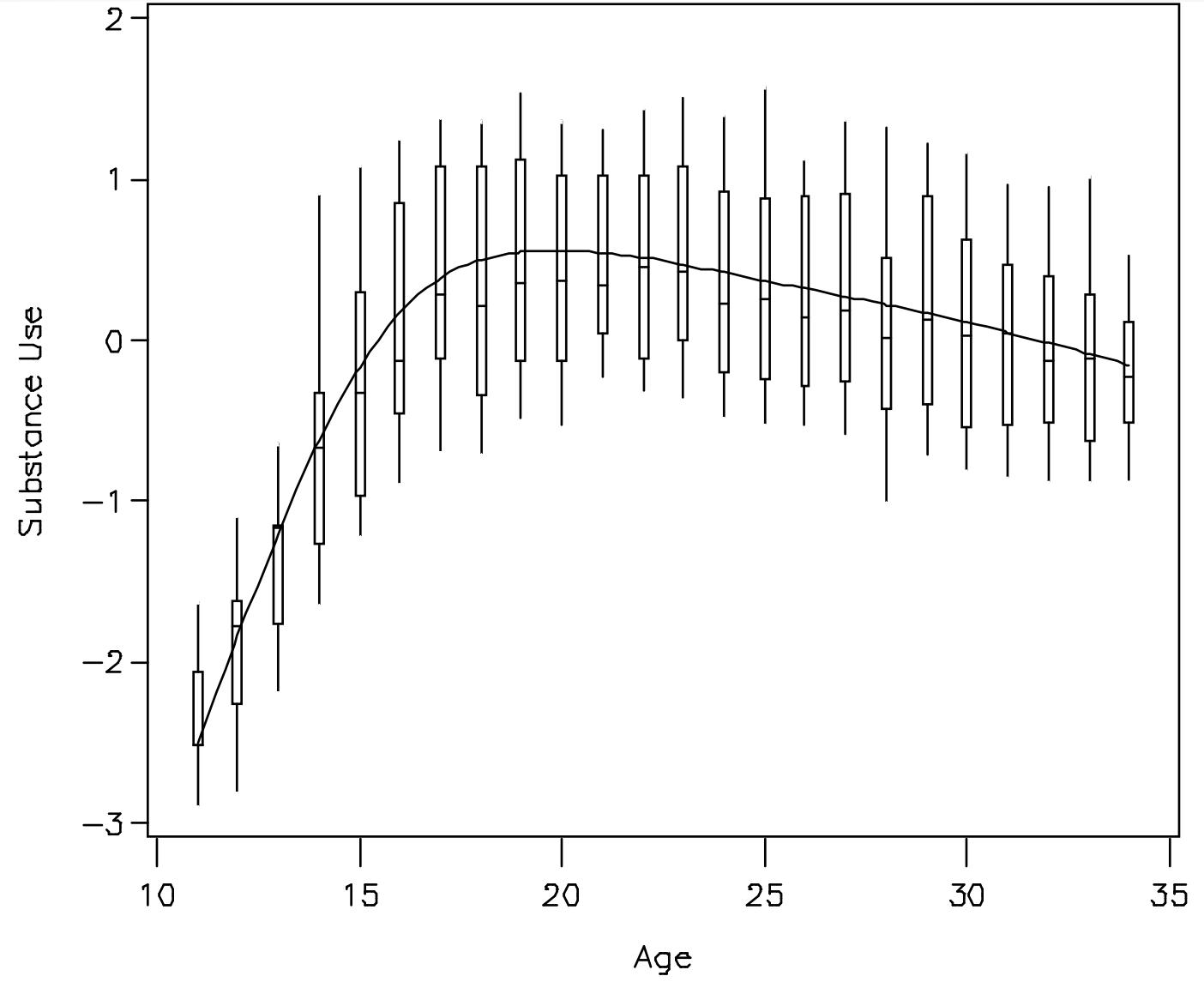
Scoring: Depression



Measurement: Substance Use



Scoring: Substance Use



Applications of IDA to "Big Data"

- Depends on how define "big data"
- Work thus far focused on "long" data
  - o ***concatenating data*** for additional **cases**
  - o may be less relevant for big data
- Instead consider creating "wide" data
  - o ***merging data*** for additional **measures** on cases
- But wide data much more challenging
  - o uncertainty via probabilistic matching of cases
- Great promise, but much work needed