MONASH University

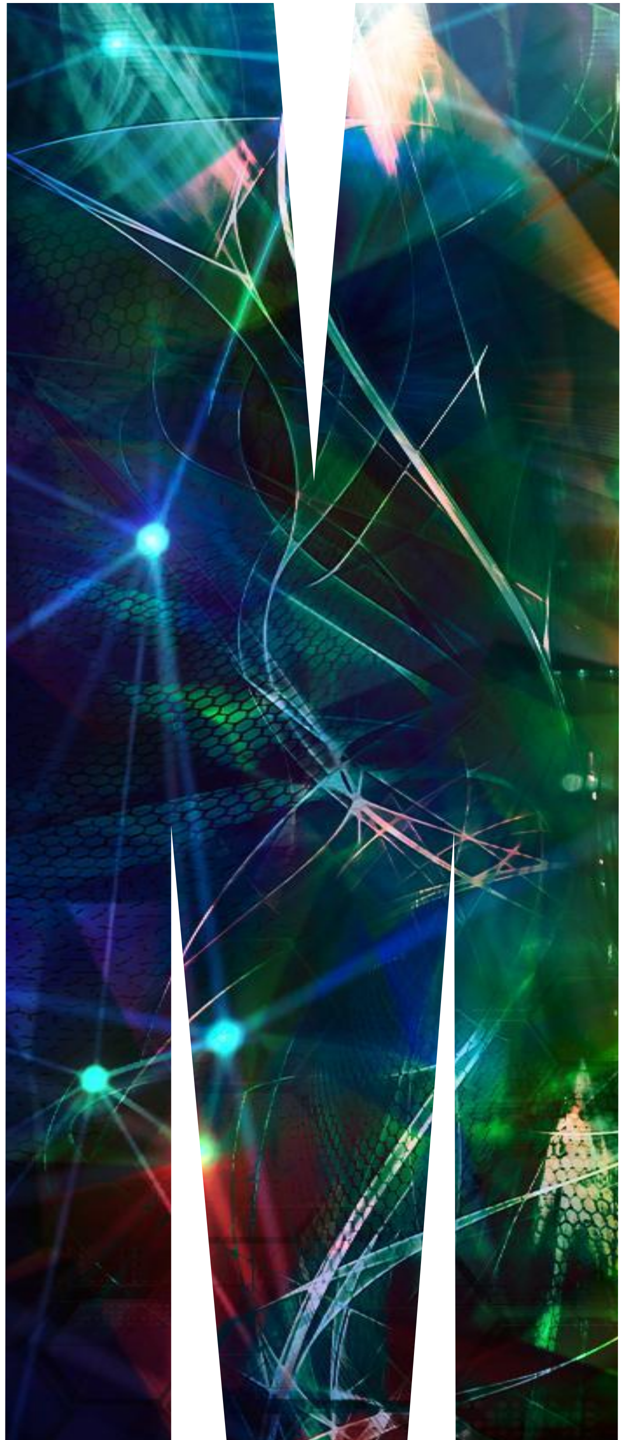
Longitudinal Data Analysis in R

AS4SAN 2024 Conference – Melbourne, Australia

27 May 2024

Dr. Michelle Byrne

Michelle.Byrne@monash.edu



MONASH
University

MONASH UNIVERSITY recognises that its Australian campuses are located on the unceded lands of the people of the Kulin Nations, and pays its respects to their Elders, past and present.

Overview

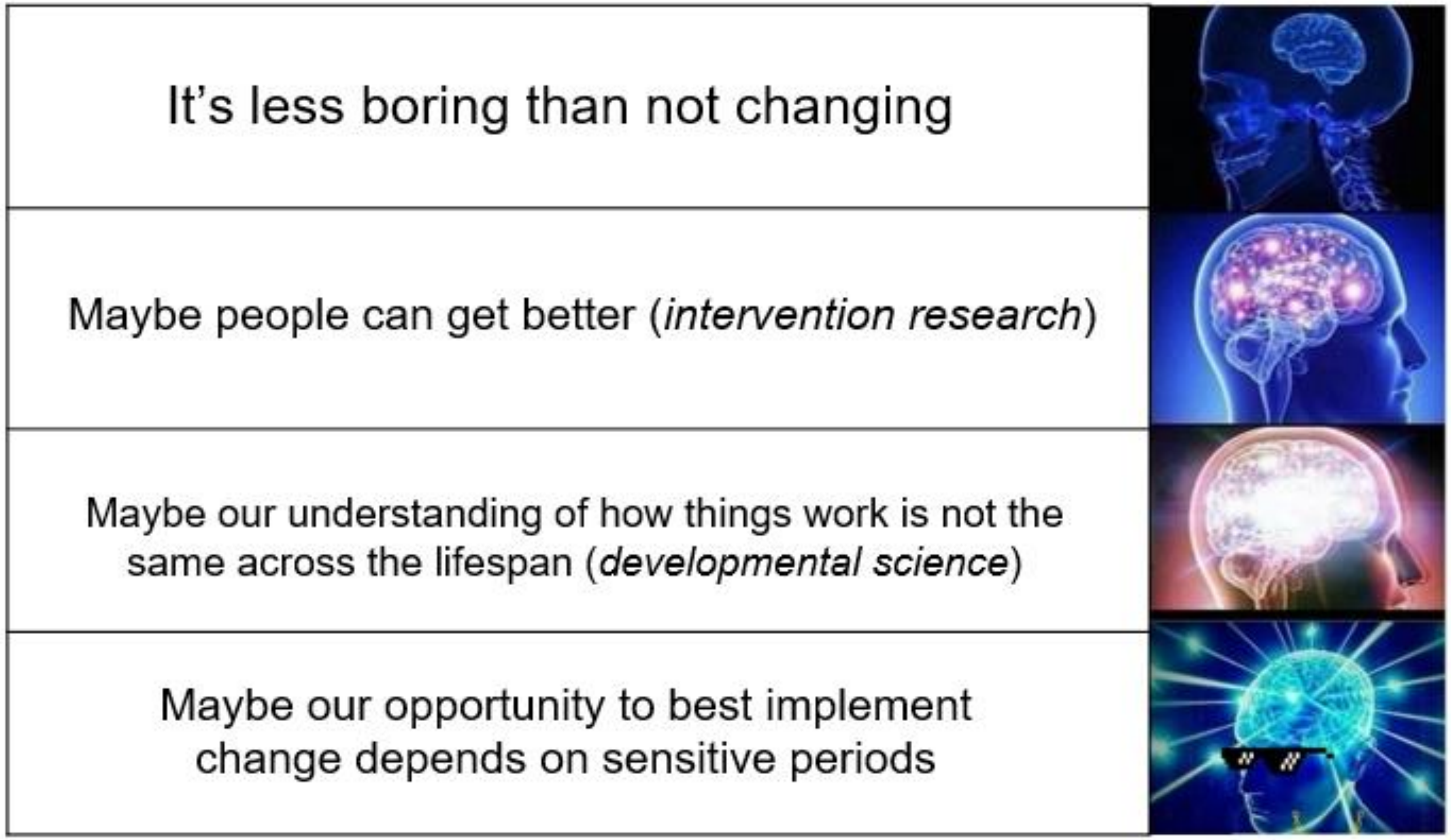
9:30 - 10:00 am	*Get to know you and your projects*
10:00 - 11:00 am	Lecture – but please ask questions • Rationale for longitudinal models • Multilevel models (including interaction/moderation) • SEM growth factor models • The shape of change: polynomials and splines
11:00 - 11:10 am	*Break*
11:10 am-12:00 pm	• Hands-on examples of MLM & SEM with code companion • Work on other models or your own data

Part 1 – Rationale for longitudinal models

Why do we use longitudinal models?

Let’s assume people can change

Why do we care about change?



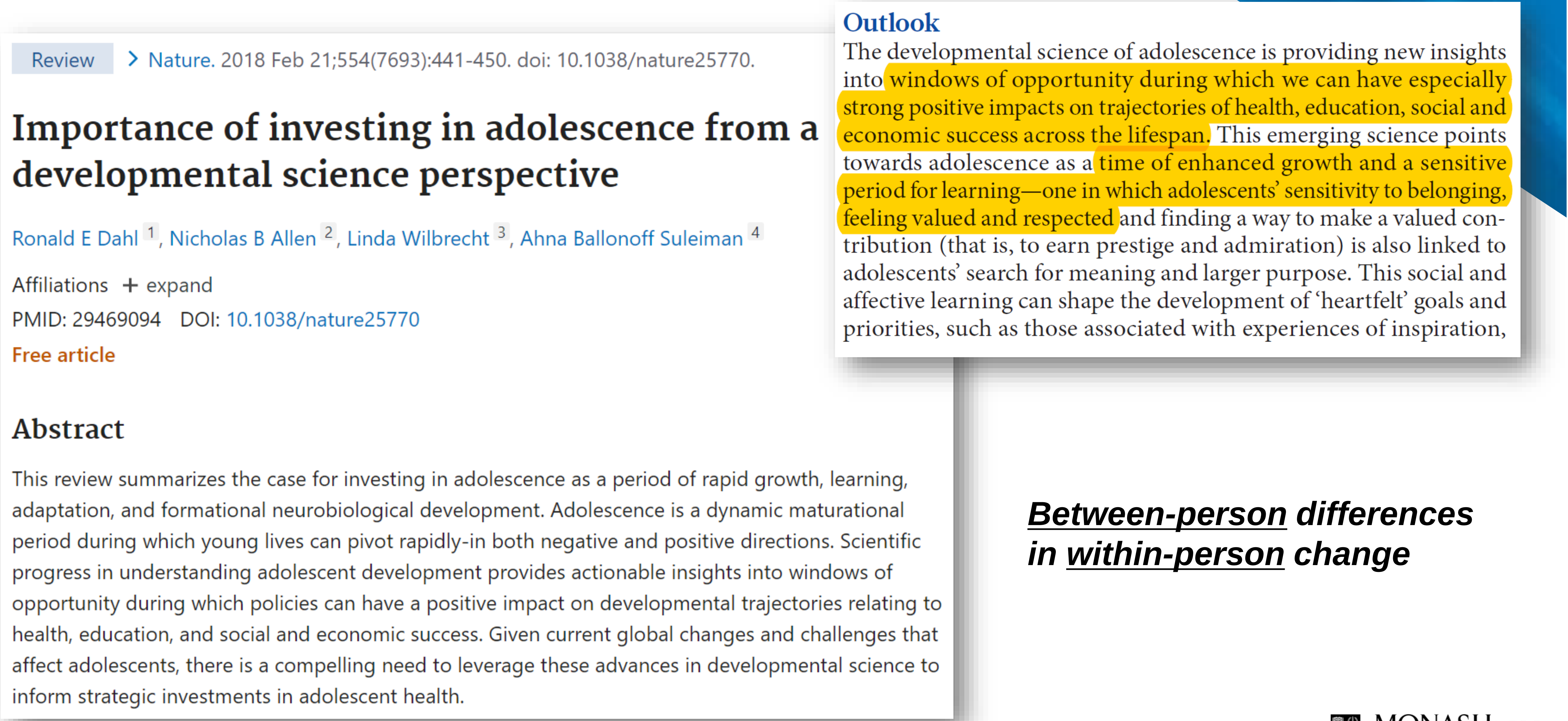
It’s less boring than not changing

Maybe people can get better (*intervention research*)

Maybe our understanding of how things work is not the same across the lifespan (*developmental science*)

Maybe our opportunity to best implement change depends on sensitive periods

Inter-individual vs. intra-individual



Review > Nature. 2018 Feb 21;554(7693):441-450. doi: 10.1038/nature25770.

Importance of investing in adolescence from a developmental science perspective

Ronald E Dahl [1], Nicholas B Allen [2], Linda Wilbrecht [3], Ahna Ballonoff Suleiman [4]

Affiliations + expand

PMID: 29469094 DOI: 10.1038/nature25770

Free article

Abstract

This review summarizes the case for investing in adolescence as a period of rapid growth, learning, adaptation, and formational neurobiological development. Adolescence is a dynamic maturational period during which young lives can pivot rapidly-in both negative and positive directions. Scientific progress in understanding adolescent development provides actionable insights into windows of opportunity during which policies can have a positive impact on developmental trajectories relating to health, education, and social and economic success. Given current global changes and challenges that affect adolescents, there is a compelling need to leverage these advances in developmental science to inform strategic investments in adolescent health.

Outlook

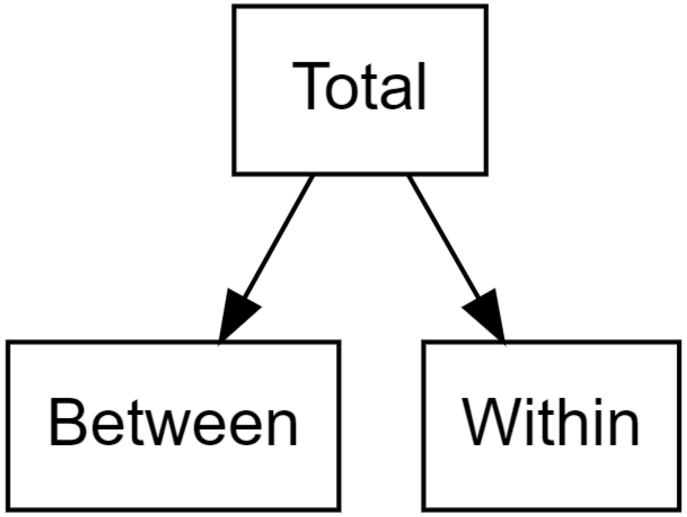
The developmental science of adolescence is providing new insights into windows of opportunity during which we can have especially strong positive impacts on trajectories of health, education, social and economic success across the lifespan. This emerging science points towards adolescence as a time of enhanced growth and a sensitive period for learning—one in which adolescents' sensitivity to belonging, feeling valued and respected and finding a way to make a valued contribution (that is, to earn prestige and admiration) is also linked to adolescents' search for meaning and larger purpose. This social and affective learning can shape the development of 'heartfelt' goals and priorities, such as those associated with experiences of inspiration,

https://www.nature.com/articles/nature25770

***Between-person** differences*
***in within-person** change*

MONASH University

Between and Within Effects

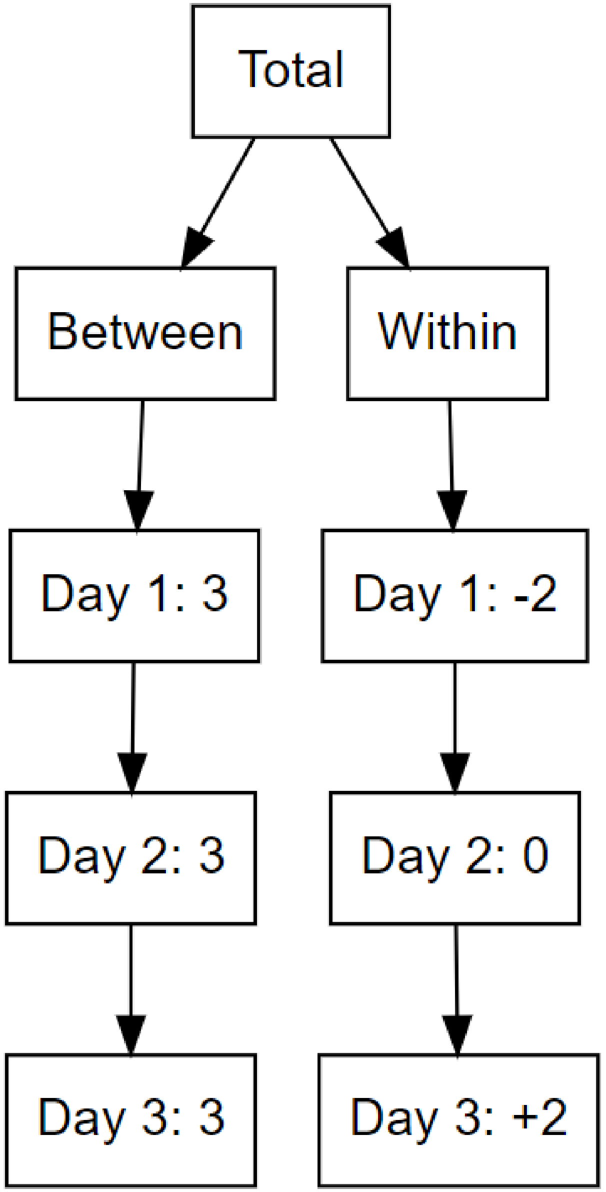


Actual scores:

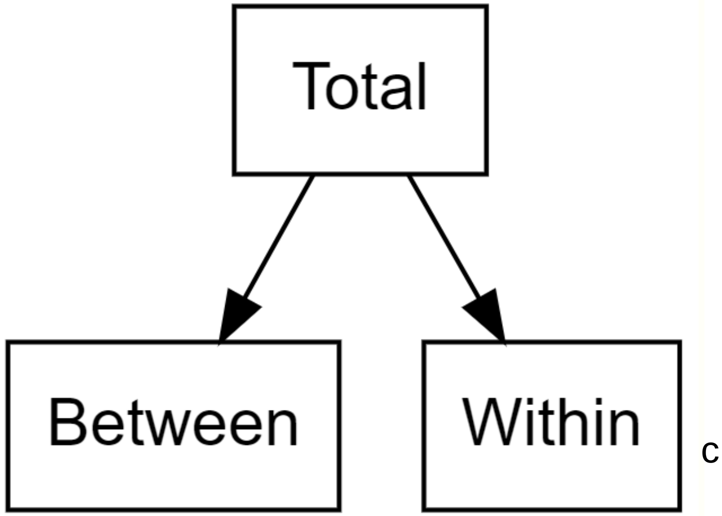
1

3

5



GROWTH CURVE MODELS (time trends, time paths, growth curves, latent trajectories)



(Within-person change across time)

How? Increasing, decreasing, flat, linear, nonlinear?

Why? What are the predictors, outcomes, group differences in trajectories?

Simpler models for change

<u>Repeated measures ANOVA</u>

- Think of group as time point
- Must have even time points per person
- Bad for missing data

<u>Residualized change scores</u>

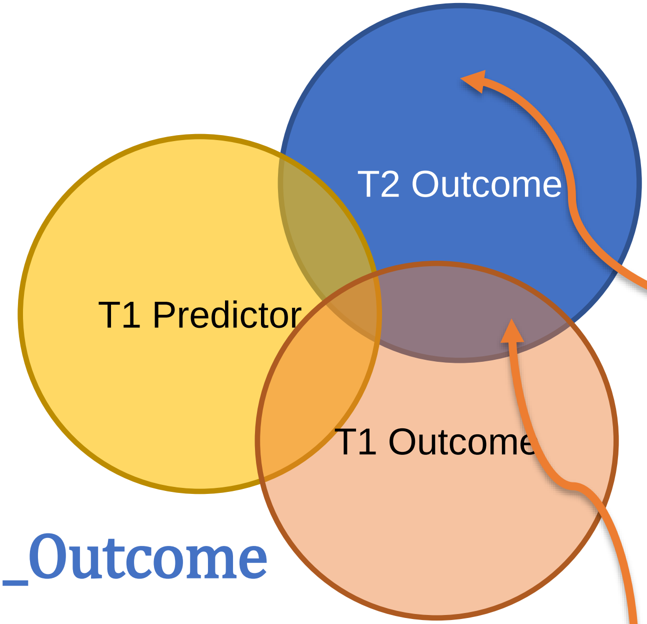
Time2_Outcome ~ T1_Outcome + e

Time2_Outcome ~ Time1_Predictor + Time1_Outcome

<u>Difference scores</u>

$$\underbrace{\text{T2\_Outcome} - \text{T1\_Outcome}}_{\Delta\text{Outcome}} \sim \text{T1 Predictor}$$

- Ok for two time points (& can't model nonlinear)
- Can't include time-varying covariates

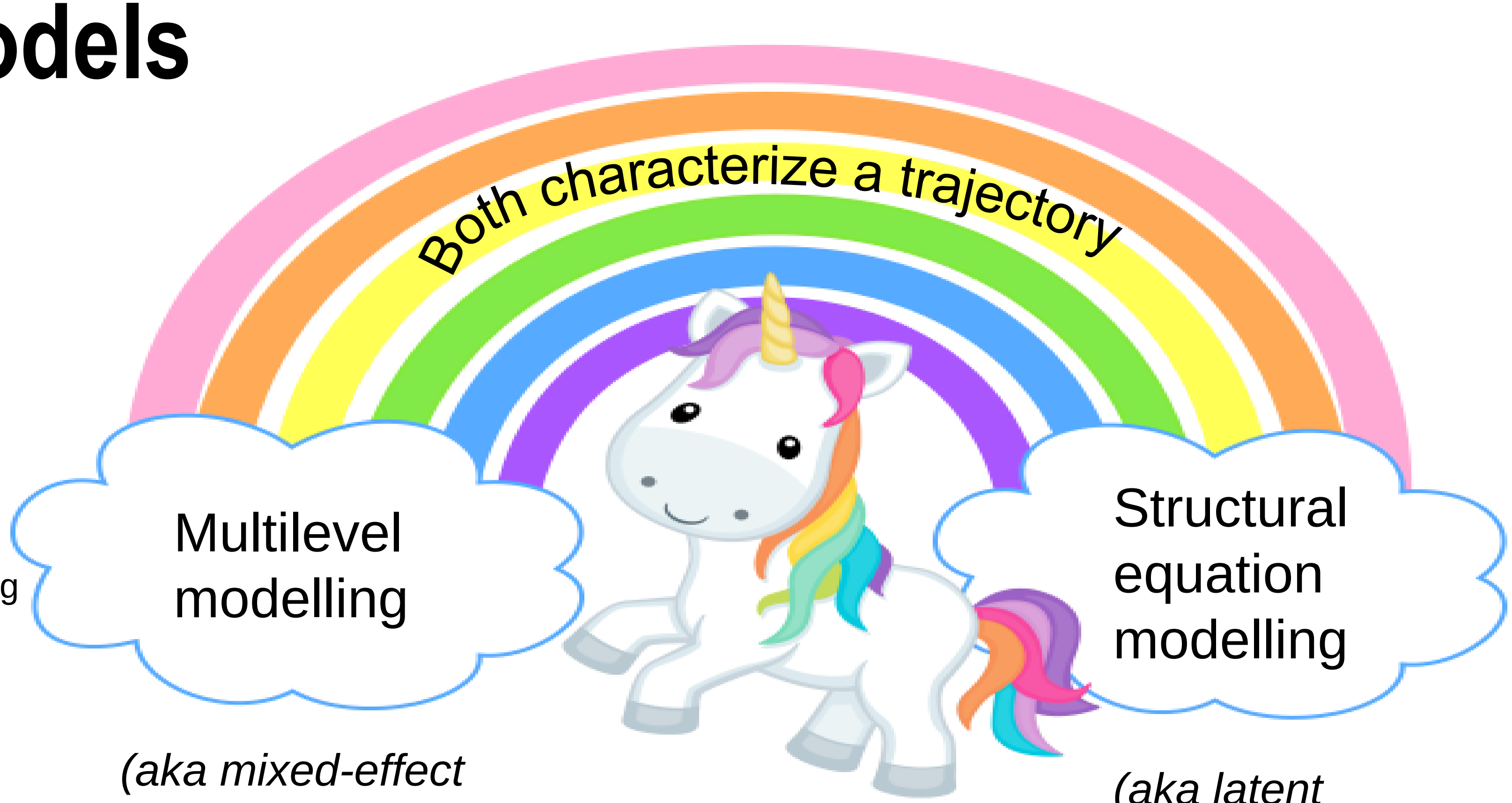


Leftover variance in T2 Outcome not accounted for by T1 Outcome essentially *is* the change

Control for, or take out the variance accounted for by, T1 outcome

Growth models

Options (for now):



Both characterize a trajectory

Bonus: Can add other levels of nesting/clustering

Multilevel modelling

(aka mixed-effect models, or linear mixed models if they’re linear)

Structural equation modelling

(aka latent curve models)

Bonus: Can remove some measurement error

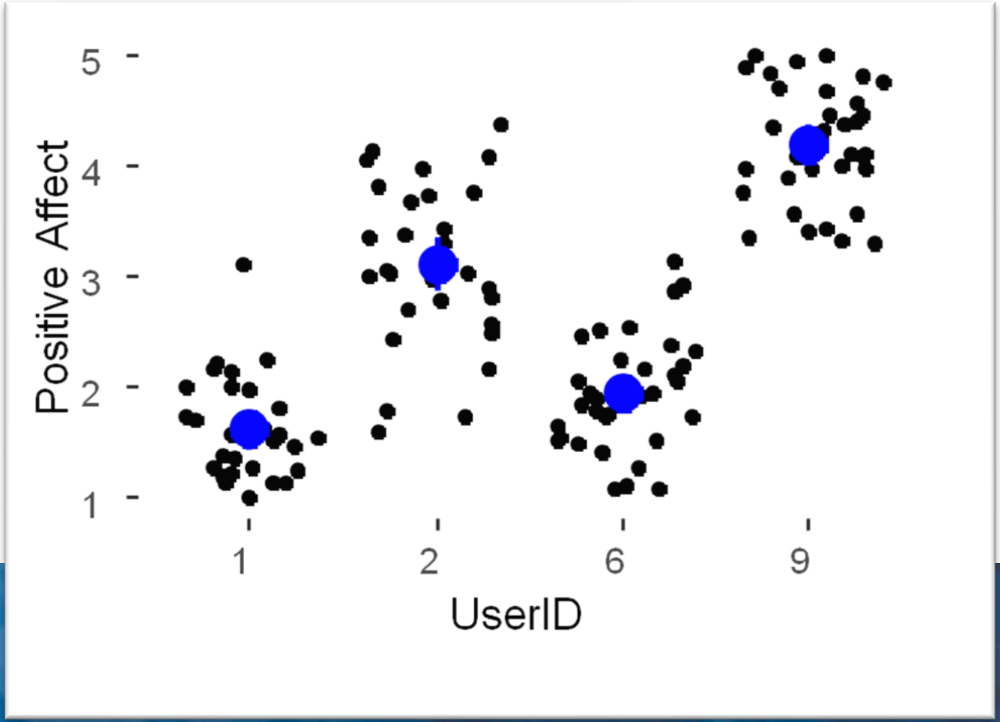
Part 2 – Multilevel and SEM growth models

Assumptions

- ~~**Independence:**~~

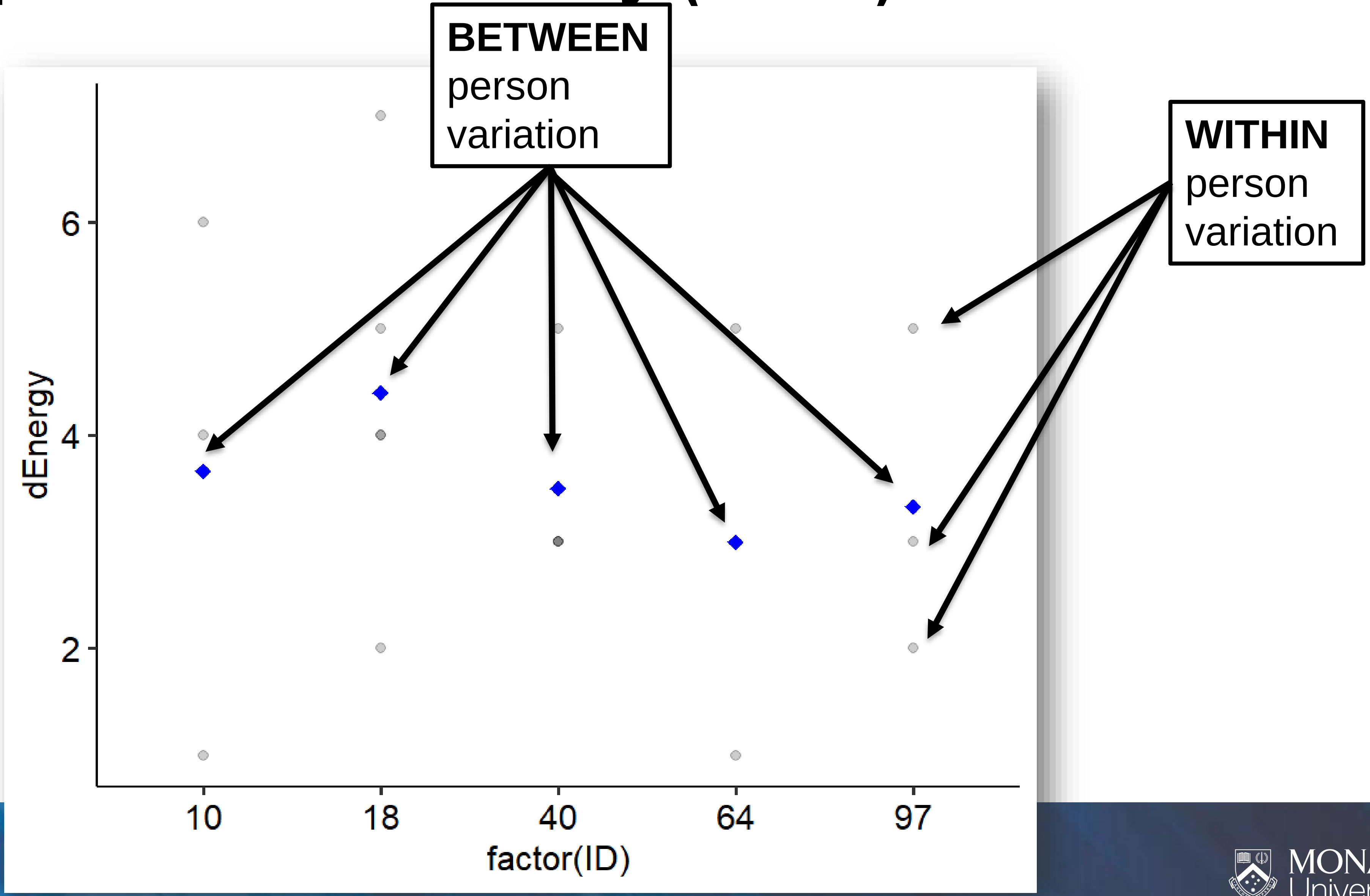
~~All values of the outcome should come from a different person~~

Linear regression **assumes** that observations are independent of each other. However, this is not always the case. **Linear mixed models** are an approach to regression that allows us to relax the assumption that our observations are independent.



Visualize repeated measures by (a few) IDs

(Follow along: 2.1)

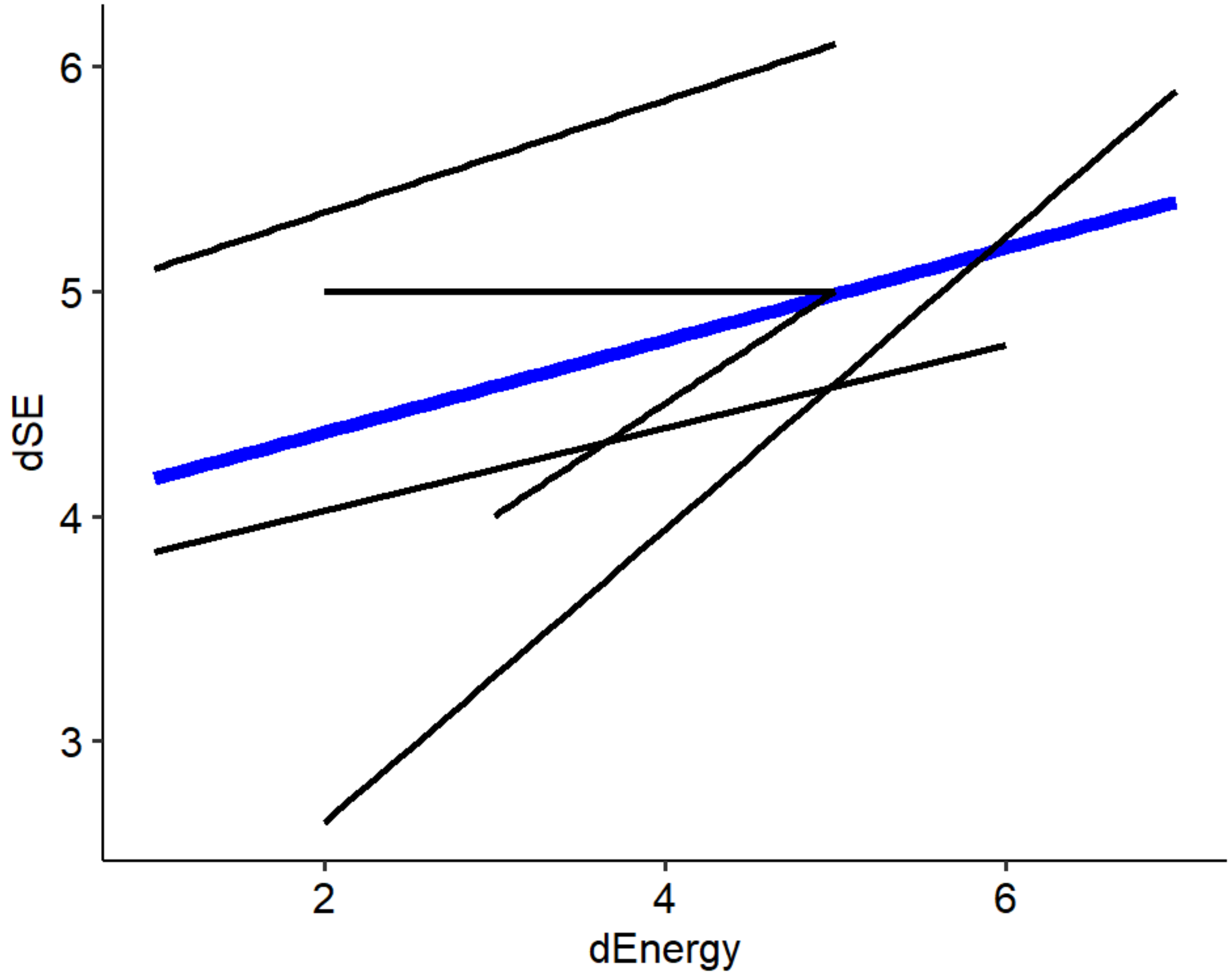


BETWEEN person variation

WITHIN person variation

Single regression line = Fixed effects

- Slope and intercept are the same for everyone
- Fixed effects are assumed to be identical for every participants
- Necessary if only one observation per participant
- Repeated measures can show different coefficients (slopes and intercepts) per participants = random effects



Linear Mixed Effects Models (LMMs) : Random vs Fixed Effects

(Follow along: 2.2)

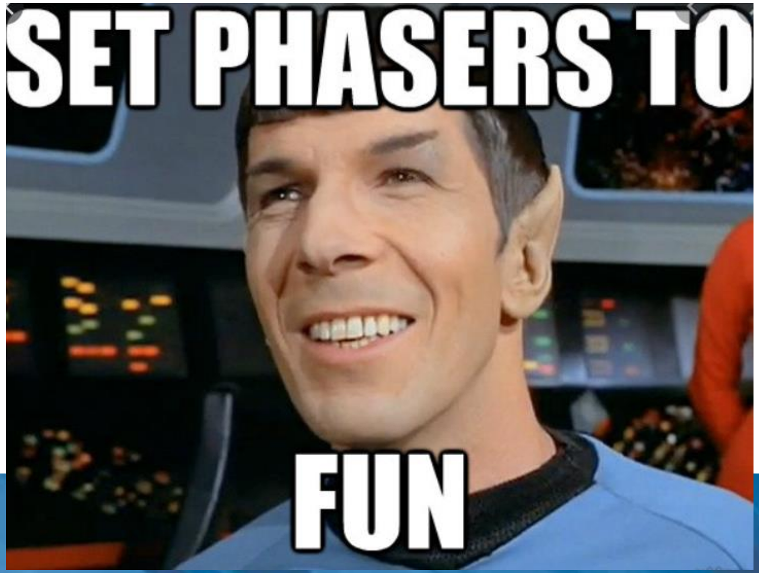
= multilevel models (MLM) or hierarchical linear models (HLM)

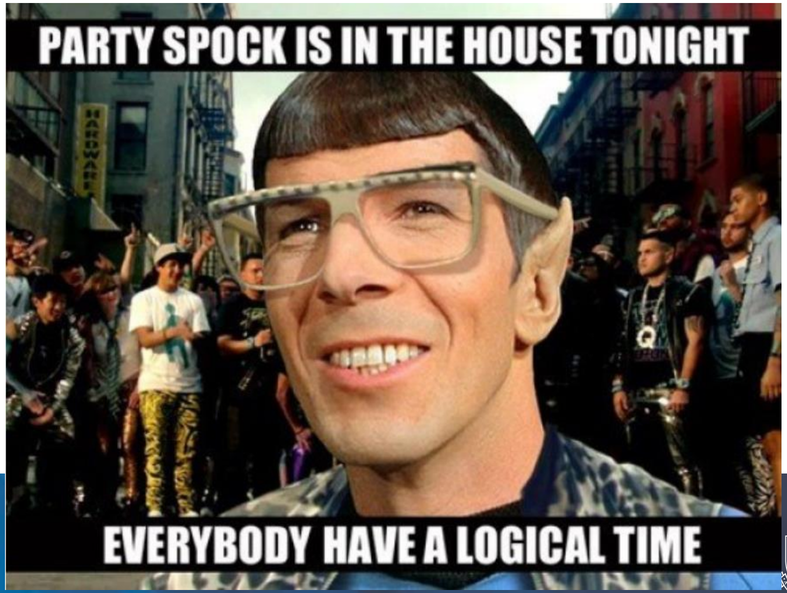
Include:

Fixed effects (regression coefficients identical for everyone)

+

Random effects (regression coefficients that vary randomly for each participant)



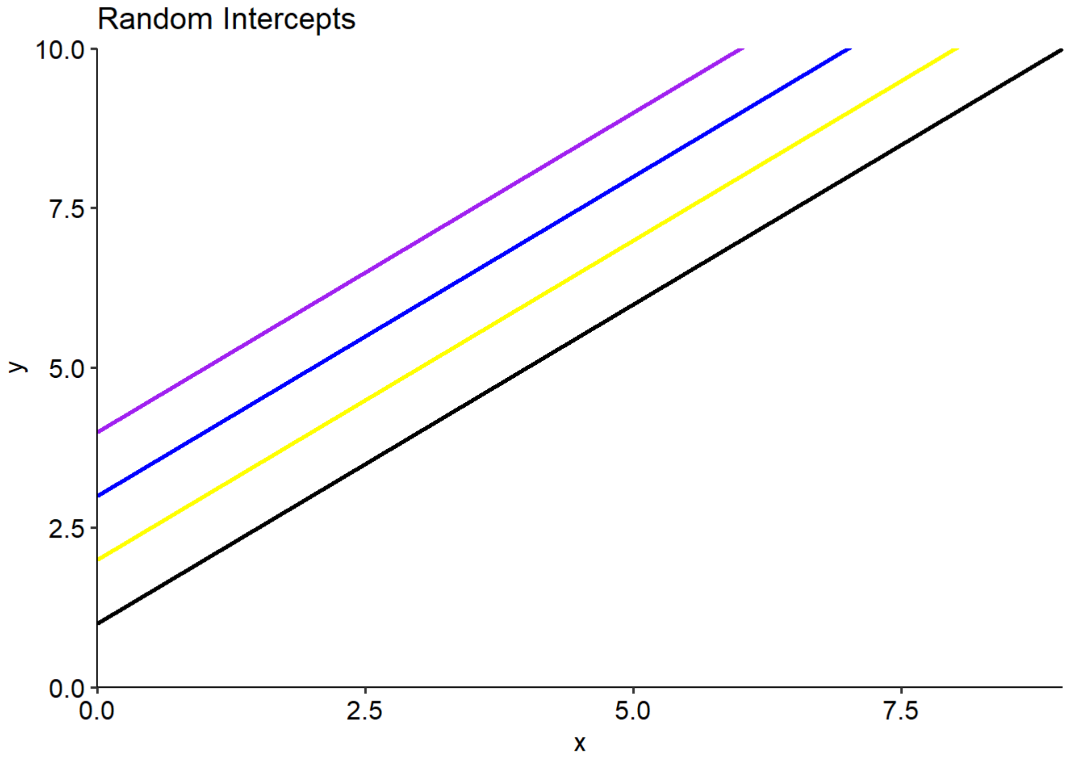

PARTY SPOCK IS IN THE HOUSE TONIGHT
EVERYBODY HAVE A LOGICAL TIME

Random effects

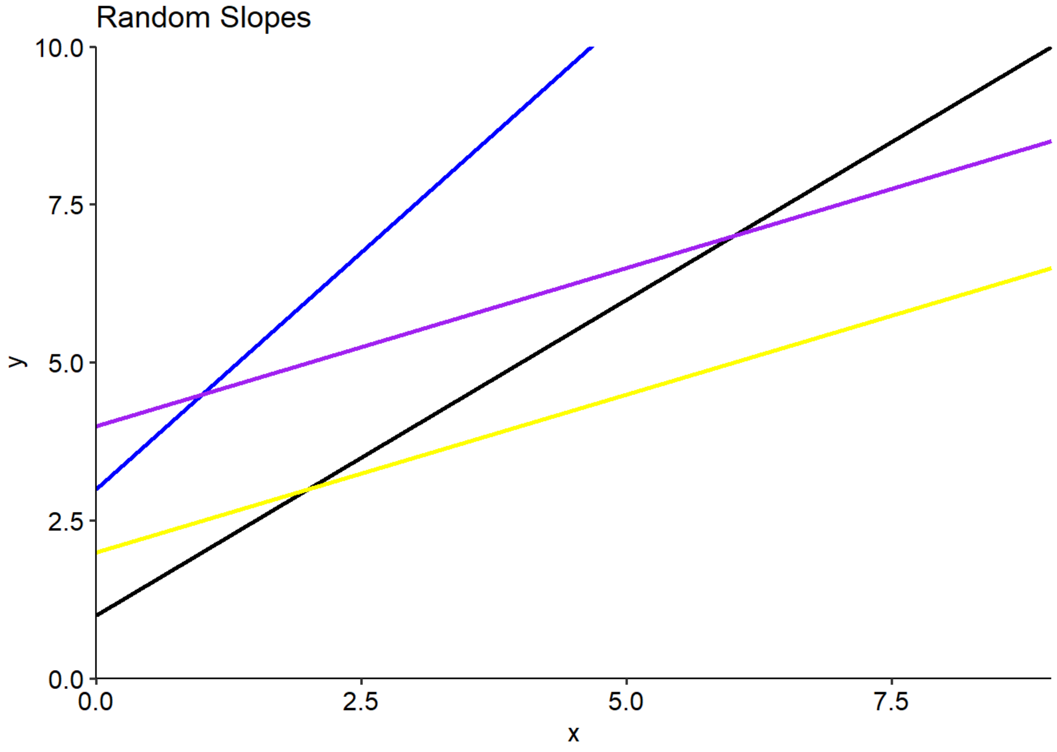
- Linear models have:

❑ An intercept

❑ A slope



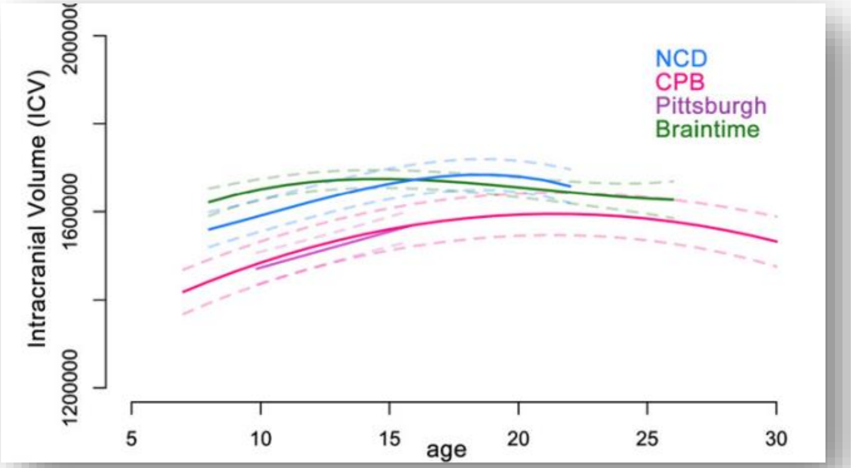
```
m0 <- lmer(y ~ x + (1 | ID), data = dm)
```



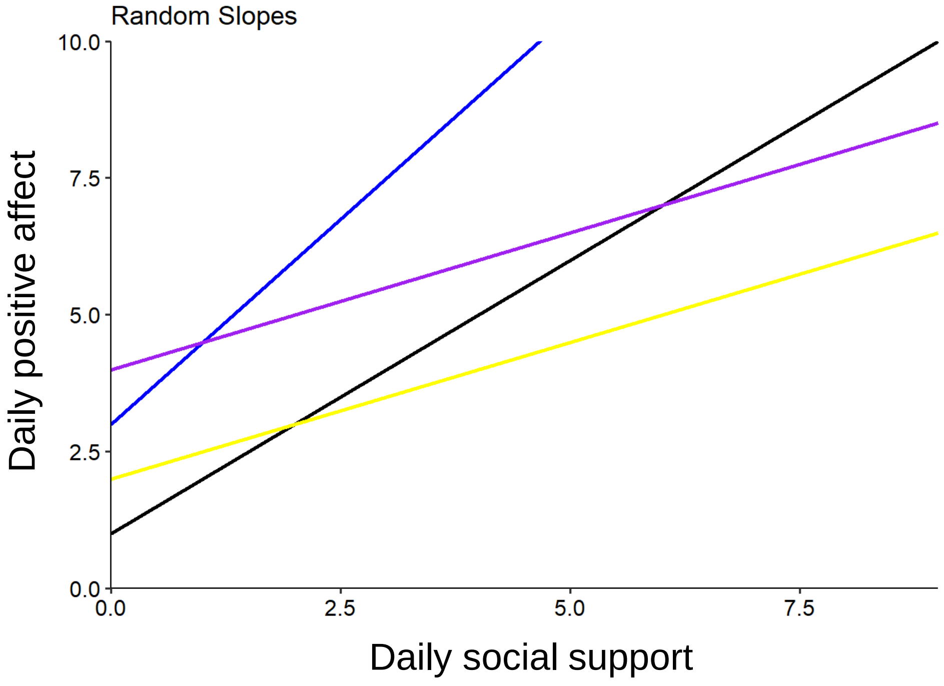
```
m1 <- lmer(y ~ x + (x | ID), data = dm)
```

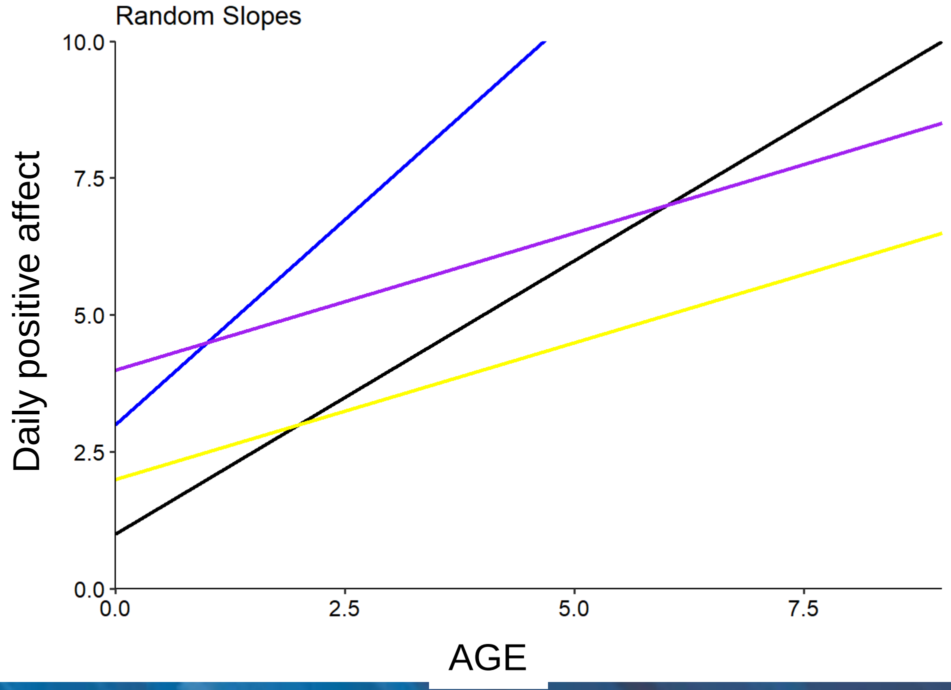
Multilevel growth models

- Have fixed and random effects
- Estimate a trajectory, or slope across some time variable



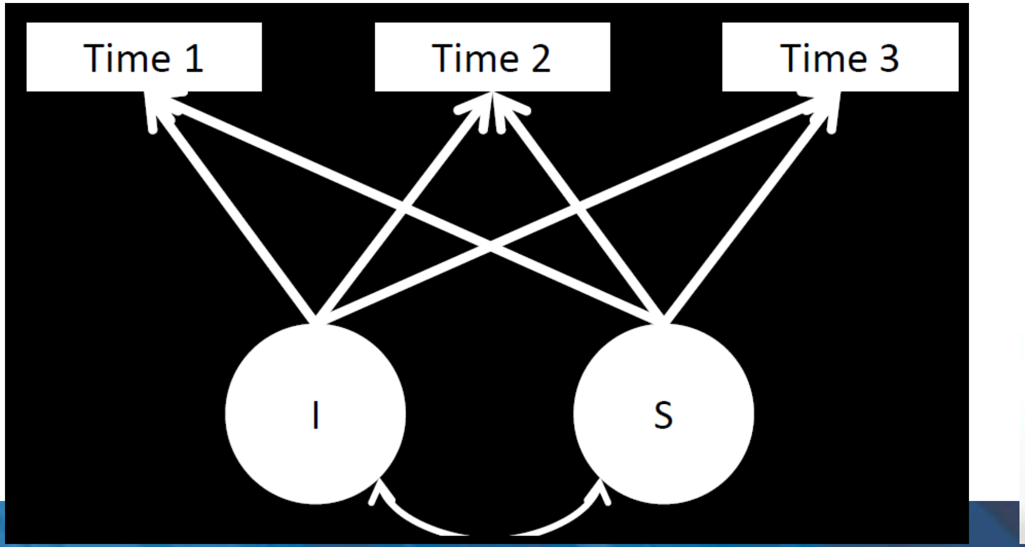
Mills, et al. (2016) NeuroImage 141, 273-81





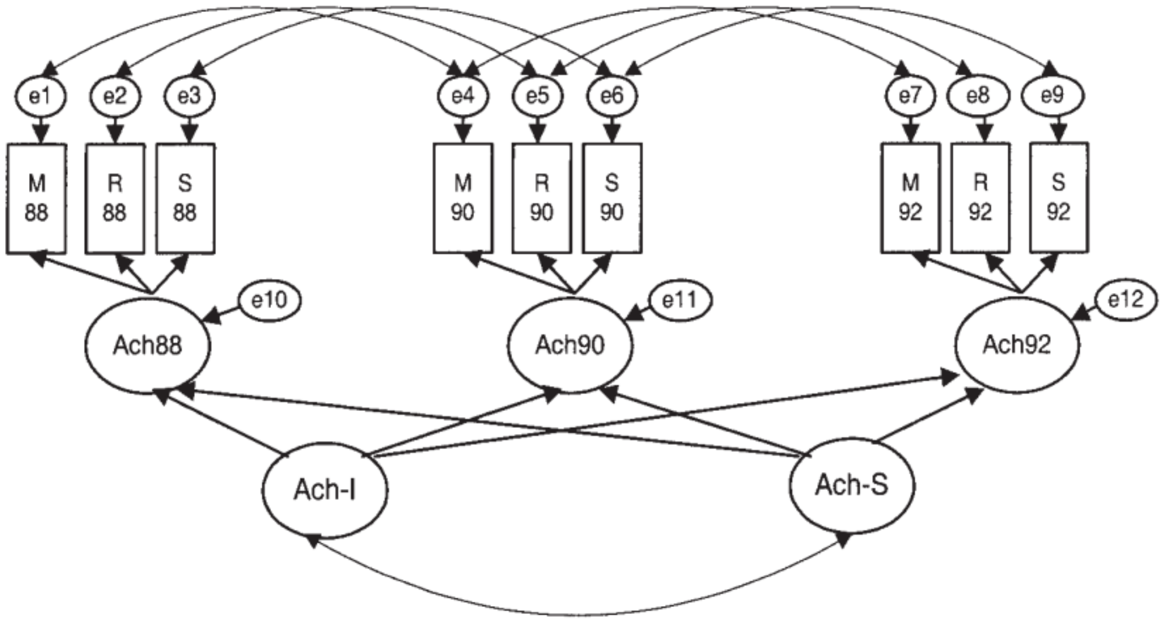
Structural equation models

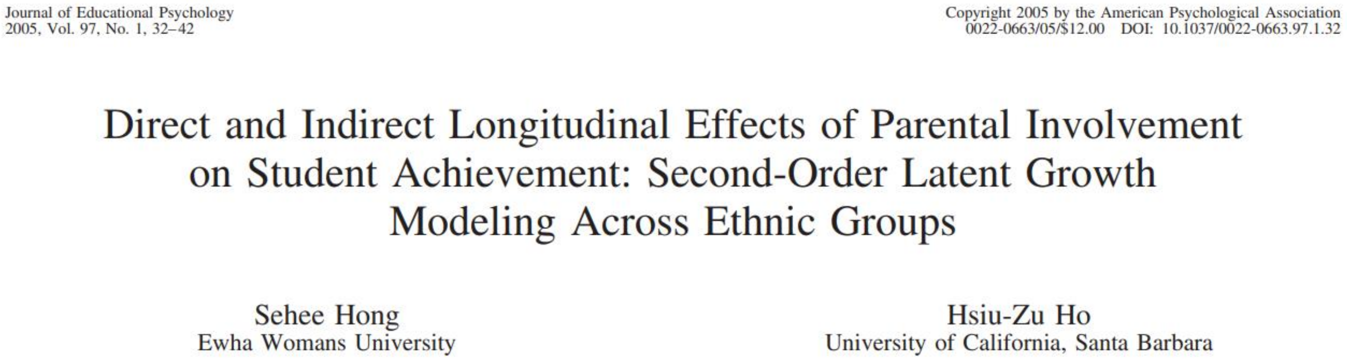
- We still have random effects in latent variable modelling
- They are conceptualized as continuous latent variables, called growth factors
- What is a latent variable? One that is ~~magic~~ not observed, but ~~made up~~ inferred from a model that includes multiple observed “indicators” (for example, a bunch of questions on a questionnaire make up a combined scale).
- Each person still gets their own intercept and slope, just like in MLM
- We just put them in special circles with arrows onto the time points in SEM.



Interpretation and visualization – telling the story

Example 1: What is the initial status and subsequent growth in student achievement, and is this different by ethnic group?




Journal of Educational Psychology
2005, Vol. 97, No. 1, 32–42

0022-0663/05/$12.00 DOI: 10.1037/0022-0663.97.1.32

Direct and Indirect Longitudinal Effects of Parental Involvement on Student Achievement: Second-Order Latent Growth Modeling Across Ethnic Groups

Sehee Hong
Ewha Womans University

Hsiu-Zu Ho
University of California, Santa Barbara

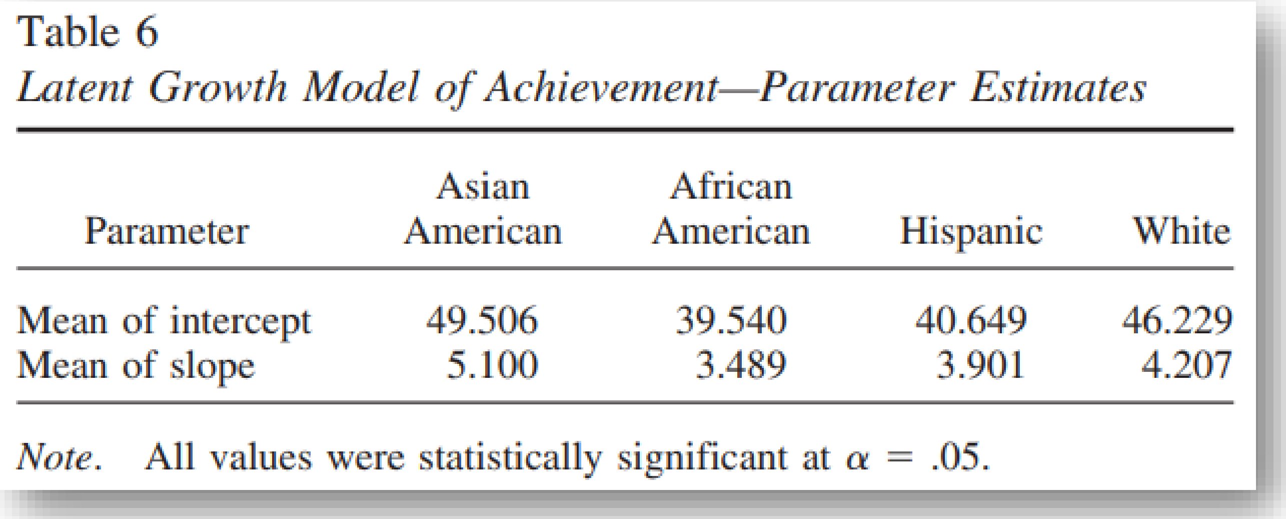


Table 6
Latent Growth Model of Achievement—Parameter Estimates

Parameter	Asian American	African American	Hispanic	White
Mean of intercept	49.506	39.540	40.649	46.229
Mean of slope	5.100	3.489	3.901	4.207

Note. All values were statistically significant at $\alpha = .05$.

What sample size do I need

A very basic rule is never more variables than observations per person.

Sample size for linear mixed models in sjstats (by Daniel Lüdecke): https://strengejacke.github.io/sjstats/reference/samplesize_mixed.html

LGMs with small sample sizes (goes into how different estimators may affect parameters): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7928428/

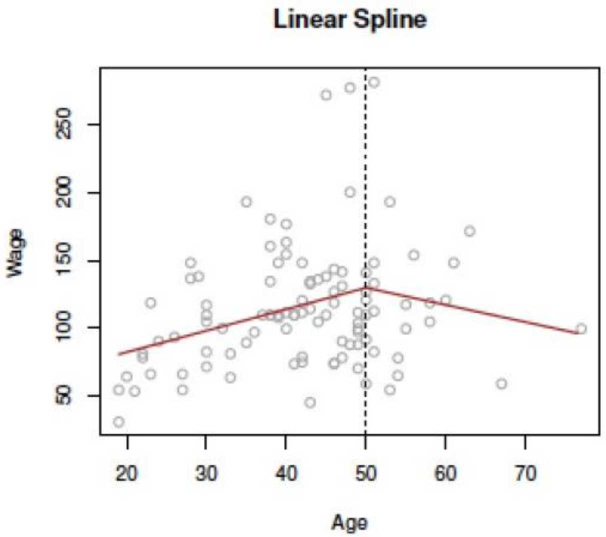
NIH webinar recording: https://prevention.nih.gov/education-training/methods-mind-gap/choosing-sample-sizes-multilevel-and-longitudinal-studies-analyzed-linear-mixed-models

Part 3 – The shape of change

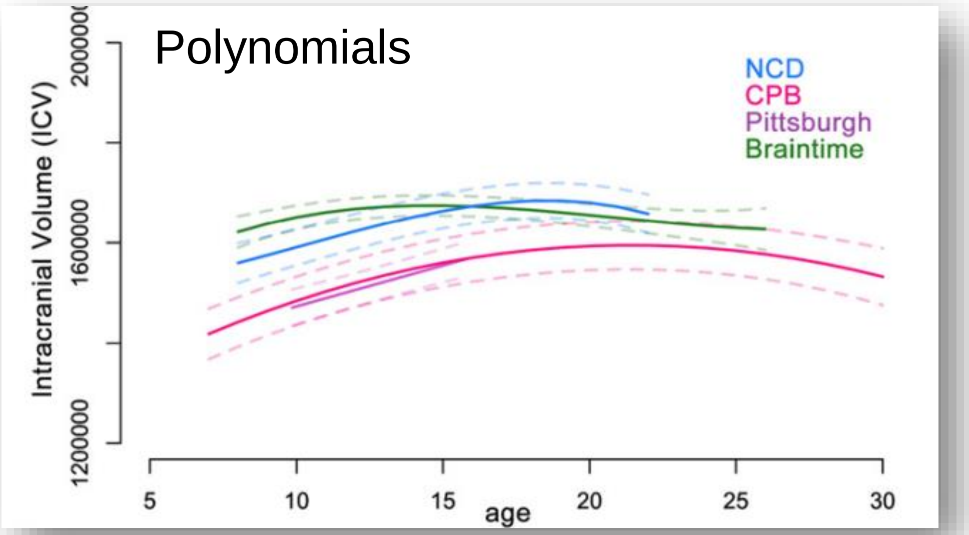
The shape of development

What shape should the growth model be?

- Informed by theory AND visualisation (and fit stats)
- Polynomials: linear, quadratic, cubic
- Piecewise and splines (generalized additive models)

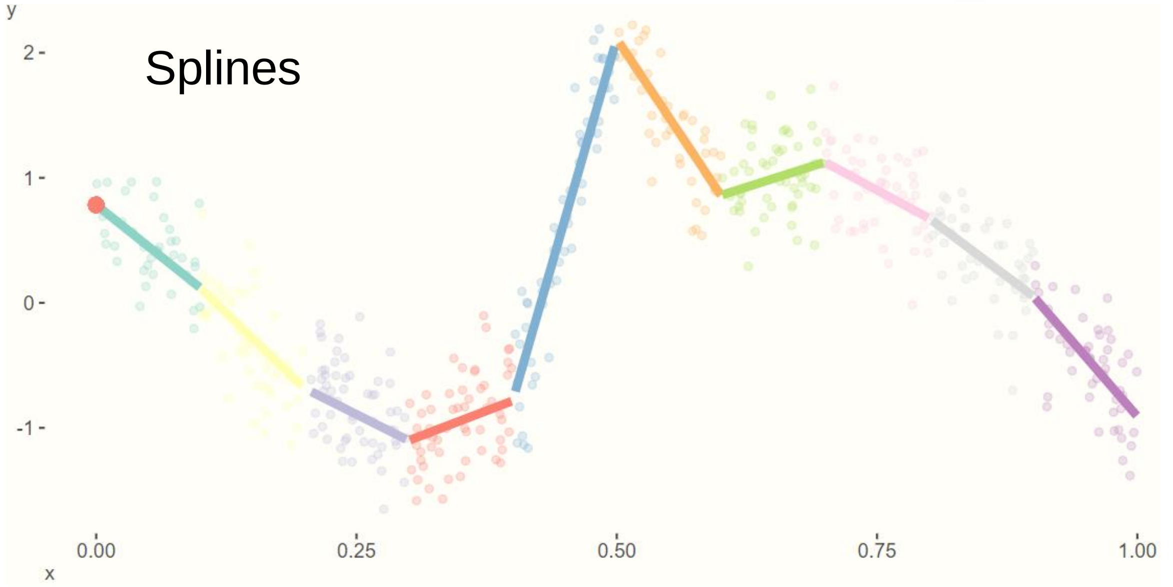


Linear Spline



Polynomials

Mills, et al. (2016) NeuroImage 141, 273-81

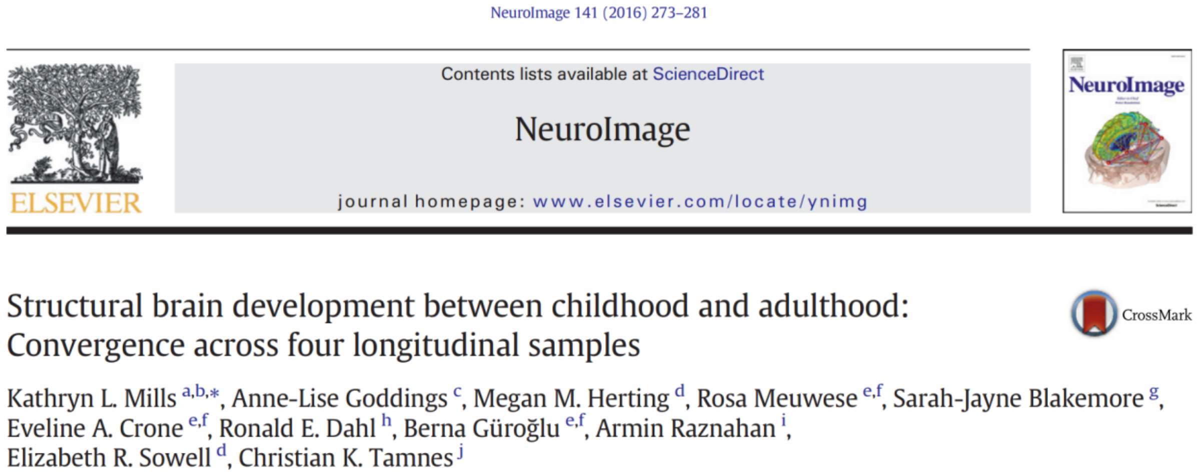


Splines

https://m-clark.github.io/generalized-additive-models/technical.html#a-detailed-example

Focus on polynomials

Example 2: How does cranial volume and brain size (ICV and WBV) change across development?



NeuroImage 141 (2016) 273–281

Contents lists available at ScienceDirect

NeuroImage

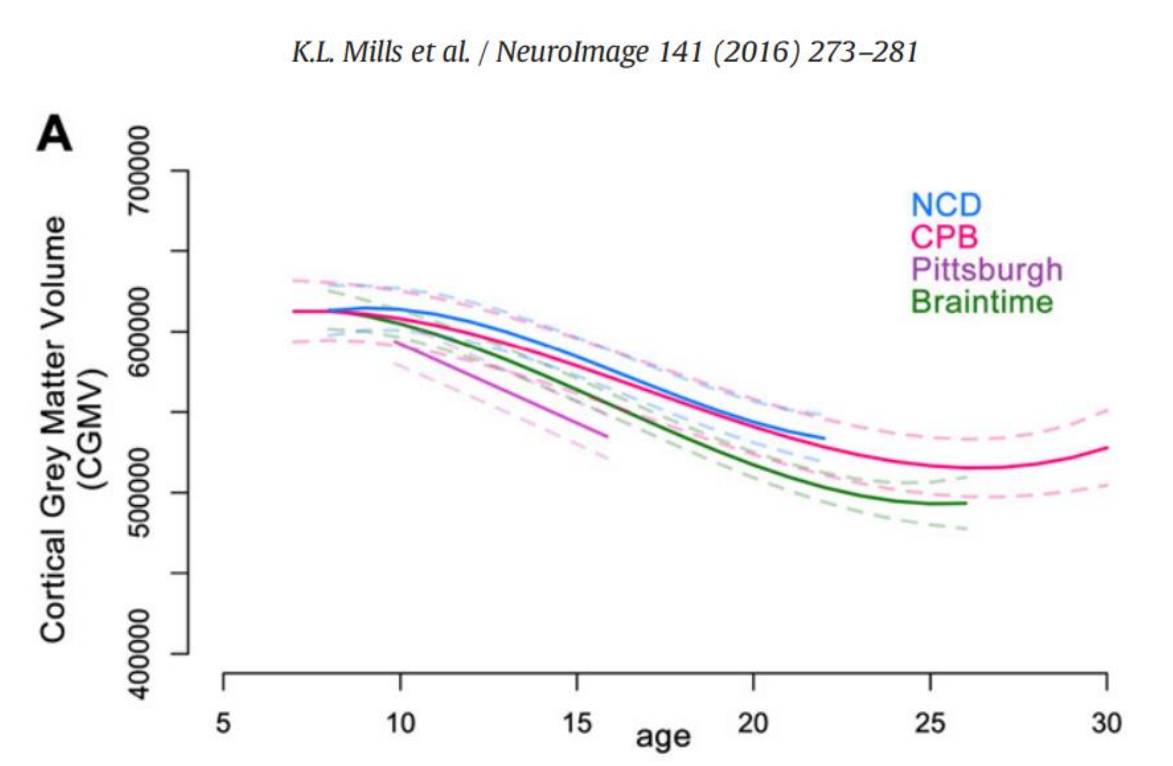
journal homepage: www.elsevier.com/locate/ynimg

ELSEVIER

Structural brain development between childhood and adulthood: Convergence across four longitudinal samples

CrossMark

Kathryn L. Mills [a,b,*], Anne-Lise Goddings [c], Megan M. Herting [d], Rosa Meuwese [e,f], Sarah-Jayne Blakemore [g], Eveline A. Crone [e,f], Ronald E. Dahl [h], Berna Güroğlu [e,f], Armin Raznahan [i], Elizabeth R. Sowell [d], Christian K. Tamnes [j]



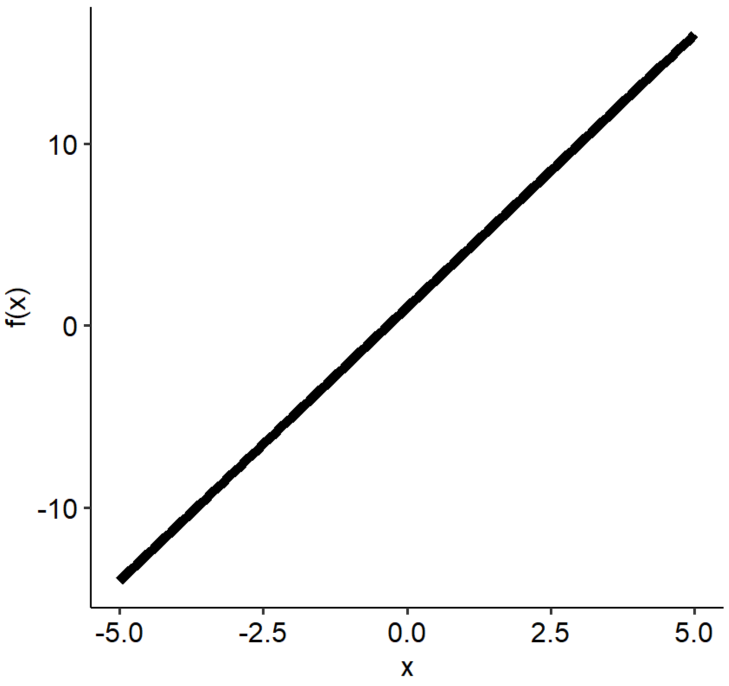
K.L. Mills et al. / NeuroImage 141 (2016) 273–281

1. Linear model: Volume $=$ Intercept $+ \alpha(\text{age})$
2. Quadratic model: Volume $=$ Intercept $+ \alpha(\text{age}) + \beta(\text{age}^2)$
3. Cubic model: Volume $=$ Intercept $+ \alpha(\text{age}) + \beta(\text{age}^2) + \gamma(\text{age}^3)$.

Mills, et al. (2016) NeuroImage 141, 273-81

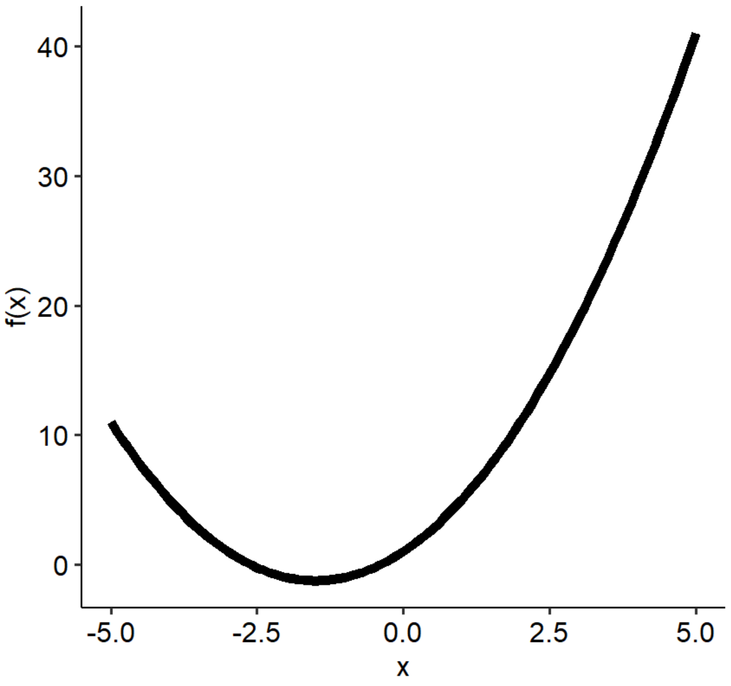
Follow along: 6.4

Polynomials



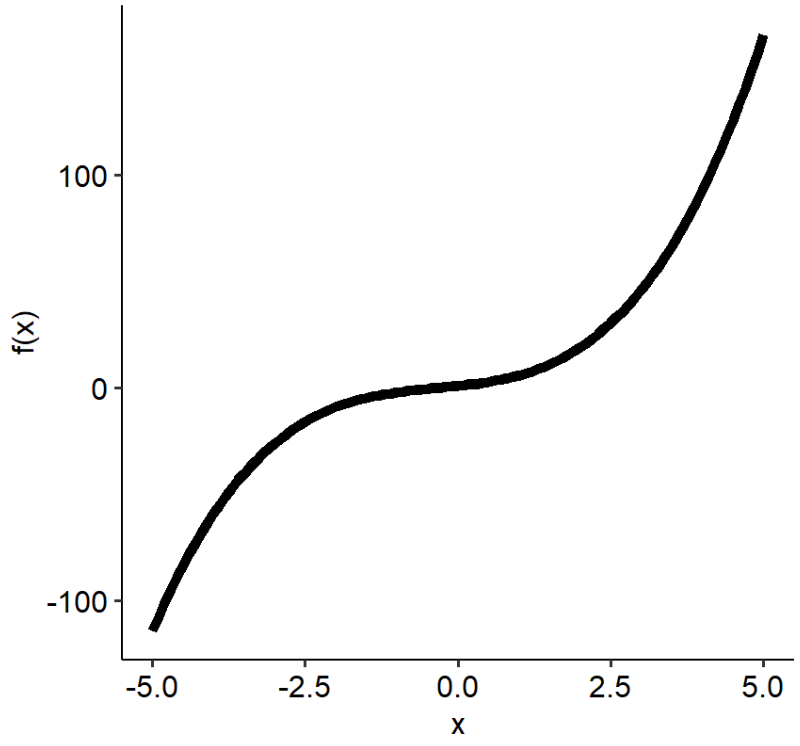
Degree 1

x



Degree 2

x^2



Degree 3

x^3

Follow along: 6.4

Polynomial random intercept models

```
m0 <- lmer(Mood ~ 1 + (1 | ID),
             data = dm, REML = FALSE)
m1 <- lmer(Mood ~ poly(wave, 1) + (1 | ID),
             data = dm, REML = FALSE)
m2 <- lmer(Mood ~ poly(wave, 2) + (1 | ID),
             data = dm, REML = FALSE)
m3 <- lmer(Mood ~ poly(wave, 3) + (1 | ID),
             data = dm, REML = FALSE)
```

Part 4 – Code longitudinal change

Codebook

https://e-m-mccormick.github.io/static/longitudinal-primer/

1. Install packages: "nlme", "lme4", "lmerTest", and "lavaan“

(full packages for everything in the codebook:

```
install.packages(c("nlme", "lme4", "lmerTest", "tidyr", "dplyr",
"sjPlot", "ggplot2", "visreg",
"lavaan","semPlots","broom","kableExtra"))
```

2. Download datasets (we’ll mostly work with the “Executive Function” dataset:

https://e-m-mccormick.github.io/static/longitudinal-primer/07-datasets.html

(set your working directory to where you download these)

3. Go to: https://e-m-mccormick.github.io/static/longitudinal-primer/02-canonical.html

- We will start with the multilevel model section

Extra Stuff

Further analysis – Predictors and Outcomes

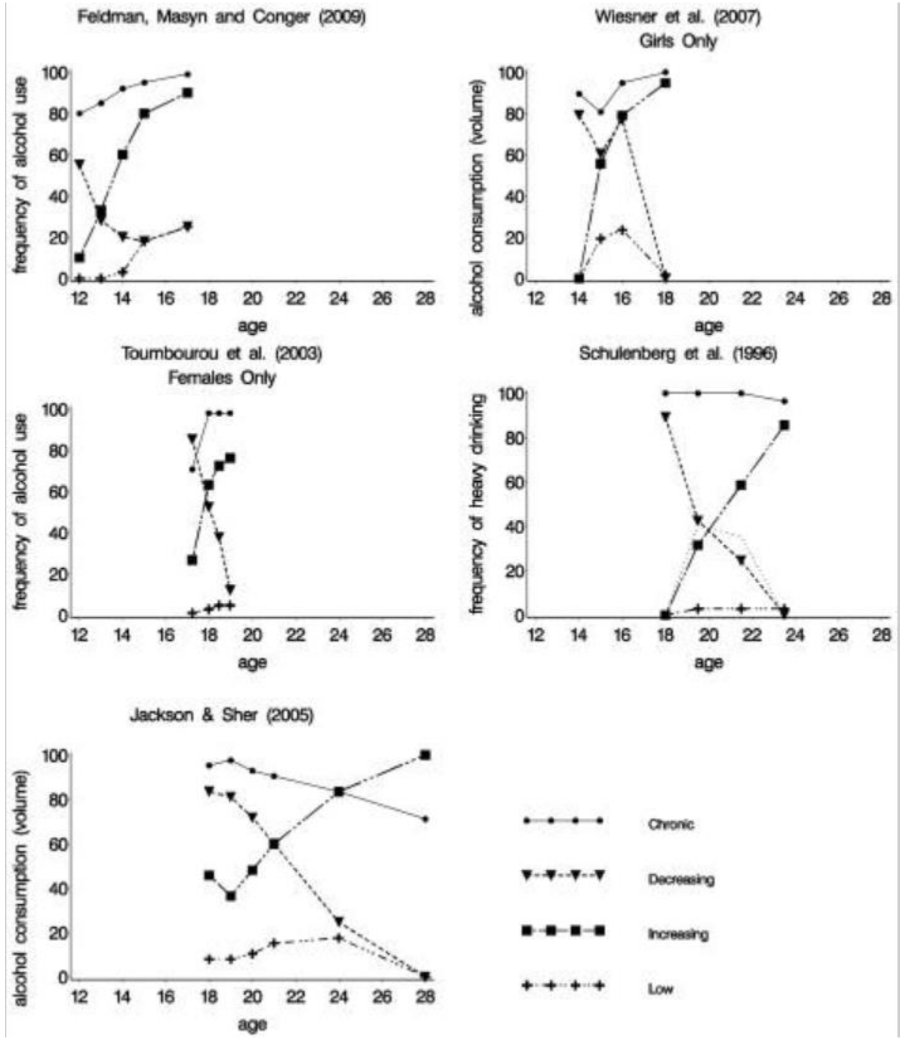
Growth mixture models

- combination of latent class analysis (LCA) and growth model – also called latent class growth analysis*
- Also reduces variance when grouping people (ok if clinically or theoretically justified)

Instead, keep the variance from everyone's individual slope or intercept (latent or random)

Latent profile analysis (LPA). Also creates/generates profiles.

Latent transition analysis (LTA). Another longitudinal version of LCA – but looks to see if group membership changes over time.

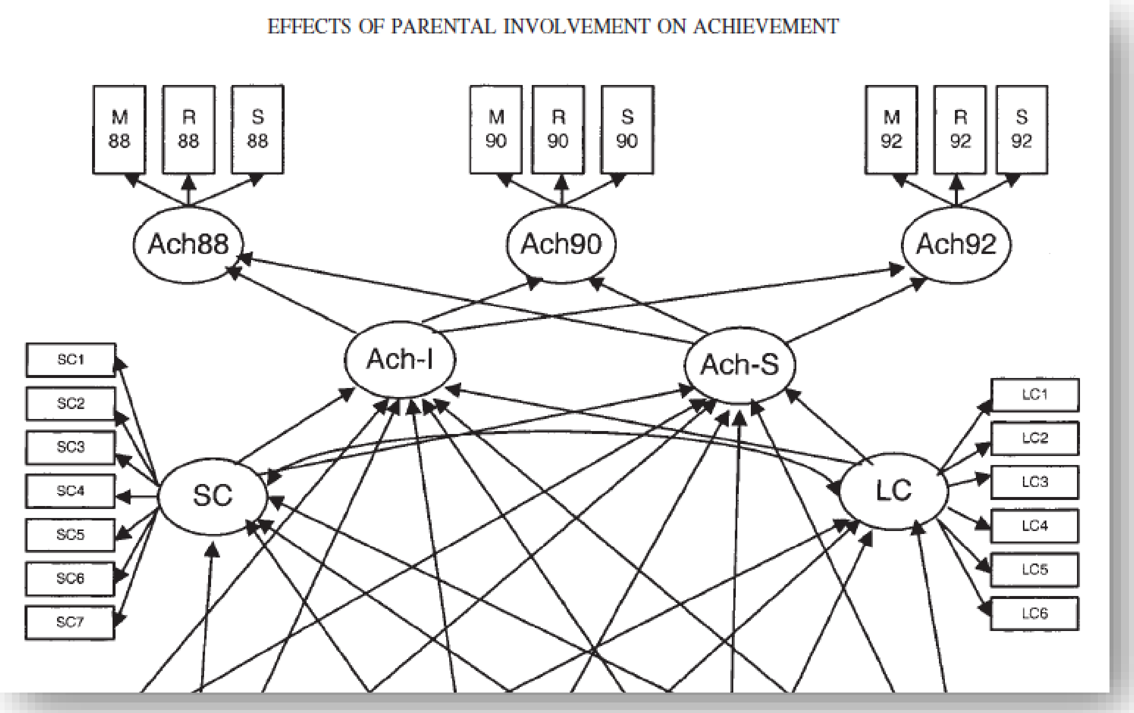


From Sher, et al. (2011), doi: 10.1037/a0021813

* *For differences between these and another one, longitudinal latent class analysis (LLCA), see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2791967*

Further analysis – Predictors and Outcomes

Warning: Using latent or random effects to subsequently predict other observed variables within the same sample is likely overfitting.
Solution - training and holdout samples, or replication studies. Or include the predictors within the model



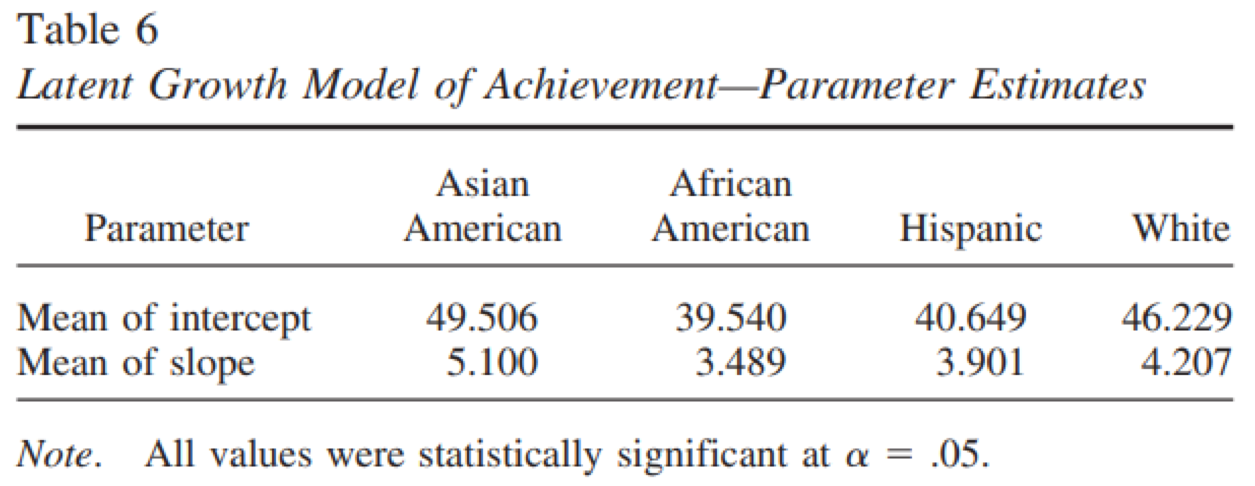


Table 6
Latent Growth Model of Achievement—Parameter Estimates

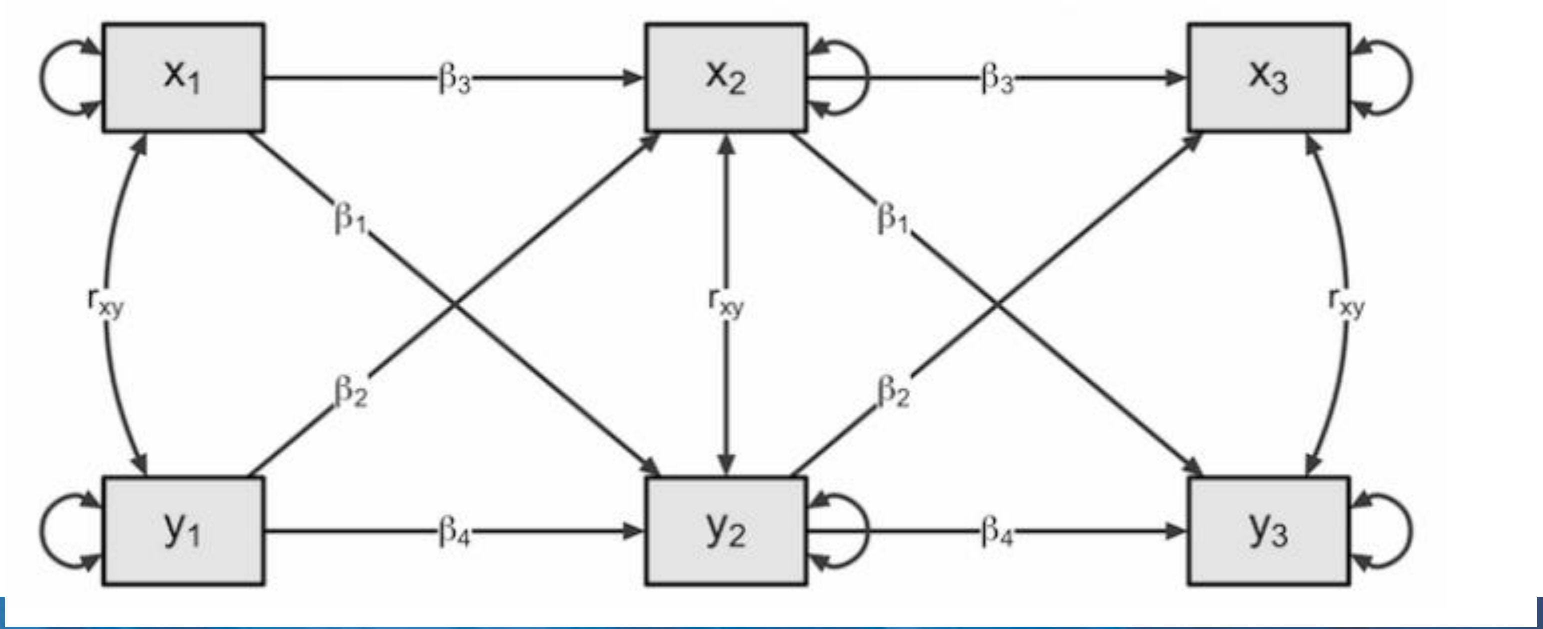
Parameter	Asian American	African American	Hispanic	White
Mean of intercept	49.506	39.540	40.649	46.229
Mean of slope	5.100	3.489	3.901	4.207

Note. All values were statistically significant at $\alpha = .05$.

Cross lagged panel model

- An SEM approach that accounts for bidirectional or reciprocal relationships between variables measured longitudinally

- Essentially it's like looking at path models in more than one direction at once
- You can technically do this with just two time points (but it's barely identified)*.
- Important assumptions – see : https://mikewk.com/clpm.pdf



** And with a RI-CLPM which is the best practice now, you must have at least 3 time points*

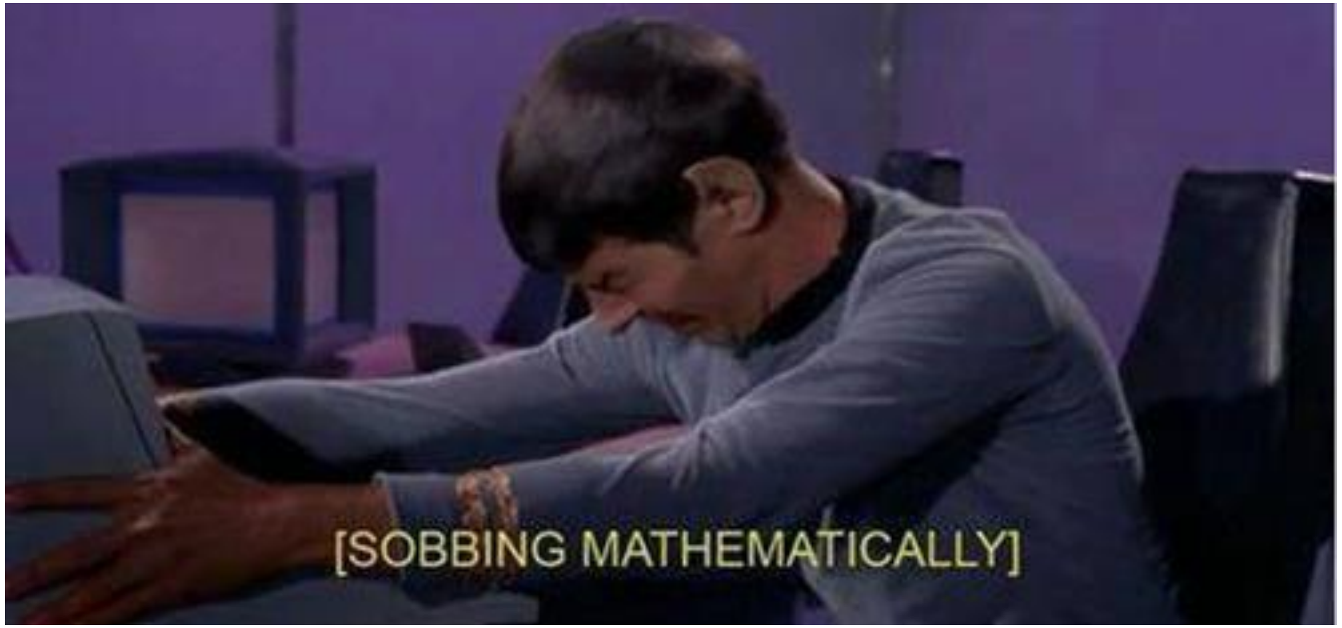
Extra – Missing Data

MONASH
University

Missing Data Makes Me Sad

If we only use complete cases (i.e., listwise deletion):

1. Missing data cause a loss of efficiency and makes everyone sad
2. Results from the non-missing data may be biased and that’s a waste of time and also sad



Why it matters

T.H. Matta, et al. Developmental Cognitive Neuroscience 33 (2018) 83–98

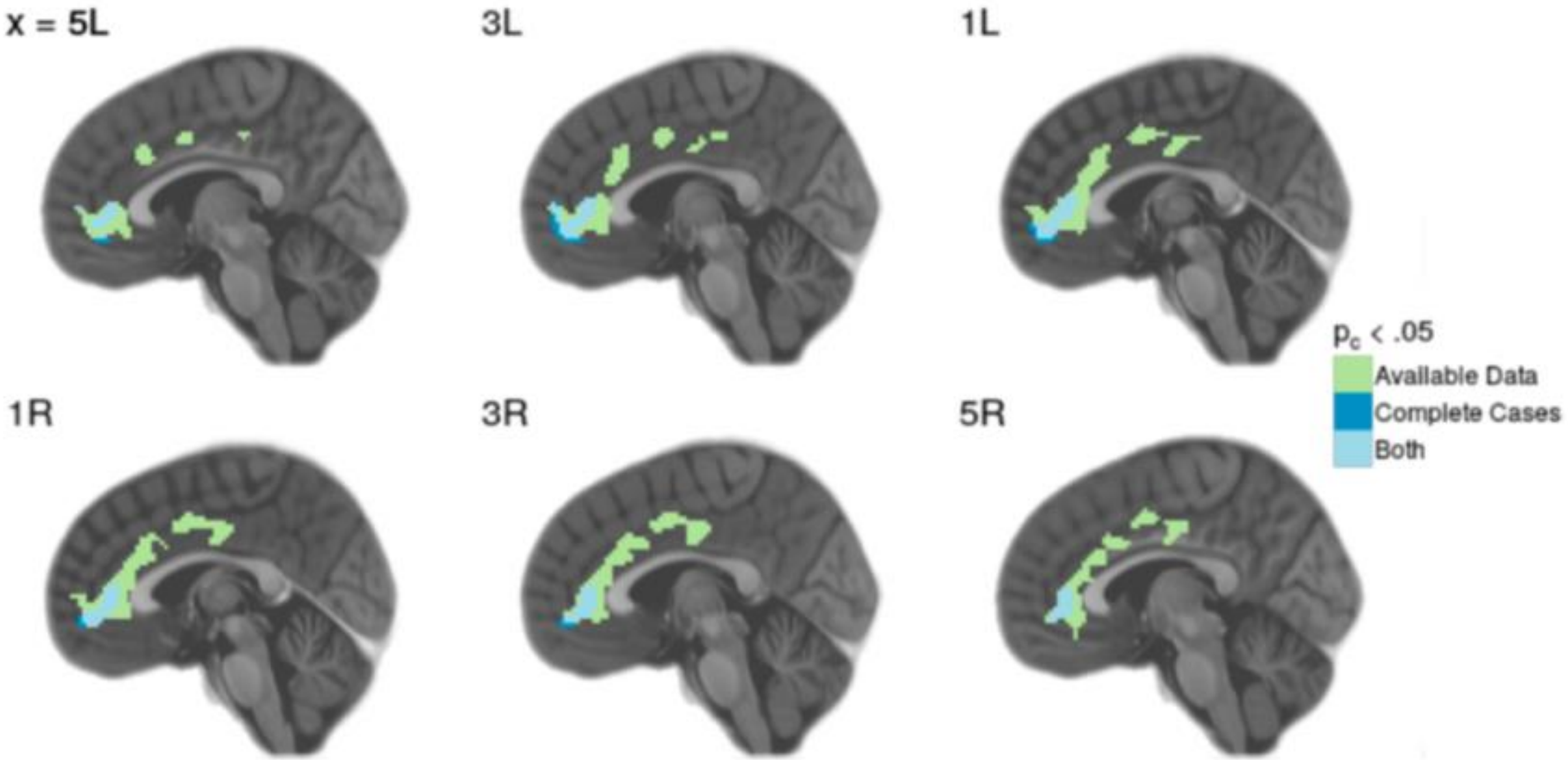
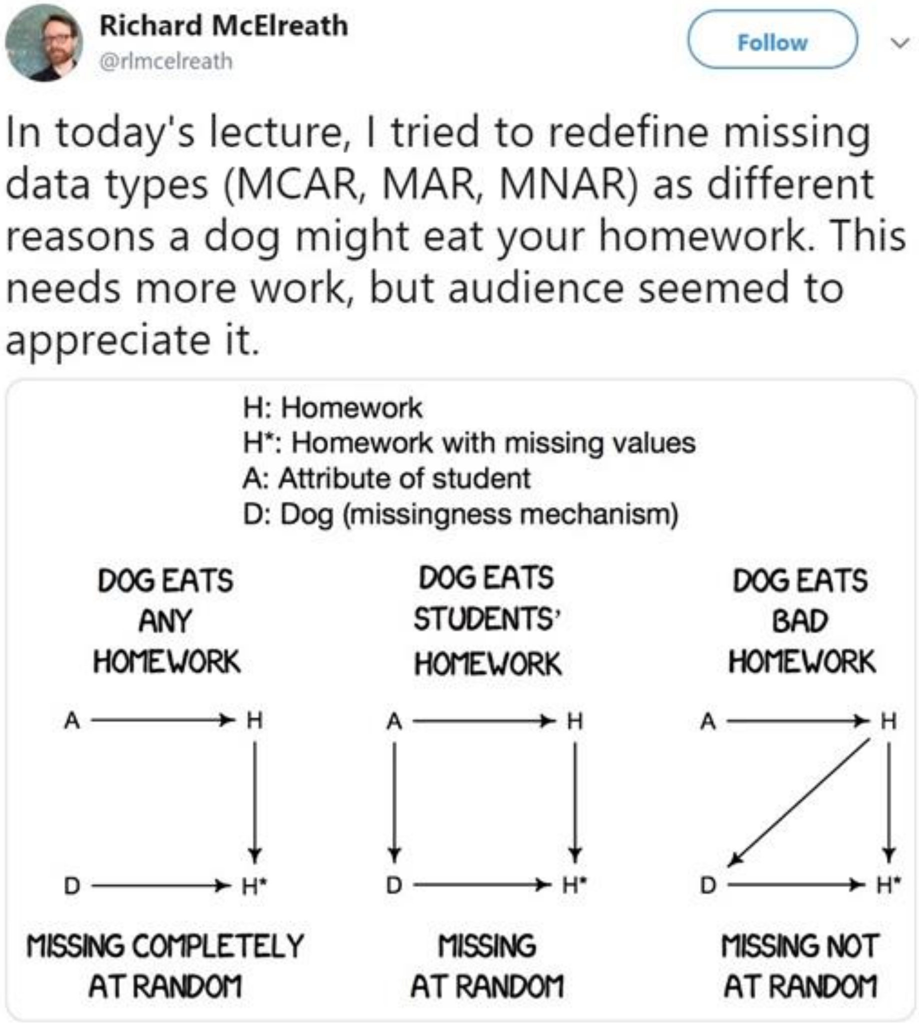


Fig. 3. Significant clusters identified in both available data and complete case analysis is indicated in blue, while significant clusters identified in the available data analysis only are indicated in green. Slice labels indicate the MNI coordinate along which the slice was acquired. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Missing data type	**Assumptions**	**Conditions for unbiasedness**
Missing completely at random (MCAR): missingness is entirely independent of our outcome of interest (*y*) and any covariates *y* depends on	missingness is (1) not dependent on observations of Y that we don't have (2) not dependent on observations of Y that we do have, and (3) not dependent on covariates X, that Y is dependent on.	Unbiased results using complete case analysis, or all available data in maximum likelihood, multiple imputation.
Covariate-dependent MCAR: missingness is only dependent on the values of variables that affect *y*	(1) and (2); dropping assumption 3.	Unbiased if covariate is included in the models for maximum likelihood analysis or imputation.
Missing at random (MAR): missingness (for a particular participant or unit) is independent of the unobserved values of *y,* i.e., depends only on the values of *y* (for that unit) that we were able to collect	(1); dropping assumptions 2 & 3.	Unbiased estimates only if all available data are used in a maximum likelihood or multiple imputation framework.
Missing not at random (MNAR): missingness is dependent on the unobserved values of *y*, i.e., the values of the missing data depend on the outcome values that we were not able to collect.	No assumptions made.	Biased estimates (sensitivity analyses should be performed).
Take-away: It's best to choose an estimation method that allows you to assume the data are MAR and do sensitivity analyses under the assumption that the data are actually MNAR.		

Missing Data Mechanisms



Richard McElreath
@rlmcelreath

In today's lecture, I tried to redefine missing data types (MCAR, MAR, MNAR) as different reasons a dog might eat your homework. This needs more work, but audience seemed to appreciate it.

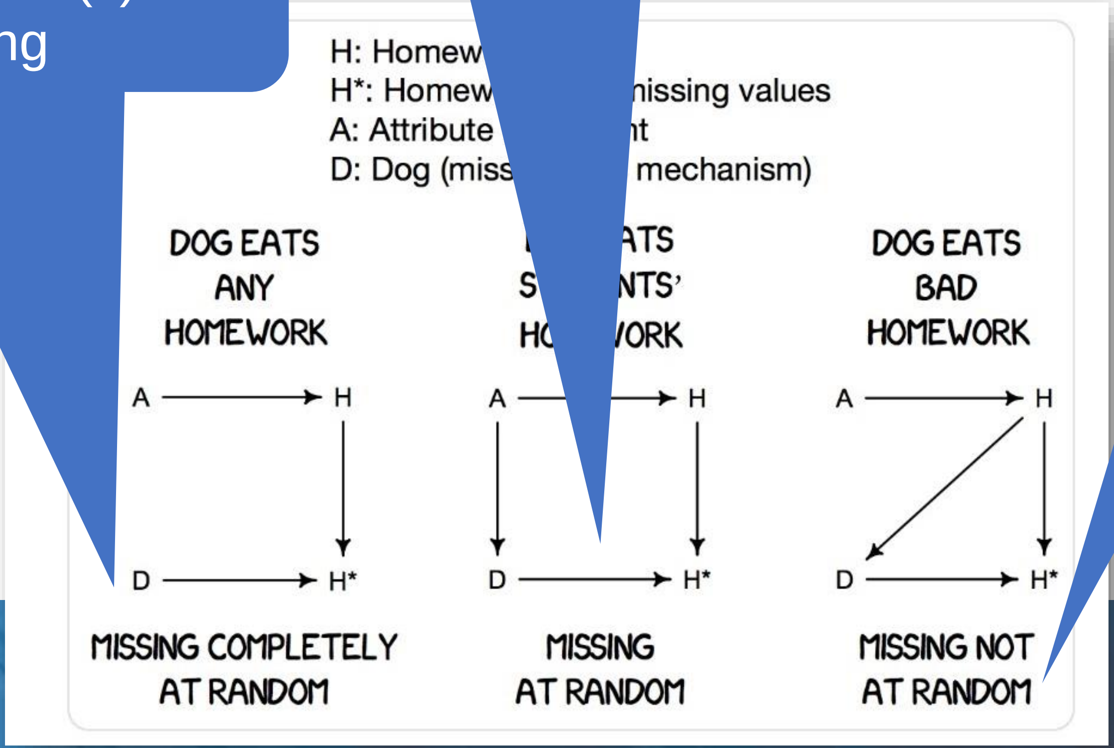
H: Homework
H*: Homework with missing values
A: Attribute of student
D: Dog (missingness mechanism)

Missing Data Mechanisms

possible to recover unbiased estimates if the right other variables are present.

cannot recover unbiased estimates

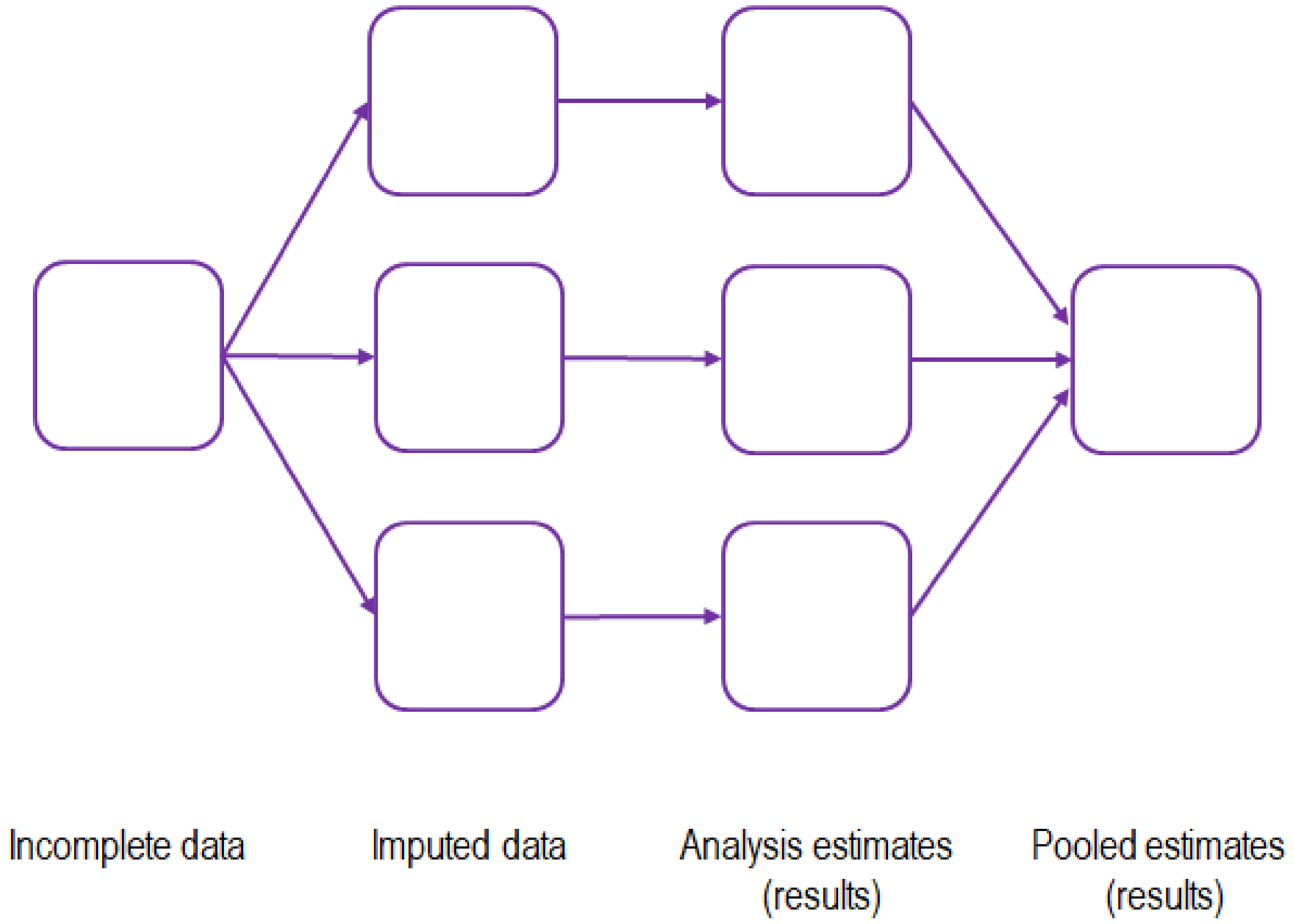
listwise deletion will yield unbiased estimates of the true parameter(s) if the data had not been missing



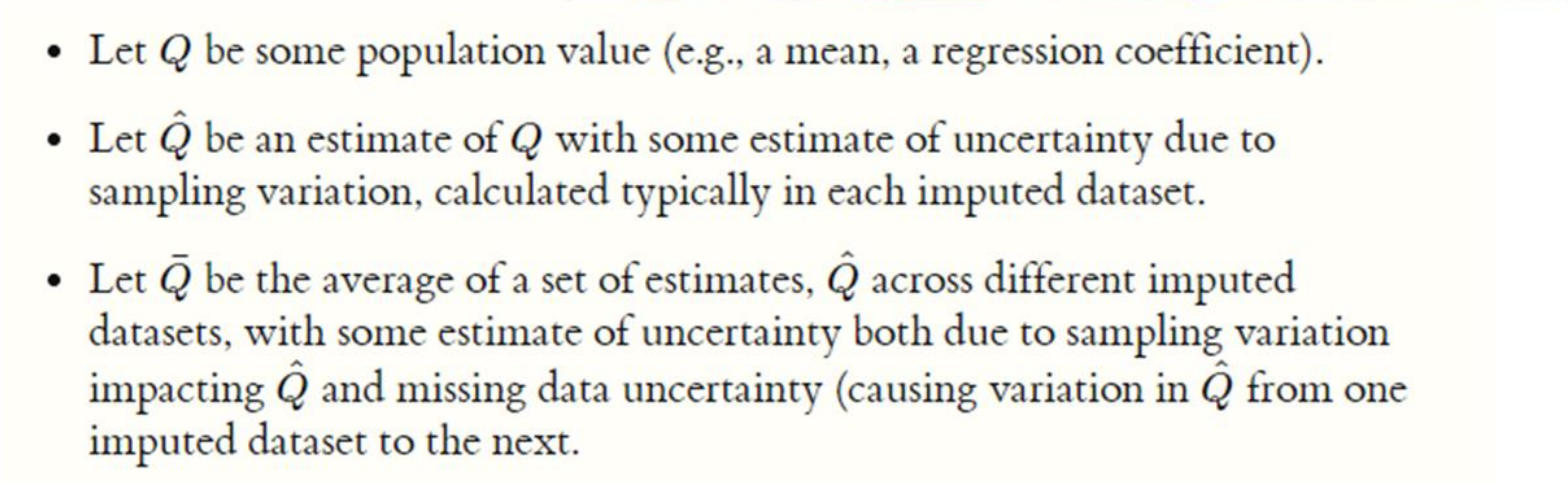
Recommendations:

1. Always include any variables upon which *y* theoretically (not empirically) depends.
2. Carefully consider the missing data mechanism. Is the missing outcome variable conditionally dependent on some other variable that *did* get collected (MAR or covariate-dependent MCAR), or is it assumed that the value of the outcome variable (if we did know it) is the reason for the missingness (MNAR)?
3. It is impossible to empirically rule out MNAR. Sensitivity analyses can be performed to assess the effect of missing observations had they been collected. Tutorials: Coertjens et al., 2017; Leurent et al., 2018; Resseguier et al., 2011.
4. Always conduct analyses with all available data using maximum likelihood or multiple imputation methods. In rare cases, ML methods may not be available in which case multiple-imputation may be helpful.

A very crash course in Multiple Imputation (MI)

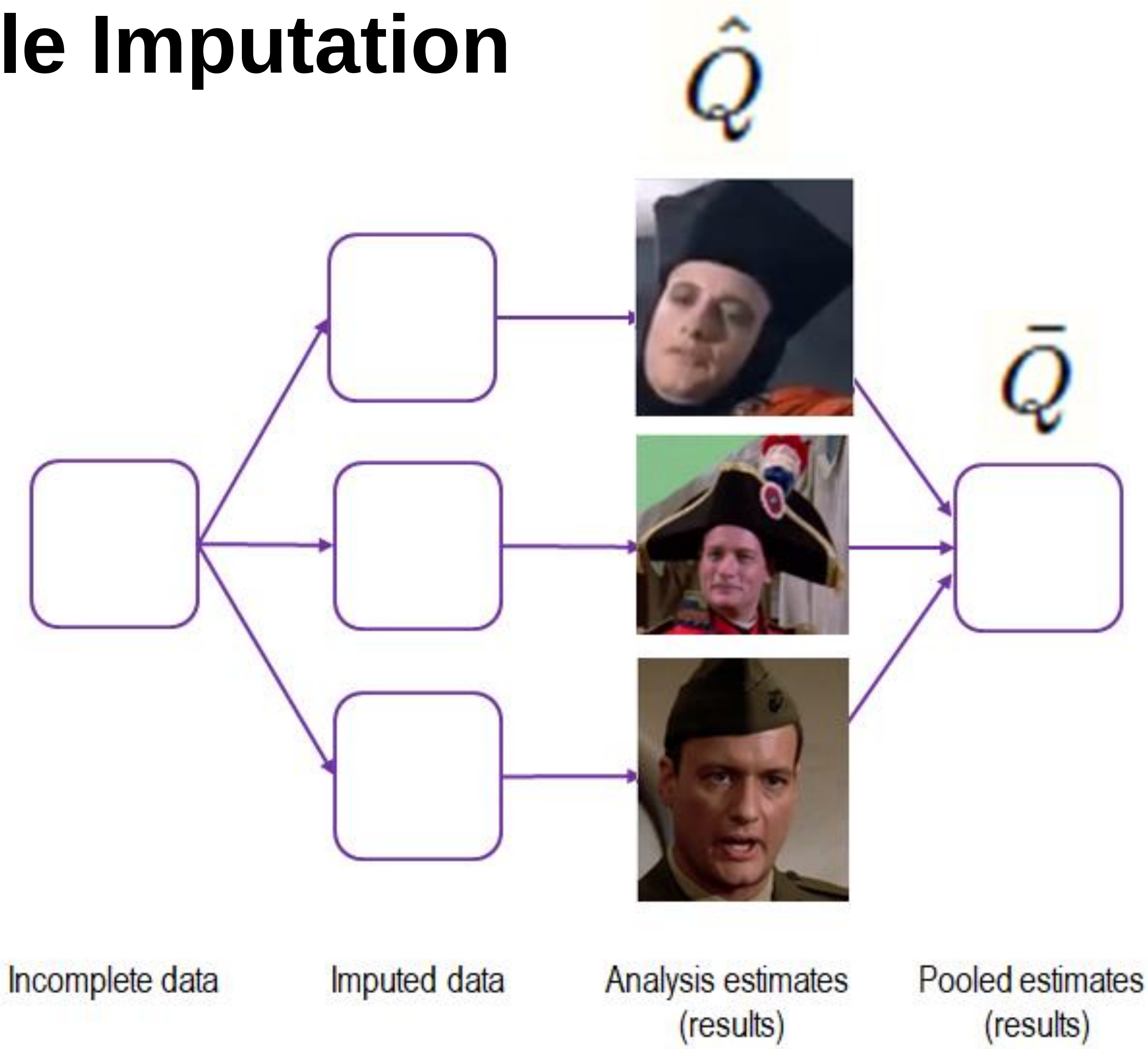


Multiple Imputation



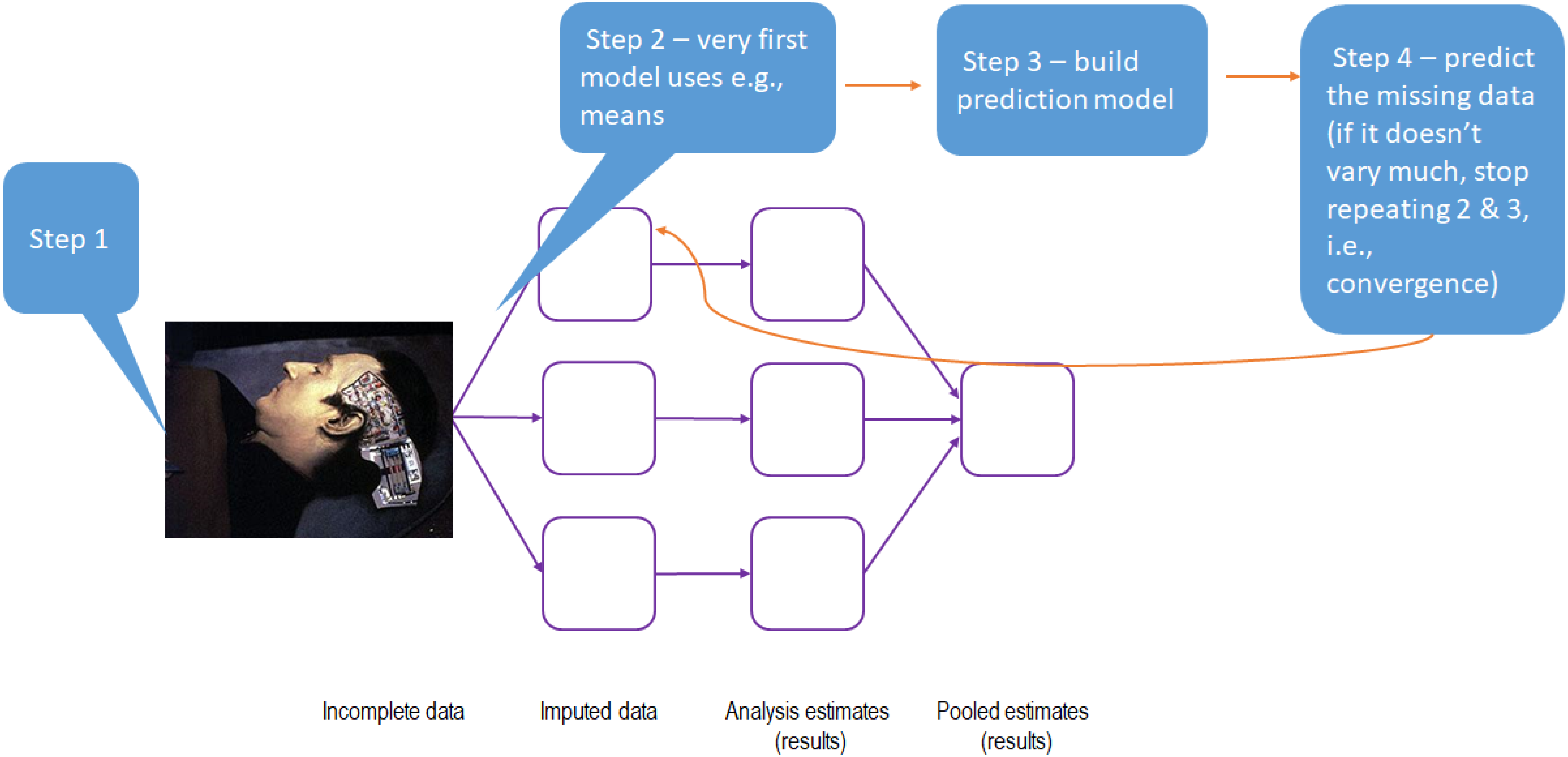
- Let Q be some population value (e.g., a mean, a regression coefficient).
- Let $\hat{Q}$ be an estimate of Q with some estimate of uncertainty due to sampling variation, calculated typically in each imputed dataset.
- Let $\bar{Q}$ be the average of a set of estimates, $\hat{Q}$ across different imputed datasets, with some estimate of uncertainty both due to sampling variation impacting $\hat{Q}$ and missing data uncertainty (causing variation in $\hat{Q}$ from one imputed dataset to the next.

Multiple Imputation



$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i$$

Multiple imputation STEPS



Step 1

Step 2 – very first model uses e.g., means

Step 3 – build prediction model

Step 4 – predict the missing data (if it doesn’t vary much, stop repeating 2 & 3, i.e., convergence)

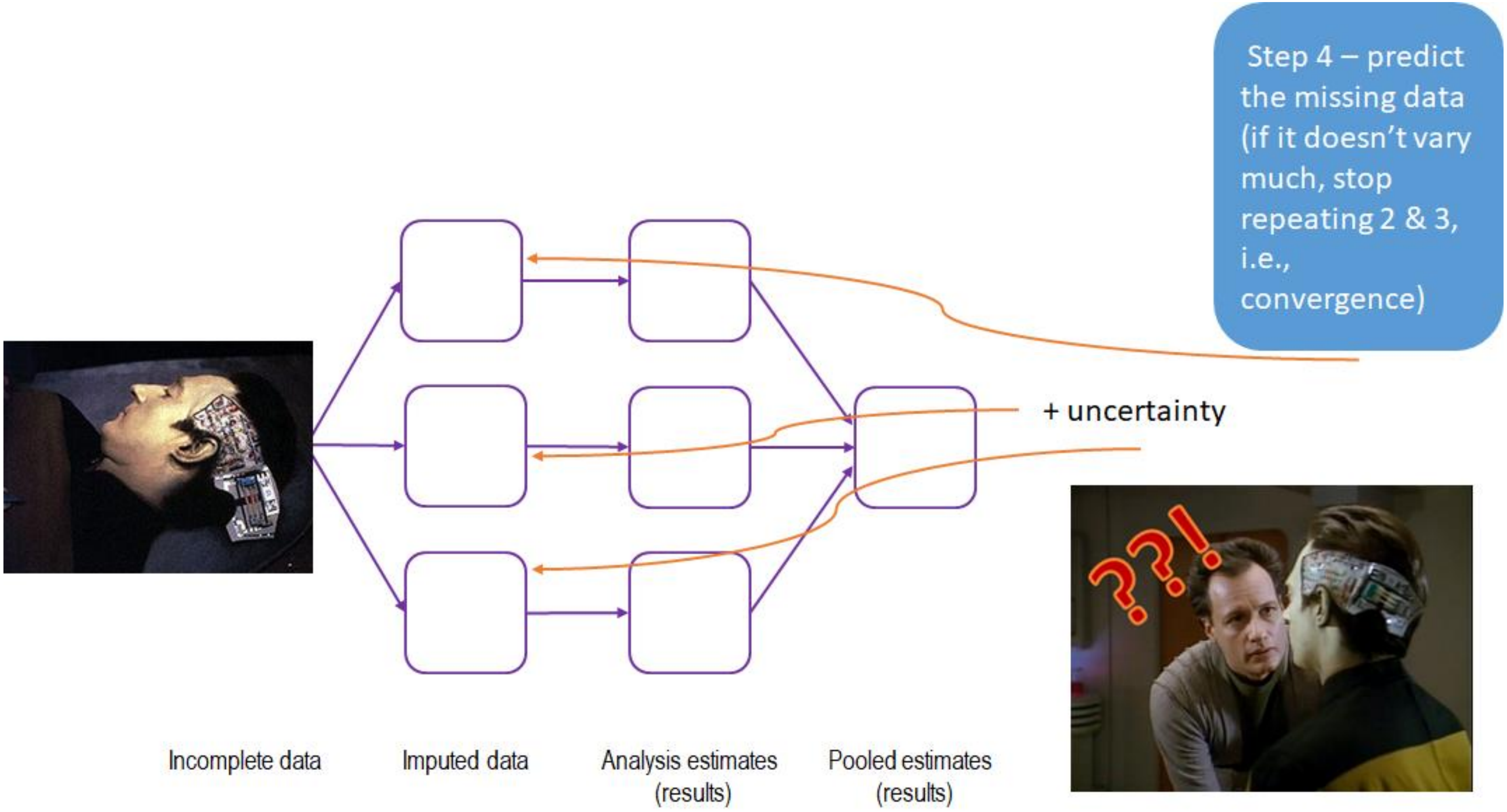
Incomplete data

Imputed data

Analysis estimates (results)

Pooled estimates (results)

Multiple imputation STEPS



Step 4 – predict the missing data (if it doesn't vary much, stop repeating 2 & 3, i.e., convergence)

+ uncertainty

Incomplete data

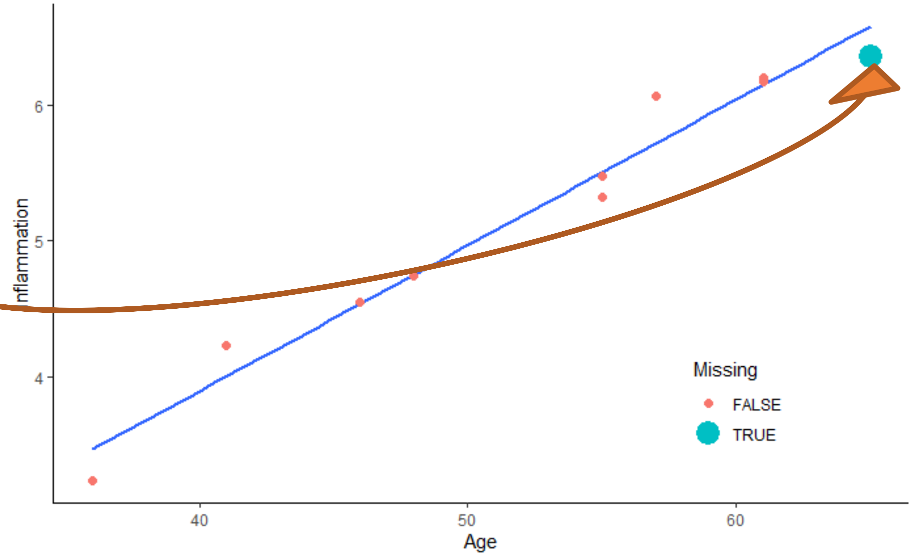
Imputed data

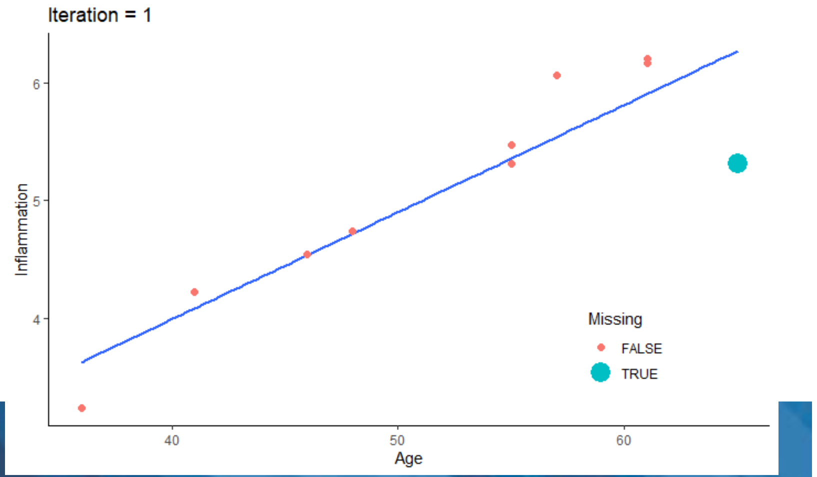
Analysis estimates (results)

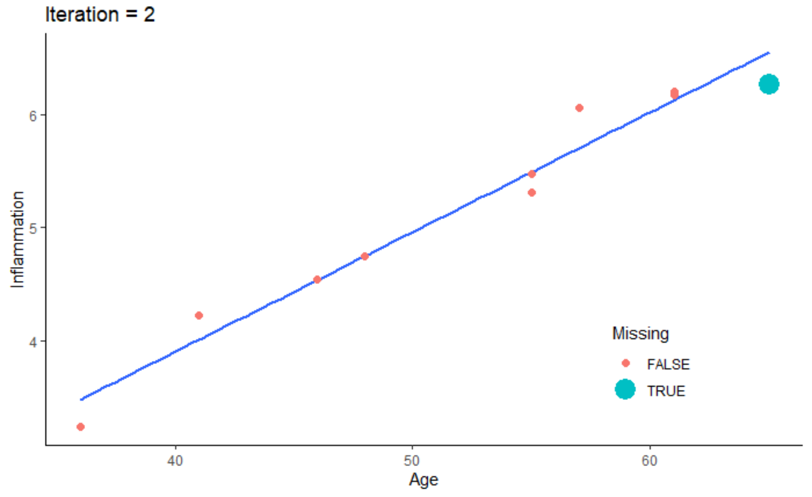
Pooled estimates (results)

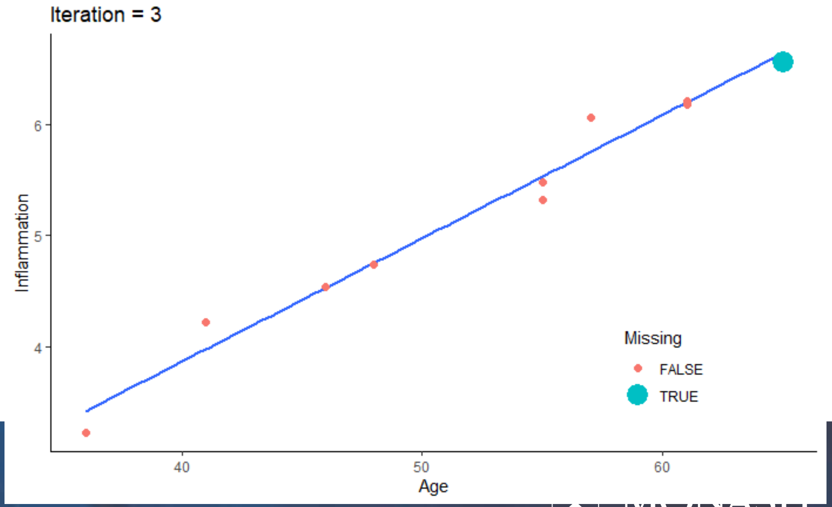
Prediction models to get imputed data

- You don’t usually do this yourself but here’s how it works
- Start with a data set and make one missing:
- Remove it, replace it with something simple like the median, and then use the regression line in a series of iterations getting that imputed data point closer and closer to the regression line:









Graphs from Dr Josh Wiley: https://joshuawiley.com/

MI vs other missing data approaches

- Imputation means that you get one or more new datasets with real actual values in place of the missing data
- Other approaches ARE MAGICAL (they estimate the parameters as if the data is not missing but you cannot extract imputed data).
- These include FIML, EM, WLSMV and others which are often used in a latent framework