Big Data: Is it for me?

What it is, how to use it

Henrik Daae Zachrisson

Apr 07 2021, Society for Research in Child Development Biannial Conference

The presentation is
supported by grant # 818425
from the Eurpoean Research Council
ERC-CoG-2018

UiO : University of Oslo

What I'll cover

- What is big data?
- Why consider using big data?
- Where to find them?
- What does it take?

What is big data?

Definition

- Volume, velocity, variety of data streams (Laney, 2001; Gilmore, 2016)
  - High volume
    - n or t or storage
  - High velocity (intensity)
    - Data generation or measurement
  - High variety
    - Various data sources
    - Complex data structure
- Little consensus on definition (Favaretto et al., 2020)

Illurstation: flickr.com

What is big data?

Data sources

- Dynamic digital data, e.g.,
  - Social media data
  - Meta-data from applications/games
- Large secondary datasets, linked
  - Administrative
  - Surveys

What is big data?

Data sources

- Dynamic digital data, e.g.,
  - Social media data
  - Meta-data from applications/games

- Large secondary datasets, linked
  - Administrative
  - Surveys

What is big data?

Methodology

- Machine learning
- Data mining techniques
  - See Grimm et al., 2020, for integrations with longitudinal modeling
- «Conventional» methodology
  - Secondary data analyses in big (complex) data sets
    - See e.g., Davis-Kean et al., 2015; 2017; opportunityinsights.org

What is big data?

Methodology

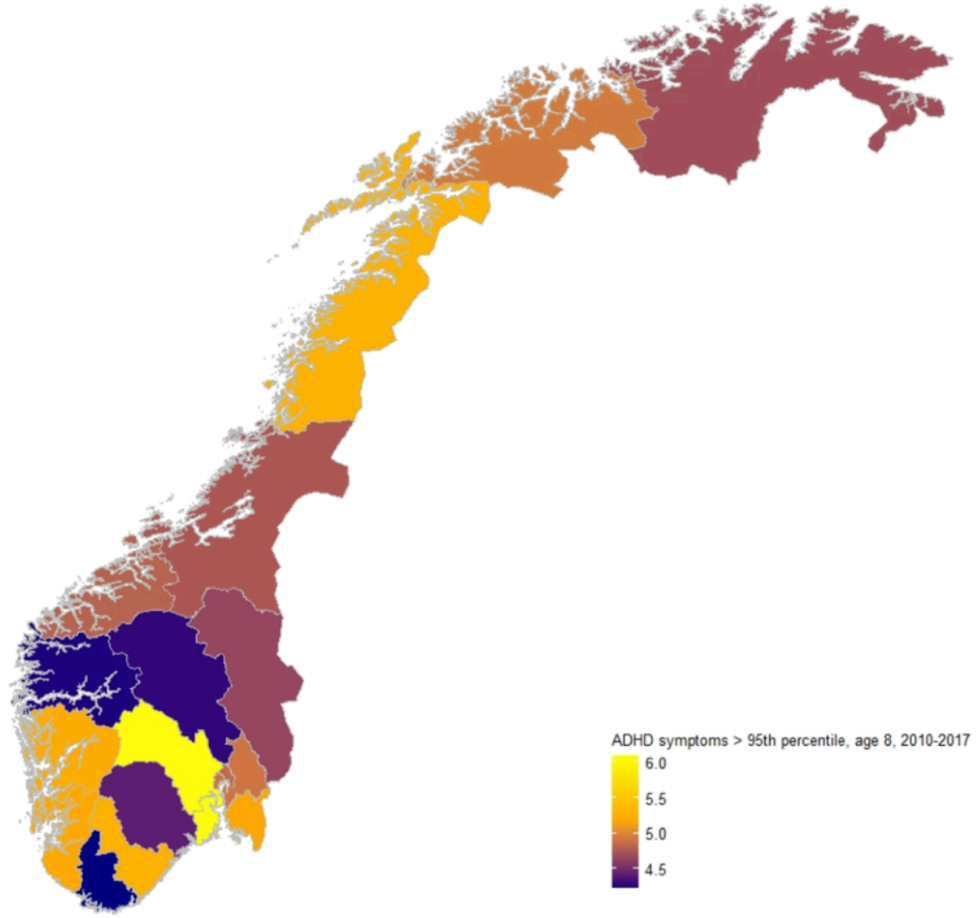
- Machine learning
- Data mining techniques
  - See Grimm et al., 2020, for integrations with longitudinal modeling

- «Conventional» methodology
  - Secondary data analyses in big (complex) data sets
    - See e.g., Davis-Kean et al., 2015; 2017; opportunityinsights.org

What is big data?

Examples of large secondary datasets

- Administrative records (local/national/international)
  - Education—test scores
  - Demographics
- Surveys
  - Large health/social surveys
  - Compilations of smaller surveys
- Linkage of data
  - Administrative and survey



An example of big data

The Norwegian Mother Father and Child Cohort Study (MoBa)

+ Administrative records = EQOP (uio.no/eqop)

- 109,000 pregnancies (95,000 mothers; Magnus et al., 2016)
- 1999-2009
- Questionnaires (across childhood) & genetic data
  - E.g., health, behavior, language, temperament, parenting/ home, parental physical & mental health

  Linked with

- Demographics, income, education, residency, for the entire population of Norway since 1970´s

Why consider using big data?

It's fun: Creativity & Opportunity

- Creativity in theory and design
    - Know the context of your data
        - Quasi experiments
        - Children in context
            - (e.g., differential growth)
- Statistical power
    - Complex interactions
    - Small subgroups

Why consider using big data?

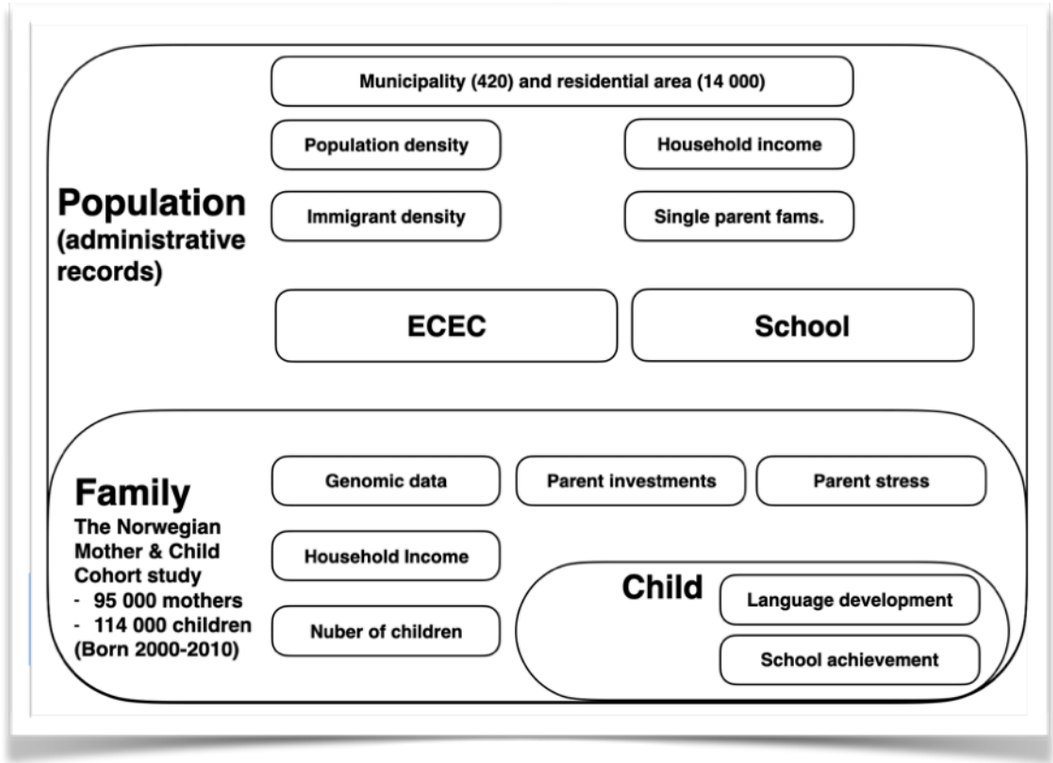
It's demanding: Forces openness & rigour

- You're stuck with what's in the data
  - Attention to design and measurement
- Encourages open science practice and thinking
  - Preregistration for secondary data analyses (OSF)
- Rigour
  - Be aware of limitations
  - Sensitivity & robustness checks
  - Others will check!

Why consider using big data?

Theory

- Socio-Ecological theory
  - Children (and data) in context
- Transactions
  - Person-context
  - Interplay between contextual levels
- Think big and not so WEIRD?
  - Developmental processes across contexts
- Does not exclude other theories!



Why consider using big data?

Method—External validity

- Data on the population or representative samples
  - (Often) multiple cohorts & longitudinal data
  - Geographical variation
  - Multiple countries
- Empirical opportunities
  - Multiple groups
  - Multiple levels of analyses

Why consider using big data?

Method—External validity in smaller datasets

- Sampling frame and generalizability of smaller samples
- Smaller datasets (no linkage)
    - Contextualize
        - Wider context of space & time
- Link big data with your (smaller) data
    - School-level, administrative, data
    - Demographic data

Why consider using big data?

Method—Internal validity

- Design opportunities
  - Quasi experimental methods
    - Difference-in-difference
    - Instrumental variables
    - Sibling models
- Know the context of your data
  - Socioeconomic changes/differences
  - Policies/reforms

Why consider using big data?

Motivational example: Macro level

Linking job loss, inequality, mental health, and education

Elizabeth O. Ananat[1], Anna Gassman-Pines[1], Dania V. Francis[2], Christina M. Gibson-Davis[1]
+ See all authors and affiliations

Science 16 Jun 2017:
Vol. 356, Issue 6343, pp. 1127-1128
DOI: 10.1126/science.aam5347

- Job losses at state level across time: Bureau of Labor Statistics
- Educational mobility, college enrollment: Equality of Opportunity Project
- Achievement 8th grade: NAEP
- Adolescent suicidal ideation: Youth Risk Behavior Survey (CDC)
- Household income: American Community Service

Why consider using big data?

Motivational example: Program evaluation

AERA Open
January-March 2018, Vol. 4, No. 1, pp. 1–16
DOI: 10.1177/2332858418756598

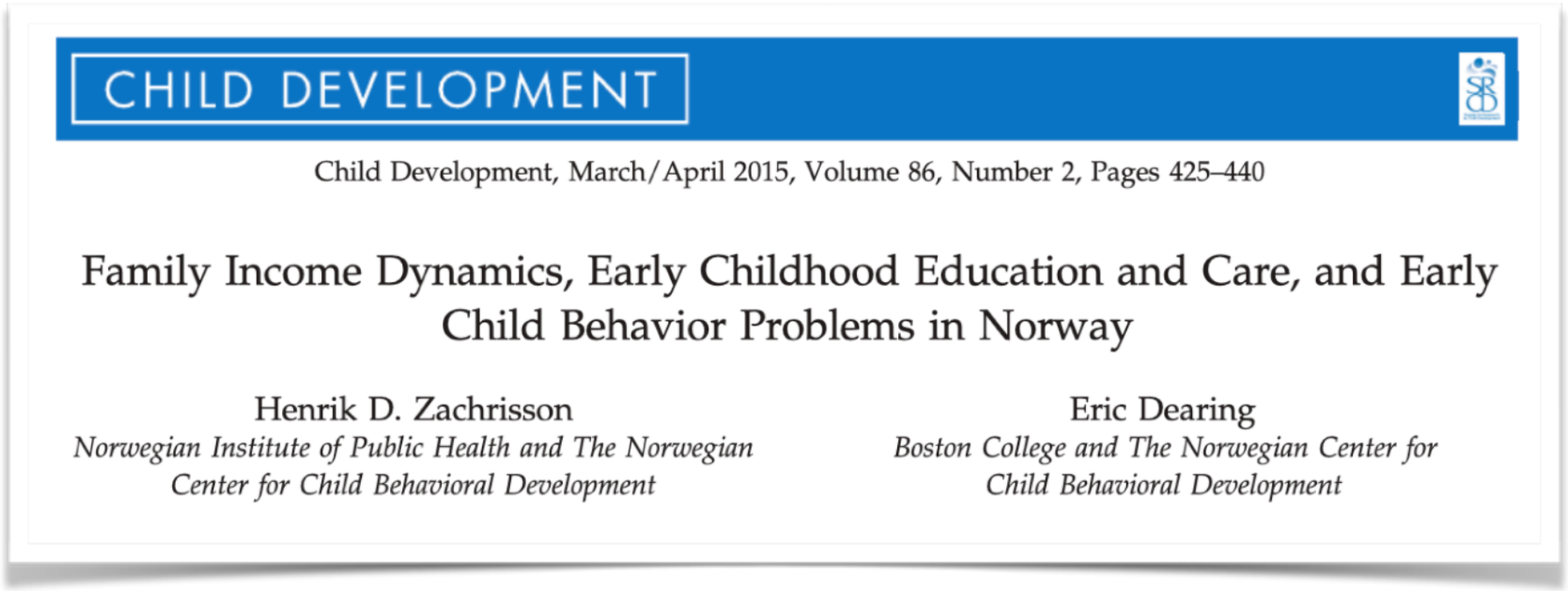
Estimating the Consequences of Norway's National Scale-Up of Early Childhood Education and Care (Beginning in Infancy) for Early Language Skills

Dearing, E., Zachrisson, H.D., Mykletun, A. & Toppelberg, C.O.

- MoBa health survey (6 birth cohorts)
  - Covering child care expansion in Norway
- Tax records (annual)

Why consider using big data?

Motivational example: Context/family level



CHILD DEVELOPMENT

Child Development, March/April 2015, Volume 86, Number 2, Pages 425–440

Family Income Dynamics, Early Childhood Education and Care, and Early Child Behavior Problems in Norway

Henrik D. Zachrisson
Norwegian Institute of Public Health and The Norwegian Center for Child Behavioral Development

Eric Dearing
Boston College and The Norwegian Center for Child Behavioral Development

- MoBa health survey
- Tax records (annual)

Why consider using big data?

Motivational example: developmental processes

CHILD DEVELOPMENT

Child Development, March/April 2013, Volume 84, Number 2, Pages 560–573

Action Imitation at 1½ Years Is Better Than Pointing Gesture in Predicting Late Development of Language Production at 3 Years of Age

Imac M. Zambrana, Eivind Ystrom, and Synnve Schjølberg
Norwegian Institute of Public Health

Francisco Pons
University of Oslo

- MoBa health survey

Where to find them?

Datasets collected for research, examples

- Overviews
  - Secondary data (Davis-Kean et al., 2015)
    - Data archives (e.g., icpsr; Murray Research Archive)
      - Large Soc Sci & Health surveys (Gilmore, 2016)
        - Hosted by govenment entities (e.g., Current Population Survey; Millenium Cohort Study)
        - Hosted by NGOs (e.g., AddHealth)
        - Measure-based datasets (e.g., MacArthur CDI)
- ILSA (TIMSS, PISA) & Stanford Education Data Archive
- opportunityinsights.org

Where to find them?

Administrative and evaluation data, examples

- Administrative records (national, regional)
  - Demographics, socioeconomics
  - Education & Health (e.g., state DoE-data)

- Data banks
  - UNICEF (www.unicef.org/statistics)
  - World Bank (microdata.worldbank.org)

What does it take?

Data aquisition

- Open source data
  - The search...
  - Read papers in related fields

- Restricted sources
  - Administrative work & diplomacy
  - Secure approvals
  - Arrange lincage

illustration: ripcurl.com

What does it take?

Data management

- Not all data are ready to analyze the way you want
  - Restructuring, combining, creating
    - Basic programming skills (codes to repeat procedures)
- Tips
  - Do not create a «master» dataset
    - Files specific to a set of analyses
  - Look for codes online on data management
    - Supplementary material for papers (econ/sociology)

What does it take?

Measurement

- Relevant measures
  - Often brief or abbrevated
  - Not always best choice of measure
- Often not best choice of timing of data collection
  - Developmentally appropriate
    - Precise about implications for conclusions

What does it take?

Measurement

- Psychometric care
  - Measurement models—often in short forms
    - Structure of scales (CFA)
      - Consistent with full formats?
    - Where are items most sensitive? (IRT)
      - Health surveys often use screenings
        - Most sensitive at the lower end of the distribution

What does it take?

Measurement

- External validation of scales
  - Short form vs long form in other data (Zachrisson et al., 2013)
    - Correlate subsets with full scale in other data

- Construction of new measures
  - External expert evaluations of items (e.g., Zambrana et al., 2013)
    - Mapping items to constructs with empirical validation

Big Data: is it for me?

Conclusion

- Untapped opportunites?
  - Theoretical extensions
    - Strengthen validity
      - External
      - Internal
- Creativity & rigour
  - Theory
  - Data/design

Thank you for your attention!

henrikdz@uio.no **www.uio.no/eqop**

Thanks to Imac M. Zambrana, Eric Dearing, Monica Melby-Lervåg, & Arne Lervåg for valuable comments!

References

Ananat, E. O., Gassman-Pines, A., Francis, D. V., & Gibson-Davis, C. M. (2017). Linking job loss, inequality, mental health, and education. *Science, 356*(6343), 1127-1128. doi:10.1126/science.aam5347

Davis-Kean, P. E., & Jager, J. (2017). From small to Big: Methods for incormorating large scale data into developmental science. *Monographs of the Society for Research in Child Development, 82*(2), 31-45. doi:https://doi.org/10.1111/mono.12297

Davis-Kean, P. E., Jager, J., & Maslowsky, J. (2015). Answering Developmental Questions Using Secondary Data. *Child Development Perspectives, 9*(4), 256-261. doi:https://doi.org/10.1111/cdep.12151

Dearing, E., Zachrisson, H. D., Mykletun, A., & Toppelberg, C. O. (2018). Estimating the consequences of Norway's national scale-up of early childhood education and care (beginning in infancy) for early language skills. *AERA Open, 4*(1), 1-16. doi:10.1177/2332858418756598

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLOS ONE, 15*(2), e0228987. doi:10.1371/journal.pone.0228987

Gilmore, R. O. (2016). From big data to deep insight in developmental science. *WIREs Cognitive Science, 7*(2), 112-126. doi:https://doi.org/10.1002/wcs.1379

Grimm, K. J., et al. (2020). Big data in developmental psychology. Big Data in Psychological Research. S. E. Woo, L. Tay and R. W. Proctor, American Psychological Association.

Laney, D. 3D data management: controlling data volume, velocity, and variety (Tech. Rep.). META Group, February 2001. Available at: http://blogs.gartner.com/ doug-laney/files/2012/01/ ad949-3D-Data-Management- Controlling-Data-Volume-Velocity-and-Variety.pdf.

Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Daltveit, A. K. i., . . . Knudsen, G. P. (2016). Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *International journal of epidemiology*, dyw029.

Zachrisson, H. D., Dearing, E., Lekhal, R., & Toppelberg, C. O. (2013). Little evidence that time in child care causes externalizing problems during early childhood in Norway. *Child Dev, 84*(4), 1152-1170. doi:10.1111/cdev.12040

Zachrisson, H. D., & Dearing, E. (2015). Family income dynamics, early childhood education and care, and early child behavior problems in Norway. *Child Dev, 86*(2), 425-440. doi:10.1111/cdev.12306

Zambrana, I.M., Ystrom, E., Schjølberg, S. and Pons, F. (2013), Action Imitation at 1½ Years Is Better Than Pointing Gesture in Predicting Late Development of Language Production at 3 Years of Age. Child Dev, 84: 560-573. https://doi.org/10.1111/j.1467-8624.2012.01872.x