For whom M L rolls?

Sense and feasibility

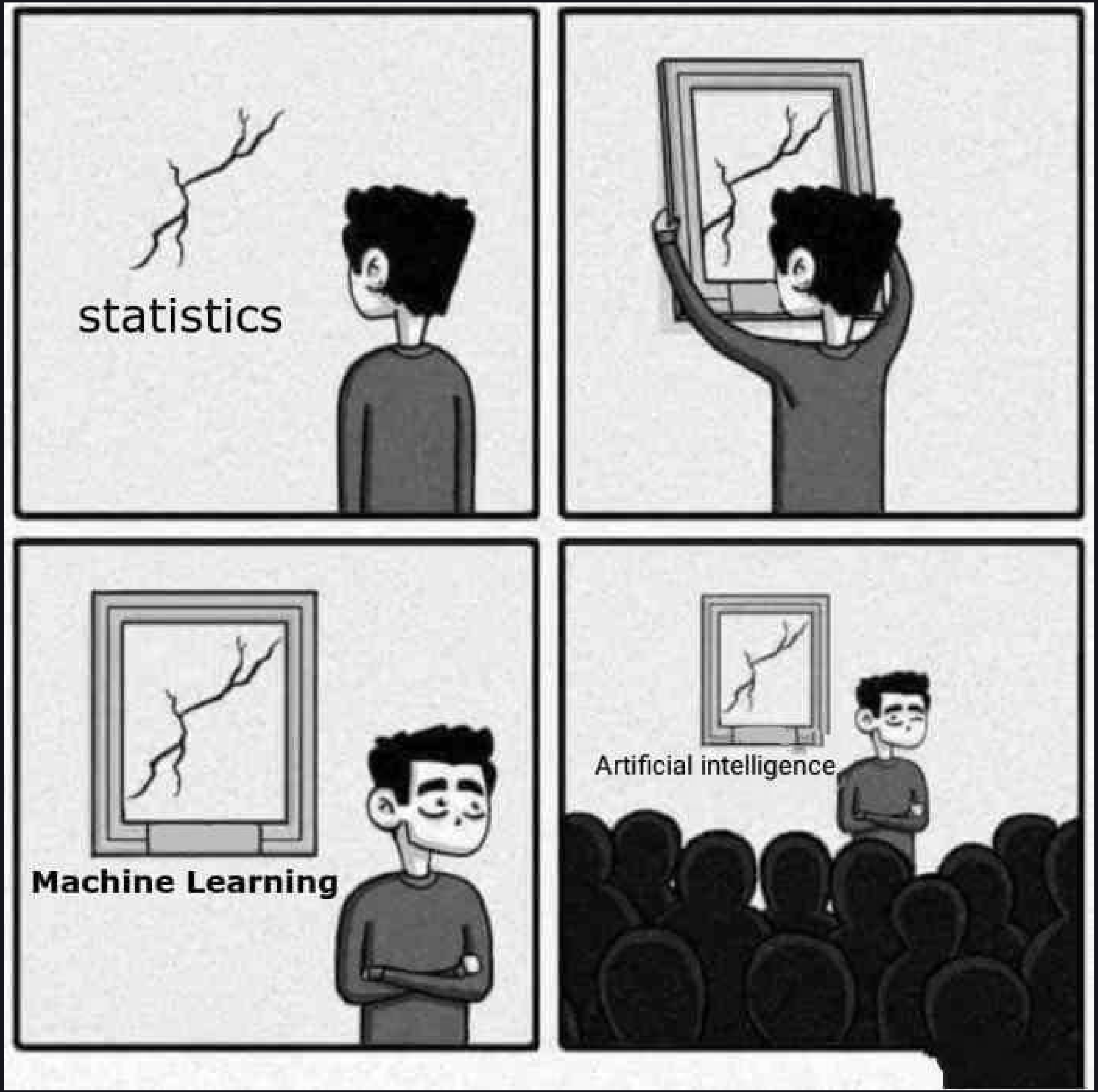
Claus Thorn Ekstrøm
UCPH Biostatistics
ekstrom@sund.ku.dk

DES, May 20th 2021

@ClausEkstrom

Sorry!

Can Machine Learning Assist Epidemiologists in Drawing Causal Inference?



Comic by sandserif

Studies Show By Kim Tingley

If randomized control trials or observational studies alone can't say whether coffee is good for our hearts, maybe machine learning can help.

NY Times Magazine, March 24th, 2021

Excerpt from NC on Health Research Ethics

Protocol: *The full dataset will be analyzed using supervised and unsupervised machine learning methods to identify associations and patterns in radiological diagnoses that traditional statistical models cannot identify.*

These associations can be used to explain combinations of factors, where patients are potentially unnecessesary scanned.

It is not possible to make a power calculation for this study since there are more factors in play when research is done with machine learning algorithms.

Proponents

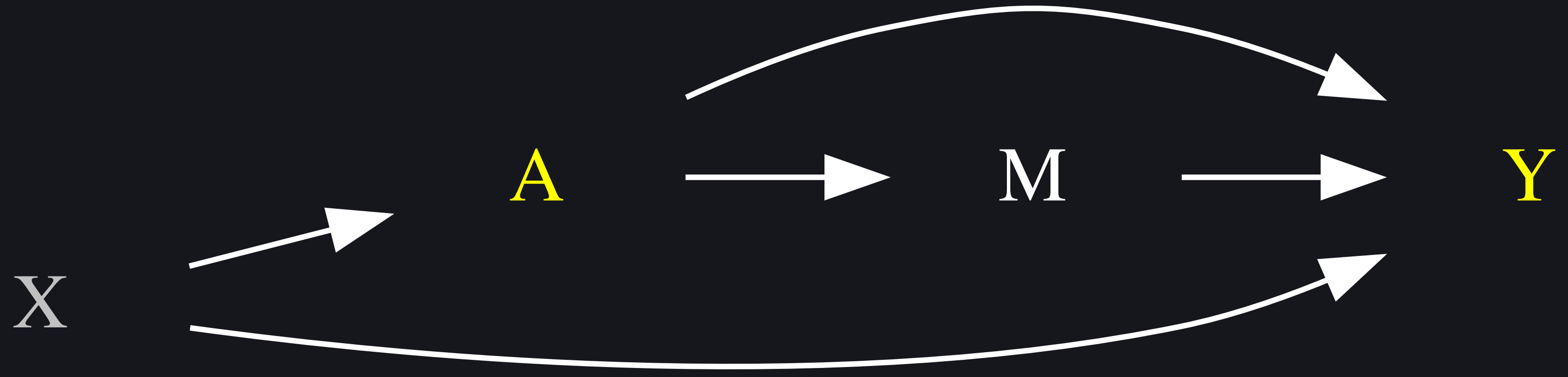
The magic of ML methods:

- Allow the data to speak for themselves
- Better
- More flexible
- Have fewer assumptions

- Random forest
- Neural networks
- Penalized regression
- Gradient boosting
- Logistic regression
- Algorithms

But what about causality?

Causal Inference and Directed Acyclic Graphs



Read causal relationships. Can we identify causal effects? Confounders, colliders, conditional independence.

Assumptions untestable. *"Let the DAG be given …"*

"Let the data speak ..."

- Massive data
- "Hunt for patterns"

$$A \longrightarrow Y$$

$$A \longleftarrow Y$$

Confounders, colliders, ...

Danish registry data

- What variables to include?
- Time? Non-equidistant measures.

  How long since I last visited my GP?

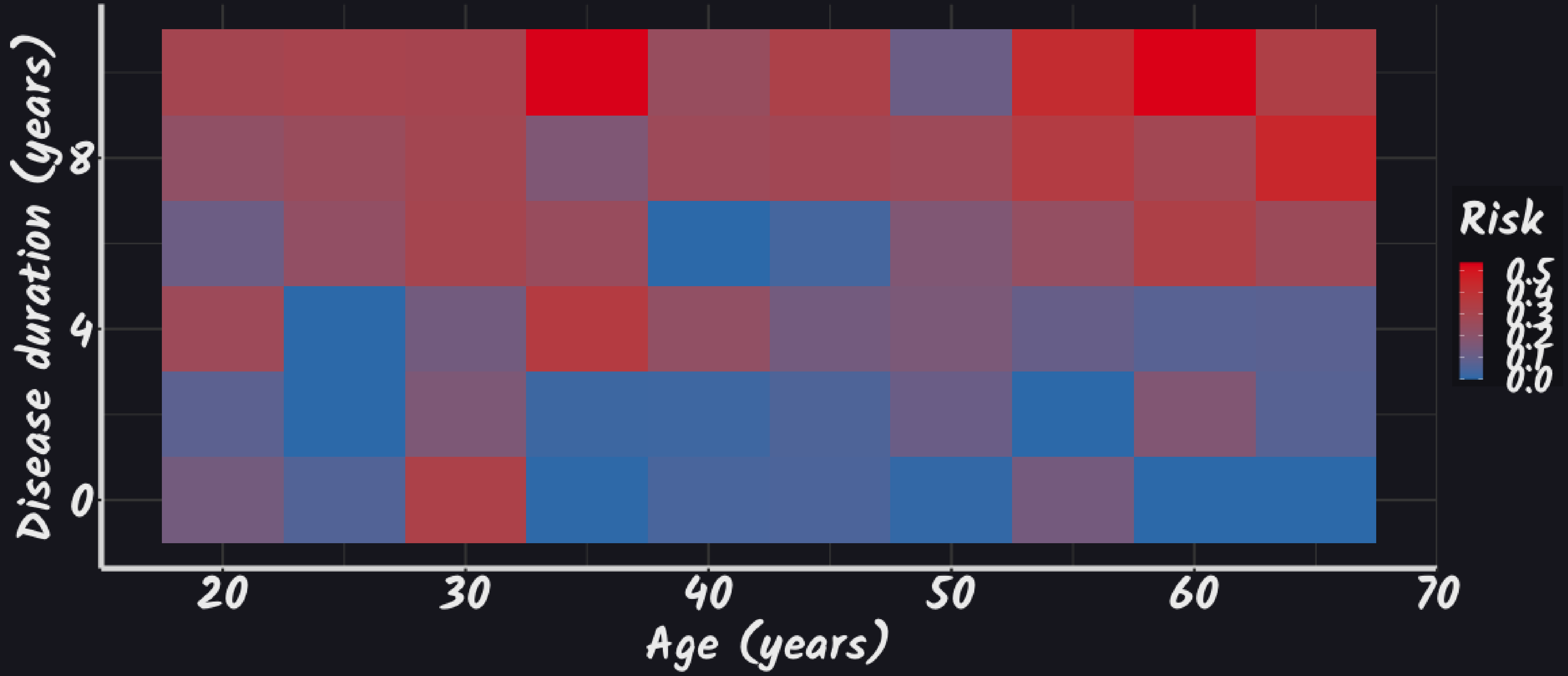
ML is more flexible

Yes - by choice.

Could achieve the same with traditional models.

Price: interpretability

Non-continuous risk prediction



Where can ML play a critical role in CI?

Estimating causal effects with ML

$$\mathbb{E}(Y|A, X) = \beta_0 + \beta_1 A + \beta_2 X$$

Estimating causal effects with ML

$$\mathbb{E}(Y|A,X) = \text{"ML"}$$

Average Treatment Effect (ATE)

$$\mathbb{E}_X[\mathbb{E}(Y|A=1,X) - \mathbb{E}(Y|A=0,X)]$$

with estimator

$$\frac{1}{N}\sum_{i=1}^{N}[\widehat{\mathbb{E}}(Y|A=1,X_i) - \widehat{\mathbb{E}}(Y|A=0,X_i)]$$

Machine Learning *g*-formula algorithm

1. Estimate $\mathbb{E}(Y|A, X)$ using our machine learning tool. Even better: an ensemble tool
2. Set $A = 1$ for all observations and predict outcomes for all
3. Set $A = 0$ for all observations and predict outcomes for all

$$\frac{1}{N}\sum_{i=1}^{N}[\widetilde{\mathbb{E}}(Y|A=1, X_i) - \widetilde{\mathbb{E}}(Y|A=0, X_i)]$$

To interpret *causally* (average causal treatment effect) we still need the standard causal assumptions *and* proper models.

Causal discovery / structure learning

Let the DAG be given …

Use ML to discover causal relationships from observational data.

PC algorithm identifies conditional independencies among the variables.

PC algorithm

Input: a set of variables.

Output: *completed partially directed acyclic graph* (CPDAG).

Assumptions:

- The set of observed variables is sufficent
  - All common causes present in the dataset
  - Extensions that account for latent variables do exist!
- The distribution of the observed variables is faithful to a DAG

Spirtes & Glymour (1991). *An algorithm for fast recovery of sparse causal graphs*. Social Science Computer Review.

PC algorithm 2

There is an edge $A - Y$ if and only if A and Y are dependent conditional on every possible subset of the other variables.

$A \perp Y$? $A \perp Y | X$? $A \perp Y | M$? $A \perp Y | X, M$?

Number of tests? Prone to statistical mistakes? ML for (conditional) independence testing? Time?

After skeleton: Orient triplets $X - Y - Z$ as $X \to Y \leftarrow Z$ iff X and Z are dependent conditional on every set containing Y.

Or use additional information

ACCEPTED MANUSCRIPT

Data-Driven Model Building for Life Course Epidemiology

Anne H Petersen, Merete Osler, Claus T Ekstrøm

American Journal of Epidemiology, kwab087, https://doi.org/10.1093/aje/kwab087

Published: 29 March 2021 **Article history** ▾

Cite Permissions Share ▾

Abstract

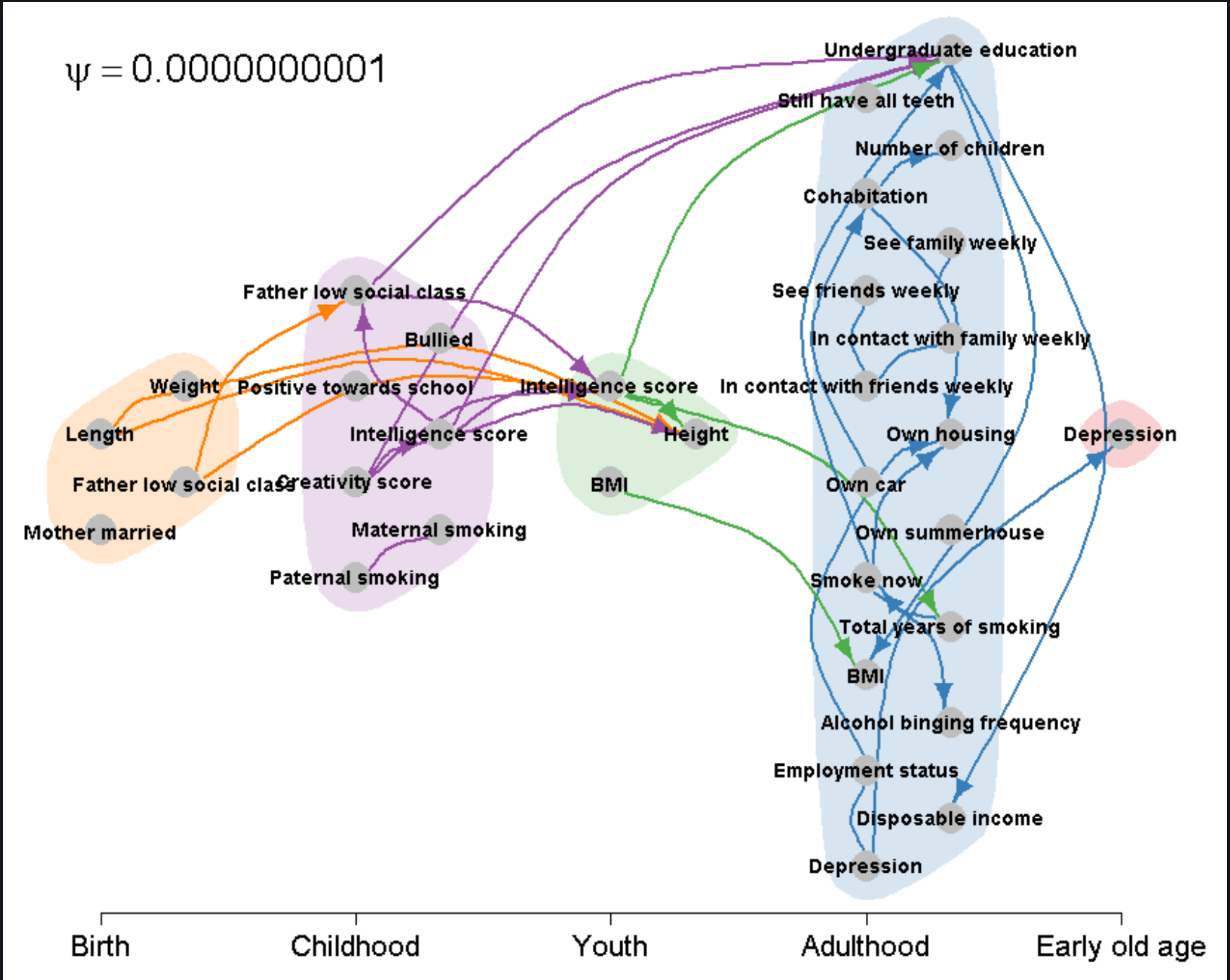
Life course epidemiology is useful for describing and analyzing complex

Use *temporal information* to help orient edges (Temporal PC).

Metropolit Cohort

- Danish men born in 1953. Followed from birth until 65 yo.
- Surveys at age 12 and 51. Extensive administrative register data from the Danish national registers. $N = 2928$.
- Consider 33 variables measured in 5 periods over the life course: birth, childhood (age approximately 12), youth (age 18-30), adulthood (age approximately 51), and early old age (age approximately 65).
- Outcome: clinical depression.

Osler et al. *Cohort profile: the Metropolit 1953 Danish male birth cohort*. International Journal of Epidemiology



ψ = 0.0000000001
Undergraduate education
Still have all teeth
Number of children
Cohabitation
See family weekly
See friends weekly
In contact with family weekly
In contact with friends weekly
Own housing
Own car
Own summerhouse
Smoke now
Total years of smoking
BMI
Alcohol binging frequency
Employment status
Disposable income
Depression
Depression
Father low social class
Bullied
Positive towards school
Intelligence score
Creativity score
Maternal smoking
Paternal smoking
Weight
Length
Father low social class
Mother married
Intelligence score
Height
BMI
Birth
Childhood
Youth
Adulthood
Early old age

Summary

No inherent benefits for ML wrt causal inference.

Useful in *combination* with existing framework(s) for causal inference.
But no free lunch.

- Machine learning provides a useful alternative/addendum to modeling.
- Ideas in ML force us out of the old go-to techniques.
- Improved algorithms can perhaps make approaches feasible.

We still need to **think**. Field-knowledge is ever-more crucial.