Causal discovery from medical data: challenges and opportunities

Tom Claassen
Institute for Computer and Information Sciences,
Radboud University Nijmegen

AICPM workshop – 9 September, 2022

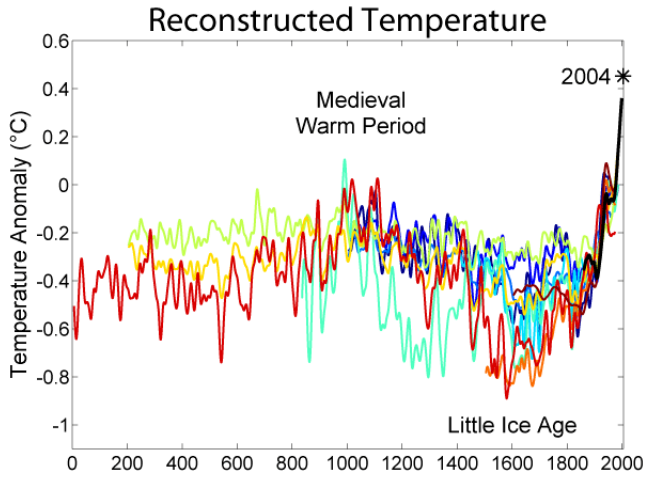
Radboud University Nijmegen

Outline

1. **Causal discovery algorithms**
2. Medical data and example applications
3. The Gap … and the quest to bridge it

Many important research questions are rooted in causality

benefits of exercise and healthy nutrition



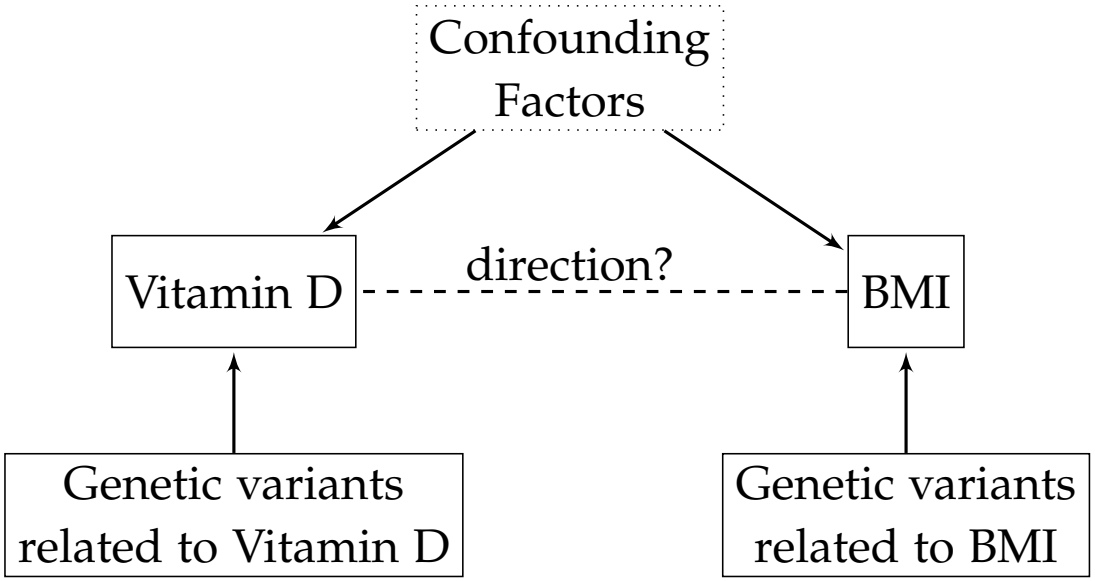
human activity and climate change

racial and gender bias in AI

Covid vaccine efficacy

Why causality?

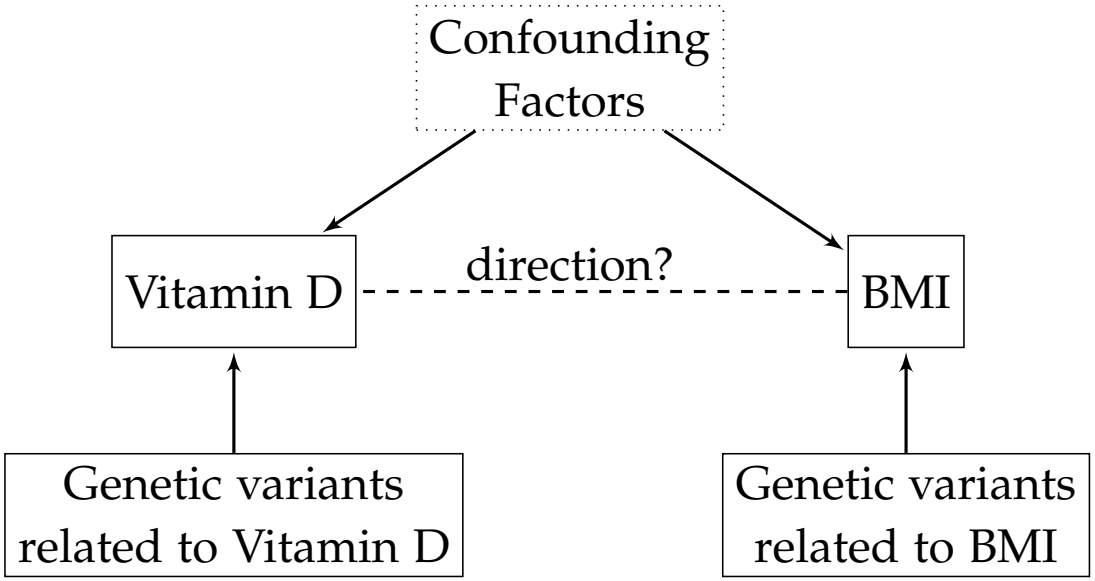
- '*correlation is not causation*' …
- … but in science we assume things happen for a reason,
- causal models explain *why* things happen
- with correct model => *predict effect* of interventions (changes)



from: Bucur, I. G., Claassen, T., & Heskes, T. (2020). Inferring the direction of a causal link and estimating its effect via a Bayesian Mendelian randomization approach. Statistical methods in medical research, 29(4), 1081-1111.

Personalized medicine

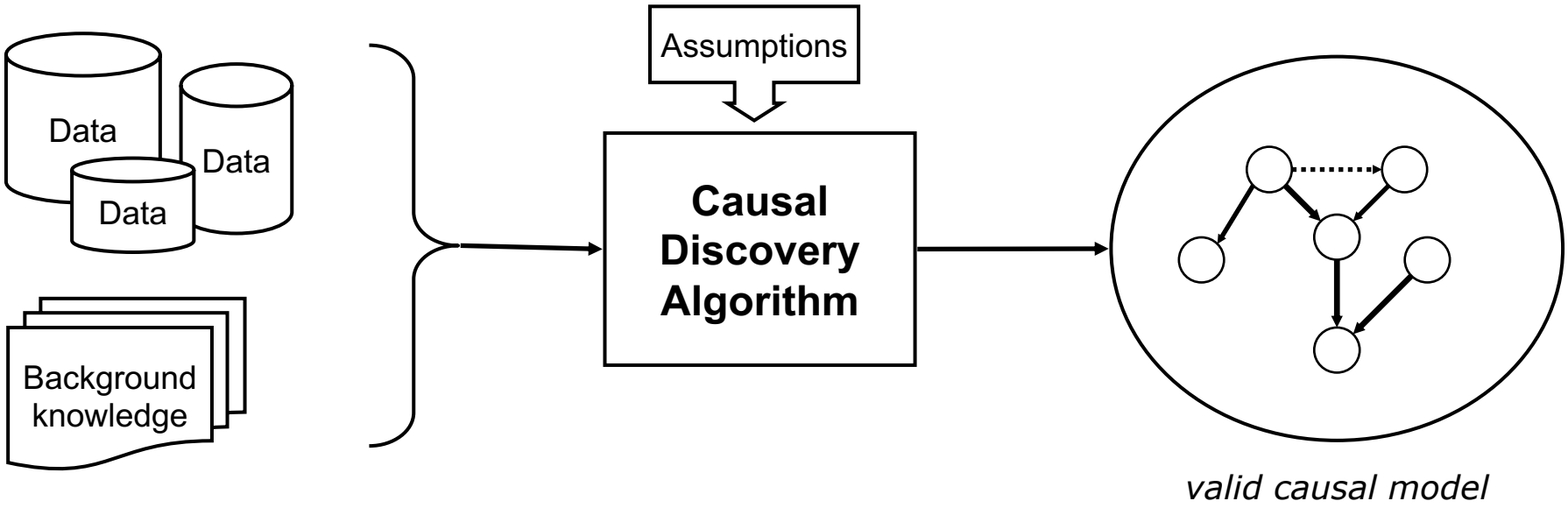
- for different individuals different treatments may be more effective / desirable in order to obtain a target effect
- focus on (weighted) CATE (*conditional* average treatment effect)
- relies on knowing the causal model



from: Bucur, I. G., Claassen, T., & Heskes, T. (2020). Inferring the direction of a causal link and estimating its effect via a Bayesian Mendelian randomization approach. Statistical methods in medical research, 29(4), 1081-1111.

How do we get the correct causal model?

- knowing the *structure* of the causal model is key
- early 1990's: theoretical breakthrough … we *can* learn very good causal networks from purely observational data!



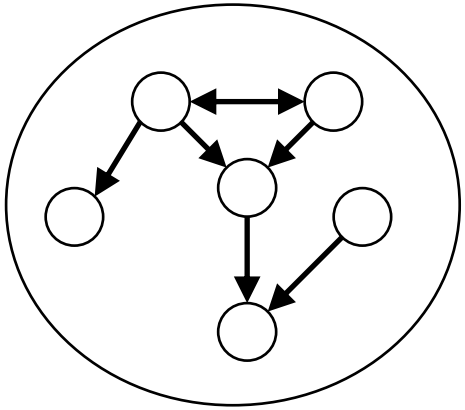
Two main paradigms

⇒ constraint-based algorithms

- score-based algorithms

Key assumptions

Ground truth causal model



graphical criterion

d-separation

generate

Observations

Data

Assumptions

causal Markov

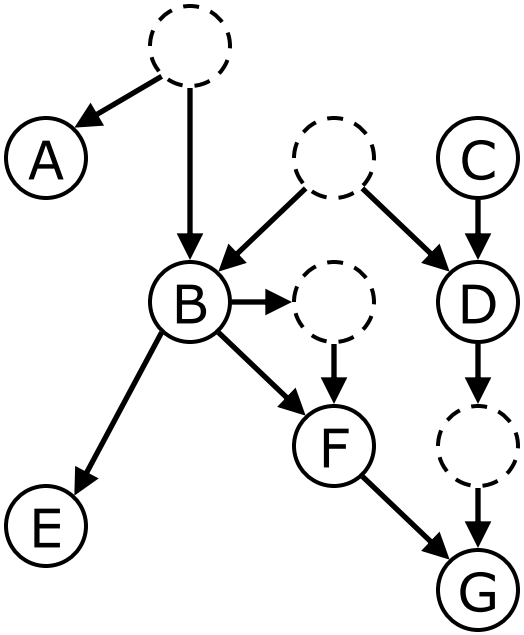
faithfulness

statistical test

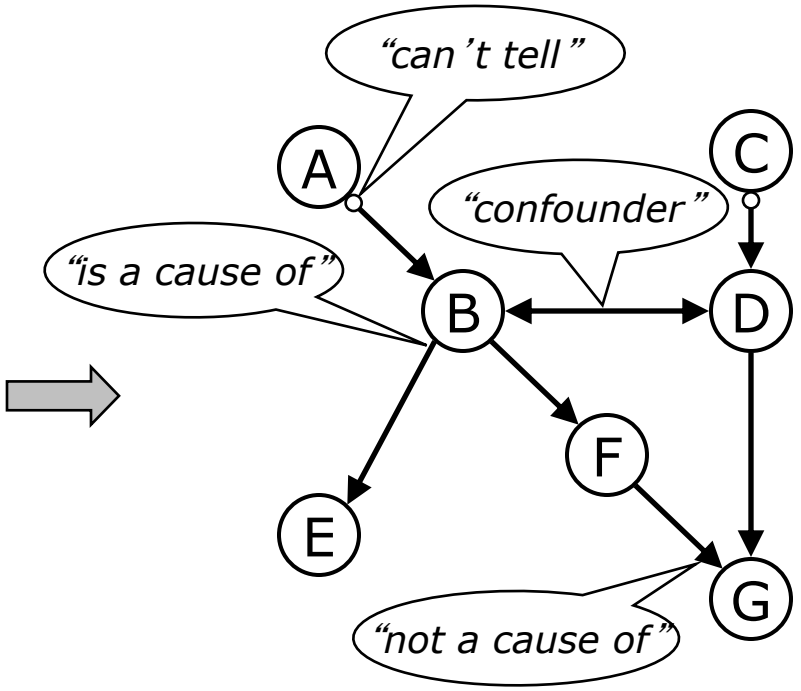
(conditional) independencies

Equivalence classes

- different causal structures can entail the same independence constraints
- no way to distinguish: indicate what we do know => 'equivalence class'



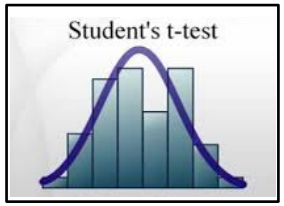
*underlying **causal DAG** G*



*inferred **causal model** P*
(Partial Ancestral Graph)

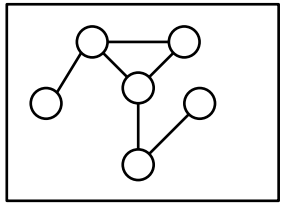
Typical constraint-based causal discovery

preprocess data, feature selection

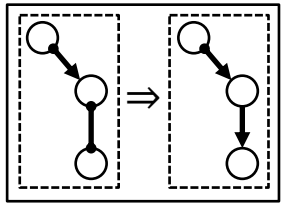


suitable test, missing data

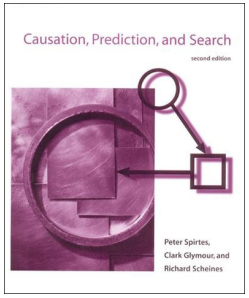
search independence constraints

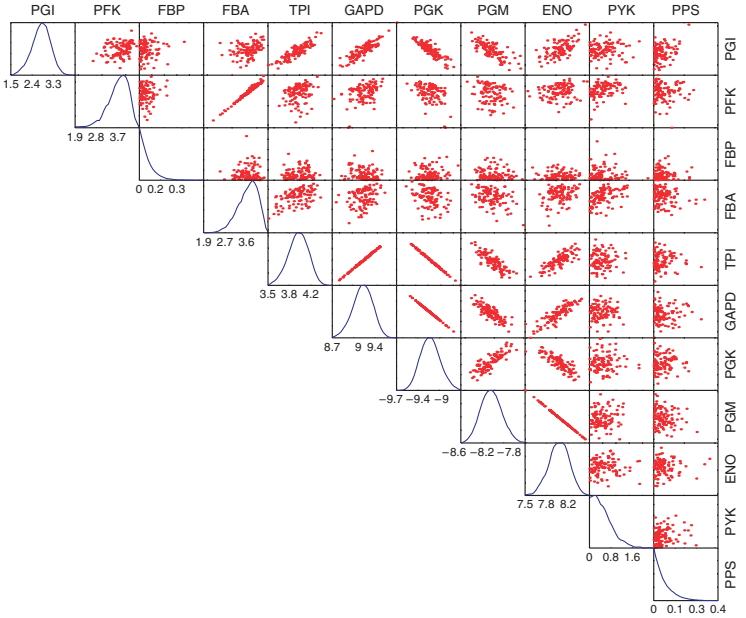


build skeleton

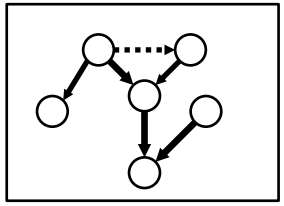


causal orientation rules





available experimental data



output causal model

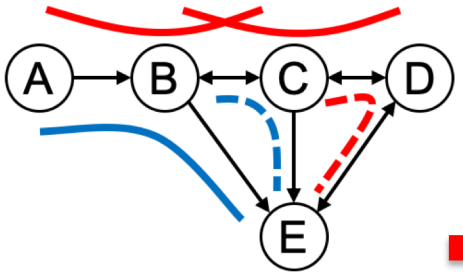
wrong when assumptions do not hold

Sneak peak demo: anytime-anywhere FCI+

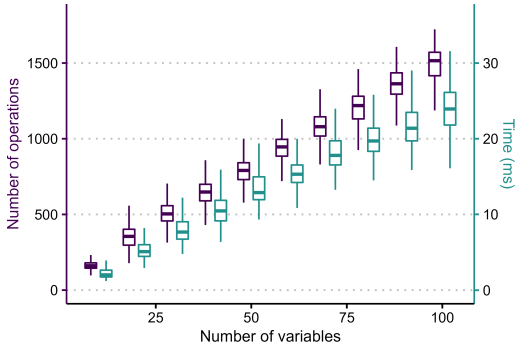
- extension to Fast Causal Inference algorithm *(Spirtes et al. 2000; Zhang 2008)*
- sound & complete under confounders and selection bias
- aimed at handling networks with high-density regions
- ‘live updating’ => true ‘anytime’ algorithm!

Example score-based causal discovery: GPS

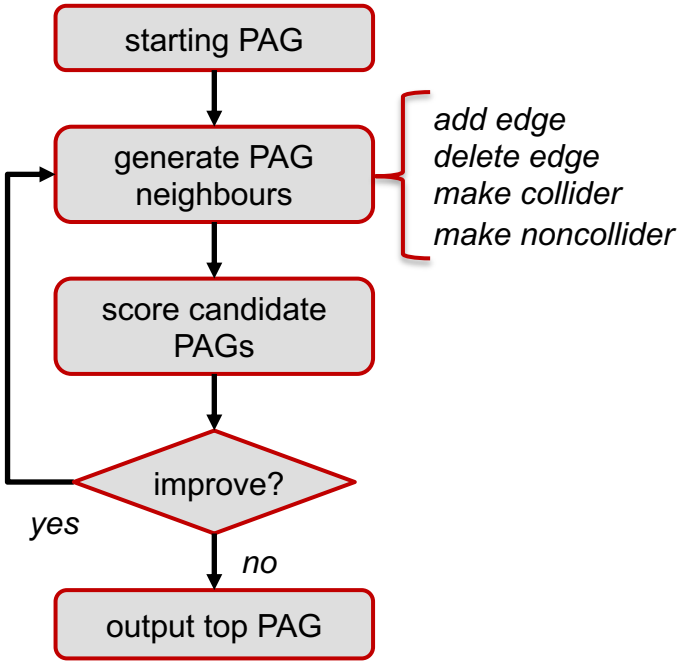
- novel characterization of Markov Equivalence Classes (MECs) for MAGs
- linear time-complexity to establish Markov equivalence for sparse graphs
- basis for *four operators* to move between equivalence classes
- resulting in *Greedy PAG search* (GPS) algorithm for score-based causal discovery
- MAG likelihood score with BIC penalty



MAG with collider and noncollider triples with order



empirical complexity MAG-to-MEC



GPS algorithm

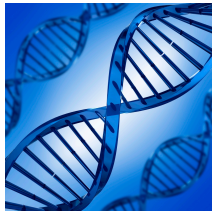
from: Claassen, T., & Bucur, I. G. (2022). Greedy equivalence search in the presence of latent confounders. 38th UAI conference

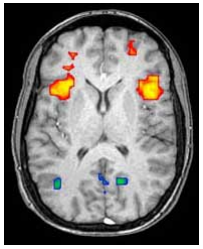
Outline

1. Causal discovery algorithms
2. **Medical data and example applications**
3. The Gap … and the quest to bridge it

The challenge - medical data characteristics

- extremely diverse, non-standard types and distributions,
- highly complex interactions, lots of unknowns


genes


fMRI


questionnaires


social network


family status


medication

comorbid traits


expert diagnosis


conduct disorder / aggression

- low sample sizes, often with lots of missing data (easily >30%)
- data from different studies under different contexts using different metrics
- multiple observations at multiple time points
- lack of ground truth / experimental validation (RTC)

Case - Heritability factors in adult ADHD

ADHD - Attention Deficit Hyperactivity Disorder

- hyperactivity/impulsivity
- inattention (attention deficit, concentration problems)

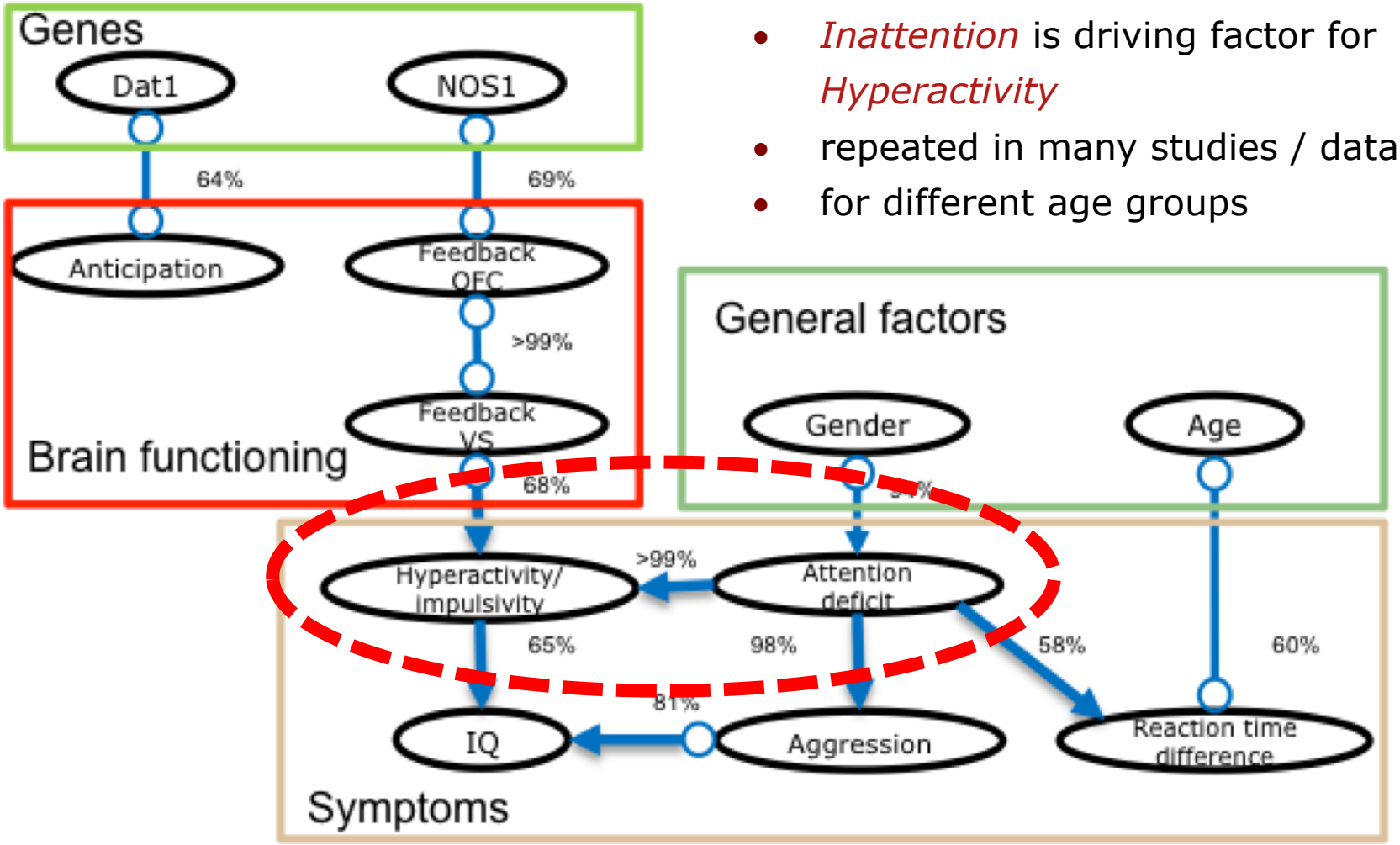
Known

- highly heritable,
- often co-occurring with other traits like ASD (autism)

Unknown

- role hyperactivity/impulsivity vs. inattention in ADHD?
- reasons behind comorbidity?

Case - Heritability factors in adult ADHD

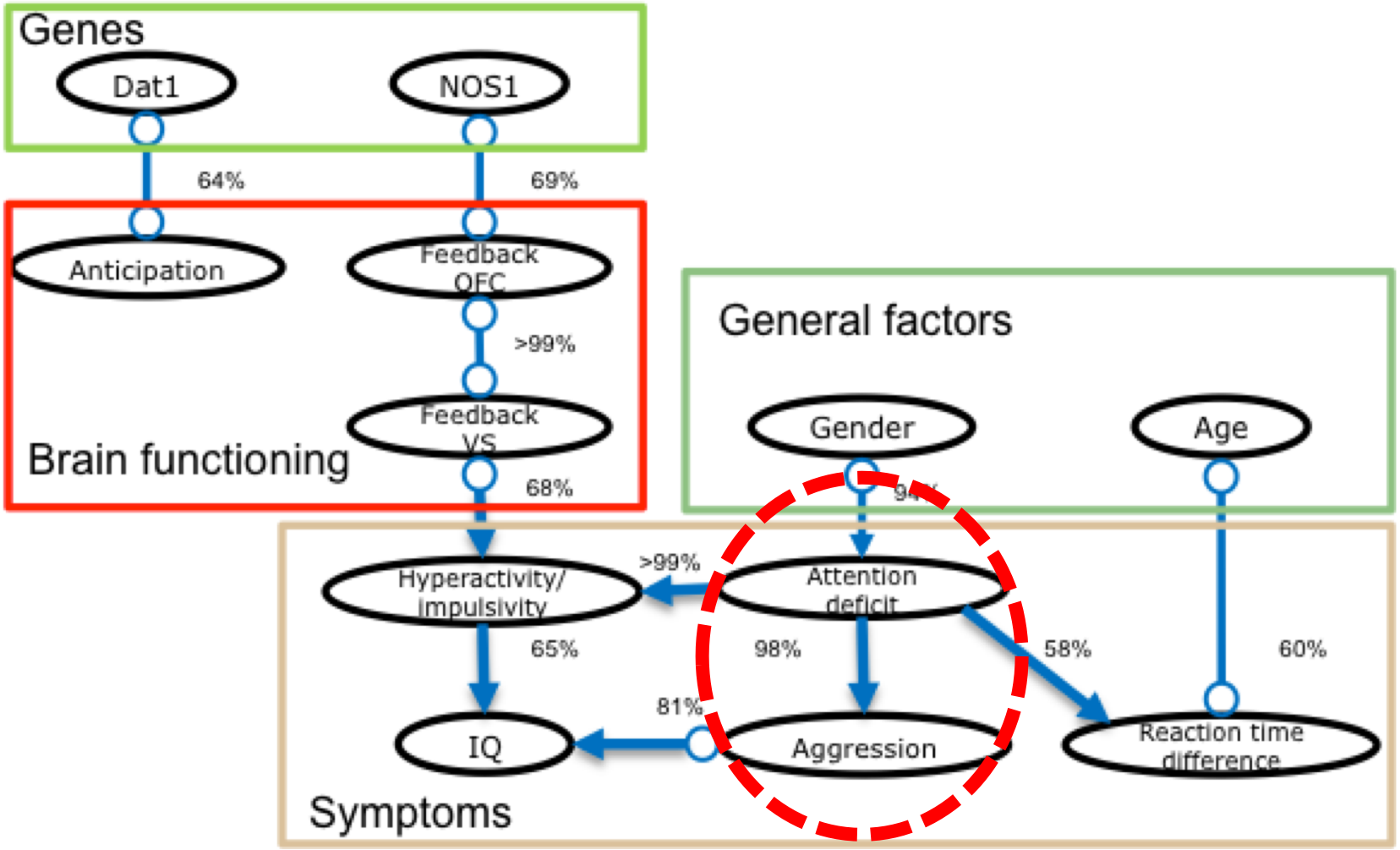


Key find

- *Inattention* is driving factor for *Hyperactivity*
- repeated in many studies / data sets
- for different age groups

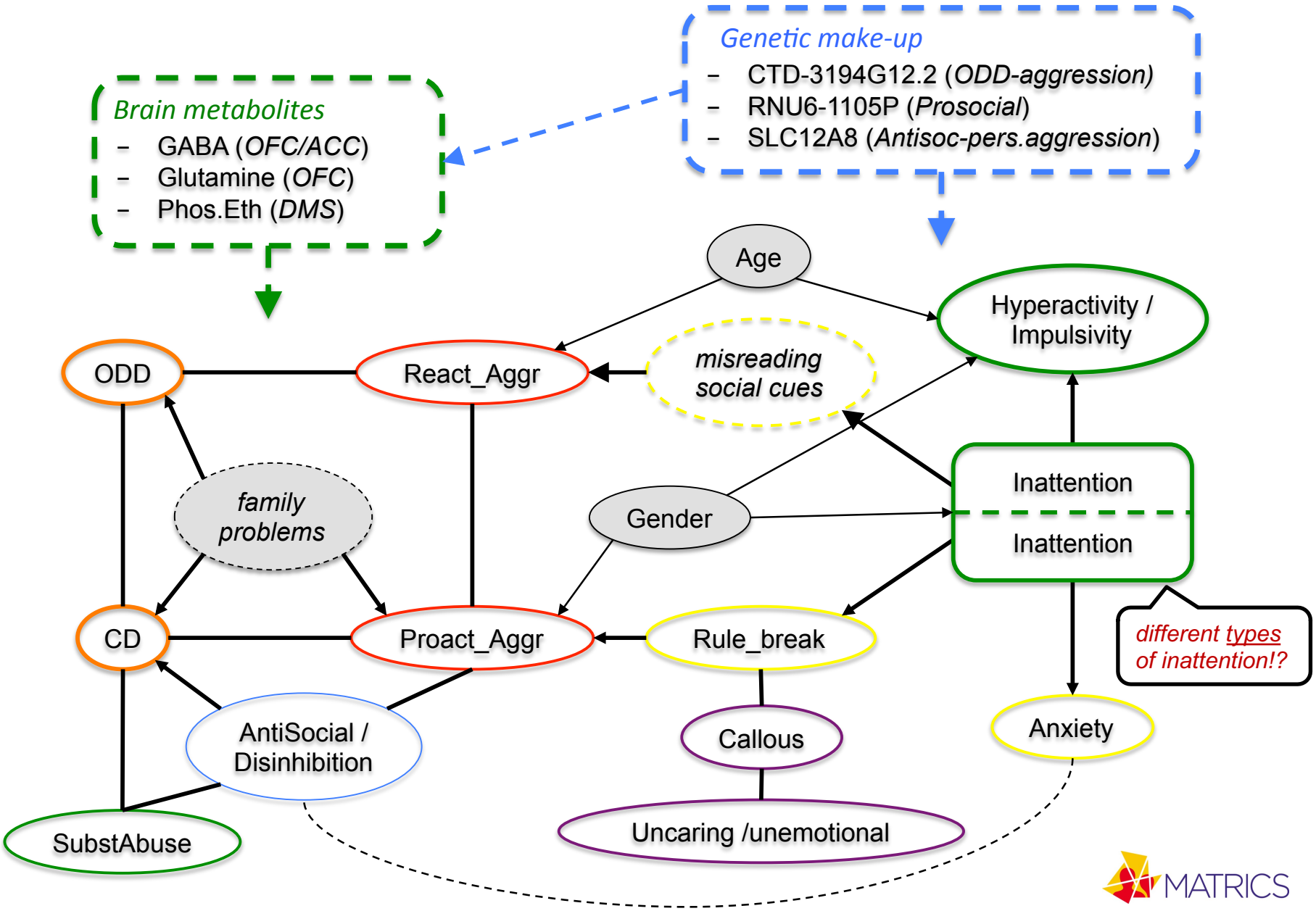
(1) E.Sokolova et al. "Causal discovery in an adult ADHD data set suggests indirect link between DAT1 genetic variants and striatal brain activation during reward processing", American Journal of Medical Genetics Part B: Neuropsychiatric Genetics (2015).

Inattention also as driving factor for aggression?



(1) E.Sokolova et al. "Causal discovery in an adult ADHD data set suggests indirect link between DAT1 genetic variants and striatal brain activation during reward processing", American Journal of Medical Genetics Part B: Neuropsychiatric Genetics (2015).

MATRICS: overview causal model link Inattention - Aggression



Unraveling mechanisms of vascular function with causal discovery

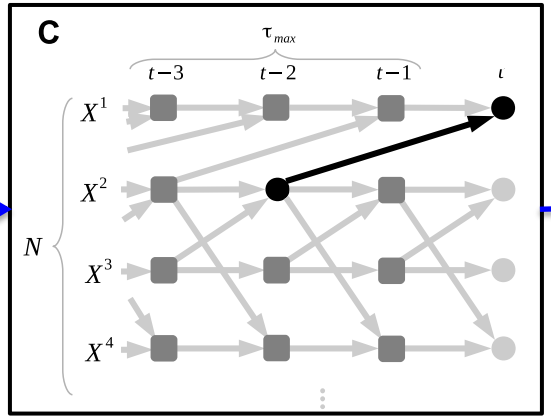
- AI for Health project with Radboudumc
- together with Mirthe van Diepen (PhD)

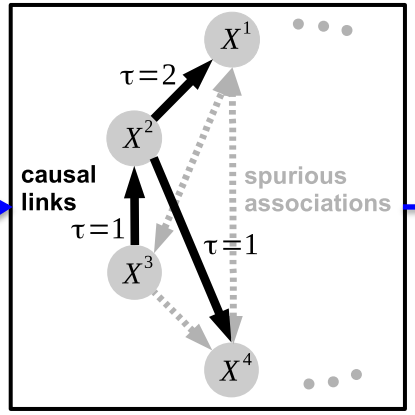
external validation

Vascular function ⟷ brain health outcome

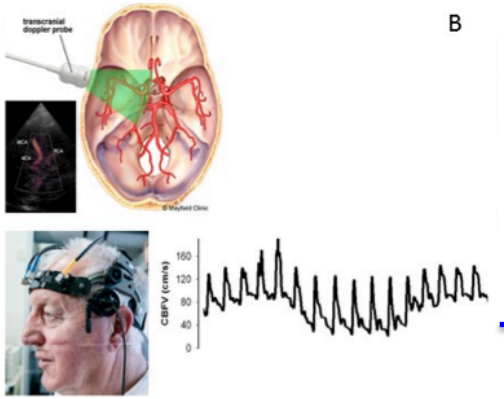
- Alzheimer's disease & vascular surgery
- disentangle confounding variables from key mechanisms
- improve clinical prognosis

non-invasive measure





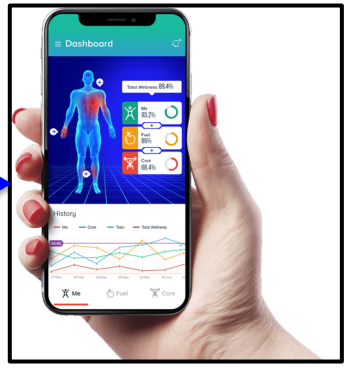
non-stationary causal process reconstruction



age, blood pressure recordings, cerebral flow

Personalized treatment effects

- optimal treatment identification
- expected short-term outcomes
- long term risk assessment



Outline

1. Causal discovery algorithms
2. Medical data and example applications
3. **The Gap … and the quest to bridge it**

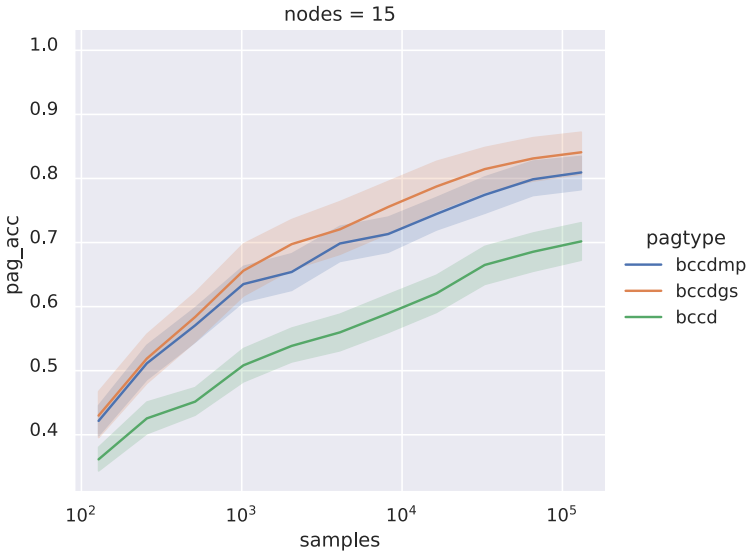
The elephant in the room …

- "So why are your fancy causal methods not in widespread use?"

The Gap

The problem is really, really difficult …

- lots of interest … but available methods unsuitable in practice
- rely on unrealistic assumptions (*no confounding, acyclicity, linear Gaussian*)
- difficulty combining all available data / information
- available statistical tools fall short



Research focus off

- 'ivory tower syndrome'
- we're not answering the right questions (*models changing over time, subtyping*)
- not squeezing out everything we can

Experimental mismatch

- unawareness of impact of certain experimental design choices (*balancing data sets, 'standardizing' variables, etc.*)
- sample sizes based on statistical power for T-tests are too low
- interpretation of causal model can be difficult (e.g. 'SES -> Age' ??)

Call to arms

If we are serious about bringing causal methods to the medical world

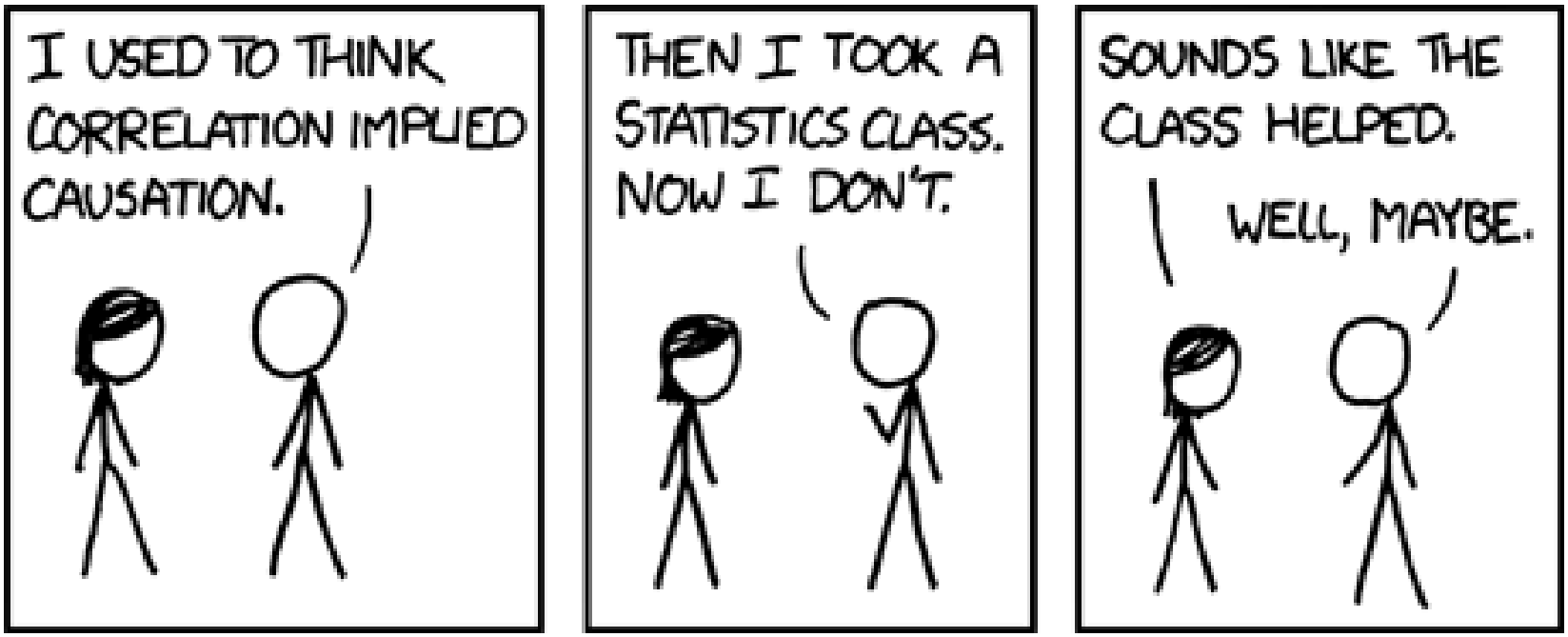
- ban the 'causal sufficiency' assumption (!)
- listen to medical practitioners for their needs
- develop robust, user-friendly algorithms, fully equipped to handle real-world data (*incl. non-linear interactions, missing data, backgr. knowledge*)
- validate causal model predictions

Active research lines

- include feedback / cycles (a.k.a. '*retire the Causal DAG*')
- develop framework for systems changing over time (*ageing, disease progression*)
- create (simulated) realistic benchmark data sets with known ground truth

Interested to join in?

- both medical/health practitioners and/or researchers in causality
- contact me: Tom.Claassen@ru.nl



Randall Munroe, www.xkcd.org

Thank you!