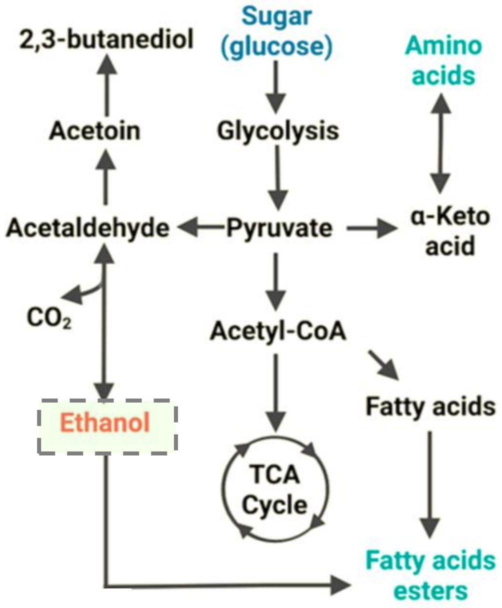
Classical Improvements to Modern Machine Learning

Shiva Kaul <skkaul@cs.cmu.edu>

FERMENTED FOODS

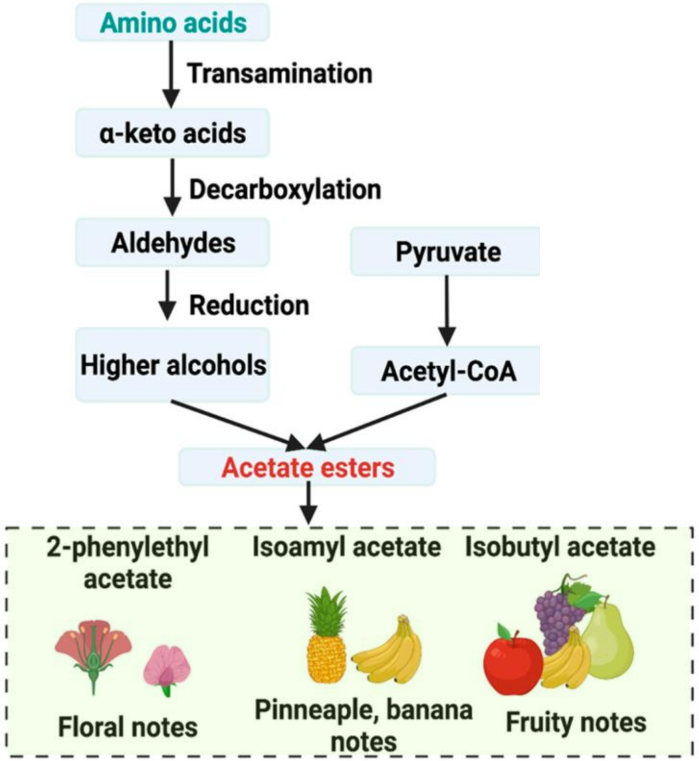
Gutiérrez-Ríos et al. (2022)



Sugar metabolism

WIN-WIN

Inhibit pathogens, add flavor compounds



Protein metabolism

Utilitarian

- Nutritious
- Long-lasting
- Easy to prepare

Hedonic

- Tastes good

Utilitarian

- Nutritious
- Long-lasting
- Easy to prepare

Hedonic

Tastes good •

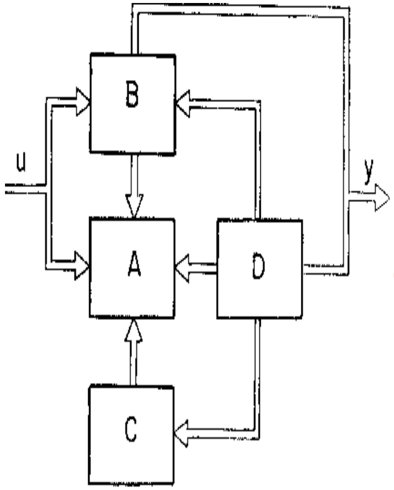
FERMENTED FOODS

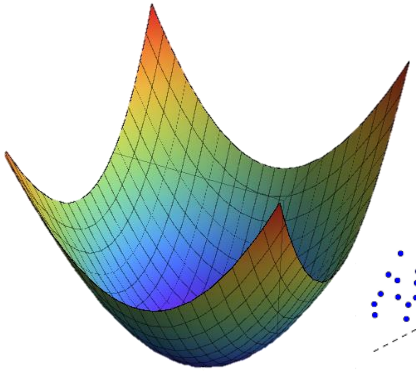
WIN-WIN

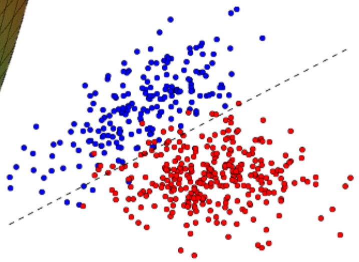
Inhibit pathogens, add flavor compounds

Classical

- Efficient
- Safe (reliable, robust, interpretable)
- Easy to analyze



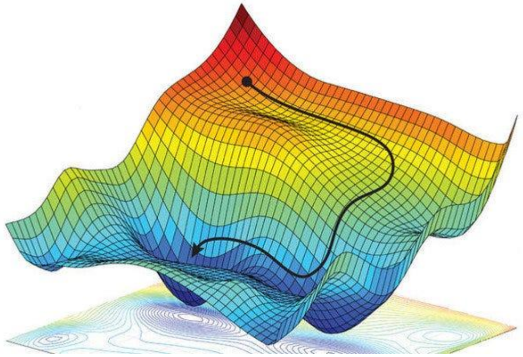


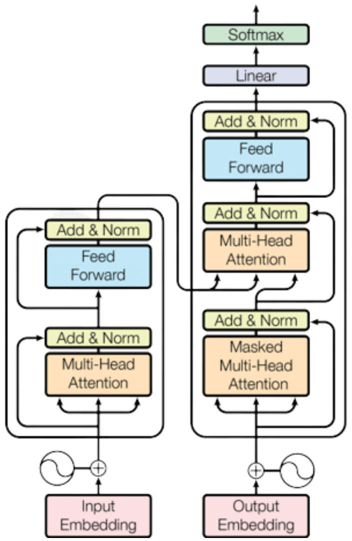


Modern

- Accurate

OpenAI



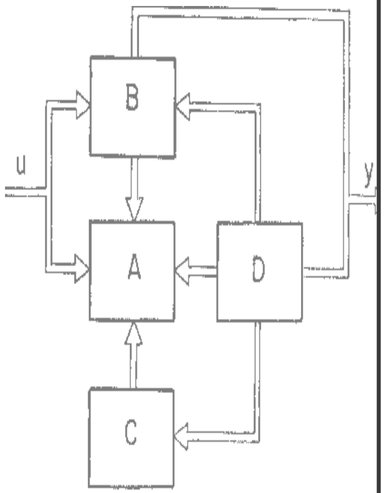


Classical

- Efficient
- Safe (reliable
- Easy to analy

Modern

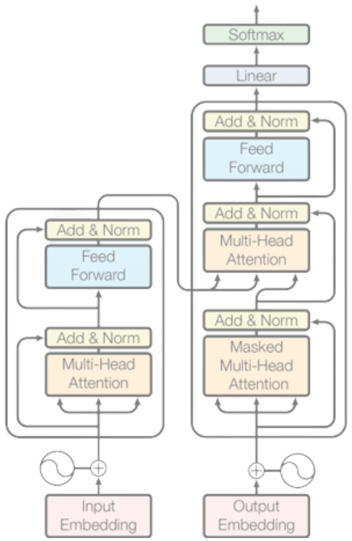
Accurate •



THIS TALK

Syntheses between classical and modern machine learning

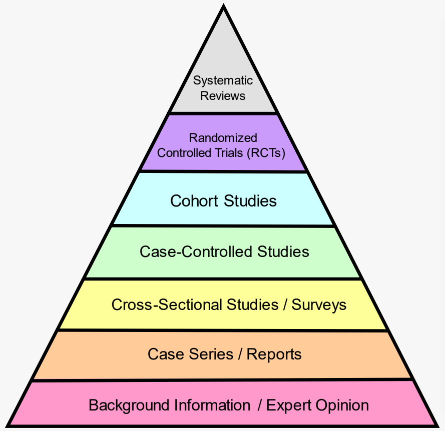
What are the* WIN-WIN *mechanisms?



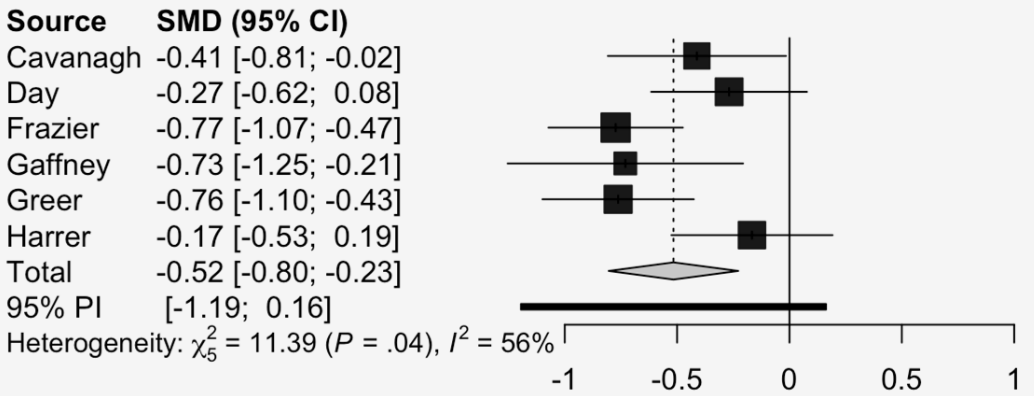
Classical

- Efficient
- Safe (reliable
- Easy to analy

THIS TALK



Systematic Reviews
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies
Cross-Sectional Studies / Surveys
Case Series / Reports
Background Information / Expert Opinion



Source	SMD (95% CI)
Cavanagh	-0.41 [-0.81; -0.02]
Day	-0.27 [-0.62; 0.08]
Frazier	-0.77 [-1.07; -0.47]
Gaffney	-0.73 [-1.25; -0.21]
Greer	-0.76 [-1.10; -0.43]
Harrer	-0.17 [-0.53; 0.19]
Total	-0.52 [-0.80; -0.23]
95% PI	[-1.19; 0.16]

Heterogeneity: $\chi^2_5 = 11.39$ ($P = .04$), $I^2 = 56\%$

Meta-Analysis

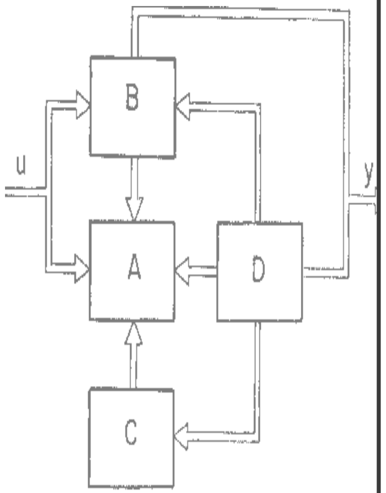
$$f\left(\texttt{What is the most famous cheese in France?}\right) = \texttt{It is arguably Camembert.}$$

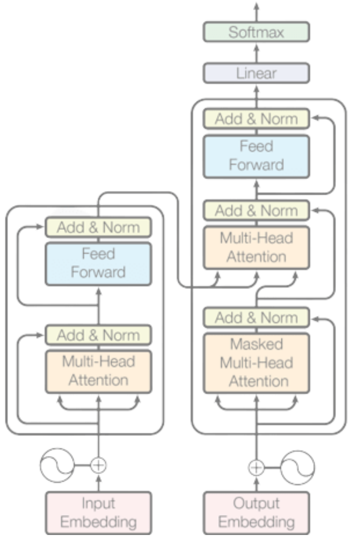
Sequence Models

What are the ***WIN-WIN*** *mechanisms?*

Modern

- Accurate





CLAIM

Meta-analysis is an interesting machine learning problem.

- Use large datasets to dramatically improve causal inferences and patient outcomes
- Scientific question-answering is an interesting unsolved problem for LLMs
- A beautiful statistical problem which exposes key challenges in uncertainty quantification

Meta-Analysis

Effect

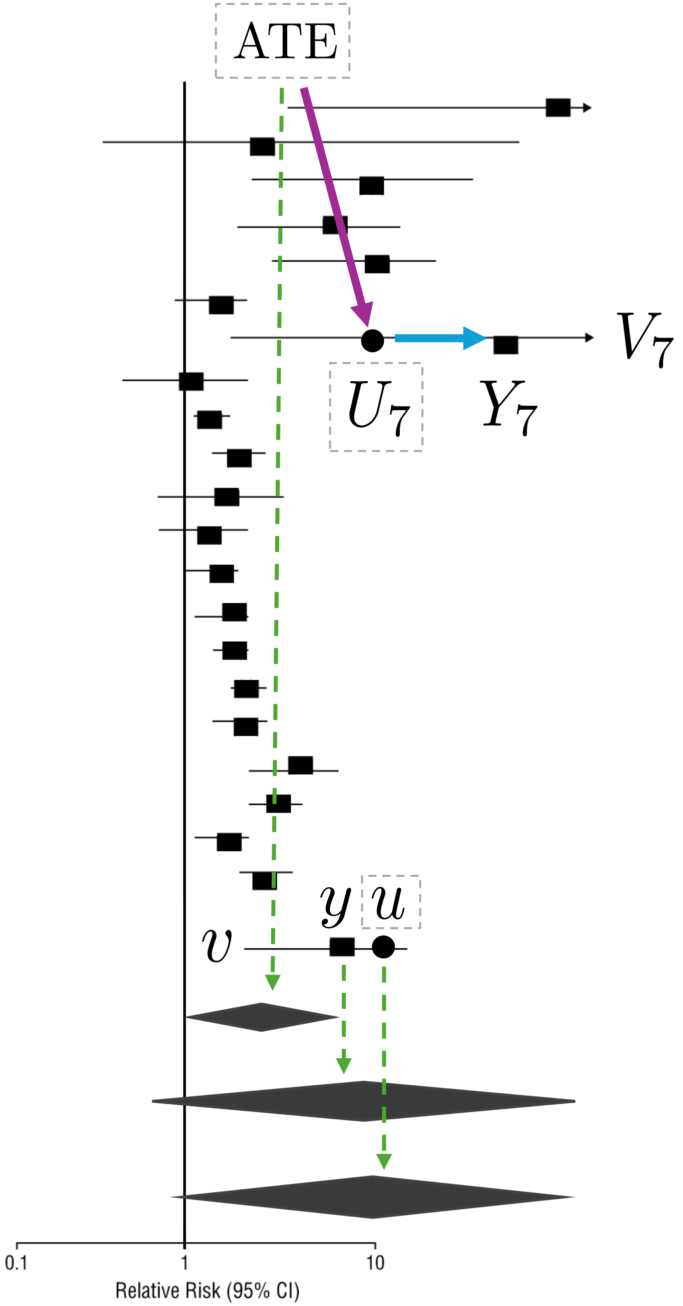
$$U_i = \text{ATE} + N(0, \nu)$$

between-trial heterogeneity

$$Y_i = U_i + N(0, V_i)$$

Observed effect

within-trial variance



Letelier et al. (2003)

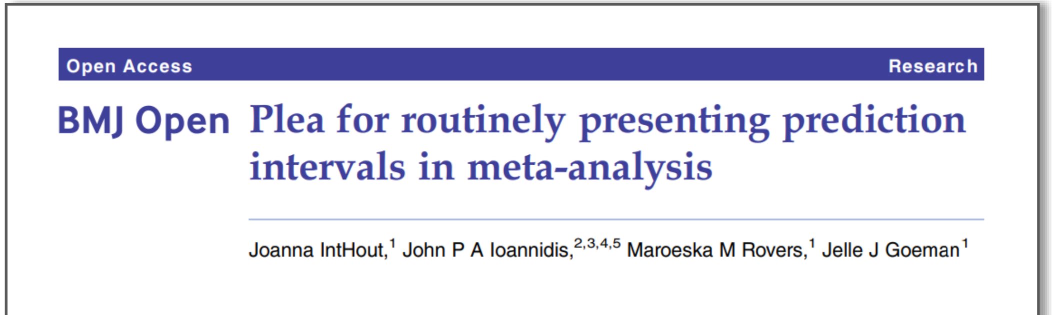
Galperin et al[29] (2000) 33.7 (2.08-546.00) 95
Bianconi et al[28] (2000) 2.04 (0.19-22.00) 83
Villani et al[11] (2000) 4.75 (1.60-14.00) 120
Hohnloser et al[3] (2000) 3.13 (1.5-6.70) 203
Natale et al[25] (2000) 5.12 (2.60-10.00) 85
Cowan et al[16] (1986) 1.11 (0.78-1.58) 34
Noc et al[17] (1990) 18.00 (1.17-276.00) 24
Capucci et al[18] (1992) 0.77 (0.37-1.62) 40
Cochrane et al[19] (1994) 1.15 (0.91-1.44) 30
Hou et al[21] (1995) 1.29 (0.97-1.72) 39
Kondili et al[22] (1995) 1.33 (0.71-2.47) 42
Donovan et al[20] (1994) 1.05 (0.69-1.60) 64
Galve et al[23] (1996) 1.13 (0.84-1.52) 100
Kontoyannis et al[24] (1998) 1.42 (1.08-1.85) 42
Bellandi et al[26] (1999) 1.41 (1.15-1.72) 120
Kochiadakis et al[12] (1999) 1.46 (1.19-1.78) 204
Cotter et al[27] (1999) 1.43 (1.15-1.8) 100
Peuhkurinen et al[30] (2000) 2.45 (1.49-4.02) 62
Vardas et al[31] (2000) 2.01 (1.55-2.6) 208
Joseph and Ward[32] (2000) 1.32 (095-1.80) 75
Cybulski et al[33] (2001) 1.87 (1.37-2.55) 160

[Future]

95% CI

95% PI for y

95% PI for u



Open Access Research

BMJ Open **Plea for routinely presenting prediction intervals in meta-analysis**

Joanna IntHout,[1] John P A Ioannidis,[2,3,4,5] Maroeska M Rovers,[1] Jelle J Goeman[1]

Meta-Analysis

Features *Effect* *Variance*

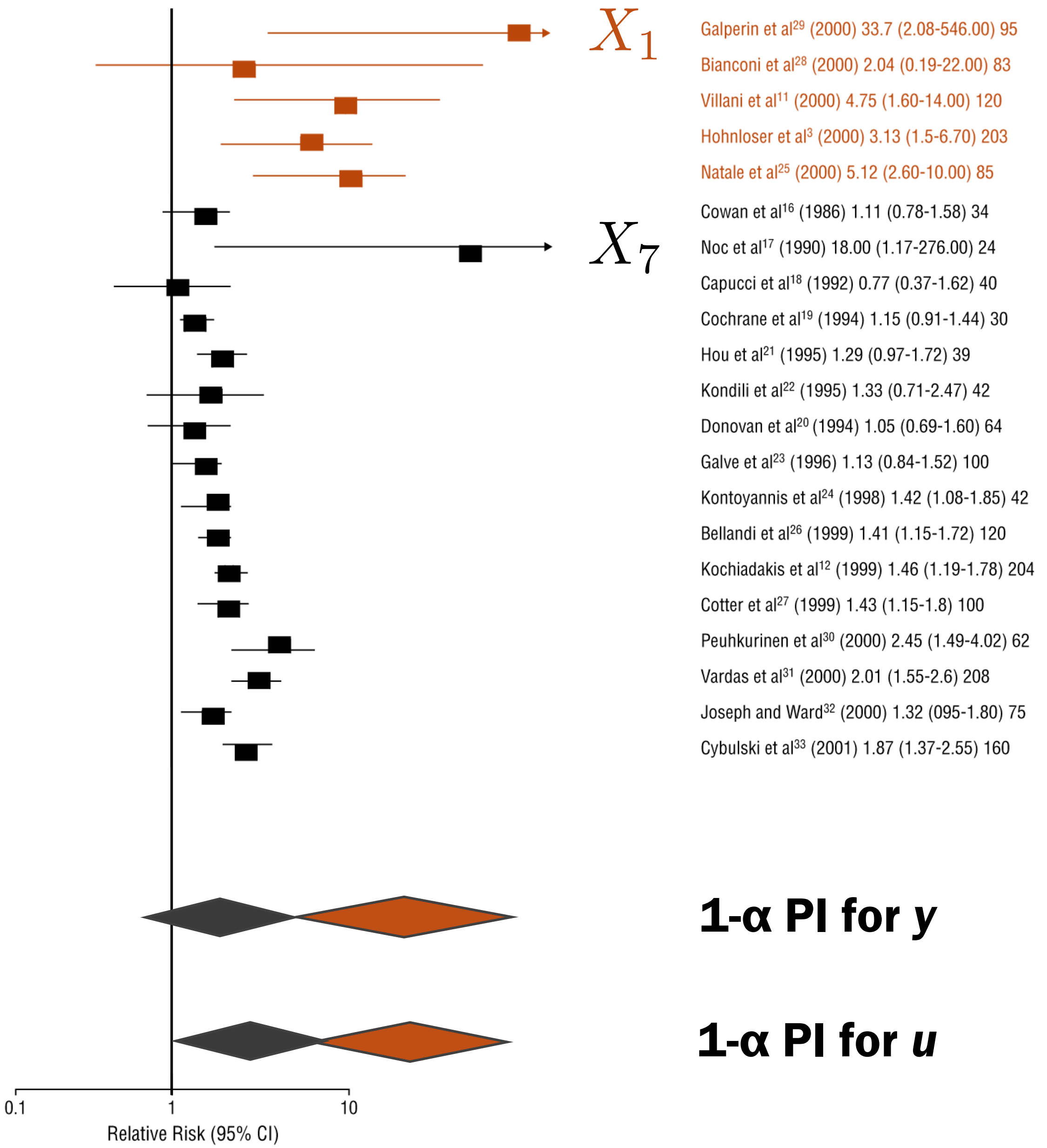
$$(X_i,\ U_i,\ V_i)\ \sim\ \mathbb{P}$$

$$Y_i = U_i + N(0, V_i)$$

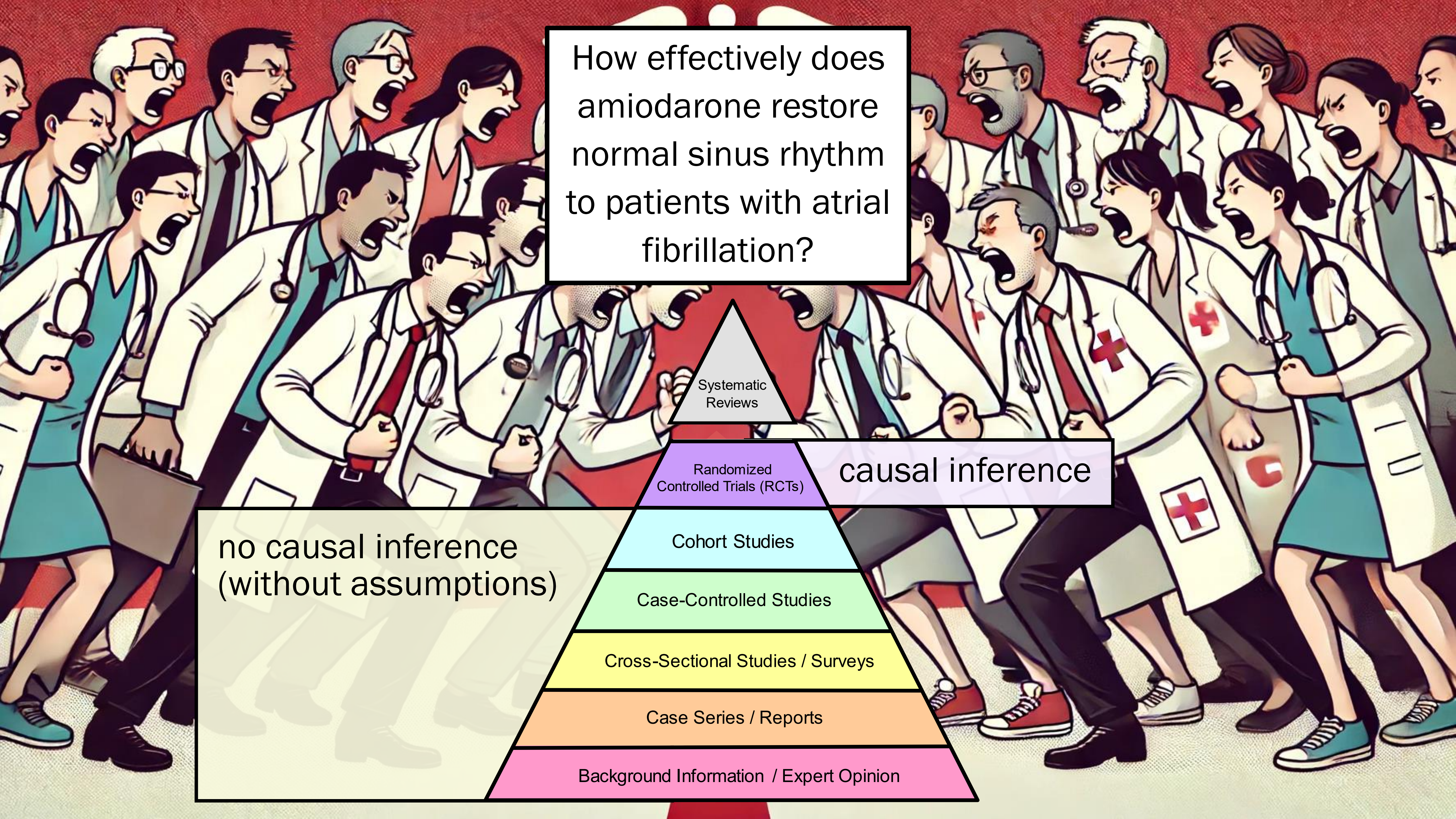
Observed effect

$$1-\alpha \leq \mathbb{P}(y \in C(x,v))$$

$$1-\alpha \leq \mathbb{P}(u \in C(x))$$



How effectively does amiodarone restore normal sinus rhythm to patients with atrial fibrillation?



Systematic Reviews

Randomized Controlled Trials (RCTs)

Cohort Studies

Case-Controlled Studies

Cross-Sectional Studies / Surveys

Case Series / Reports

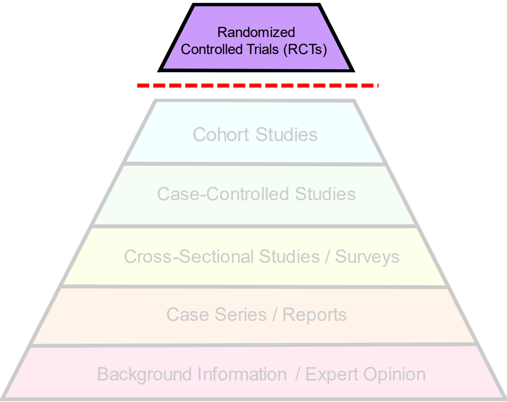
Background Information / Expert Opinion

causal inference

no causal inference (without assumptions)

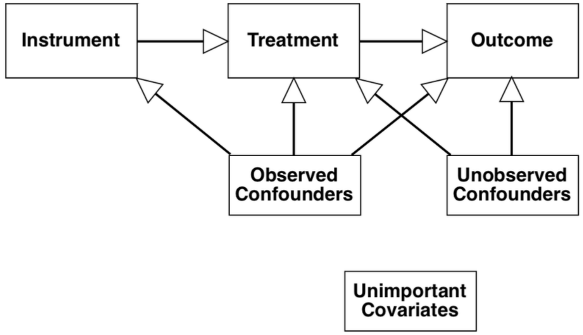
Trusted data

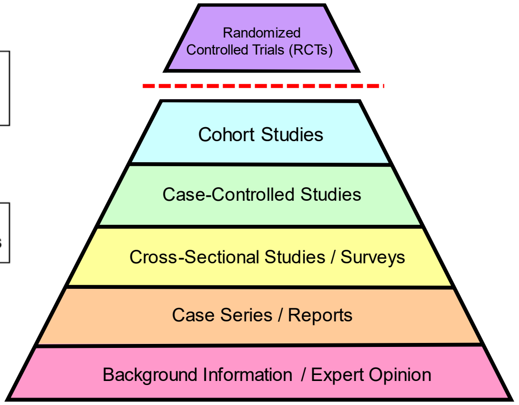
- Rigorous and unbiased
- Loose predictions


Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies
Cross-Sectional Studies / Surveys
Case Series / Reports
Background Information / Expert Opinion

Untrusted data

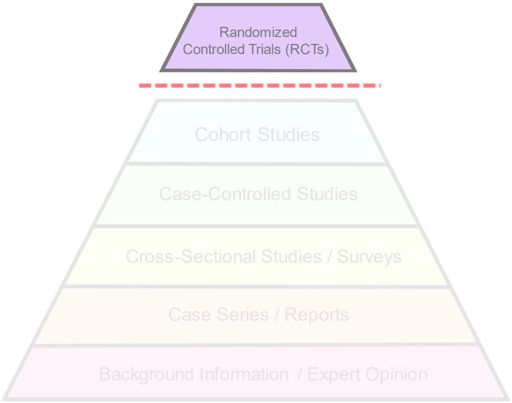
- Need strong assumptions
- Tight predictions



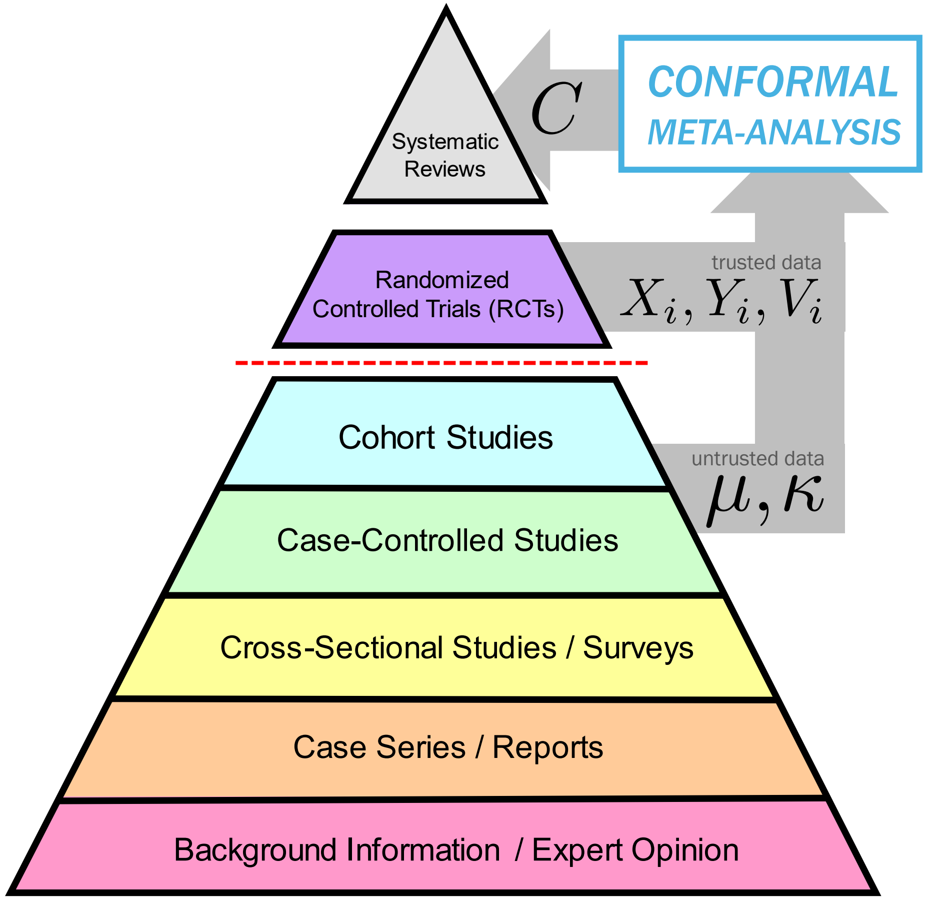

Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies
Cross-Sectional Studies / Surveys
Case Series / Reports
Background Information / Expert Opinion

Trusted data

- Rigorous and unbiased
- Loose predictions


Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies
Cross-Sectional Studies / Surveys
Case Series / Reports
Background Information / Expert Opinion

CONFORMAL META-ANALYSIS

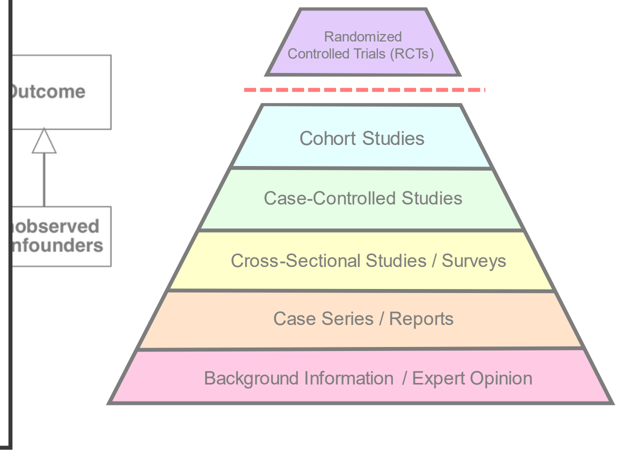


WIN-WIN

Assumption-free inclusion of untrusted prior

Untrusted data

- eed strong assumptions
- Tight predictions


utcome
observed
nfounders
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies
Cross-Sectional Studies / Surveys
Case Series / Reports
Background Information / Expert Opinion

$$0.90 \;\le\; \mathbb{P}\,(r^* \text{ among lowest 20 of } R_i)$$

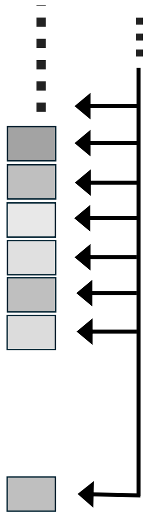
$$C(x,v) = \{y : r \text{ among lowest 20 of } R_i\}$$

→

$$0.90 \le \mathbb{P}(y^* \in C(x,v))$$

Conformal Prediction

$n = 21$

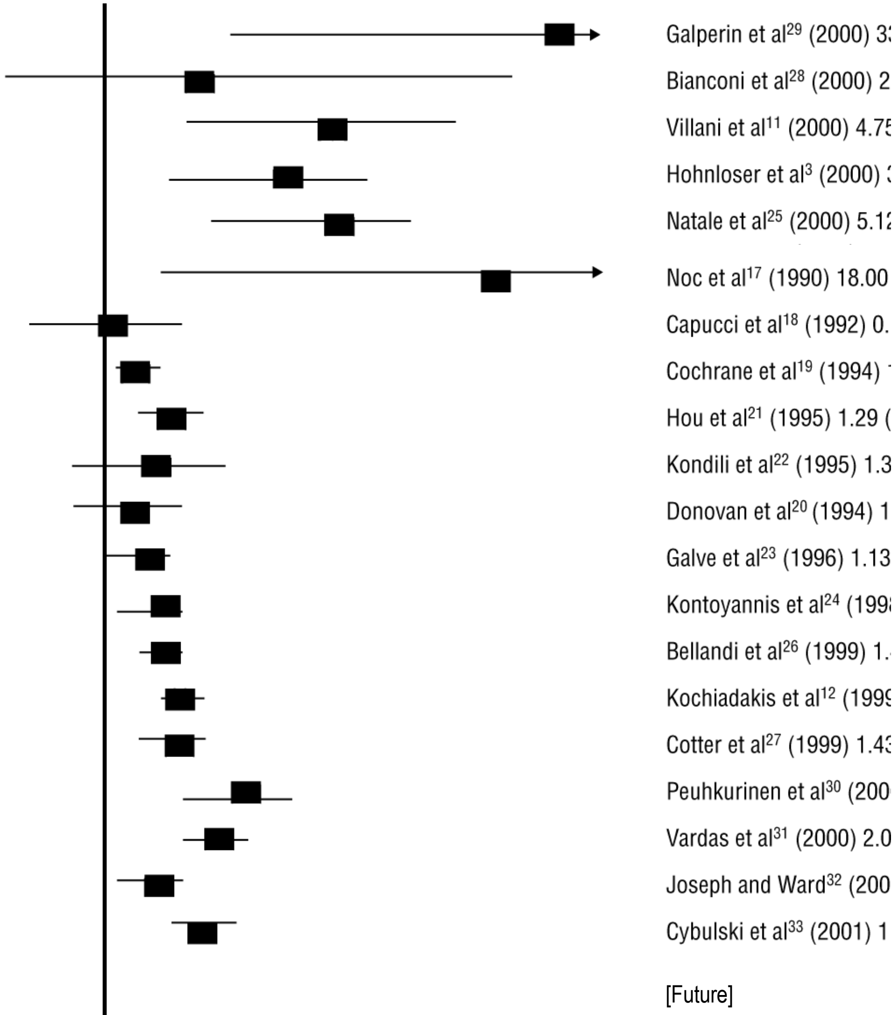


exchangeable, so rank of r^* is uniform among R_i

Past (X_i, Y_i, V_i) has residual R_i

Future (x, y^*, v) has residual r^*

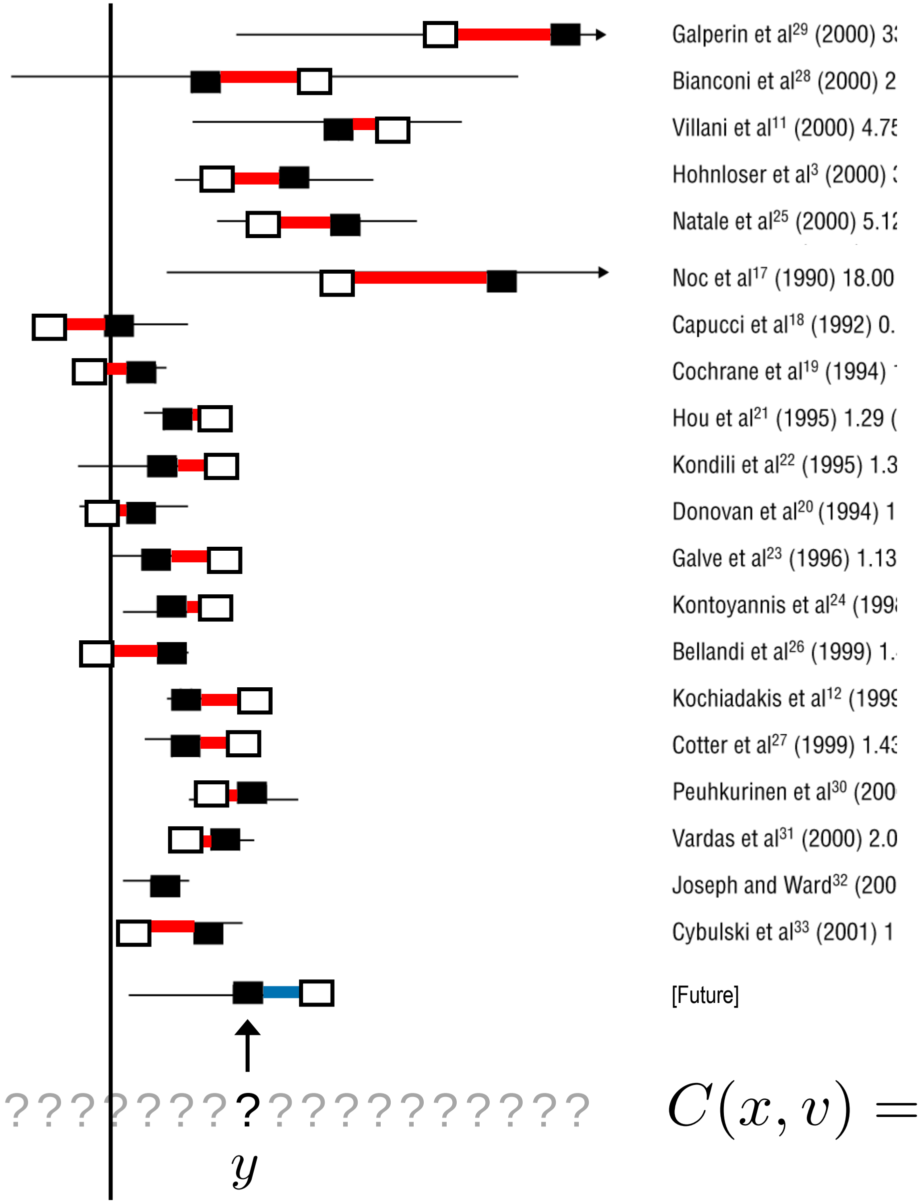
(Hypothetical) Future (x, y, v) has residual r



90% PI for y ???????????????????? $C(x, v) = \{y : r \text{ among lowest } 20 \text{ of } R_i\}$

train on everything for exchangeability

Prior μ, κ —— $[Y; y]$ — $[V; v]$ —— $[X; x]$ $\rightarrow$ **Posterior** $\hat{\mu}, \hat{\kappa}$ $\rightarrow$ **Residuals** R_i, r

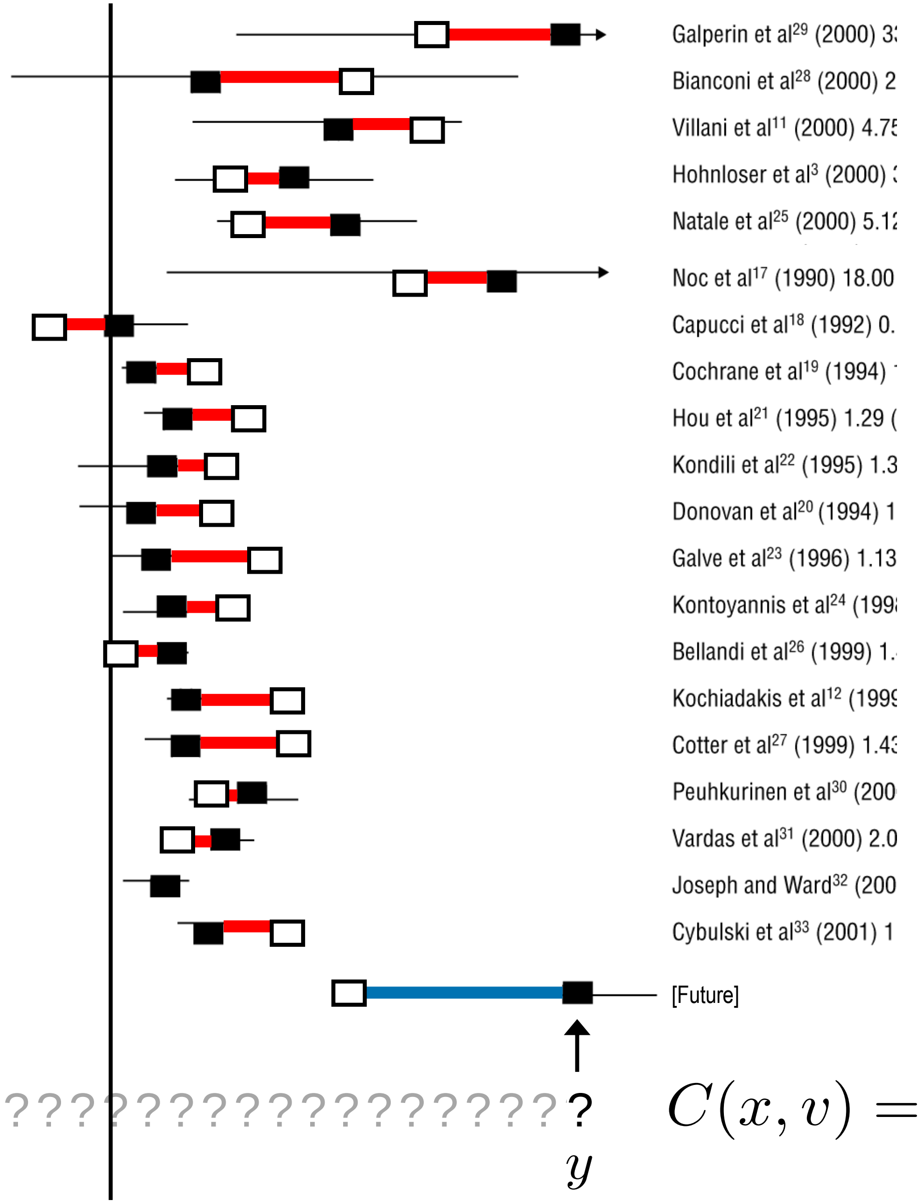


90% PI for y

$$C(x, v) = \{y : r \text{ among lowest 20 of } R_i\}$$

train on everything for exchangeability

Prior μ, κ —— $[Y; y]$ —— $[V; v]$ —— $[X; x]$ → **Posterior** $\hat{\mu}, \hat{\kappa}$ → **Residuals** R_i, r

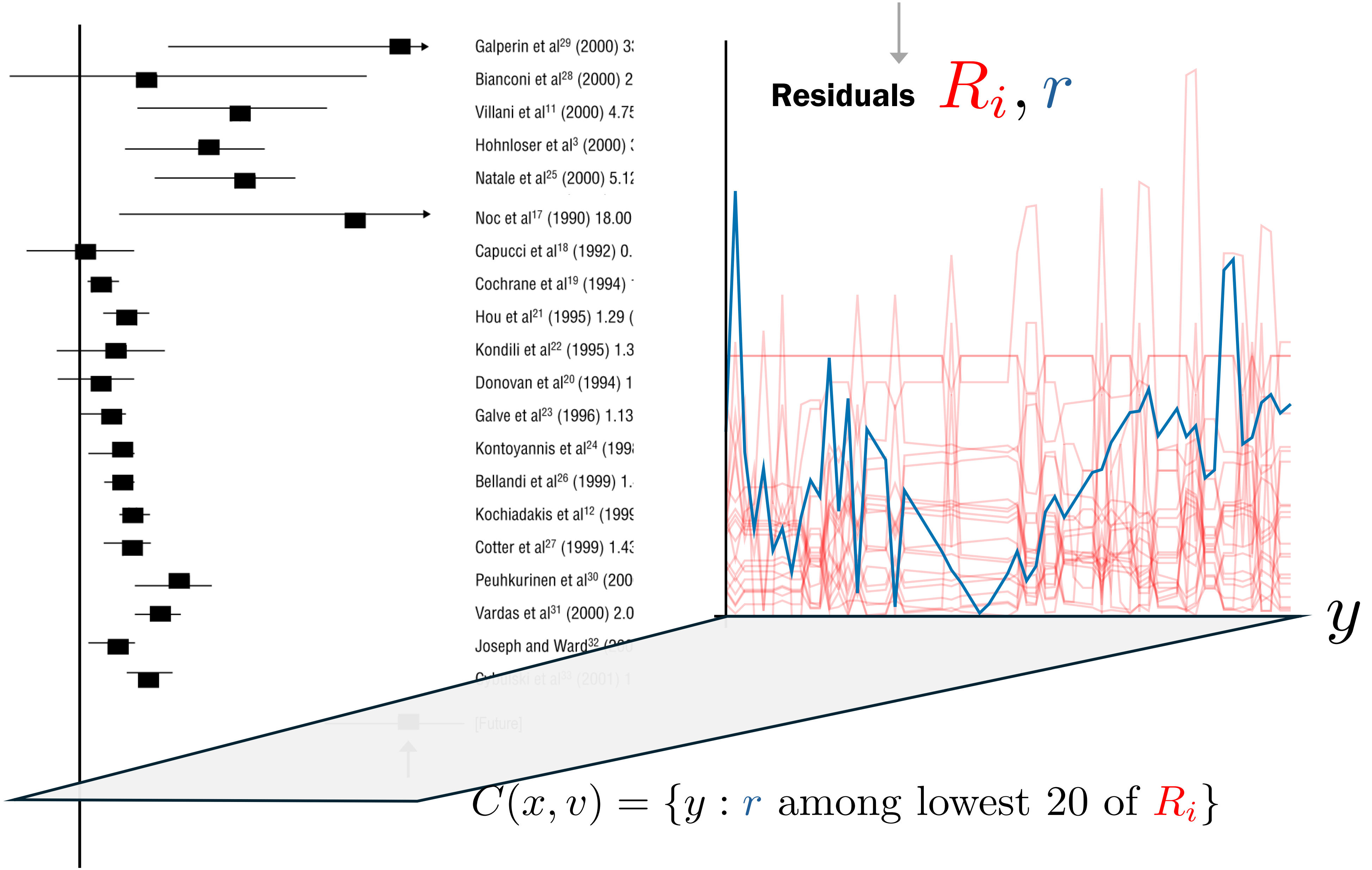


90% PI for y

$$C(x, v) = \{y : r \text{ among lowest 20 of } R_i\}$$

train on everything for exchangeability

Prior μ, κ —— $[Y;y]$ —— $[V;v]$ —— $[X;x]$ $\longrightarrow$ **Posterior** $\hat{\mu}, \hat{\kappa}$



Galperin et al[29] (2000) 3
Bianconi et al[28] (2000) 2
Villani et al[11] (2000) 4.75
Hohnloser et al[3] (2000)
Natale et al[25] (2000) 5.12
Noc et al[17] (1990) 18.00
Capucci et al[18] (1992) 0.
Cochrane et al[19] (1994)
Hou et al[21] (1995) 1.29 (
Kondili et al[22] (1995) 1.3
Donovan et al[20] (1994) 1
Galve et al[23] (1996) 1.13
Kontoyannis et al[24] (199
Bellandi et al[26] (1999) 1.
Kochiadakis et al[12] (199
Cotter et al[27] (1999) 1.43
Peuhkurinen et al[30] (200
Vardas et al[31] (2000) 2.0
Joseph and Ward[32] (200
Cybulski et al[33] (2001) 1
[Future]
y

Residuals R_i, r

90% PI for y

$$C(x,v) = \{y : r \text{ among lowest 20 of } R_i\}$$

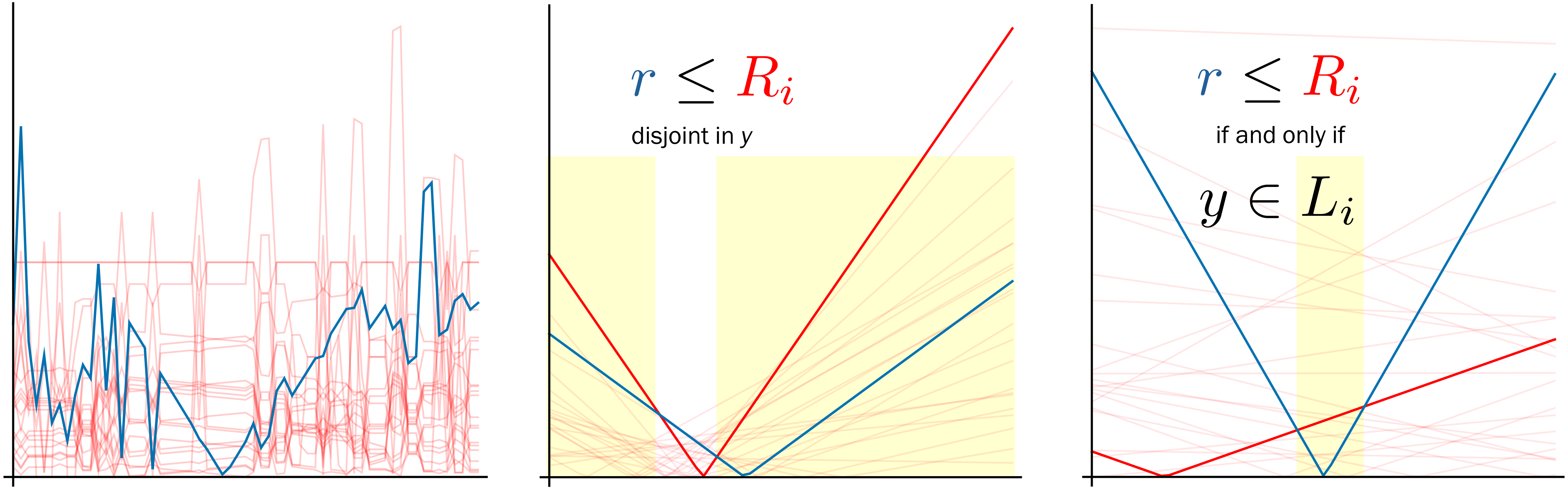
CHALLENGES

1. Full conformal prediction is intractable

   (*n* is small, so cannot split the data)

2. Also want interval for *u*, not just $y = N(u,v)$

Kaul and Gordon (2024)



Focus on **linear smoothers**

like kernel ridge regression (KRR)

$$R_i = \ldots|A_i y + B_i|\ldots \quad r = \ldots|ay + b|\ldots$$

residuals are convex in *y*

Ensure **idiocentricity**

changing *y* affects *r* more than any R_i

$$|a| > |A_i| \quad \Longleftarrow \quad \lambda \geq \max_x \kappa(x, x)$$

for linear smoothers

easy to ensure for KRR

*Tolerate **approximation**

$$C(x, v) \subseteq \left[\begin{array}{l} \text{2nd lowest} \\ \text{left end of } L_i \end{array}, \begin{array}{r} \text{2nd highest} \\ \text{right end of } L_i \end{array} \right]$$

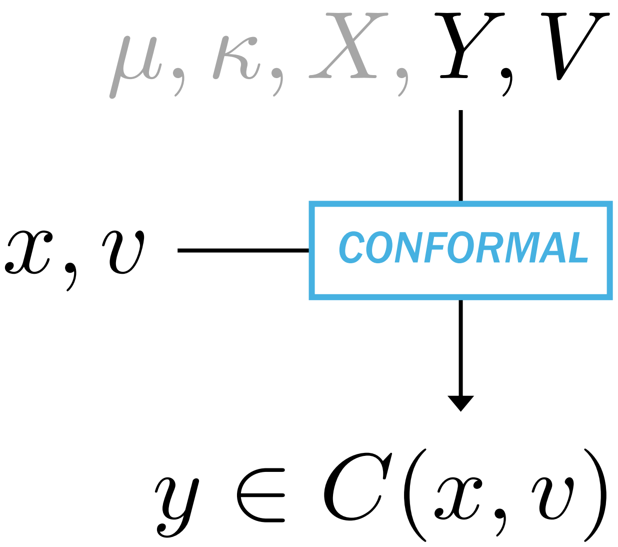
CHALLENGES

1. Full conformal prediction is intractable

   ...but not for idiocentric linear smoothers.

2. Also want interval for *u*, not just $y = N(u,v)$

Kaul and Gordon (2024)



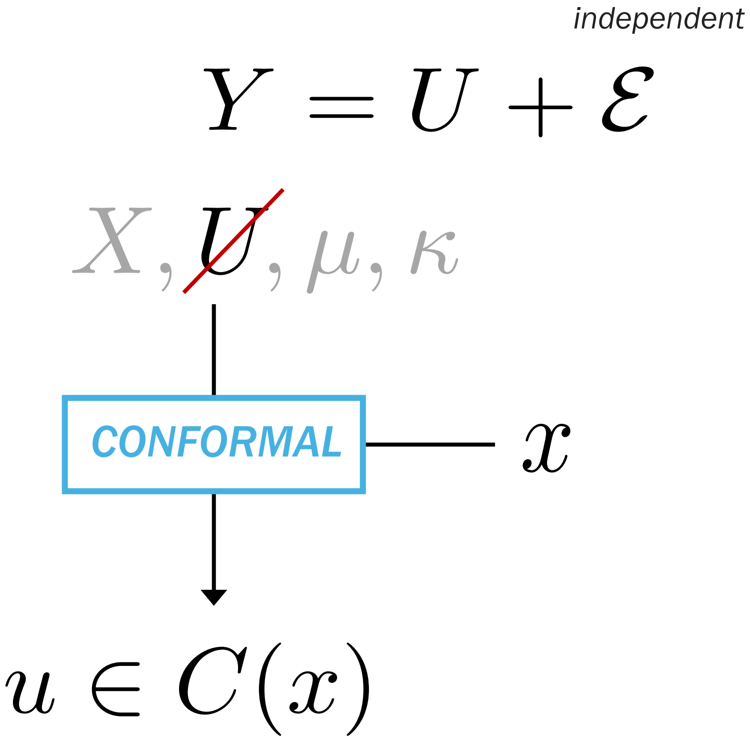
μ, κ, X, Y, V

x, v

CONFORMAL

$y \in C(x, v)$

VS



independent

$$Y = U + \mathcal{E}$$

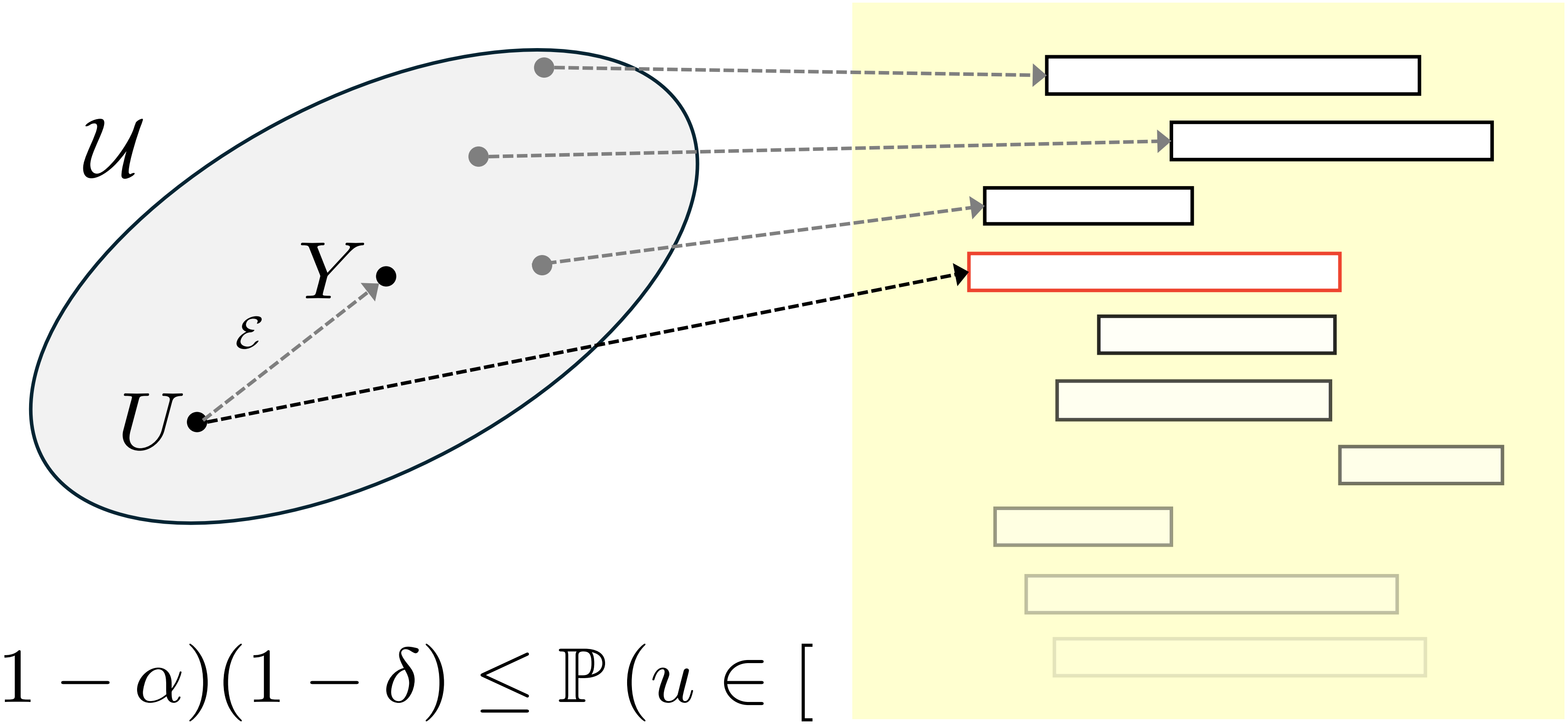
$X, \cancel{U}, \mu, \kappa$

CONFORMAL

x

$u \in C(x)$

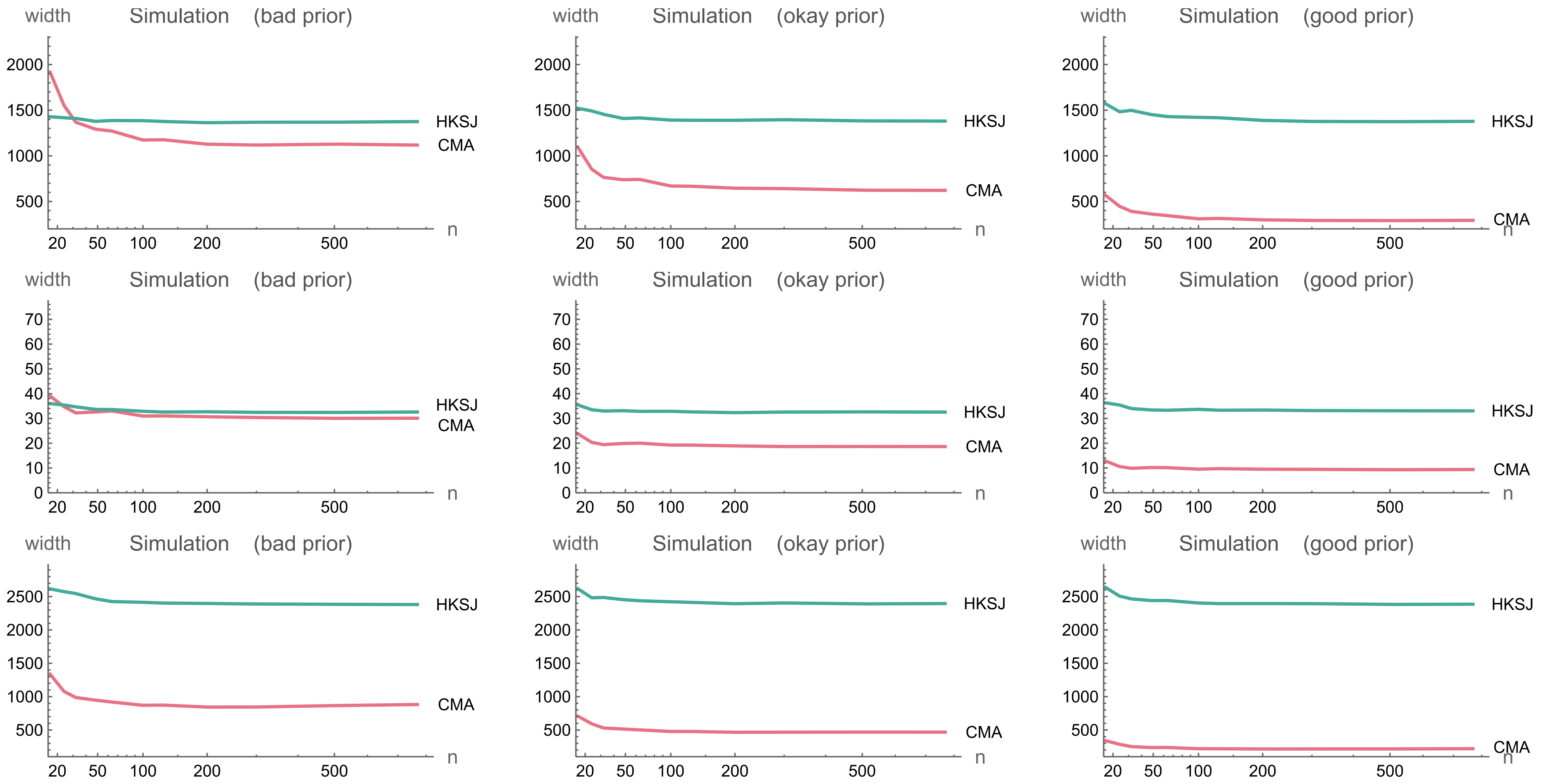
$$\mathbb{P}_{\mathcal{E}}(U \in \mathcal{U}) \geq 1 - \delta \qquad \bigcup \{C(x; \widehat{U}) : \widehat{U} \in \mathcal{U}\}$$



$$(1 - \alpha)(1 - \delta) \leq \mathbb{P}\,(u \in [\qquad])$$

Exploit independence of noise $\mathcal{E}$

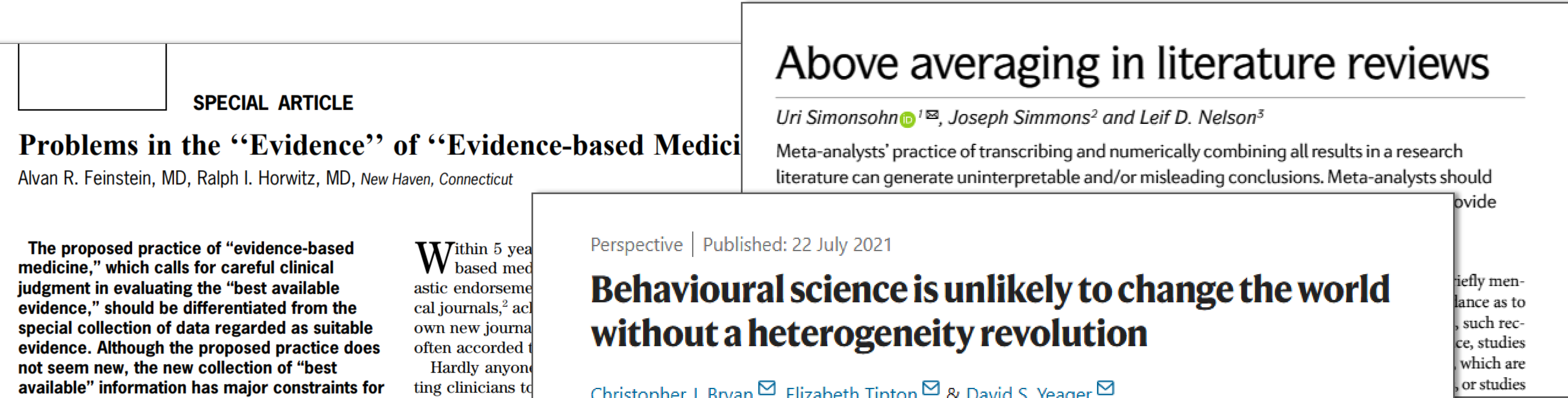
Idiocentricity → tightly bound outer interval



width
Simulation (bad prior)
2000
1500
1000
500
20
50
100
200
500
n
HKSJ
CMA
width
Simulation (okay prior)
2000
1500
1000
500
20
50
100
200
500
n
HKSJ
CMA
width
Simulation (good prior)
2000
1500
1000
500
20
50
100
200
500
n
HKSJ
CMA
width
Simulation (bad prior)
70
60
50
40
30
20
10
0
20
50
100
200
500
n
HKSJ
CMA
width
Simulation (okay prior)
70
60
50
40
30
20
10
0
20
50
100
200
500
n
HKSJ
CMA
width
Simulation (good prior)
70
60
50
40
30
20
10
0
20
50
100
200
500
n
HKSJ
CMA
width
Simulation (bad prior)
2500
2000
1500
1000
500
20
50
100
200
500
n
HKSJ
CMA
width
Simulation (okay prior)
2500
2000
1500
1000
500
20
50
100
200
500
n
HKSJ
CMA
width
Simulation (good prior)
2500
2000
1500
1000
500
20
50
100
200
500
n
HKSJ
CMA

CONTRIBUTIONS

- Formulated meta-analysis as an interesting machine learning problem
- Simplified full conformal prediction for idiocentric linear smoothers
- Addressed statistical/algorithmic challenges in handling noise



SPECIAL ARTICLE

Problems in the “Evidence” of “Evidence-based Medici

Alvan R. Feinstein, MD, Ralph I. Horwitz, MD, *New Haven, Connecticut*

The proposed practice of “evidence-based medicine,” which calls for careful clinical judgment in evaluating the “best available evidence,” should be differentiated from the special collection of data regarded as suitable evidence. Although the proposed practice does not seem new, the new collection of “best available” information has major constraints for

Within 5 yea
based med
astic endorseme
cal journals,[2] ac
own new journa
often accorded t
Hardly anyon
ting clinicians t

Above averaging in literature reviews

Uri Simonsohn[1], Joseph Simmons[2] and Leif D. Nelson[3]

Meta-analysts’ practice of transcribing and numerically combining all results in a research literature can generate uninterpretable and/or misleading conclusions. Meta-analysts should
ovide

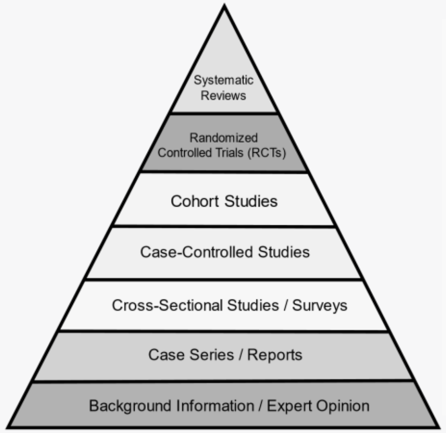
iefly men-
lance as to
, such rec-
ce, studies
which are
, or studies

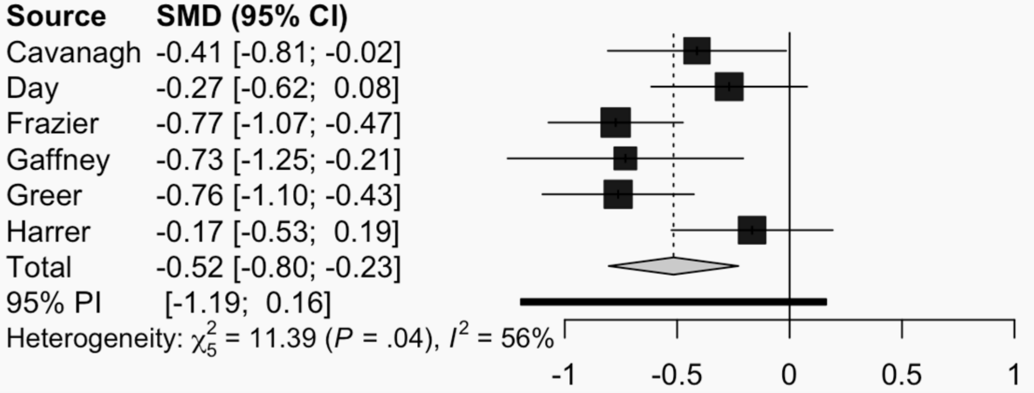
Perspective | Published: 22 July 2021

Behavioural science is unlikely to change the world without a heterogeneity revolution

Christopher J. Bryan, Elizabeth Tipton & David S. Yeager

THIS TALK


Systematic Reviews
Randomized Controlled Trials (RCTs)
Cohort Studies
Case-Controlled Studies
Cross-Sectional Studies / Surveys
Case Series / Reports
Background Information / Expert Opinion



Source	**SMD (95% CI)**
Cavanagh	-0.41 [-0.81; -0.02]
Day	-0.27 [-0.62; 0.08]
Frazier	-0.77 [-1.07; -0.47]
Gaffney	-0.73 [-1.25; -0.21]
Greer	-0.76 [-1.10; -0.43]
Harrer	-0.17 [-0.53; 0.19]
Total	-0.52 [-0.80; -0.23]
95% PI	[-1.19; 0.16]

Heterogeneity: $\chi^2_5 = 11.39$ (P = .04), I^2 = 56%

Meta-Analysis

$f($ What is the most famous cheese in France? $) =$ It is arguably Camembert.

Sequence Models

Linear

- Efficient (*O(T)* memory) and
- Fast (*O(log T)* parallel time via scans)
- Unexpressive

$$h_t = A_t h_{t-1} + B_t x_t$$

(Time-varying) Linear dynamical system

Nonlinear

- Inefficient (*O(T^2)* memory) or
- Slow (*O(T)* parallel time)
- Expressive

$$h = \psi(Q(x)K(x))V(x)$$

Attention

$$\overset{(n)}{h_t} = \rho(\overset{(n,n)}{A_t}\overset{(n)}{h_{t-1}} + \overset{(n,d)}{B_t}\overset{(d)}{x_t})$$

Recurrent neural network

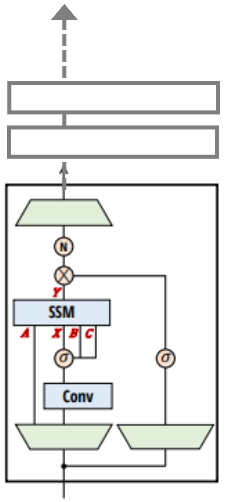
Linear

- Efficient ($O(T)$ mem
- Fast ($O(\log T)$ para
- Unexpressive

Nonlinear

- ($O(T^2)$) memory) or
- $O(T)$ parallel time)
- Expressive

STATE-SPACE MODELS



INTERSPERSED NONLINEARITIES

$$\overset{(n)}{h_t} = \overset{(n)}{a_t} \cdot \overset{(n)}{h_{t-1}} + \overset{(n)}{b_t}\overset{(1)}{x_t}$$

Mamba and Mamba-2 [Dao and Gu, 2024]

WIN-WIN

Linear (i.e. efficient) across time, nonlinear along depth

$$K(x))V(x)$$

$$h_t = A_t h_{t-1} + B_t x_t$$

(Time-varying) Linear dynamical system

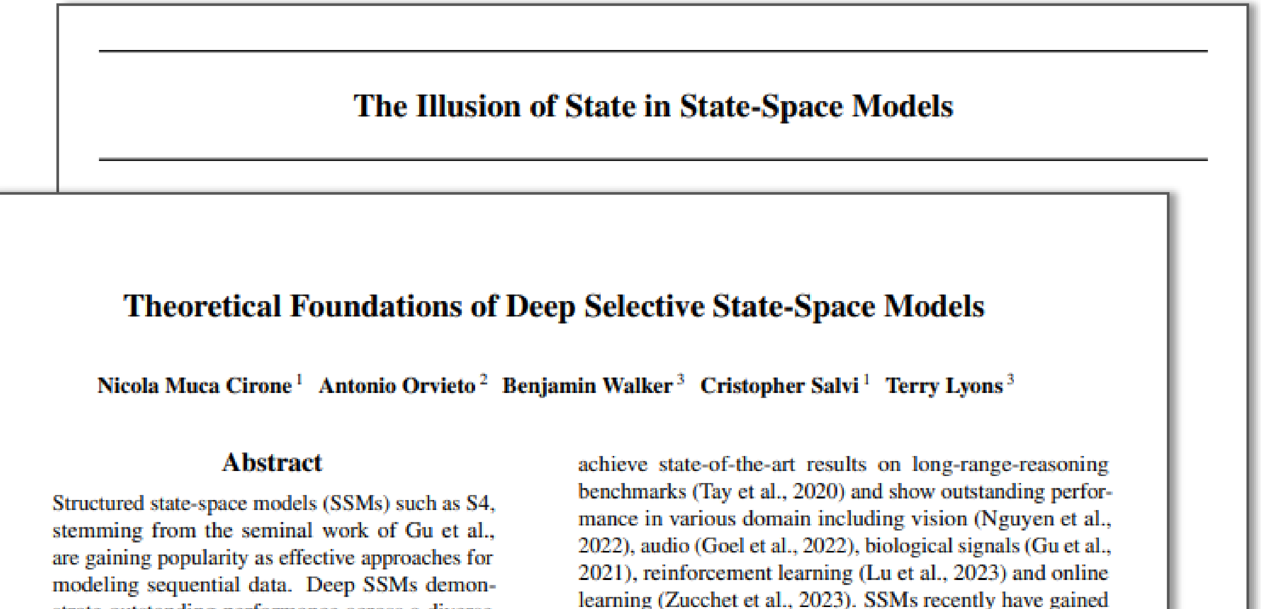
$$\overset{(n)}{h_t} = \rho(\overset{(n,n)}{A_t}\overset{(n)}{h_{t-1}} + \overset{(n,d)}{B_t}\overset{(d)}{x_t})$$

Recurrent neural network

Nonlinearity ~~across time~~ along depth via iterated local corrections *[Kaul 2020]*

Goal: approximate nonlinear RNN by a stack of linear systems, with nonlinearity along only depth

Theory: understand power of depth



The Illusion of State in State-Space Models

Theoretical Foundations of Deep Selective State-Space Models

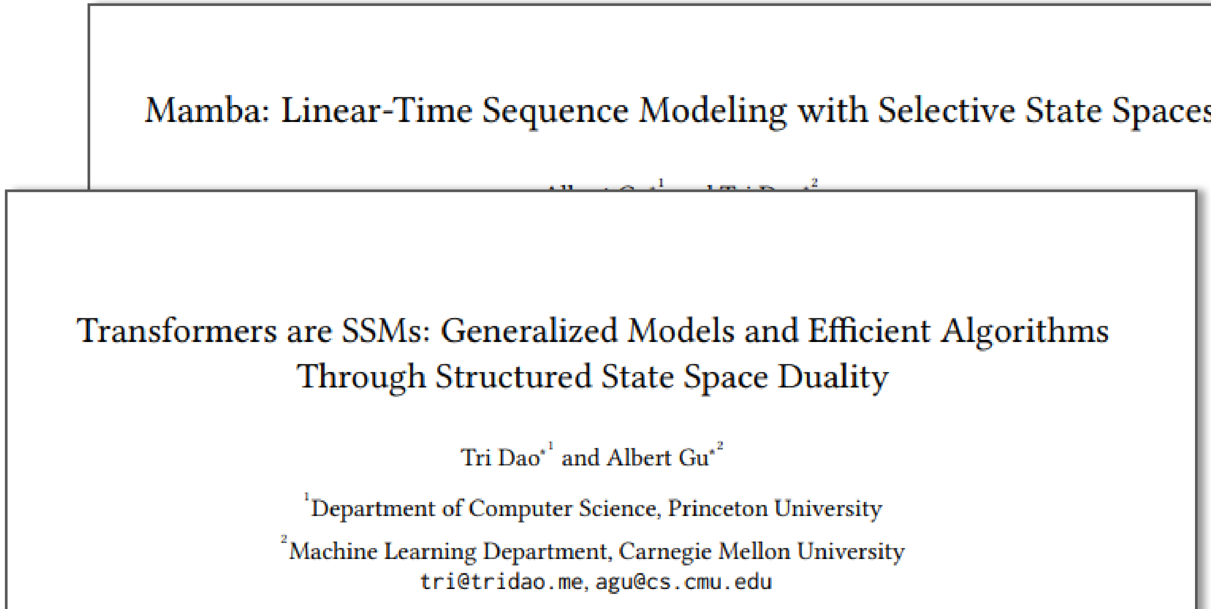
Nicola Muca Cirone[1] Antonio Orvieto[2] Benjamin Walker[3] Cristopher Salvi[1] Terry Lyons[3]

Abstract

Structured state-space models (SSMs) such as S4, stemming from the seminal work of Gu et al., are gaining popularity as effective approaches for modeling sequential data. Deep SSMs demon-

achieve state-of-the-art results on long-range-reasoning benchmarks (Tay et al., 2020) and show outstanding performance in various domain including vision (Nguyen et al., 2022), audio (Goel et al., 2022), biological signals (Gu et al., 2021), reinforcement learning (Lu et al., 2023) and online learning (Zucchet et al., 2023). SSMs recently have gained

Practice: use within new models



Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality

Tri Dao*[1] and Albert Gu*[2]

[1]Department of Computer Science, Princeton University

[2]Machine Learning Department, Carnegie Mellon University

tri@tridao.me, agu@cs.cmu.edu

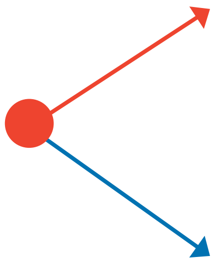
Nonlinearity ~~across time~~ along depth via iterated local corrections *[Kaul 2020]*

ASSUMPTION

$\rho(0)/0 \searrow 1$

$$s_0^{(1)} = s_0 = h_0$$

If a state is correct...



$$h_1 = \rho(a_1 \cdot h_0 + b_1 x_1)$$

$$s_1 = a_1 \cdot s_0 + b_1 x_1$$

Then its next-state multiplier is correct...

$$k_1 = \frac{\rho(a_1 \cdot s_0 + b_1 x_1)}{a_1 \cdot s_0 + b_1 x_1} = \frac{\rho(a_1 \cdot h_0 + b_1 x_1)}{a_1 \cdot h_0 + b_1 x_1} = \frac{h_1}{a_1 \cdot h_0 + b_1 x_1}$$

So, in the next layer, the next state becomes correct.

$$\begin{aligned} s_1^{(1)} &= k_1 \cdot (a_1 \cdot s_0^{(1)} + b_1 x_1) \\ &= k_1 \cdot (a_1 \cdot h_0 + b_1 x_1) = h_1 \end{aligned}$$

Nonlinearity ~~across time~~ along depth via iterated local corrections *[Kaul 2020]*

ASSUMPTION

$\rho(0)/0 \searrow 1$

$$h_t = \rho(a_1 \cdot h_{t-1} + b_1 x_1)$$

$$s_0^{(1)} = s_0 = h_0$$

$$s_t^{(0)} = a_t \cdot s_{t-1}^{(0)} + b_t x_t$$

If a state is correct...

Then its next-state multiplier is correct...

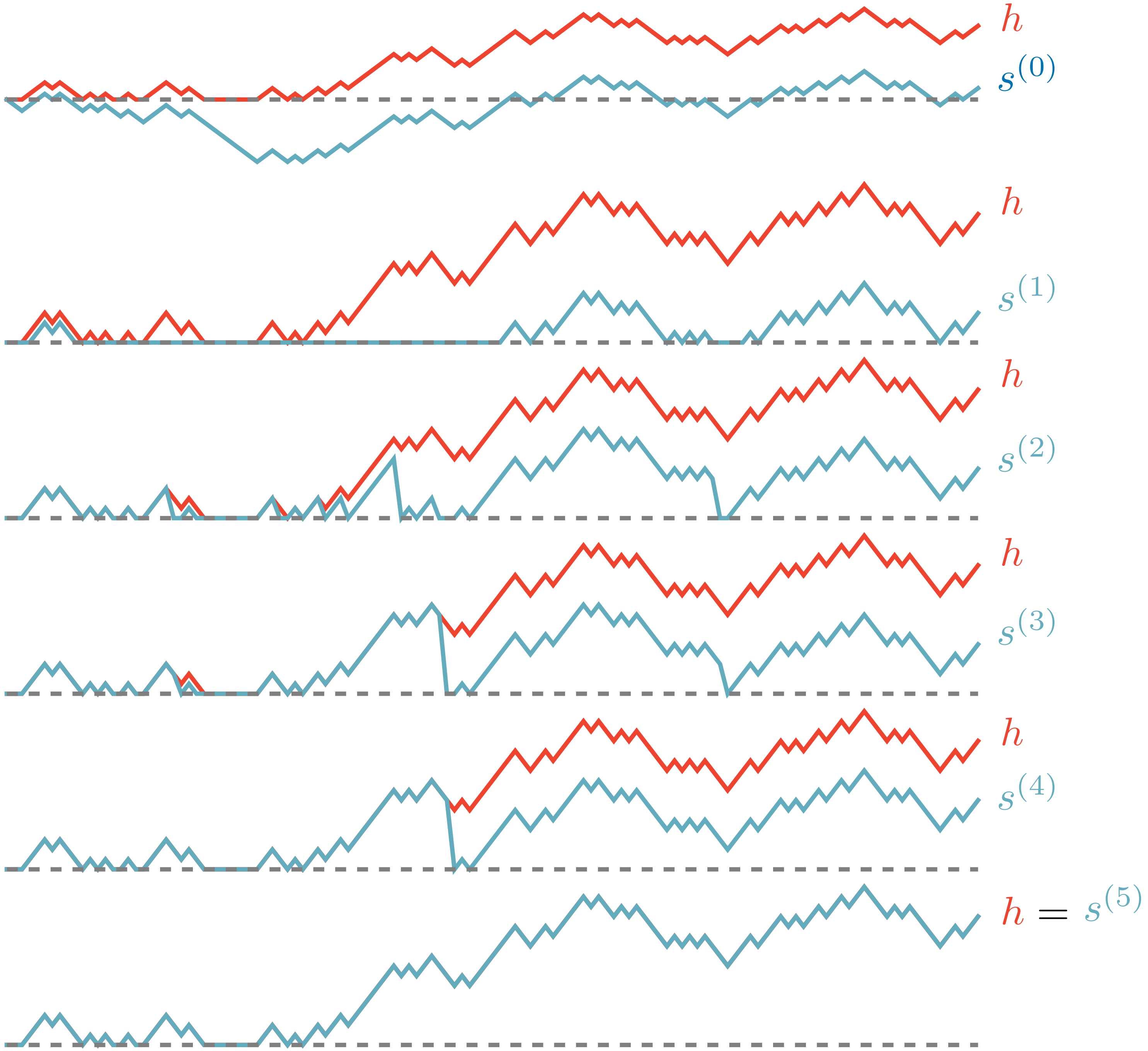
$$k_t^{(i)} = \frac{\rho(a_t s_{t-1}^{(i-1)} + b_t x_t)}{a_t s_{t-1}^{(i-1)} + b_t x_t} \qquad k_i^{(i)} = \frac{h_i}{a_i h_{i-1} + b_i x_i}$$

So, in the next layer, the next state becomes correct.

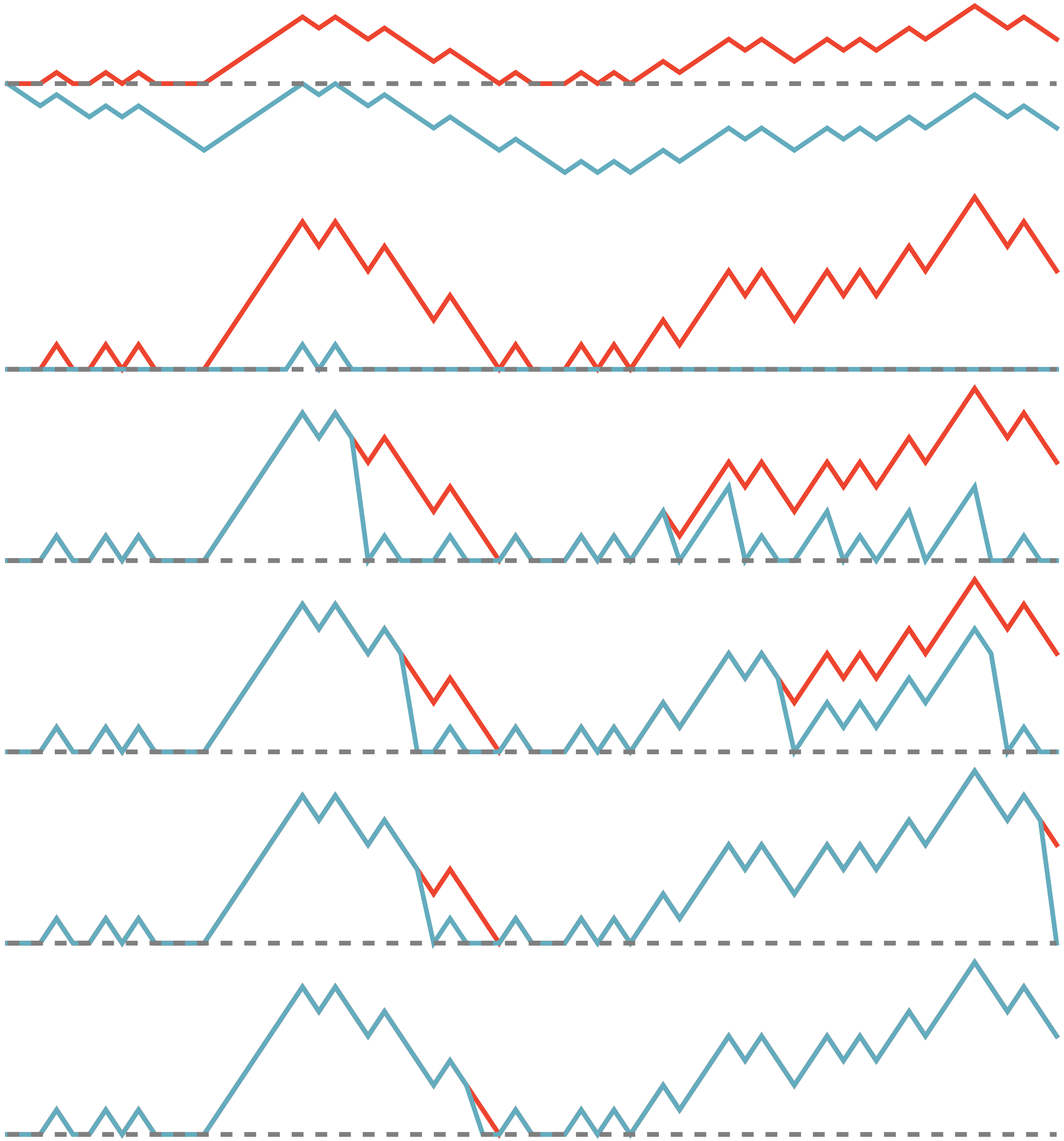
$$s_t^{(i)} = k_t^{(i)} \cdot (a_t \cdot s_{t-1}^{(i-1)} + b_t x_t)$$

$$s_i^{(i)} = k_i^{(i)} \cdot (a_i \cdot h_{i-1} + b_i x_i) = h_i$$

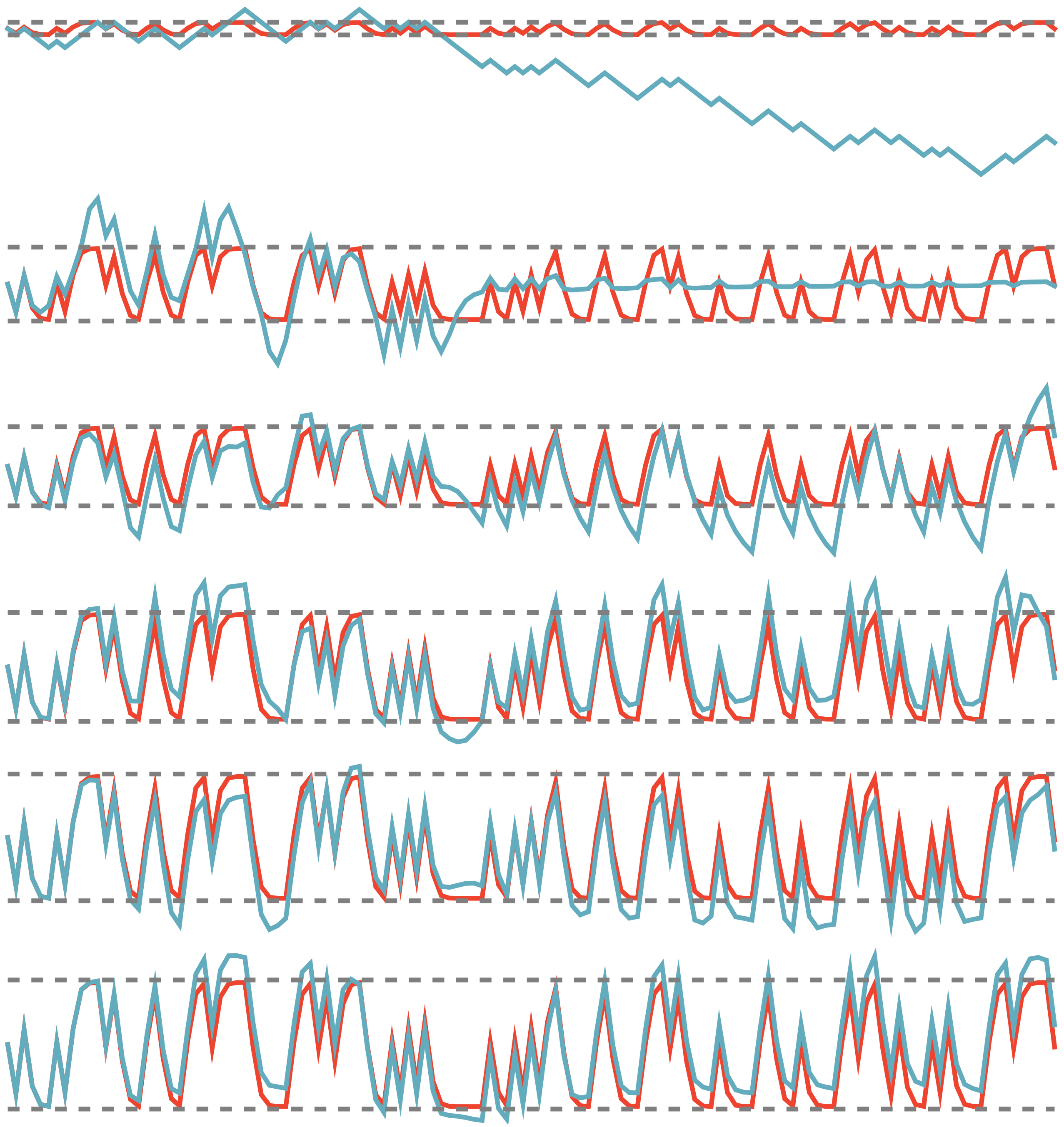
$\rho = \mathrm{ReLU}$



$\rho = \mathrm{ReLU}$



$\rho = \tanh$



$\rho = \mathrm{clip}$