Data science to ask questions in mental health

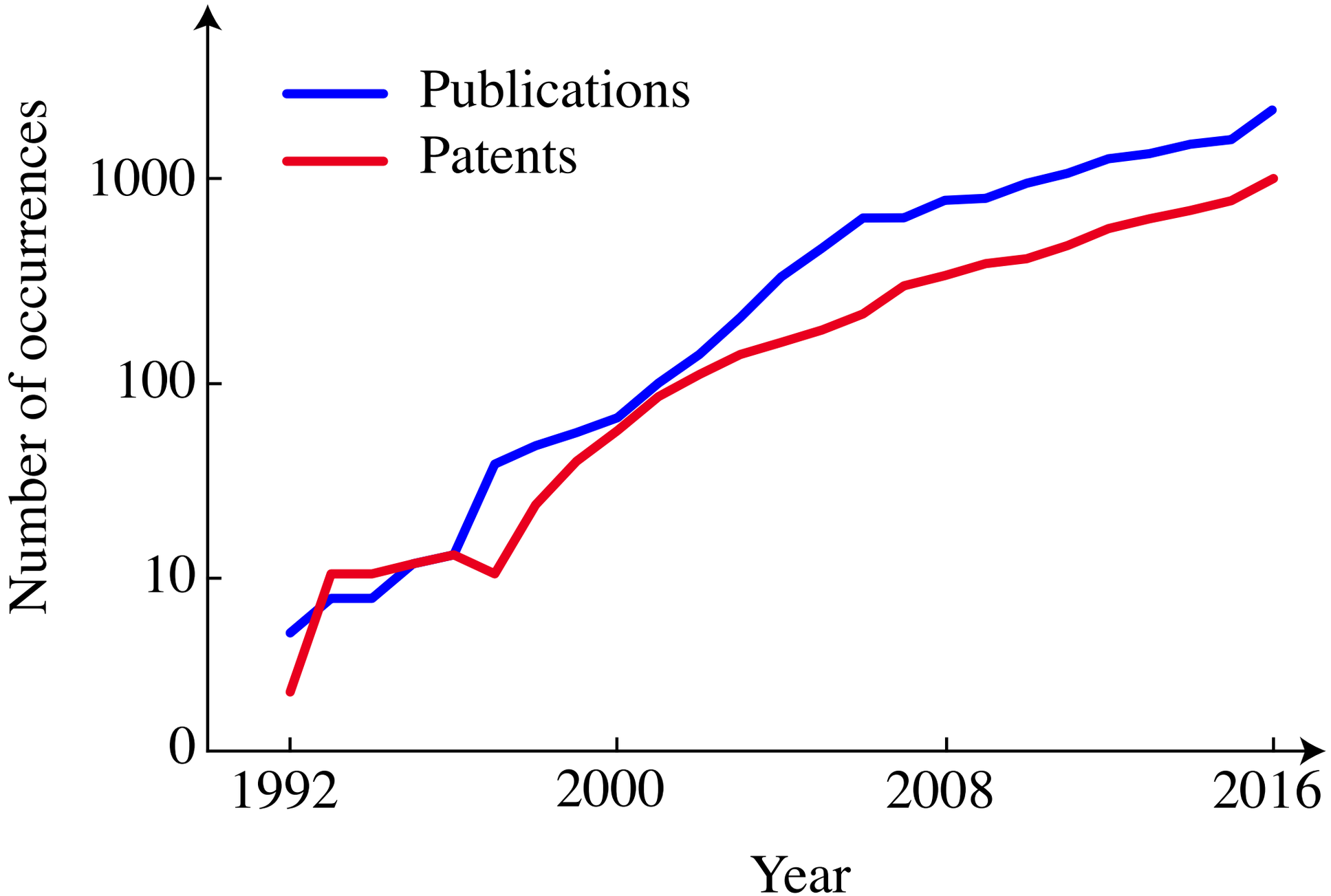
@kordinglab

Shameless plug: Please read *10 simple rules for structuring papers*

Outline

- I) What ML is used for
- II) ML settings, diagnostics and typical uses
- III) Four ways of doing it wrong
- IV) An aside: video based approaches
- V) Causality FTW

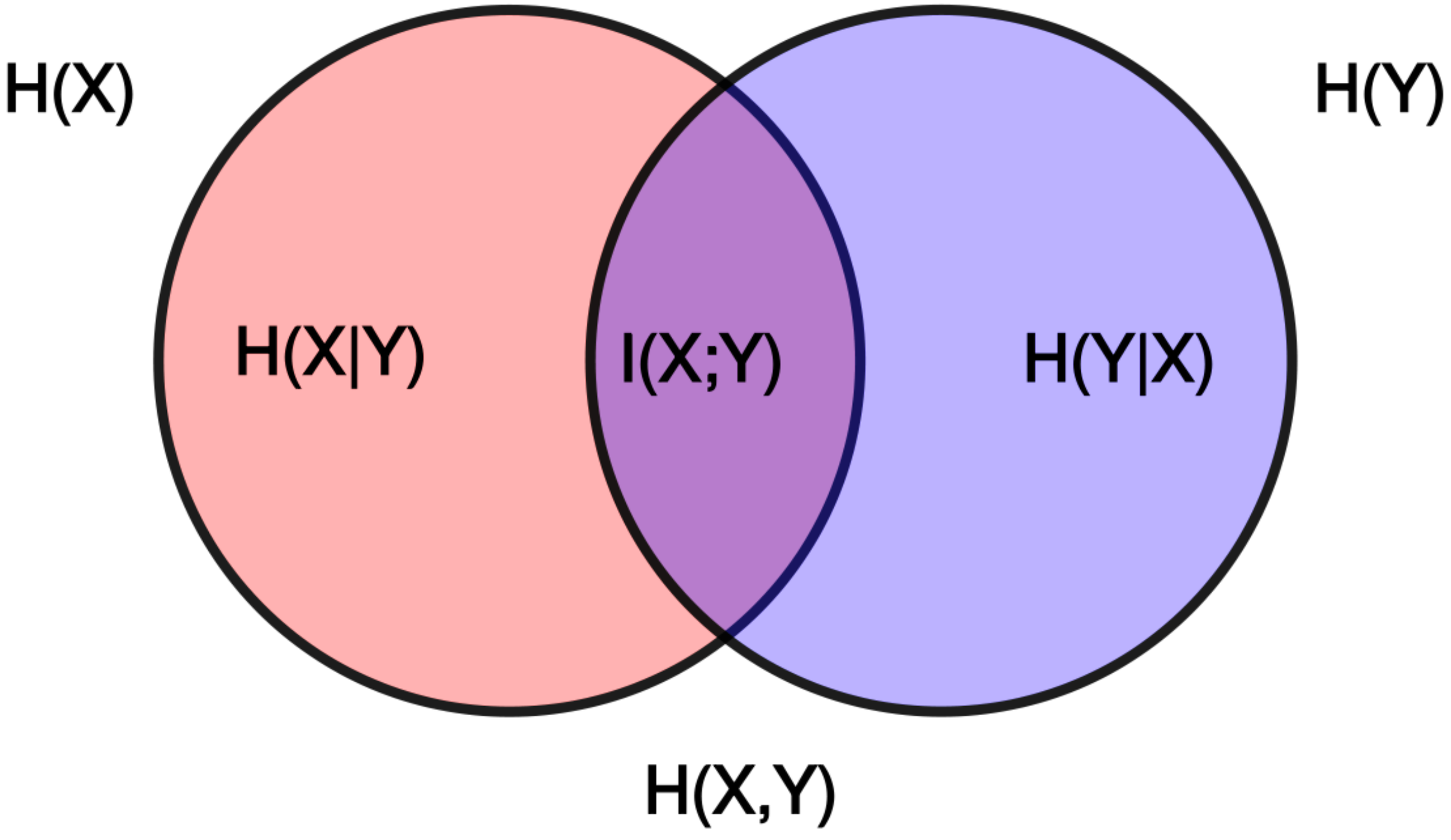
I: ML is getting popular in biomedical science



Solve real problems

Depression estimates from mobile phones (with Mohr)

Understand data



e.g. Bialek

Provide a benchmark

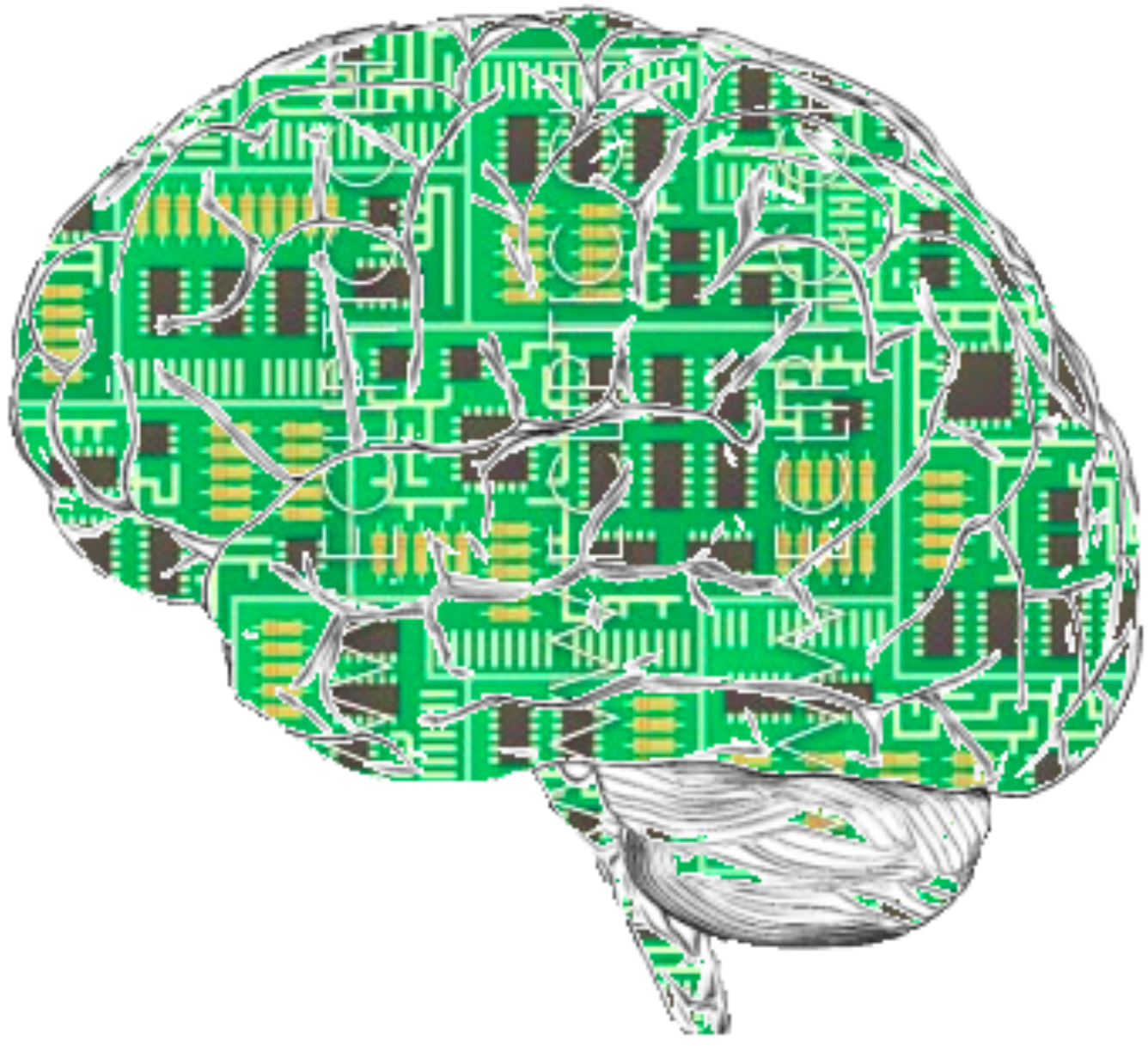
Your Model

Machine learning

Being better than another model does not make a model true.

See Jonas and Kording, Could a neuroscientist understand a microprocessor 2017

Model for brain



see Marblestone, Wayne, Kording, 2017

Model for disease

- Solutions
- Fitting
- Bayes
- …
- Deep learning

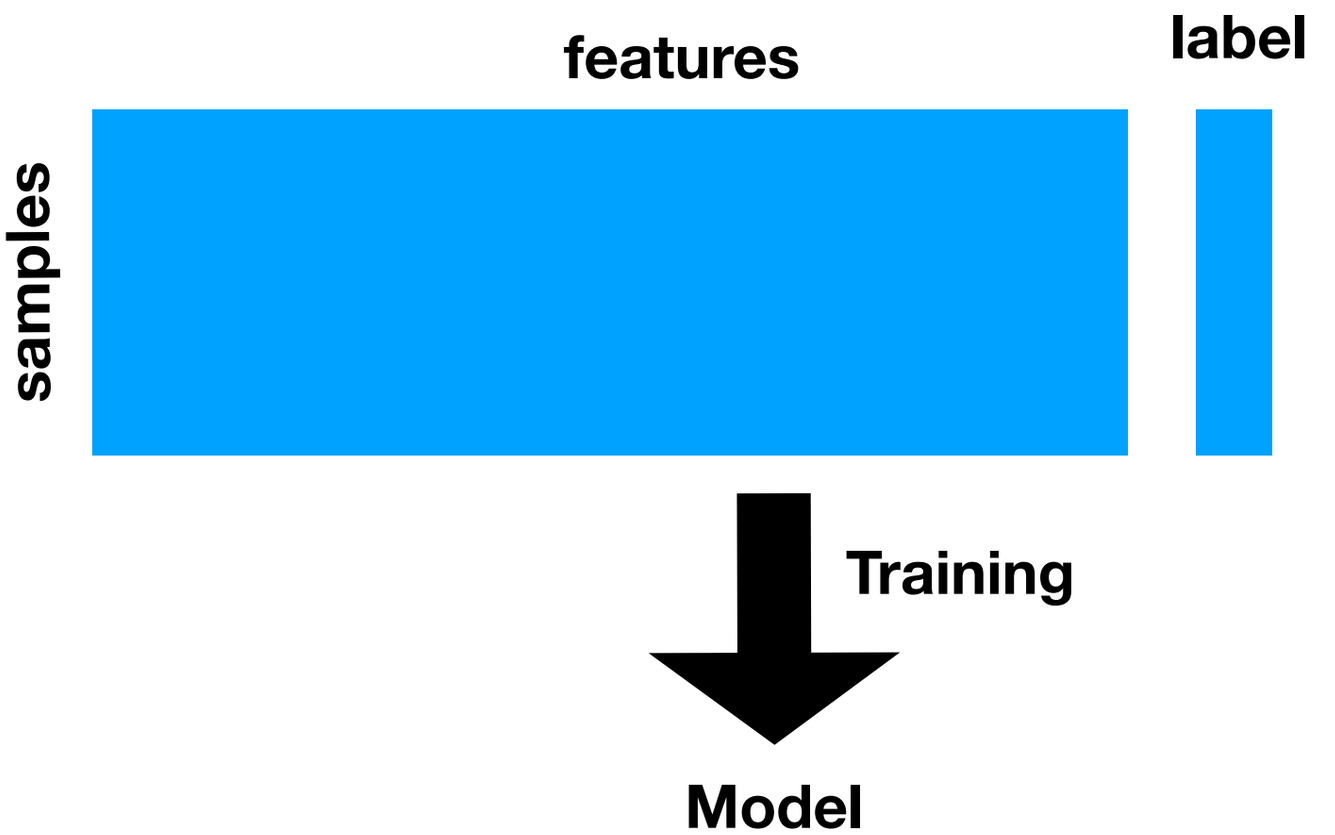
II: Two approaches towards diagnostics

- Measure the right thing
  - e.g. identify antibodies, viral RNA etc
- Measure a lot of stuff (ubiquitous)
  - Google searches (e.g. Flu)
  - Locations
  - New media use
  - Accelerations
  - Etc
- And then get at the relevant stuff through machine learning

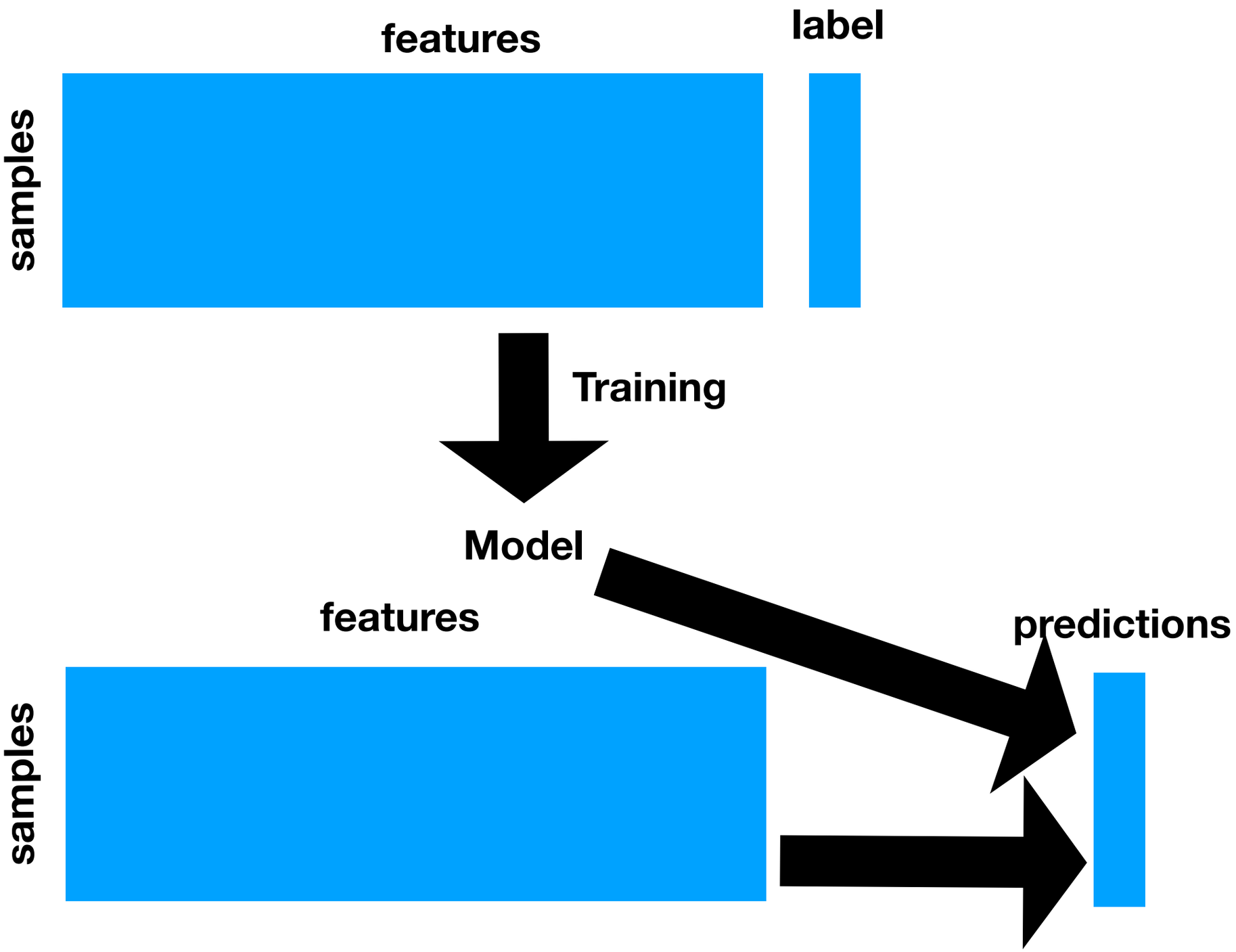
Workflow

- Produce data where we know the correct diagnostic
- Train a machine learning system
- Test that our machine learning system works
- Use it to make cheaper/better diagnostics

Typical Supervised ML setting



Typical Supervised ML setting

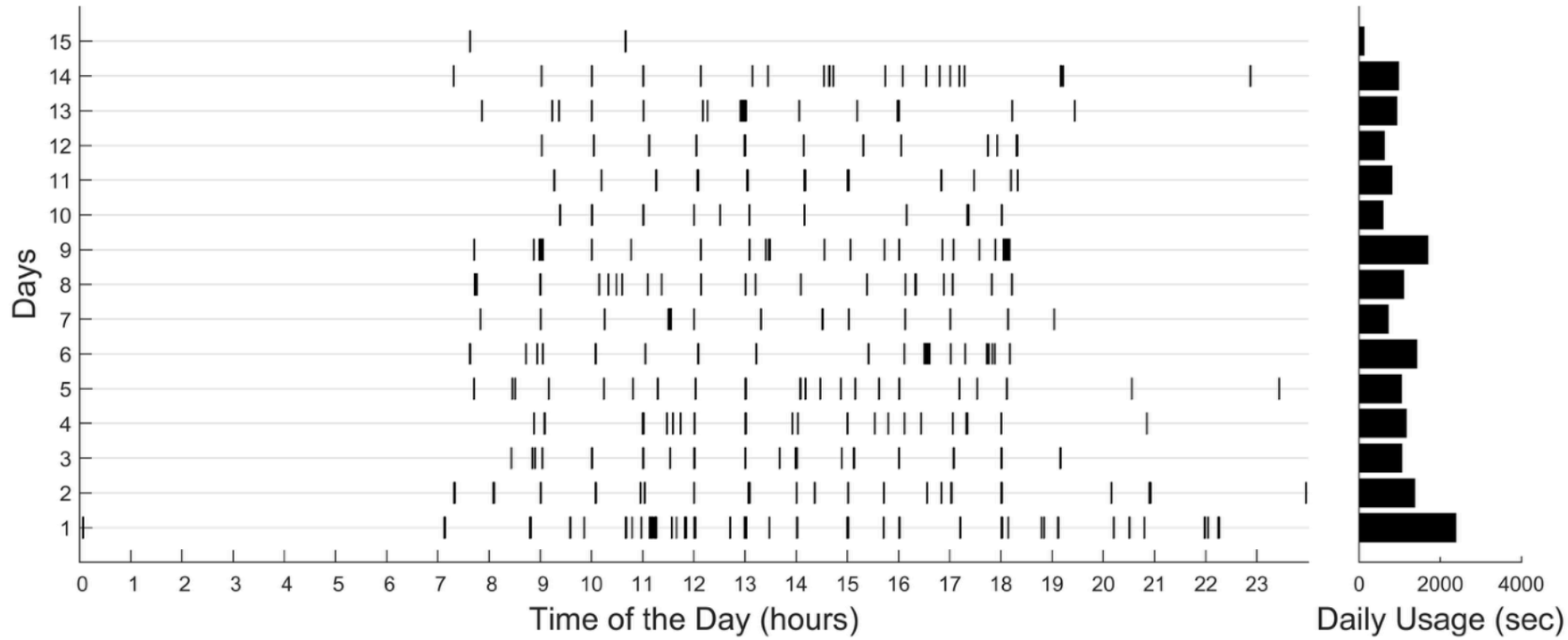


A typical example: PHQ9 from phone sensors

Phone sensors, truly ubiquitous

- Accelerometer/ Magnetometer/ Barometer
- Brightness sensor
- GPS
- Screen/ Keyboard
- Microphone

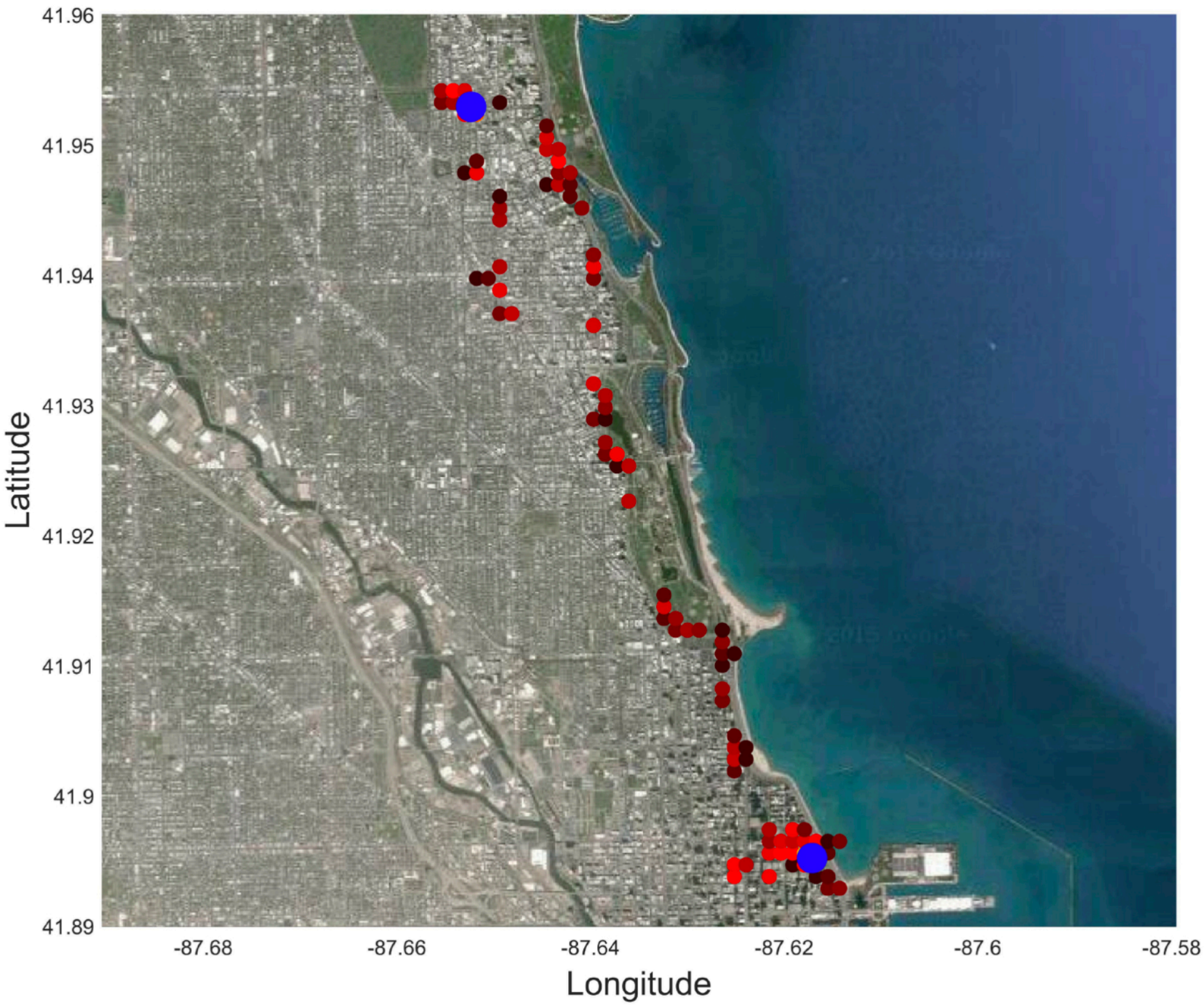
Phone use



With Lonini, Jayaraman

People use their phones all the time

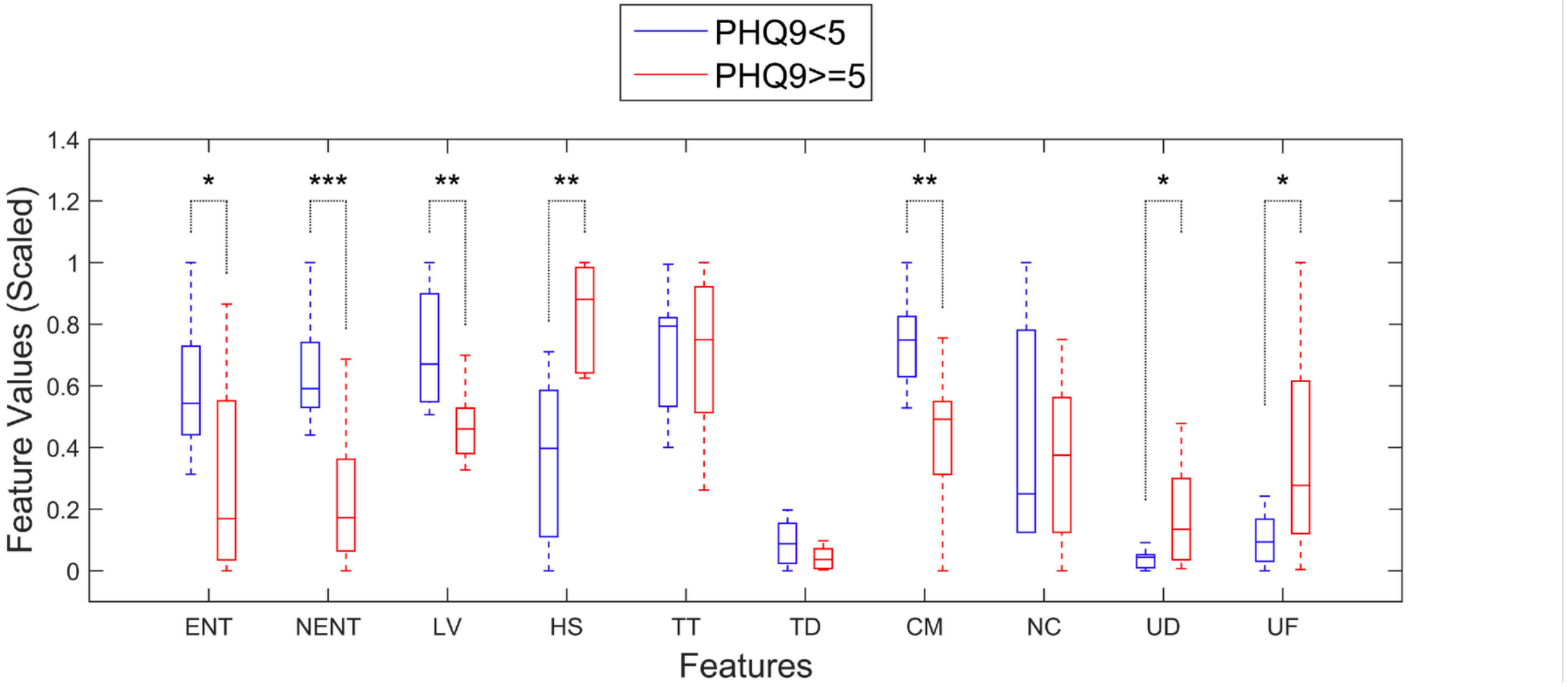
GPS data


41.96
41.95
41.94
41.93
41.92
41.91
41.9
41.89
Latitude
-87.68
-87.66
-87.64
-87.62
-87.6
-87.58
Longitude

Extract GPS Features

- Location Variance
- Number of clusters
- Entropy
- Home Stay
- Circadian Movement
- Transition time

Correlated with PHQ9



Combine them with trivial machine learning!

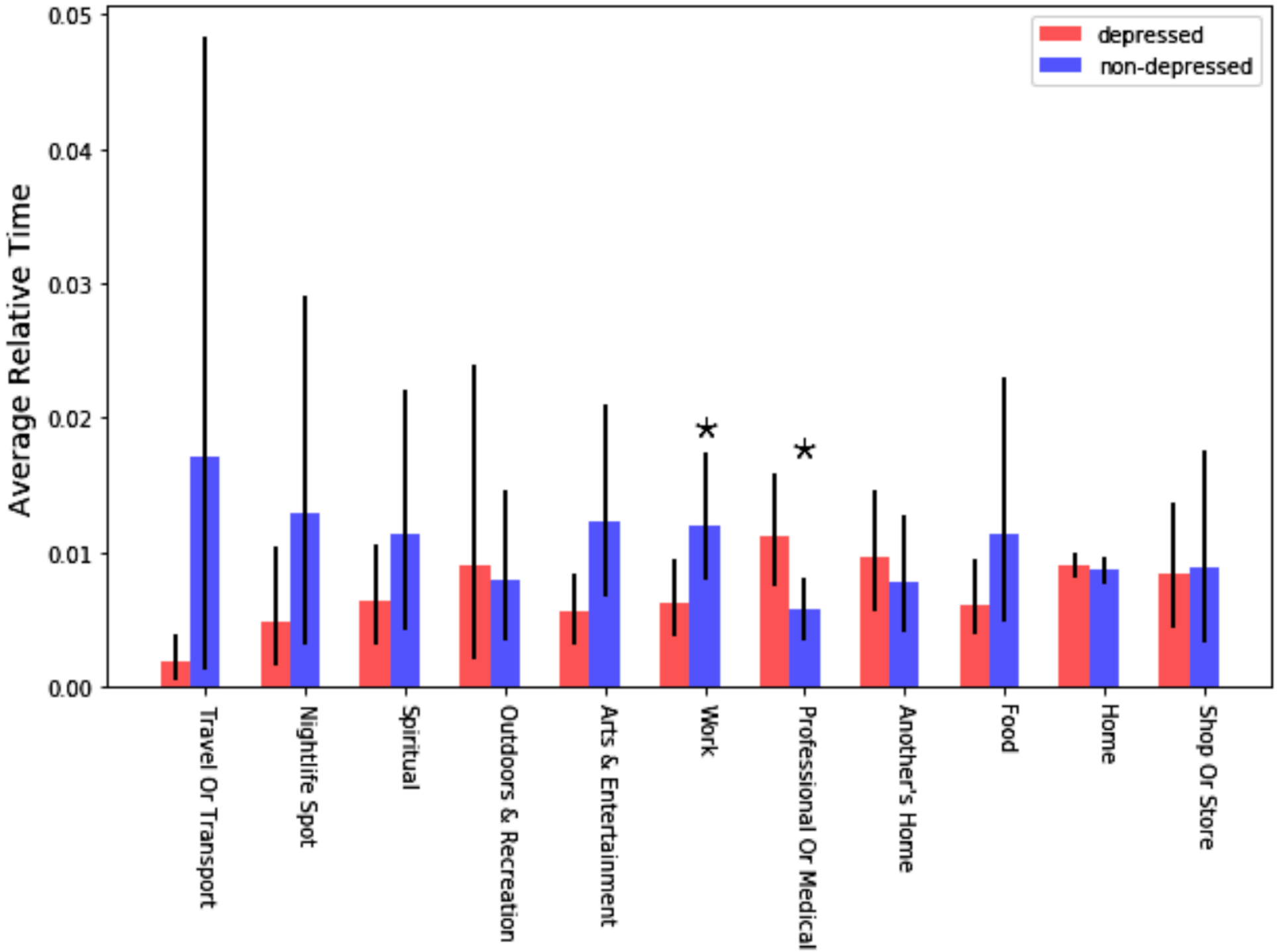
$$P(\text{Depressive Symptoms}) = g(b_0 + b_1F_1 + b_2F_2 + \ldots + b_nF_n)$$

While looking for small b

Somewhat can predict mood

	Classification (PHQ9<5 vs PHQ9≥5)			PHQ9 score estimation
Training features	% mean accuracy (SD)	% mean sensitivity	% mean specificity	Mean NRMSD (SD)
Usage duration	74.2 (3.4)	64.0	83.9	0.268 (0.018)
Usage frequency	68.6 (4.1)	56.4	79.6	0.249 (0.013)
All	65.7 (4.9)	55.7	74.9	0.273 (0.019)

Semantic location



Saeb… Kording Mohr

How to do good ML

- SVM/SVR
- kNN
- xgBoost
- Random Forest
- GLM
- Stacking!

This is what all the ML courses teach

Use Auto-ML instead

- Approaches are sufficiently standard that this part can easily be automated, e.g. auto-SKlearn, auto-WEKA

- Implication: knowledge about details of ML techniques will become less relevant for biomedical scientists

Result

- AutoML (autosklearn, Freiburg) is almost always better than published results
- AutoML is usually better than our own results
- It is literally three lines of code

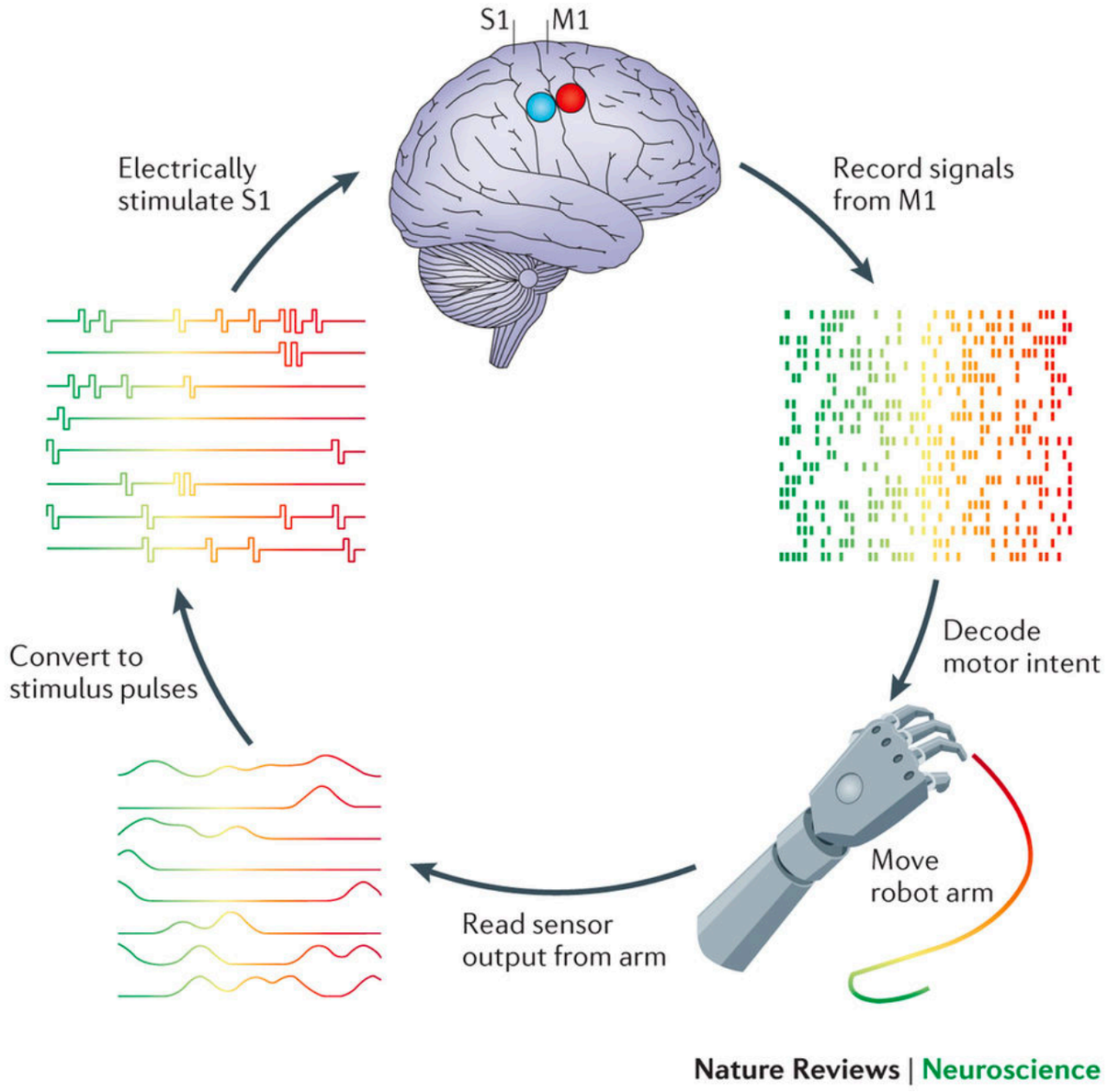
Auto-sklearn is good

model	features	accuracy	macro f1	weighted f1
majority baseline	N/A	0.5714	0.1818	0.4156
random forest	age/gender	0.5714	0.1818	0.4156
random forest	comm	0.6667	0.4795	0.6254
random forest	comm + age/gender	0.6667	0.4750	0.6225
random forest	comm + demo + loc	0.6762	0.4744	0.6326
auto-sklearn	age/gender	0.5714	0.1818	0.4156
auto-sklearn	comm	0.6571	0.4731	0.6195
auto-sklearn	comm + age/gender	0.6905	0.5488	0.6654
auto-sklearn	comm + demo + loc	**0.7095**	**0.5519**	**0.6806**

Relationship prediction, with Lyle Ungar, Tony Liu

Example uses of ML in Neuroscience

Encoding

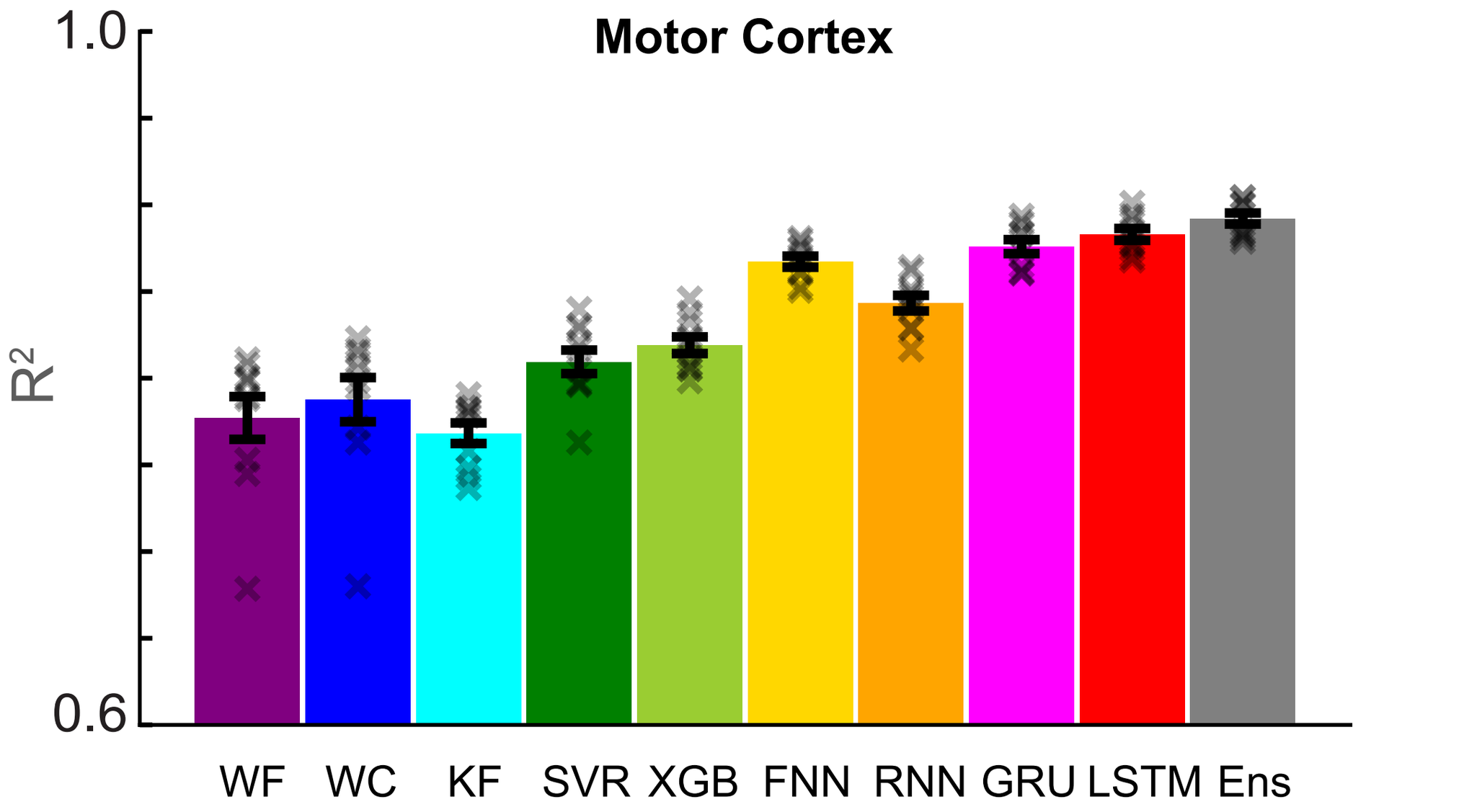


Decoding

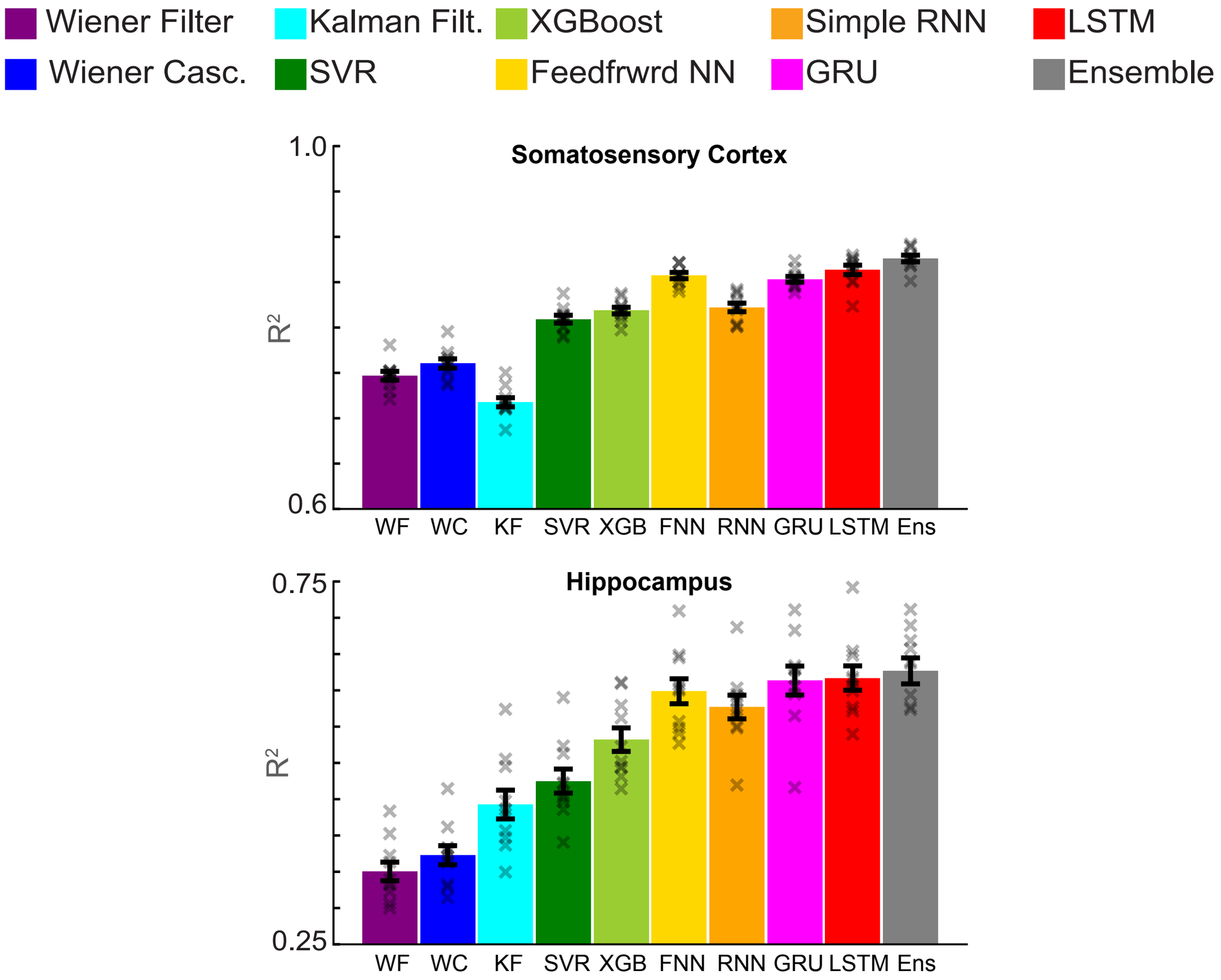
Bensmaia, Miller, 2014

Decoding (Neurons-> movement)

Wiener Filter
Kalman Filt.
XGBoost
Simple RNN
LSTM
Wiener Casc.
SVR
Feedfrwrd NN
GRU
Ensemble

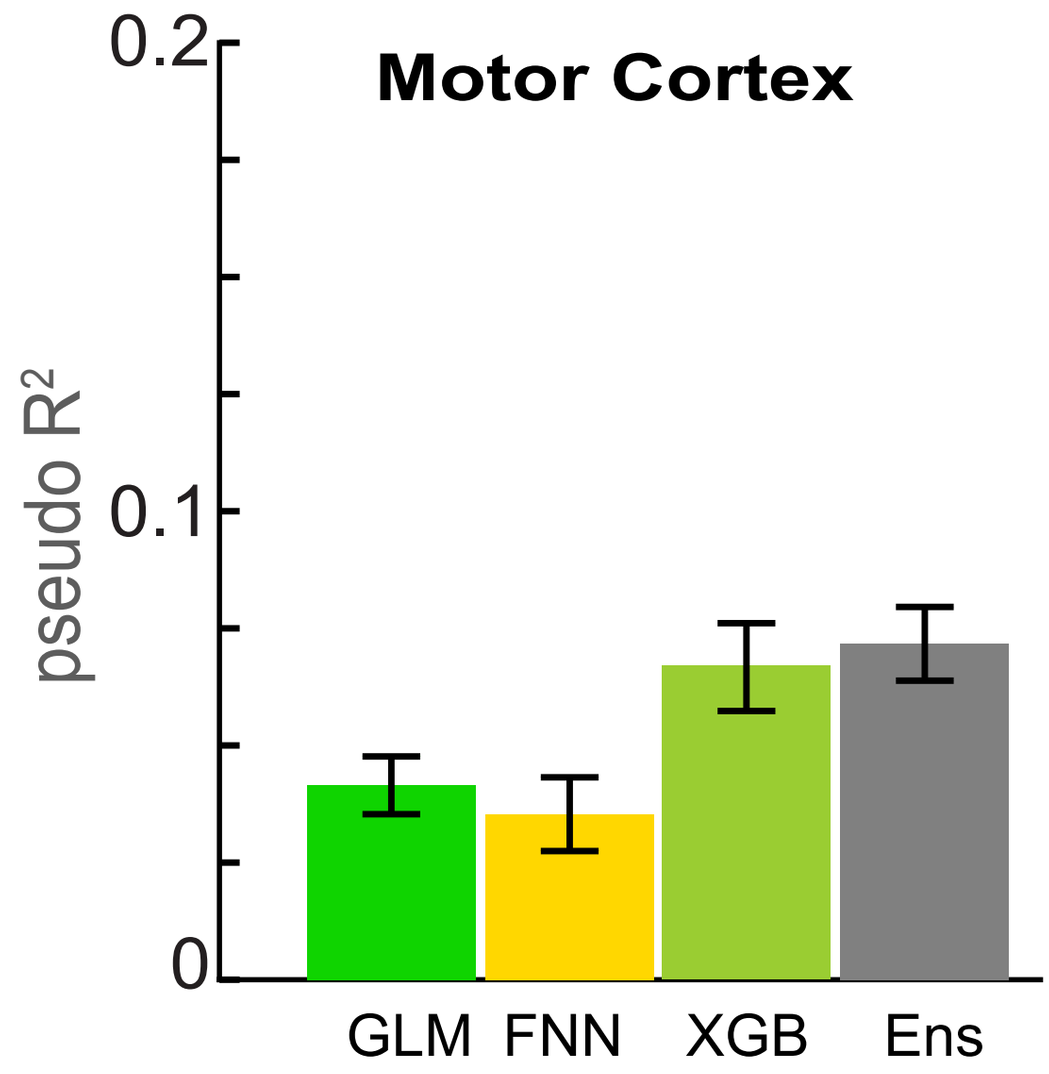


Finding generalizes

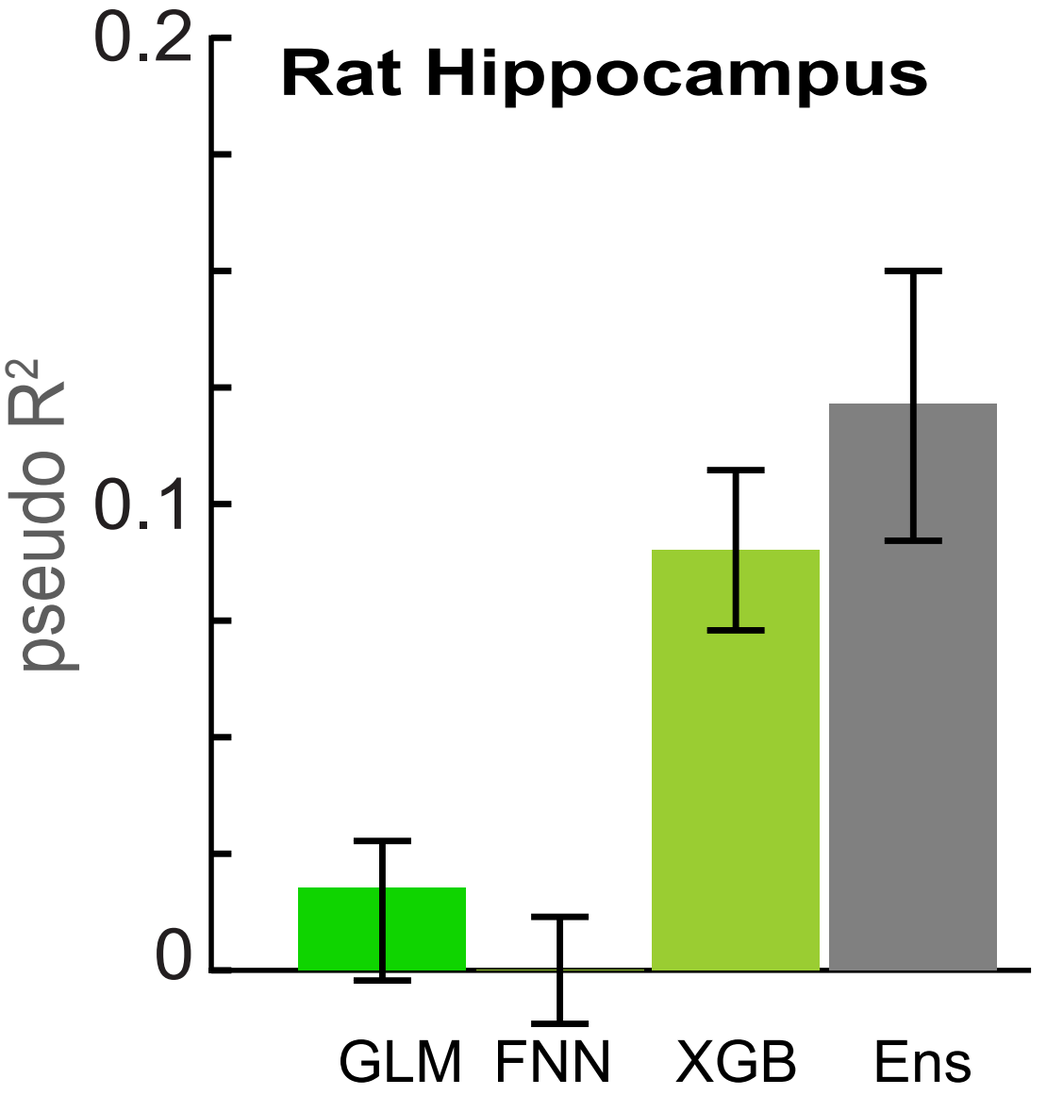


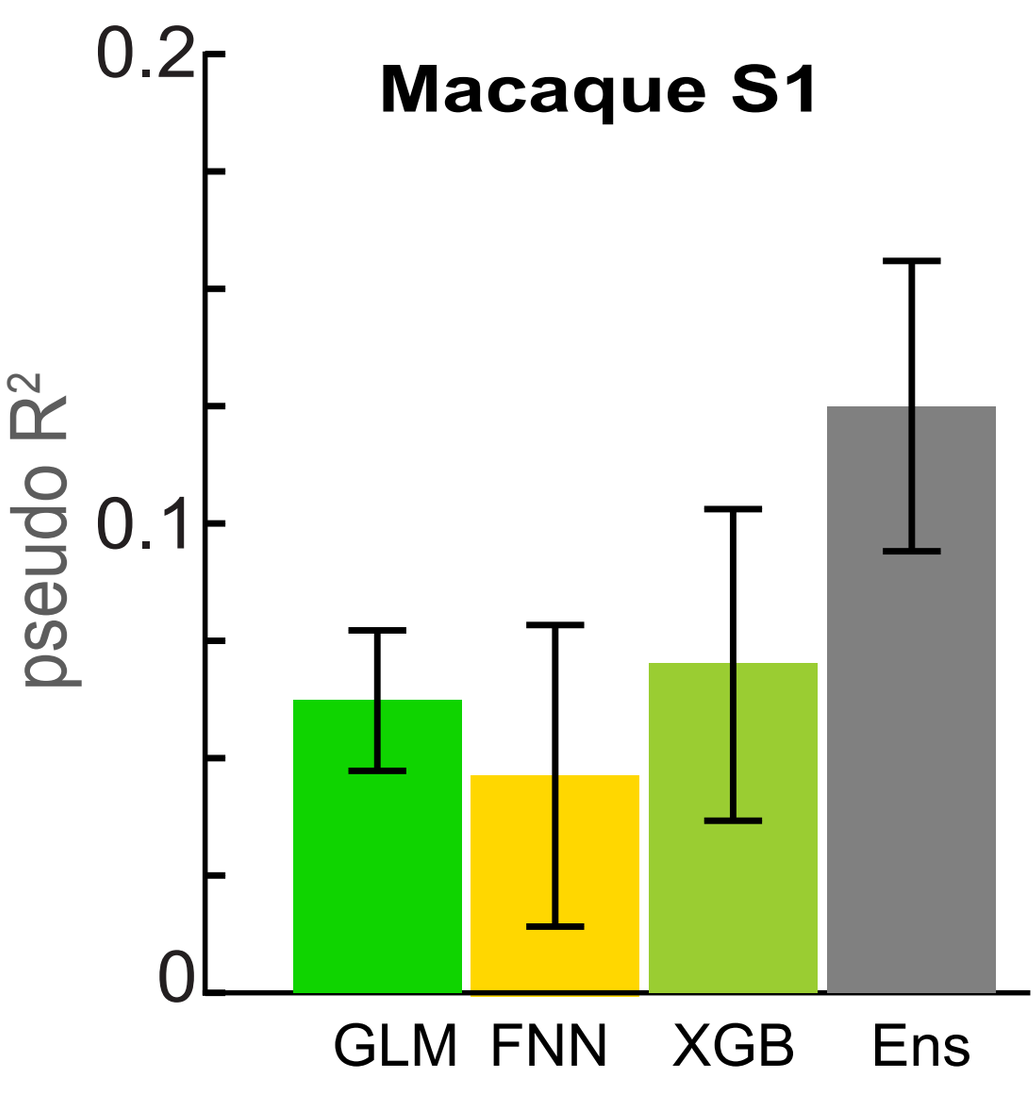
Encoding (movements->neurons)

GLM Feedfrwrd NN XGBoost Ensemble



Finding Generalizes





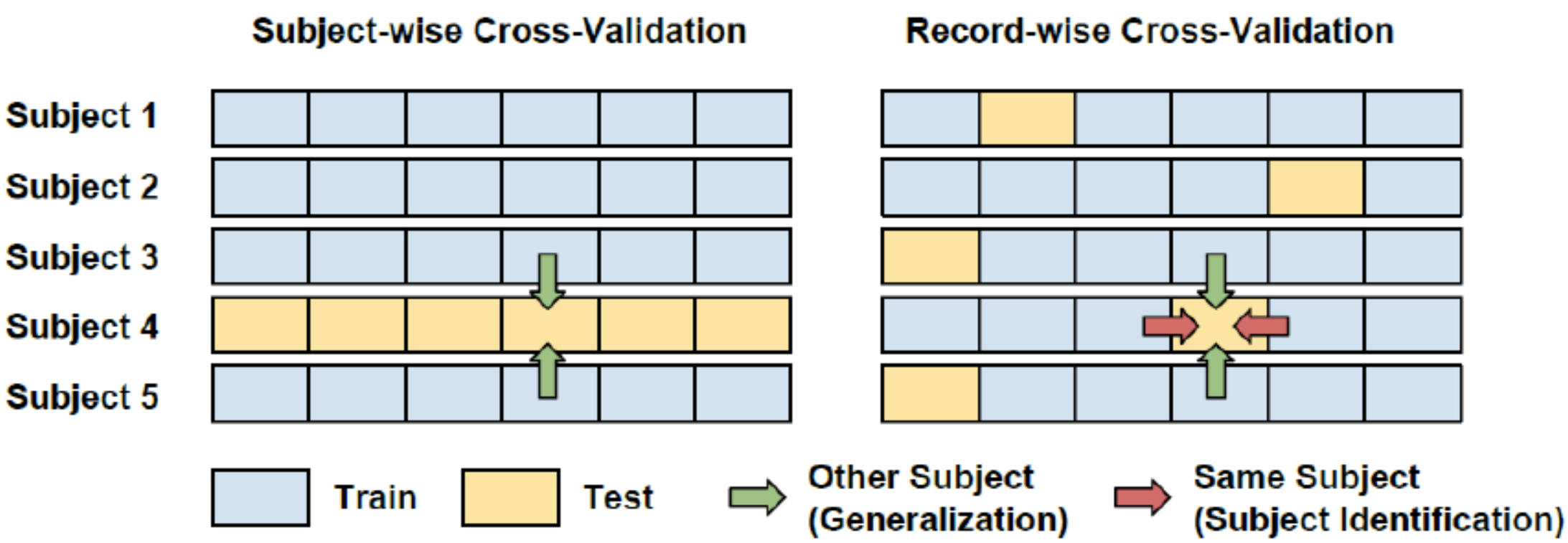
III: The four deadly sins of machine learning

- (1) Wrong question
- (2) Wrong way of assessing quality
- (3) Wrong way of comparing
- (4) Wrong way of managing

(1) Wrong question

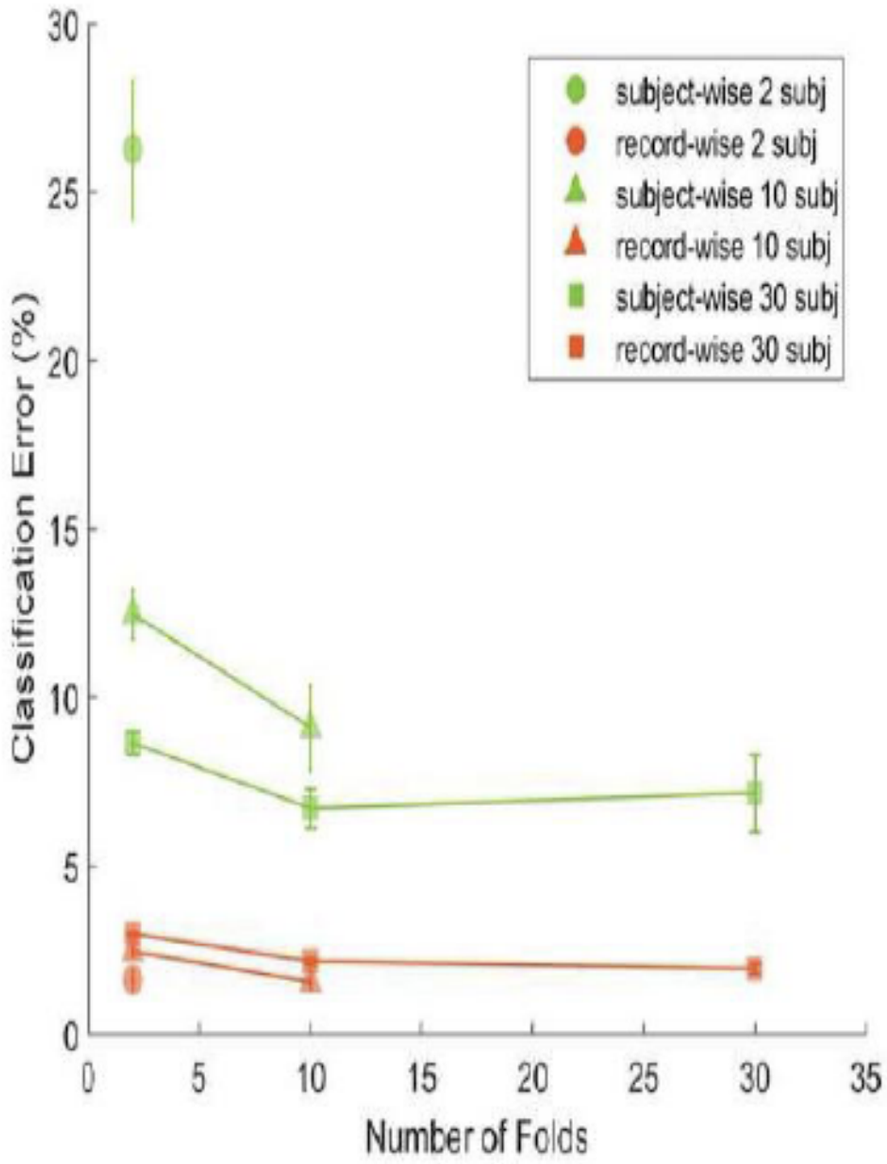
- Most ML people are in CS
- Little knowledge about medicine
- Often ask medically irrelevant question

(2) Wrong way of assessing Quality e.g. bad crossvalidation

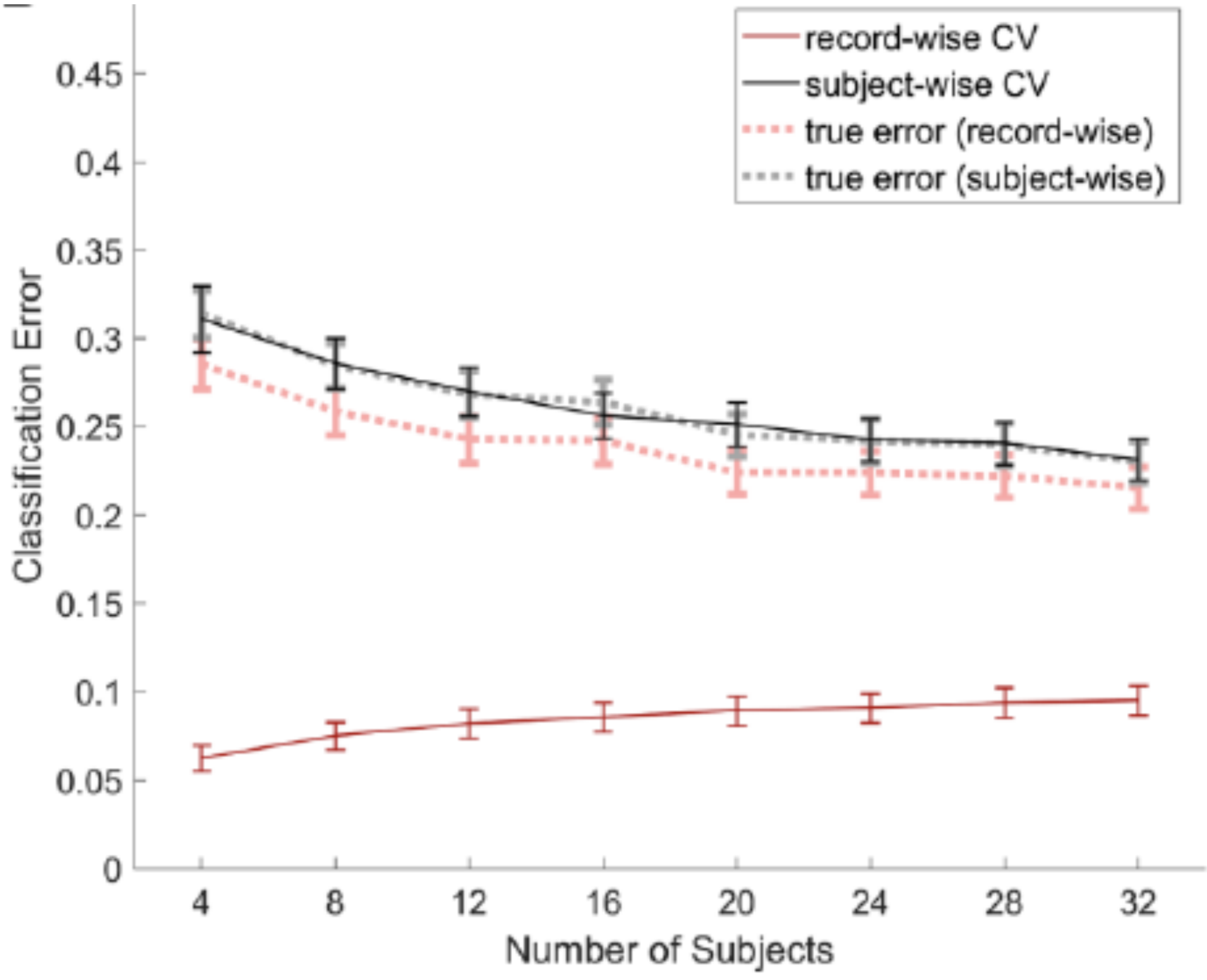


With Lonini, Saeb, Mohr, Jayaraman

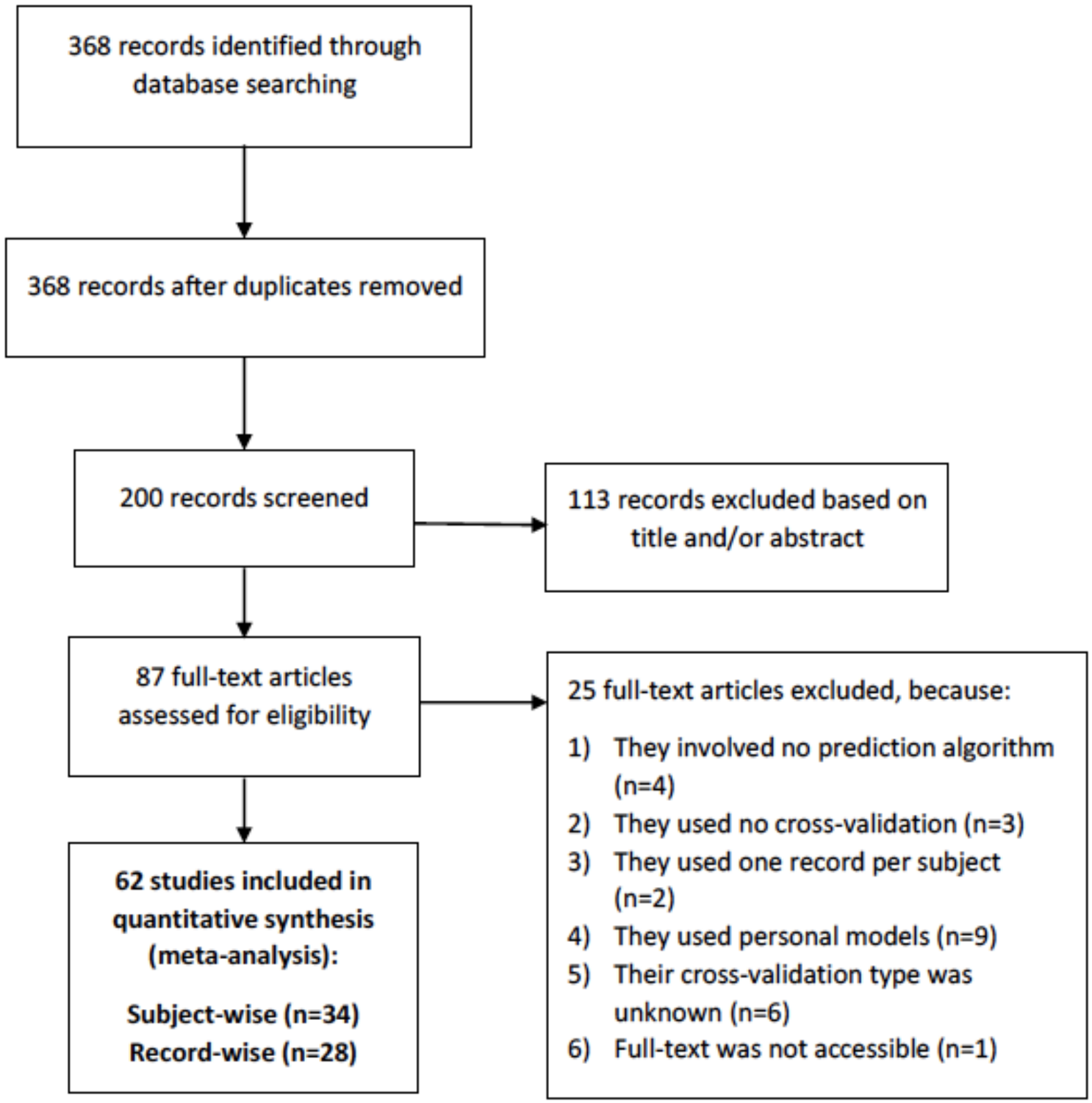
Cheating works



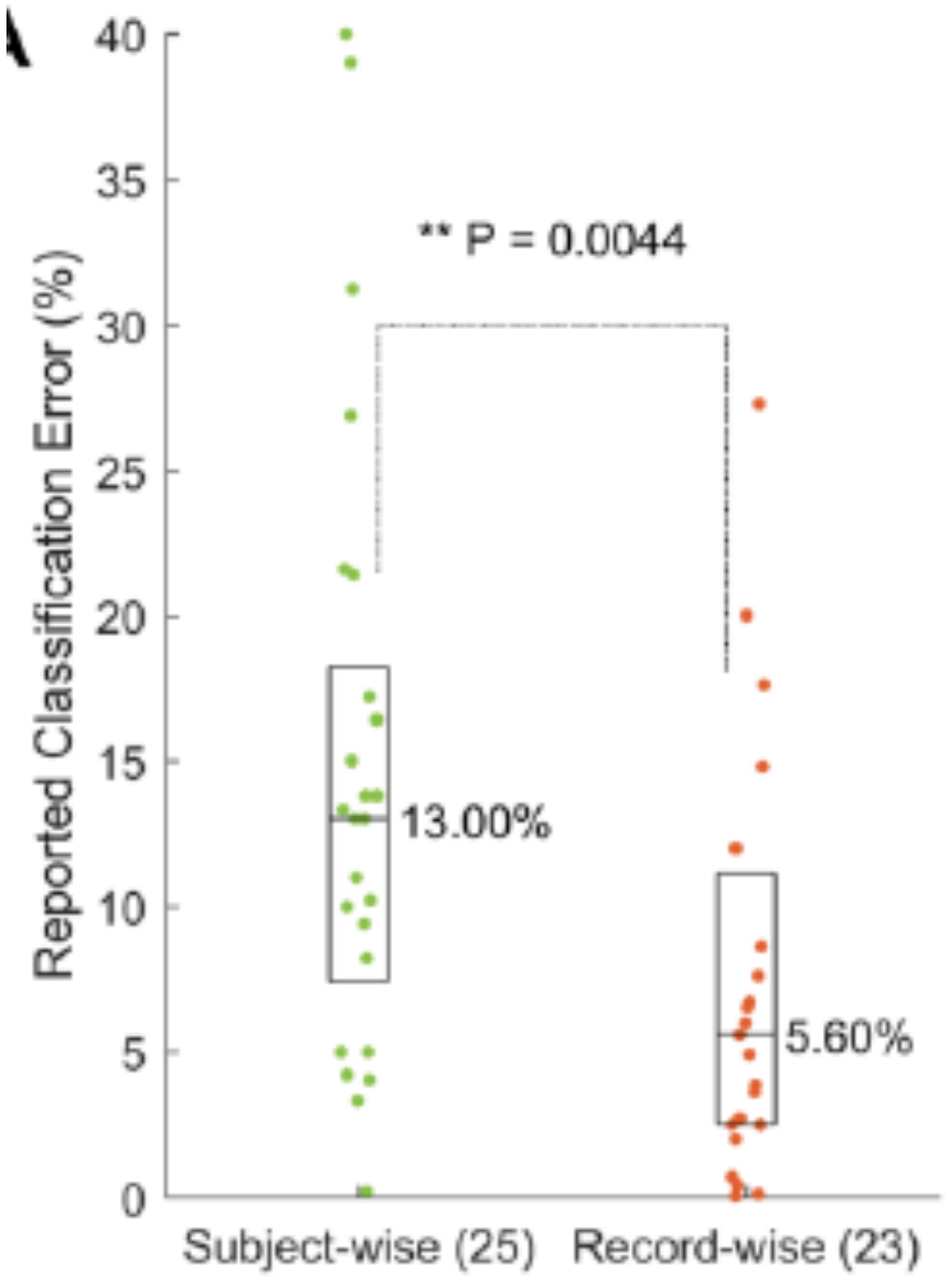
Massive overconfidence



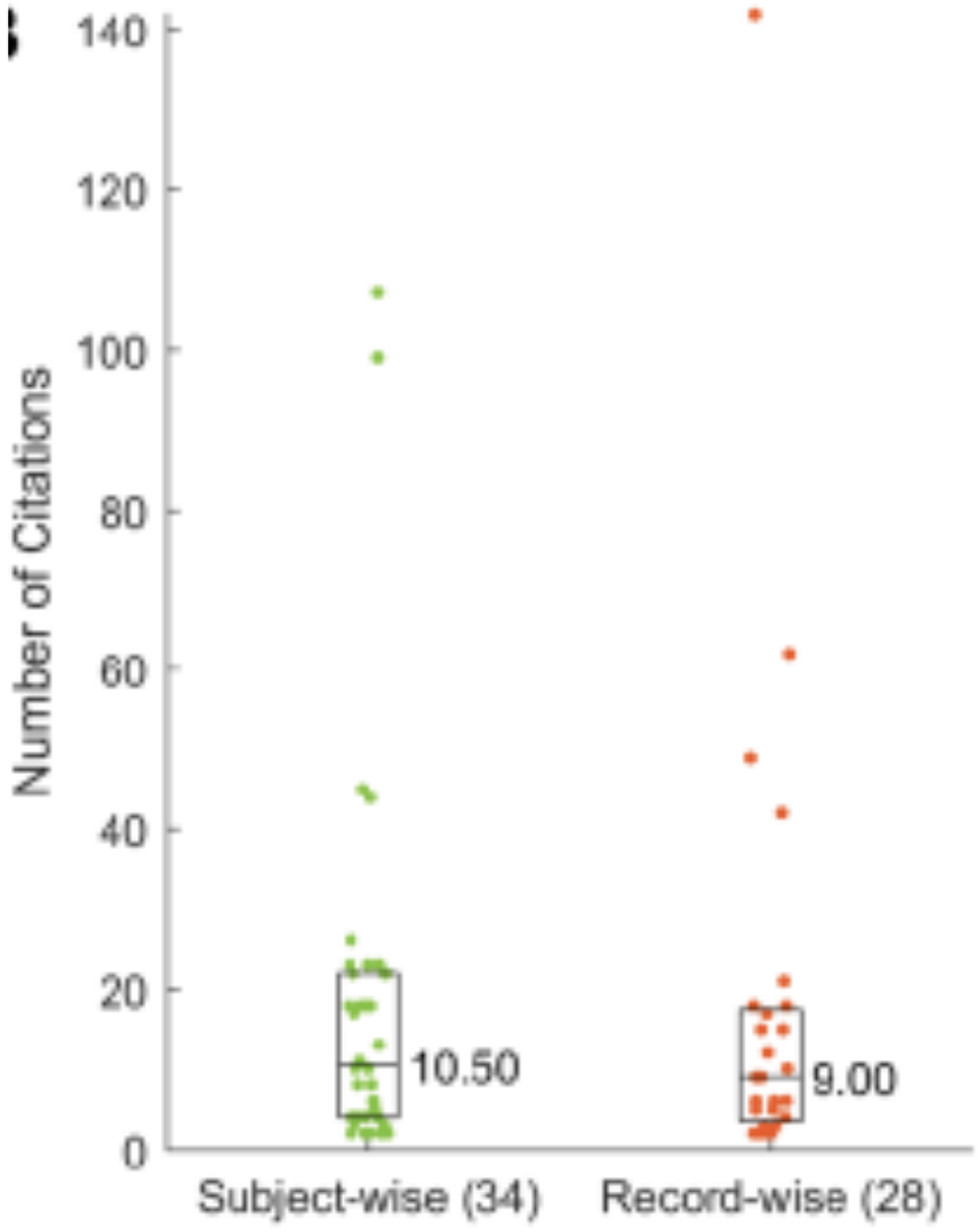
Literature review



Cheating helps



No one cares

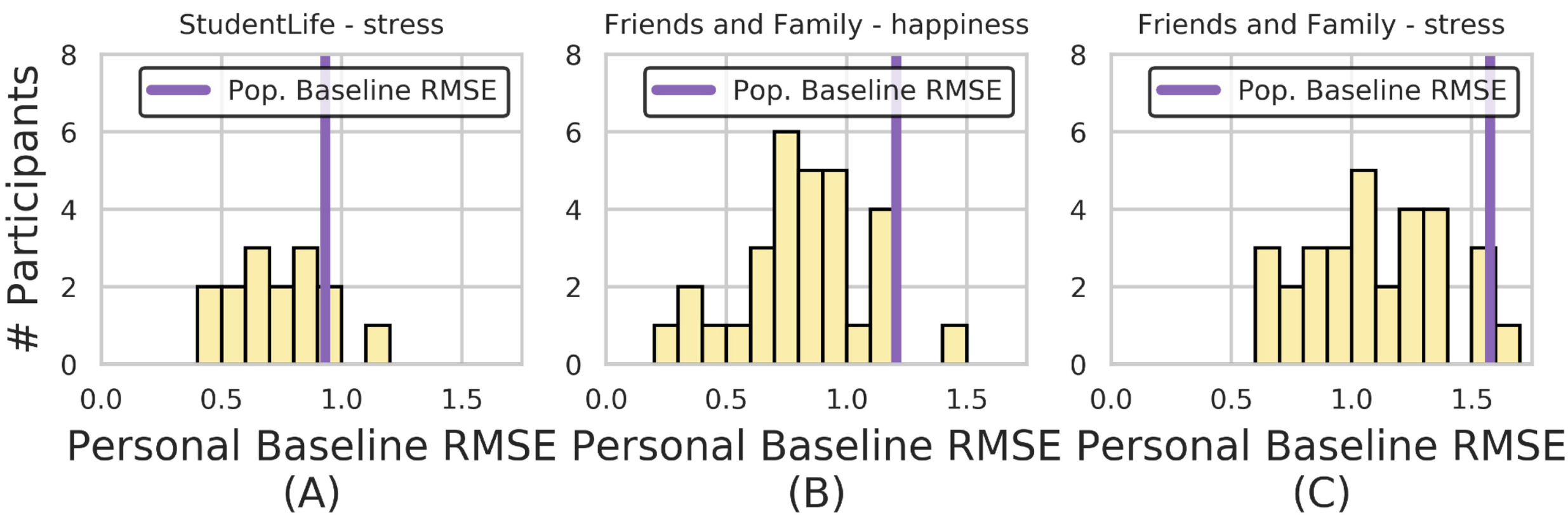

140
120
100
80
60
40
20
0
Number of Citations
10.50
9.00
Subject-wise (34)
Record-wise (28)

(3) Wrong way of comparing e.g. personal baselines

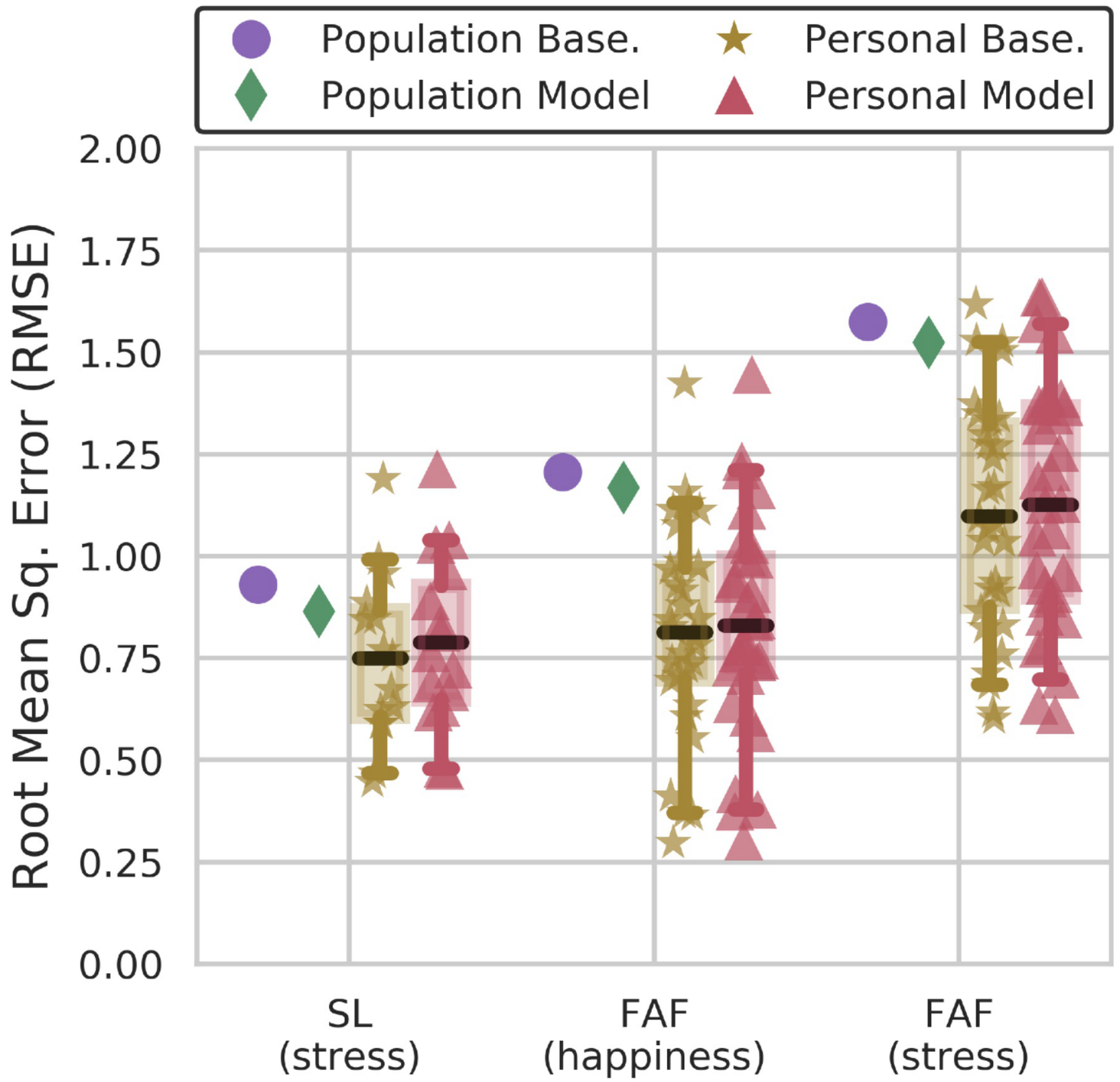
- Variance explained

W Orianna Demasi, Ben Recht

Personal vs group baselines



Machine learning often does not help

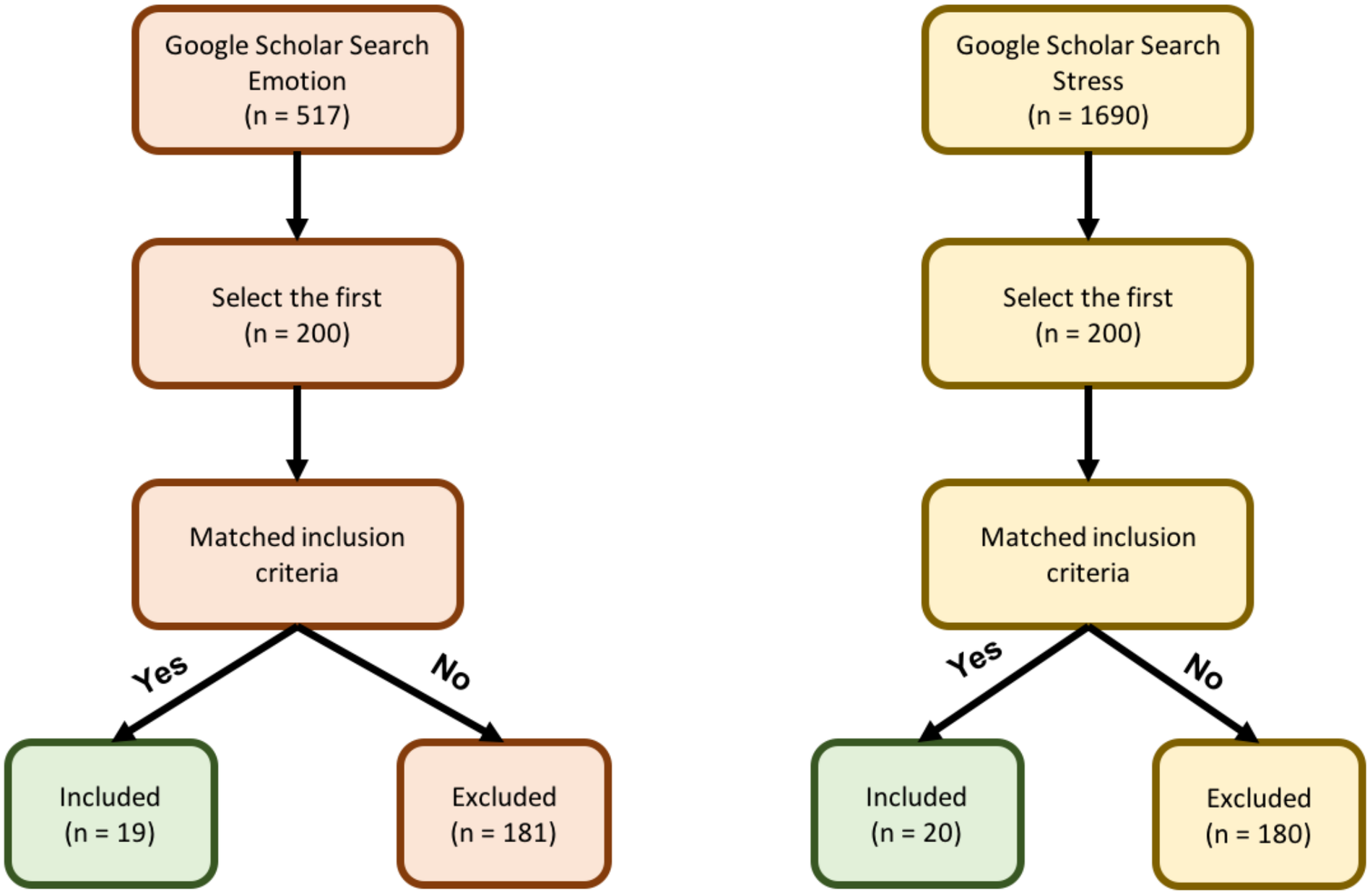


User lift

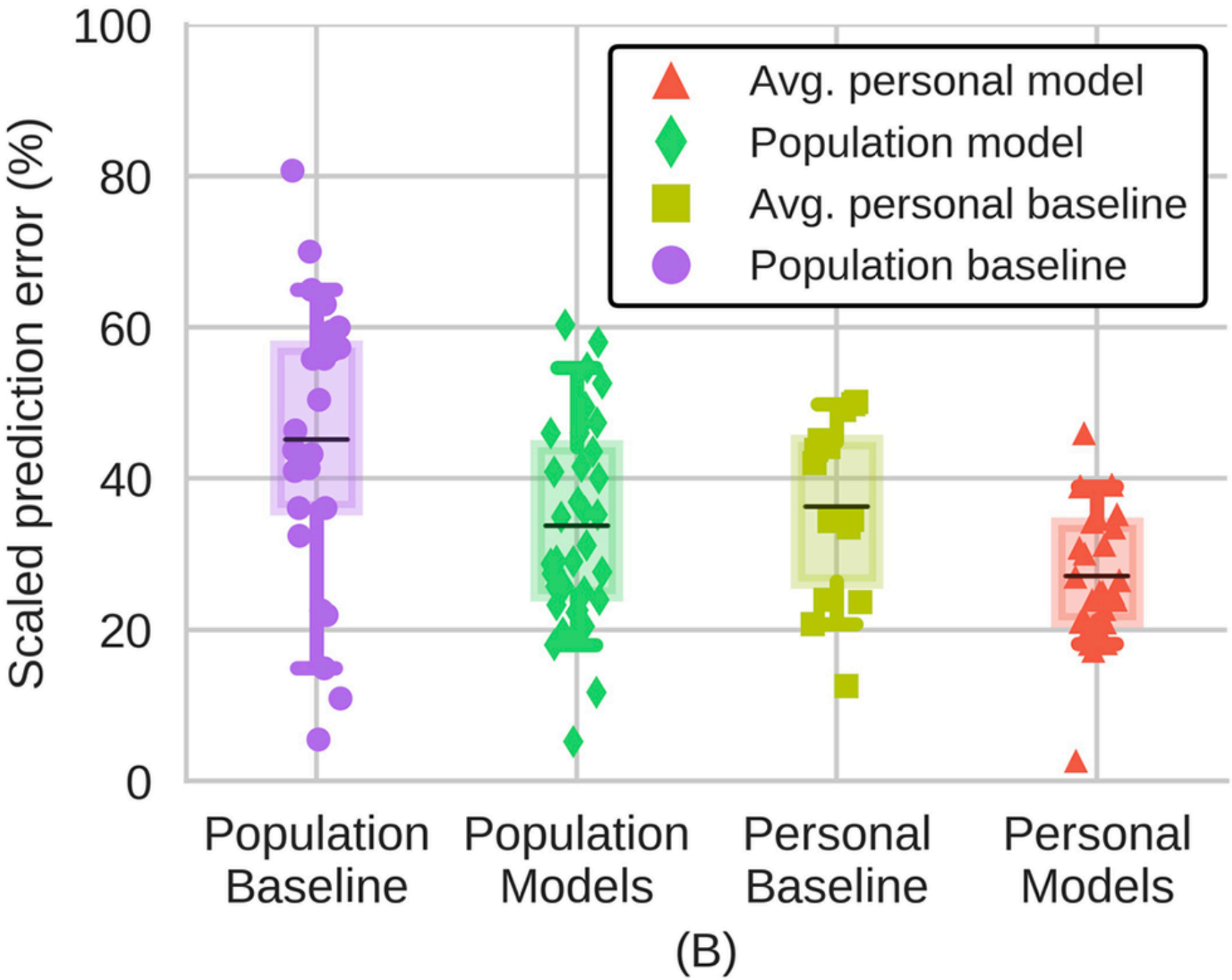
Dataset	Problem	Model	Avg. Personal Baseline Error	Avg. Personal Model Error	Avg. User Lift (Error)	p-value
SL—Stress	binary	Log.Reg.	29.19%	29.09%	**0.10**	.481
FaF—Happiness	binary	SVM(rbf)	16.51%	18.67%	**-2.17**	.967
FaF—Stress	binary	SVM(rbf)	25.17%	23.35%	**1.82**	.240
SL—Stress	regression	Elastic Net	0.75	0.78	**-0.03**	.988
FaF—Happiness	regression	Elastic Net	0.81	0.83	**-0.02**	.999
FaF—Stress	regression	Elastic Net	1.10	1.13	**-0.03**	1.000

https://doi.org/10.1371/journal.pone.0184604.t001

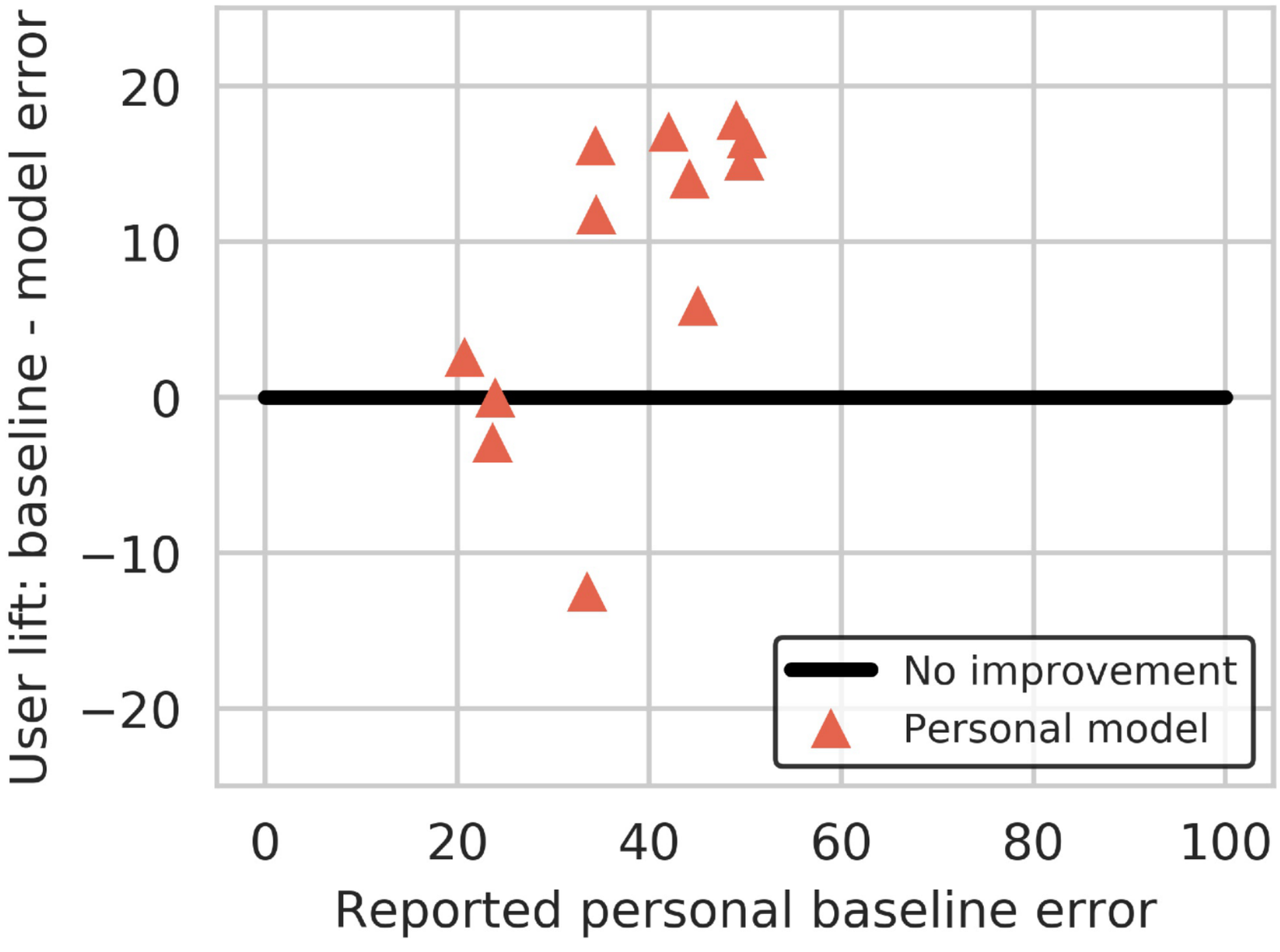
Literature review



Machine learning often does not help



Does ML even help?



(4) Wrong way of managing

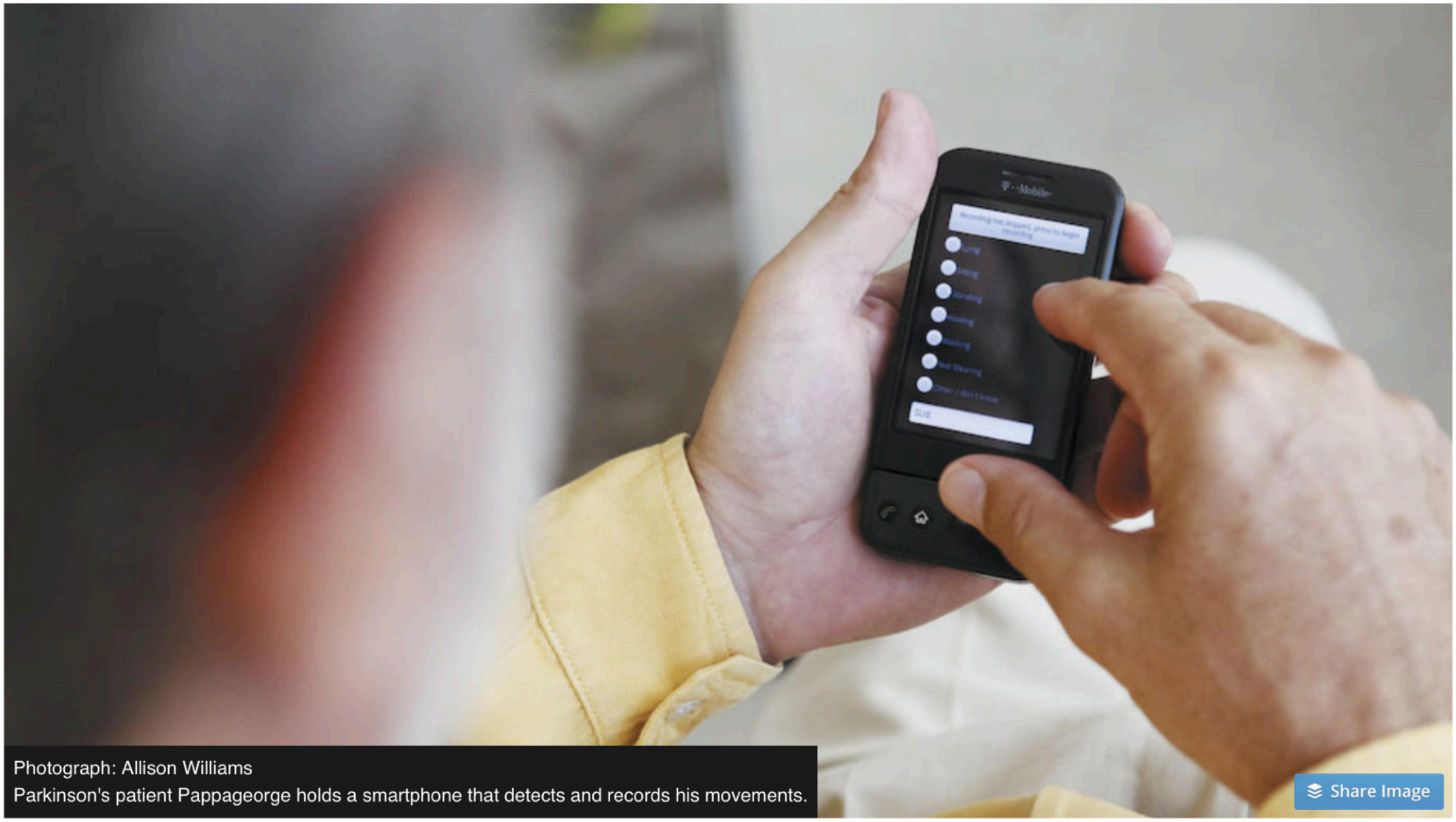
- Get data
- Give half of it to your ML collaborators
- Lock the other half away
- Get their algorithm
- Then test performance on the parts they have not seen

The many ways of leakage

- By not cross validating
- By cross validating wrongly
- By shared recruitment strategy
- By trainee

An app to track Parkinson’s disease

Can the technology behind cell-phone bowling change the lives of Parkinson’s patients?



Photograph: Allison Williams
Parkinson's patient Pappageorge holds a smartphone that detects and records his movements.

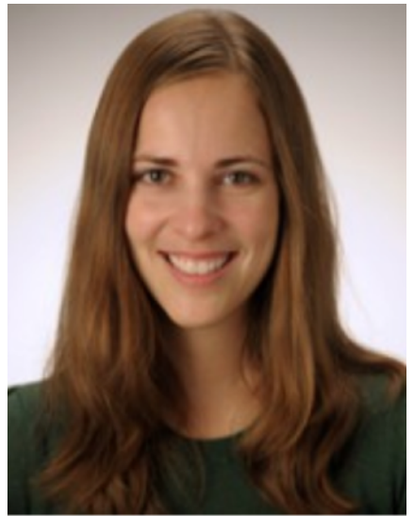
Share Image

IV) Towards computer vision-based automated infant neuromotor disorder diagnosis

Dr. Claire Chambers

Rachit Saluja

Wilson Torres



Dr. Laura Prosser

Dr. Michelle Johnson

Neuromotor developmental disorders cause lifelong disability and can be detected early

5 to 10% children are affected by developmental disorders (Rydz et al., 2005)

Cerebral Palsy: 2.11 per 1000 live births (Oskoui et al., 2013)
May be higher, 5 per 1000, in lower and middle income countries (Khandaker et al. 2018)

Early detection is crucial so as to maximize brain **plasticity** during treatment (Palmer, 2004)

Need for a quantified, sensitive and accessible diagnostic

Early diagnosis

Existing clinical methods (General Movements Assessment) have high specificity and are widely tested, but are:

- qualitative
- expensive
- inaccessible in resource-poor environments

Optic flow assessments:

- give only gross movement features
- not clinician interpretable

Approach

A database of ‘normative’ infant movements

Infant movements in a clinical setting

Compare to assess risk

‘Normative’ infant movements from YouTube

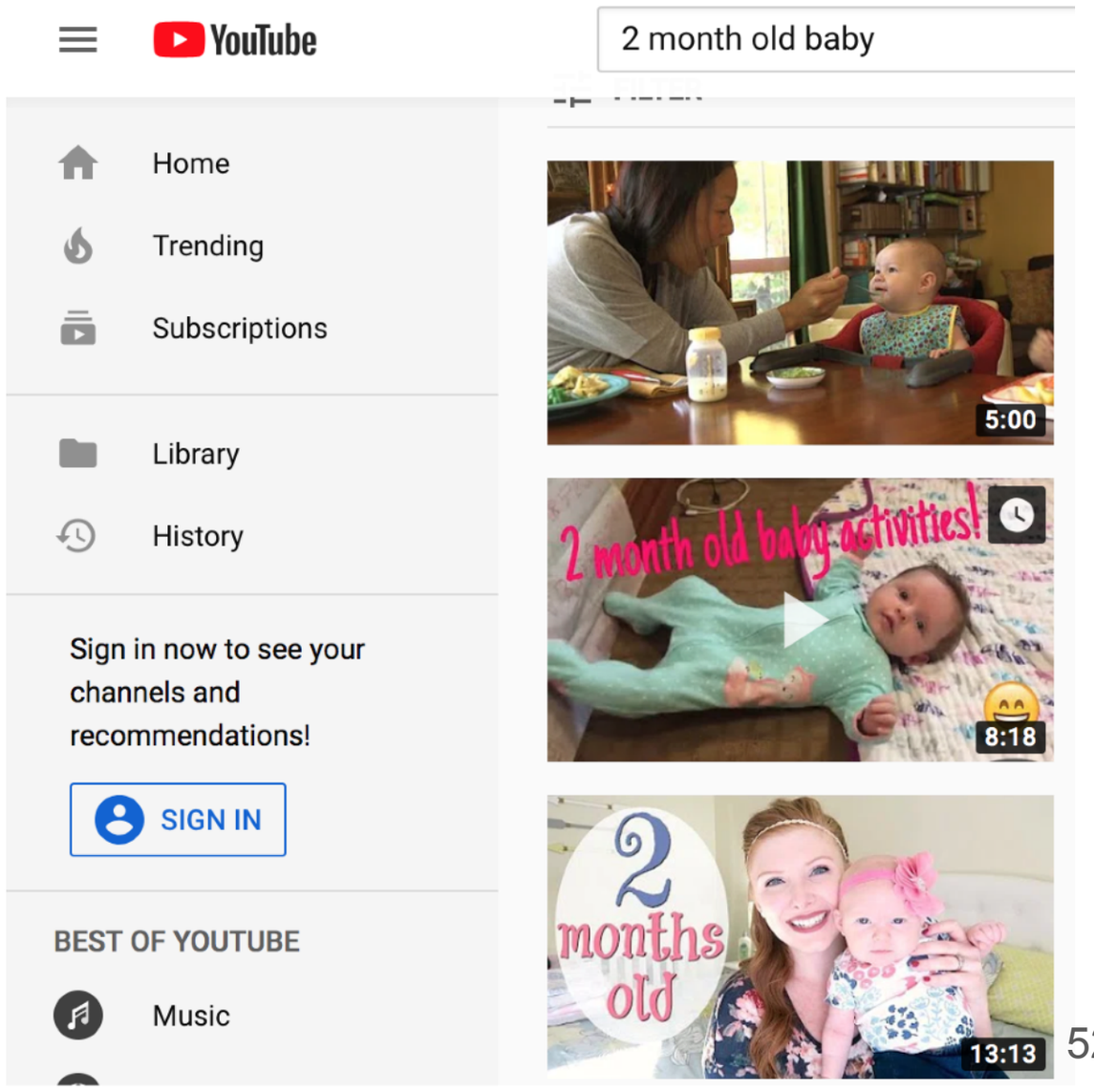
YouTube search terms such as:

- one, two, three, four, five, six months old baby
- ____ weeks old

Inclusion criteria:

- infant is non-occluded
- infants move independently
- infant body is present in the video within the frame
- Duration > 6 sec

385 videos found, and 85 included



Collecting infant movement data in a clinical setting

Data collected in the **Children's Hospital of Philadelphia**. Approved by ethics board.

Inclusion criteria:

- Infants cannot yet walk
- absence of history of cardiac, neurological or orthopedic condition
- Parents provide informed consent

GoPro camera used to record movements while in supine position.

Bayley Infant Neurodevelopment Screener (BINS) was used by clinical to assess neuromotor risk. **19** infants assessed. 5 low-risk, 9 moderate-risk, 5 high-risk.

Using computer vision-based pose estimation to extract infant pose

OpenPose (Cao et al., 2018):

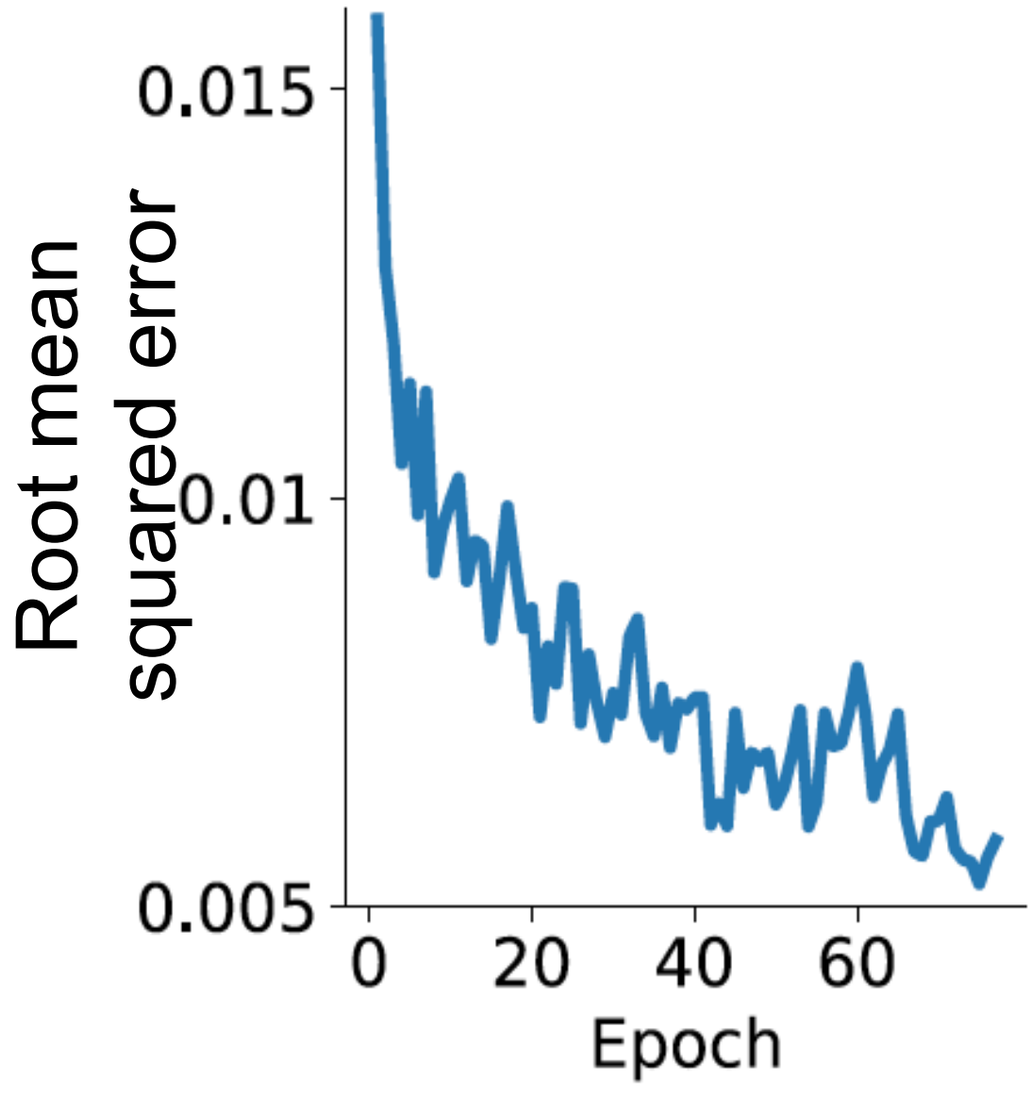
- nose, neck, ears, eyes, shoulders, elbows, wrists, hip, knees, and ankles

OpenPose initially provided messy estimates for infants because:

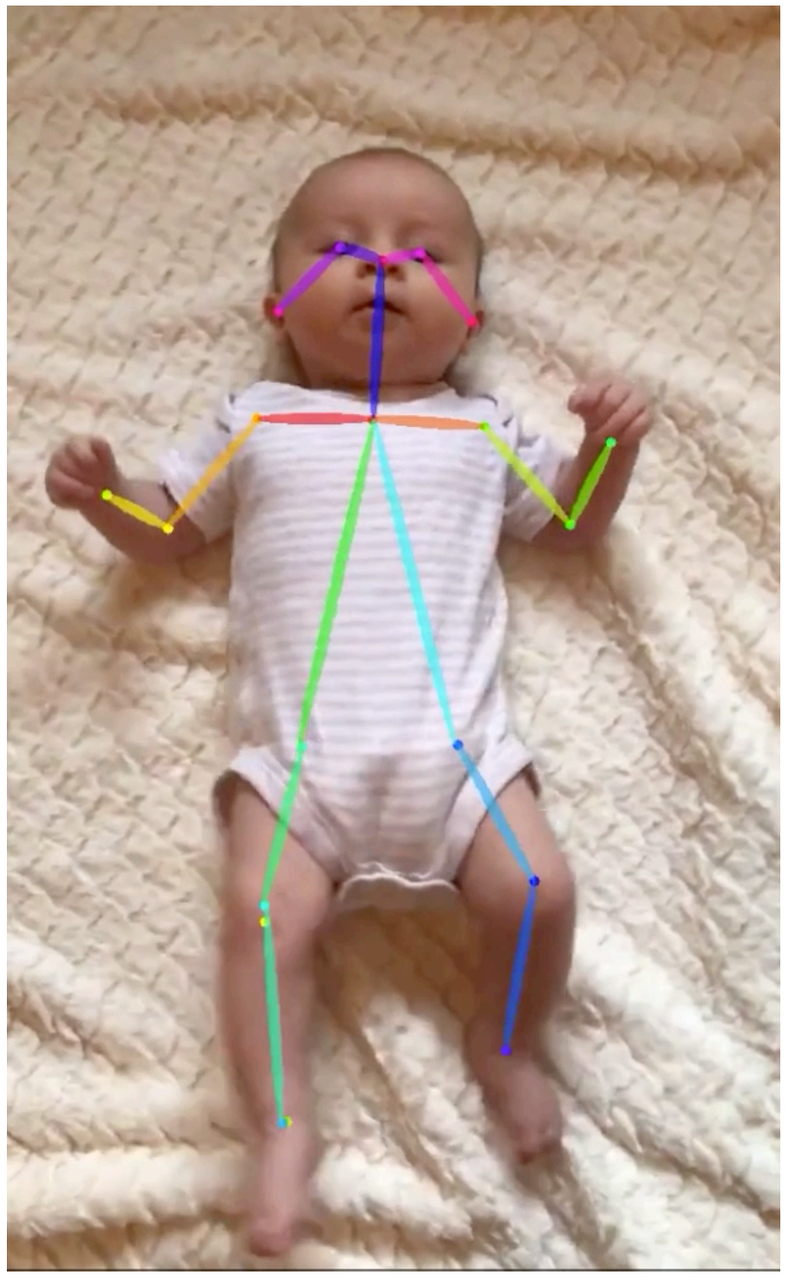
- infant body proportions are different from adults
- Such infant images are missing from the original training dataset (COCO and MPII).

OpenPose domain adaptation

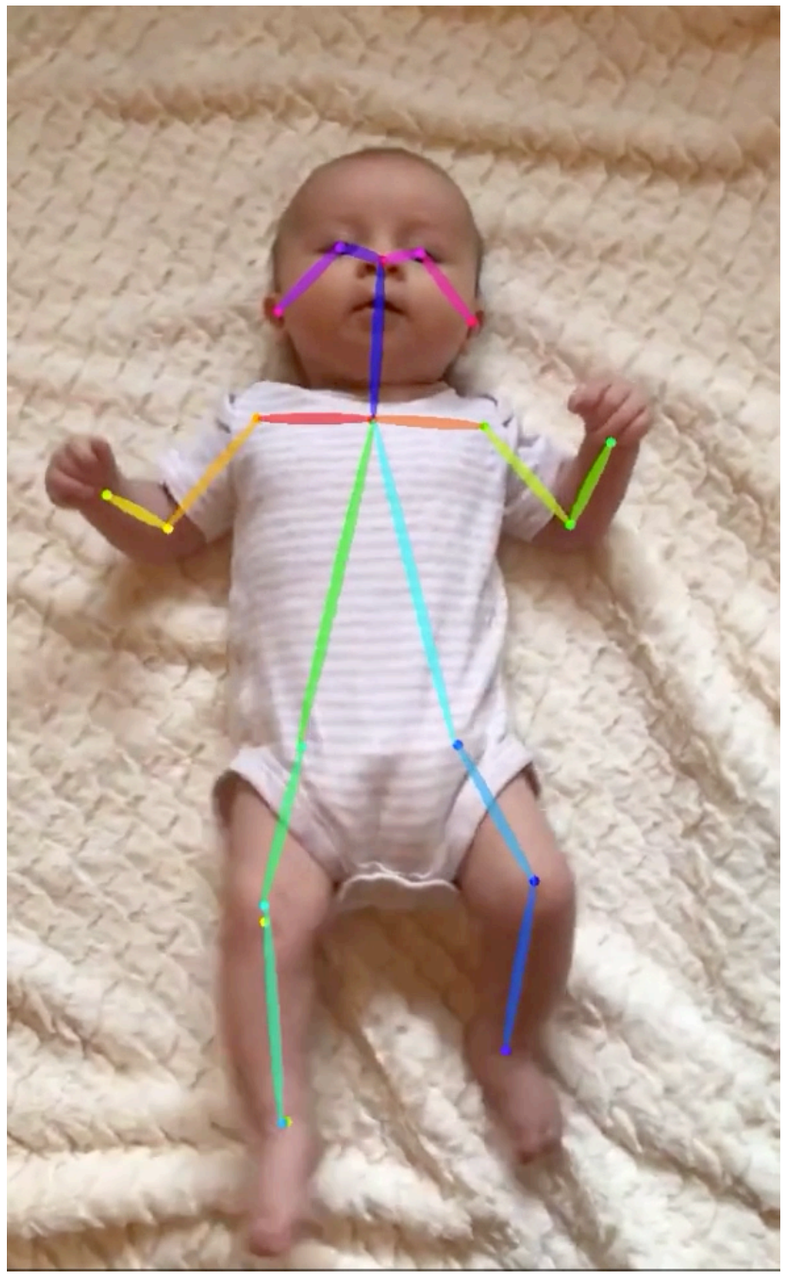
- Keypoints for ~9000 frames were manually labeled using *Vatic*
- 8003 frames in the training set and 1036 frames in the test set.
- The test frames are from videos unseen during training.
- Gradient descent for 75 iterations.
- Minimize the error relative to the ground truth manual labels.



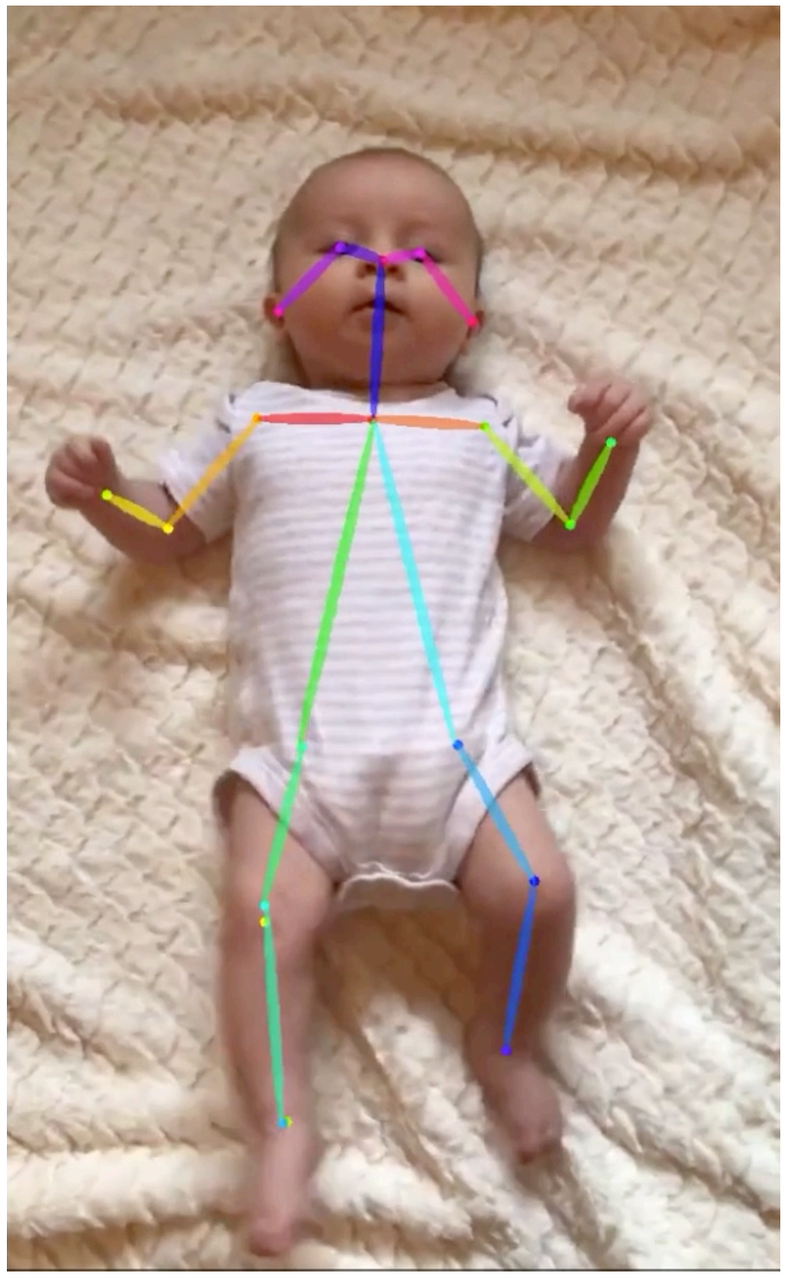
The network worked better on infants after retraining



The network worked better on infants after retraining

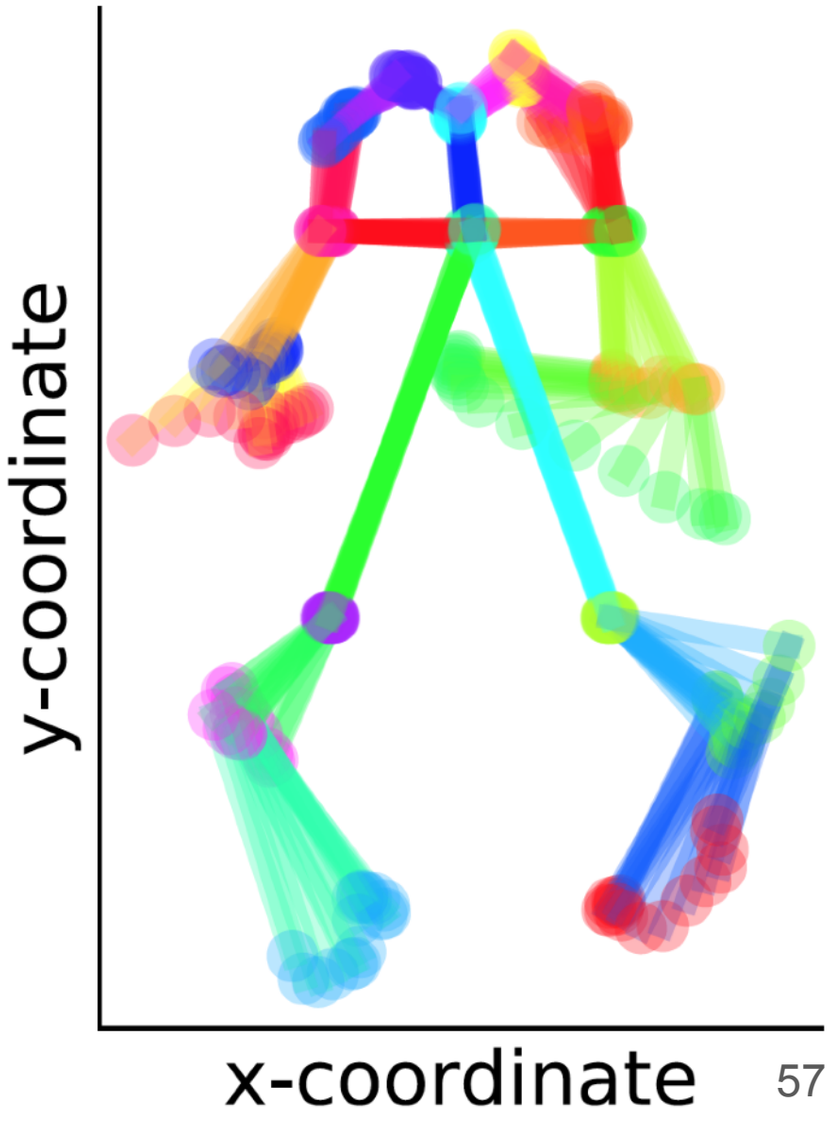


The network worked better on infants after retraining



Cleaning the infant pose raw data

- outlier removal: interpolate and drop points that are greater than two standard deviations (0.1 s bins)
- smoothing using moving average of 1 sec
- camera movements were dealt with by fixing a reference body part (trunk)
- lengths were normalized by trunk length

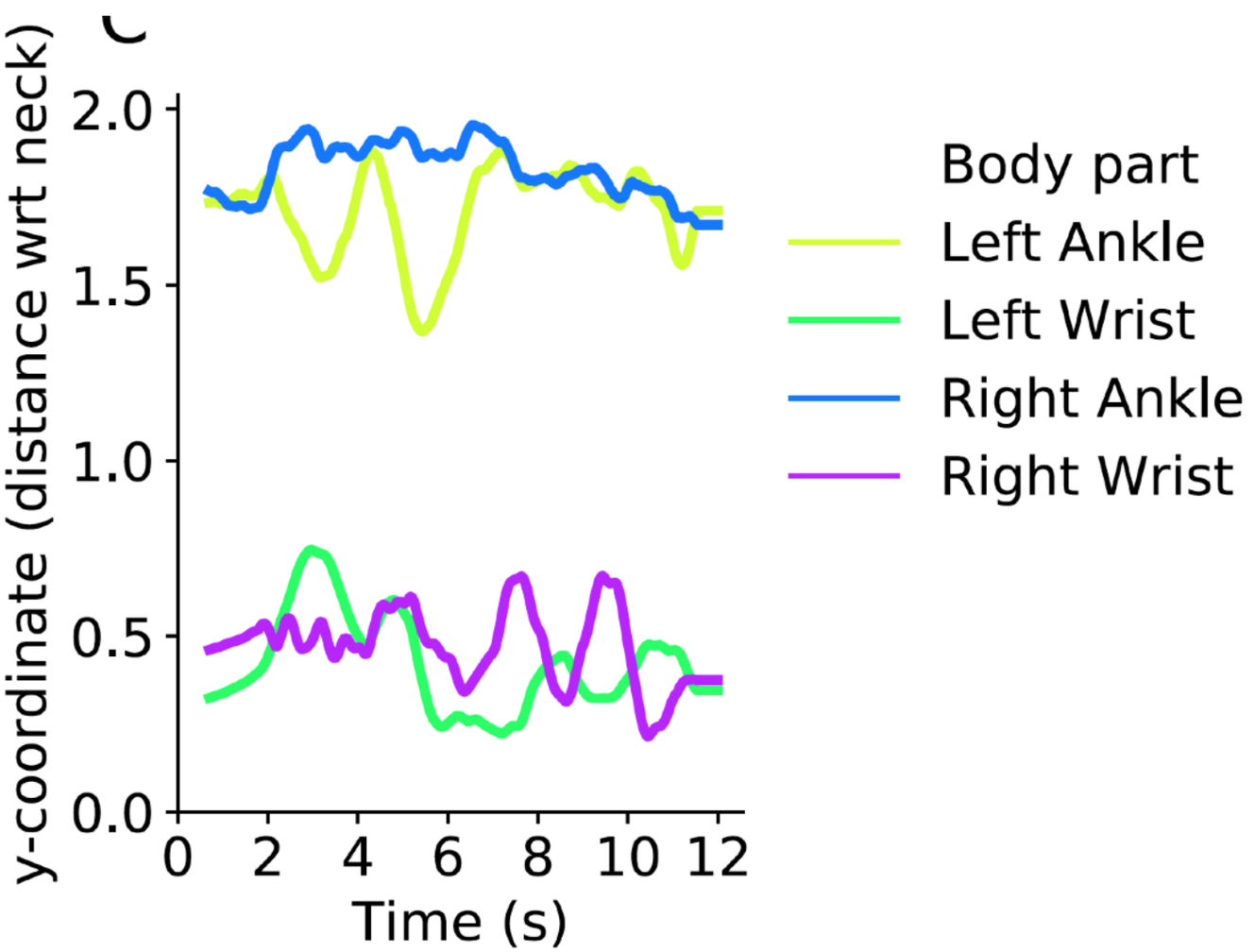


Old fashioned features

52 features in all

For the positions of the extremities (wrists/ ankles) and joint angles (elbows/knees) on both left and
right side of the body, we included:

- median position/angle
- IQR of position/angle
- median speed
- IQR of speed
- IQR of acceleration
- mean entropy
- left-right cross correlation



Naive Bayesian surprise metric

- assumes normal distribution and feature independence
- normalized the metric with respect to the ‘normative’ database
- estimate the log probability that a given infant’s movements are drawn from the ‘normative’ distribution

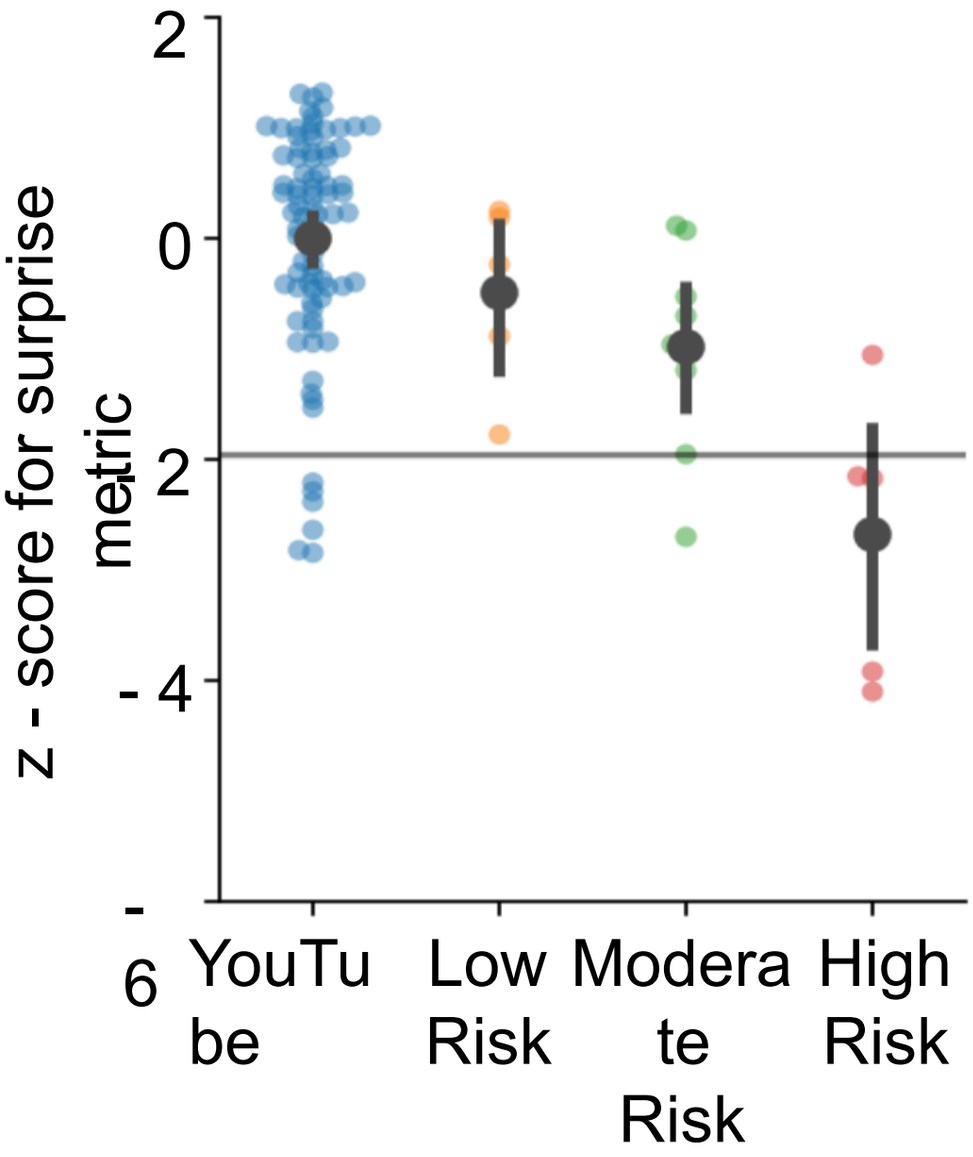
$$p\left(x_1, \dots, x_n \middle| \mu_{i,H}, \sigma_{i,H}^2\right)$$

$$= \prod_{i=1}^{n} p\left(x_i \middle| \mu_{i,H}, \sigma_{i,H}^2\right)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{i,H}^2}} e^{\frac{-(x_i-\mu_{i,H})^2}{\sigma_{i,H}^2}}$$

$$\Psi = -\ln p = -\sum_{i=1}^{n} \left(\frac{1}{2} \ln\left(2\pi\sigma_{i,H}^2\right) + \frac{(x_i - \mu_{i,H})^2}{2\sigma_{i,H}^2} \right)$$

Predicted risk corresponds to clinician-assessed risk



Chambers, Seethapathi, et al., 2019. Towards accessible computer vision-based diagnosis of infant neuromotor disorders. (in prep.)

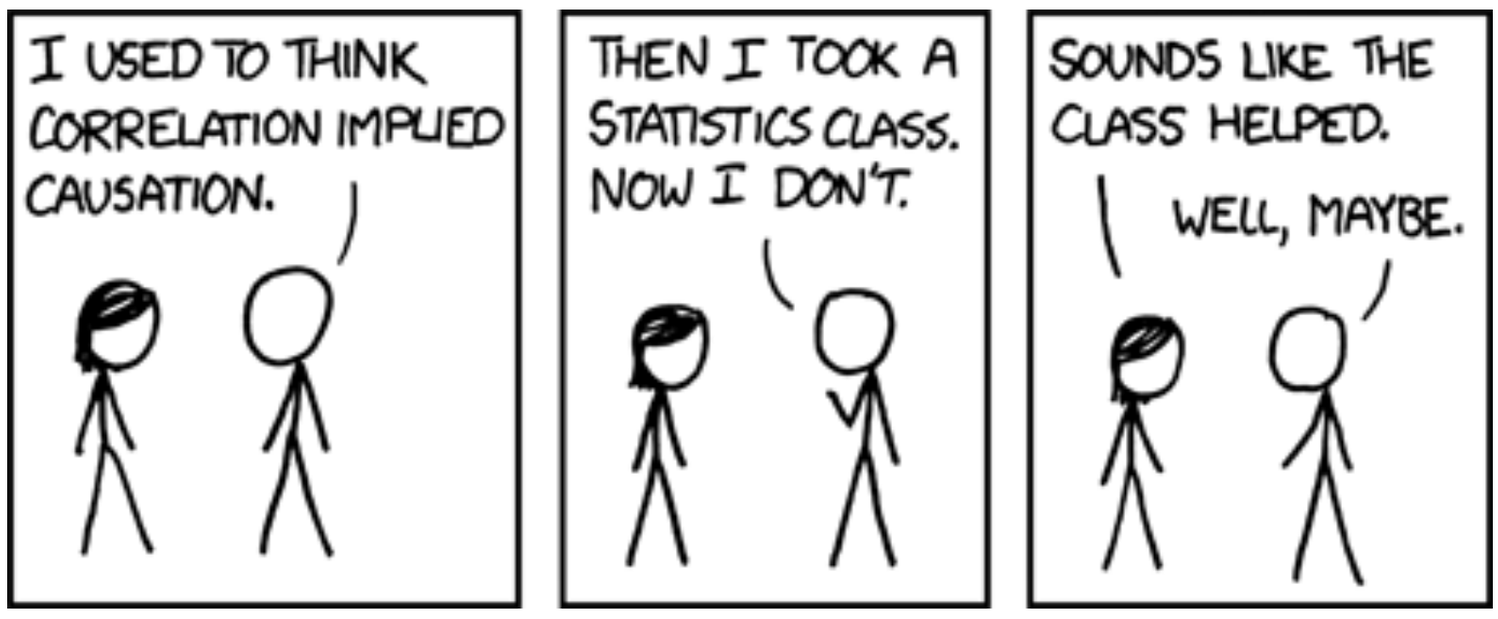
V) causality and pseudo experiments

cau·sal·i·ty

/kôˈzalədē/

noun

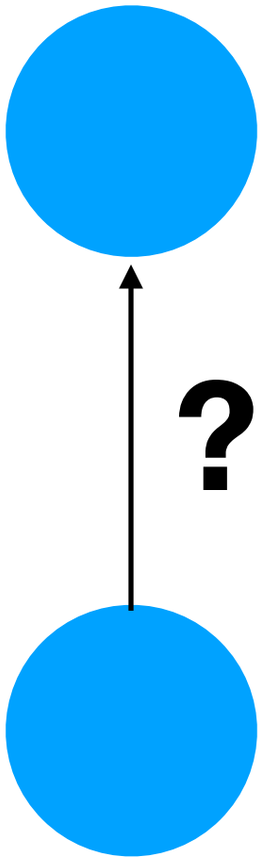
1. the relationship between cause and effect.
2. the principle that everything has a cause.



Definition of causality

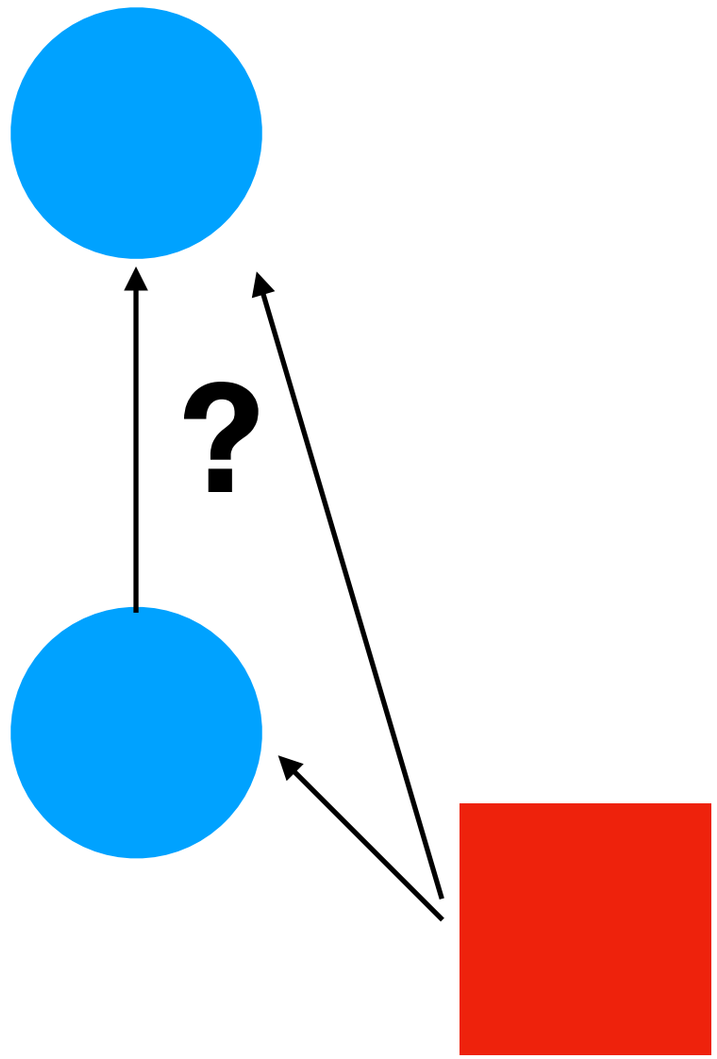
- Let a and b be events
- Causation exists if:
    - if we had changed a to a*, the probability for b would have been different

Why causality is hard: Confounding



E.g. Hormone Replacement Therapy,
Buying extra insurance

Why causality is hard: Confounding



E.g. Hormone Replacement Therapy,
Buying extra insurance

A continuum of confounding

- No confounders: e.g. atari, imagenet
- Few confounders: starcraft
- Countless confounders: Medicine
- 10^11 confounders: brains

Medicine

- Countless thresholds
- Few controllable variables
- Everything is confounded
- Big datasets
- The ultimate control problem

Simulate a trivial causal system

$$x_{t+1} = Ax_t + \epsilon$$

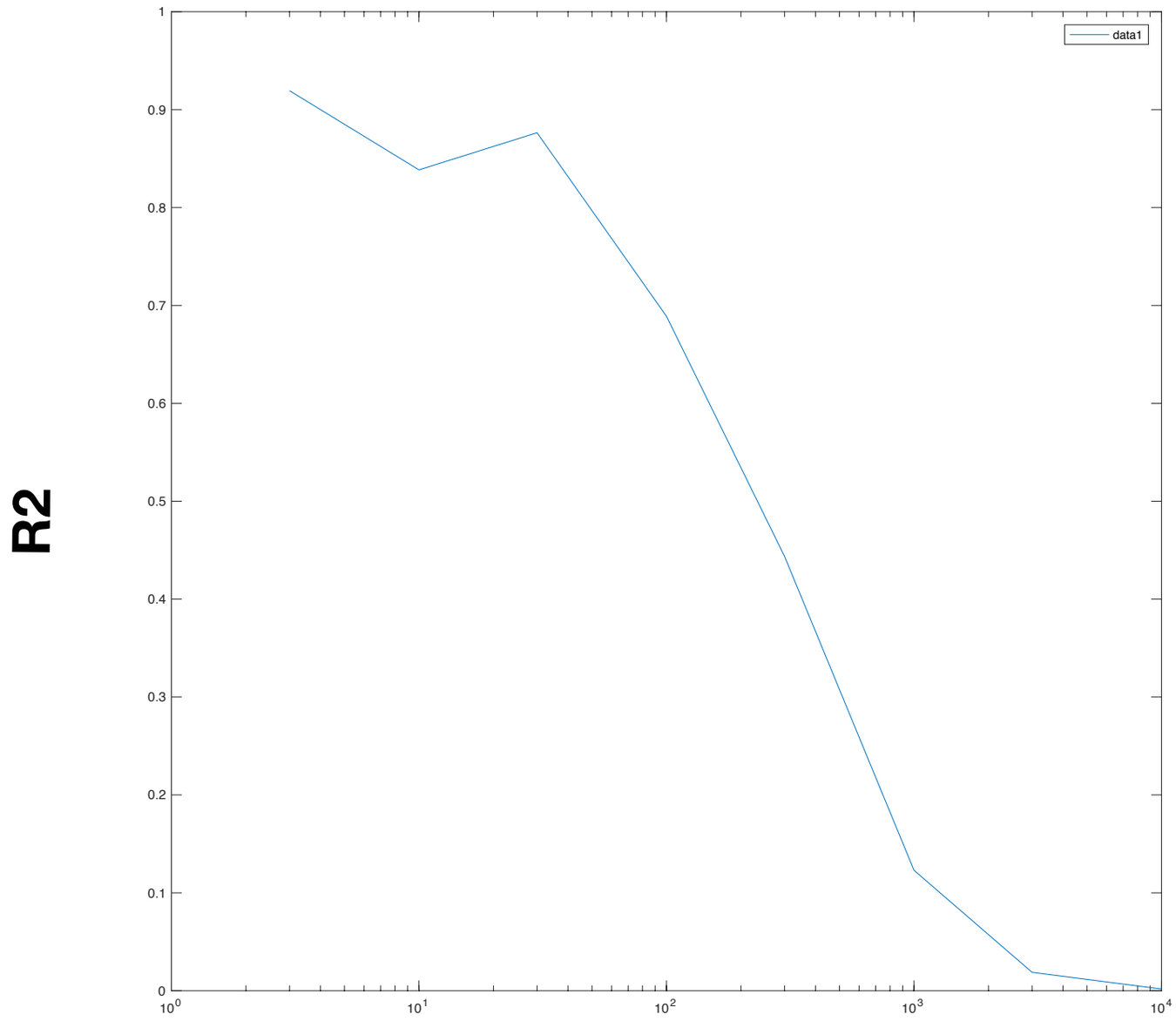
Where

$$\epsilon \sim \mathcal{N}(0, \Sigma)$$

$$\Sigma = diag(nL)$$

Choose A: sparse binary (p=.1), largest SV=.99

Delayed Correlation vs Causation



Popular solutions

- (1) Randomized perturbations (Experiments)
  - RL exploration $\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$
- (2) ML Bayesian network/ saturated structural equation model

$$p(x) = \prod p\left(x_i \mid Pa(x_i)\right)$$

- (3) Model comparisons

Popular solutions

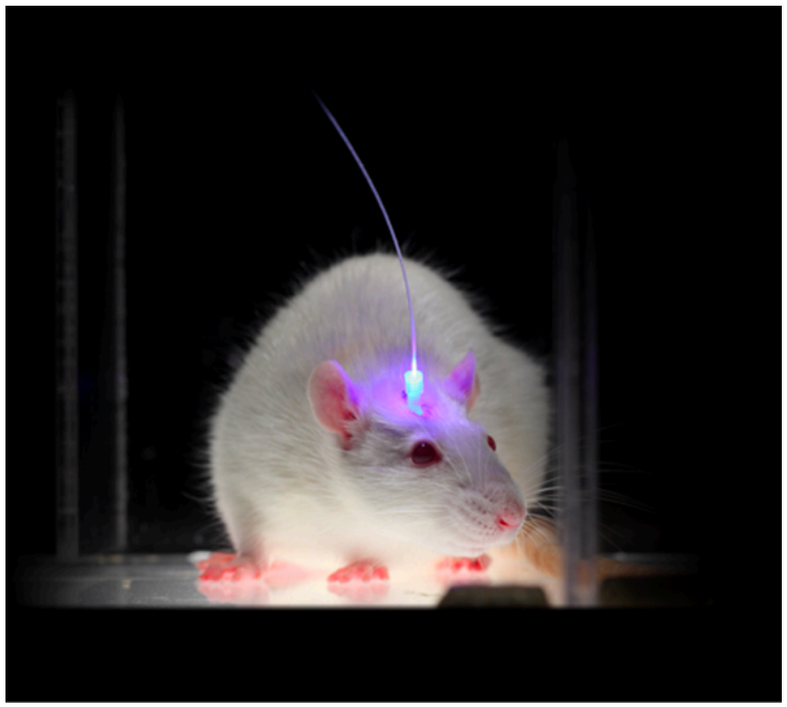
- (1) Randomized perturbations (Experiments)
  - RL exploration $\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$
- (2) ML Bayesian network/ saturated structural equation model

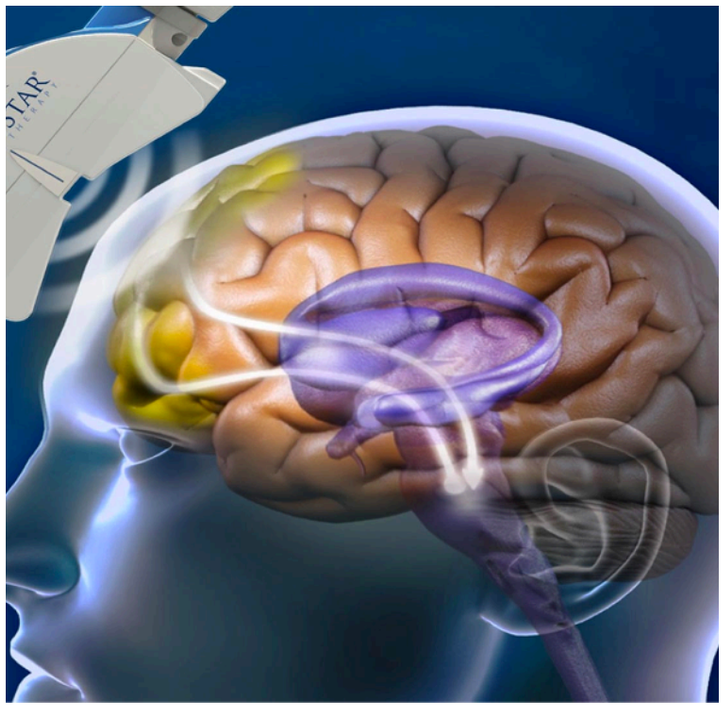
$$p(x) = \prod p\left(x_i \mid Pa(x_i)\right)$$

- (3) Model comparisons

- Quasiexperiments

Perturbations





Implicit assumptions: we randomly perturb what we care about

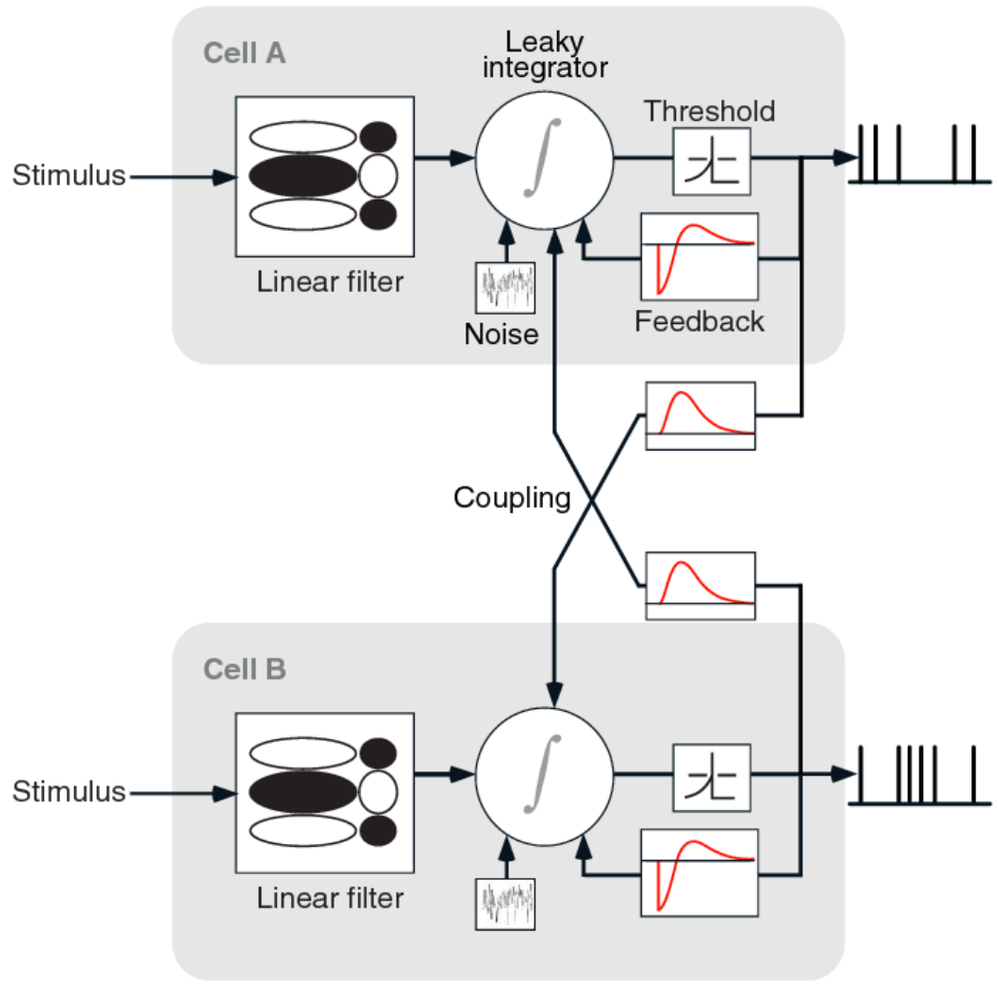
Low-D!, expensive, unethical, dangerous

Model comparison

- Have two models with distinct internal causality
- Choose the one that describes data better (p<.05)

$$AIC = 2k - 2\ln(\hat{L})$$

Saturated structural equations + DAGs



Chichilnisky

$$L = \sum \log \lambda_\theta(t_{sp}) + \int \lambda_\theta(t)dt$$

Assumptions: causal sufficiency, correct functional form, …

Paninski, Pillow, Butts, Sahani, …, yours truly

Pearl/ DAGs

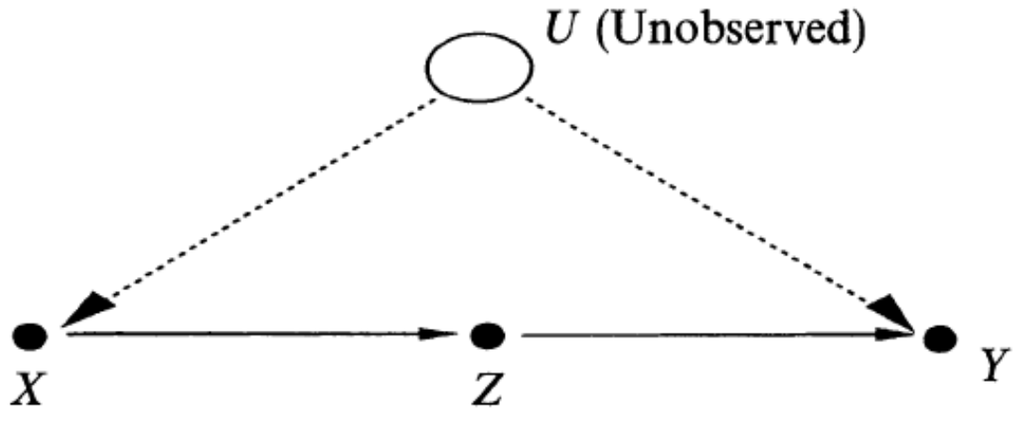


Fig. 3. A diagram representing the front-door criterion.

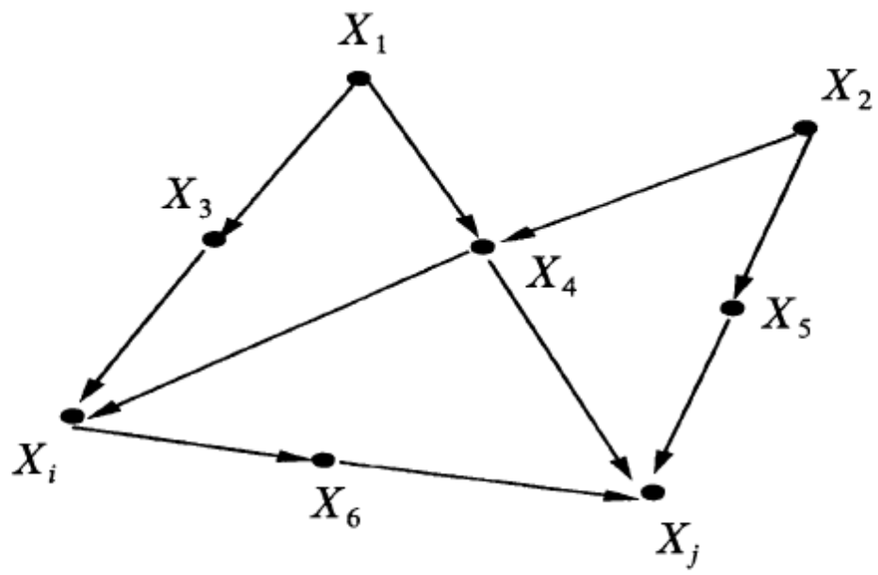
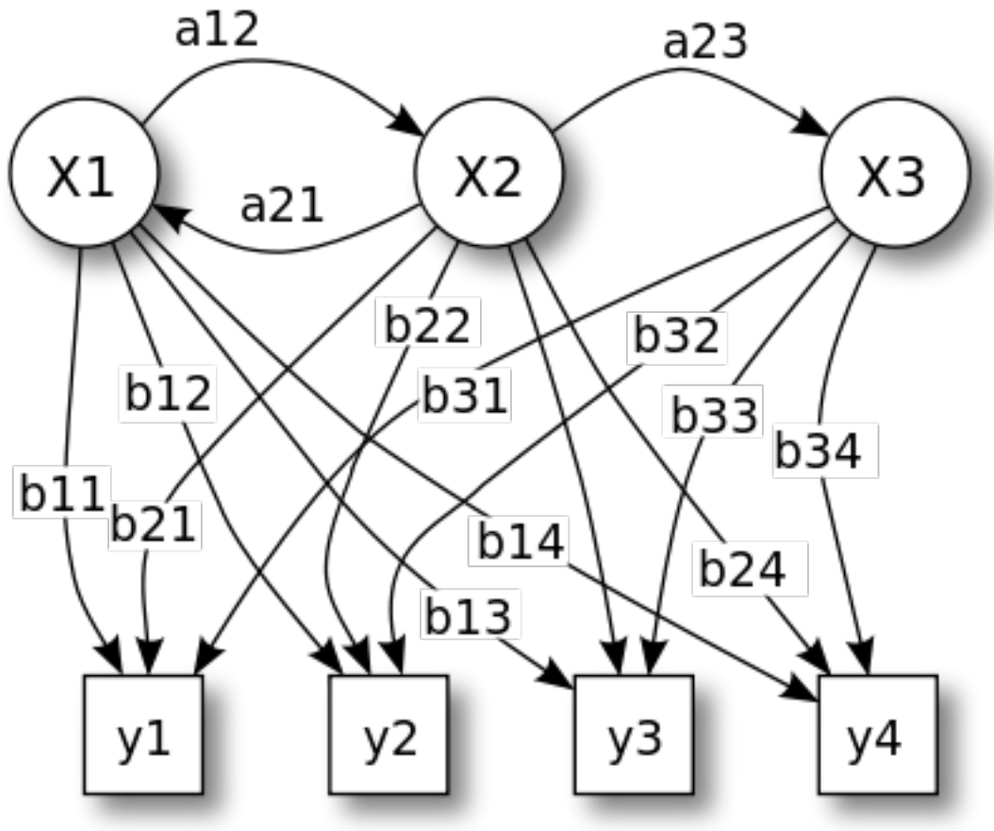
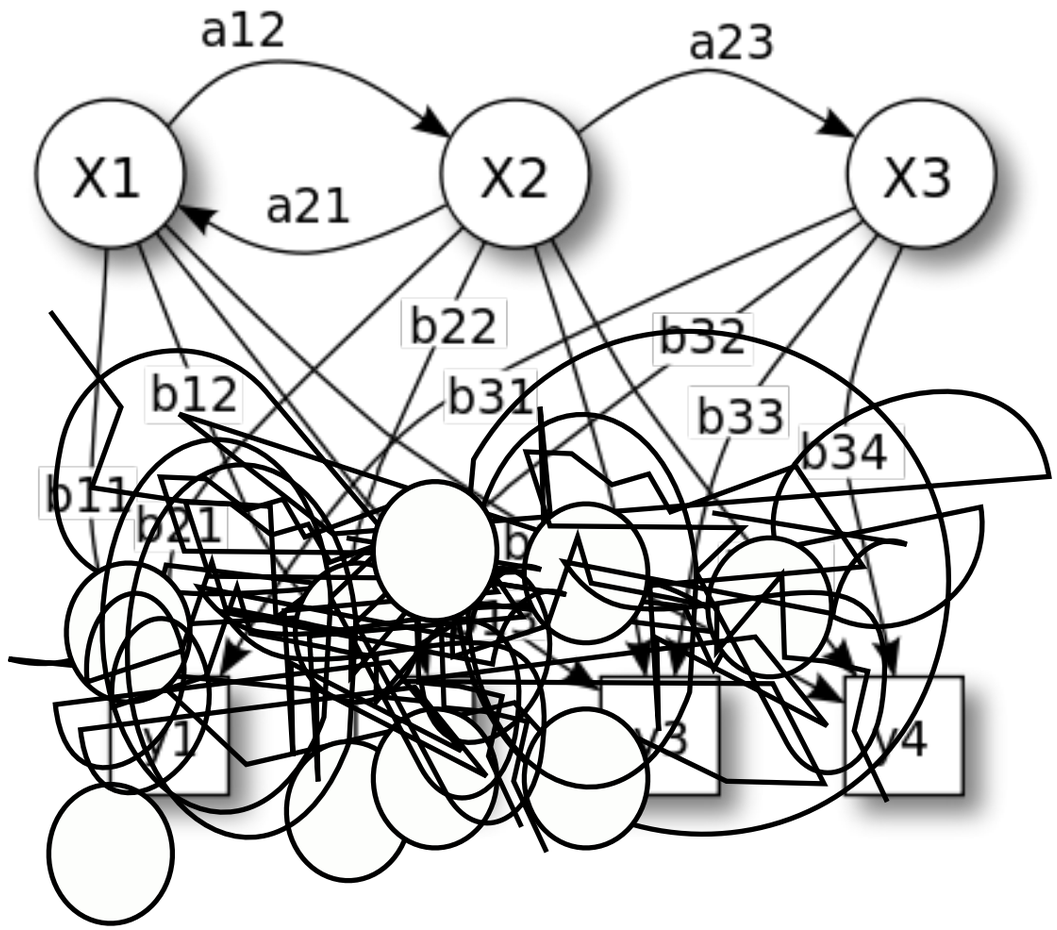


Fig. 2. A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ or $\{X_4, X_5\}$ yields a consistent estimate of $\mathrm{pr}(x_j|\check{x}_i)$.

Does the world look like this?



Or this?



Potential outcomes

Untreated $Y_i(0)$

Treated $Y_i(1)$

Rubens, Imbens, Athey

No bias in RCT

Measurement	Y_0	Y_1
1	1.2	3.7
2	3.5	9
3	2	6.3
4	3.4	6.5
5	4.1	11.1
6	3.6	4.9
...	...	...

$$TE = E\left(Y_1 - Y_0\right)$$

$$\approx \frac{1}{N}\sum Y_1(i) - Y_0(i)$$

No bias in RCT

Measurement	Y_0	Y_1
1	1.2	3.7
2	3.5	8
3	2	6.3
4	3.4	6.5
5	4.1	11.1
6	3.6	4.9
...	...	...

$$TE = E\left(Y_1 - Y_0\right)$$

$$\approx \frac{1}{N}\sum Y_1(i) - Y_0(i)$$

$$\approx \frac{1}{N_1}\sum_{S_1} Y_1(i) - \frac{1}{N_0}\sum_{S_0} Y_0(i)$$

Quasiexperiments

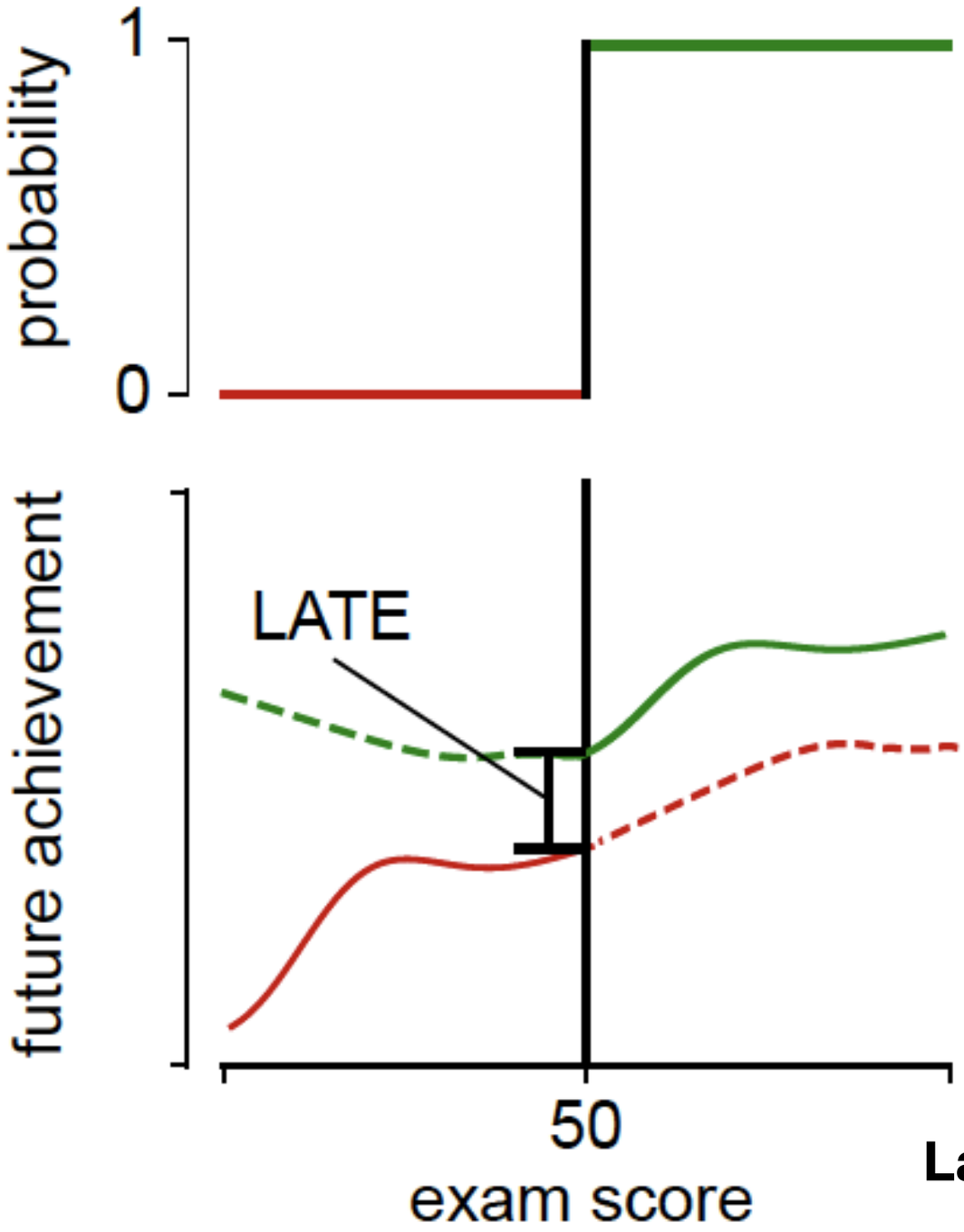
Observational — **Quasi-experiments** — **Experiments**

Causal validity

Idea: find something that is locally kinda random

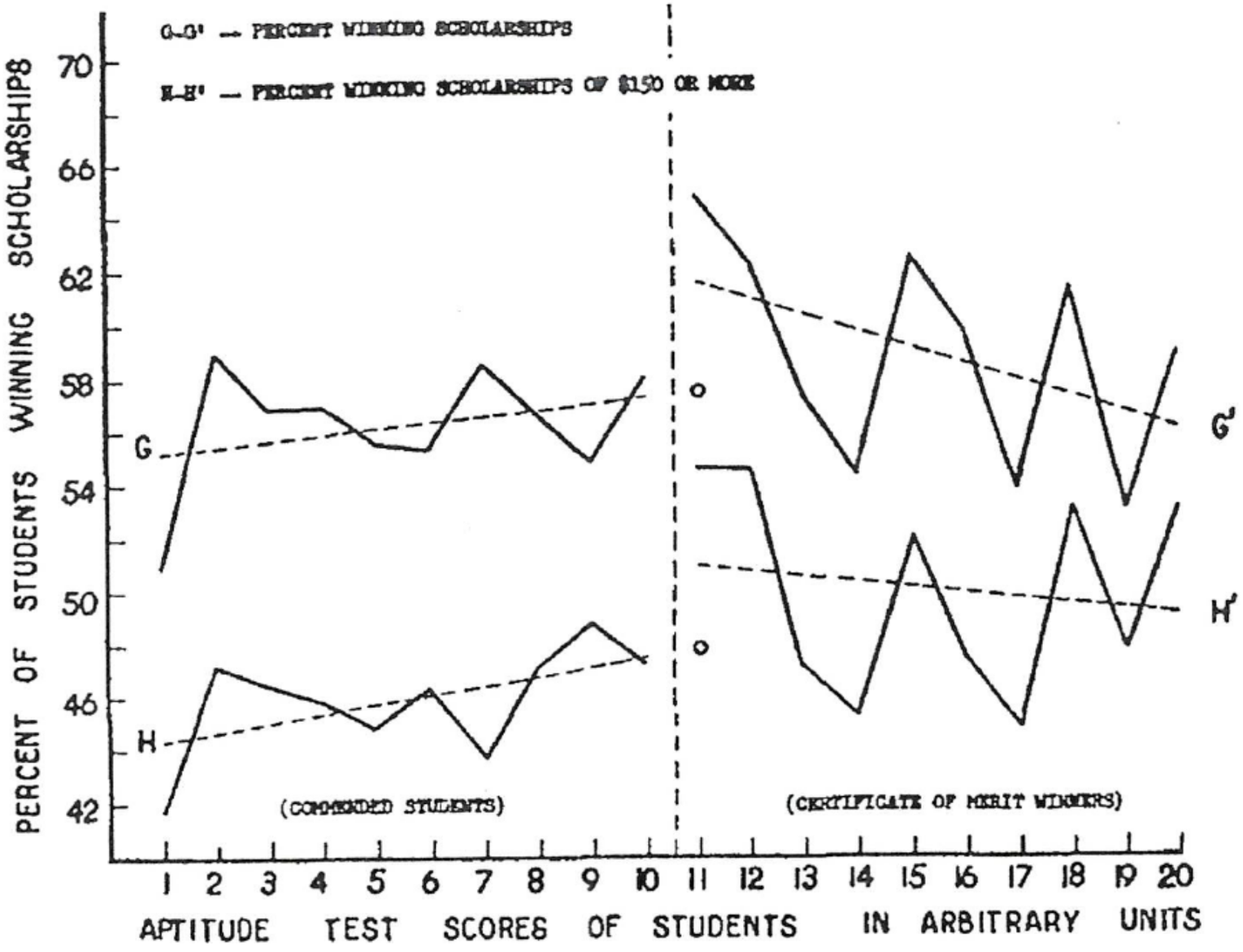
Marinescu, Lawlor, Kording, Nature Human Behavior, In press

Estimate effect of certificate of merit



Lawlor, Marinescu, Kording, NHB, in press

Does winning merit certificate help?



Thistlewaite and Campbell 1960

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 (X_i - X_c) + \hat{\beta}_3 Z_i (X_i - X_c) + e_i$$

Sanity checks

- Cheating
  - visible as discontinuity in co-variates
- Fuzziness
  - visible as smooth treatment changes

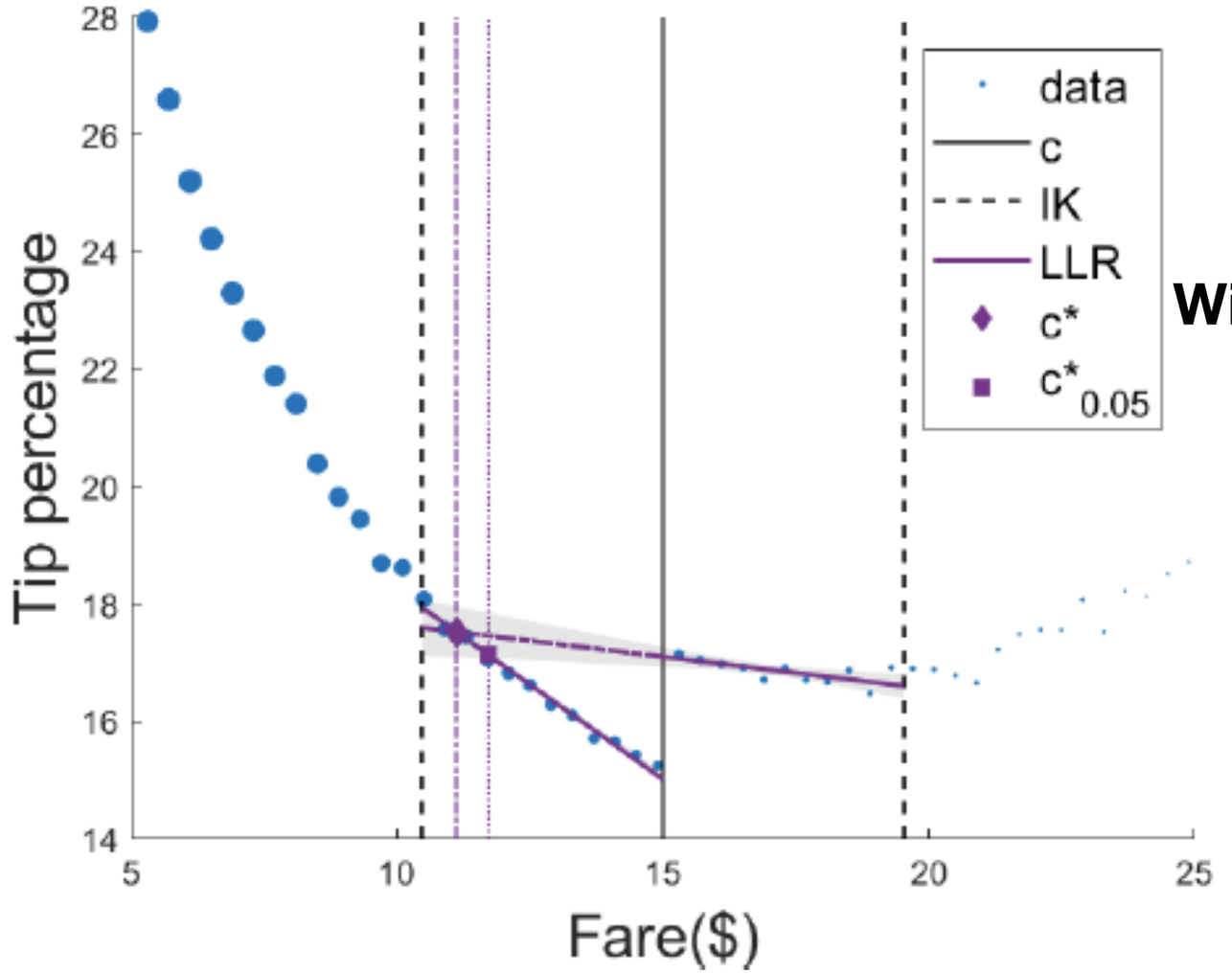
Variance of RDD estimators

- requires ~3 times as many samples as experiment

$$Var_{RDD}(\alpha_0) \propto \frac{3\sigma^2}{n_{bandwidth} p^2}$$

- how to choose bandwidth? E.g. crossvalidation

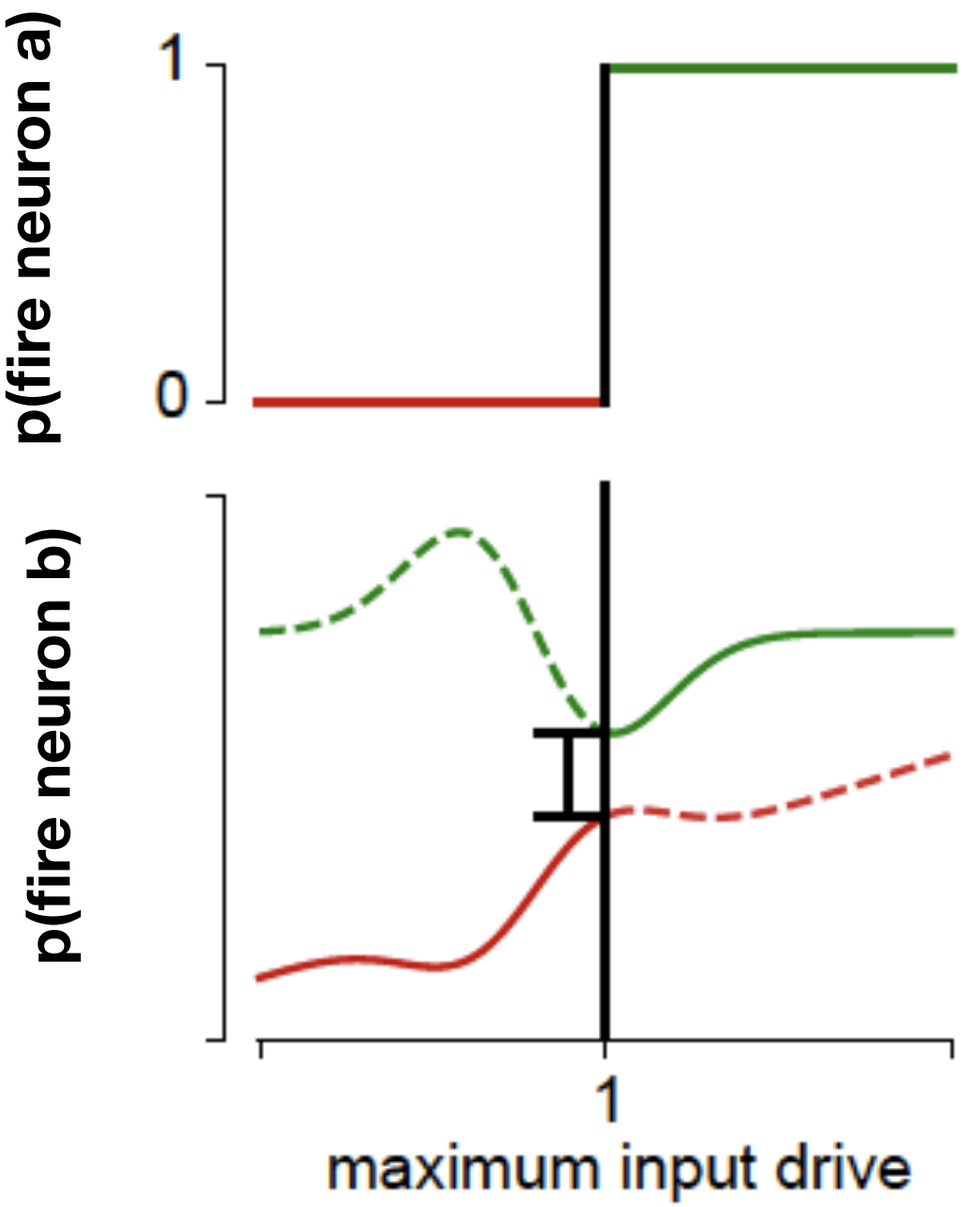
Obvious optimization problem (linear)



With Marinescu, Triantafillou, forthcoming

Reinforcement learning without Exploration

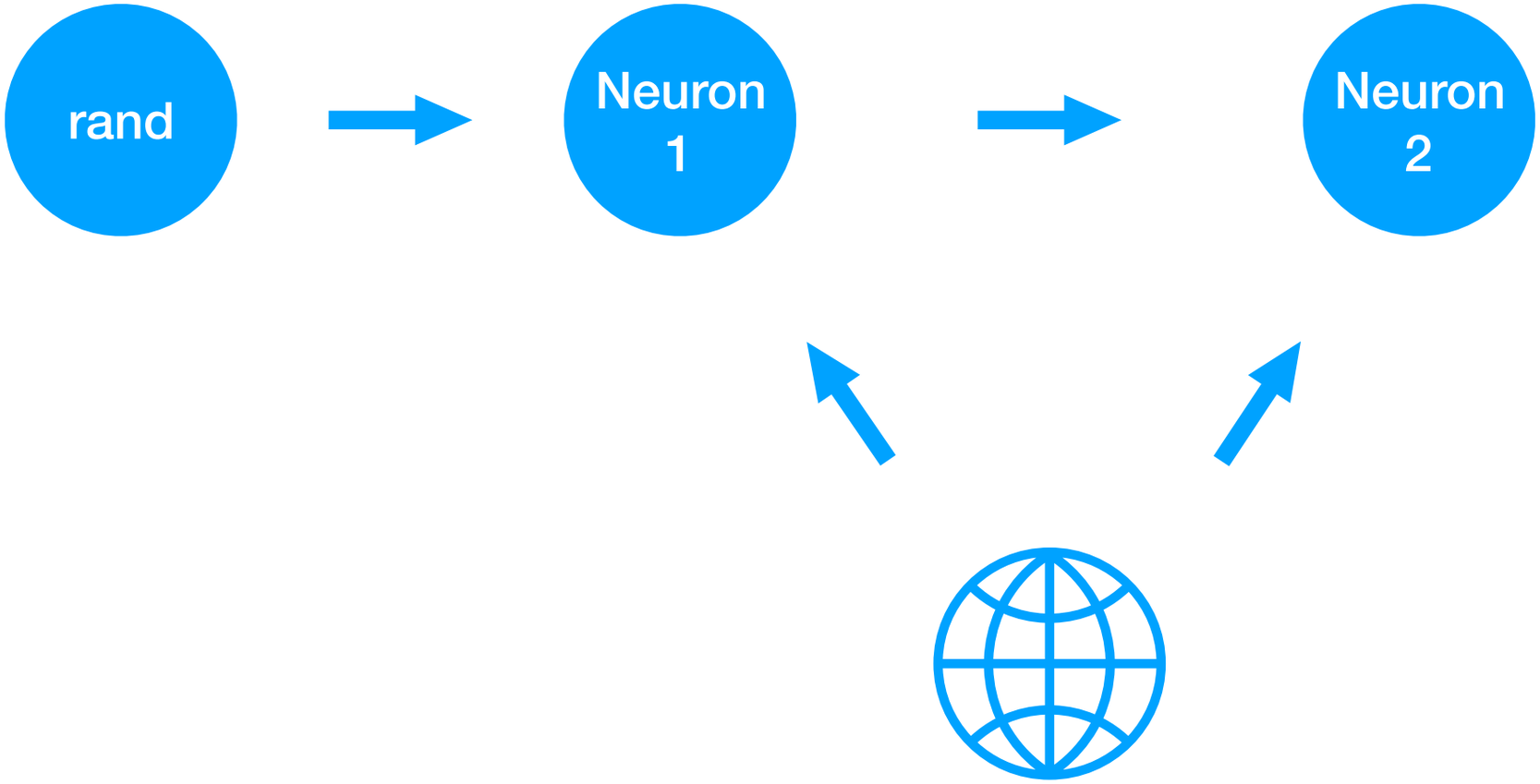
Neural data analysis: intracellular recordings



Preplanned RDD

- Often more ethical: e.g. help the poorest districts
    - instead of random
- same in medicine, apply to those who are highest risk

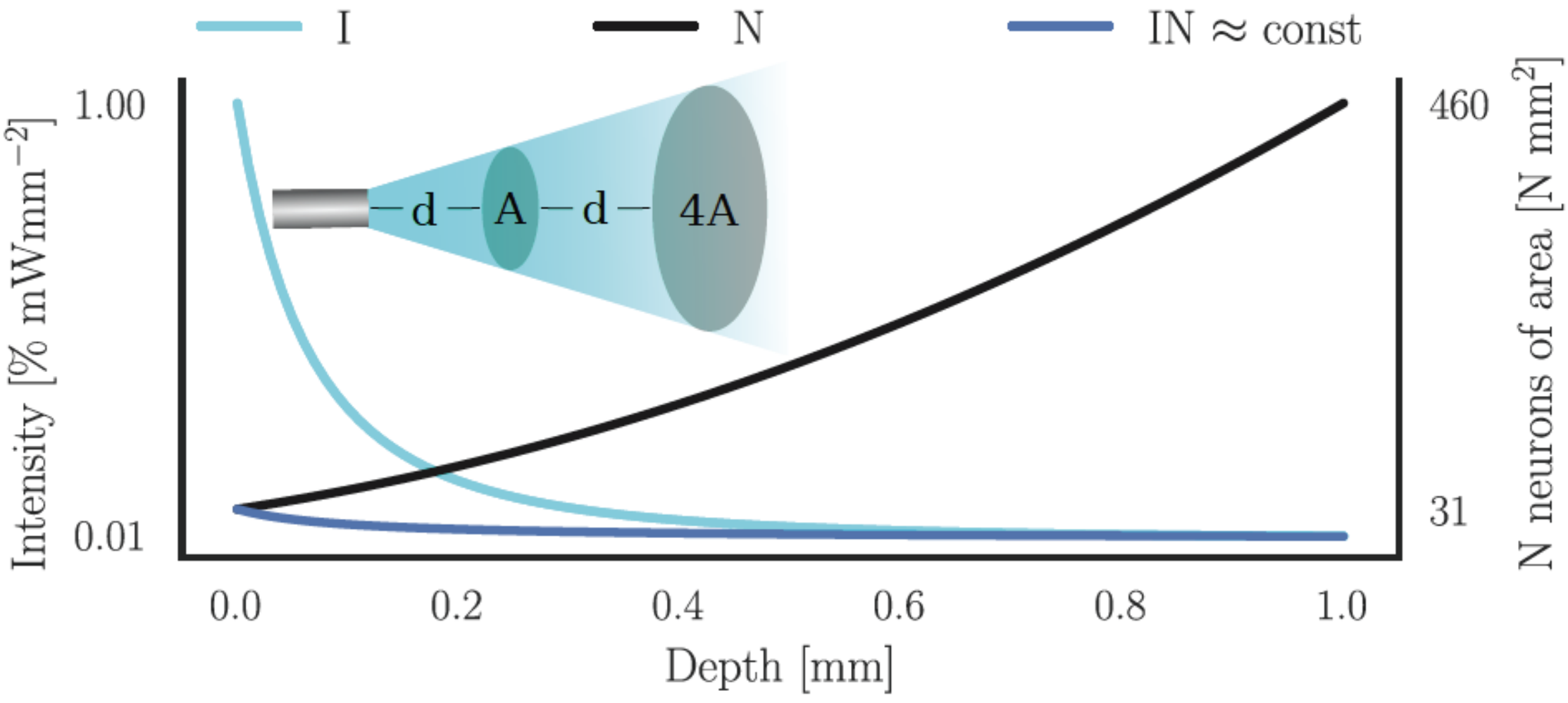
Instrumental variable



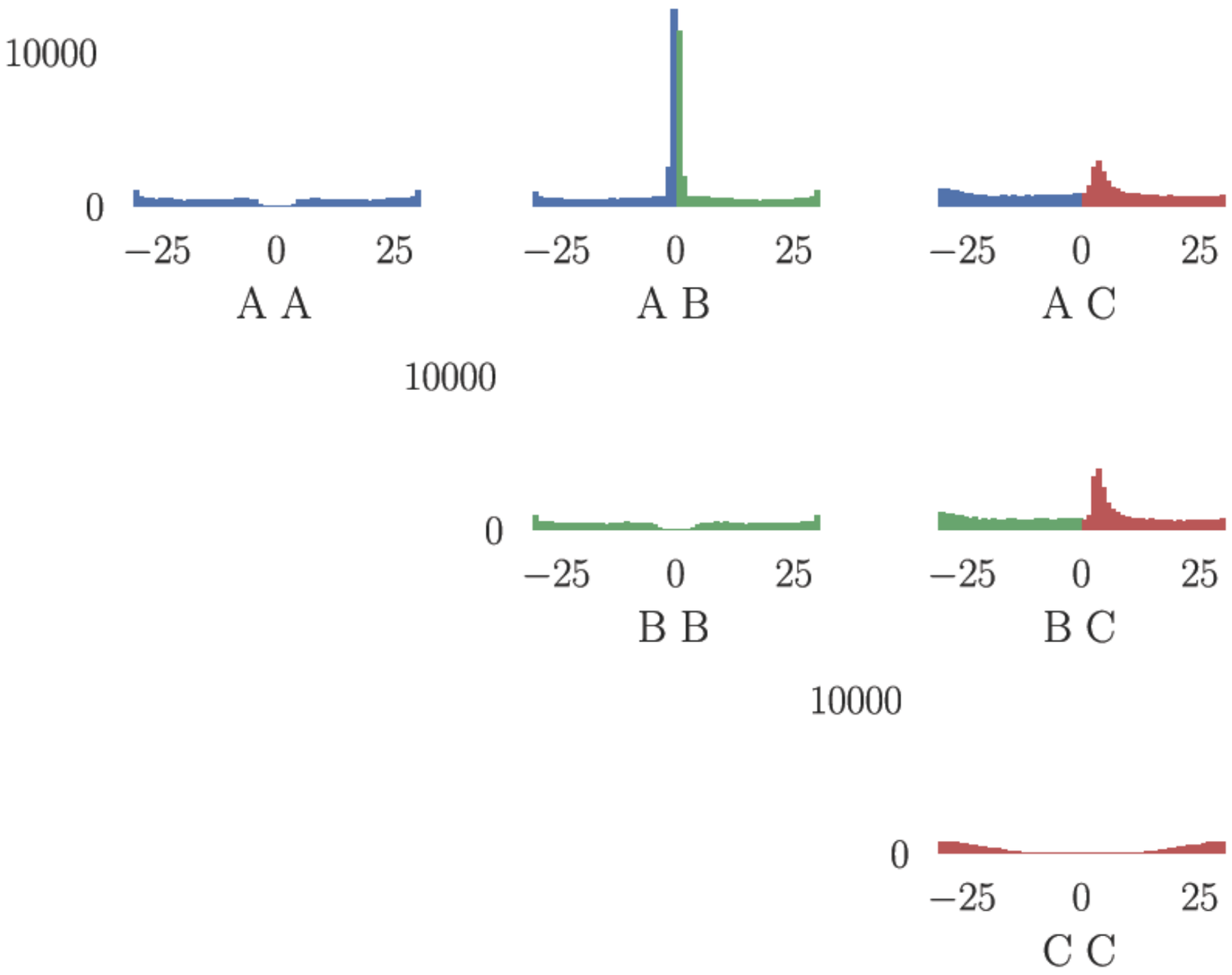
With Mikkel Lepperod

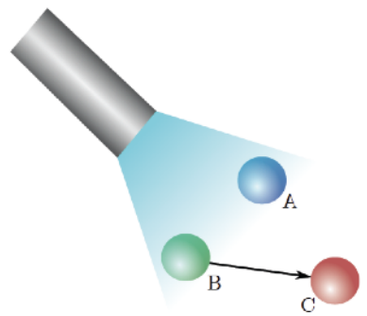
Optogenetics is not local

$I \approx 1/d^2$ $N \approx d^2$

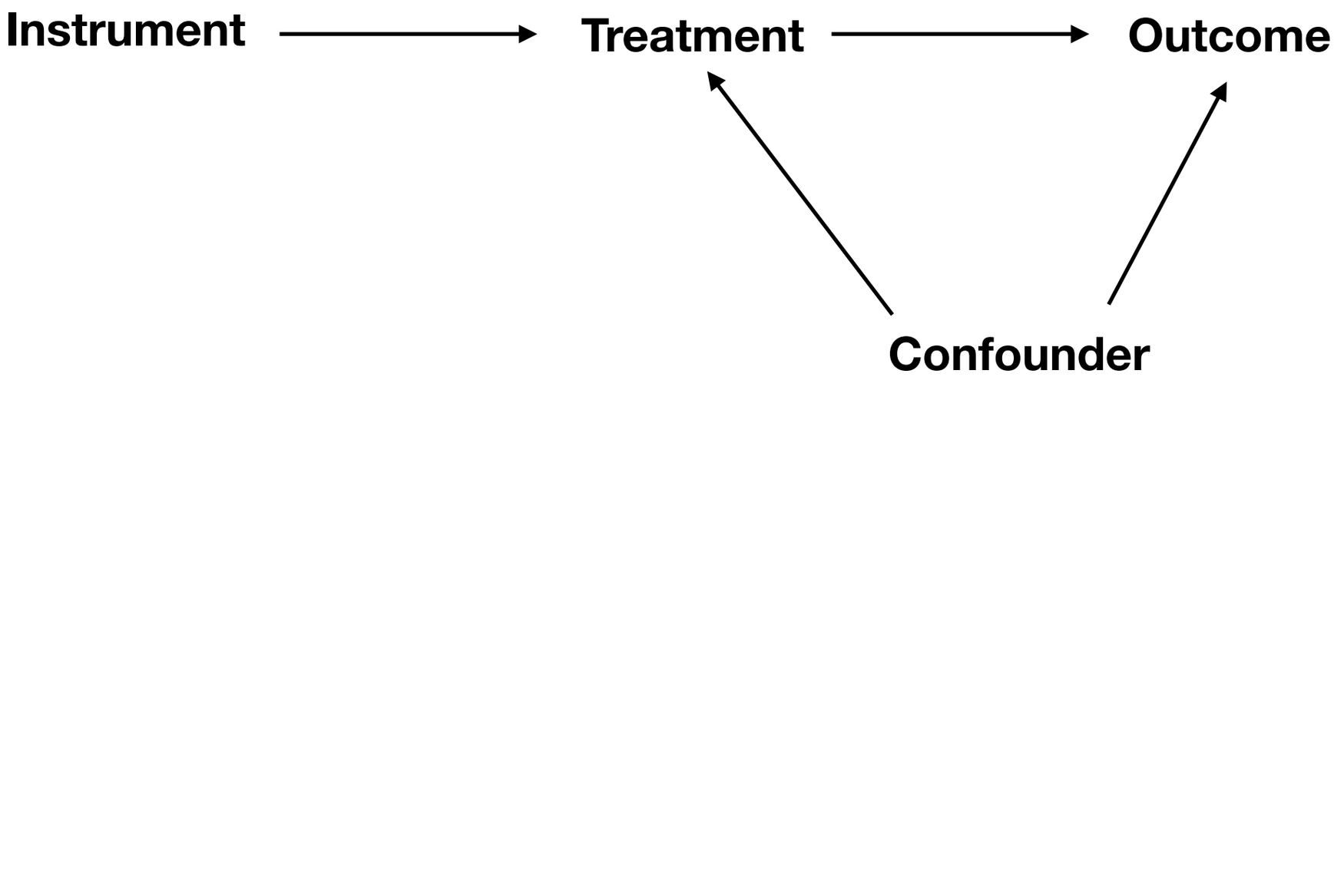


Massive confounding





Instrumental variables

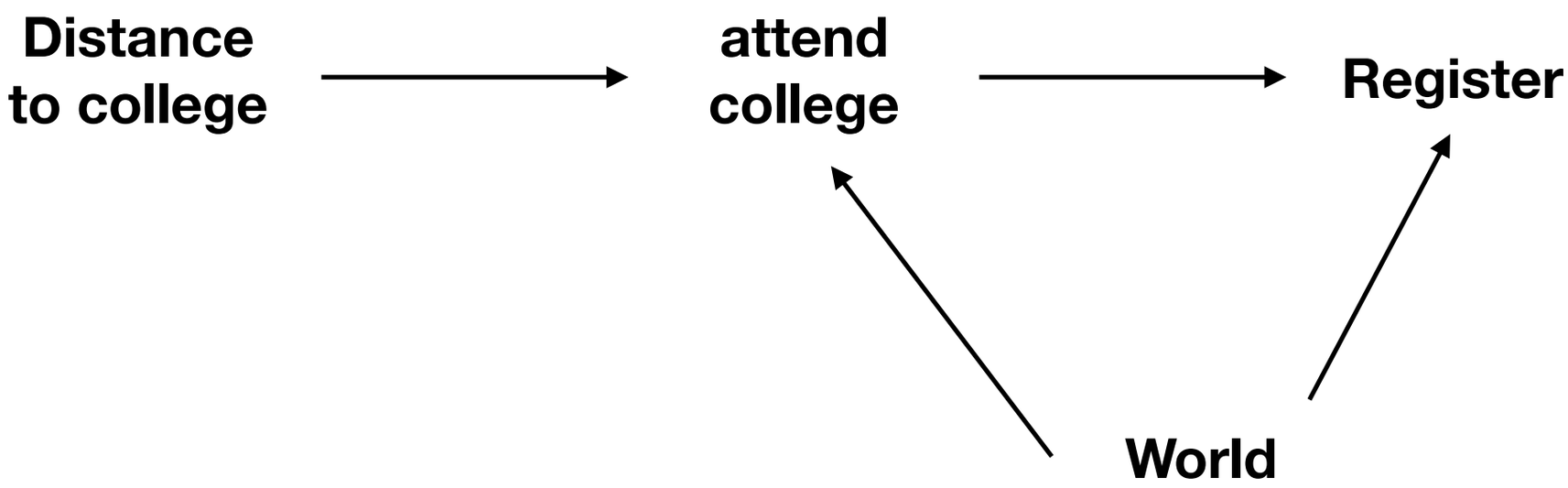


Civic engagement - college relation

- Distance to nearest college as instrument

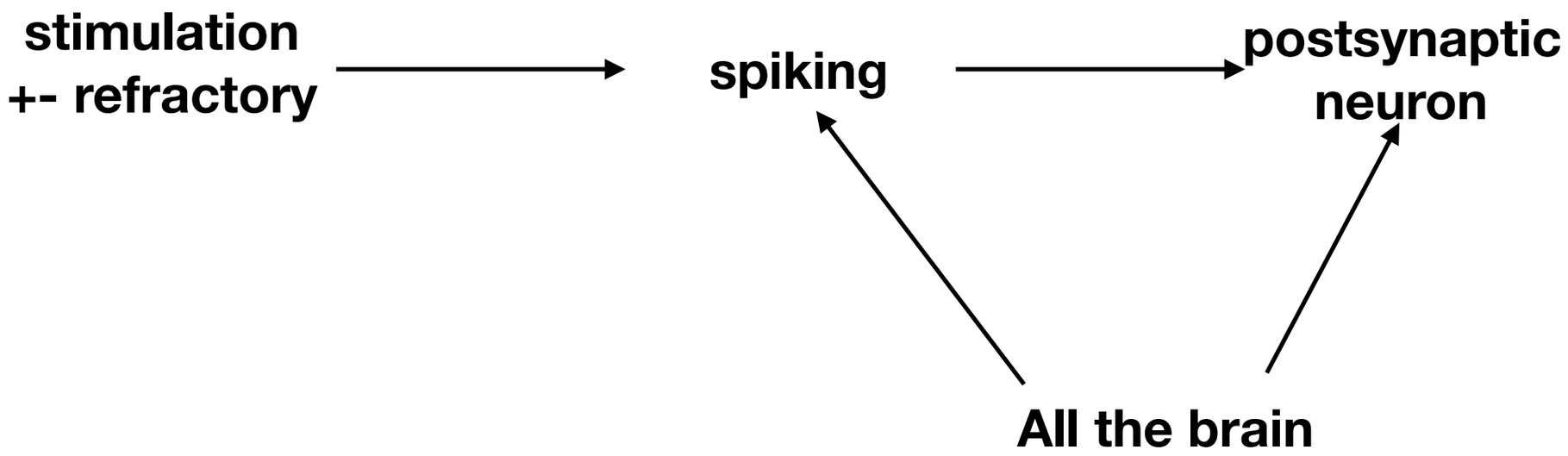
- Does it affect p(register to vote)?

Example

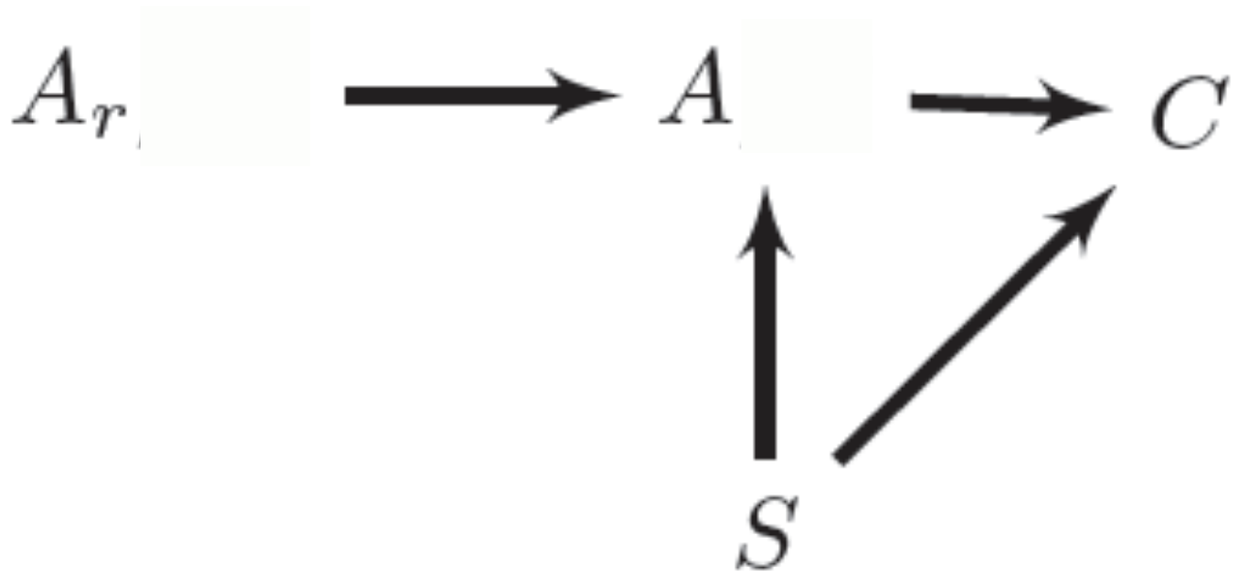


Groenwold et al

For us


stimulation
+- refractory
spiking
postsynaptic
neuron
All the brain

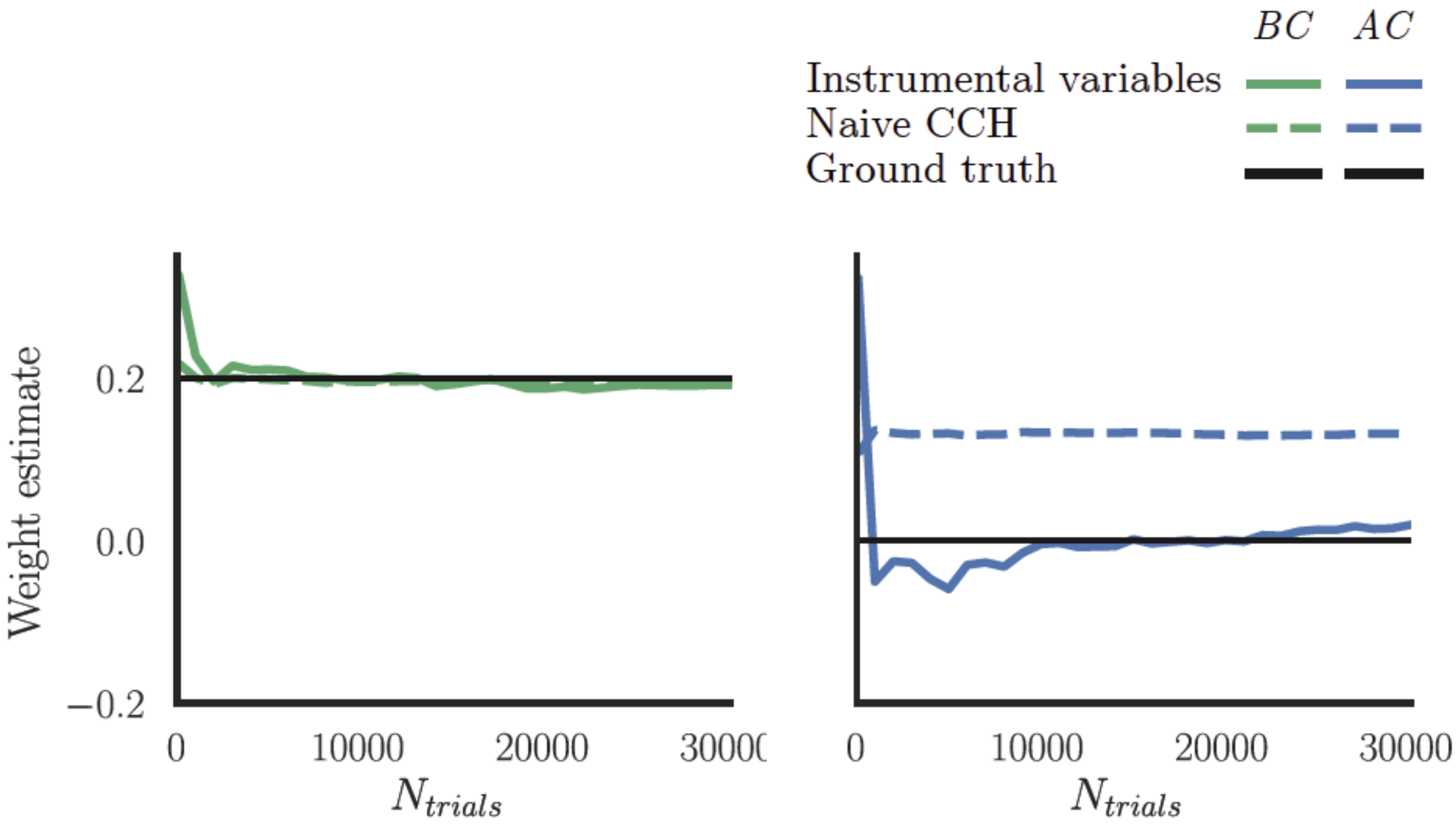
Instrumental variables



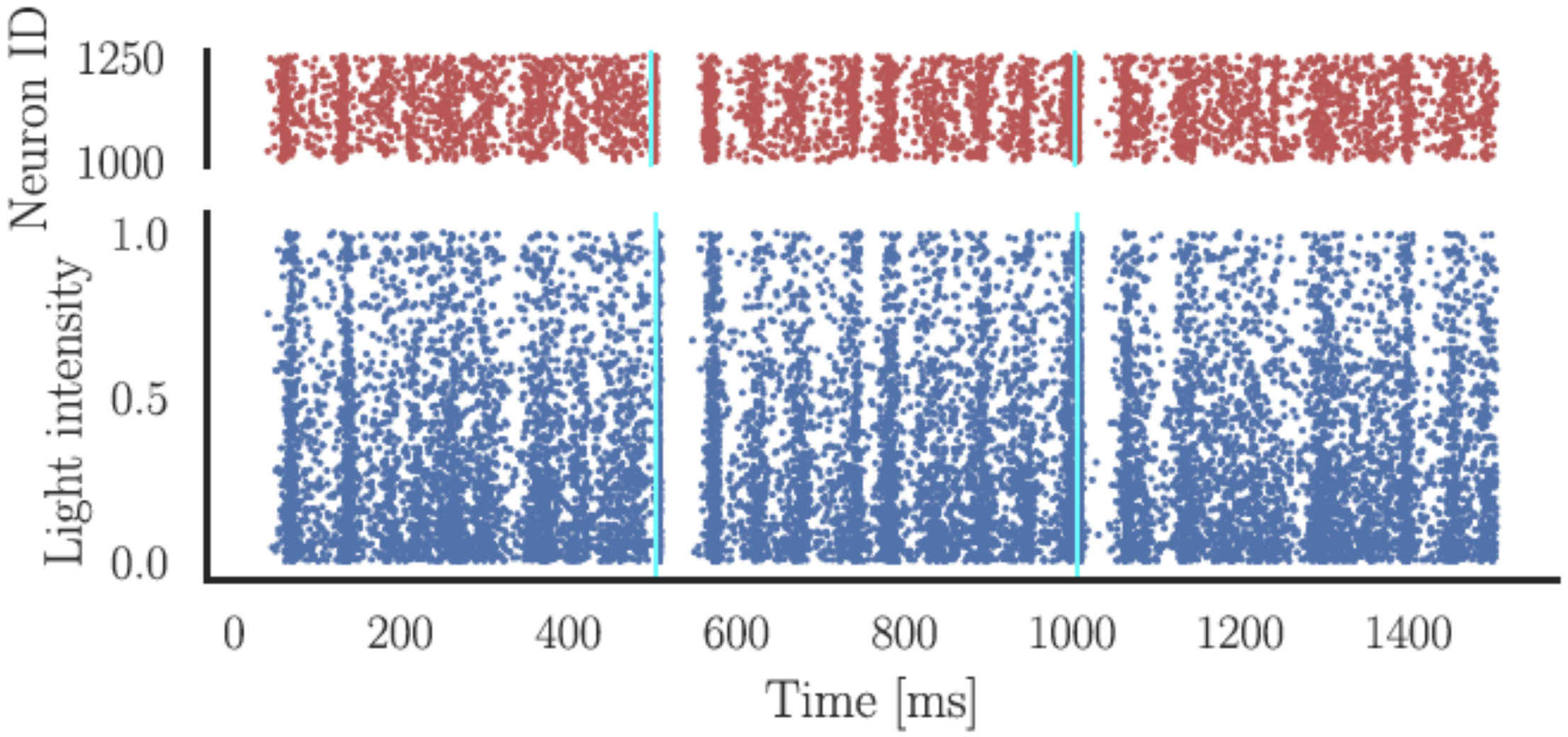
$$\theta^{IV} = \frac{E[C\,|\,A_r = 1] - E[C\,|\,A_r = 0]}{E[A\,|\,A_r = 1] - E[A\,|\,A_r = 0]}$$

Wald estimator (1940)

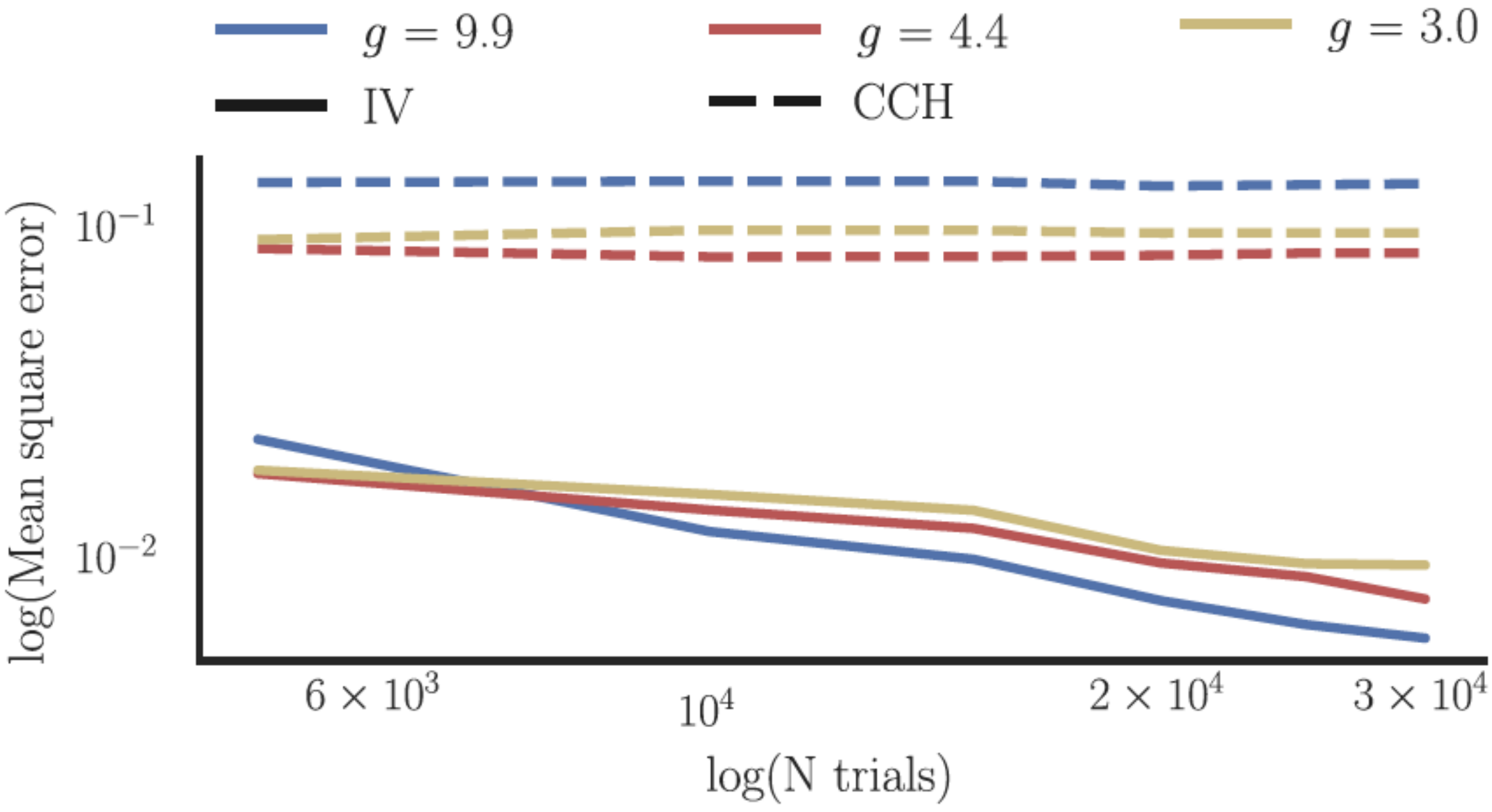
Instrumental variables



Many neurons



IV helps. A lot.



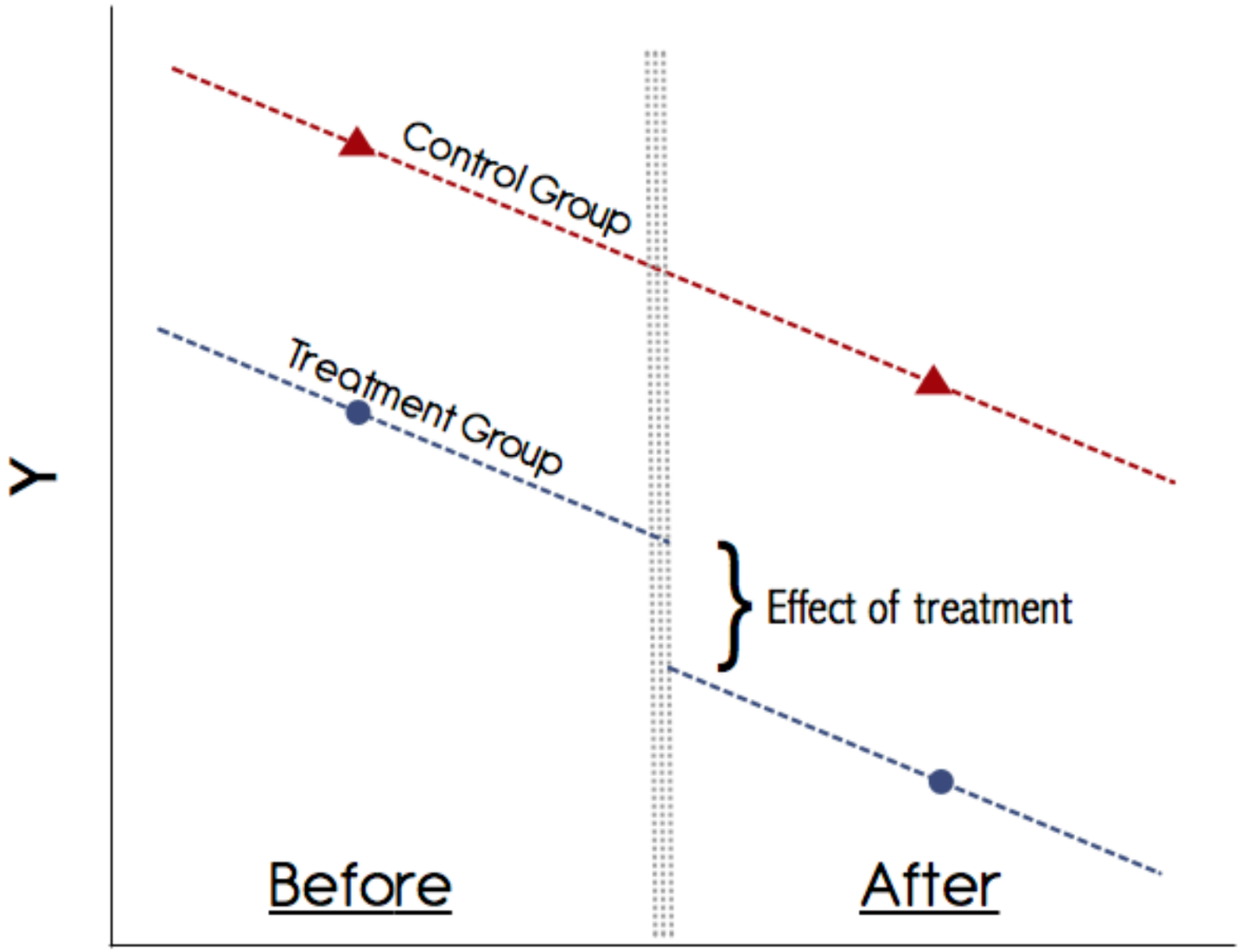
Why it matters

- Optogenetics is arguably the best causal tool we have
- But crazy hard (2p) to target individual cells
- Use causal inference tricks to cure confounding

An aside

- Medicine has
    - many thresholds
    - many random assignments (e.g. doctors)
- Confounding literally kills

One more pseudoexperiment: Diff in Diff



Caveats

The lure of causal statements: Rampant mis-inference of causality in estimated connectivity

Mehler & Kording

shoutout: Manjari Narayan (@neurostats)

Take home message

- We really mean causality when we talk about mechanism
- In many cases we provide no relevant information re causality
- Perturbations are gold standard. But do not scale
- Quasiexperiments are important set of approximation ideas

Acknowledgements

- **ML**
- Ari Benjamin
- Hugo Fernandes

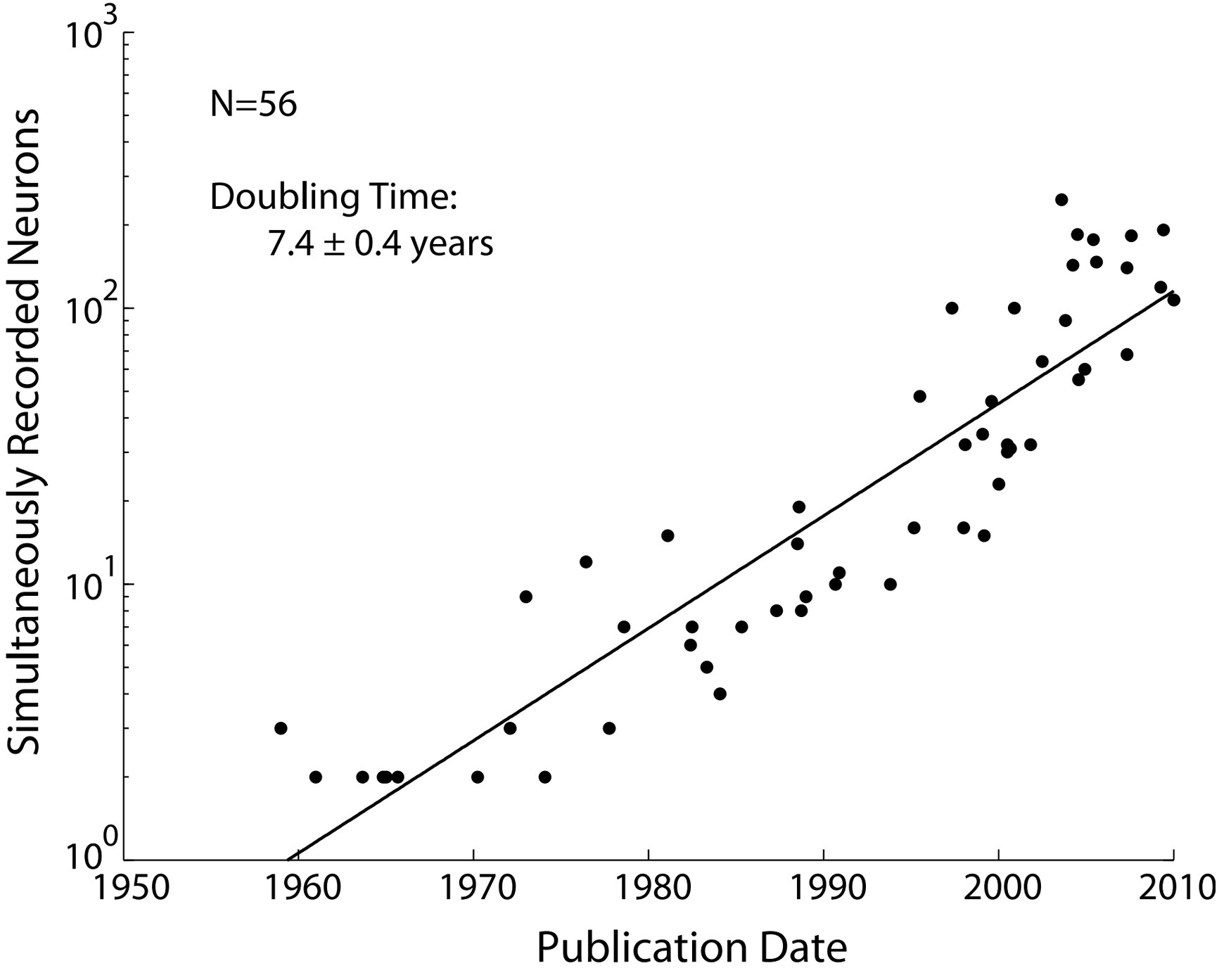
- **Video tracking**
- Claire Chambers
- Gaiqing Kong
- Julian Yarkoni
- Shaofei Wang

- **Bad ML**
- Luca Lonini
- Sohrob Saeb
- David Mohr
- Ben Recht
- Orianna Demasi

Causality
- Ioana Marinescu
- Pat Lawlor
- Mikkel Läppernd

- **Funding**
- NIH, NSF

Stevenson's Law



Getting data from brains

- Typing: 100 bits/s record, 20 bits/s me
- Eye movement: 20 bits/s
- EEG: .5 bits/s
- EMG Hand movement BMI: 2bits/s

- Dancing? 200 muscles*8bits/muscle*100/s
  =160k bits/s

Take home: Standard ML

- Work really well, fast
- Challenge people to get better results with brain intuitions
- Set baseline
- Ok, lets talk about non-standard now

Machine Learning in Data Driven Medicine: how to not do it wrong

@kordinglab

UPenn

Shameless plug: Please read *10 simple rules for structuring papers*
AFAIK: Most tweeted scientific paper, ever