



University of
Zurich^{UZH}

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



Model interpretability in healthcare

Inês Pereira

March 10th, 2021

Disclaimer





Context

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD;
Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB;
Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

Research

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm

for Detection of Melanoma
in Retina

Varun Gulshan, PhD
Subhashini Venugopalan, PhD
Rajiv Raman, MS, DMRC

LETTER

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva^{1*}, Brett Kuprel^{1*}, Roberto A. Novoa^{2,3}, Justin Ko², Susan M. Swetter^{2,4}, Helen M. Blau⁵ & Sebastian Thrun⁶

Research

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm

for Detection of Diabetic Retinopathy

LETTER

Varun Gulshan, PhD
Subhashini Venugopalan, PhD
Rajiv Raman, MS, DM

doi:10.1038/nature21056

Dermatology
with deep learning

Andre Esteva^{1*}, E

News > Medscape Medical News > FDA Approvals

FDA Approves AliveCor Personal ECG Monitor for Apple Watch

Steve Stiles

November 30, 2017

Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring

Michiel Kallenberg*, Kersten Petersen, Mads Nielsen, Andrew Y. Ng, Pengfei Diao, Christian Igel, Celine M. Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, and Martin Lillholm

1038/nature21056

ICG Monitor for

Andre Esteva^{1*}, I

Apple Watch

Steve Stiles

November 30, 2017

1322

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 35, NO. 5, MAY 2016

Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring

Michiel Kallenberg*, Kersten Petersen, Mads Nielsen, Andrew Y. Ng, Pengfei Diao, Christian Igel, Celine M. Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, and Martin Lillholm

1038/nature21056

ECG Monitor for

Andre Esteva^{1*}, I

Apple Watch

IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 35, NO. 5, MAY 2016

1207

Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network

Marios Anthimopoulos, *Member, IEEE*, Stergios Christodoulidis, *Member, IEEE*, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou*, *Member, IEEE*

Unsupervised Density

Michiel Kallenberg*,
Celine M. Vachon, K

OPEN

Deep learning segmentation of major vessels in X-ray coronary angiography

Su Yang¹, Jihoon Kweon^{1,2,5*}, Jae-Hyung Roh³, Jae-Hwan Lee³, Heejun Kang¹,
Lae-Jeong Park⁴, Dong Jun Kim¹, Hyeyoung Yang¹, Jaehyeon Hur¹, Do-Yoon Kang¹,
Pil Hyung Lee¹, Jung-Min Ahn¹, Soo-Jin Kang¹, Duk-Woo Park¹, Seung-Whan Lee¹,
Young-Hak Kim^{1,5*}, Cheol Whan Lee¹, Seong-Wook Park¹ & Seung-Jung Park¹

CONVENTIONAL NEURAL NETWORK

Marios Anthimopoulos, *Member, IEEE*, Stergios Christodoulidis, *Member, IEEE*,
Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou*, *Member, IEEE*

SCIENTIFIC
REPORTS

natureresearch

for

1207

1



Geoffrey Hinton
@geoffreyhinton

...

Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

9:37 PM · Feb 20, 2020 · Twitter Web App

1,159 Retweets 614 Quote Tweets 5,172 Likes

Is it a dichotomy?

How do you define success?



Tumor segmentation

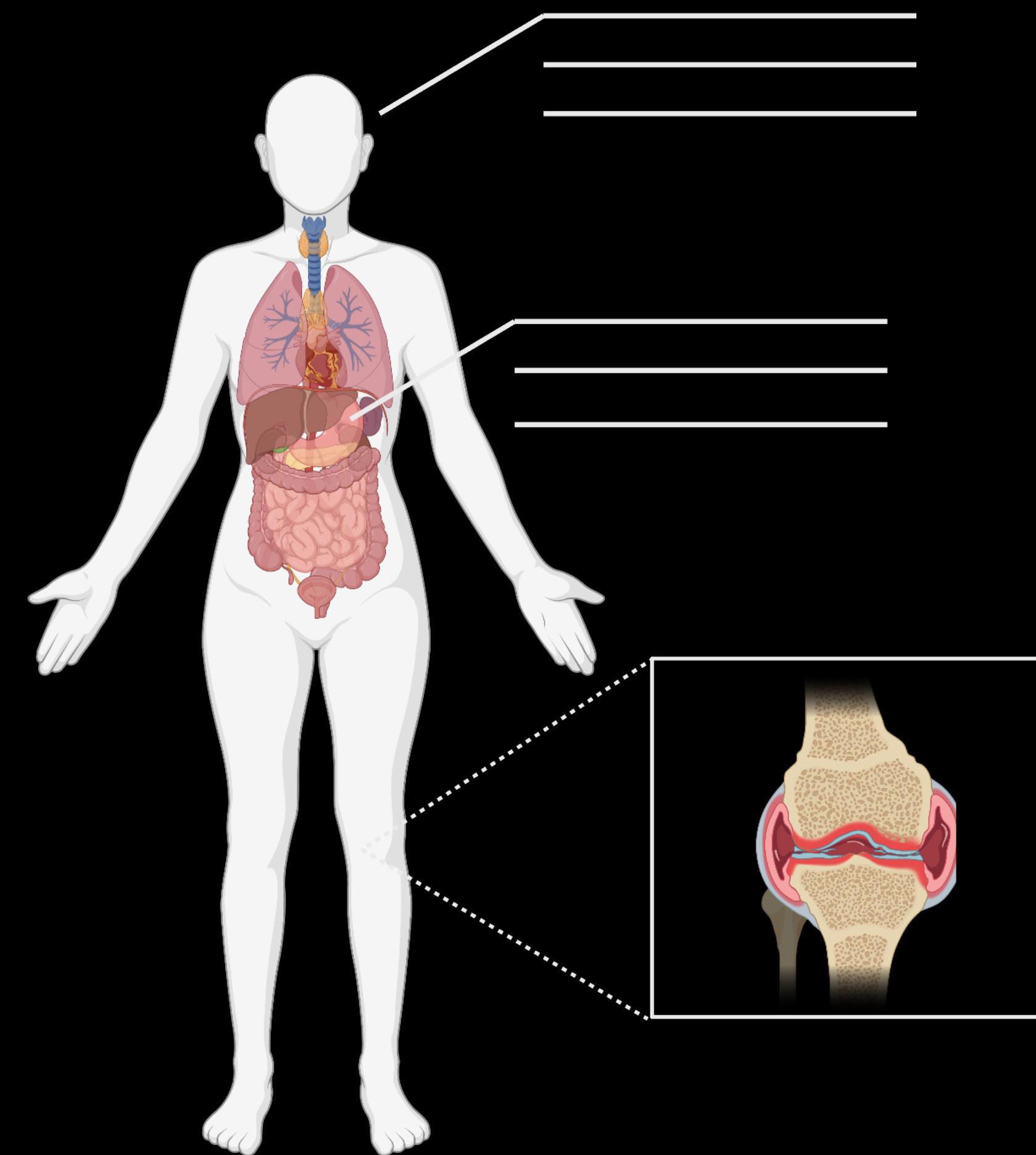


Tumor segmentation

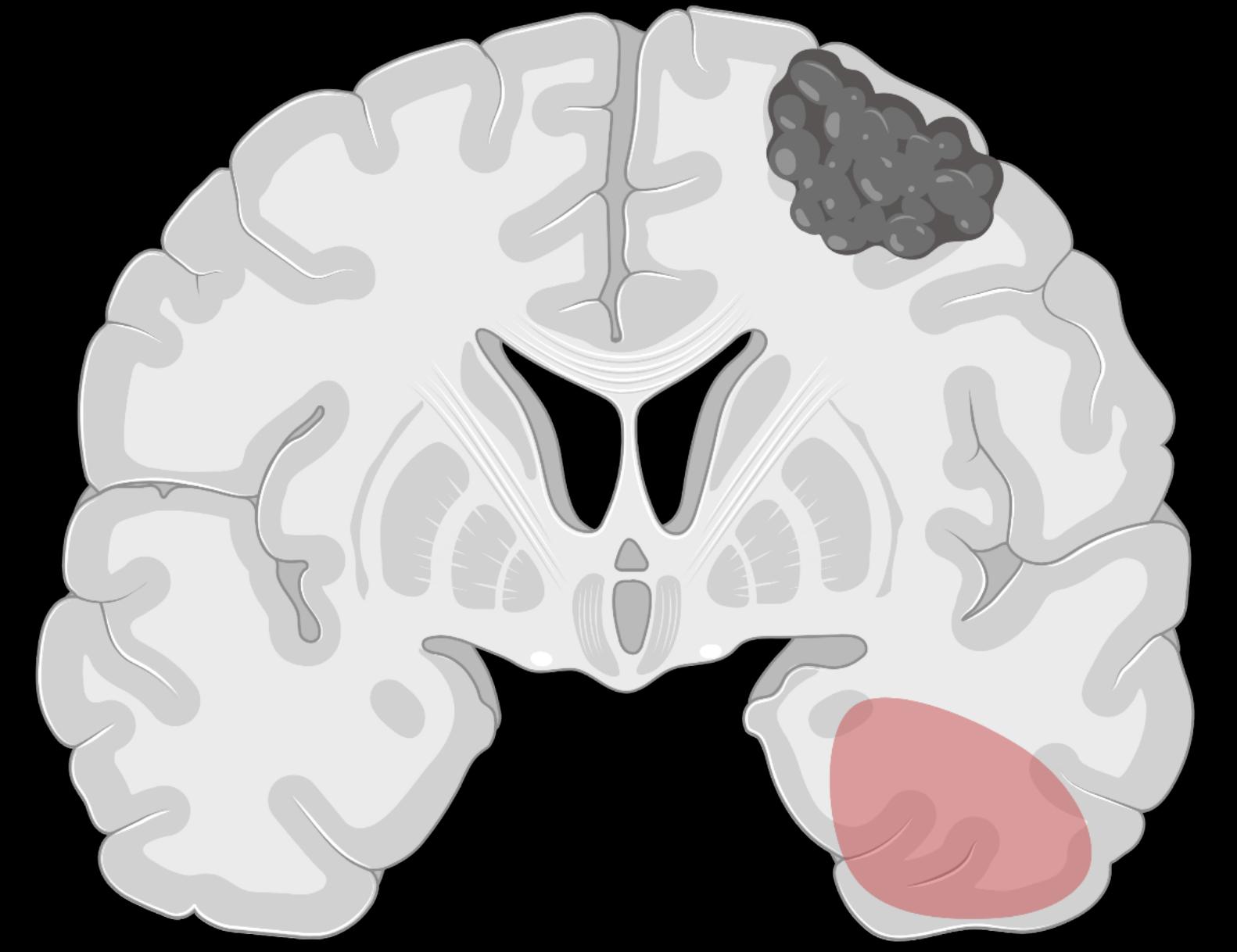
Tumor segmentation



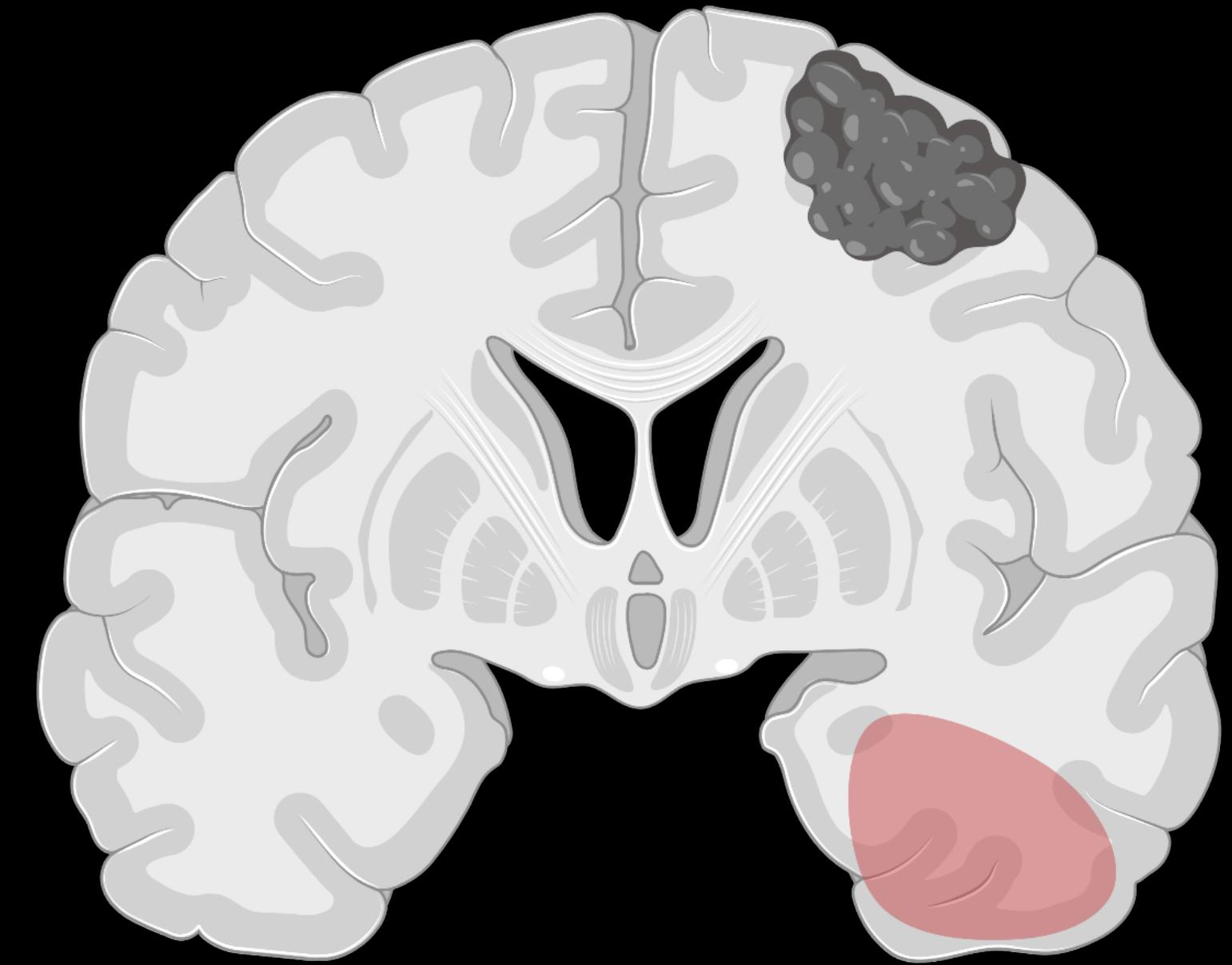
Diagnosis or prognosis



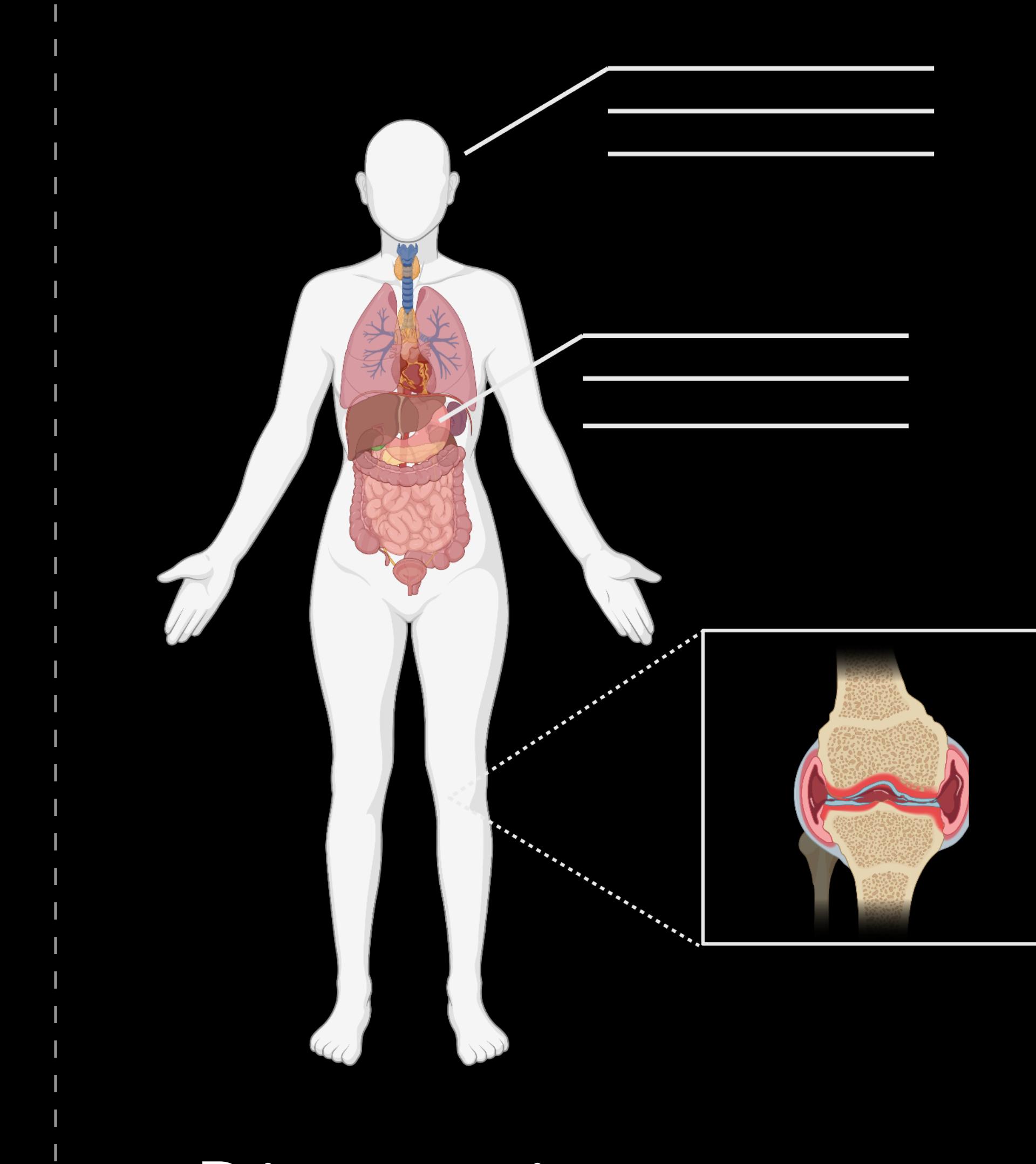
*Does the model make easily verifiable
predictions?*



Tumor segmentation



Tumor segmentation

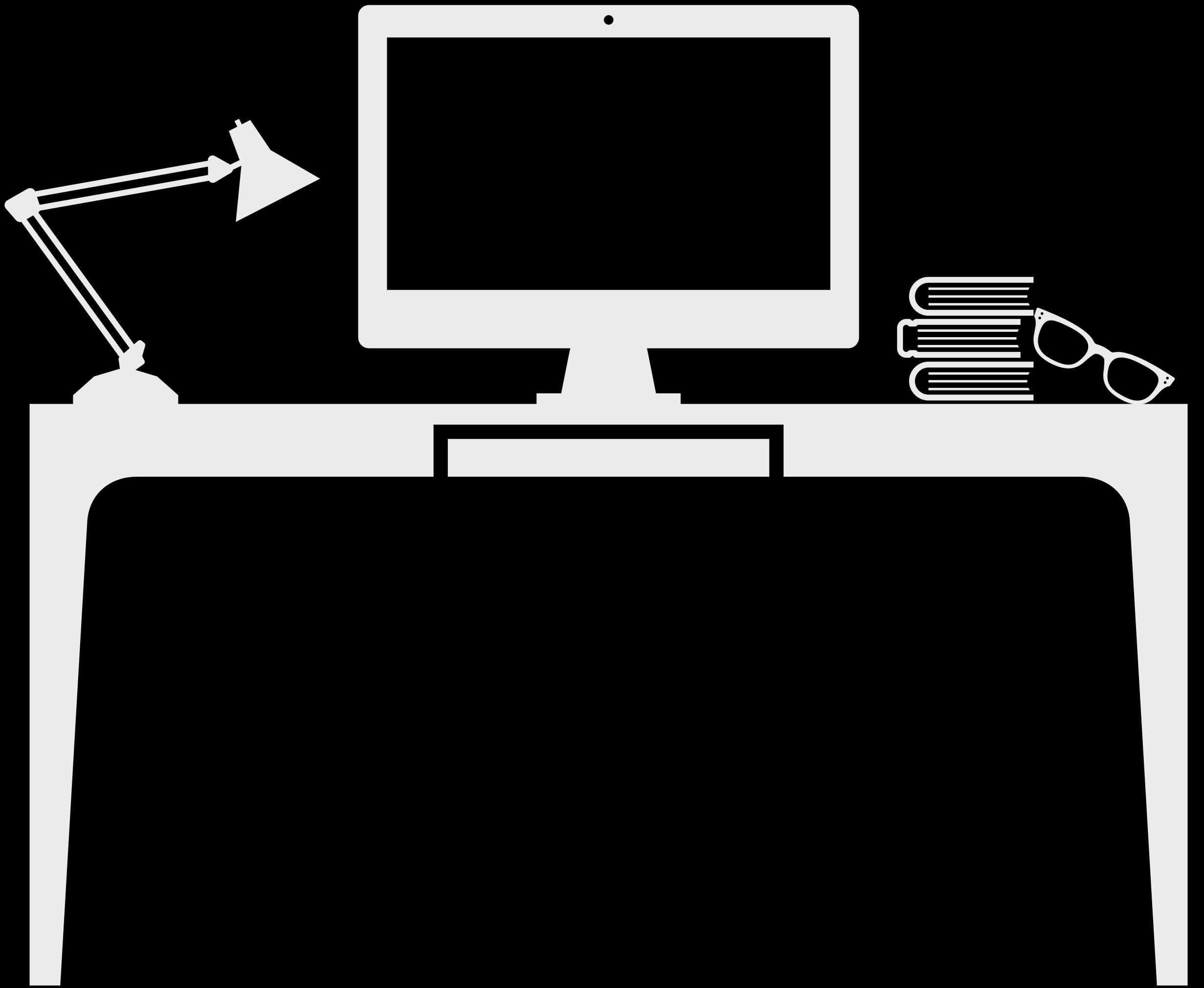


Diagnosis or prognosis

*Why are we striving for model
interpretability?*



Tumor segmentation



“Definitions”

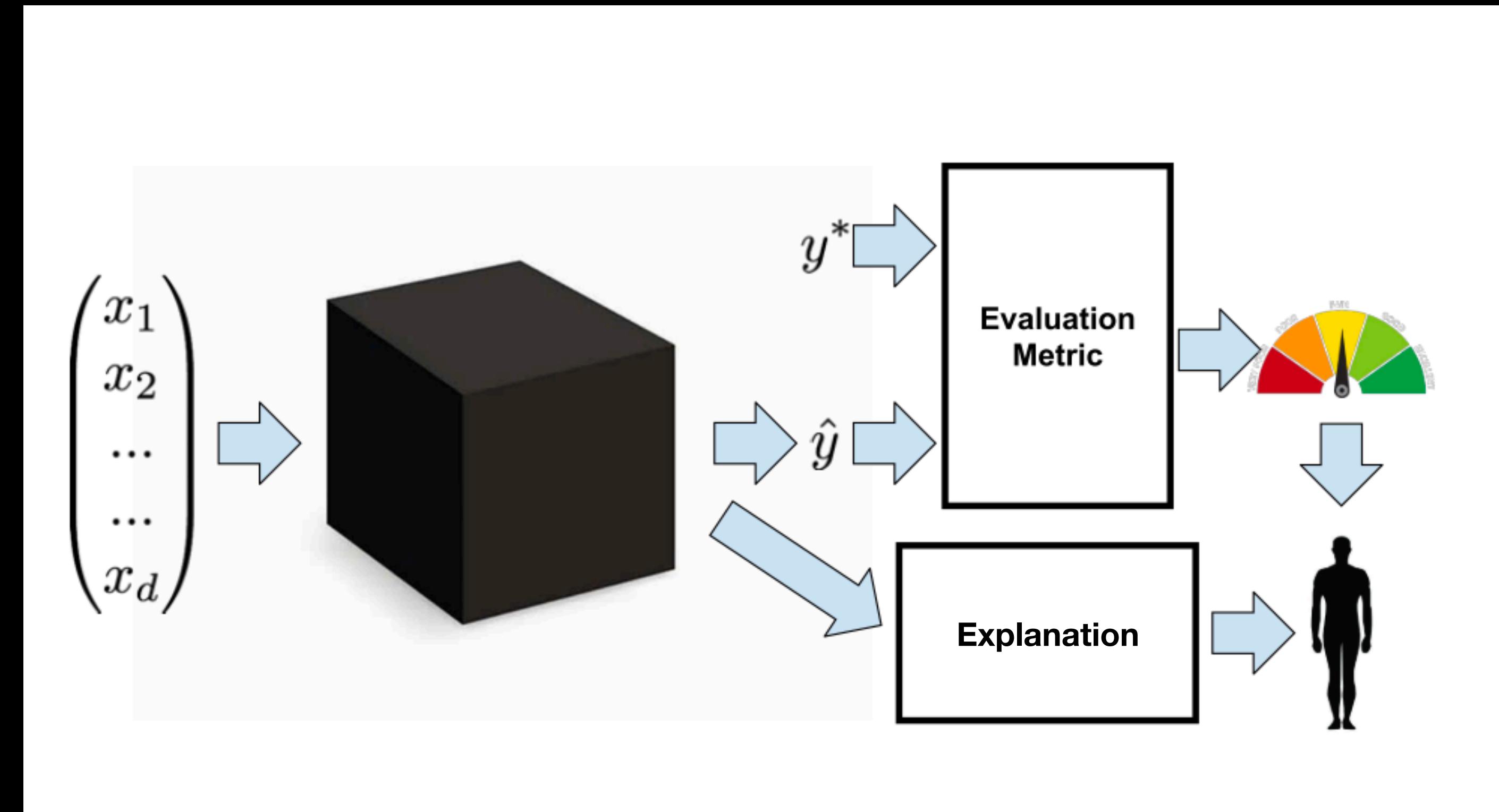
Interpretability is the degree to which a [human] observer can understand the cause of a decision.

Miller, Artificial Intelligence, 2019

Interpretability

\neq

Explainability



Adapted from figure from Lipton, *arXiv*, 2017

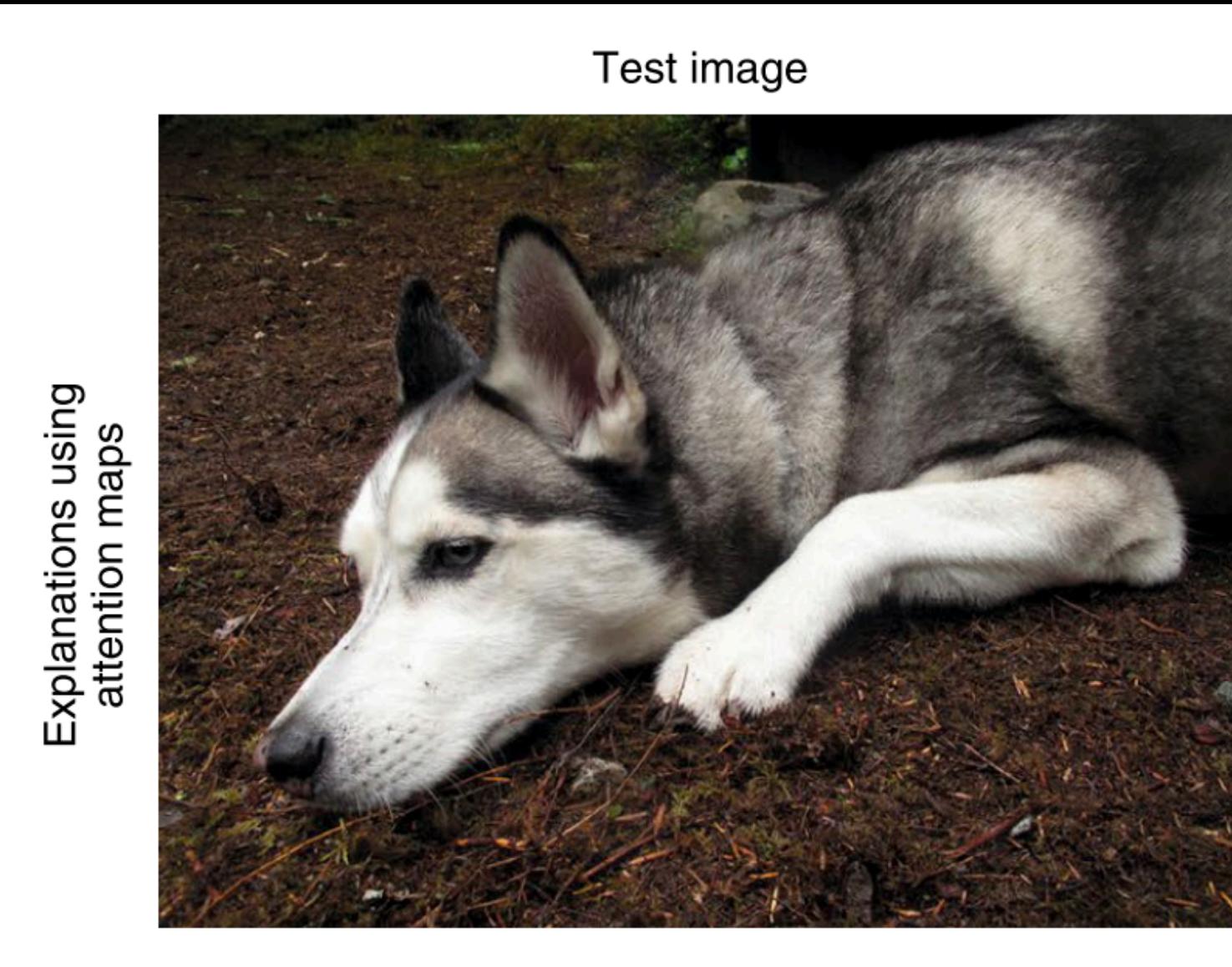


Figure from Rudin, *Nature Machine Intelligence*, 2019



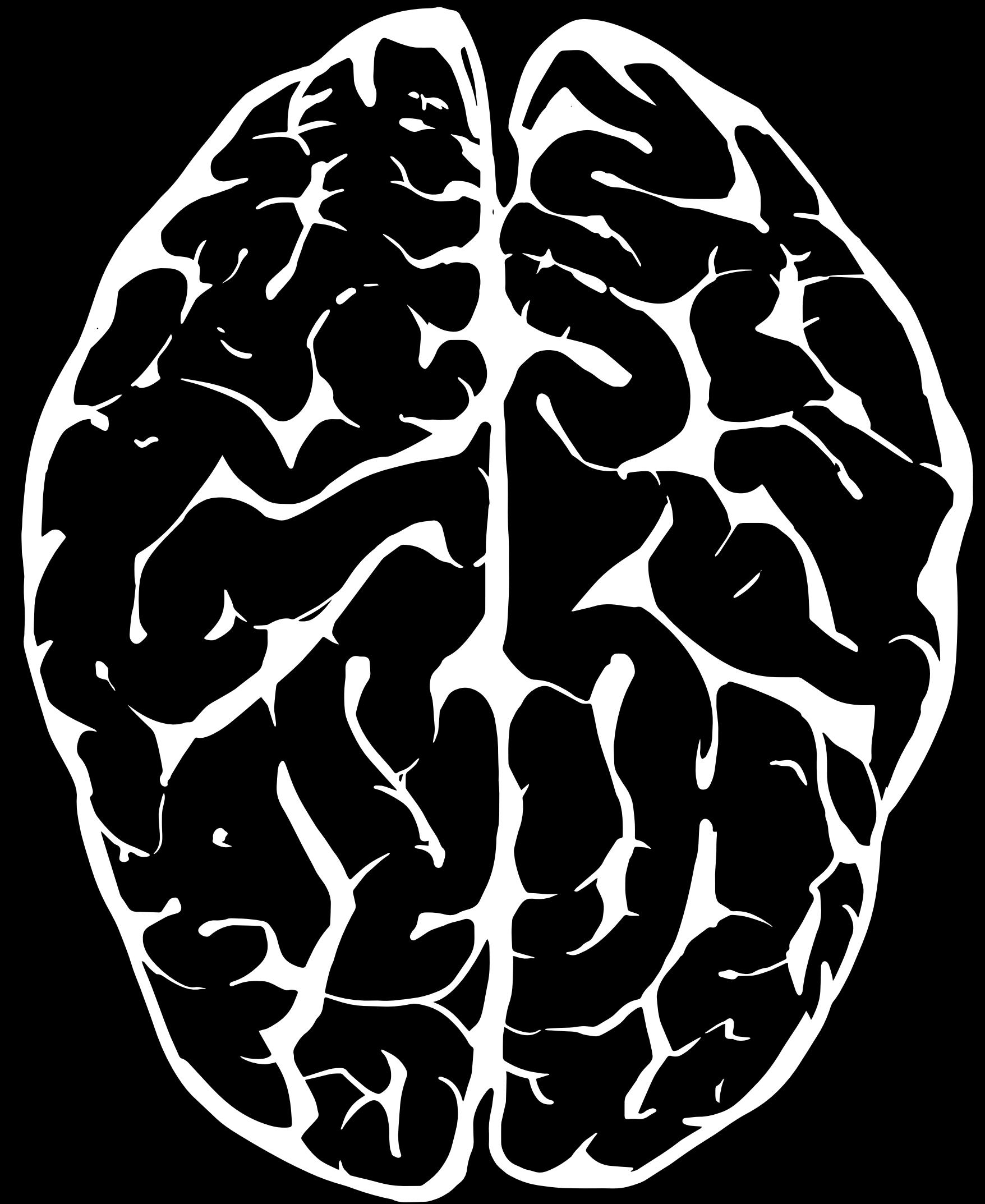
Figure from Rudin, *Nature Machine Intelligence*, 2019

The need for model interpretability

- Healthcare involves high-stakes situations
- Interpretability can help further understanding
- Healthcare is more than following algorithms

The need for model interpretability

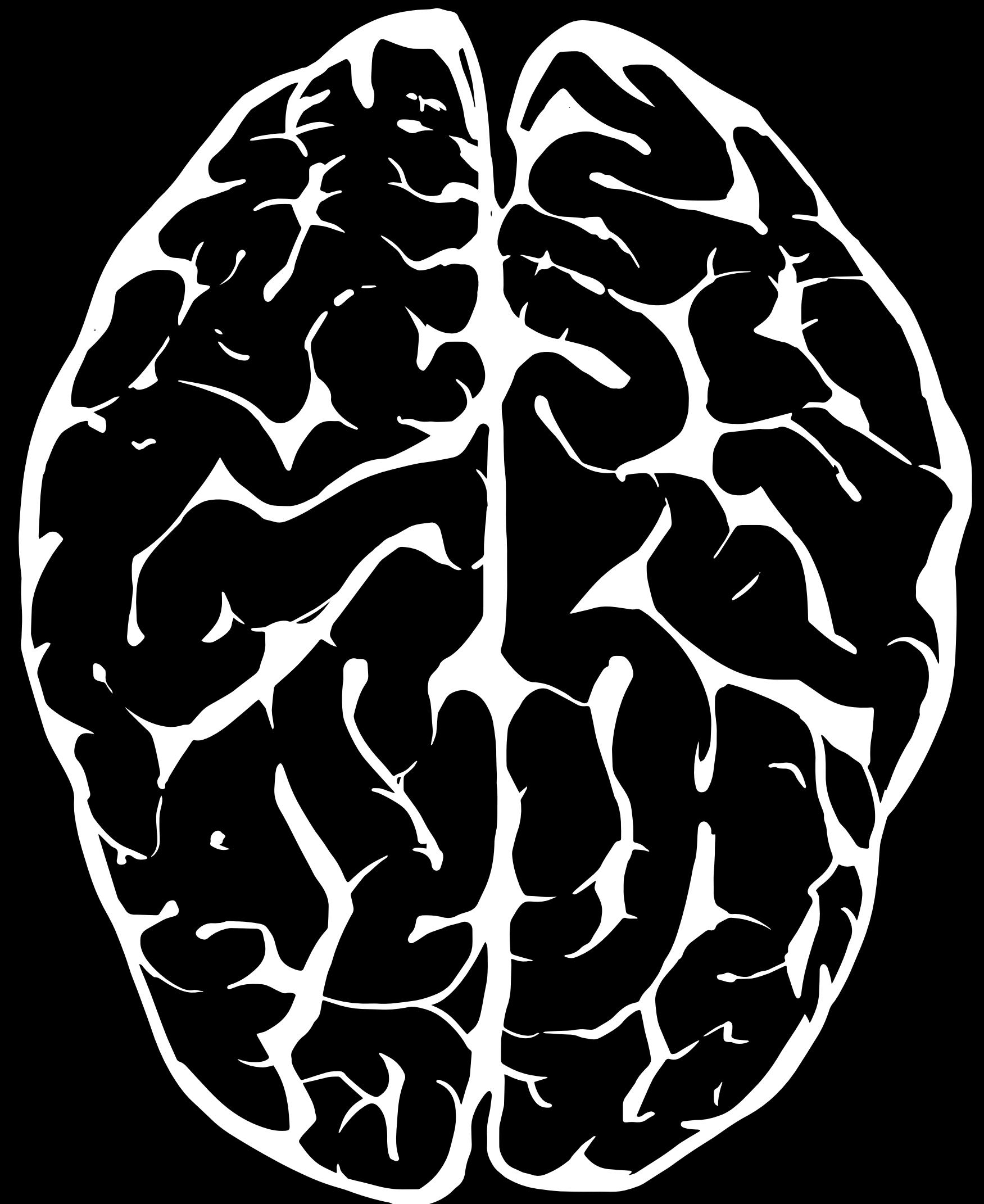
- Healthcare involves high-stakes situations
- **Interpretability can help further understanding**
- Healthcare is more than following algorithms



Bipolar disorder

Depression

**Autism spectrum
disorder**



Schizophrenia

Psychiatry

- Psychiatric disorders are broad and heterogeneous syndromes.

Messing *et al.*, "Biology of Psychiatric Disorders" in
Harrison's Principles of Internal Medicine, 20th, 2018

Psychiatry

- Psychiatric disorders are broad and heterogeneous syndromes.

Syndrome = cluster of symptoms and signs that tend to occur together

Messing *et al.*, "Biology of Psychiatric Disorders" in
Harrison's Principles of Internal Medicine, 20th, 2018

Major depressive episode (DSM-5)

- 5 (or more) of the following symptoms during the same 2-week period; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.
 1. Depressed mood (Note: In children and adolescents, can be irritable mood)
 2. Markedly diminished interest or pleasure in all, or almost all, activities
 3. Significant weight loss or weight gain, or decrease or increase in appetite nearly every day
 4. Insomnia or hypersomnia
 5. Psychomotor agitation or retardation
 6. Fatigue or loss of energy nearly every day.
 7. Feelings of worthlessness or excessive or inappropriate guilt
 8. Diminished ability to think or concentrate, or indecisiveness
 9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

What accounts for the heterogeneity?

Psychiatry

- Psychiatric disorders are broad and heterogeneous syndromes.
- Lack of well-defined neuropathology and *bona fide* biological markers

Messing *et al.*, "Biology of Psychiatric Disorders" in
Harrison's Principles of Internal Medicine, 20th, 2018

Psychiatry

- Psychiatric disorders are broad and heterogeneous syndromes.
- Lack of well-defined neuropathology and *bona fide* biological markers
- Diagnosis: clinical observations using diagnostic manuals (ICD-10, DSM-5)

Messing *et al.*, "Biology of Psychiatric Disorders" in
Harrison's Principles of Internal Medicine, 20th, 2018

Computational Psychiatry

Computational Psychiatry ?

Computational Neuroscience



Translational Neuromodeling



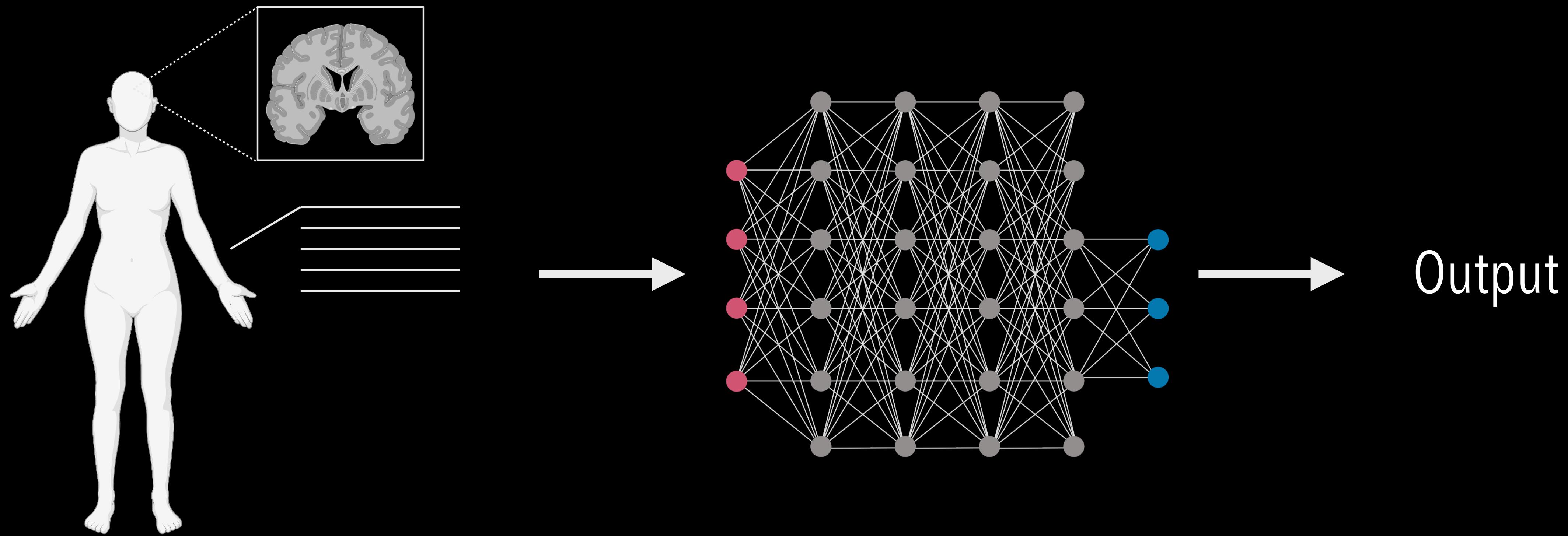
Computational Psychiatry



Computational Neurology



Computational
Psychosomatics



Generative models

Generative models

... based on
biophysical principles

Machine Learning recap

- Discriminative model: $P(Y|X = x)$
- Generative model: $P(Y, X)$
 - Definition of prior and likelihood function
 - Ability to synthesise data points

Dynamic causal modeling



ACADEMIC
PRESS

Available online at www.sciencedirect.com



NeuroImage 19 (2003) 1273–1302

NeuroImage

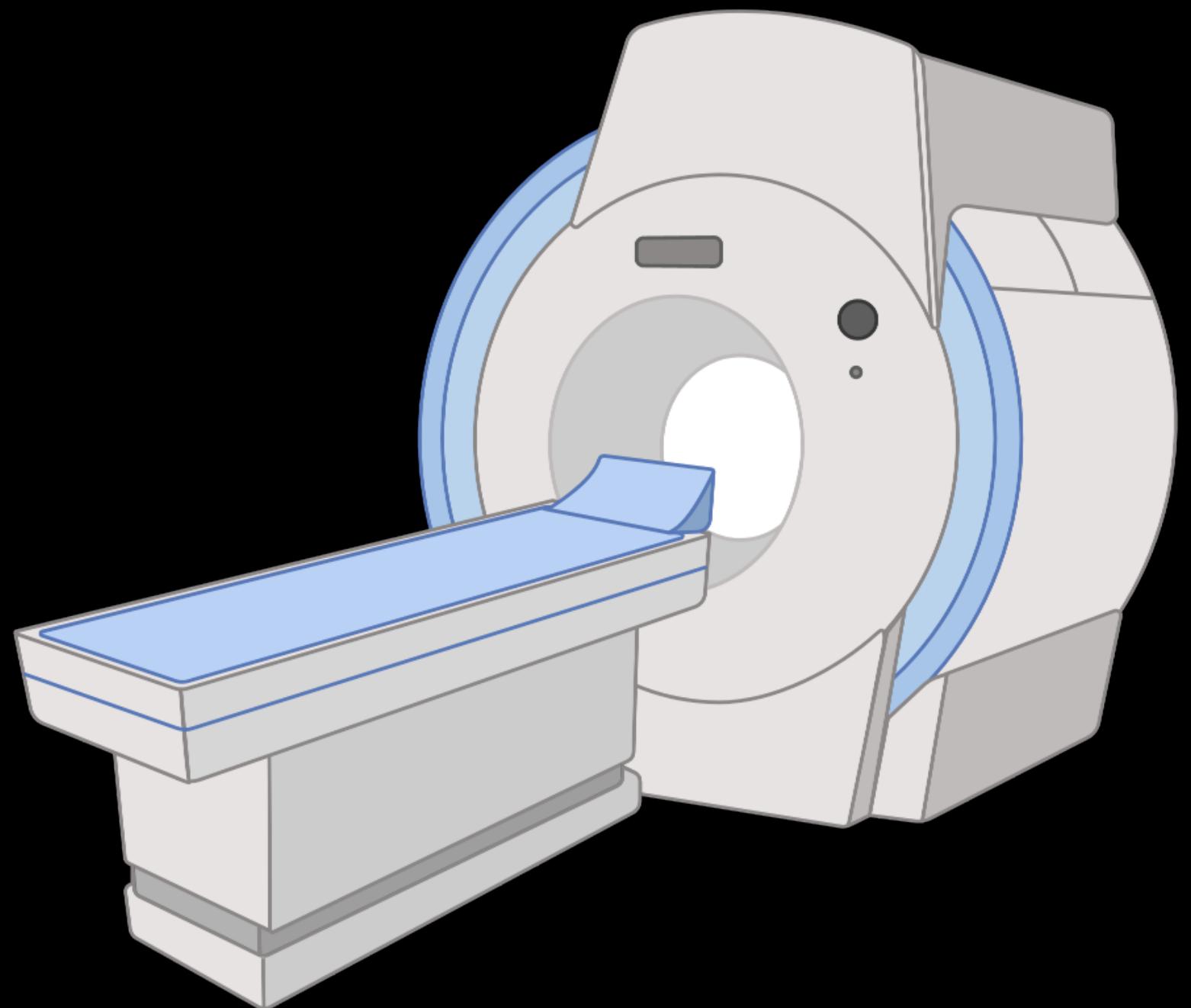
www.elsevier.com/locate/ynimng

Dynamic causal modelling

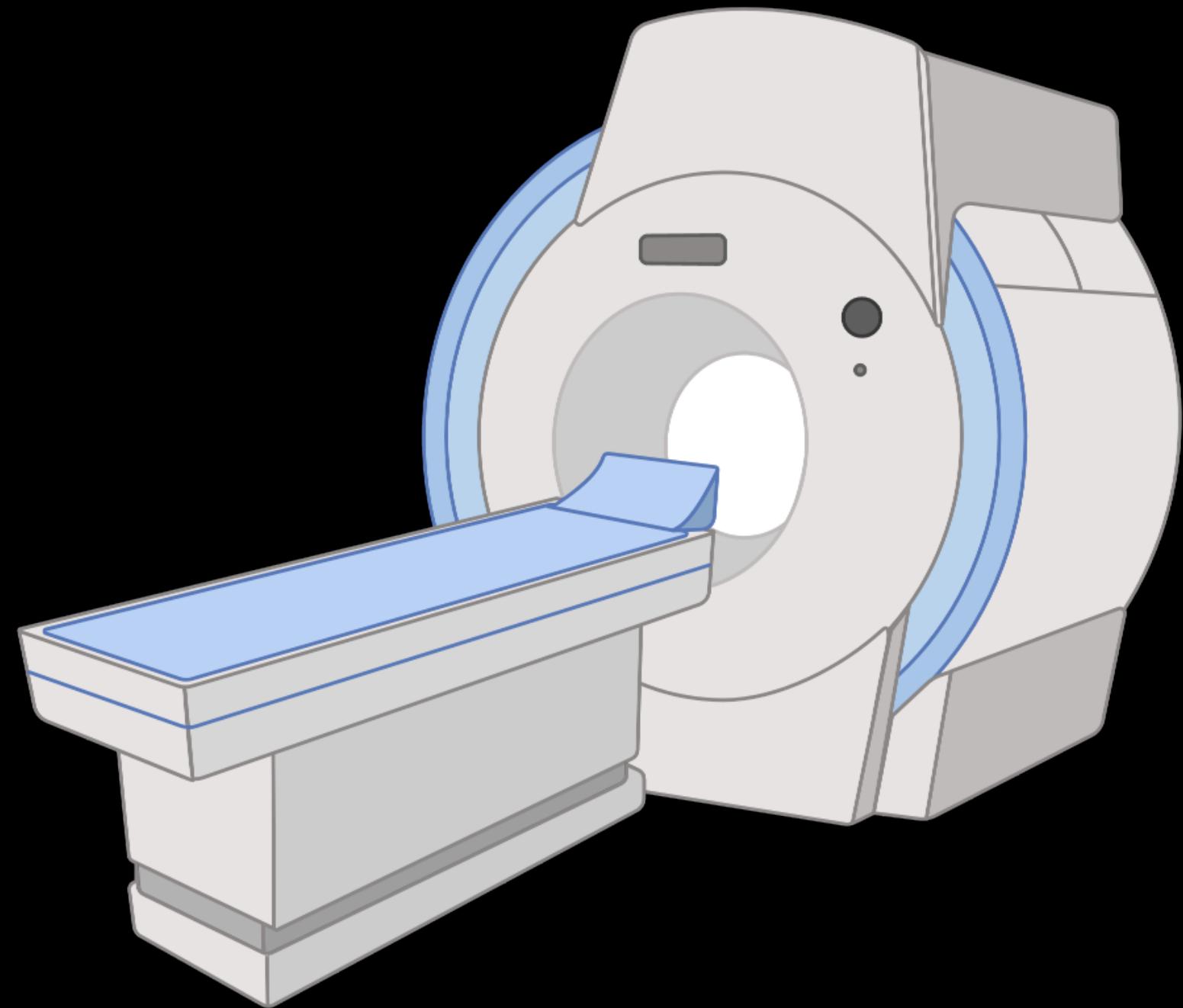
K.J. Friston,* L. Harrison, and W. Penny

The Wellcome Department of Imaging Neuroscience, Institute of Neurology, Queen Square, London WC1N 3BG, UK

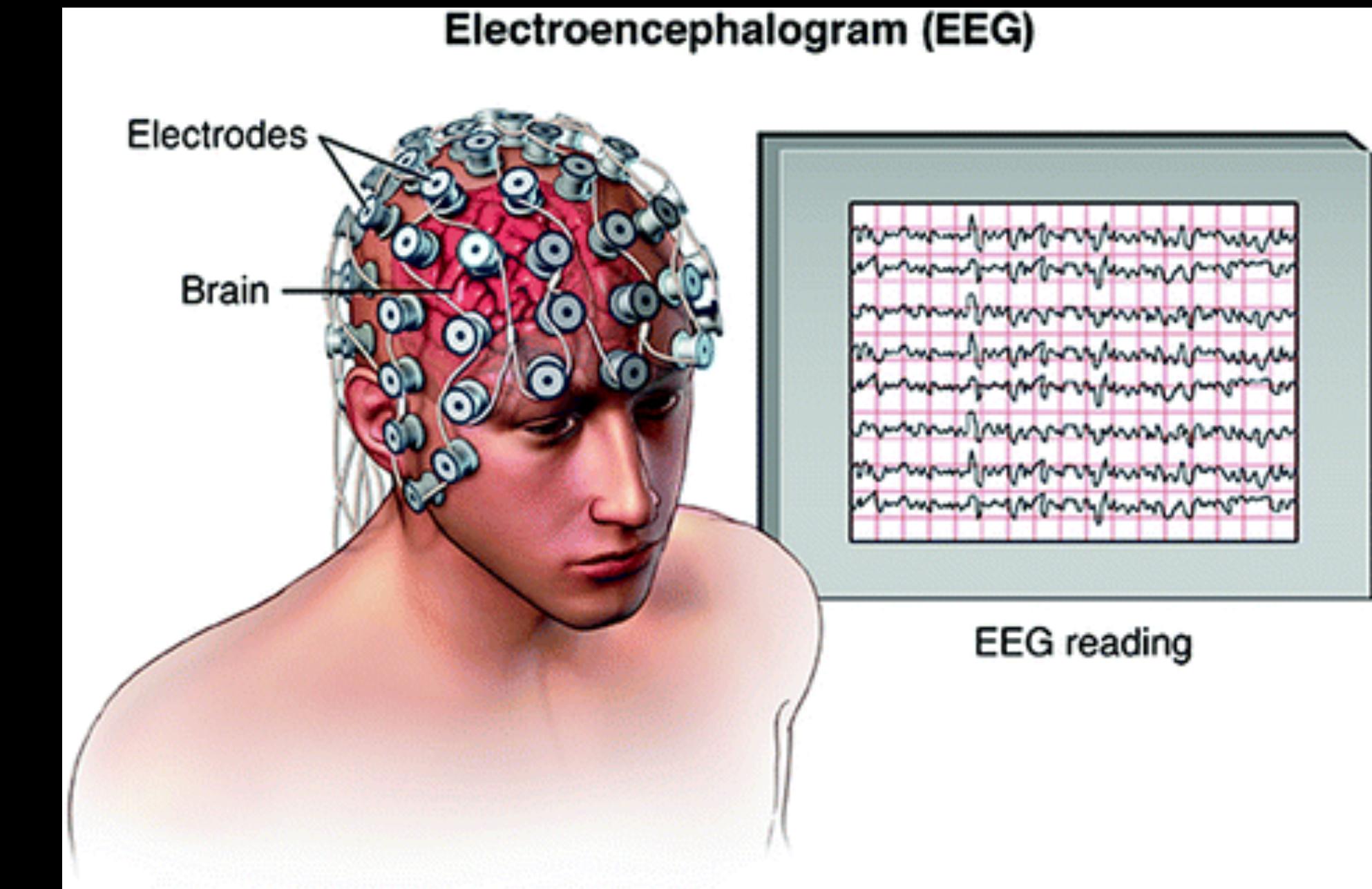
Received 18 October 2002; revised 7 March 2003; accepted 2 April 2003



Functional MRI (fMRI)



Functional MRI (fMRI)



Electroencephalography (EEG)

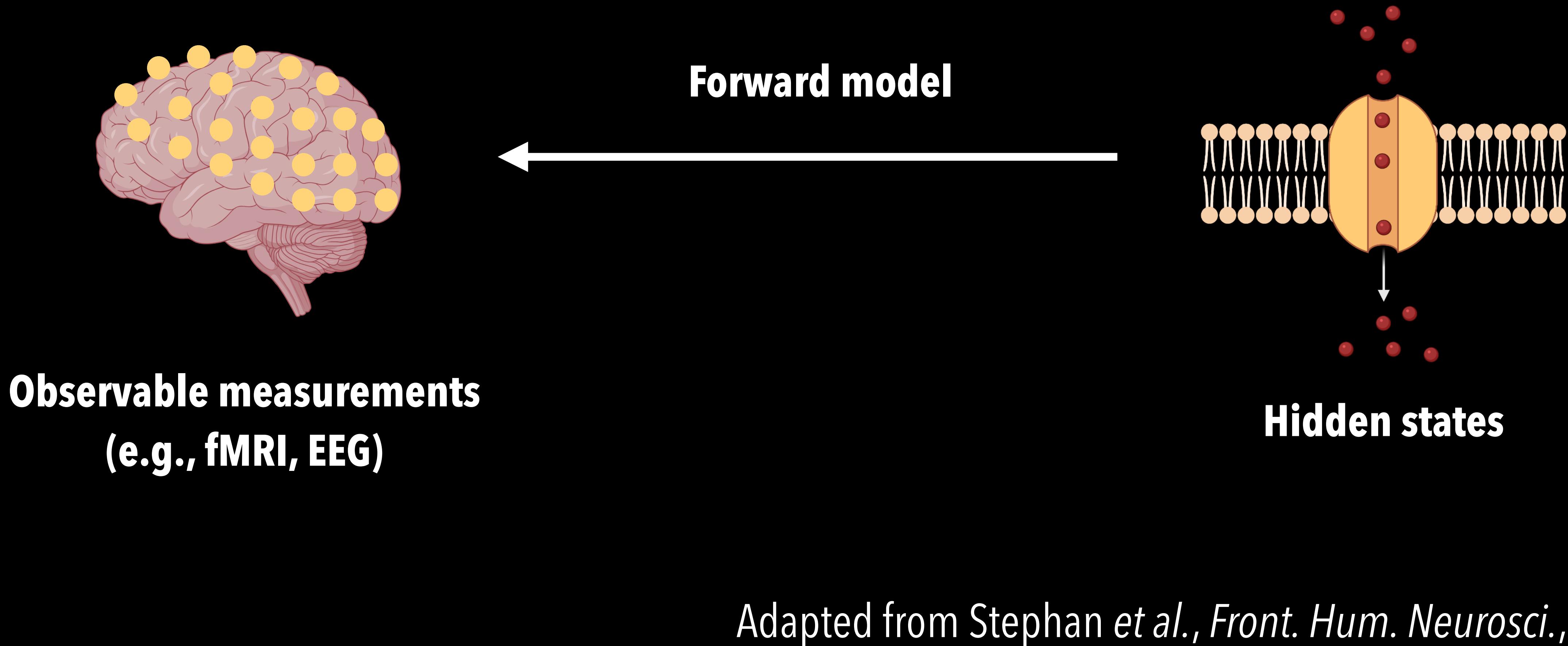
EEG image from: https://link.springer.com/chapter/10.1007/978-3-319-47653-7_1

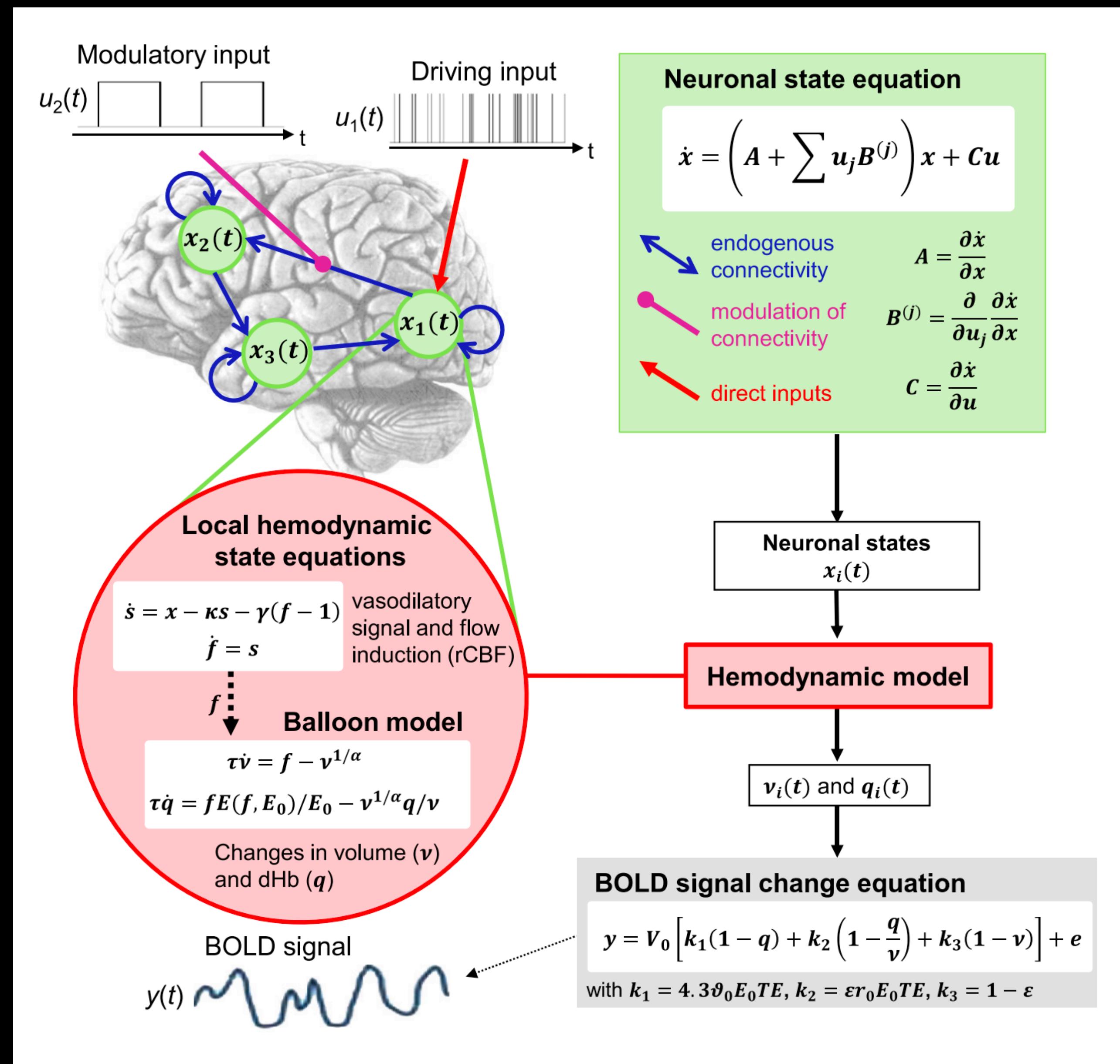
Generative model (DCM)



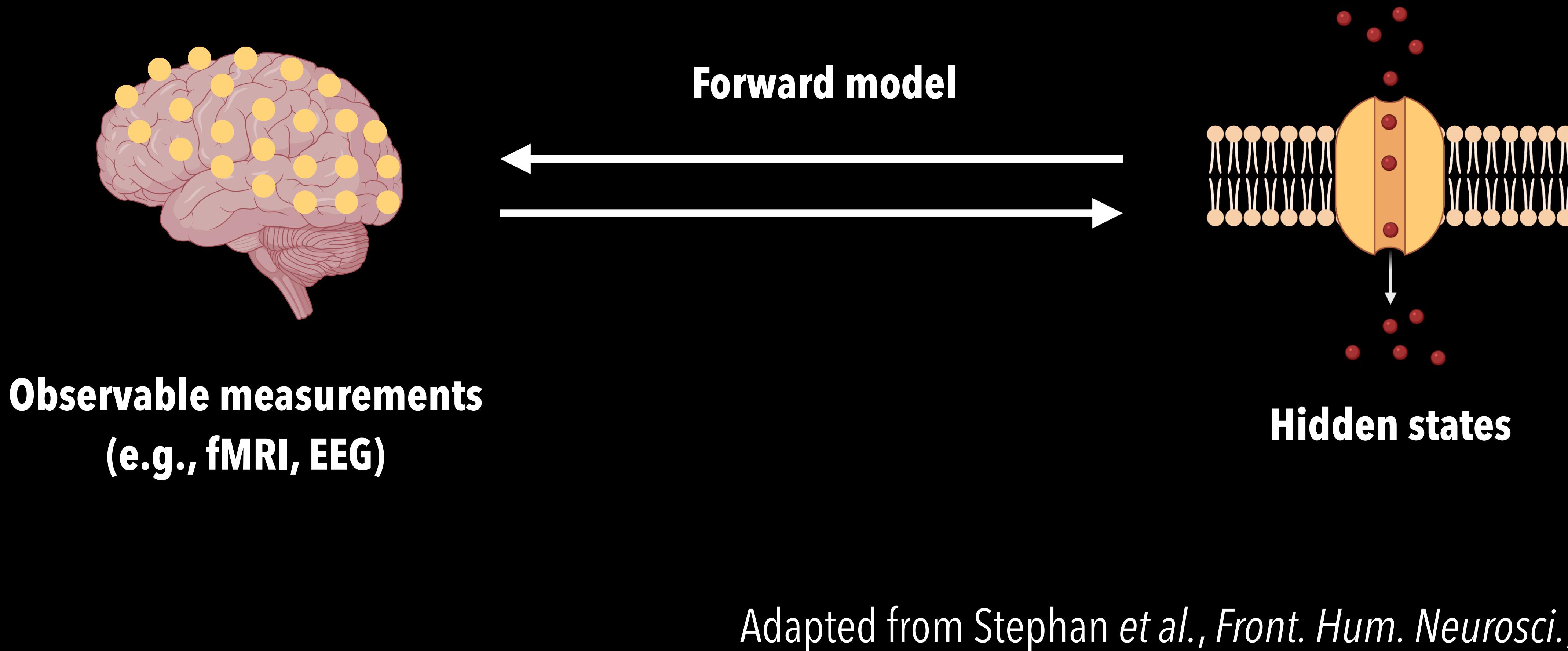
Adapted from Stephan *et al.*, *Front. Hum. Neurosci.*, 2016

Generative model (DCM)

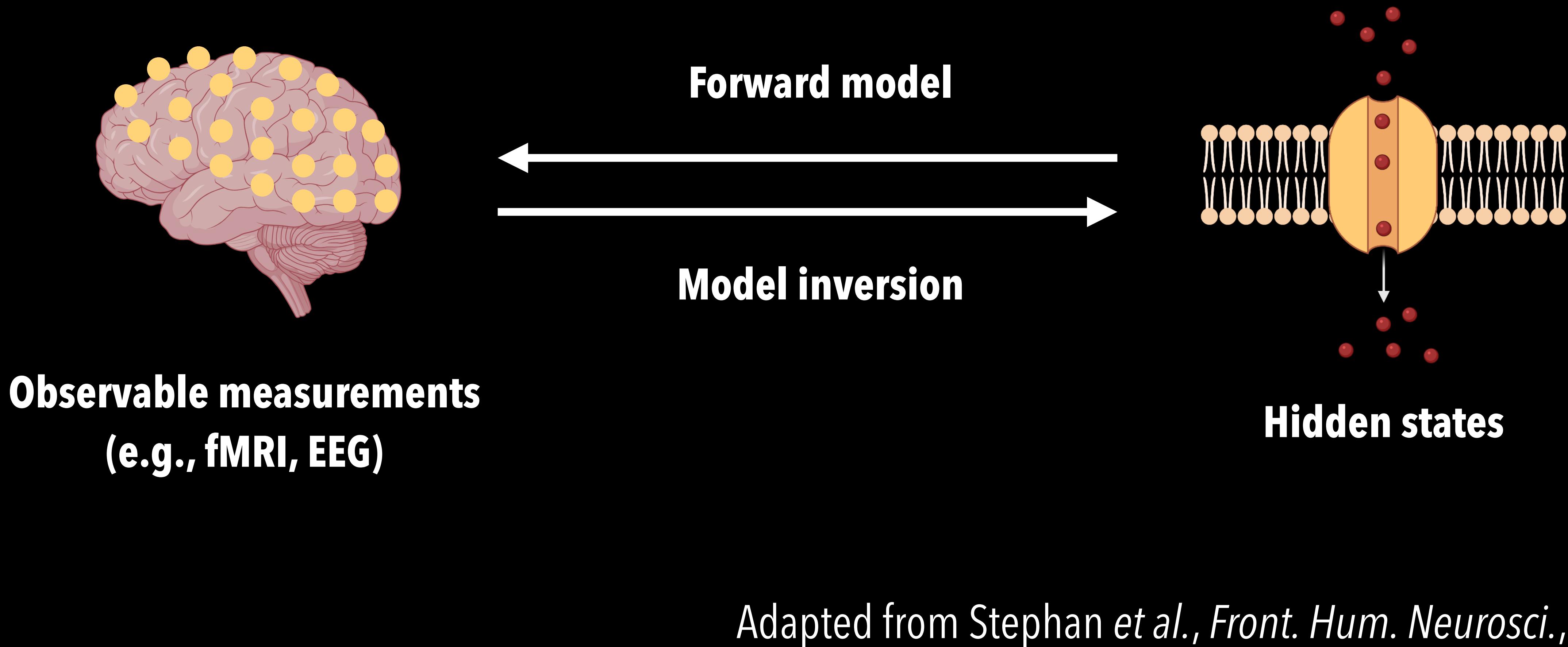


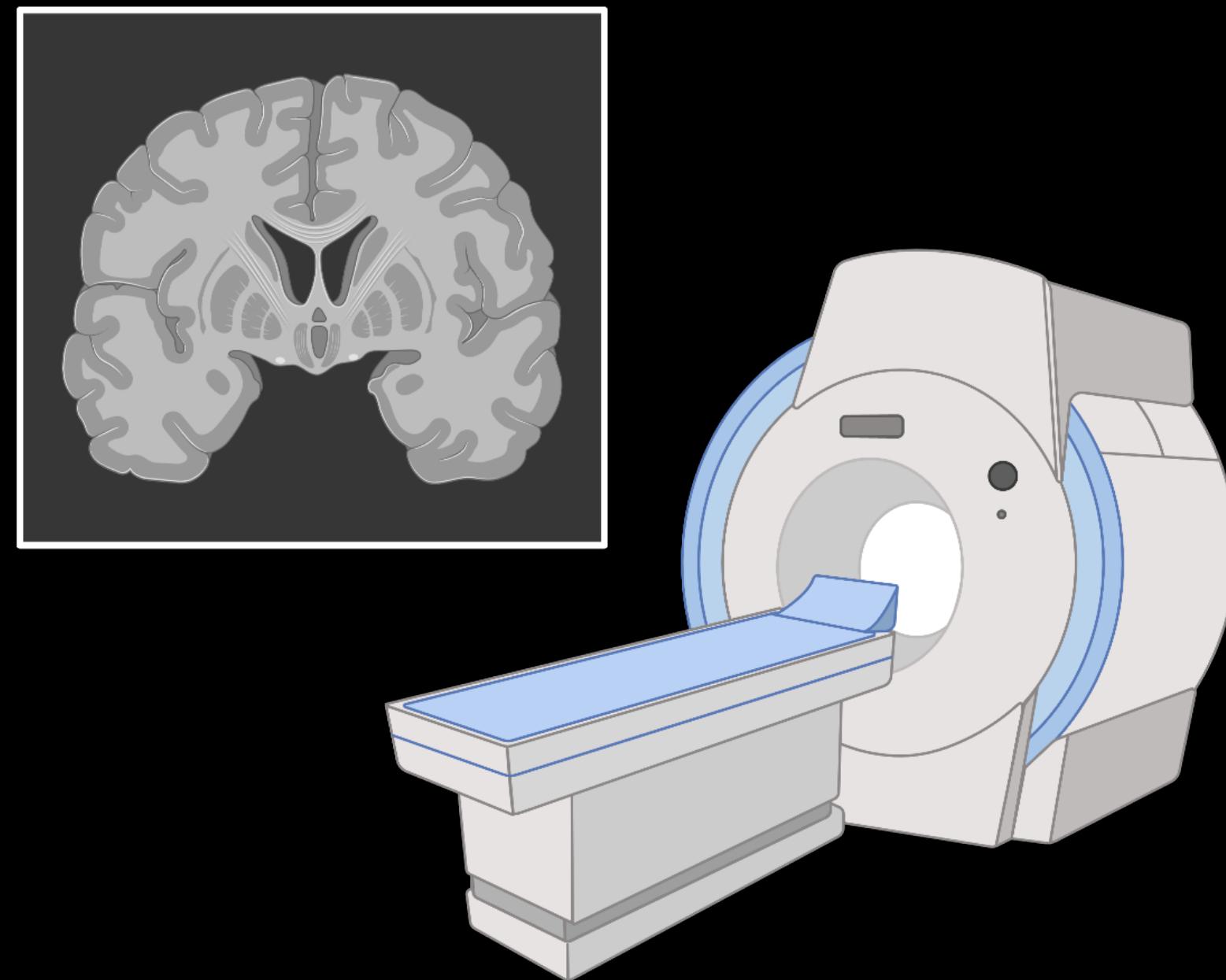


Generative model (DCM)

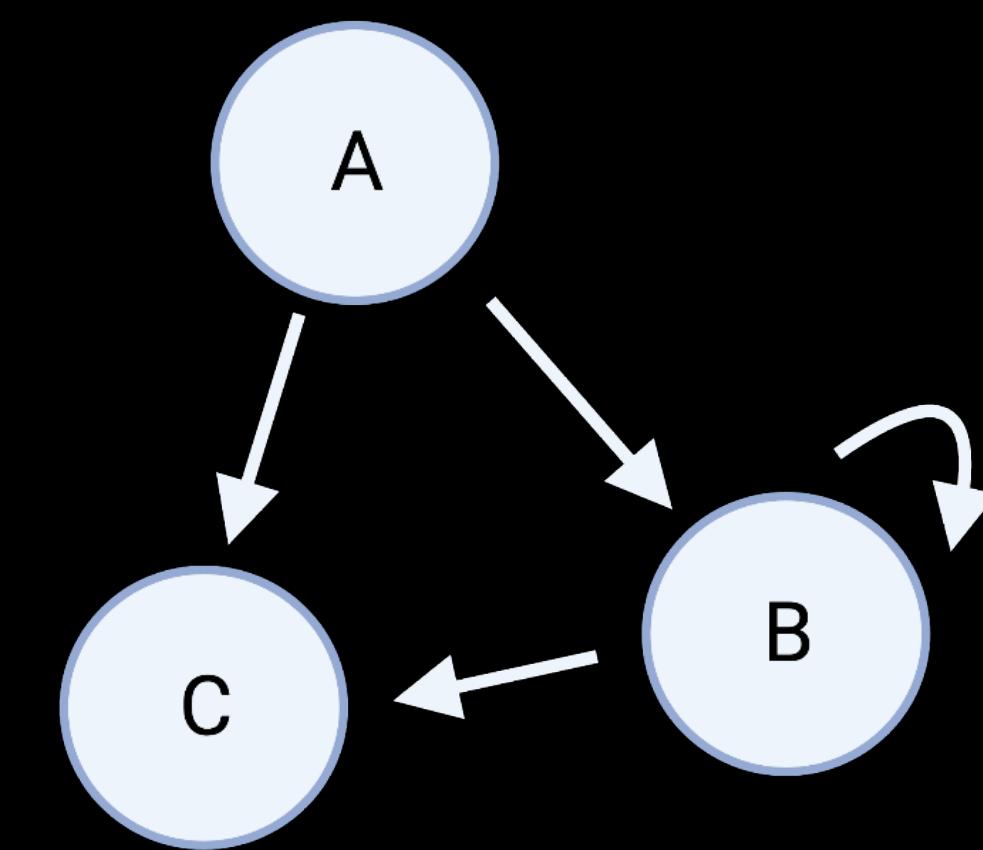


Generative model (DCM)

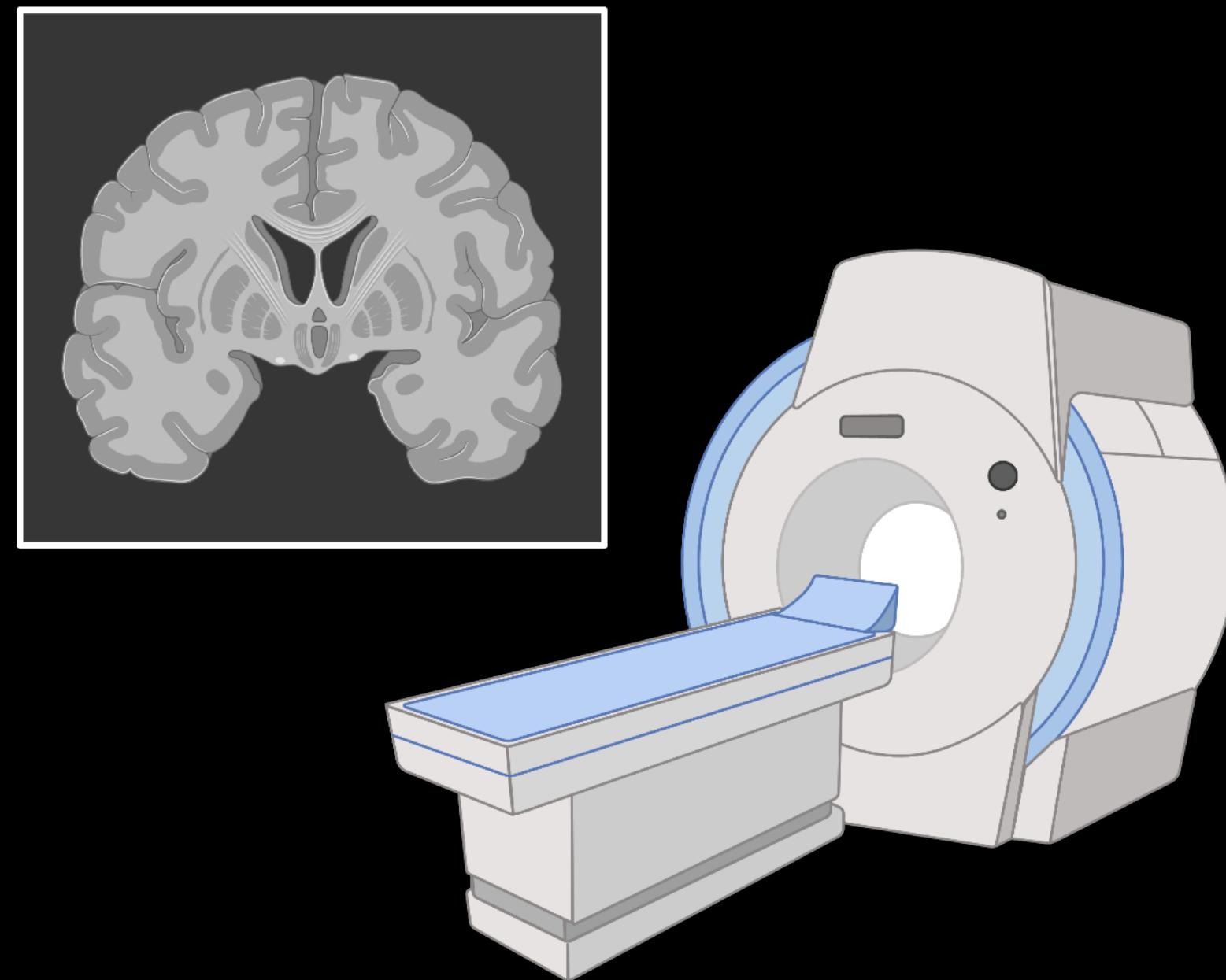




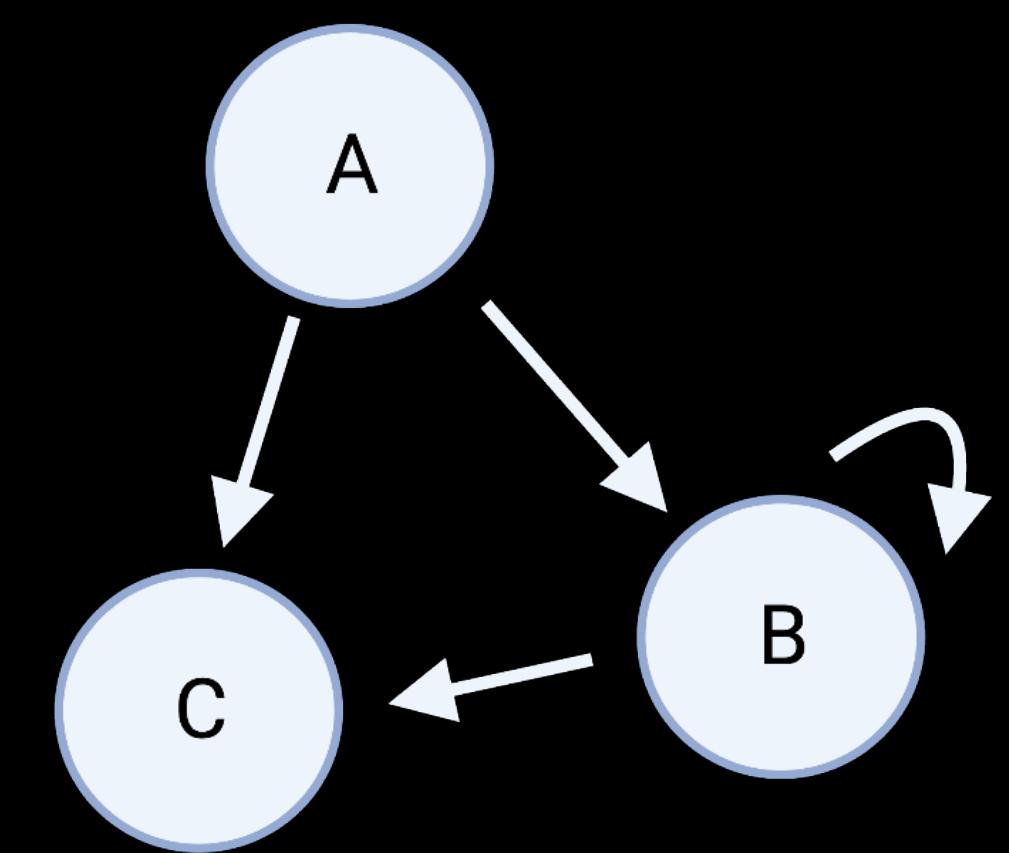
1. Collect subject data



2. Model inversion to
obtain subject-specific
parameter estimates



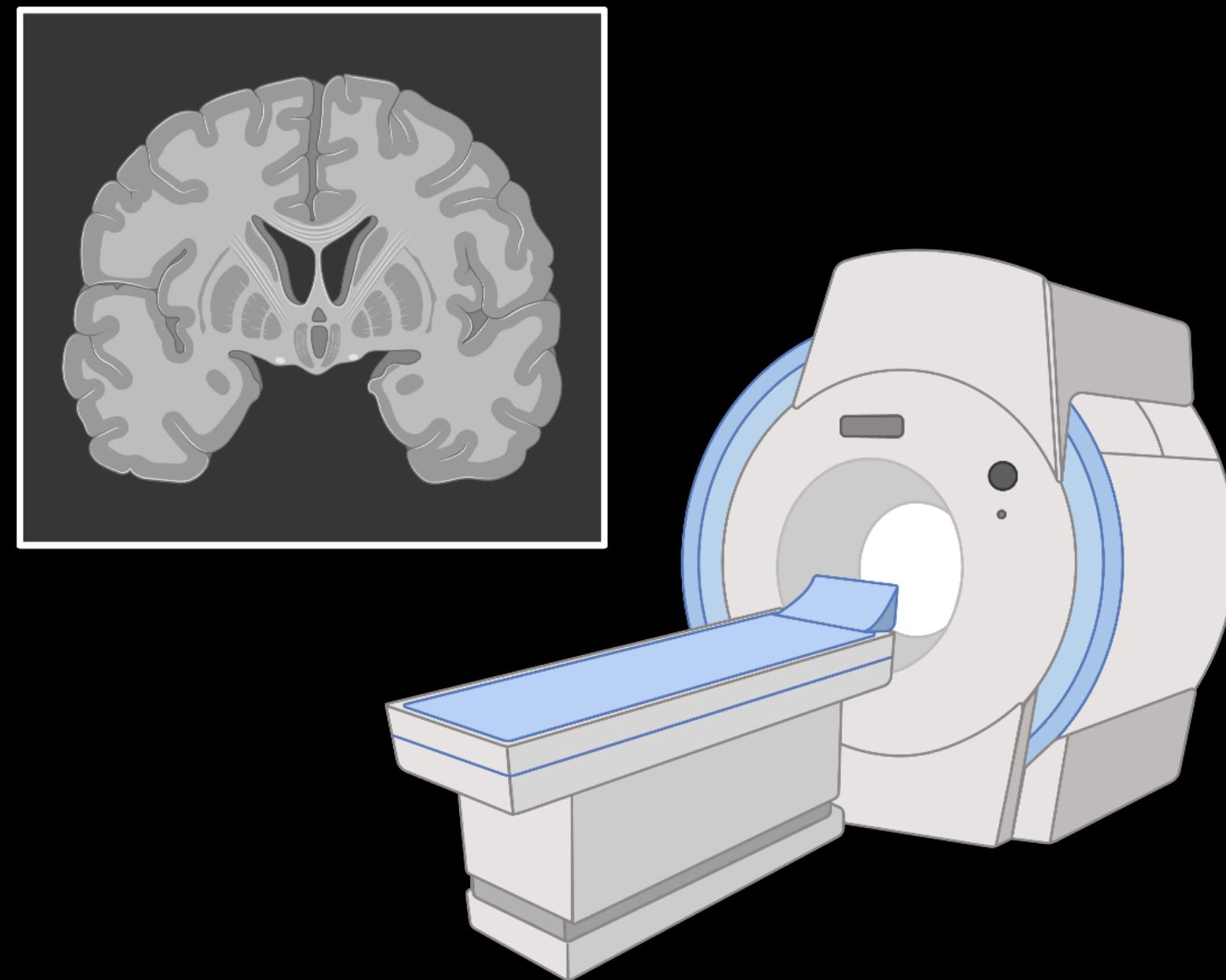
1. Collect subject data



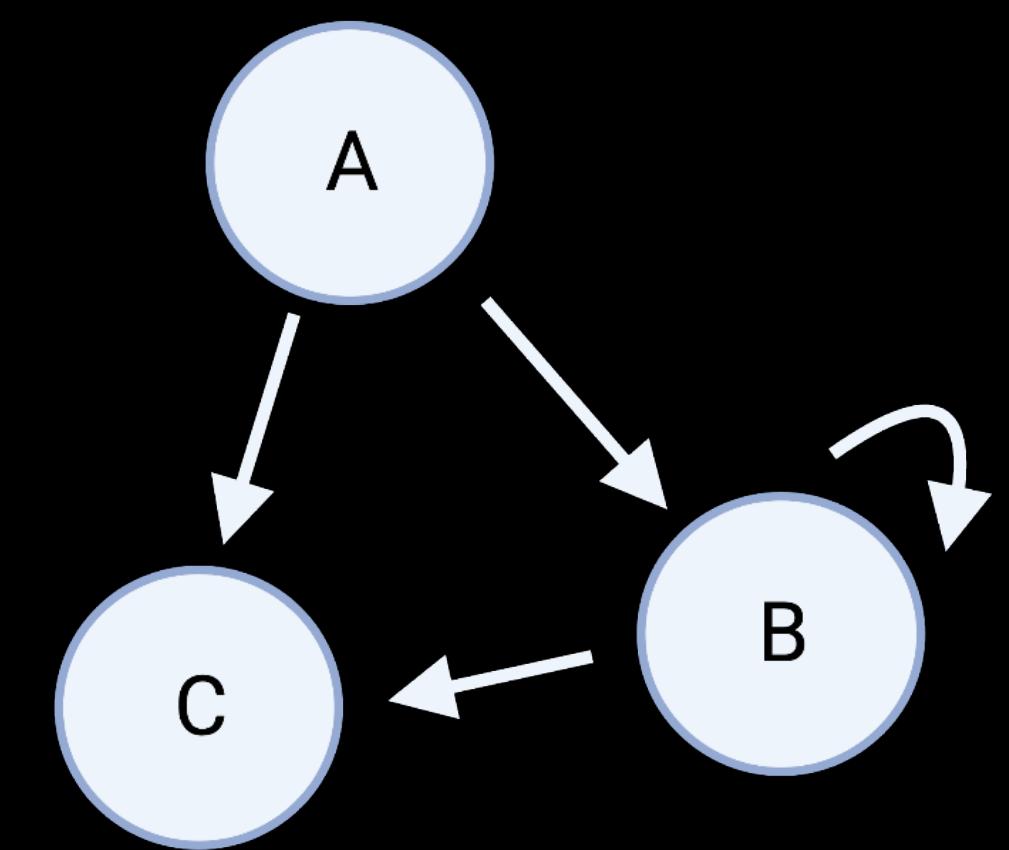
2. Model inversion to obtain subject-specific parameter estimates



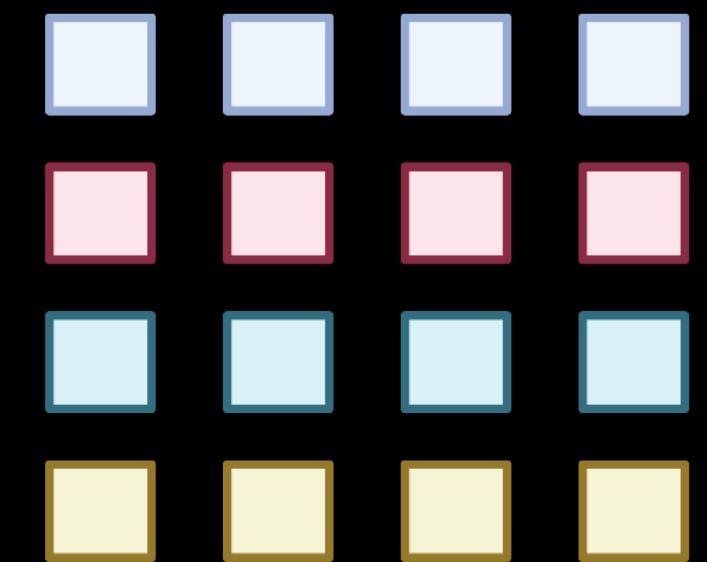
3. Use parameter estimates as new features (embedding)



1. Collect subject data

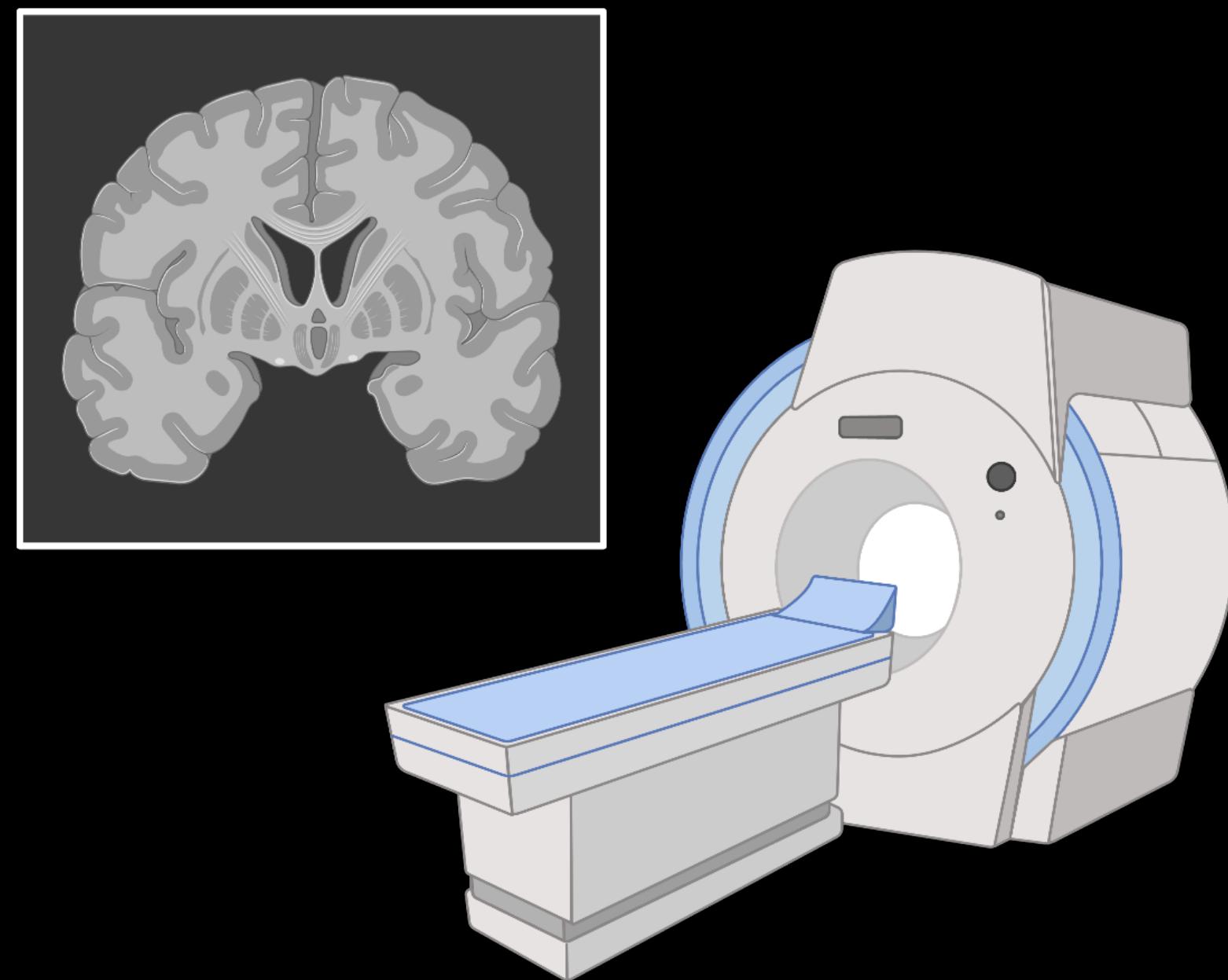


2. Model inversion to obtain subject-specific parameter estimates

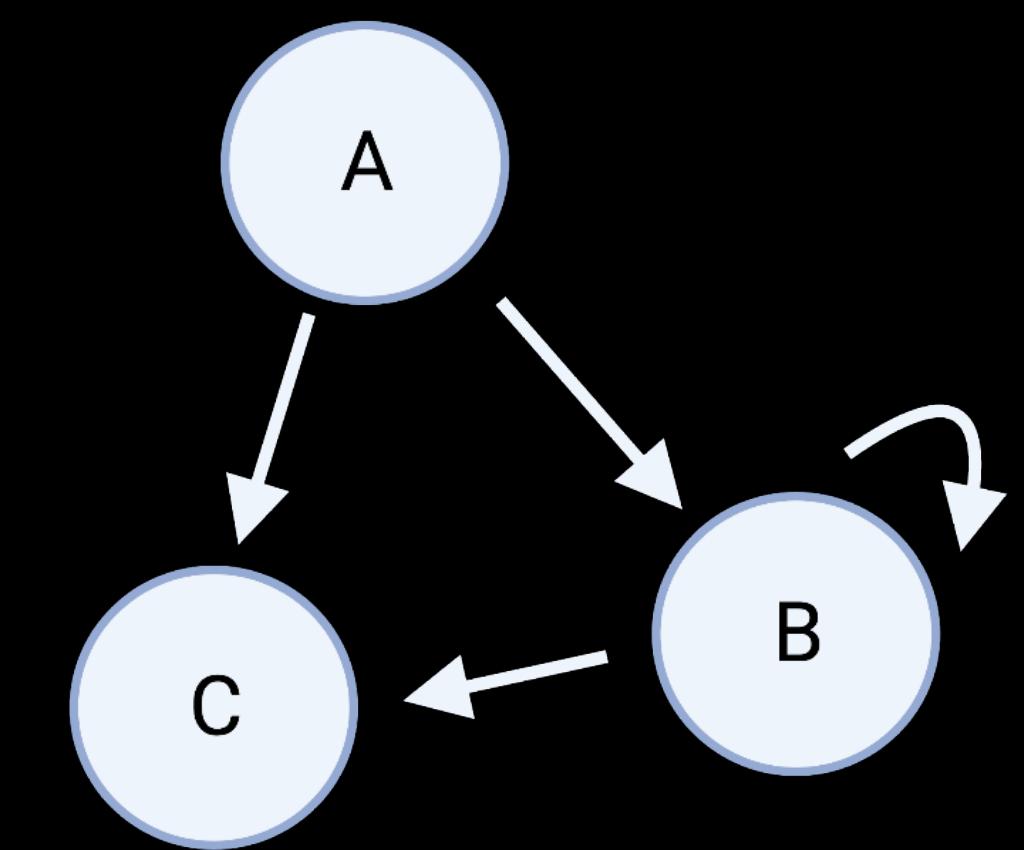


3. Use parameter estimates as new features (embedding)

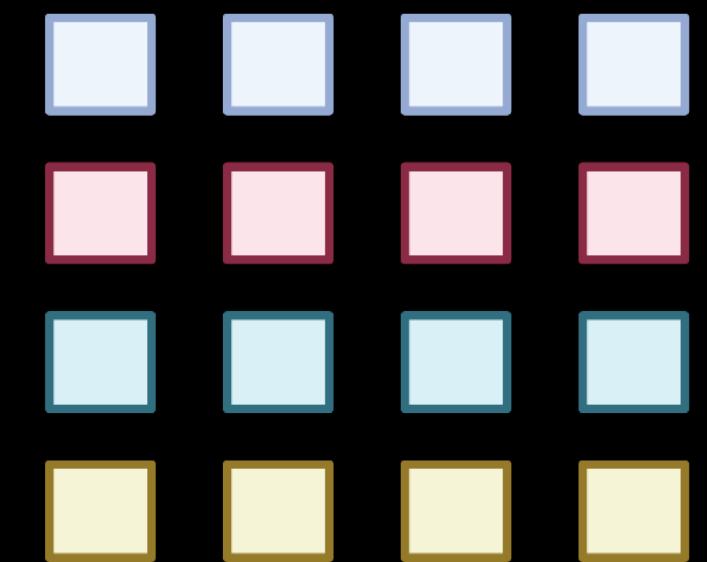
Generative embedding



1. Collect subject data



2. Model inversion to obtain subject-specific parameter estimates



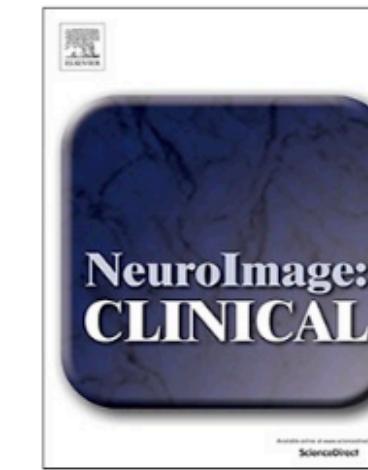
3. Use parameter estimates as new features (embedding)



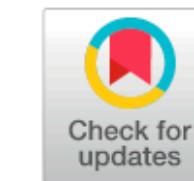
Contents lists available at [ScienceDirect](#)

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynic



Predicting individual clinical trajectories of depression with generative embedding

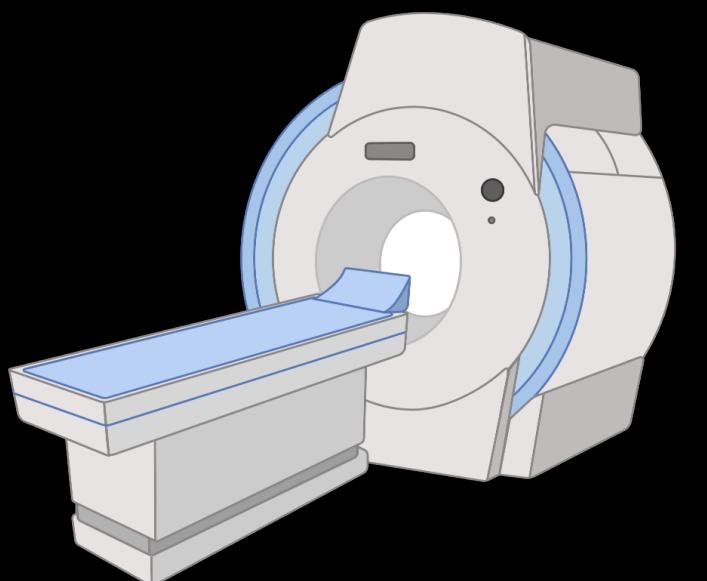


Stefan Frässle^{a,*}, Andre F. Marquand^{b,c}, Lianne Schmaal^{d,e}, Richard Dinga^f, Dick J. Veltman^f, Nic J.A. van der Wee^g, Marie-José van Tol^h, Dario Schöbi^a, Brenda W.J.H. Penninx^{f,i}, Klaas E. Stephan^{a,j,k}

Study design

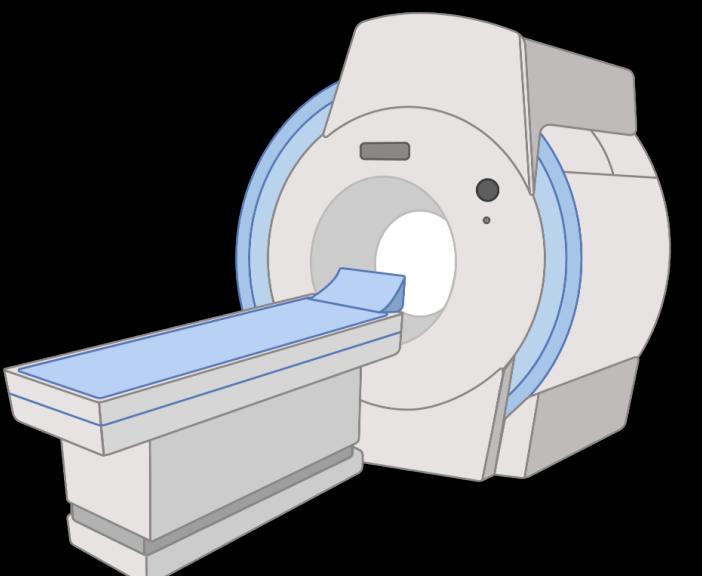
- fMRI data from 85 participants enrolled in the NEtherlands Study of Depression and Anxiety (NESDA)
- DSM-IV diagnosis of major depressive disorder
- Baseline assessment and follow up of depressive symptoms over 2 years

Event-related emotional face-perception paradigm



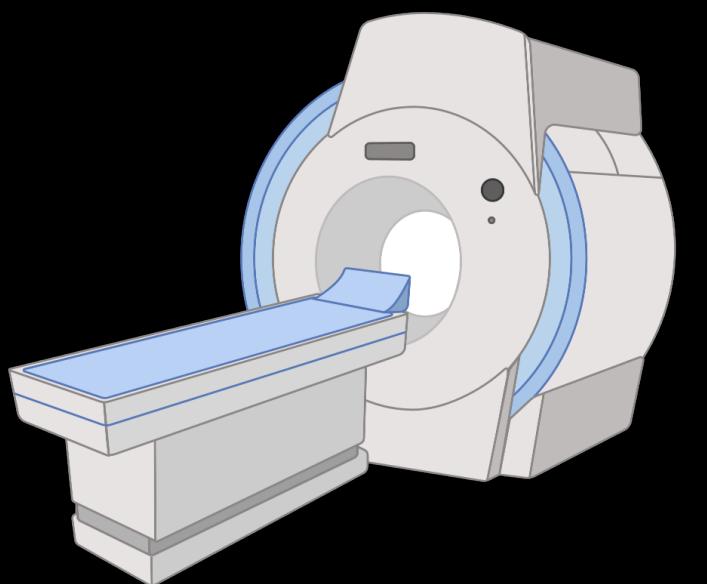
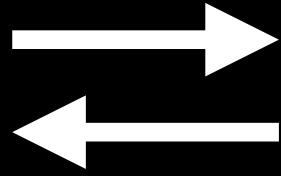
Photographs from Verpaalen *et al.*, *Cognition and Emotion*, 2019

Event-related emotional face-perception paradigm



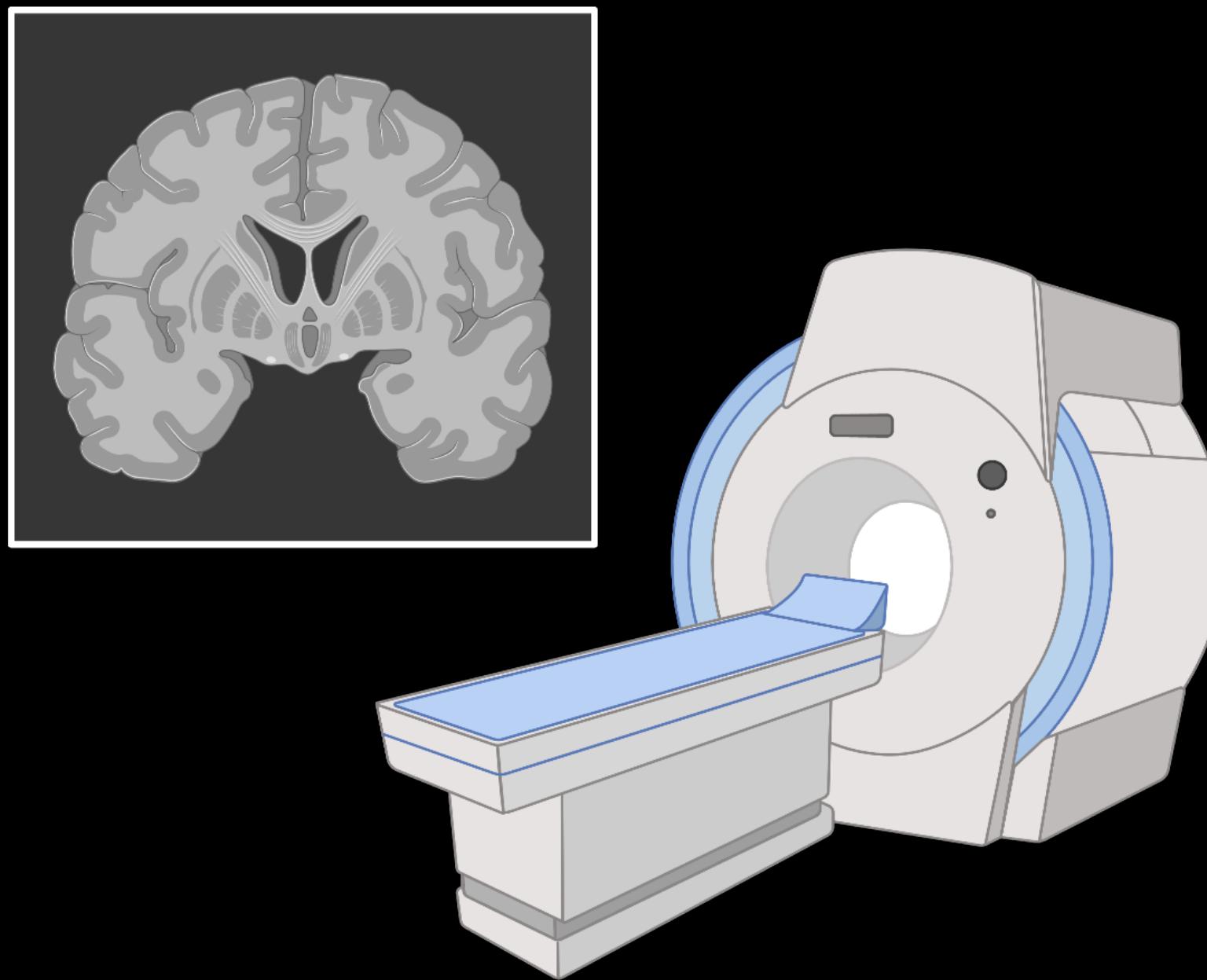
Photographs from Verpaalen *et al.*, *Cognition and Emotion*, 2019

Event-related emotional face-perception paradigm



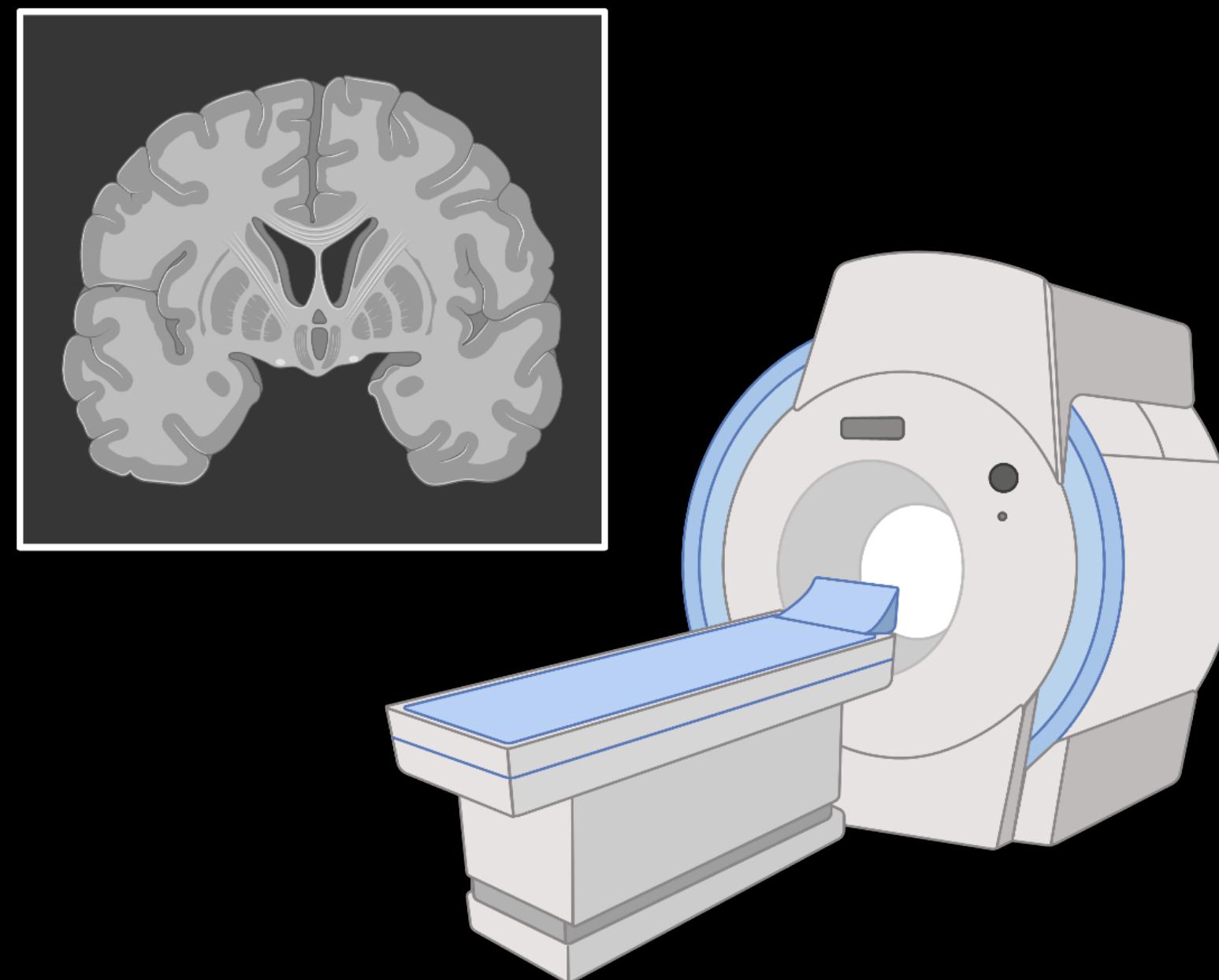
Photographs from Verpaalen *et al.*, *Cognition and Emotion*, 2019

Generative embedding

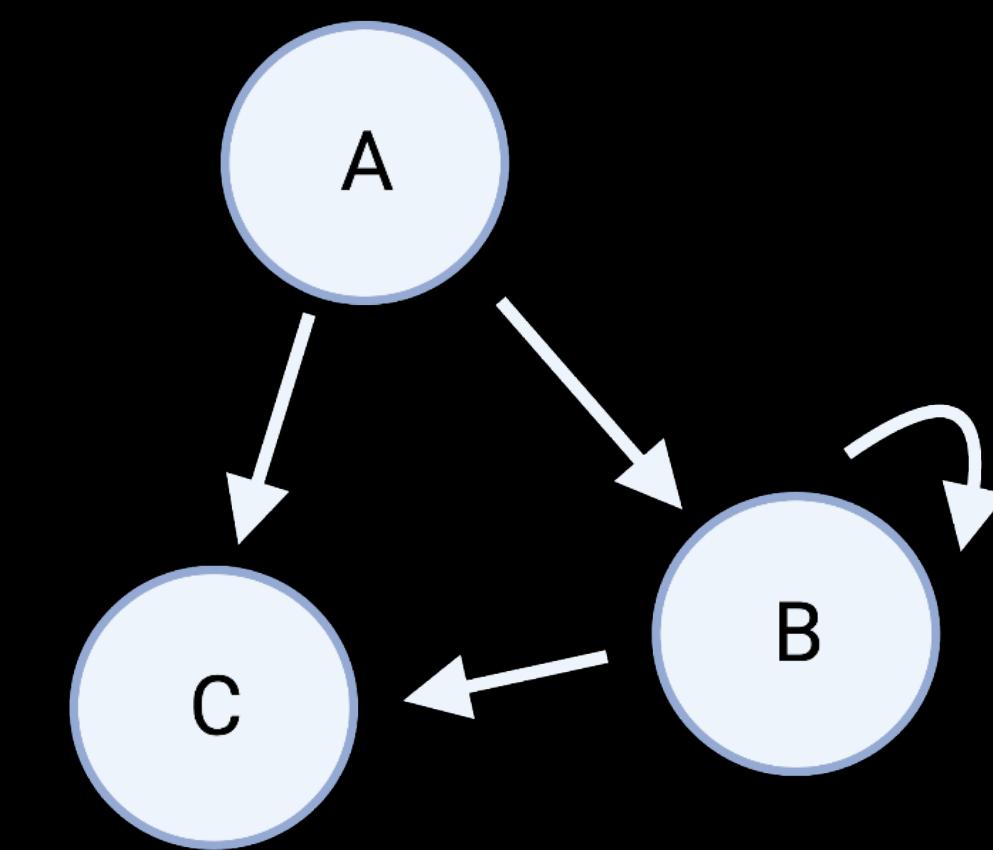


1. Collect subject data

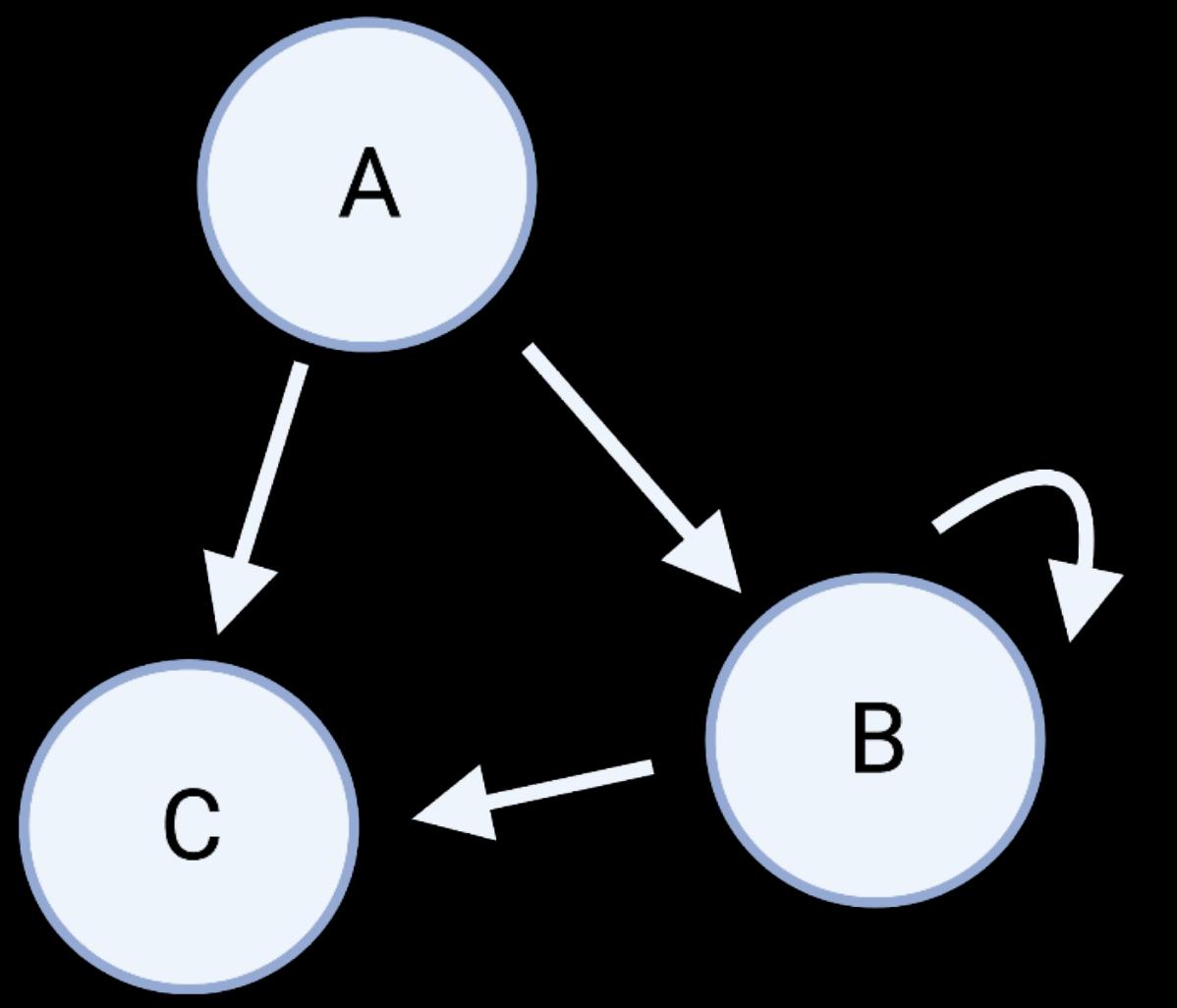
Generative embedding

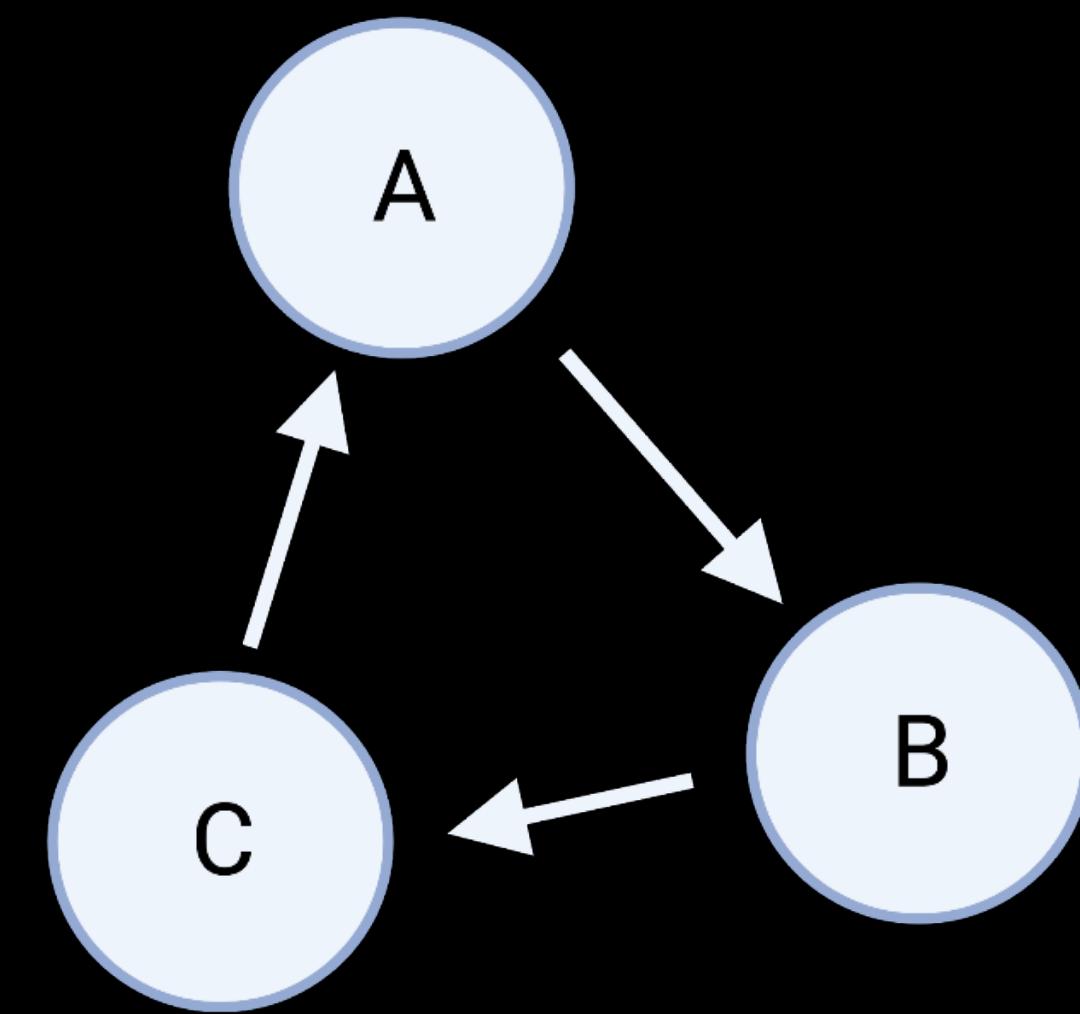
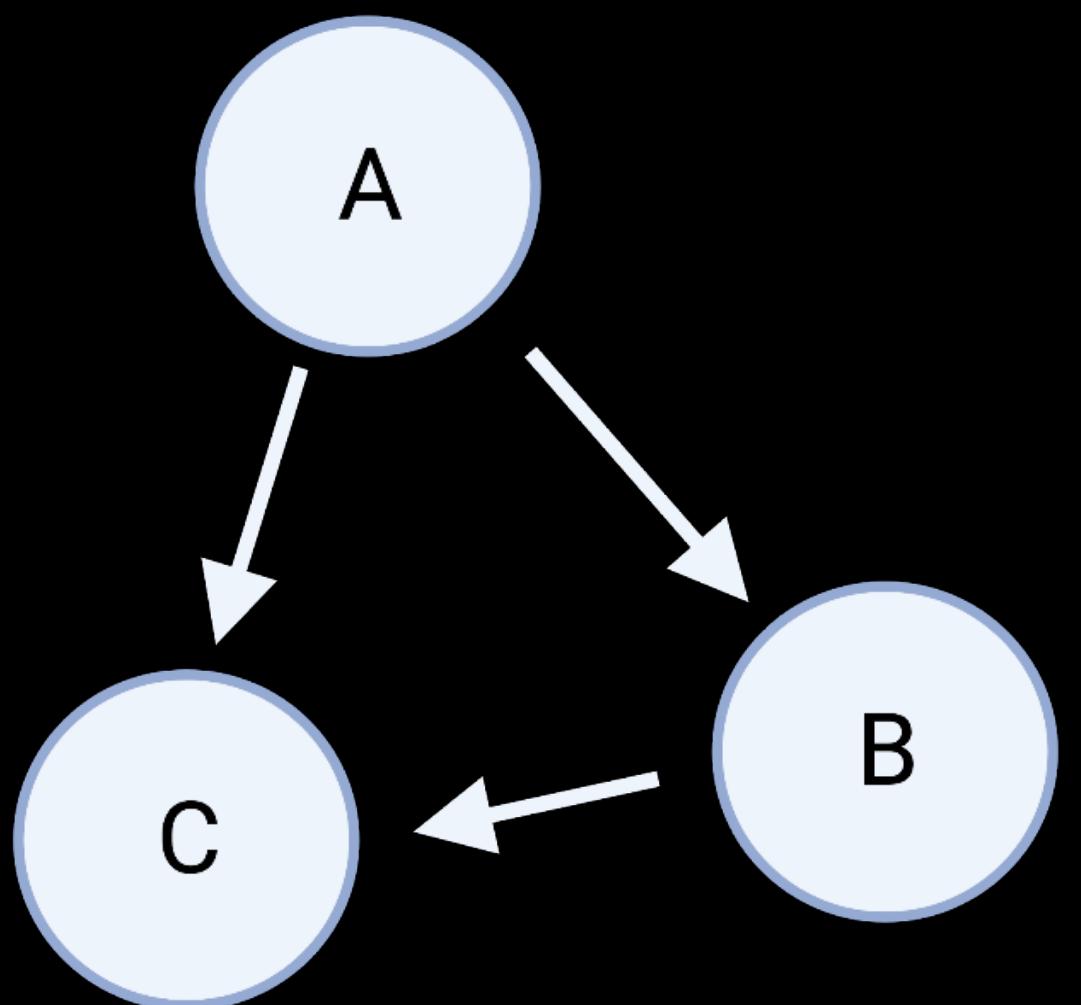
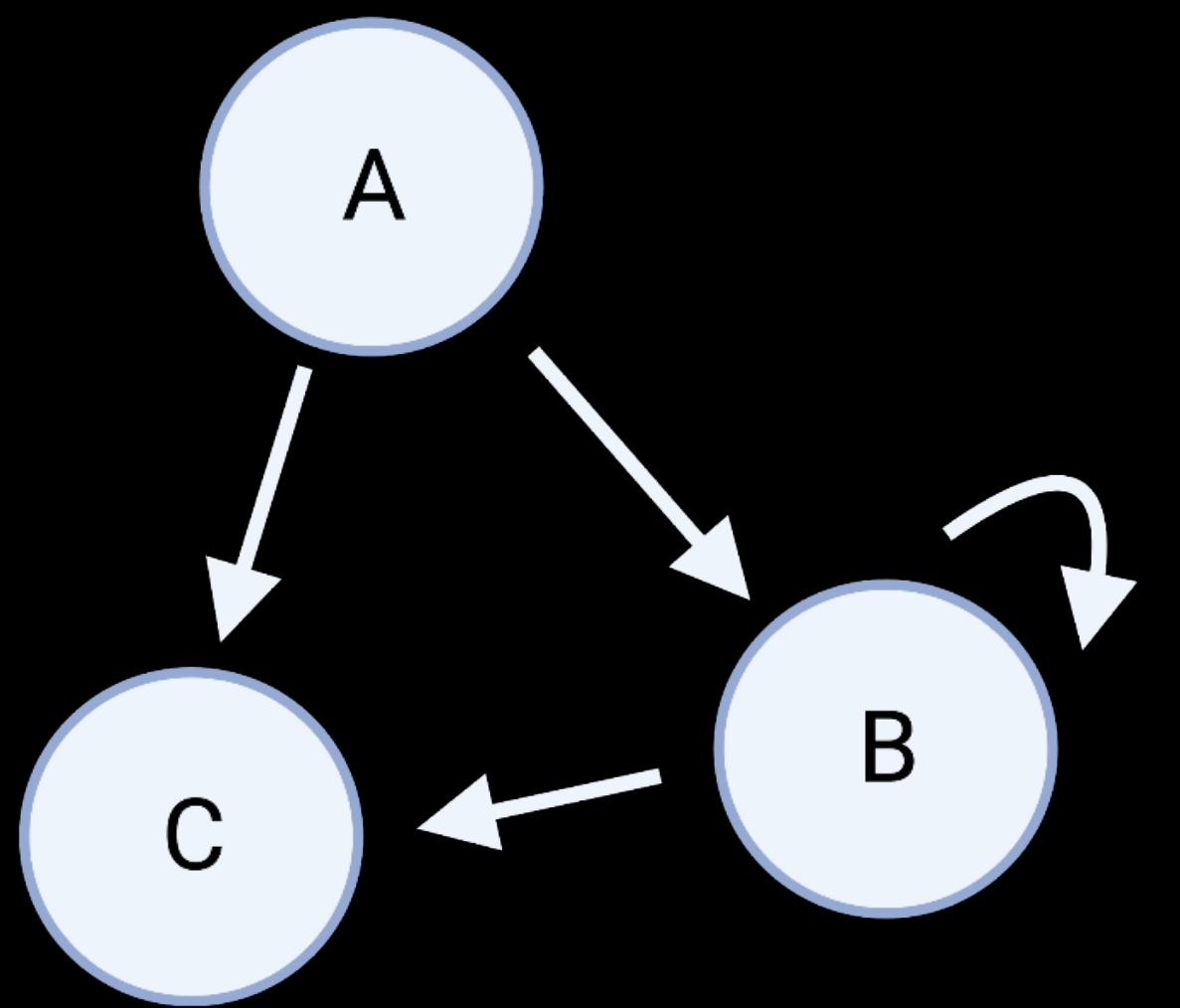


1. Collect subject data

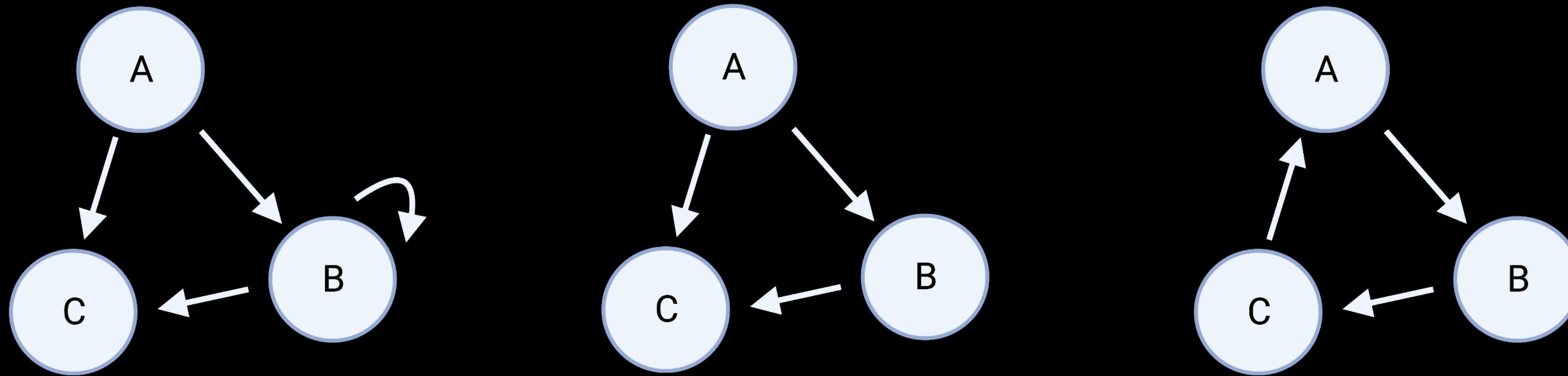


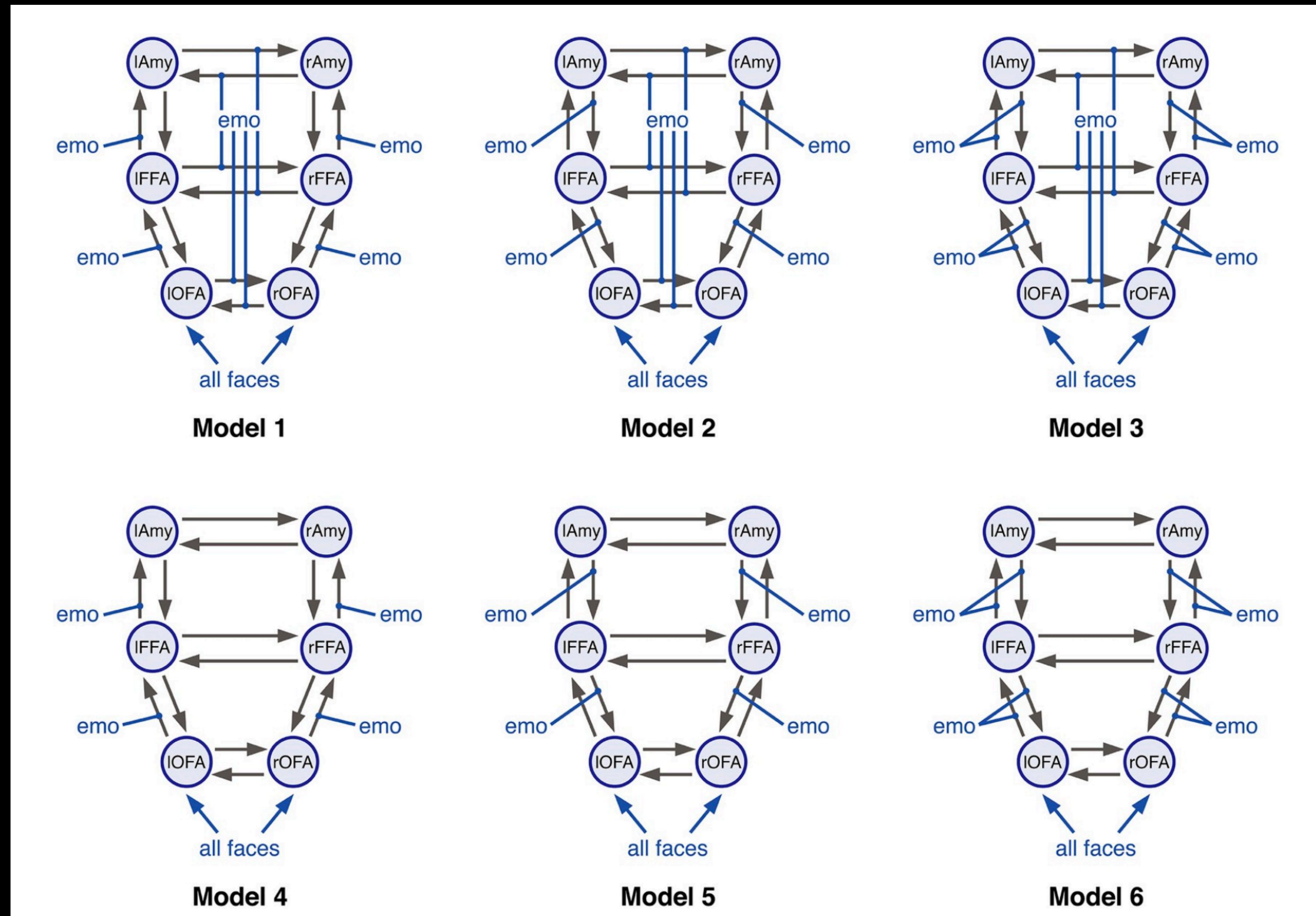
2. Model inversion to
obtain subject-specific
parameter estimates



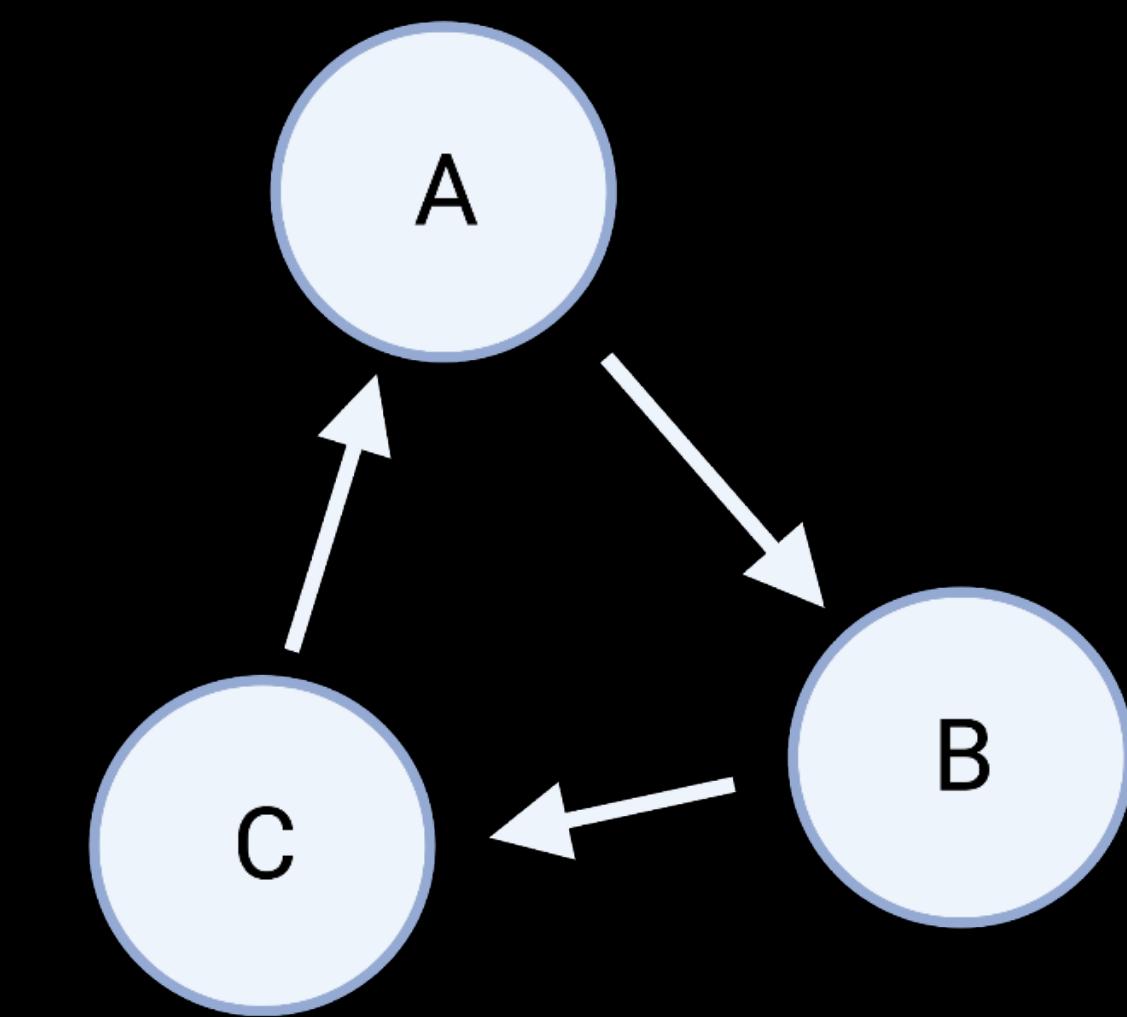
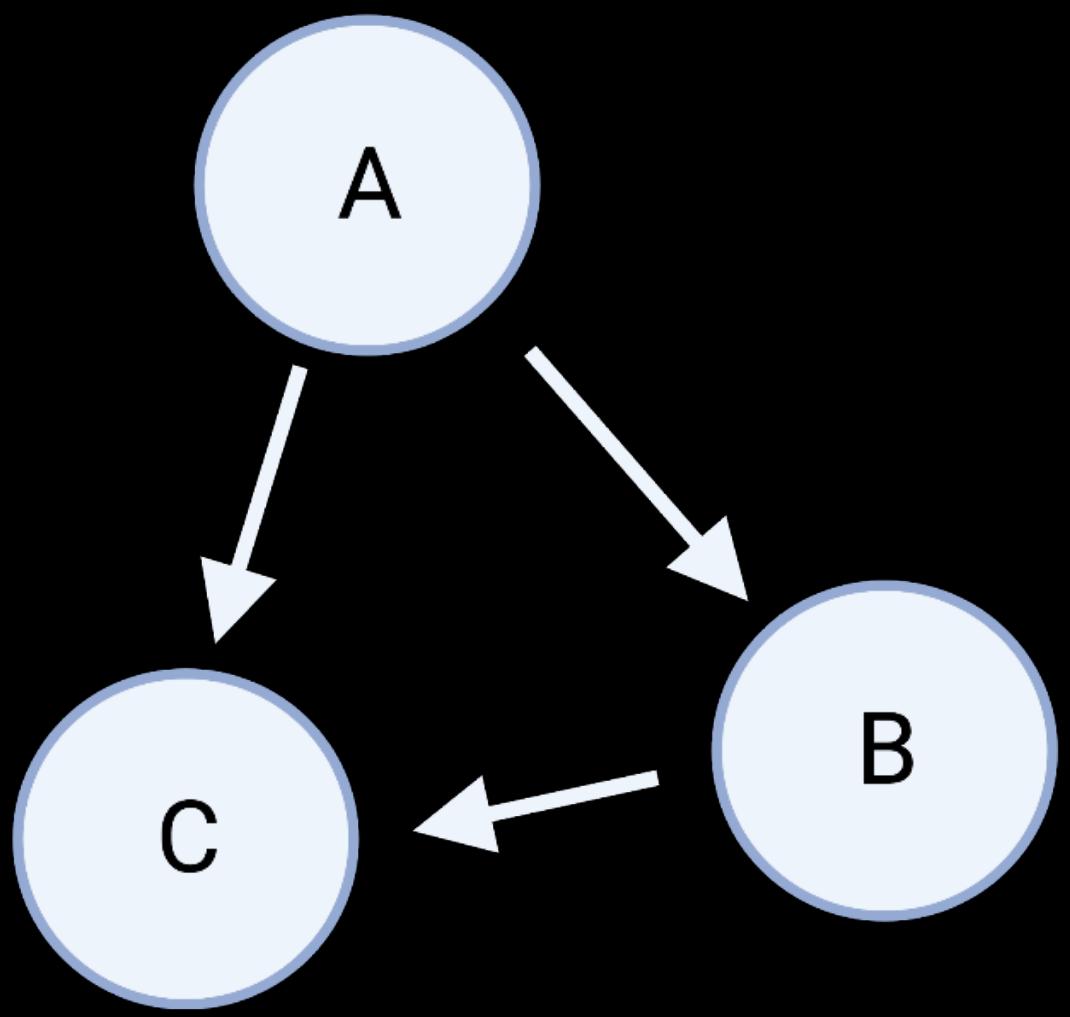
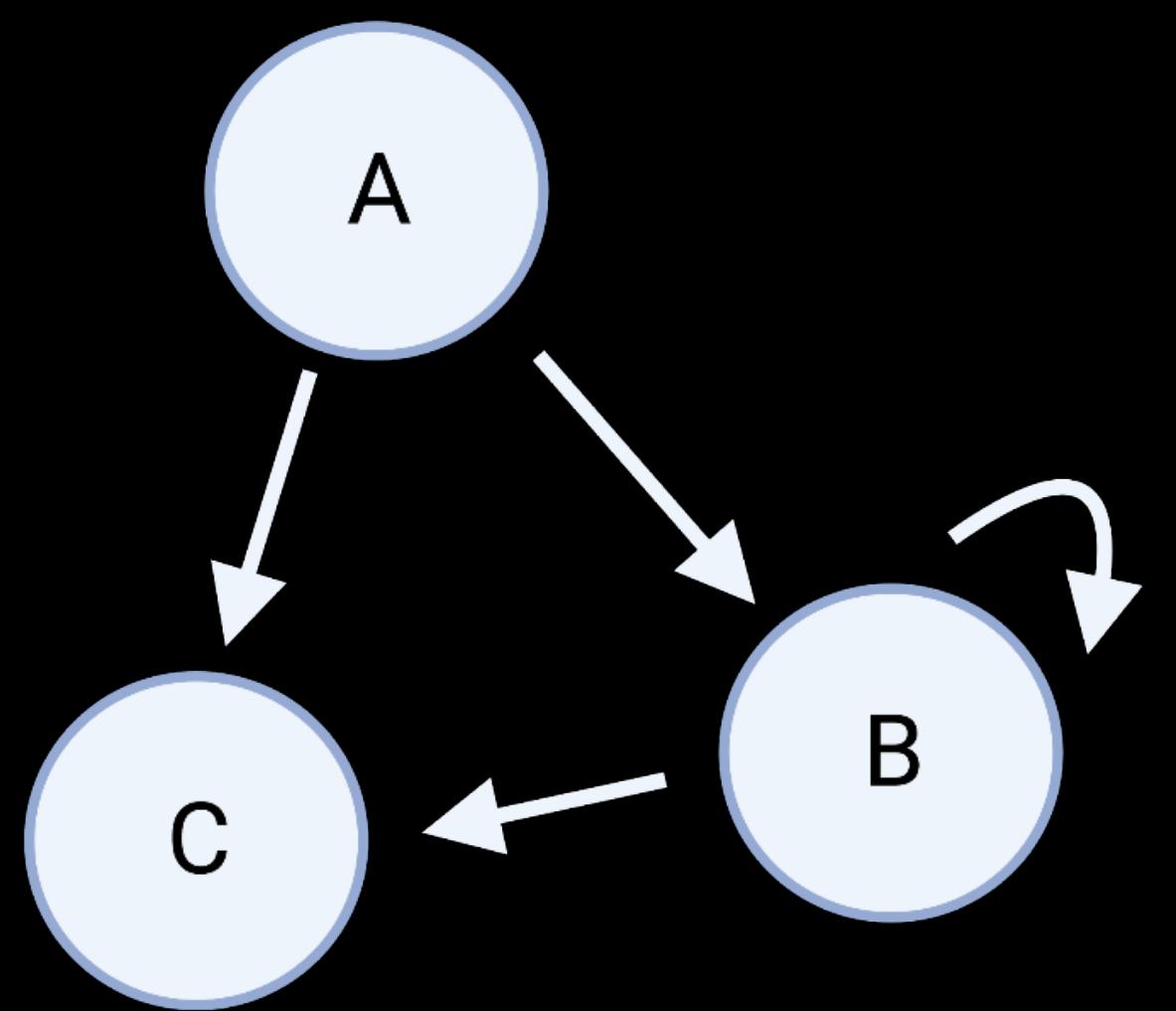


Model space

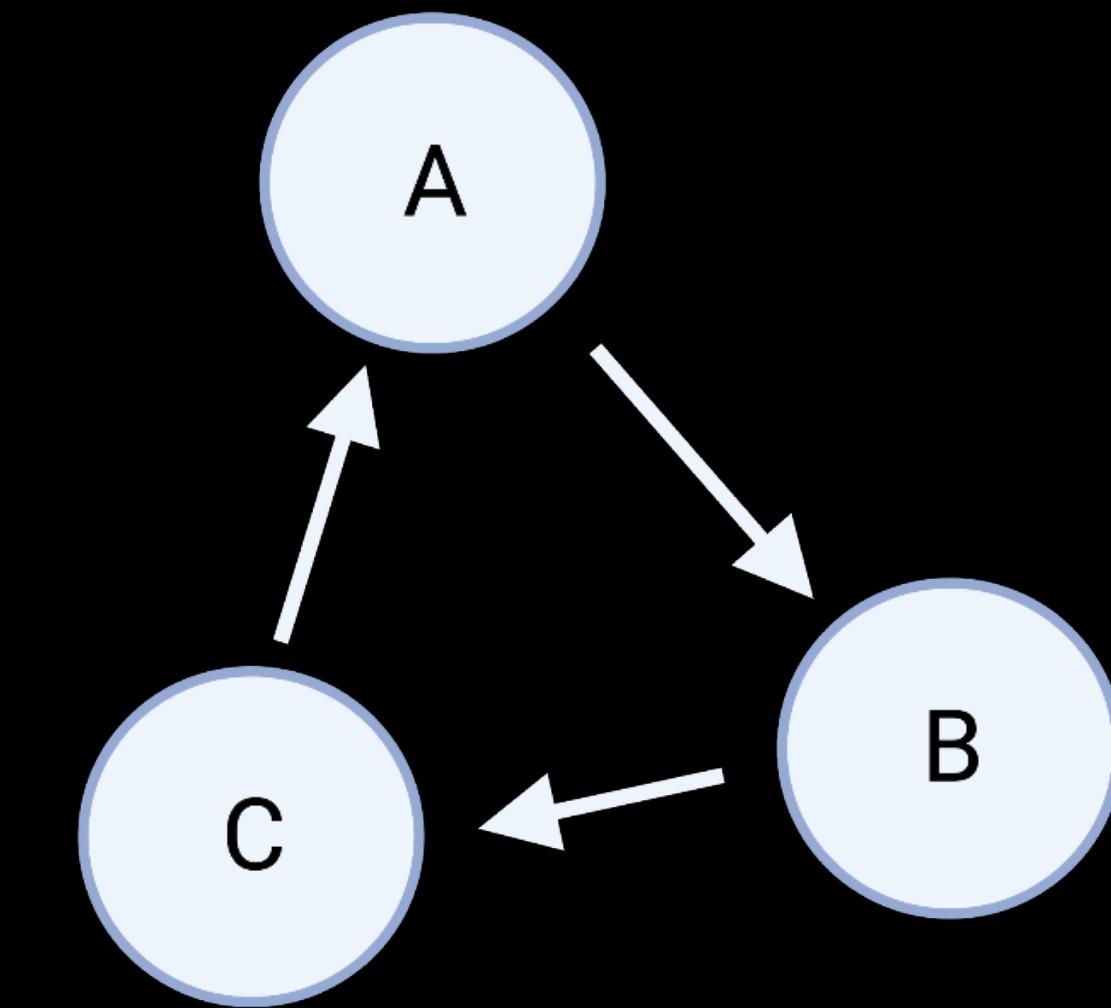
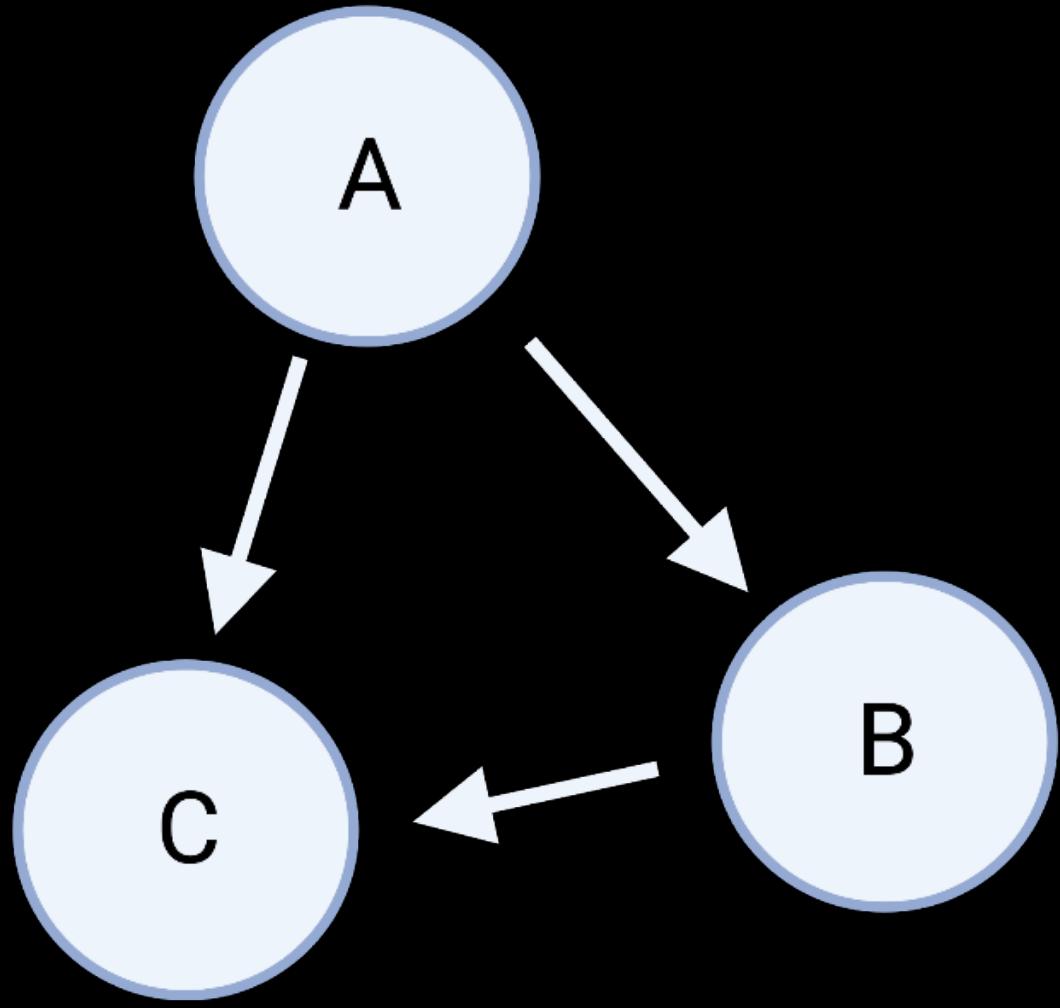
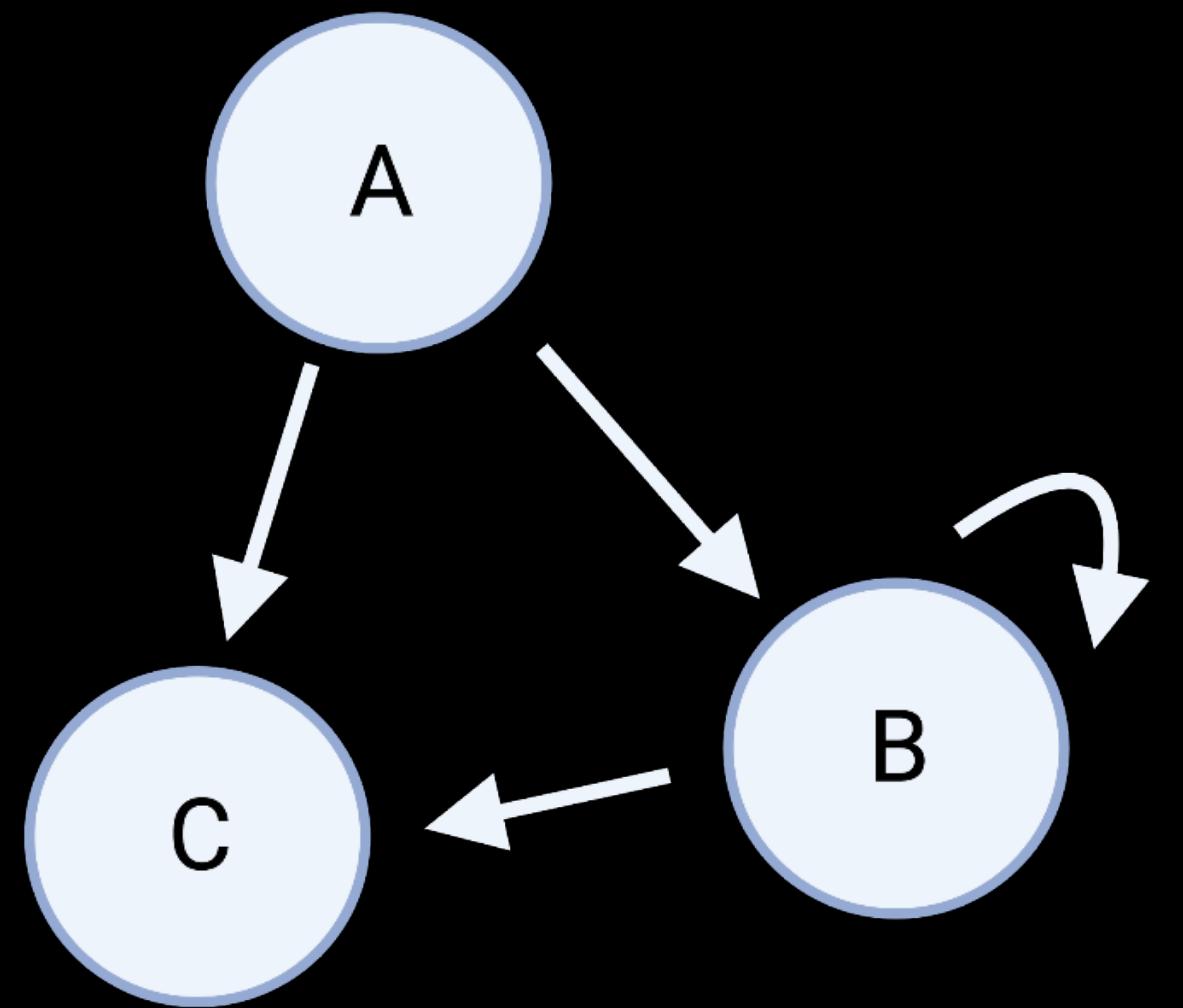




From Frässle *et al.*, *NeuroImage: Clinical*, 2020







Bayesian model averaging



Classification

- Binary pairwise classification using SVM, with nested cross-validation
- Classes:
 - Rapid remission of symptoms (REM)
 - Slow but gradual improvement of symptoms (IMP)
 - No improvement of symptoms (chronic, CHR)

Classification results

Table 2

Classification results for the generative embedding procedure. Shown are key performance measures of the classification algorithm, including: balanced accuracy, area under the curve, sensitivity (recall), specificity, positive predictive value (precision), and negative predictive value. Performance measures are shown for the three different binary classifications (i.e., CHR vs. REM, CHR vs. IMP, and IMP vs. REM).

Classification	CHR (<i>n</i> = 15) vs. REM (<i>n</i> = 39)	CHR (<i>n</i> = 15) vs. IMP (<i>n</i> = 31)	IMP (<i>n</i> = 31) vs. REM (<i>n</i> = 39)
Accuracy	0.87	0.63	0.63
Balanced accuracy	0.79	0.47	0.61
Area under the curve (AUC)	0.87	0.35	0.63
Sensitivity (recall)	0.97	0.94	0.77
Specificity	0.60	0	0.45
Positive predictive value (Precision)	0.86	0.66	0.64
Negative predictive value	0.90	0	0.61

Classification results

Table 2

Classification results for the generative embedding procedure. Shown are key performance measures of the classification algorithm, including: balanced accuracy, area under the curve, sensitivity (recall), specificity, positive predictive value (precision), and negative predictive value. Performance measures are shown for the three different binary classifications (i.e., CHR vs. REM, CHR vs. IMP, and IMP vs. REM).

Classification	CHR (<i>n</i> = 15) vs. REM (<i>n</i> = 39)	CHR (<i>n</i> = 15) vs. IMP (<i>n</i> = 31)	IMP (<i>n</i> = 31) vs. REM (<i>n</i> = 39)
Accuracy	0.87	0.63	0.63
Balanced accuracy	0.79	0.47	0.61
Area under the curve (AUC)	0.87	0.35	0.63
Sensitivity (recall)	0.97	0.94	0.77
Specificity	0.60	0	0.45
Positive predictive value (Precision)	0.86	0.66	0.64
Negative predictive value	0.90	0	0.61

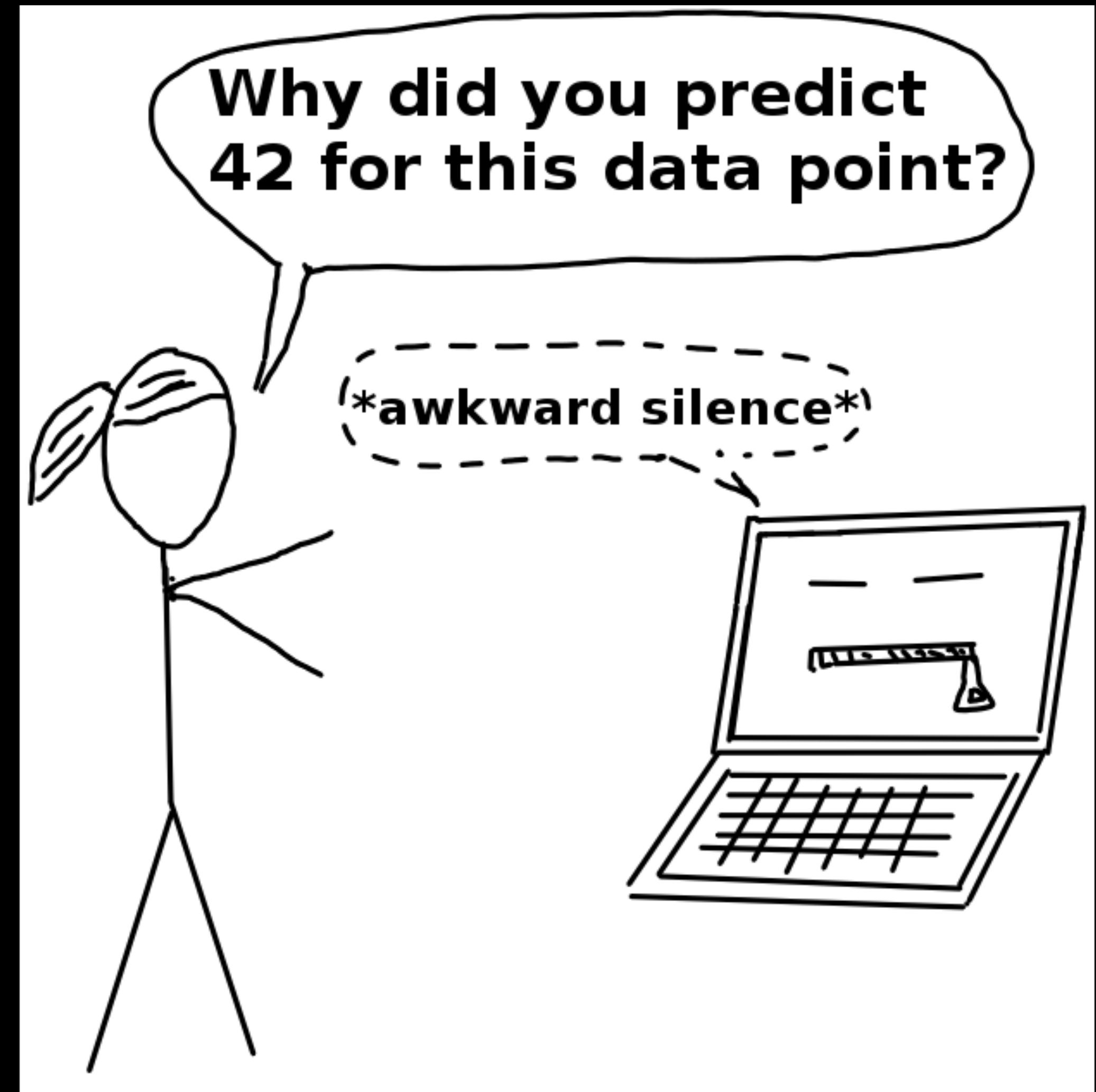


Photo by Nicola Anderson on Unsplash

In short ...

- Model interpretability is important in healthcare
- Interpretability can improve accuracy
- For Computational Psychiatry: can these models further our understanding?

Thank you for your attention!



<https://christophm.github.io/interpretable-ml-book/>

Acknowledgements

- Thank you Stefan Frässle for all your input and discussions!
- Figures created with [BioRender.com](#)