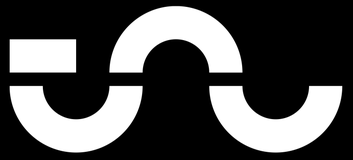
University of Zurich[UZH]

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich



Translational Neuromodeling Unit

Model interpretability in healthcare

Inês Pereira

March 10th, 2021

Disclaimer

Context

Research

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 Dete in Retin

Varun Gulshan, PhD
Subhashini Venugo
Rajiv Raman, MS, D

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 Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Research

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Dete in Retin

Varun Gulshan, PhD
Subhashini Venugo
Rajiv Raman, MS, D

LETTER

doi:10.1038/nature21056

Derma with de

Andre Esteva[1]*, I

News > Medscape Medical News > FDA Approvals

FDA Approves AliveCor Personal ECG Monitor for Apple Watch

Steve Stiles

November 30, 2017

Research

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

Andre Esteva[1]*,

ECG Monitor for Apple Watch

Steve Stiles

November 30, 2017

Research

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 Apple Watch

Andre Esteva[1]*,

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*

Research

1322 IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 35, NO. 5, MAY 2016

Unsupervi
Density

Michiel Kallenberg*,
Celine M. Vachon, K

And

r for

1207

l

Convolutional Neural Network

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

SCIENTIFIC REPORTS
nature research

OPEN

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

Su Yang[1], Jihoon Kweon[1,2,5*], Jae-Hyung Roh[3], Jae-Hwan Lee[3], Heejun Kang[1], Lae-Jeong Park[4], Dong Jun Kim[1], Hyeonkyeong Yang[1], Jaehee Hur[1], Do-Yoon Kang[1], Pil Hyung Lee[1], Jung-Min Ahn[1], Soo-Jin Kang[1], Duk-Woo Park[1], Seung-Whan Lee[1], Young-Hak Kim[1,5*], Cheol Whan Lee[1], Seong-Wook Park[1] & Seung-Jung Park[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?

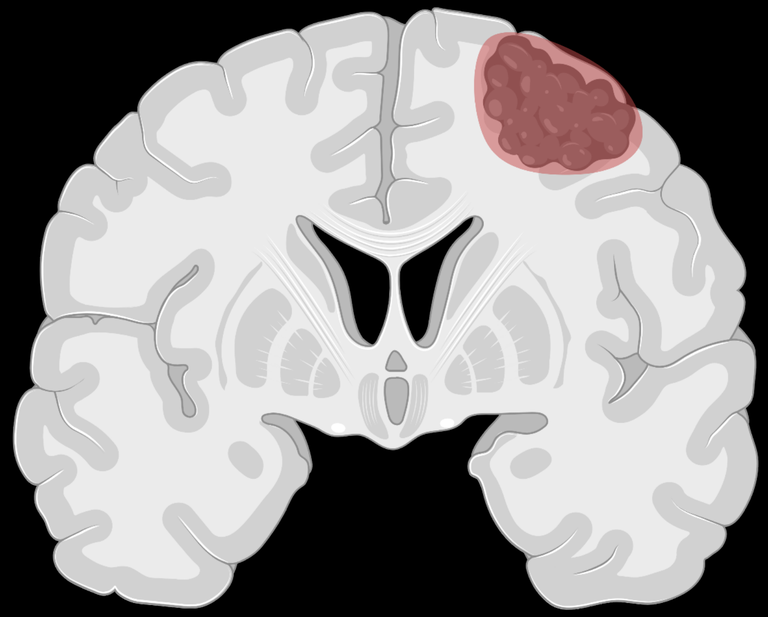
9:37 PM · Feb 20, 2020 · Twitter Web App

1,159 Retweets **614** Quote Tweets **5,172** Likes

Is it a dichotomy?

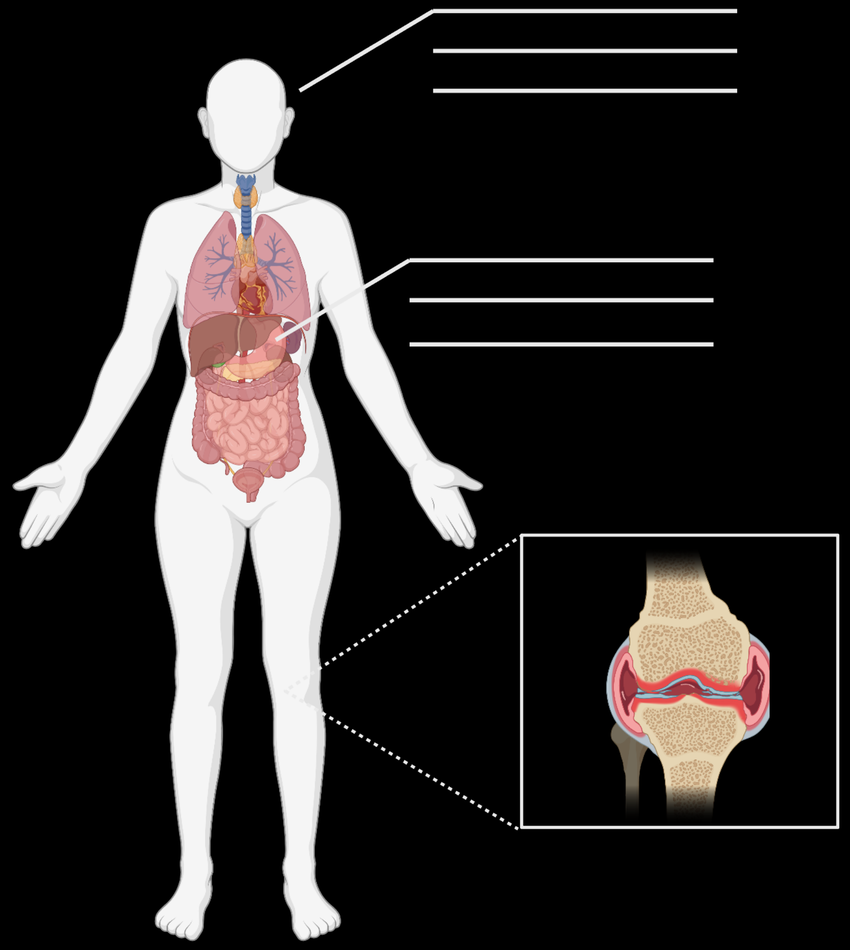
How do you define success?

Tumor segmentation



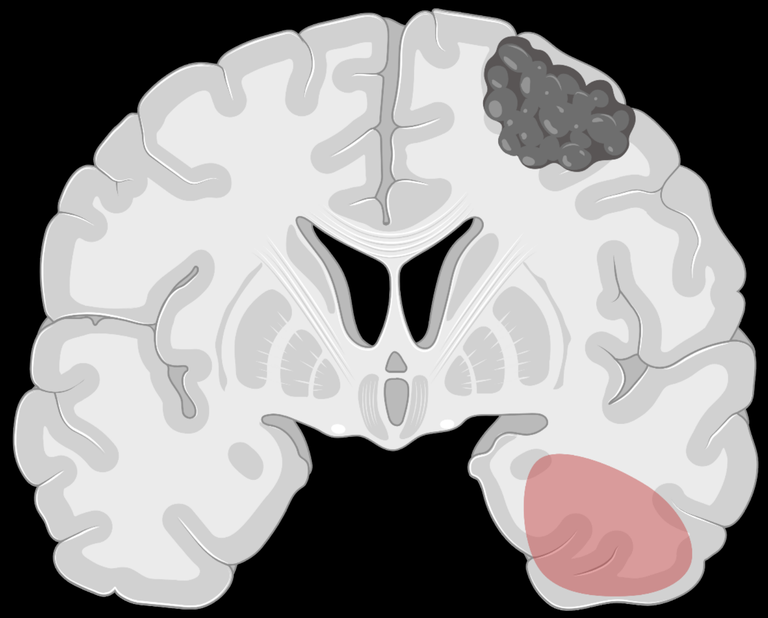
Tumor segmentation

Tumor segmentation

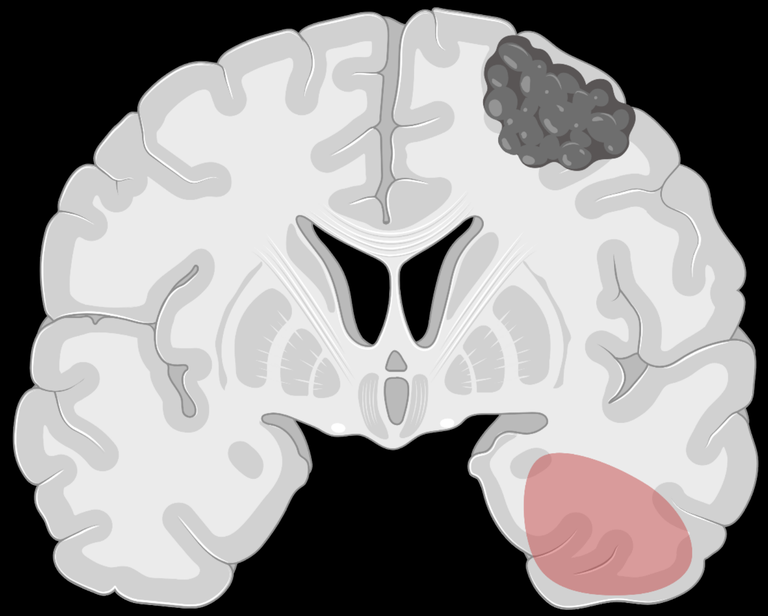


Diagnosis or prognosis

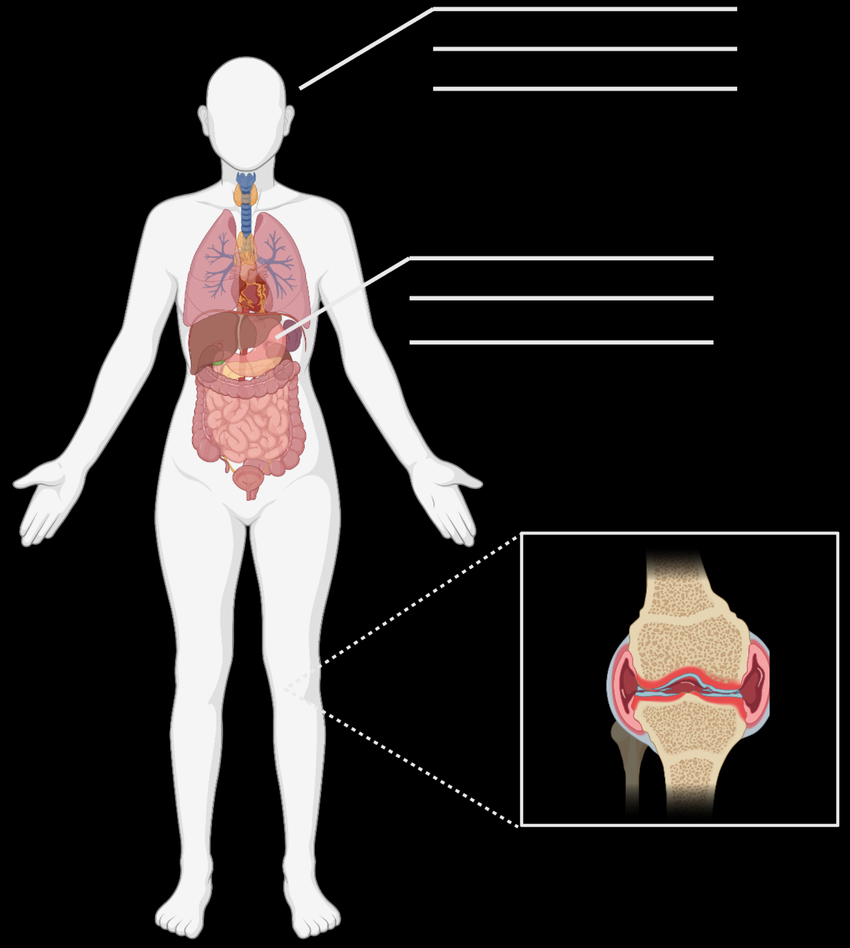
Does the model make easily verifiable predictions?



Tumor segmentation

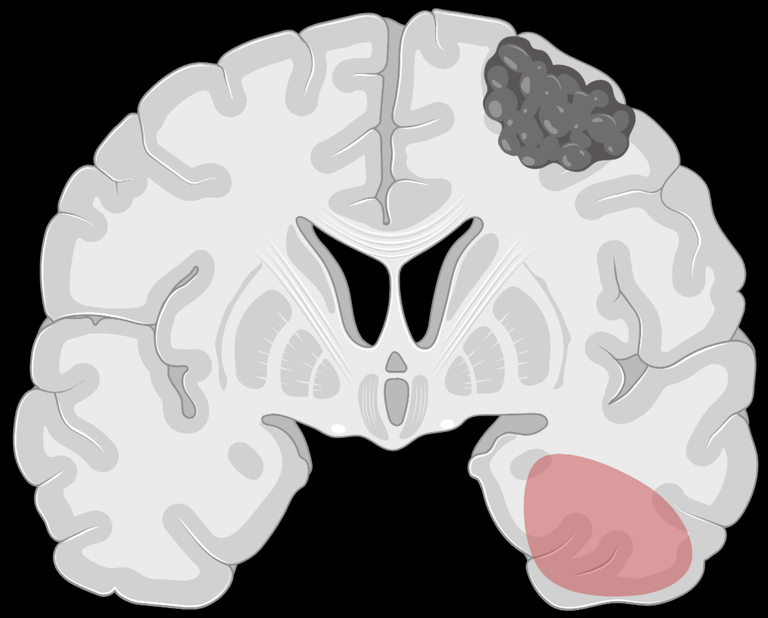


Tumor segmentation

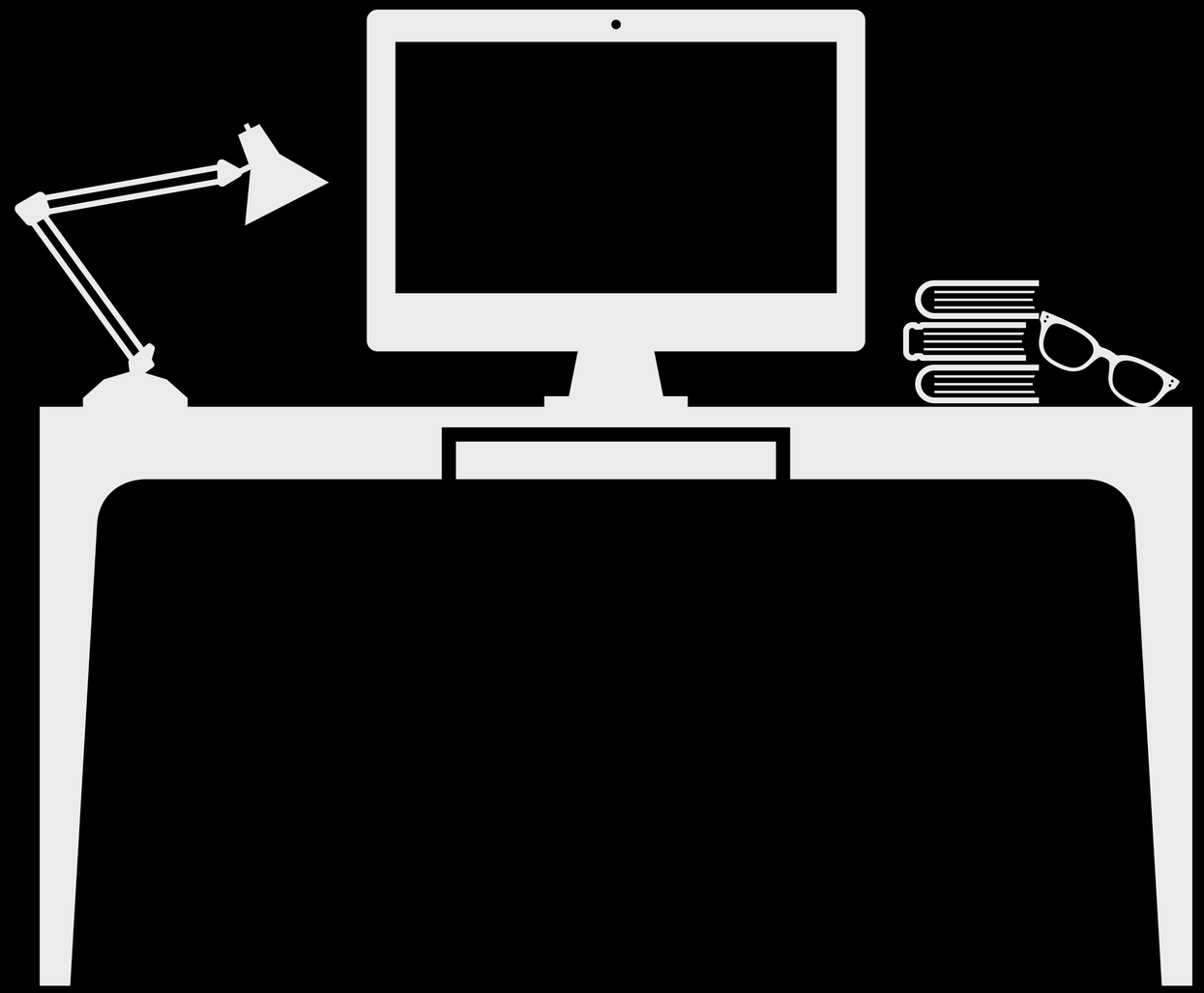


Diagnosis or prognosis

Why are we striving for model interpretability?



Tumor segmentation



“Definitions”

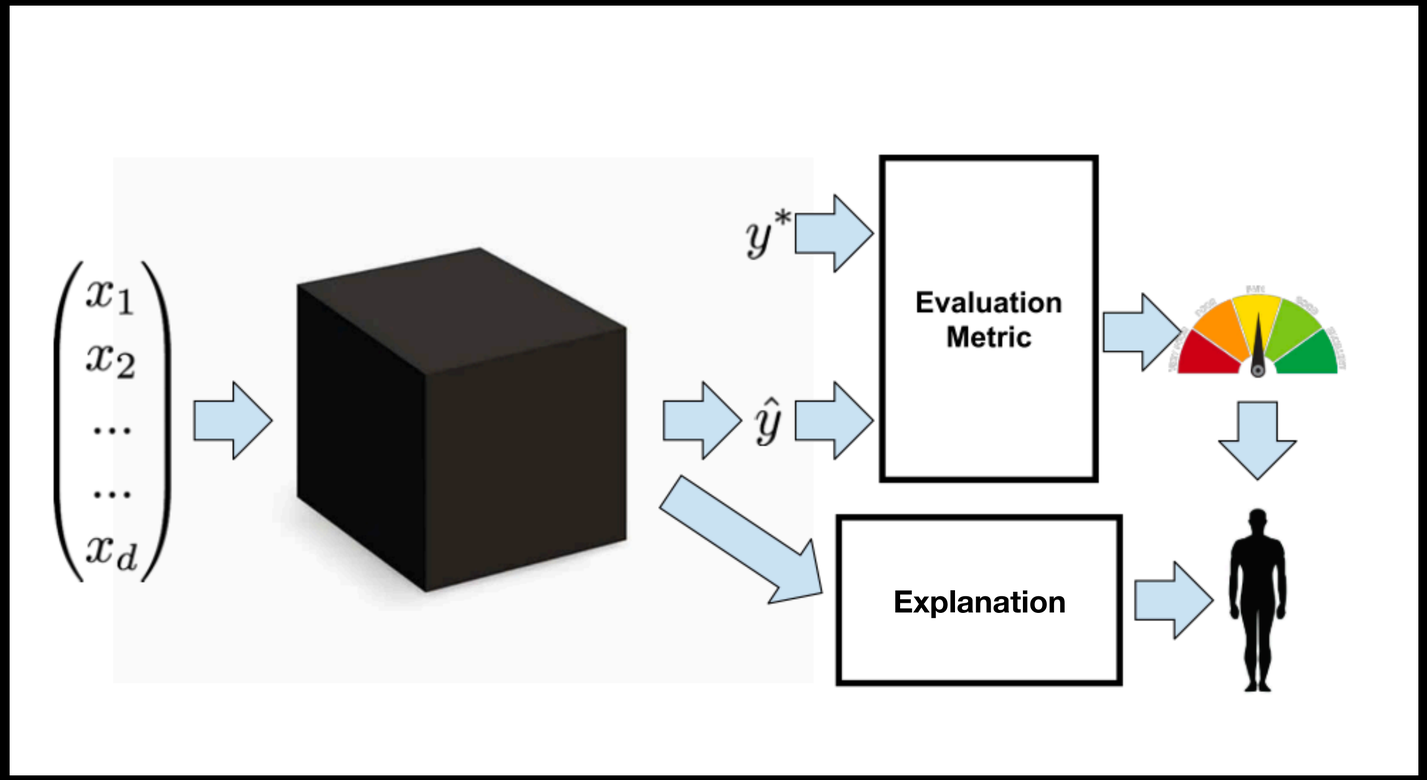
Interpretability is the degree to which a [human] observer can understand the cause of a decision.

Miller, *Artificial Intelligence*, 2019

Interpretability

$\neq$

Explainability



Adapted from figure from Lipton, *arXiv*, 2017

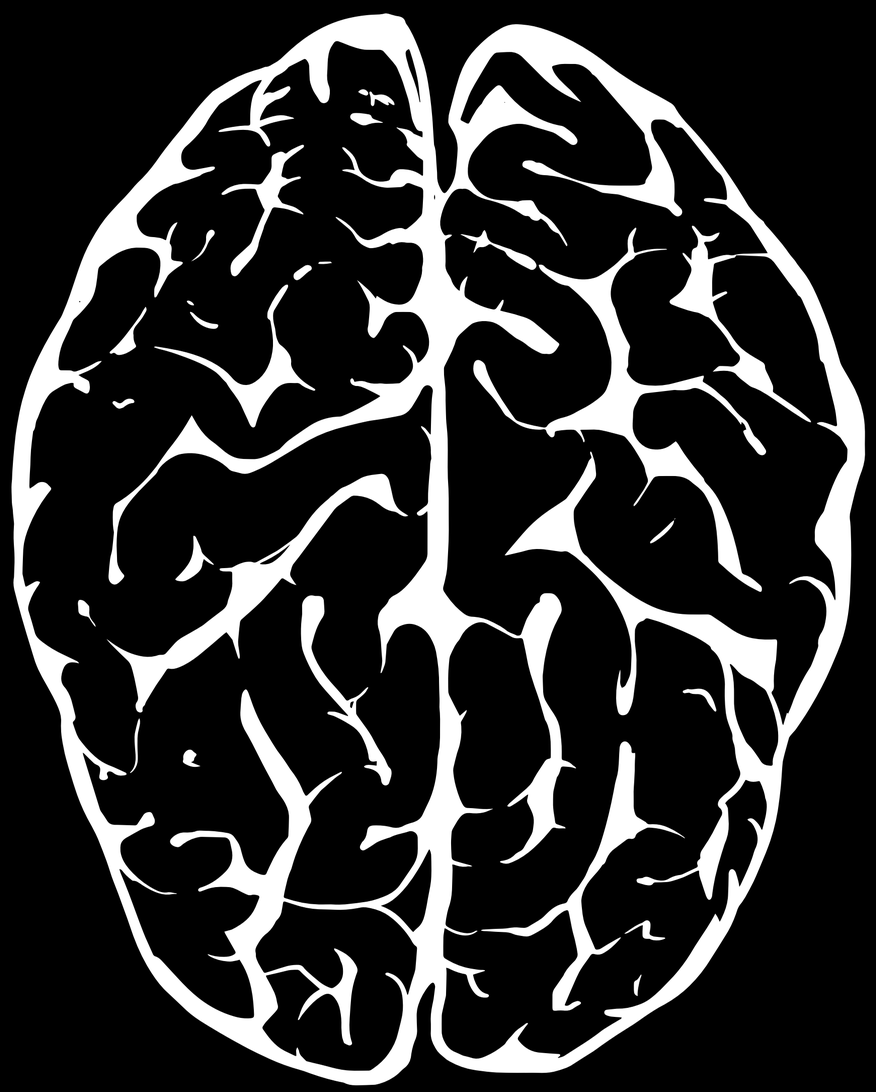
Figure from Rudin, *Nature Machine Intelligence*, 2019

The need for model interpretability

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

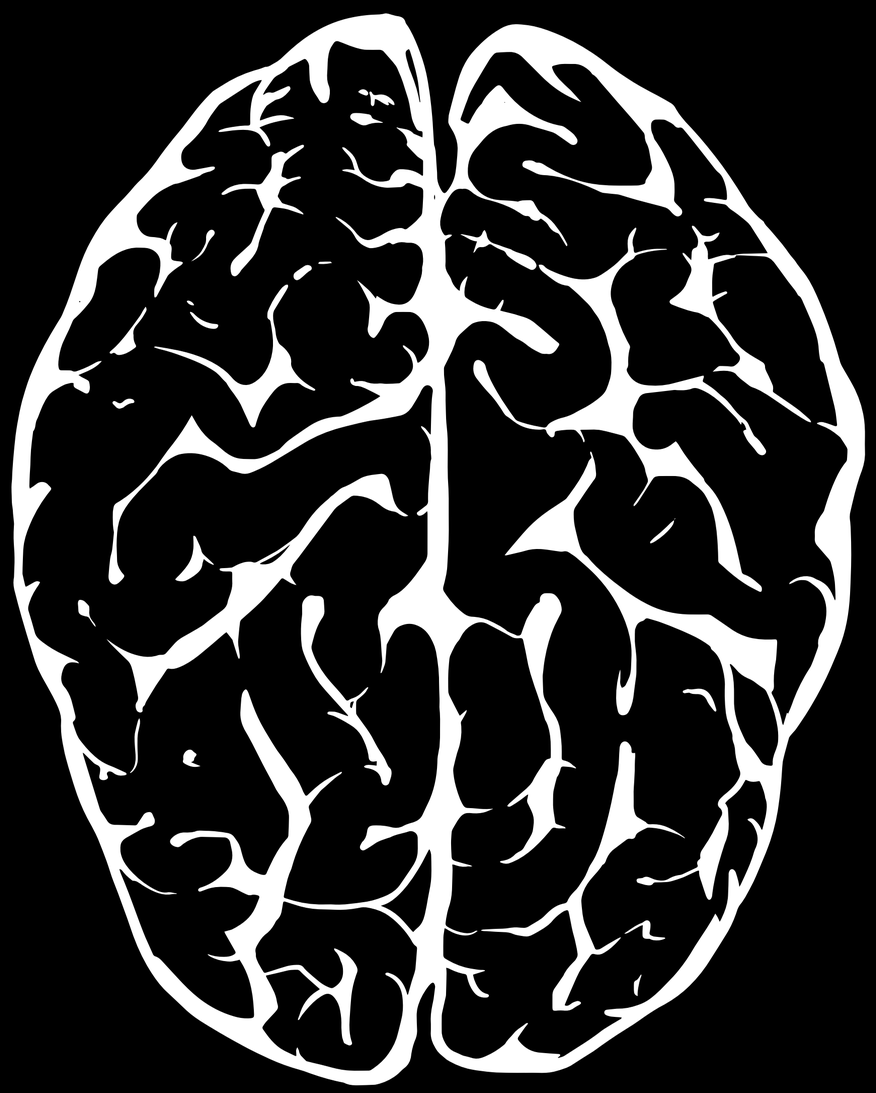
The need for model interpretability

- 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

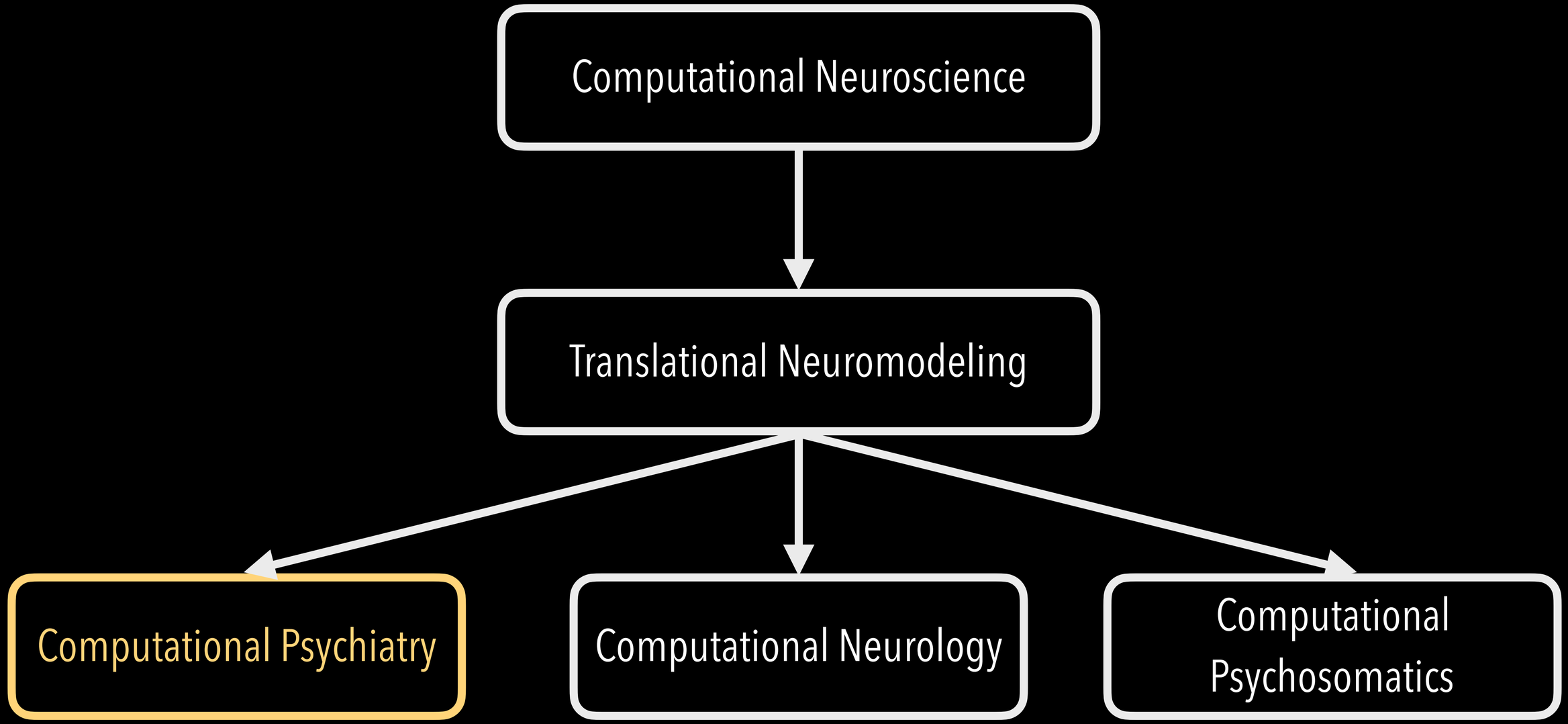
Psychiatry

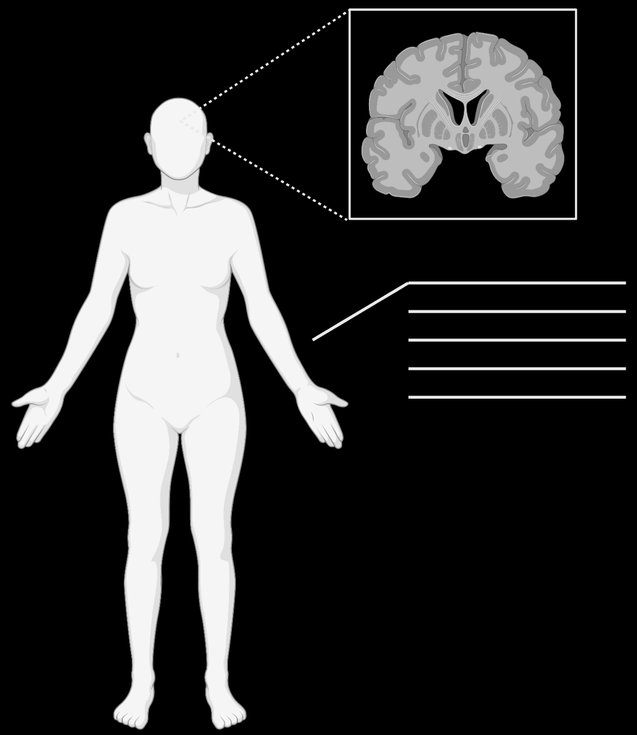
- Psychiatric disorders are broad and heterogeneous syndromes.
- Lack of well-defined neuropathology and *bona fide* biological markers
- Diagnosis: clinical observations using diagnostic manuals (ICD-10, DSM-5)

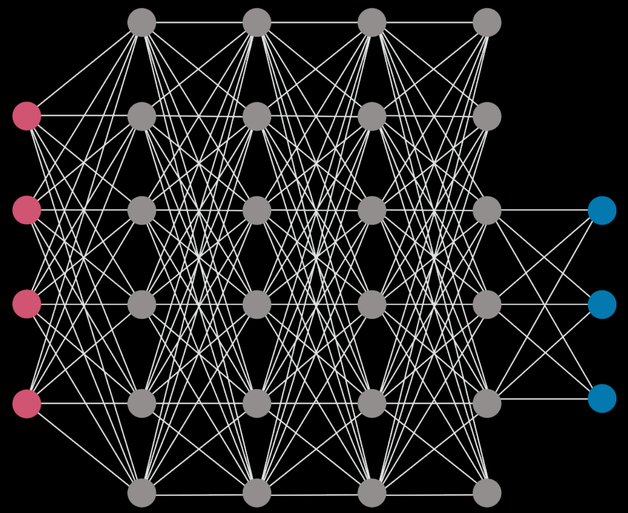
Messing *et al*., "Biology of Psychiatric Disorders" in
Harrison's Principles of Internal Medicine, 20th, 2018

Computational Psychiatry

Computational Psychiatry ?


Computational Neuroscience
Translational Neuromodeling
Computational Psychiatry
Computational Neurology
Computational Psychosomatics





Output

Generative models

Generative models

… based on
biophysical principles

Machine Learning recap

- Discriminative model: $P(Y \mid X = x)$
- 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
SCIENCE DIRECT®

NeuroImage

NeuroImage 19 (2003) 1273–1302

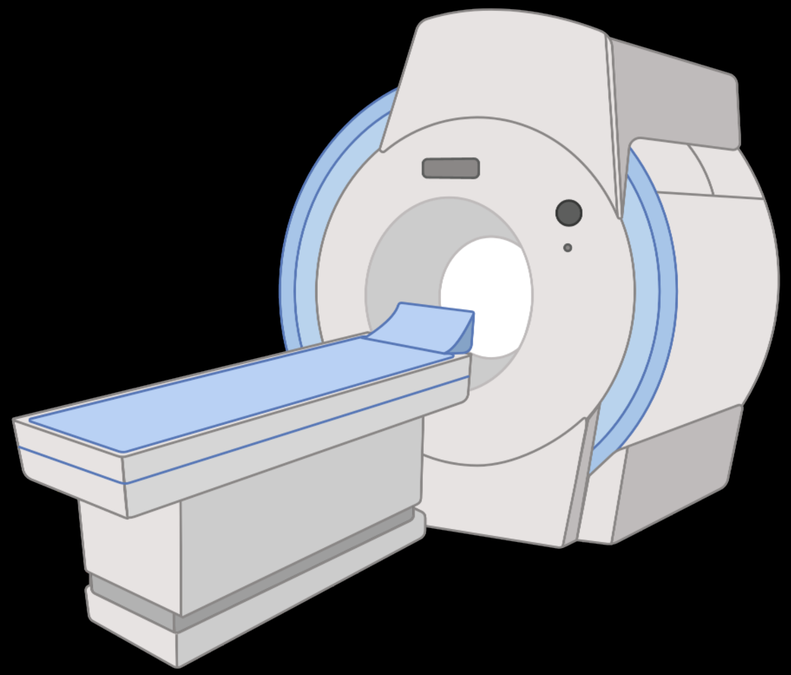
www.elsevier.com/locate/ynimg

Dynamic causal modelling

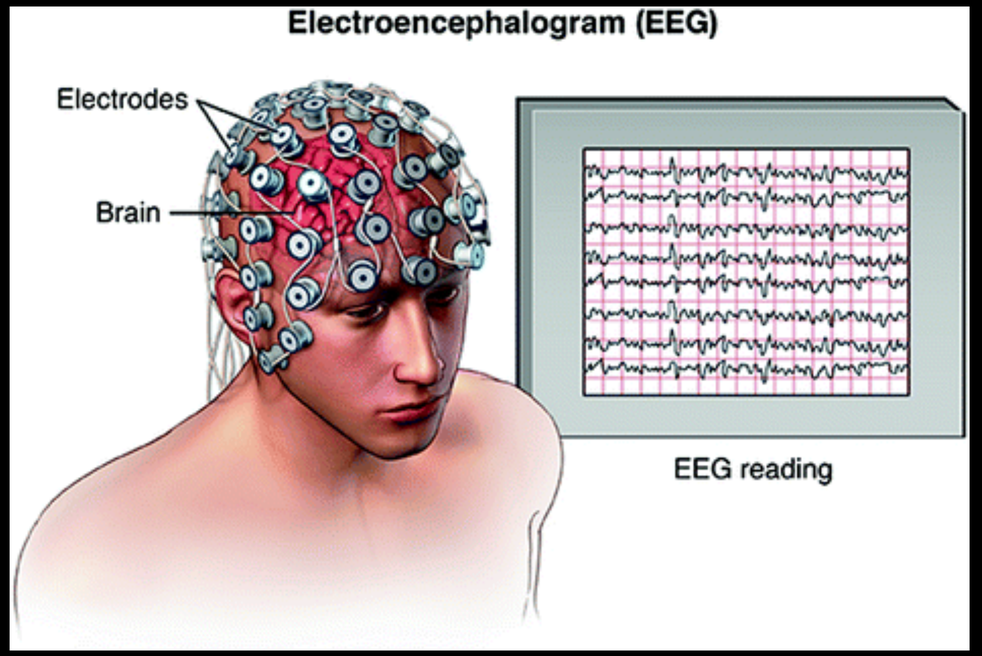
K.J. Friston,* L. Harrison, and W. Penny

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)

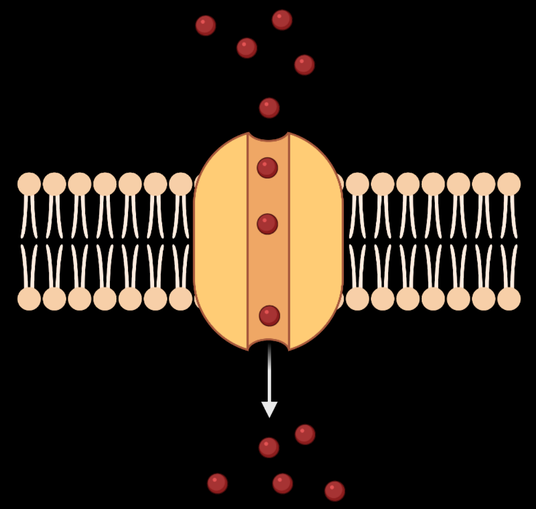


Electroencephalography (EEG)

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

Generative model (DCM)

Observable measurements (e.g., fMRI, EEG)



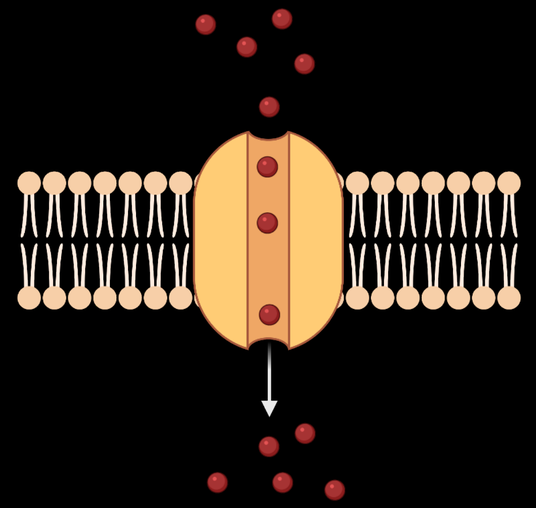
Hidden states

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

Generative model (DCM)

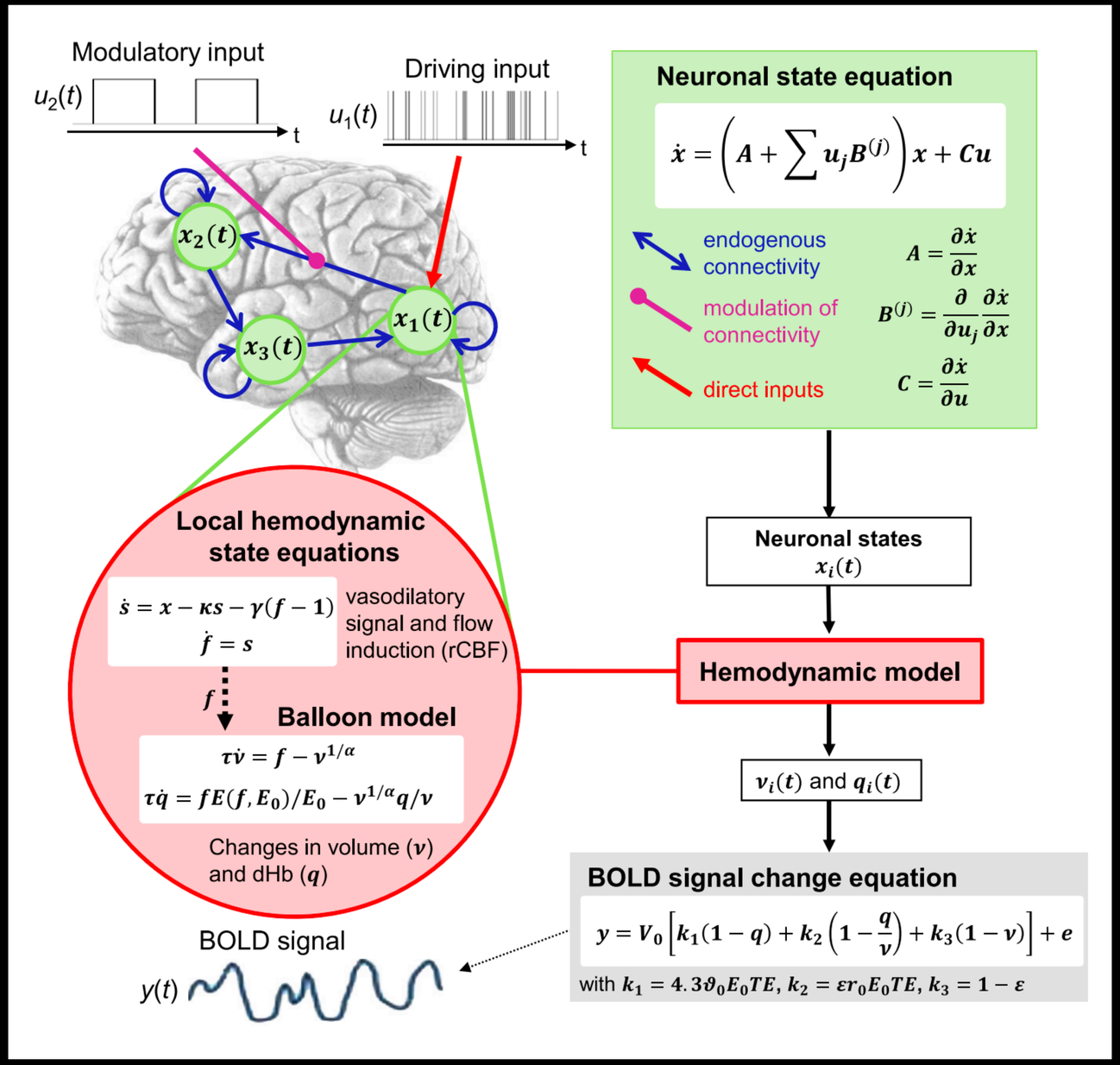
Observable measurements (e.g., fMRI, EEG)

Forward model



Hidden states

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

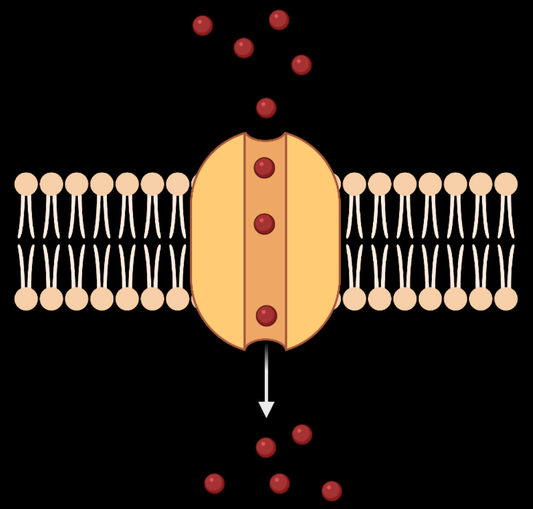


Stephan *et al.*, *Neuron*, 2015

Generative model (DCM)

Observable measurements (e.g., fMRI, EEG)

Forward model



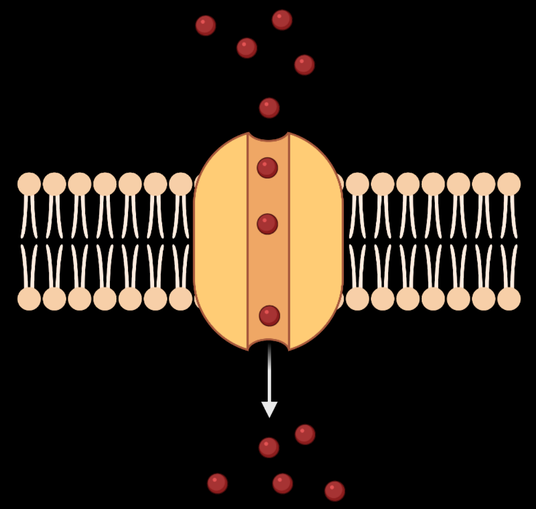
Hidden states

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

Generative model (DCM)

Forward model

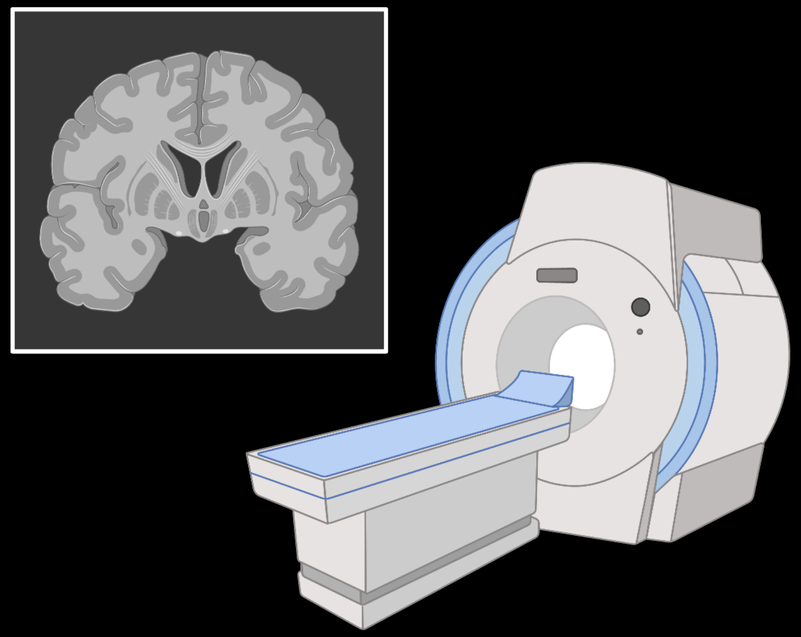
Model inversion



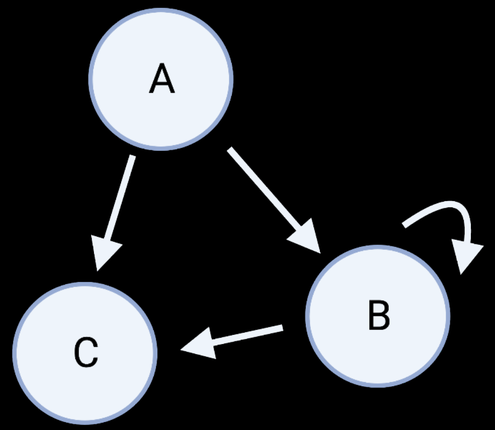
Observable measurements (e.g., fMRI, EEG)

Hidden states

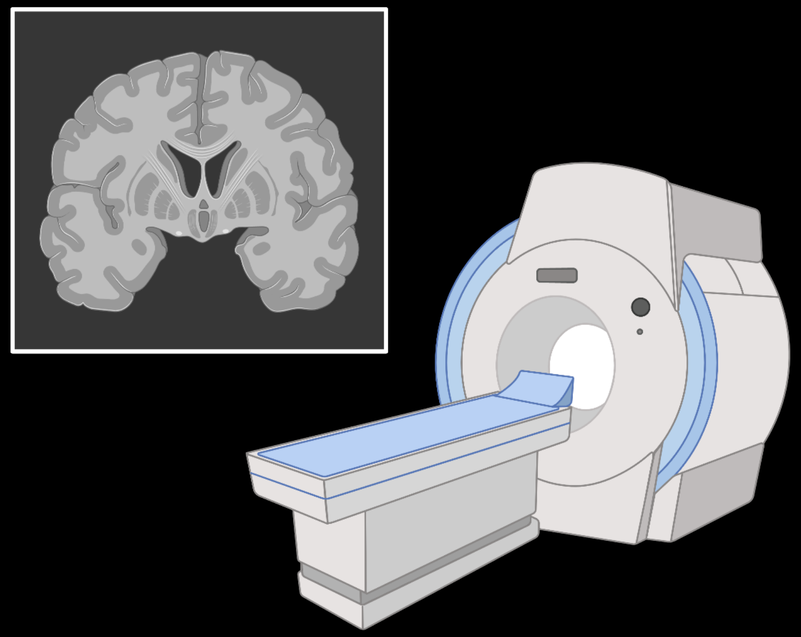
Adapted from Stephan *et al.*, *Front. Hum. Neurosci.*, 2016



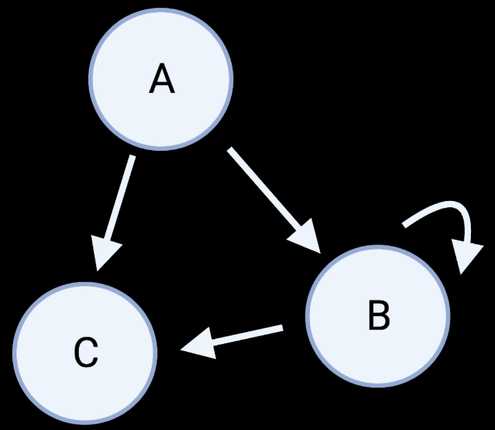
1. Collect subject data



2. Model inversion to obtain subject-specific parameter estimates

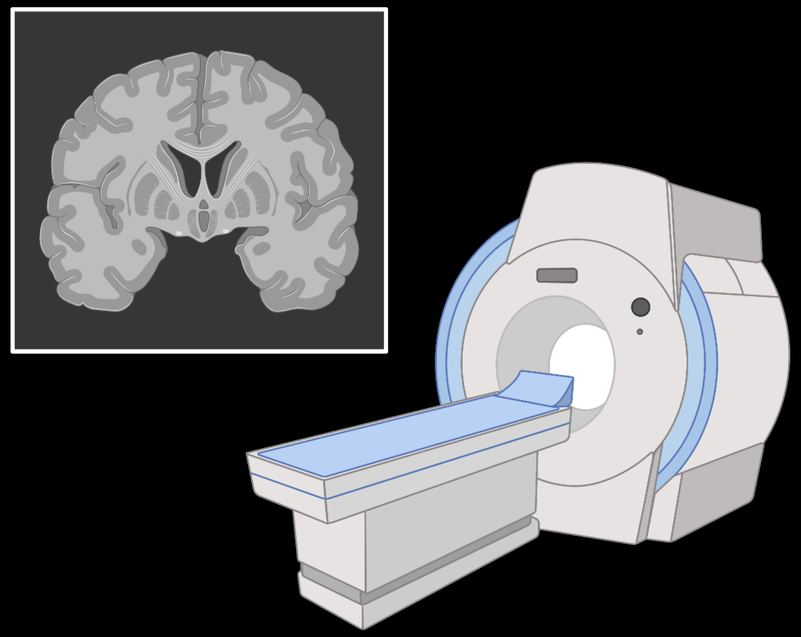


1. Collect subject data

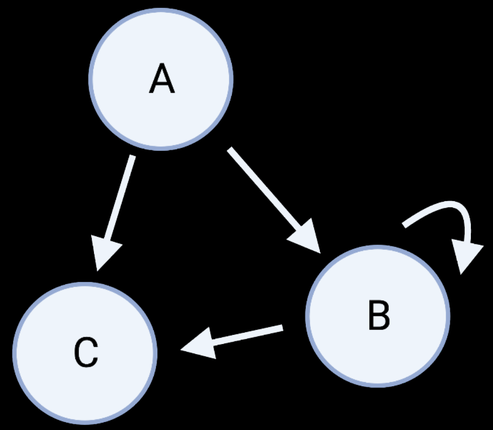


2. Model inversion to obtain subject-specific parameter estimates

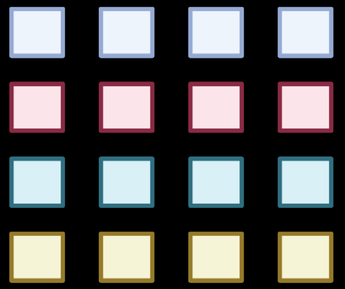
3. Use parameter estimates as new features (embedding)



1. Collect subject data

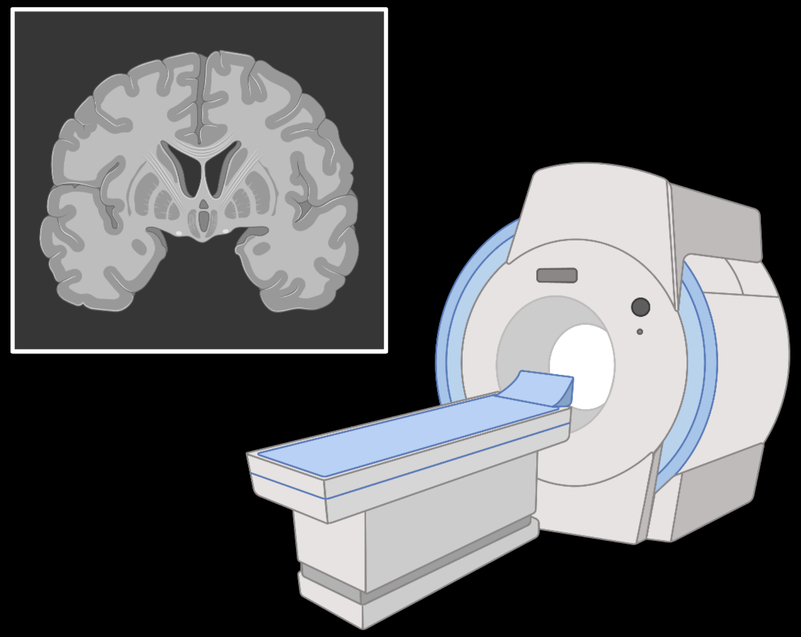


2. Model inversion to obtain subject-specific parameter estimates

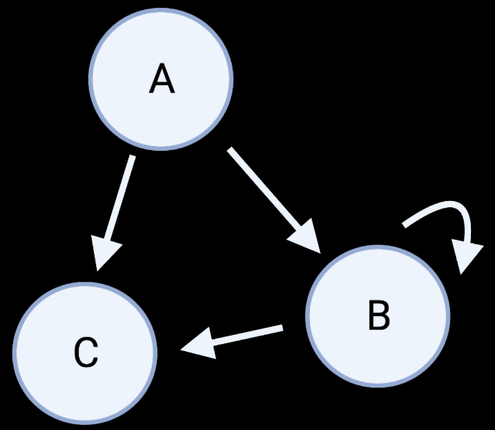


3. Use parameter estimates as new features (embedding)

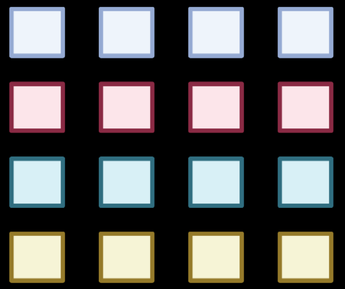
Generative embedding



1. Collect subject data



2. Model inversion to obtain subject-specific parameter estimates



3. Use parameter estimates as new features (embedding)

NeuroImage: Clinical 26 (2020) 102213

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

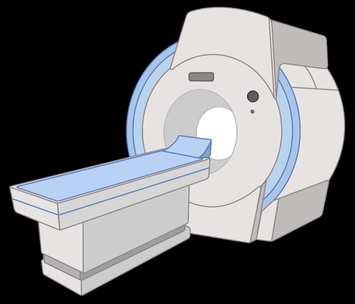
Predicting individual clinical trajectories of depression with generative embedding

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

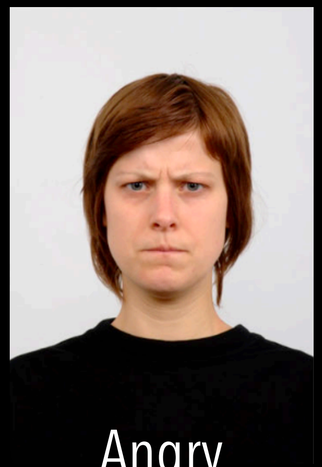
- fMRI data from 85 participants enrolled in the NEtherlands Study of Depression and Anxiety (NESDA)
- DSM-IV diagnosis of major depressive disorder
- Baseline assessment and follow up of depressive symptoms over 2 years

Event-related emotional face-perception paradigm

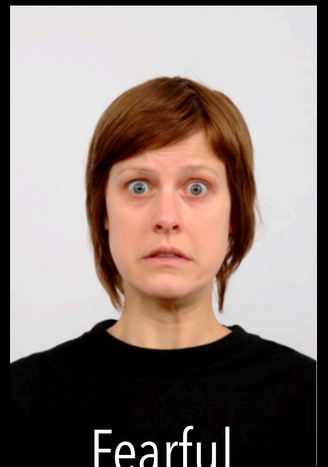


Photographs from Verpaalen *et al.*, *Cognition and Emotion*, 2019

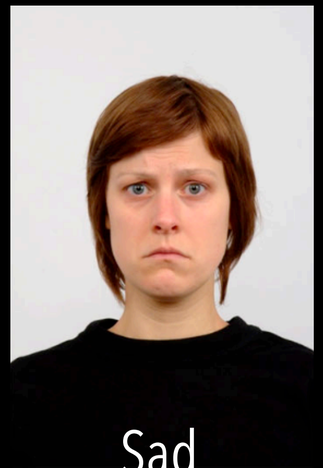
Event-related emotional face-perception paradigm



Angry



Fearful

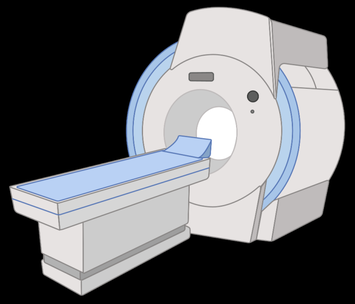


Sad

Happy

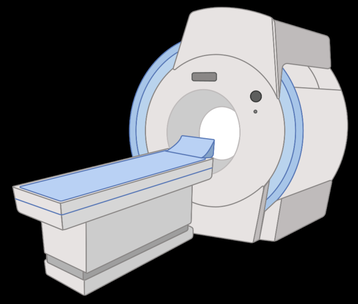
Neutral

Scrambled faces



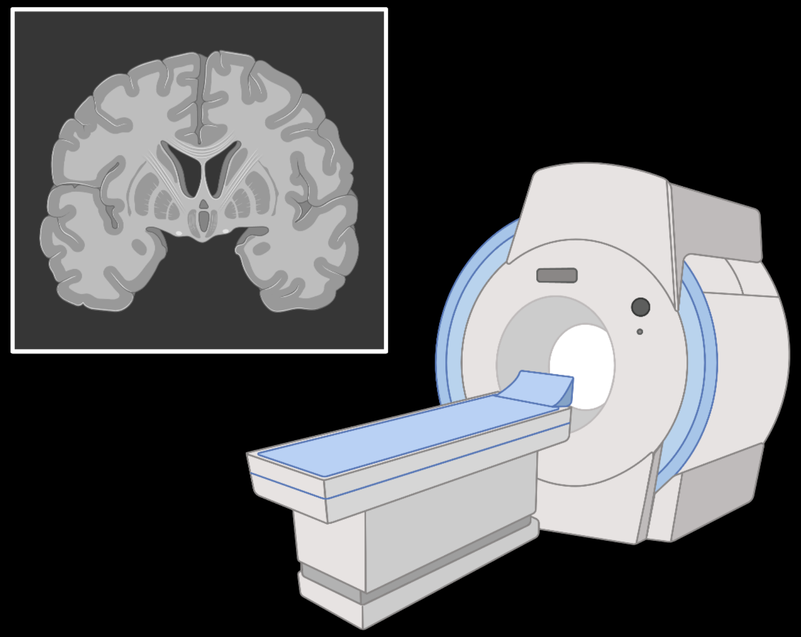
Photographs from Verpaalen *et al.*, *Cognition and Emotion*, 2019

Event-related emotional face-perception paradigm



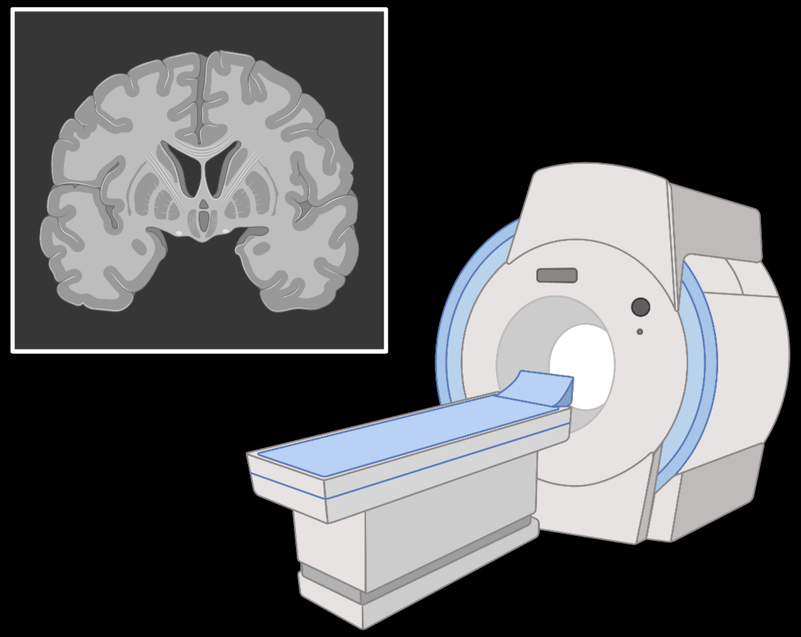
Photographs from Verpaalen *et al.*, *Cognition and Emotion*, 2019

Generative embedding

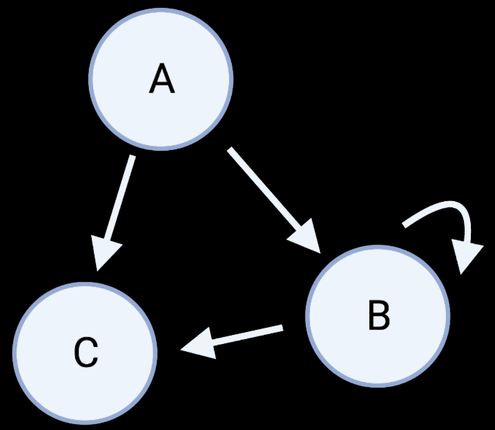


1. Collect subject data

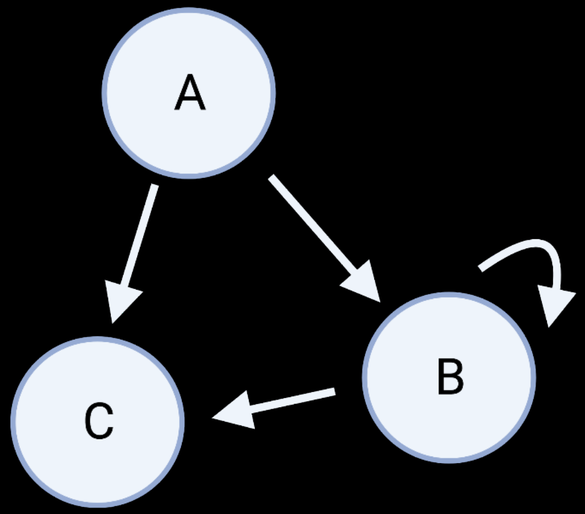
Generative embedding

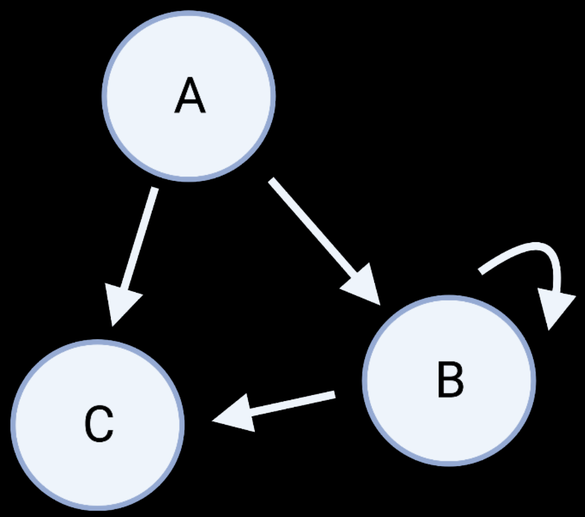


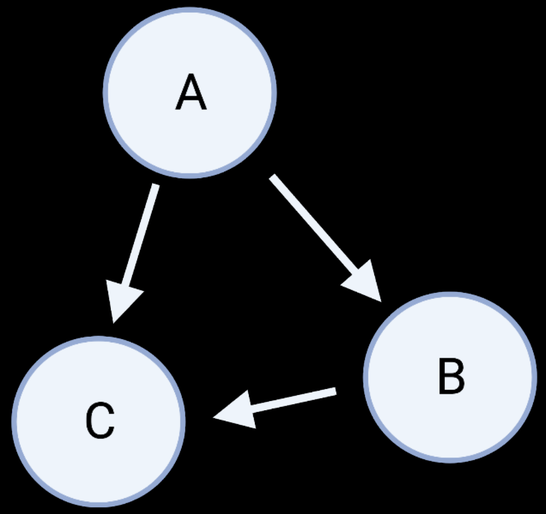
1. Collect subject data

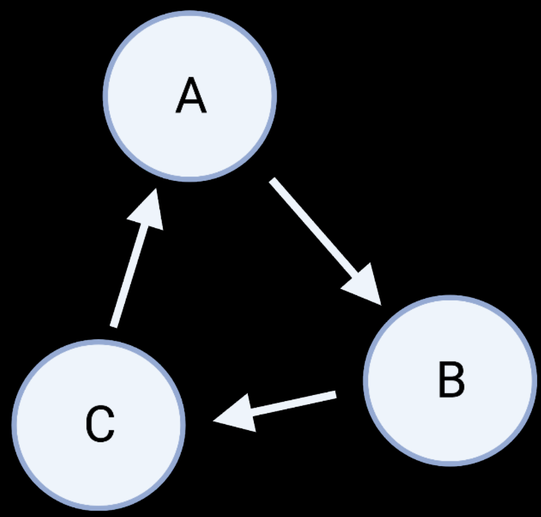


2. Model inversion to obtain subject-specific parameter estimates

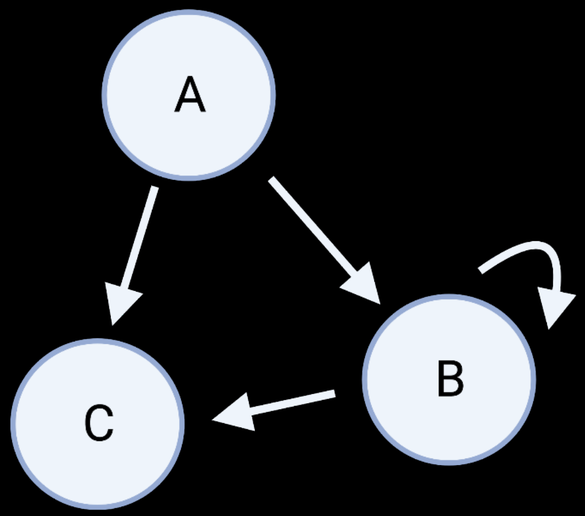

A
B
C

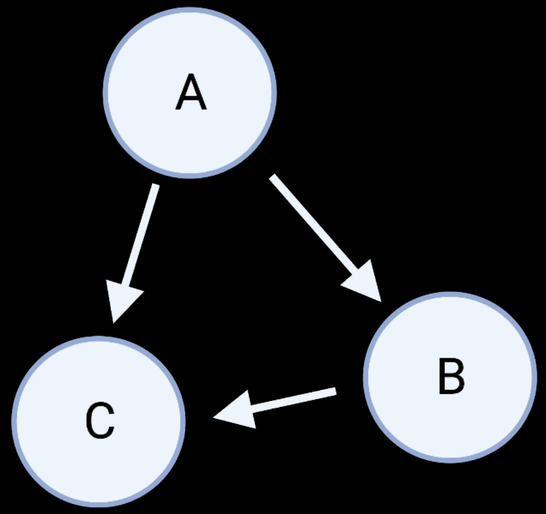

A
B
C

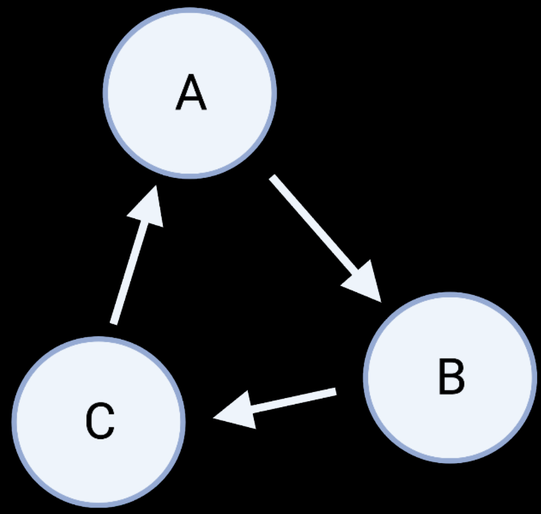

A
B
C

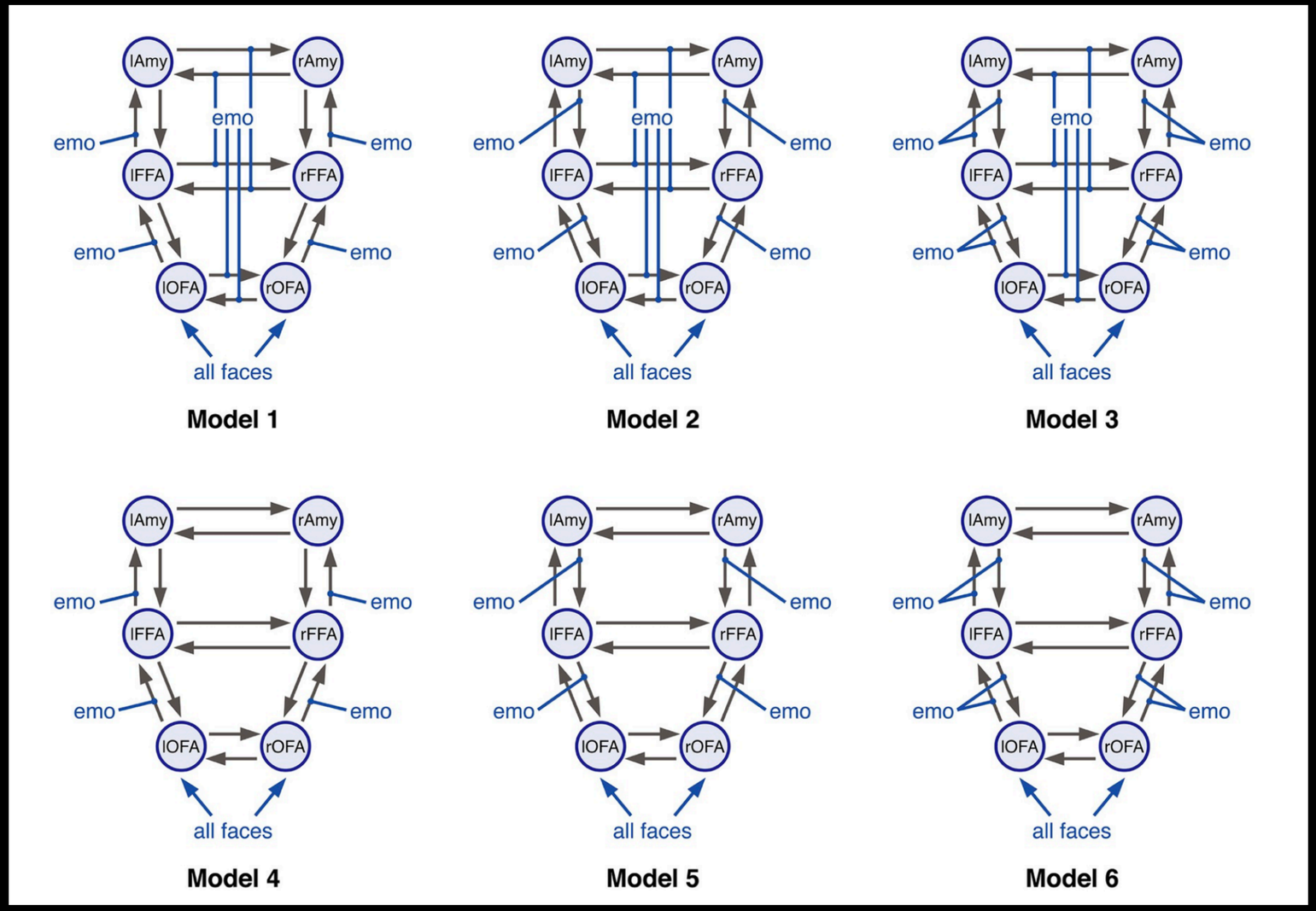

A
B
C

Model space

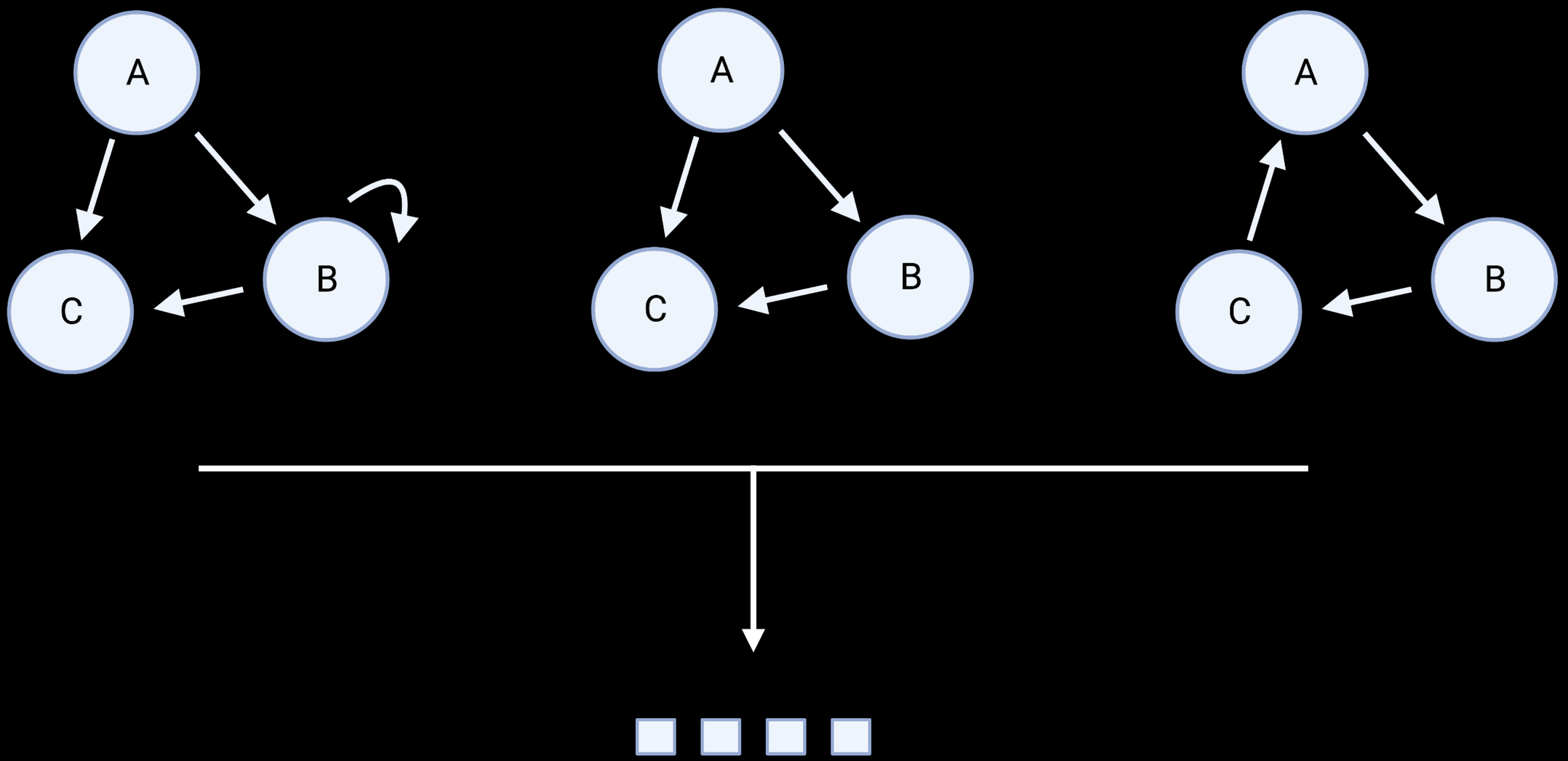


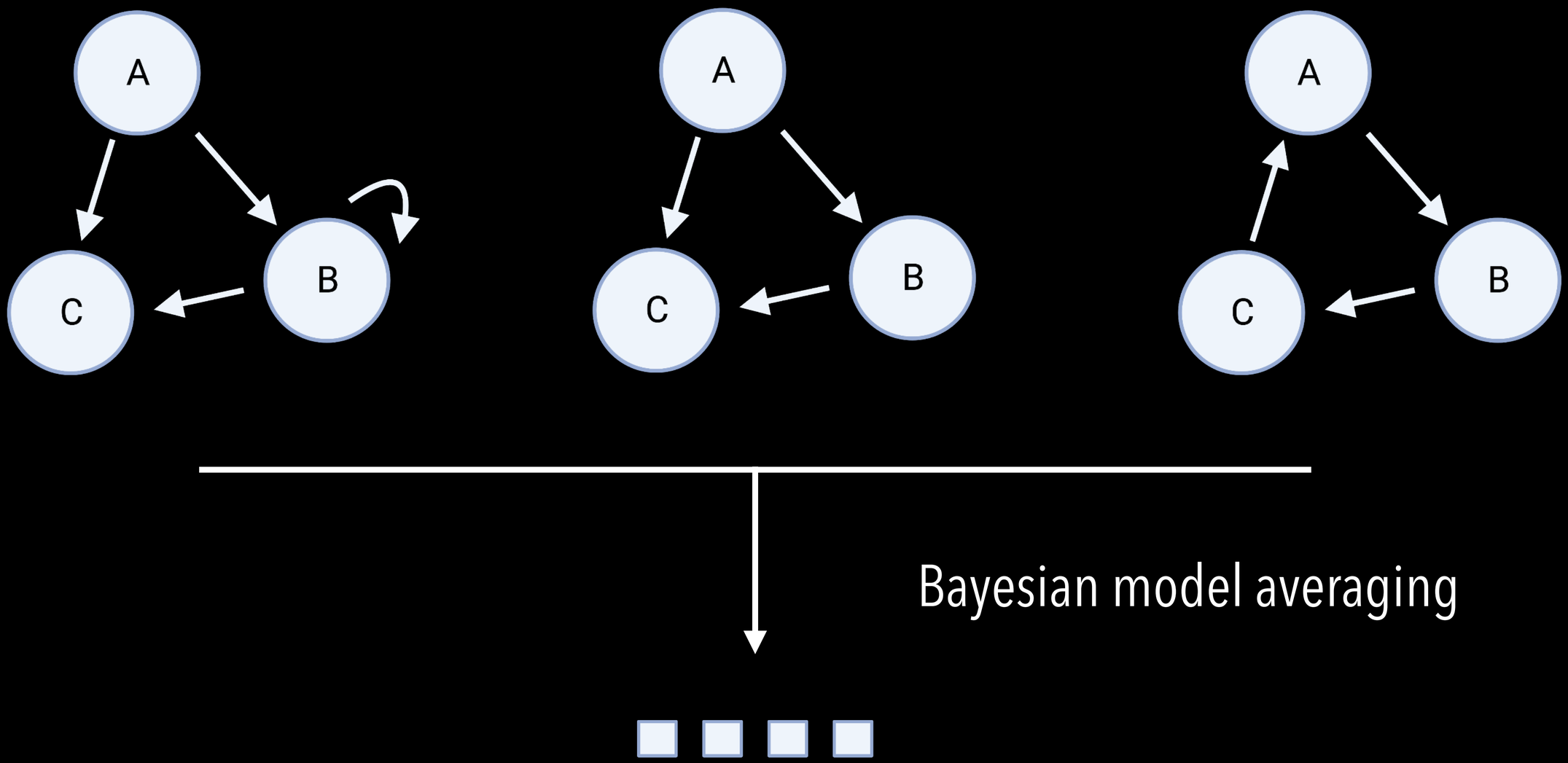






From Frässle *et al.*, *NeuroImage: Clinical*, 2020


A
B
C
A
B
C
A
B
C


A
B
C
A
B
C
A
B
C
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 (n = 15) vs. REM (n = 39)	CHR (n = 15) vs. IMP (n = 31)	IMP (n = 31) vs. REM (n = 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

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

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 (n = 15) vs. REM (n = 39)	CHR (n = 15) vs. IMP (n = 31)	IMP (n = 31) vs. REM (n = 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

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

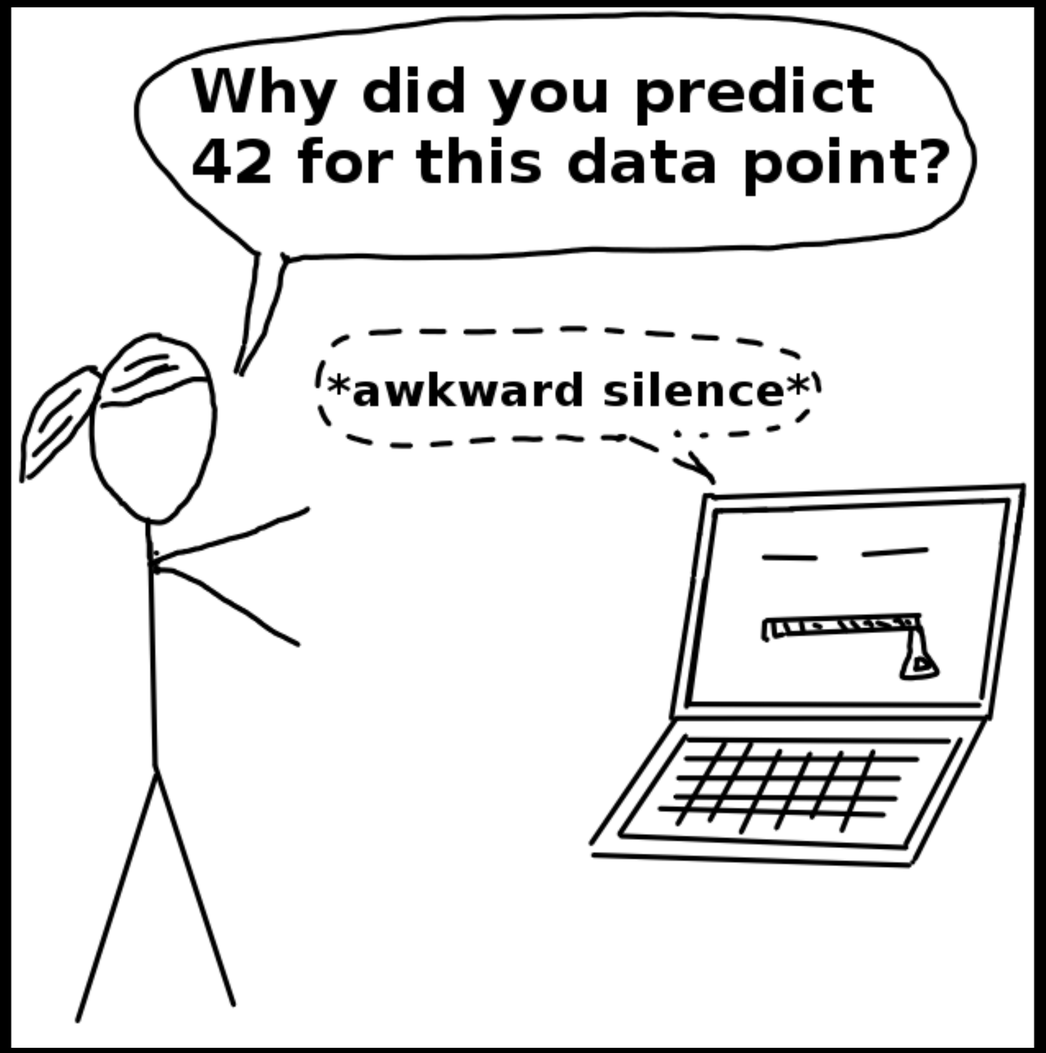
Photo by Nicola Anderson on Unsplash

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!

@nespereira_



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

Acknowledgements

- Thank you Stefan Frässle for all your input and discussions!
- Figures created with BioRender.com