UCL

From the Higgs to Huntington’s: methods for learning from data

UCL HEP seminar
24-05-19

Peter Wijeratne
MRC Skills Development Fellow

MRC London Institute of Medical Sciences

HUNTINGTON'S DISEASE CENTRE

cmic Centre for Medical Image Computing

Acknowledgements

UCL CMIC

Neil Oxtoby
Alexandra Young
Arman Eshaghi
Leon Aksman
Maura Bellio
Nonie Alexander

UCL HDC

Sarah Tabrizi
Rachael Scahill
Sarah Gregory
Eileanoir Johnson
Ed Wild
Lauren Byrne

CHDI

Cristina Sampaio
Amrita Mohan
John Warner
Dorian Pustina
Alexandra Shechtel

And all the participants of the PREDICT, TRACK and IMAGE-HD studies.

MRC London Institute of Medical Sciences

CHDI FOUNDATION

EPSRC Engineering and Physical Sciences Research Council

This talk: high level

Interested in extracting hidden information from observed data

→Bayesian methods

Two main schools of thought

Hypothesis-driven (informative priors)

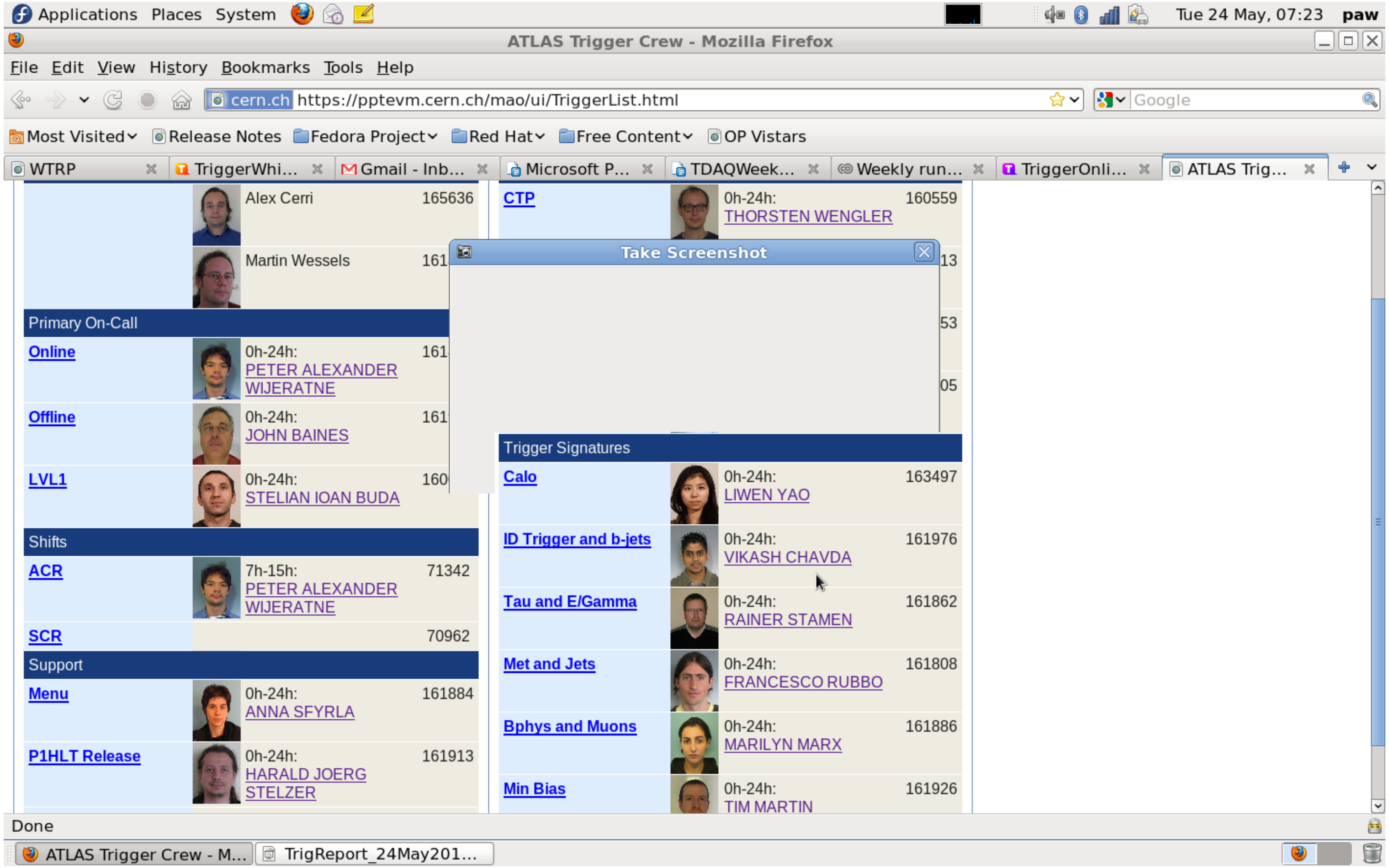
Unfolding / inverse problems – e.g. image reconstruction

Data-driven (non-informative priors)

Latent variable inference – e.g. disease progression modelling

Physics favours the former, biology the latter

My PhD: LHC Run 1 with ATLAS

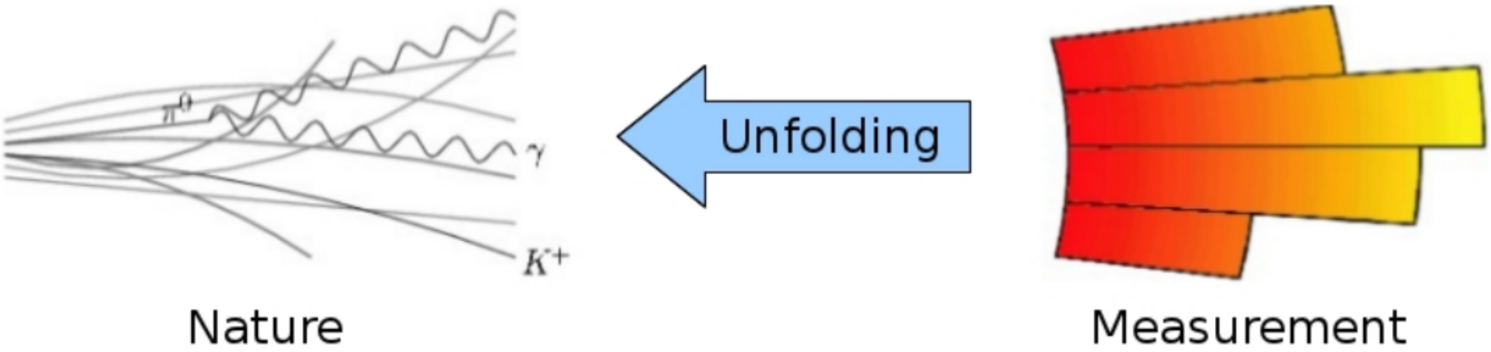


	Alex Cerri	165636
	Martin Wessels	161
Primary On-Call		
Online	0h-24h: PETER ALEXANDER WIJERATNE	161
Offline	0h-24h: JOHN BAINES	161
LVL1	0h-24h: STELIAN IOAN BUDA	160
Shifts		
ACR	7h-15h: PETER ALEXANDER WIJERATNE	71342
SCR		70962
Support		
Menu	0h-24h: ANNA SFYRLA	161884
P1HLT Release	0h-24h: HARALD JOERG STELZER	161913

CTP	0h-24h: THORSTEN WENGLER	160559
		13
		53
		05
Trigger Signatures		
Calo	0h-24h: LIWEN YAO	163497
ID Trigger and b-jets	0h-24h: VIKASH CHAVDA	161976
Tau and E/Gamma	0h-24h: RAINER STAMEN	161862
Met and Jets	0h-24h: FRANCESCO RUBBO	161808
Bphys and Muons	0h-24h: MARILYN MARX	161886
Min Bias	0h-24h: TIM MARTIN	161926

For some reason, they let me near the detector

My PhD: unfolding / inverse problems



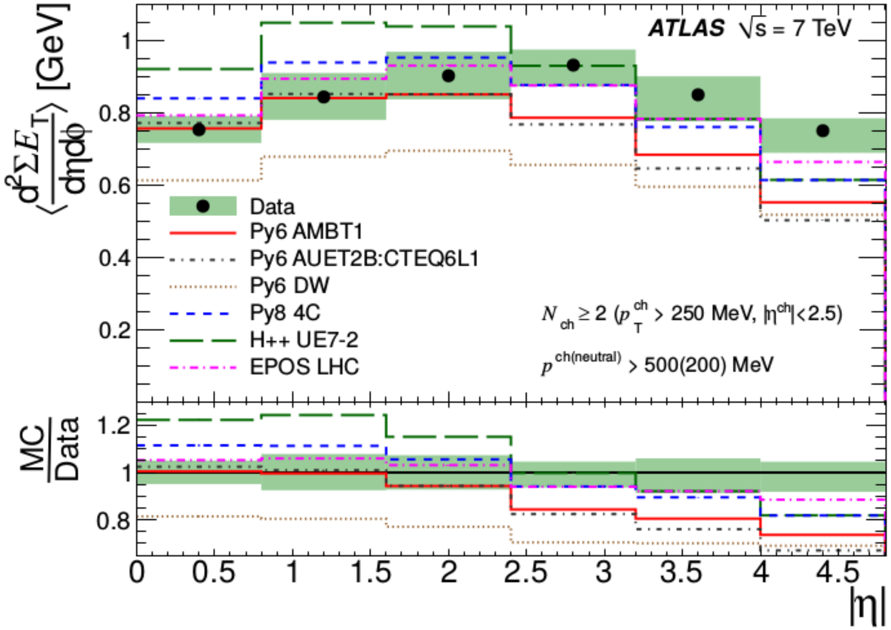
$$n(C_i^{data}) = \frac{1}{\epsilon_i}\sum_j P(T_i^{MC} \mid R_j^{MC})\, n(R_j^{data})$$

- Real data are dependent on the detector used to measure them

- Bring data back to their natural state by applying hypothesis-driven corrections derived from simulation

→”Unfolding the cause”

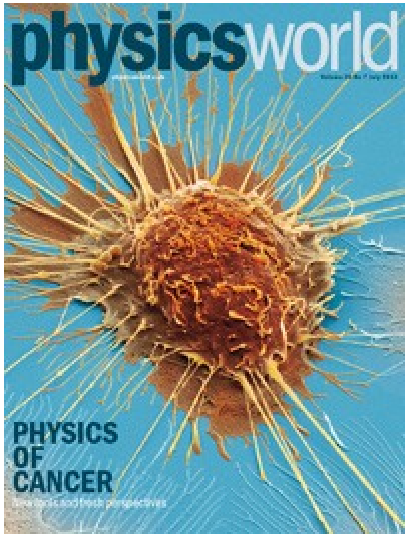
My PhD: unfolding / inverse problems



- Energy density (min bias + UE) was not modelled correctly in forward direction
    - Problem would only increase with luminosity
- We iteratively unfolded the data to compare directly with various models
- Tuned MC generators to data

PhD to postdoc

I saw this one day in 2013



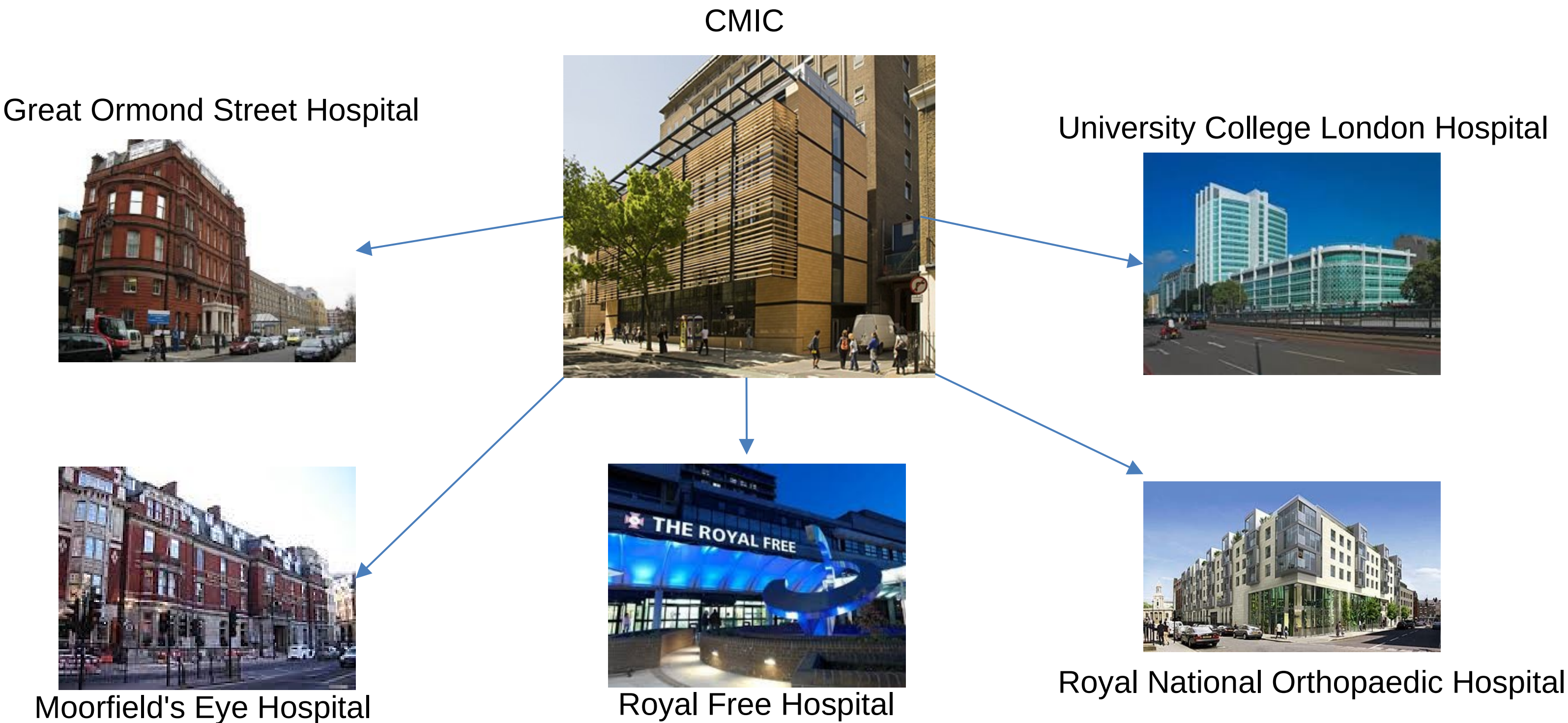
I wanted to use physics to fight cancer

I asked about for potential opportunities (thanks Simon)

I got lucky and a postdoc came up at the Centre for Medical Image Computing on jobs.ac.uk

Centre for Medical Image Computing (CMIC)

Maths, physics and engineering scientists at the interface of basic and biomedical sciences



CMIC

Great Ormond Street Hospital

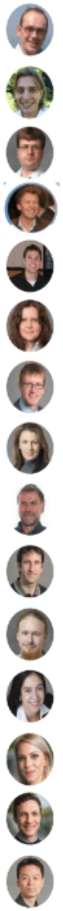
University College London Hospital

Moorfield's Eye Hospital

Royal Free Hospital

Royal National Orthopaedic Hospital

CMIC



Alexander

Altmann

Arridge

Barkhof

Barratt

Betcke

Clarkson

Drobnjak

Hawkes

Iglesias

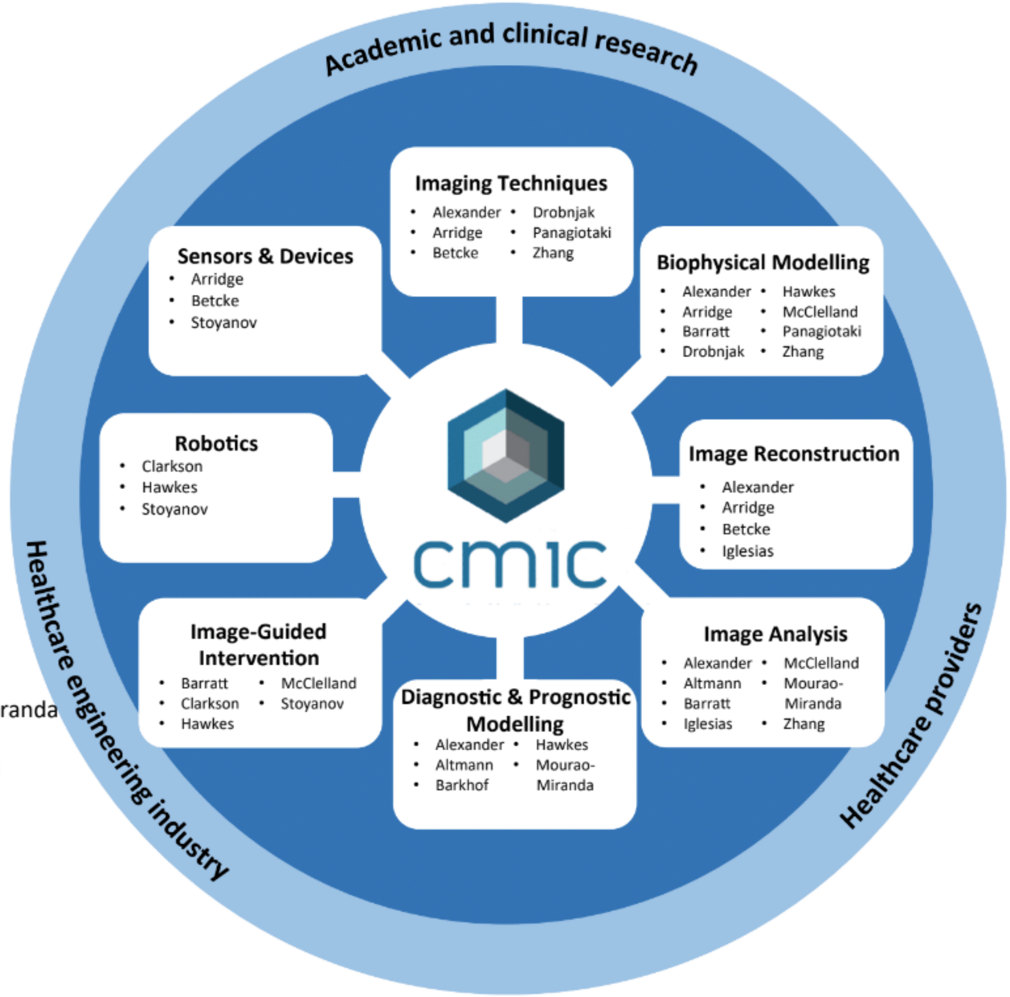
McClelland

Mourao-Miranda

Panagiotaki

Stoyanov

Zhang



Academic and clinical research

Healthcare engineering industry

Healthcare providers

cmic

Imaging Techniques

- Alexander
- Arridge
- Betcke
- Drobnjak
- Panagiotaki
- Zhang

Sensors & Devices

- Arridge
- Betcke
- Stoyanov

Biophysical Modelling

- Alexander
- Arridge
- Barratt
- Drobnjak
- Hawkes
- McClelland
- Panagiotaki
- Zhang

Robotics

- Clarkson
- Hawkes
- Stoyanov

Image Reconstruction

- Alexander
- Arridge
- Betcke
- Iglesias

Image-Guided Intervention

- Barratt
- Clarkson
- Hawkes
- McClelland
- Stoyanov

Diagnostic & Prognostic Modelling

- Alexander
- Altmann
- Barkhof
- Hawkes
- Mourao-Miranda

Image Analysis

- Alexander
- Altmann
- Barratt
- Iglesias
- McClelland
- Mourao-Miranda
- Zhang

CMIC capability-application

	Neurology and Psychiatry	Oncology	Surgery	Respiratory disease	Fetal, Neonatal, and Pediatrics	Cardiovascular disease
Imaging						
Image Analysis						
Machine Learning						
Non-imaging data science						
Robotics and vision						
Inverse Problems						
Computational Modeling						
Integrated Systems						

Slight (3 year) diversion: biophysical modelling of drug delivery

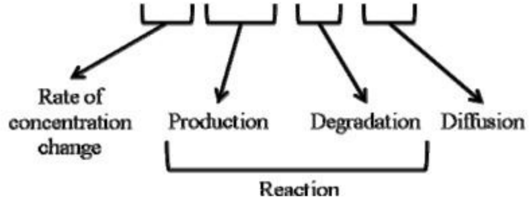
The Chemical Basis of Morphogenesis

A. M. Turing

Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, Vol. 237, No. 641. (Aug. 14, 1952), pp. 37-72.

$$\frac{\partial u}{\partial t} = F(u,v) - d_u v + D_u \Delta u$$

$$\frac{\partial v}{\partial t} = G(u,v) - d_v v + D_v \Delta v$$



Rate of concentration change

Production

Degradation

Diffusion

Reaction

Computational Modeling

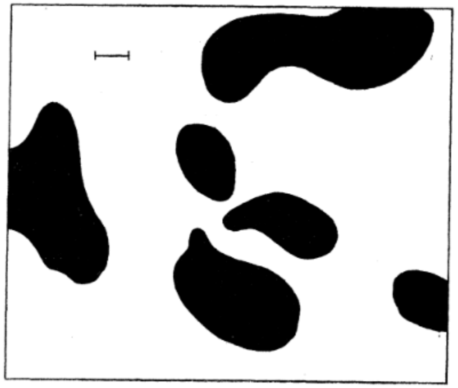
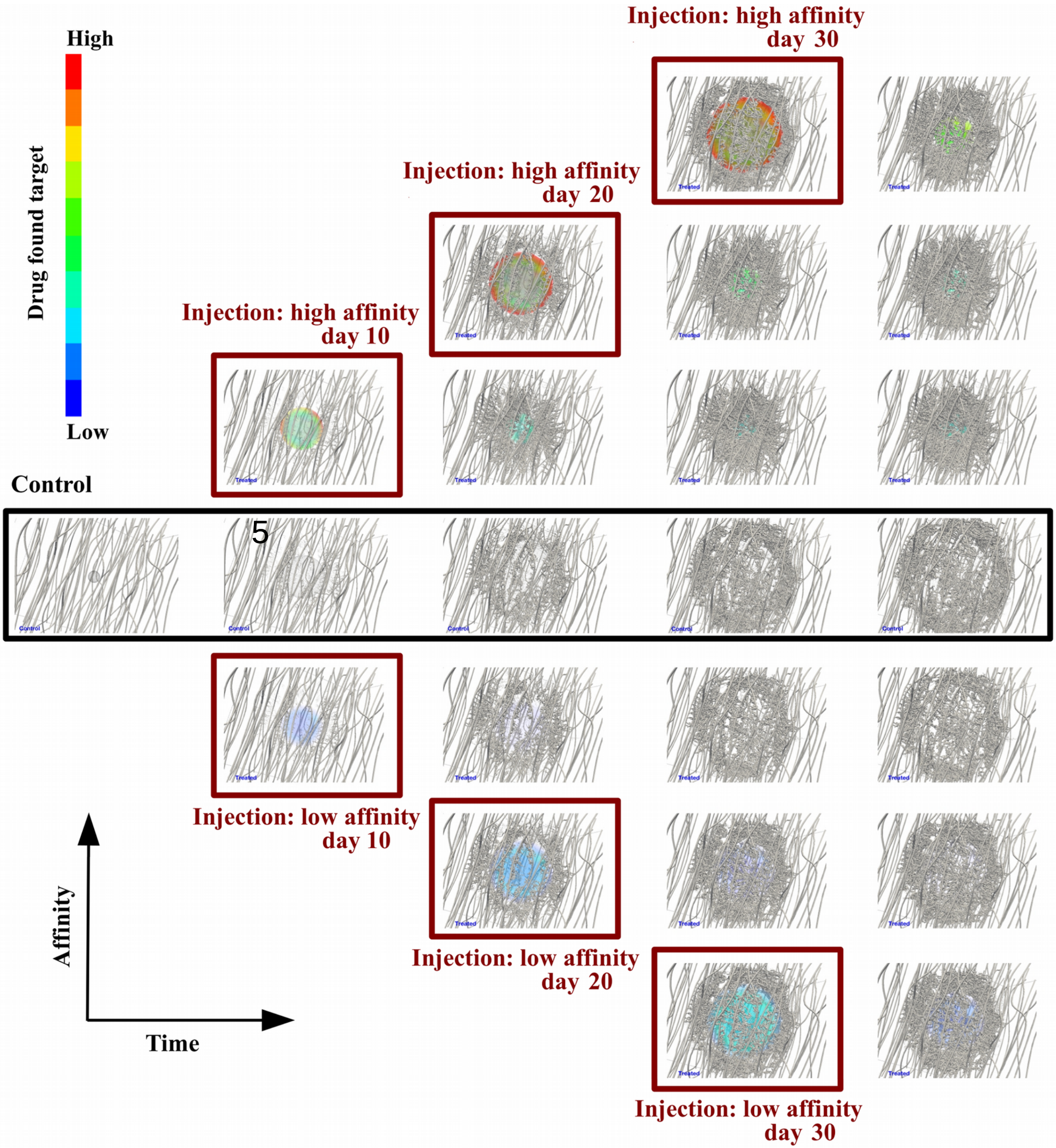


Figure 2. An example of a 'dappled' pattern as resulting from a type (*a*) morphogen system. A marker of unit length is shown. See text, §9, 11.

Slight (3 year) diversion: biophysical modelling of drug delivery


High
Drug found target
Low
Injection: high affinity day 30
Injection: high affinity day 20
Injection: high affinity day 10
Control
5
Injection: low affinity day 10
Injection: low affinity day 20
Injection: low affinity day 30
Affinity
Time

Computational Modeling

Vavourakis, Stylianopoulos, Wijeratne (2018) PLOS Comp Bio

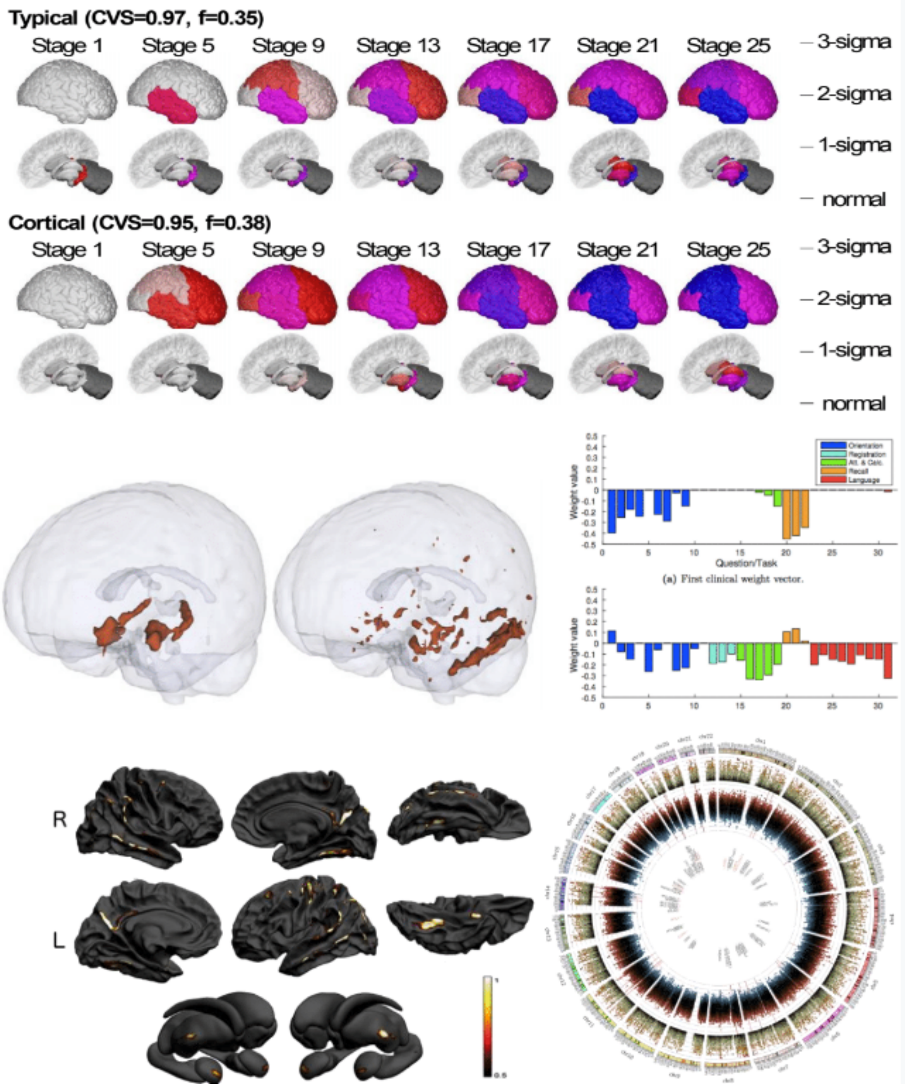
Current work: computational neurology

Image Analysis

Machine Learning

Non-imaging data science

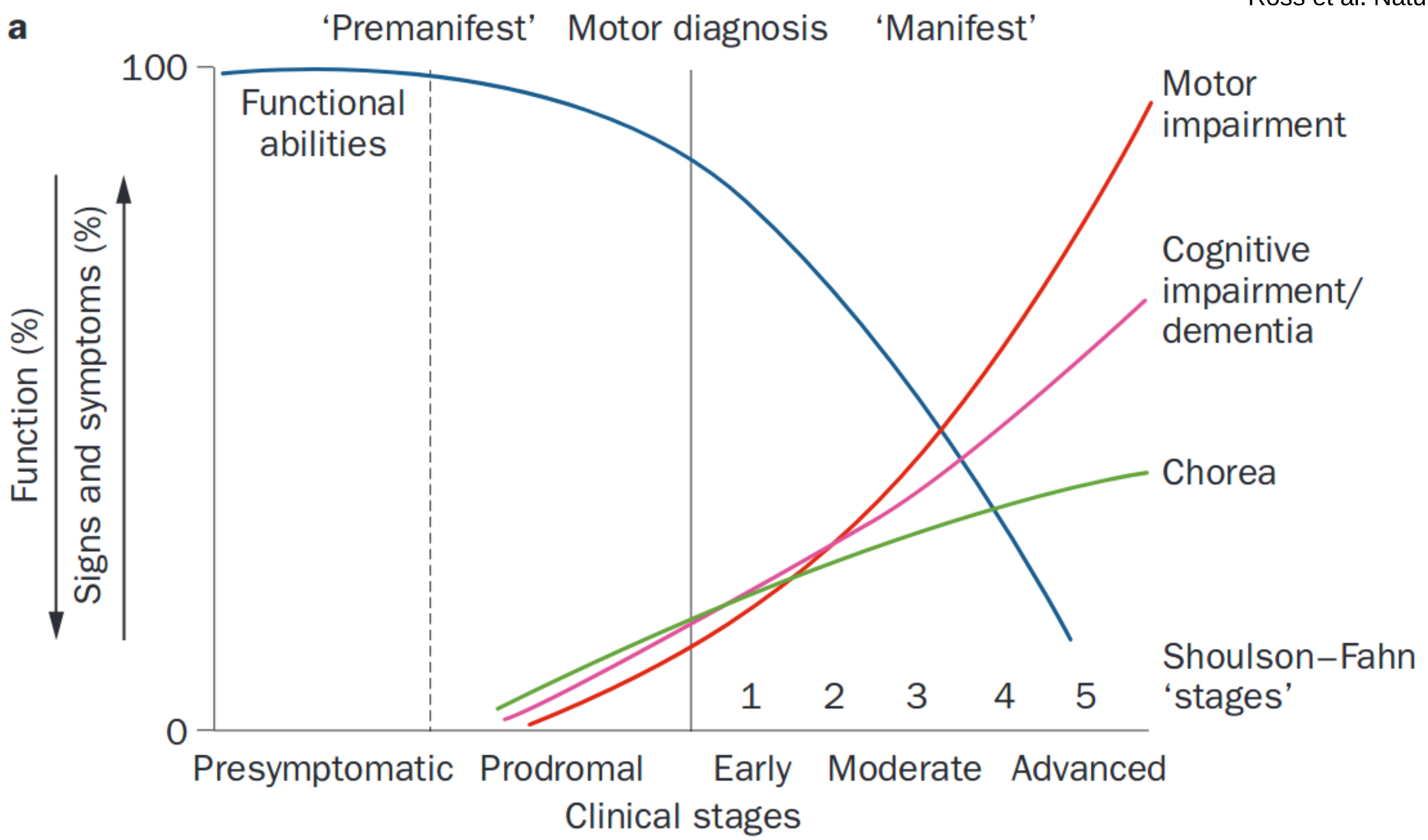
Computational Modeling



(a) First clinical weight vector.

Huntington’s disease

Ross et al. Nature Reviews 2014

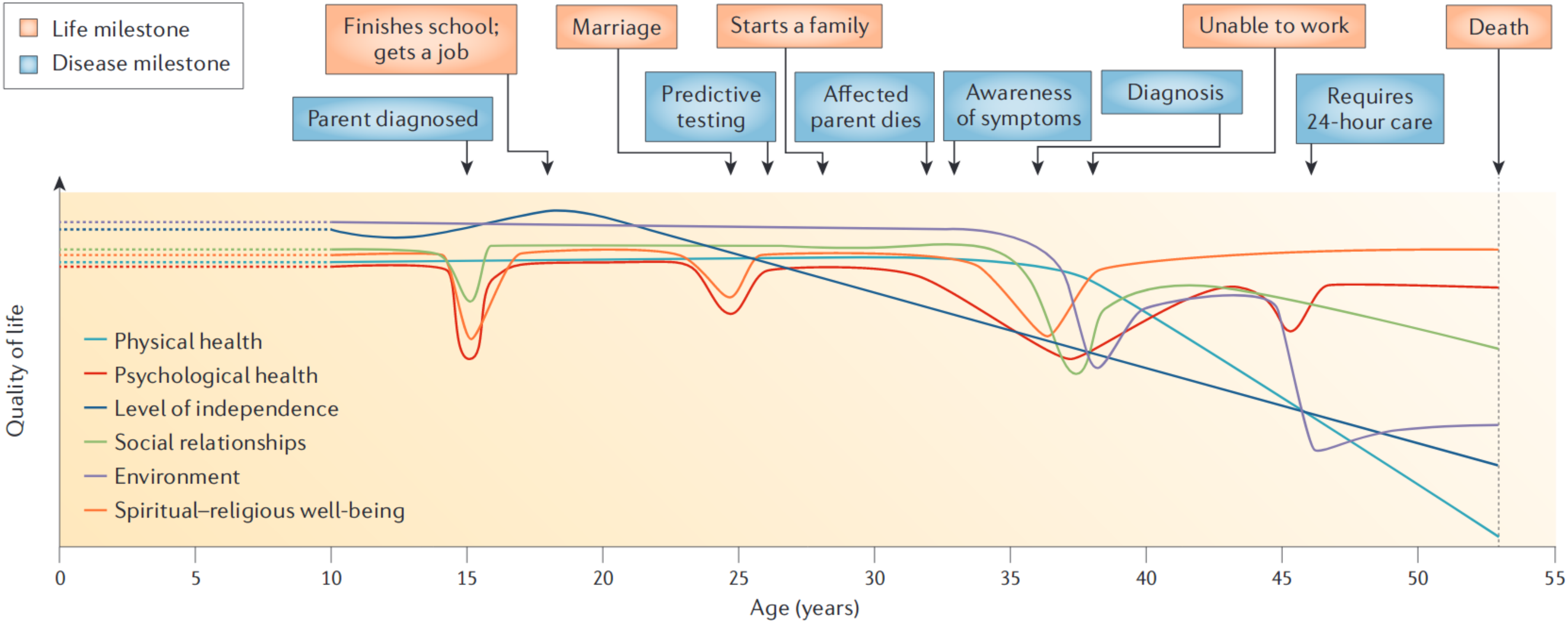


Slowly progressive, hereditary brain disease that causes changes in movement, thinking and behaviour

Autosomal dominant inheritance – 50% chance, everyone with gene will get HD

Huntington's disease

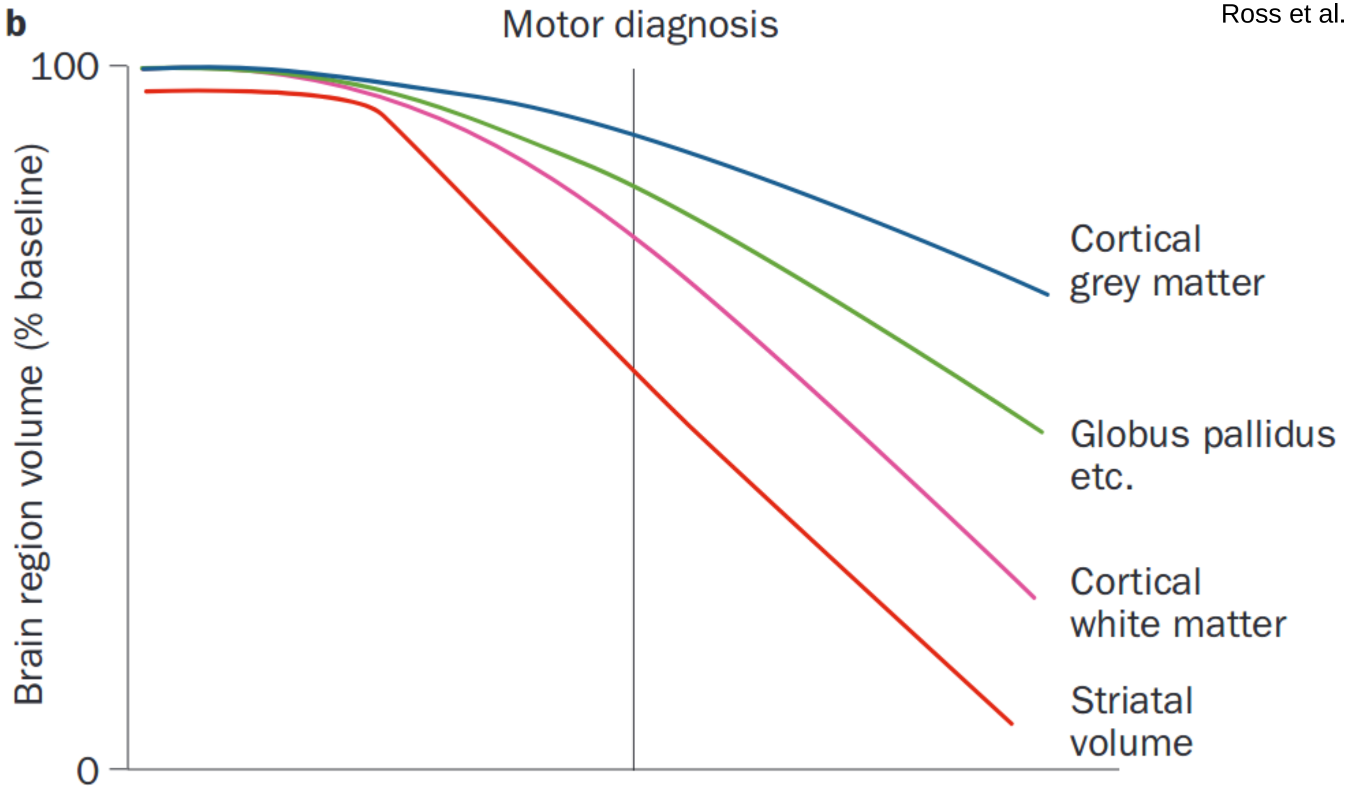
Bates et al. Nature Reviews Disease Primer. 2015



Diagnosis made at onset of movement disorder, typically with chorea and impaired voluntary movement

Huntington's disease

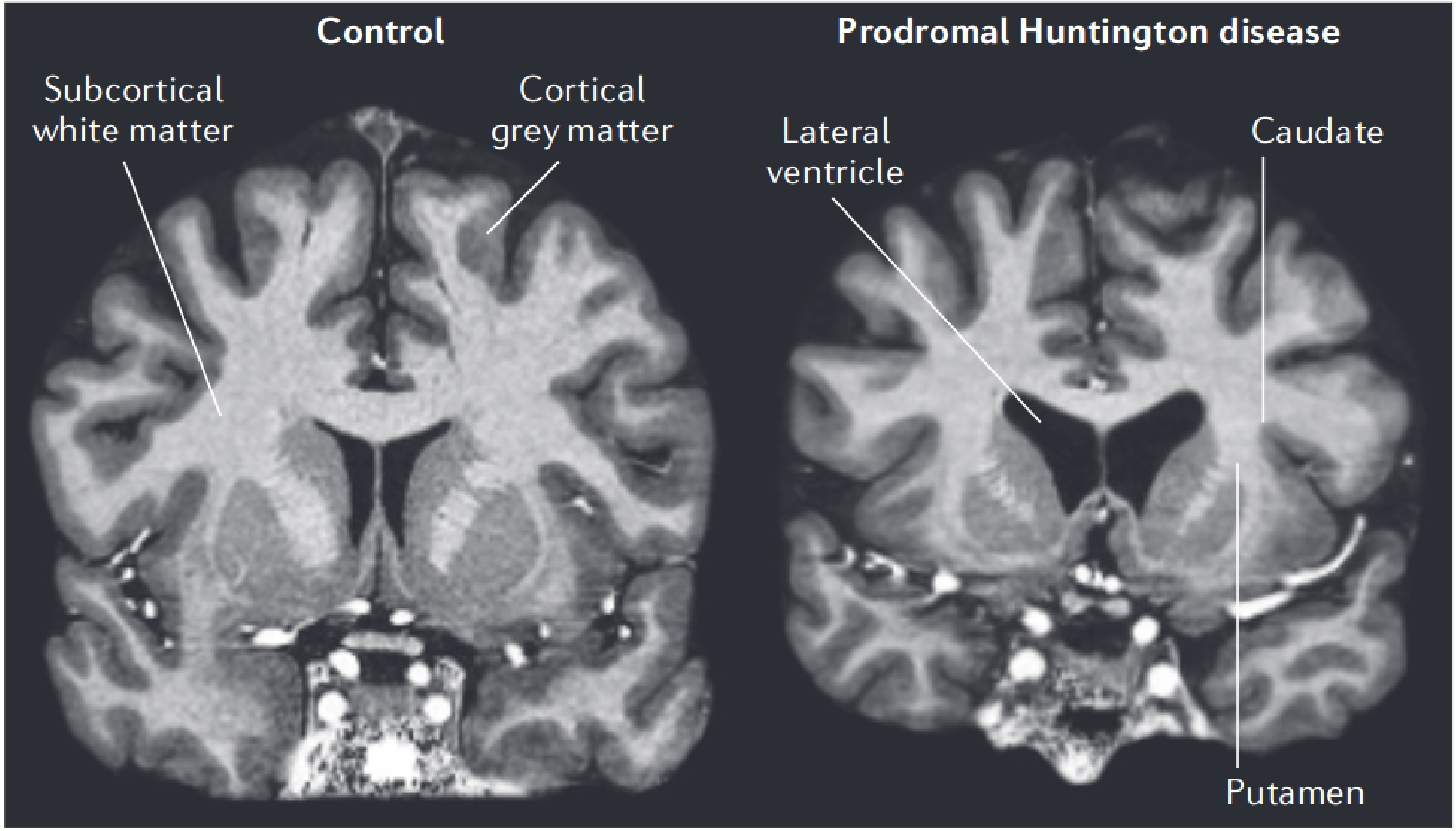
Ross et al. Nature Reviews 2014



Brain changes in HD – specific regions of the brain are atrophied

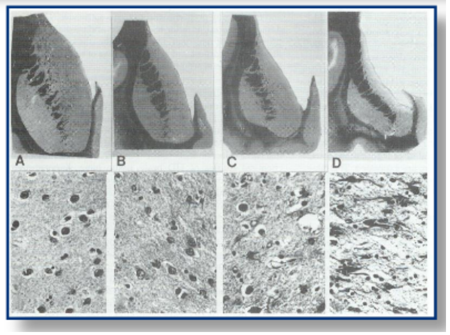
Huntington’s disease

Bates et al. Nature Reviews Disease Primer. 2015



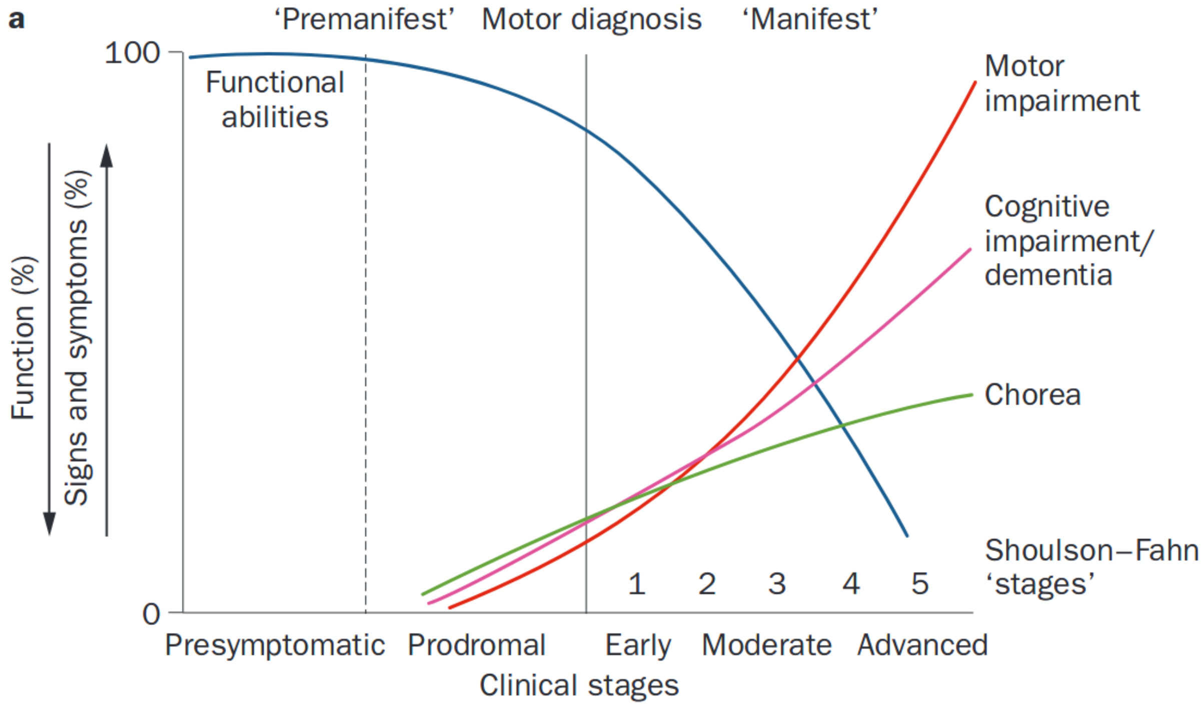
MRI provides spatial intensity measurements that depend on tissue properties

Observed changes reflected by microscopy (histology)



The problem

Can we estimate where a patient is along their disease path?

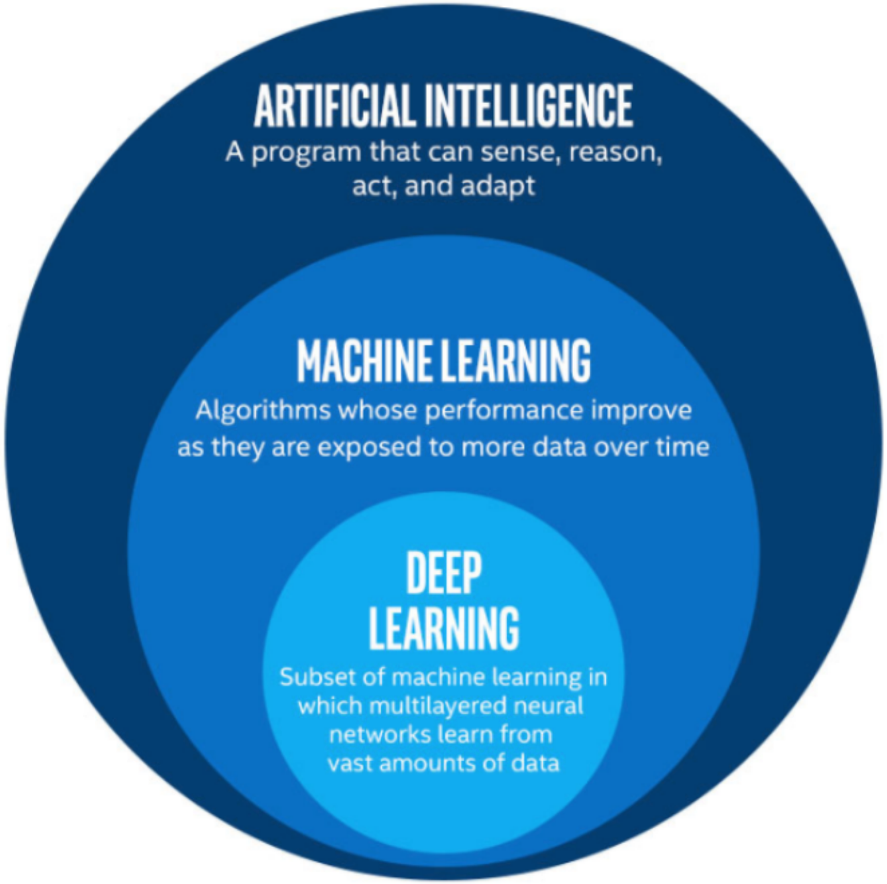


Patient stage is a latent variable – it generates the observed measurements, but is not measured directly (unlike in physics events, where we know time)

→Infer using machine learning methods

Learning and modeling

http://www.learnwebskill.com/technology


ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

Can think of machine learning as “data-driven AI”

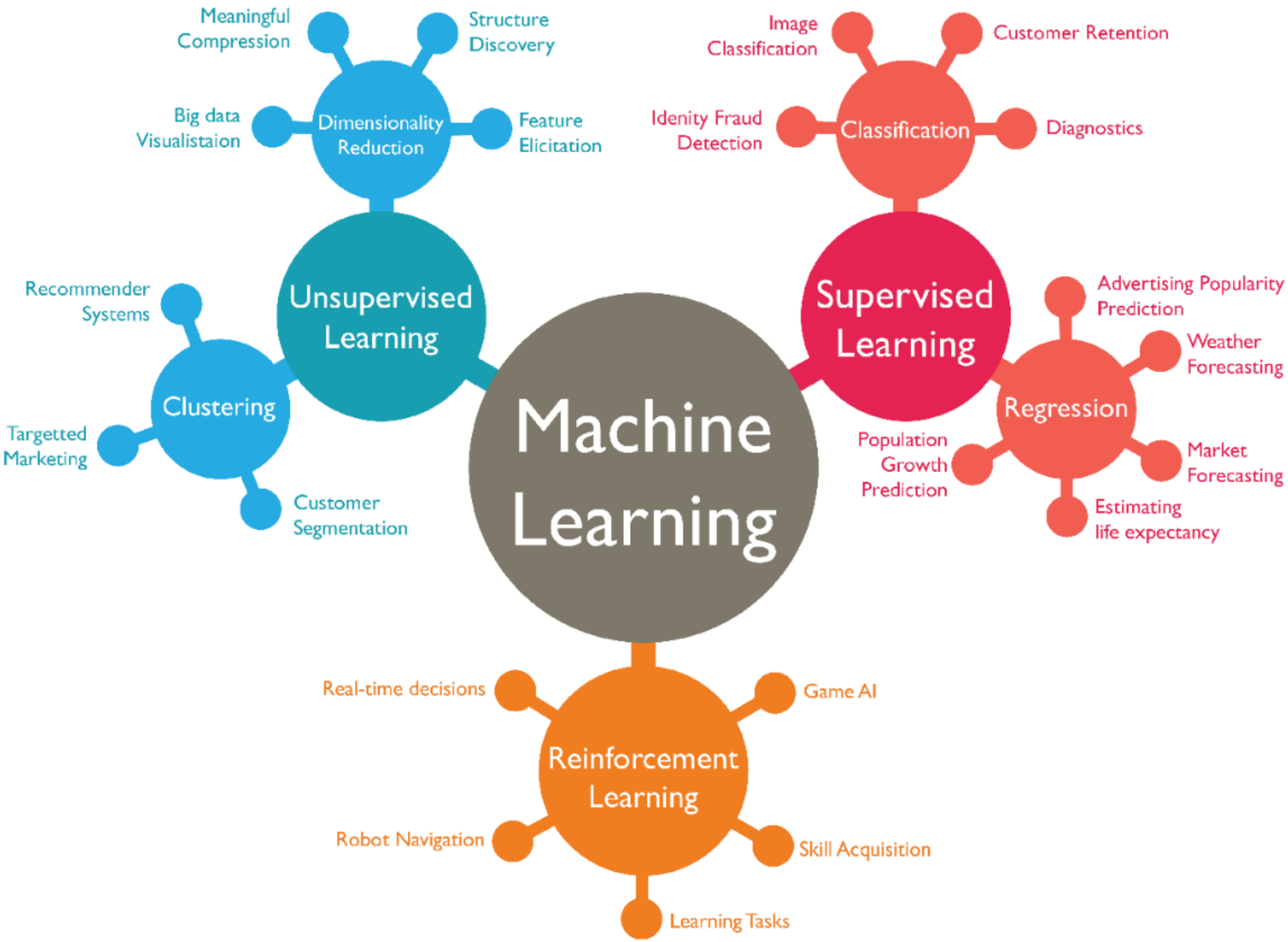
Deep learning learns its own feature space

+ improved performance over standard ML methods

- difficulty in interpretability

Learning and modelling

http://www.learnwebskill.com/technology



Learning and modelling

What machine learning does well

1. Model-free identification of trends and patterns

2. Improves with data availability

3. Requires minimal (or no) human intervention

What machine learning doesn’t do well

1. Causal mechanisms

2. Data intensive

3. Interpretability

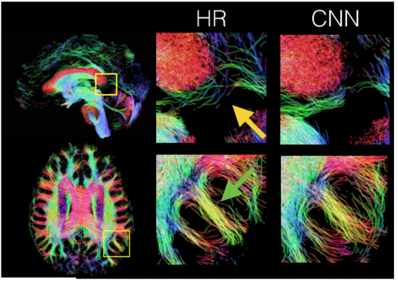
We want to diagnose and prognose patients – don’t really need to understand mechanisms

Bridging the gap

Basic sciences

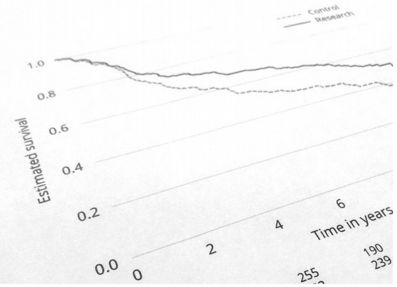
cmic
Centre for Medical Image Computing

Cluster computing



Imaging + machine learning

UCL EPSRC CDT in
Medical Imaging



Statistical methods

pond
Progression Of Neurodegenerative Disease

Clinical sciences

HUNTINGTON'S DISEASE CENTRE

Leonard Wolfson
Experimental Neurology Centre

Eisai

wellcome centre human neuroimaging

Some definitions

Biomarker: any biological measurement that tracks disease progression

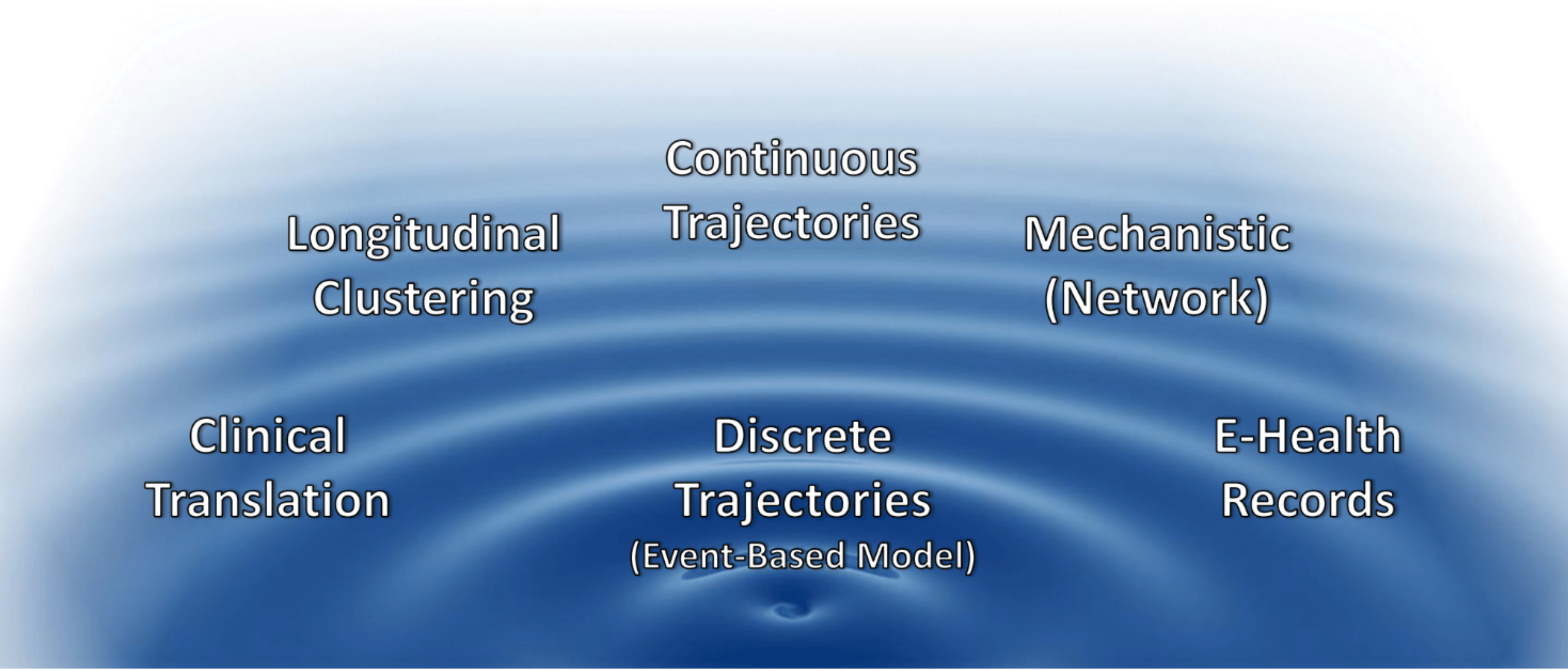
Event: transition of a biomarker from a normal to abnormal state (Markovian)

Sequence: order of events over sample of interest

Cross-sectional: data from a single time-point

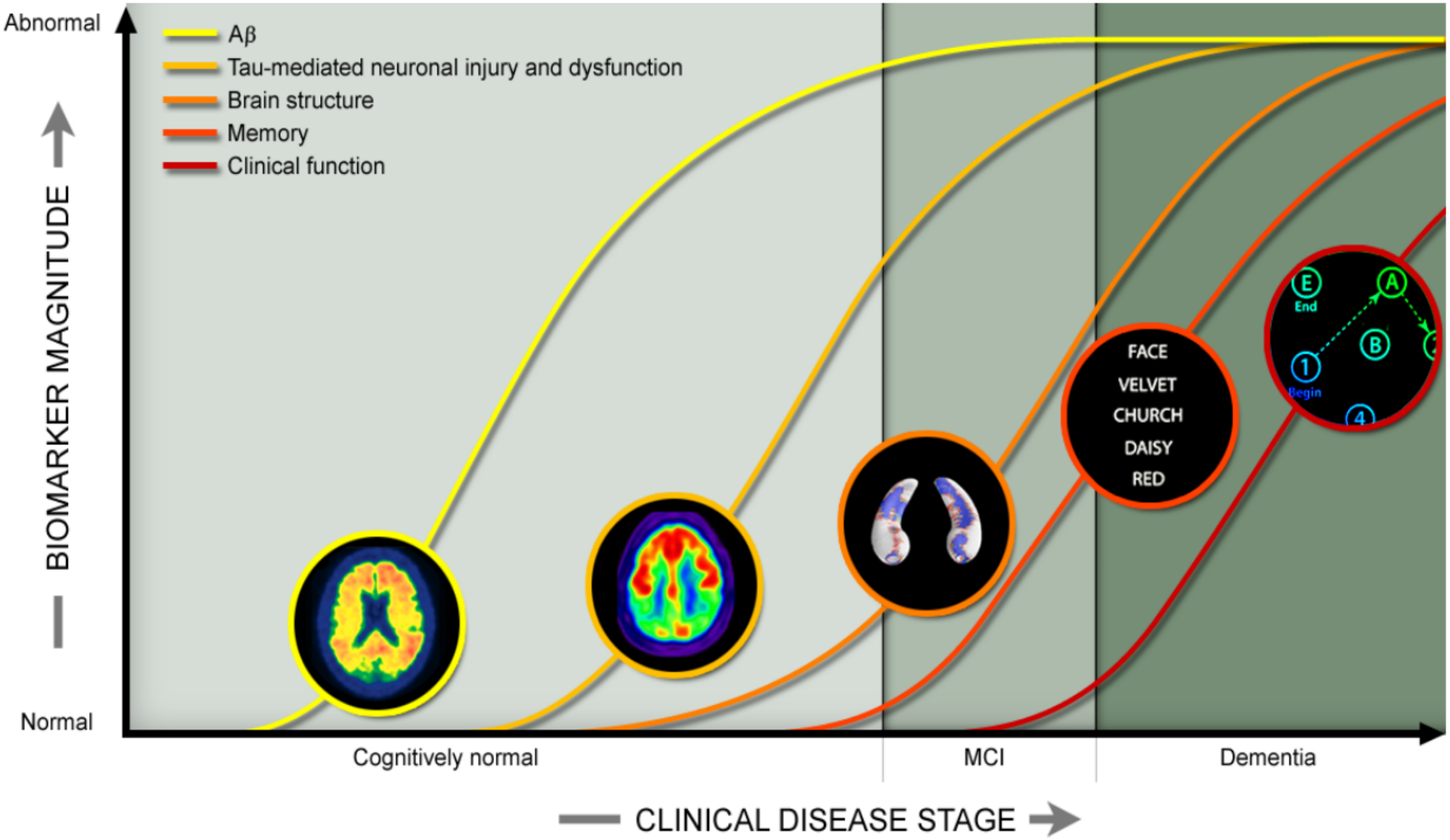
Progression of Neurological Disease (POND)

- Construct a picture of how disease plays out over time
- Express in terms of symptoms, pathologies and biomarkers
- Reconstruction must exploit cross-sectional data, where possible



High level: Disease progression modelling

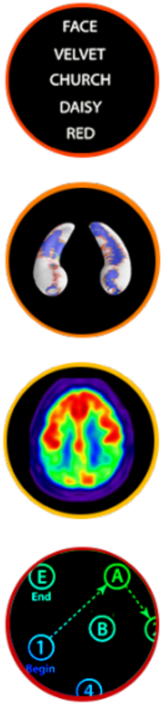
http://adni.loni.usc.edu/study-design/#background-container



A picture of how components of a disease progresses over time

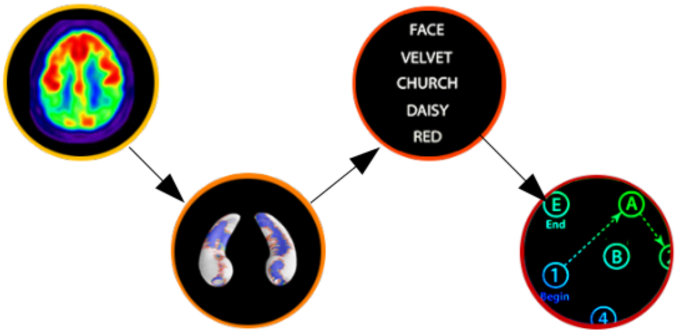
High level: Disease progression modelling

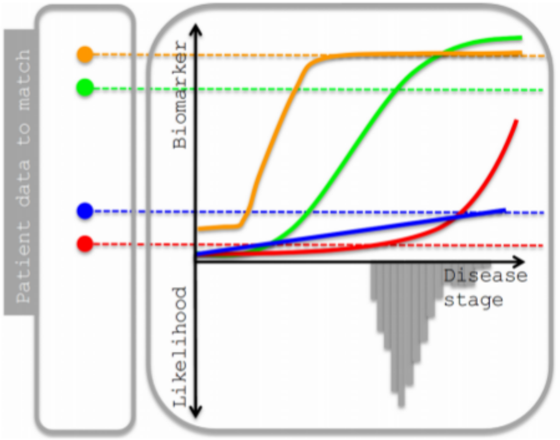
Disease progression models learn patterns of disease-related changes from data



Patient data

Machine learning





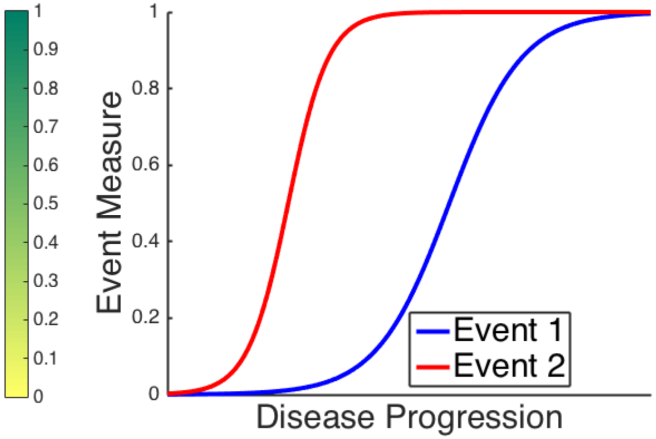
Disease progression model

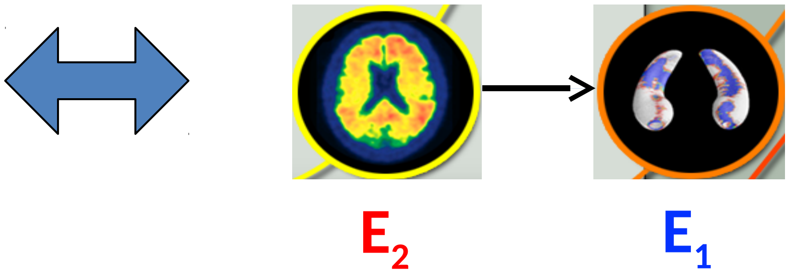
- Can use models to infer temporal ordering of changes
- Can also stage and stratify patients → clinical trial design

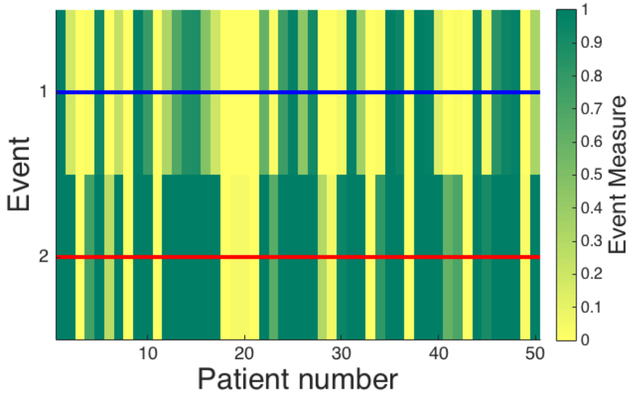
Background: Event-based model (EBM)

EBM estimates ordering of **binary events** from data – normal or abnormal

Data can be cross-sectional and any combination of types (imaging, clinical, genetic...)







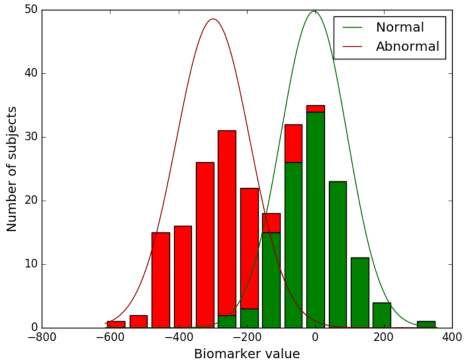
Simple example: 2 event measures

More patients have greater abnormality in Event 2 than Event 1

→ Event 2 **measurably abnormal** before Event 1

Background: Event-based model (EBM)

More formally: EBM is a generative model of observed data from unknown sequence



data uniform prior

$$P(X|S)=\prod_{j=1}^{N}\left[\sum_{k=0}^{Z}\left(P(k)\prod_{i=1}^{k}P(x_{ij}|E_i)\prod_{i=k+1}^{Z}P(x_{ij}|\neg E_i)\right)\right]$$

sequence

prob. Abnormal

prob. Normal

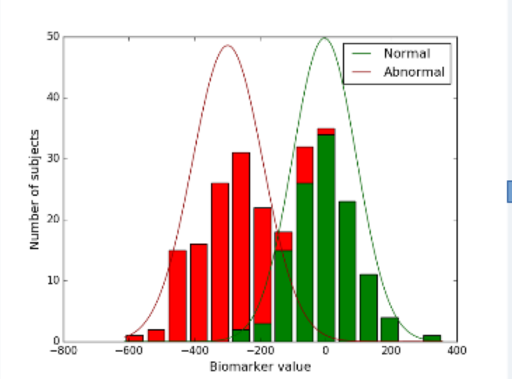
- The EBM needs likelihood distributions for normal and abnormal subjects

→ Learn directly from data

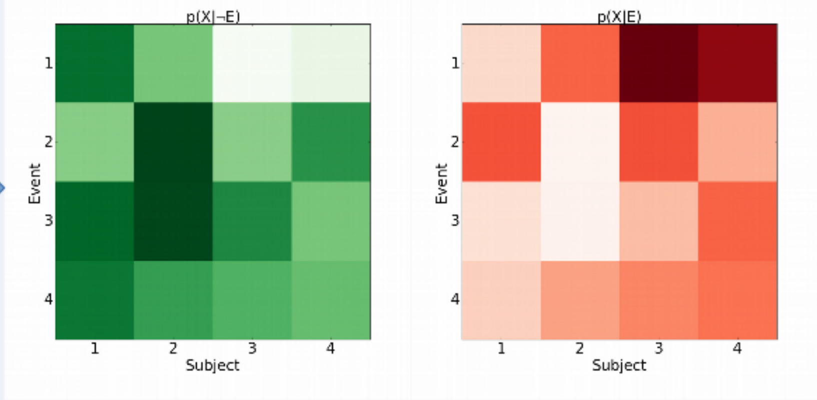
Example: Event-based model (EBM)

Event-based model

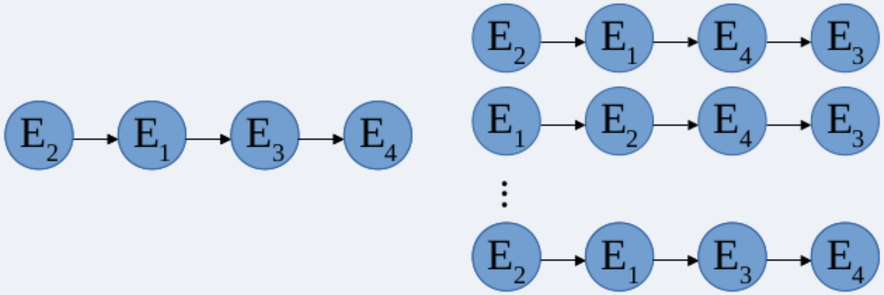
1. Fit mixture models to biomarkers

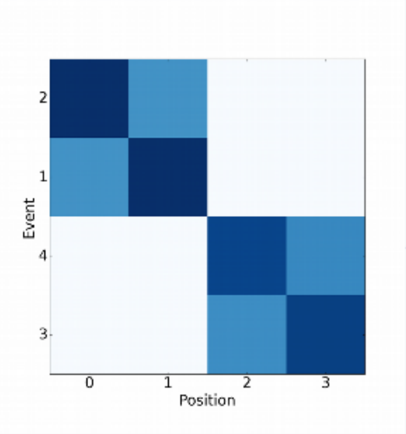


2. Calculate likelihoods of normality (event not occurred) and abnormality (event occurred)



3. Estimate most likely sequence by Markov Chain Monte Carlo sampling

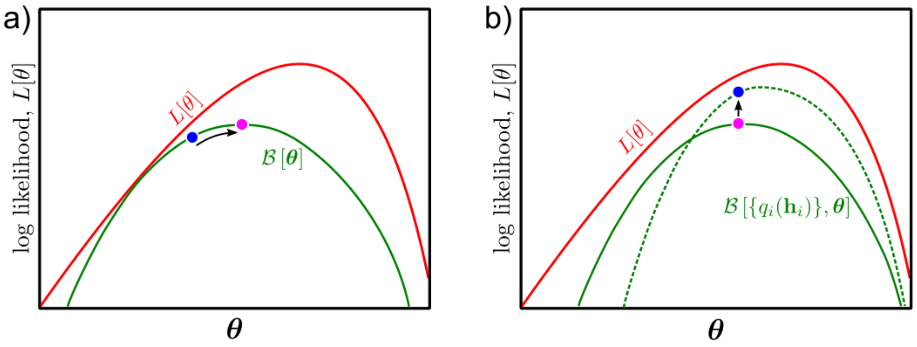

$E_2 \rightarrow E_1 \rightarrow E_3 \rightarrow E_4$
$E_2 \rightarrow E_1 \rightarrow E_4 \rightarrow E_3$
$E_1 \rightarrow E_2 \rightarrow E_4 \rightarrow E_3$
⋮
$E_2 \rightarrow E_1 \rightarrow E_3 \rightarrow E_4$



Positional variance diagram

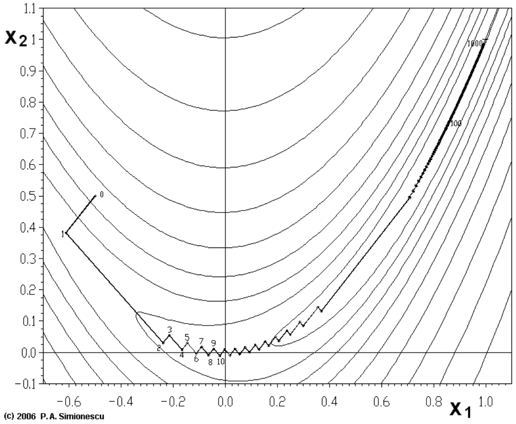
Toolkit: parameter estimation

1. Mixture model fitting
– Expectation Maximisation



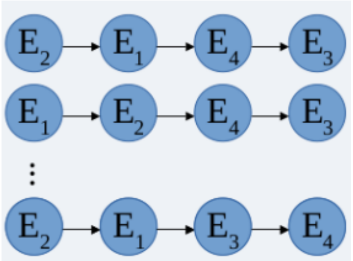
Prince, SJD. Cambridge University Press. 2012

2. Latent variable (sequence) fitting
– Gradient Ascent



wikipedia.org/wiki/gradient_descent

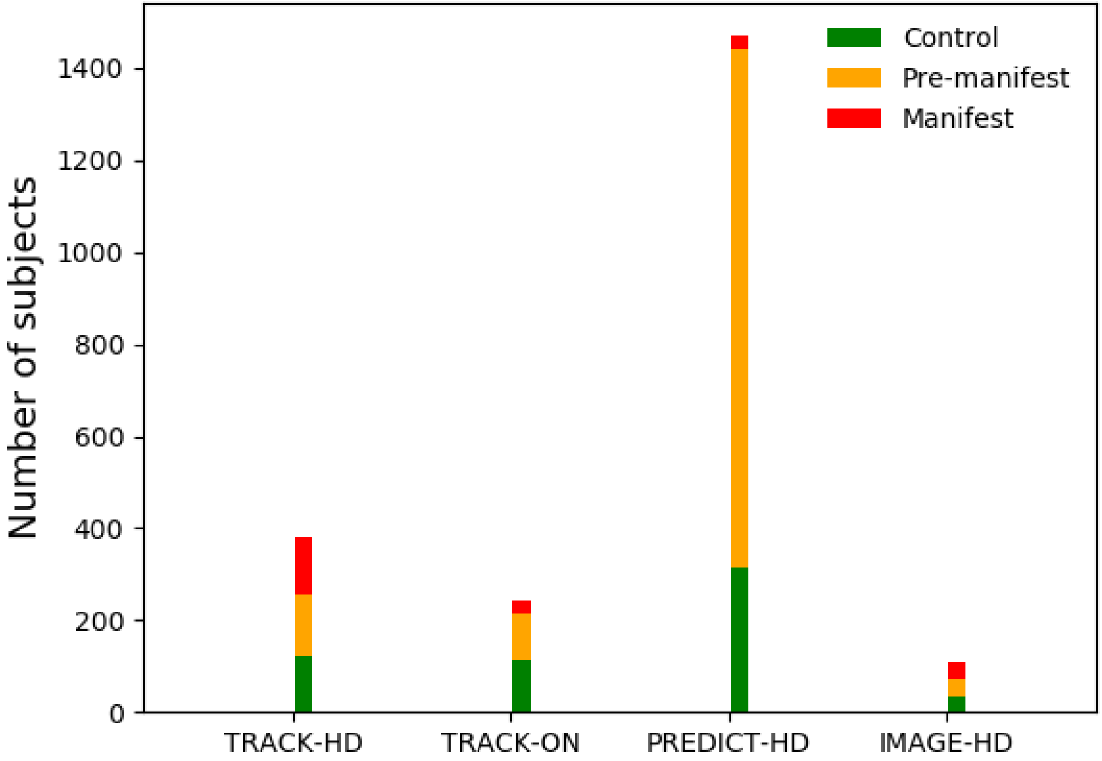
3. Uncertainty estimation
– Markov Chain Monte Carlo



E_2 → E_1 → E_4 → E_3
E_1 → E_2 → E_4 → E_3
⋮
E_2 → E_1 → E_3 → E_4

$$a = p(X \mid S')/p(X \mid S_t)$$

Methods: Data exploitation

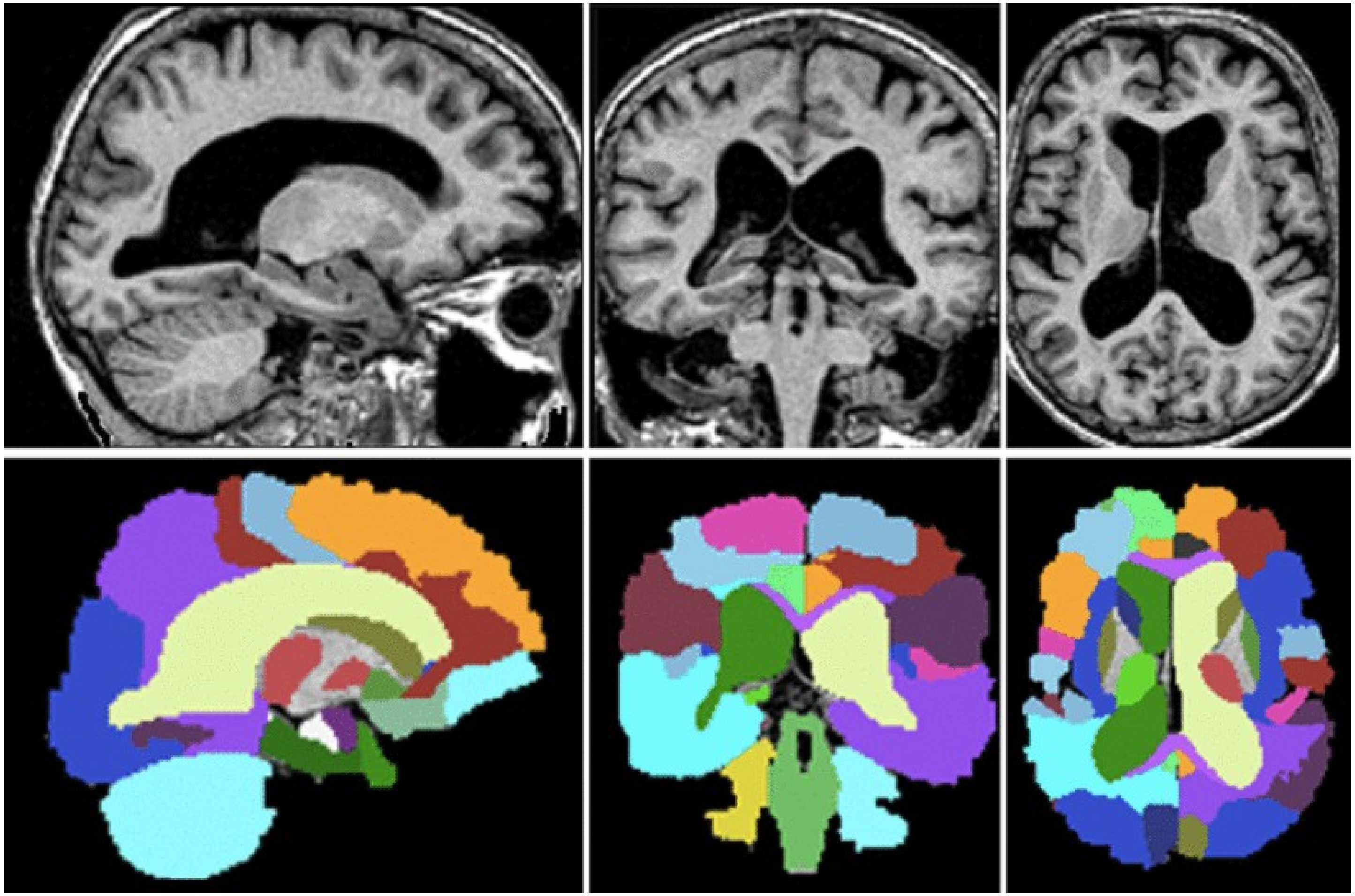


1. Build model on TRACK-HD

2. Cross-validate using PREDICT-HD and IMAGE-HD

3. Test predictive utility using TRACK-ON and PREDICT-HD

Methods: Imaging data

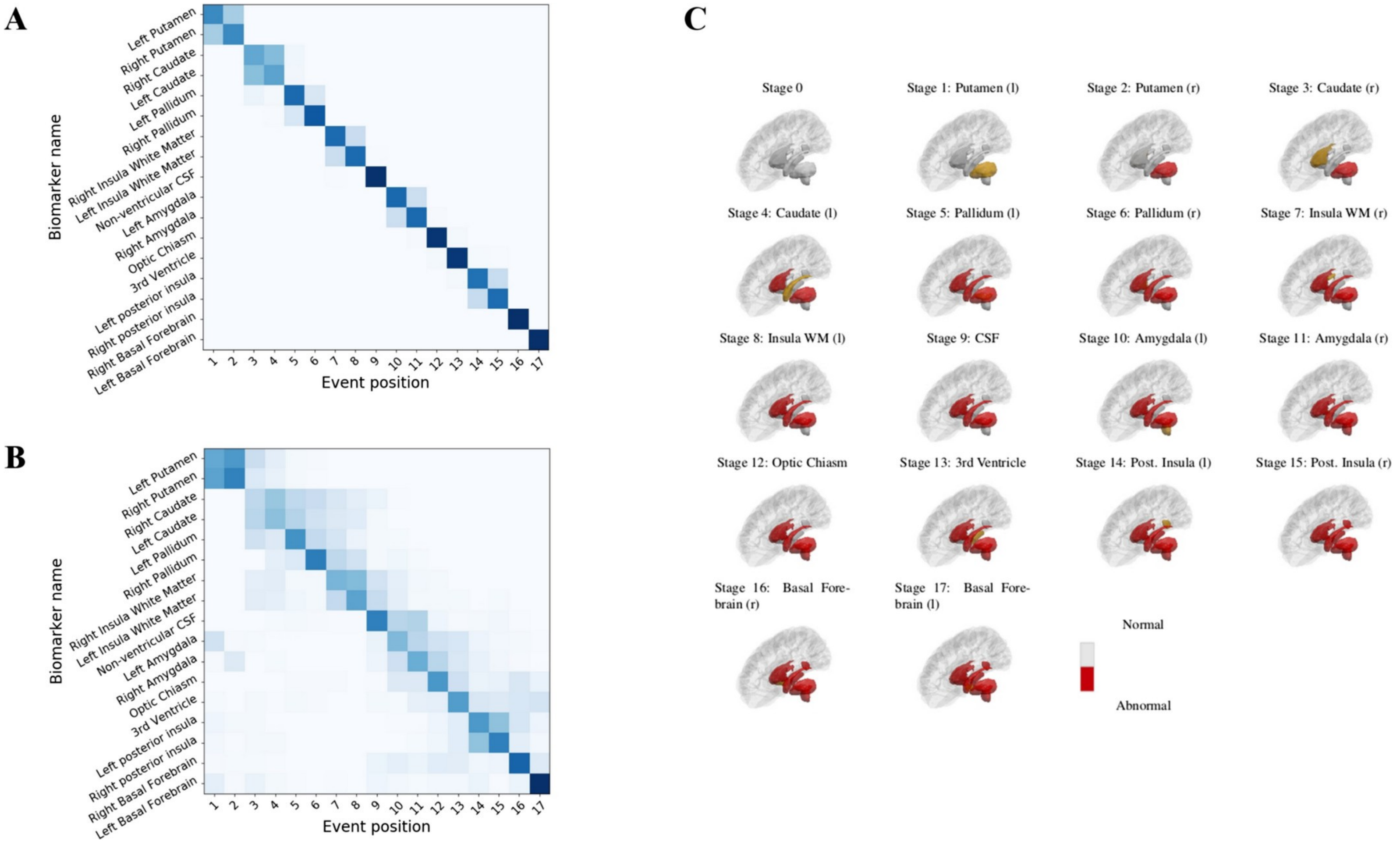


Extract regional brain volumes using Geodesic Information Flows*

→ Reduces inter-subject variability by using spatially variant graphs to connect morphologically similar subjects

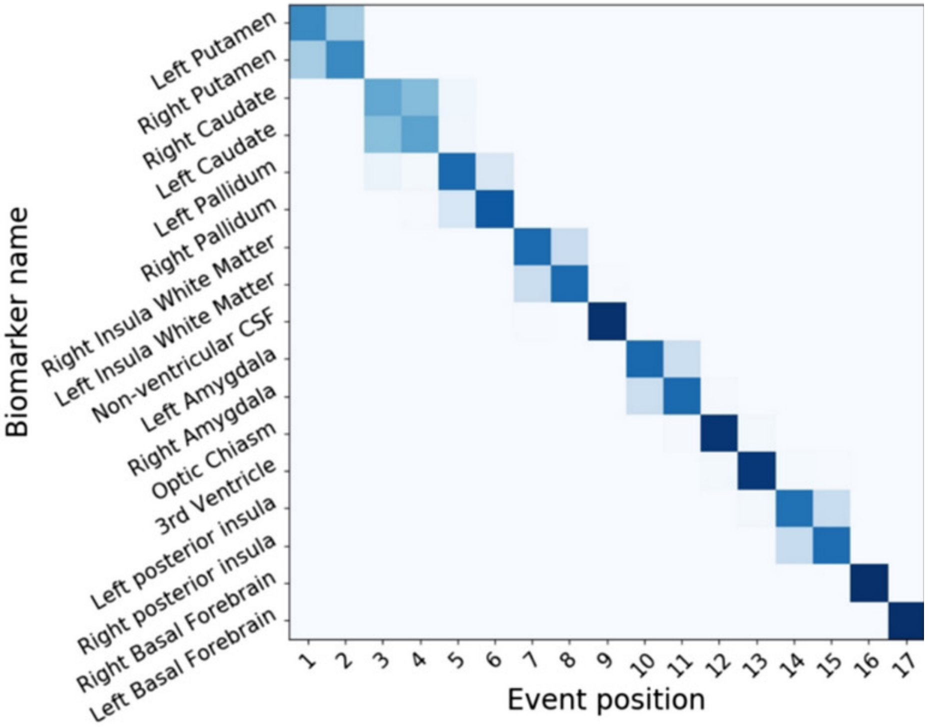
* MJ Cardoso *et al*. Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion. IEEE Transactions on Medical Imaging, 34 (2015), pp. 1976-1988, doi: 10.1109/TMI.2015.2418298

EBM in HD

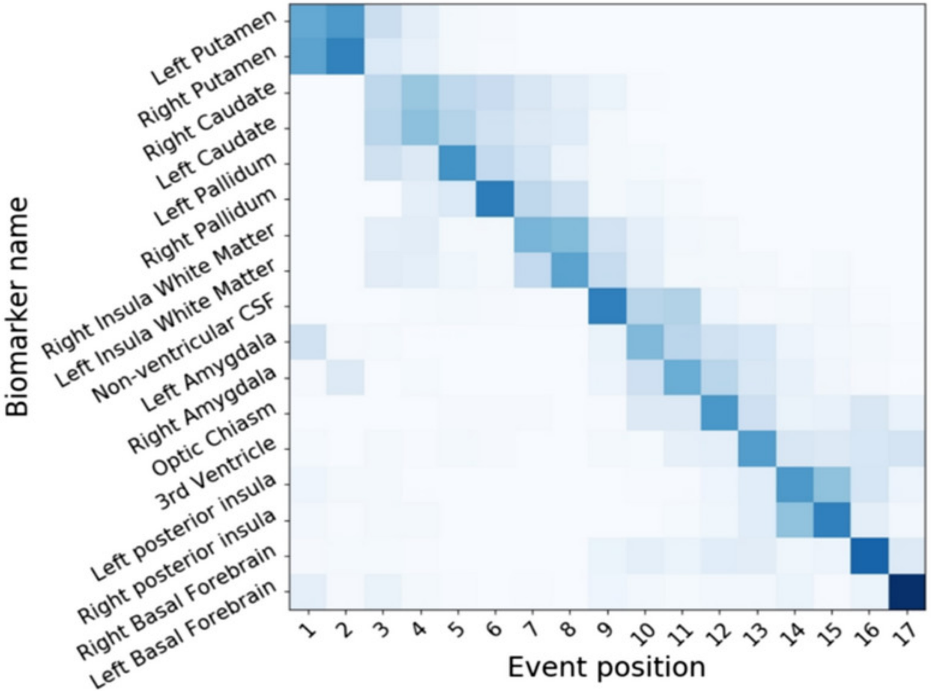

A
Biomarker name
Left Putamen
Right Putamen
Right Caudate
Left Caudate
Left Pallidum
Right Pallidum
Right Insula White Matter
Left Insula White Matter
Non-ventricular CSF
Left Amygdala
Right Amygdala
Optic Chiasm
3rd Ventricle
Left posterior insula
Right posterior insula
Right Basal Forebrain
Left Basal Forebrain
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Event position
B
Biomarker name
Left Putamen
Right Putamen
Right Caudate
Left Caudate
Left Pallidum
Right Pallidum
Right Insula White Matter
Left Insula White Matter
Non-ventricular CSF
Left Amygdala
Right Amygdala
Optic Chiasm
3rd Ventricle
Left posterior insula
Right posterior insula
Right Basal Forebrain
Left Basal Forebrain
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Event position
C
Stage 0
Stage 1: Putamen (l)
Stage 2: Putamen (r)
Stage 3: Caudate (r)
Stage 4: Caudate (l)
Stage 5: Pallidum (l)
Stage 6: Pallidum (r)
Stage 7: Insula WM (r)
Stage 8: Insula WM (l)
Stage 9: CSF
Stage 10: Amygdala (l)
Stage 11: Amygdala (r)
Stage 12: Optic Chiasm
Stage 13: 3rd Ventricle
Stage 14: Post. Insula (l)
Stage 15: Post. Insula (r)
Stage 16: Basal Fore-brain (r)
Stage 17: Basal Fore-brain (l)
Normal
Abnormal

Event sequence

Direct model fit

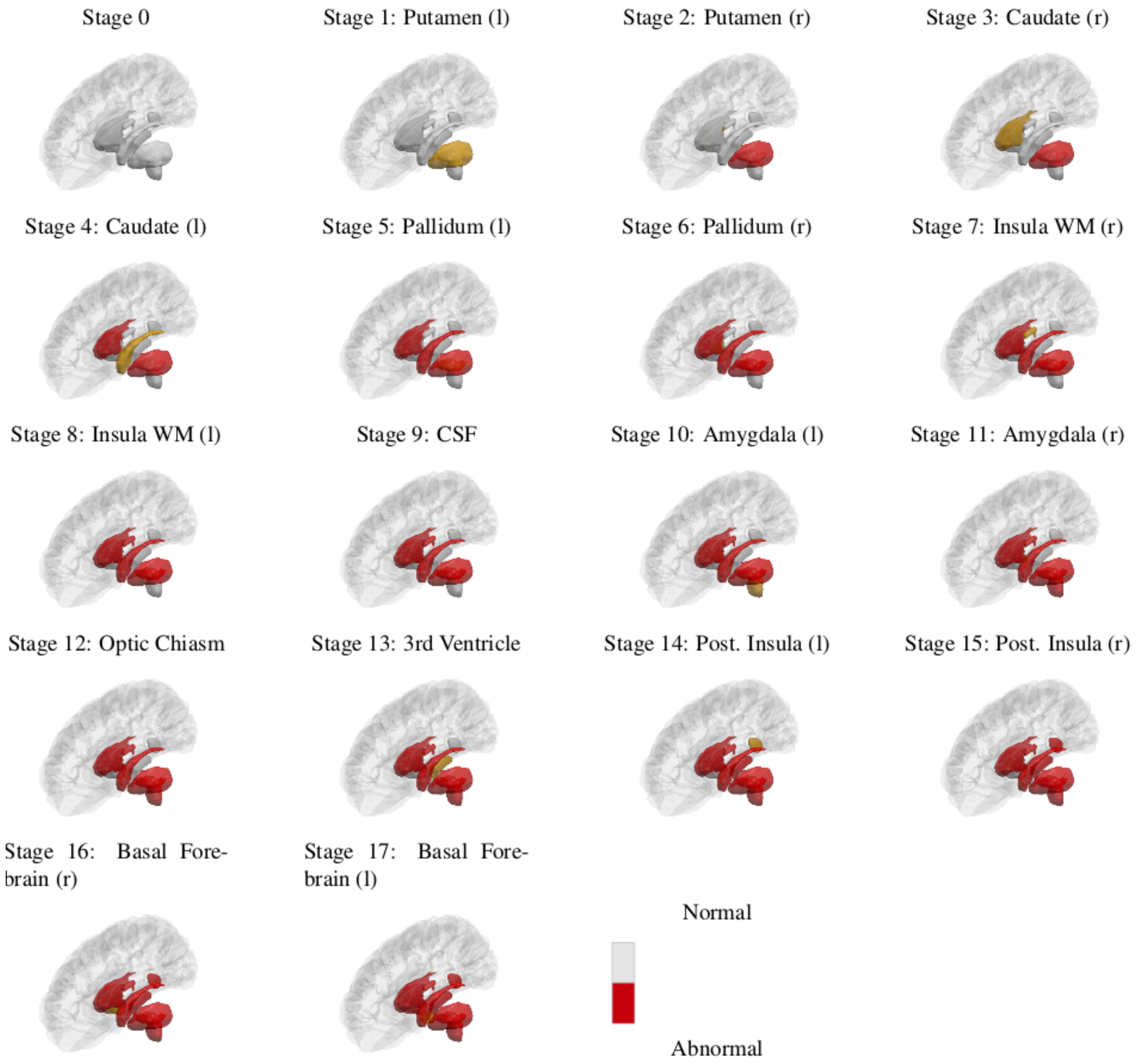


Bootstrapped model fit



- Dark diagonal components indicate strong event ordering
- Lighter indicate possible event permutations

Atrophy progression



Stage 0

Stage 1: Putamen (l)

Stage 2: Putamen (r)

Stage 3: Caudate (r)

Stage 4: Caudate (l)

Stage 5: Pallidum (l)

Stage 6: Pallidum (r)

Stage 7: Insula WM (r)

Stage 8: Insula WM (l)

Stage 9: CSF

Stage 10: Amygdala (l)

Stage 11: Amygdala (r)

Stage 12: Optic Chiasm

Stage 13: 3rd Ventricle

Stage 14: Post. Insula (l)

Stage 15: Post. Insula (r)

Stage 16: Basal Forebrain (r)

Stage 17: Basal Forebrain (l)

Normal

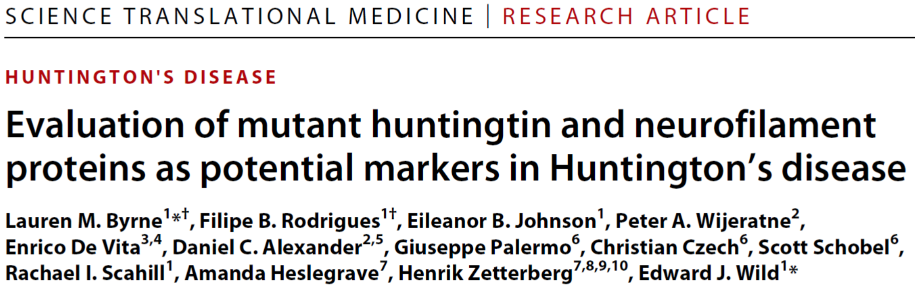
Abnormal

Central

HD progression

Peripheral

Backup: HD-CSF EBM

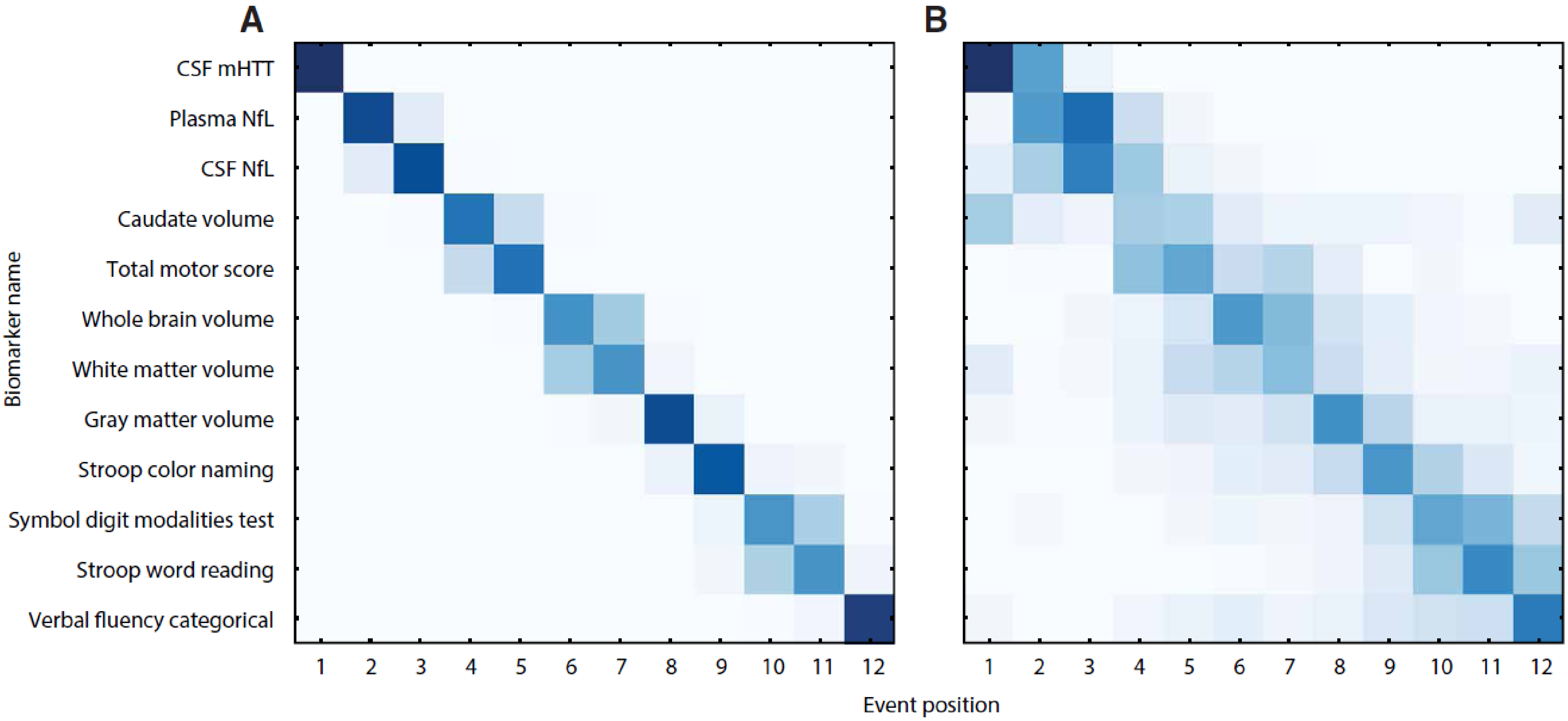


SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE

HUNTINGTON'S DISEASE

Evaluation of mutant huntingtin and neurofilament proteins as potential markers in Huntington's disease

Lauren M. Byrne[1*†], Filipe B. Rodrigues[1†], Eileanor B. Johnson[1], Peter A. Wijeratne[2], Enrico De Vita[3,4], Daniel C. Alexander[2,5], Giuseppe Palermo[6], Christian Czech[6], Scott Schobel[6], Rachael I. Scahill[1], Amanda Heslegrave[7], Henrik Zetterberg[7,8,9,10], Edward J. Wild[1]*



- Biofluid markers change before imaging and clinical markers

Results: staging

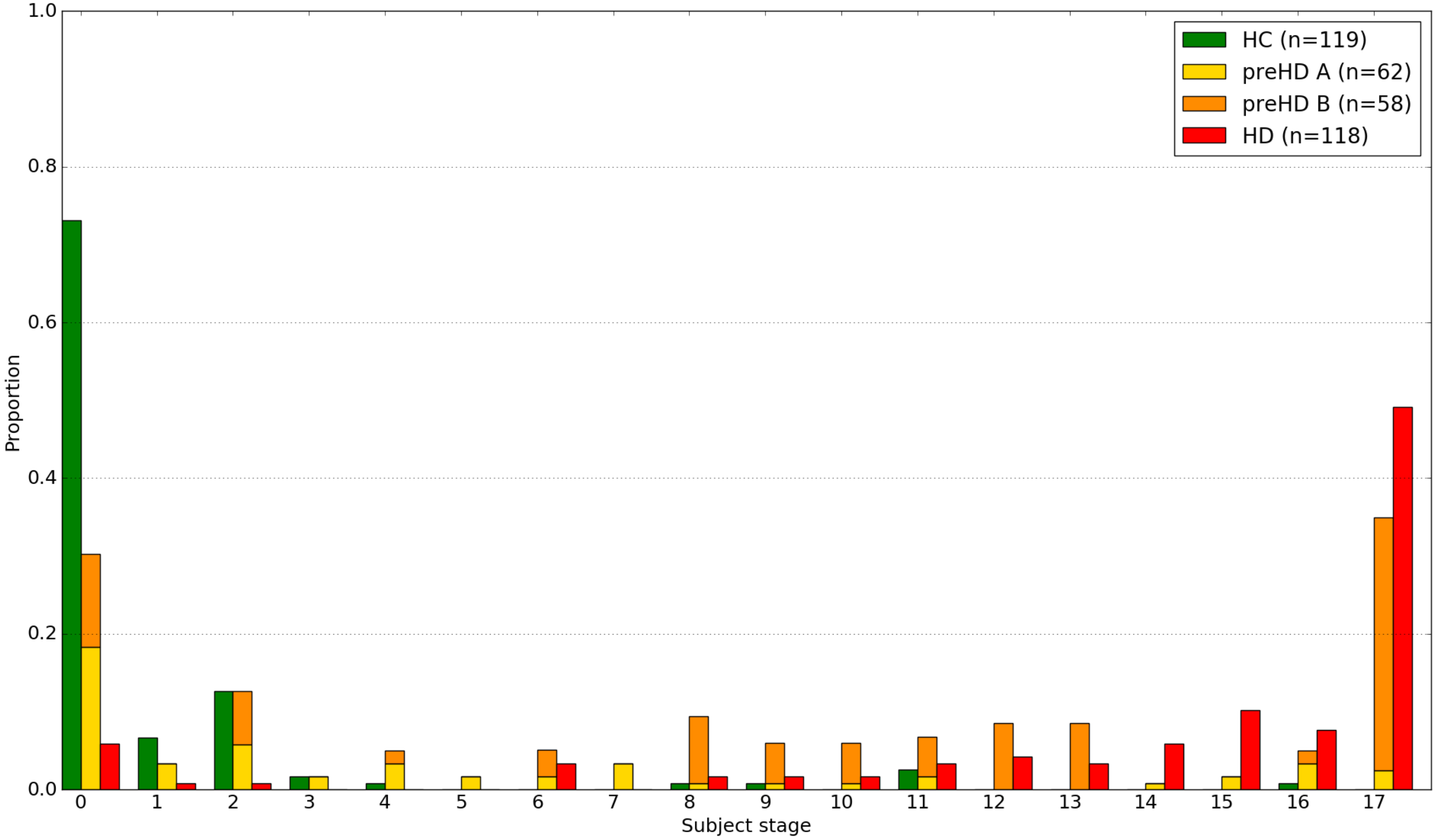
Simplest way is to take the stage that maximises the likelihood for each patient

$$argmax_k P\left(X_j \mid \bar{S}, k\right) = argmax_k P(k) \prod_{i=1}^{k} P(x_{ij} \mid E_i) \prod_{i=k+1}^{I} P(x_{ij} \mid \neg E_i)$$

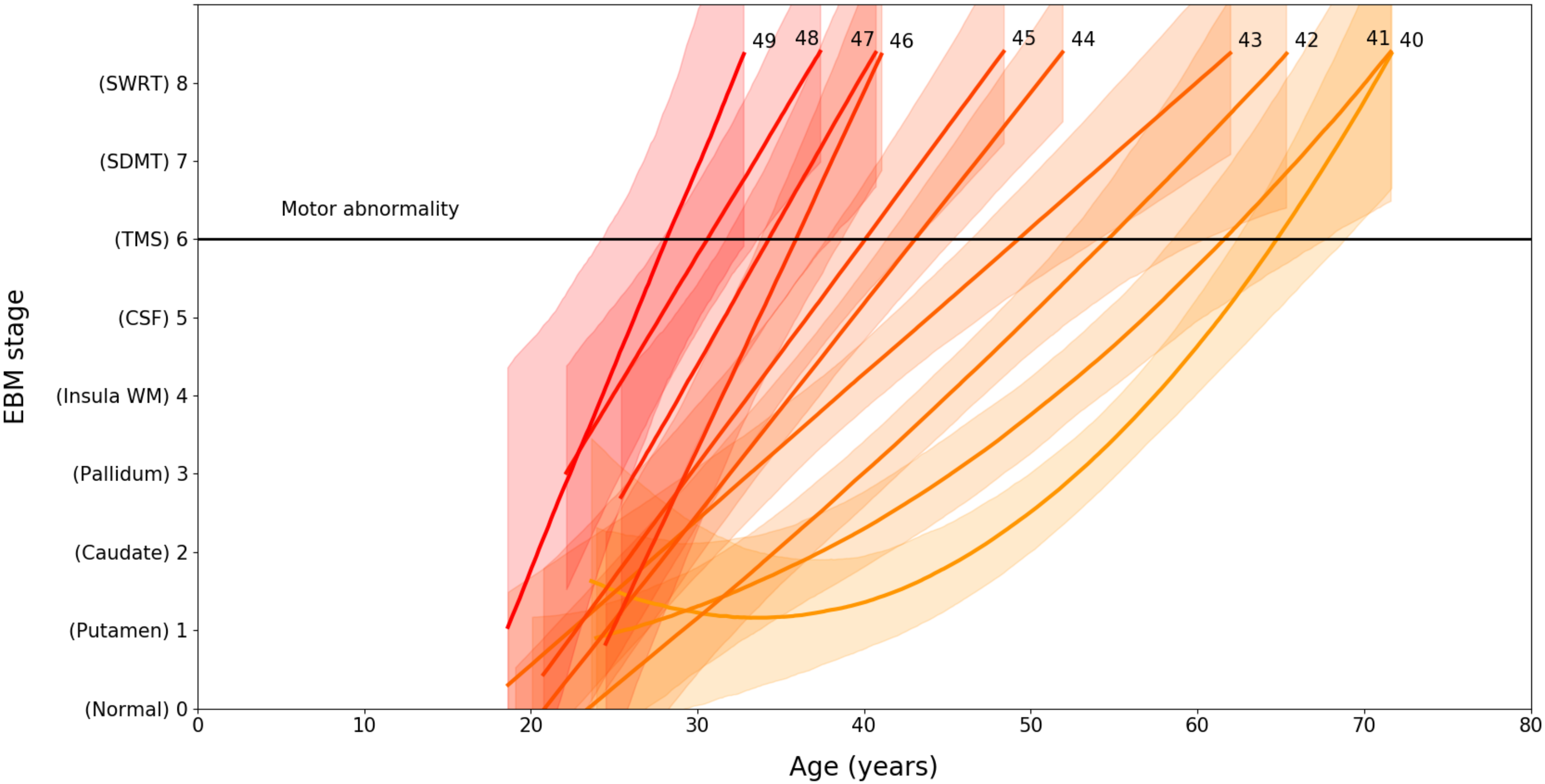
Staging patients

Simplest way is to take the stage that maximises the likelihood for each patient

$$argmax_k P(X_j | \bar{S}, k) = argmax_k P(k) \prod_{i=1}^{k} P(x_{ij} | E_i) \prod_{i=k+1}^{I} P(x_{ij} | \neg E_i)$$

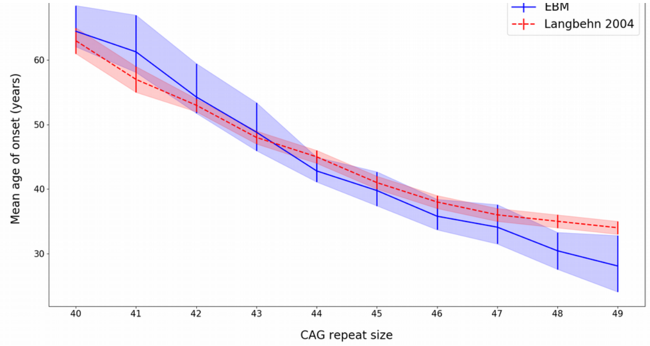


Extending EBM-HD + cross-validation



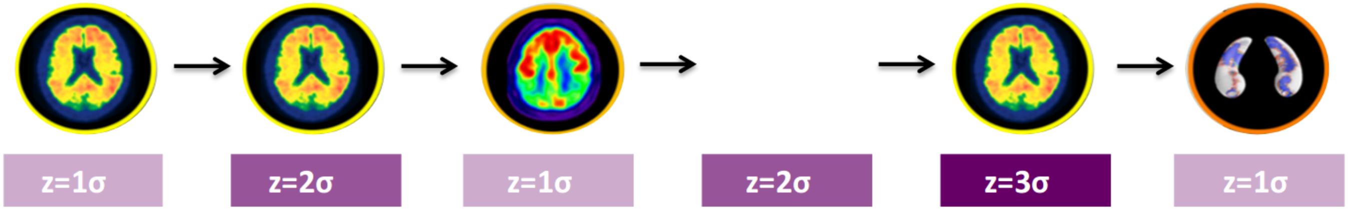
- Estimate age at event e.g.
  for CAG 40, WM atrophy at ~60 years old
  for CAG 49, WM atrophy at ~25 years old

- Age of onset agrees well with gold standard



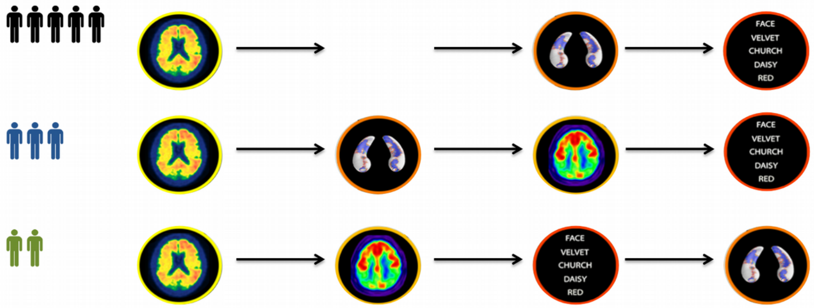
SuStaIn: Subtype and Stage Inference

1. Continuous generalisation of EBM: instead of instantaneous abnormality, markers are a linear combination of z-scores



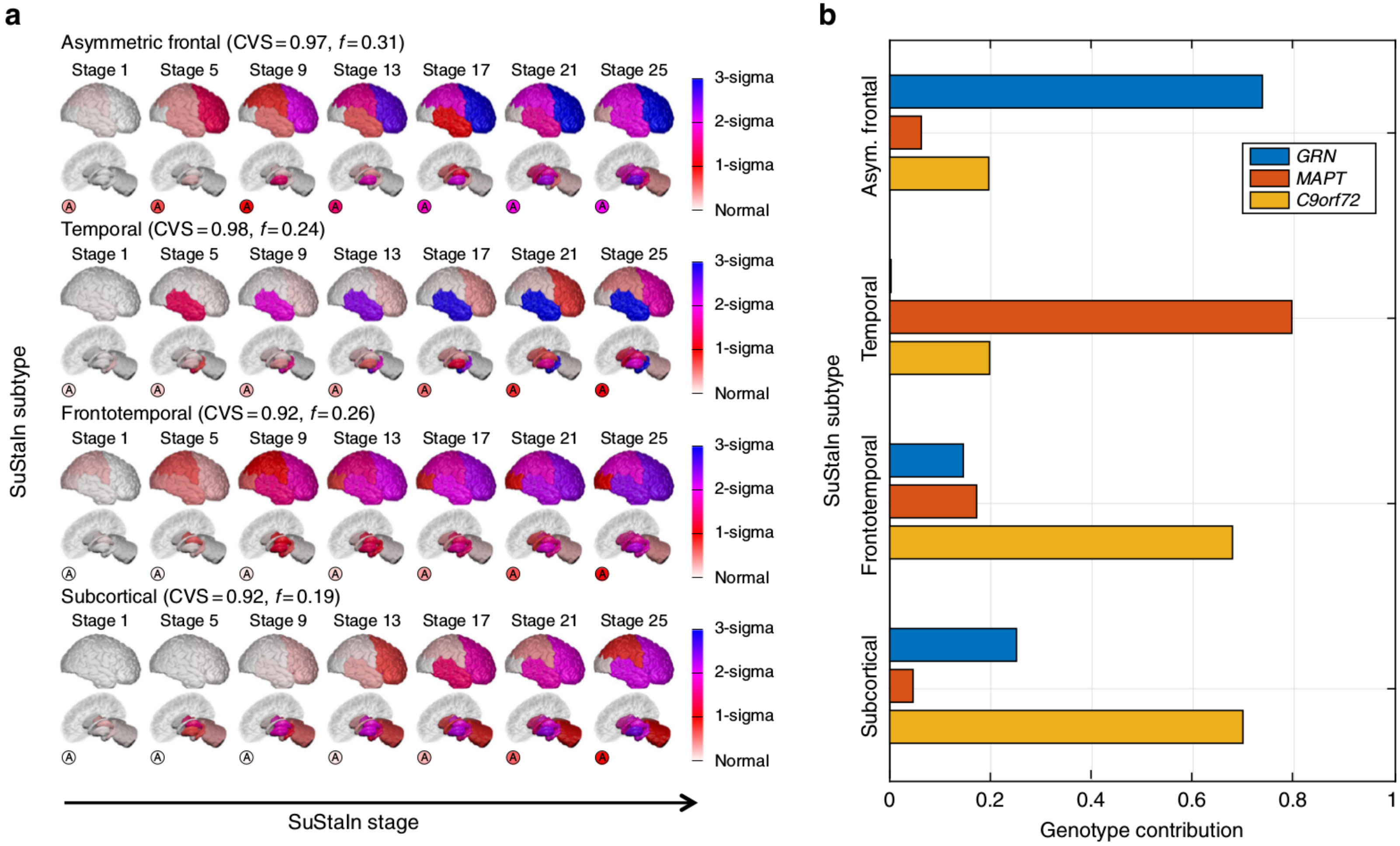
“Z-score model”

2. Total model is mixture of linear z-score models: grouped into clusters with distinct progression patterns


FACE VELVET CHURCH DAISY RED
FACE VELVET CHURCH DAISY RED
FACE VELVET CHURCH DAISY RED

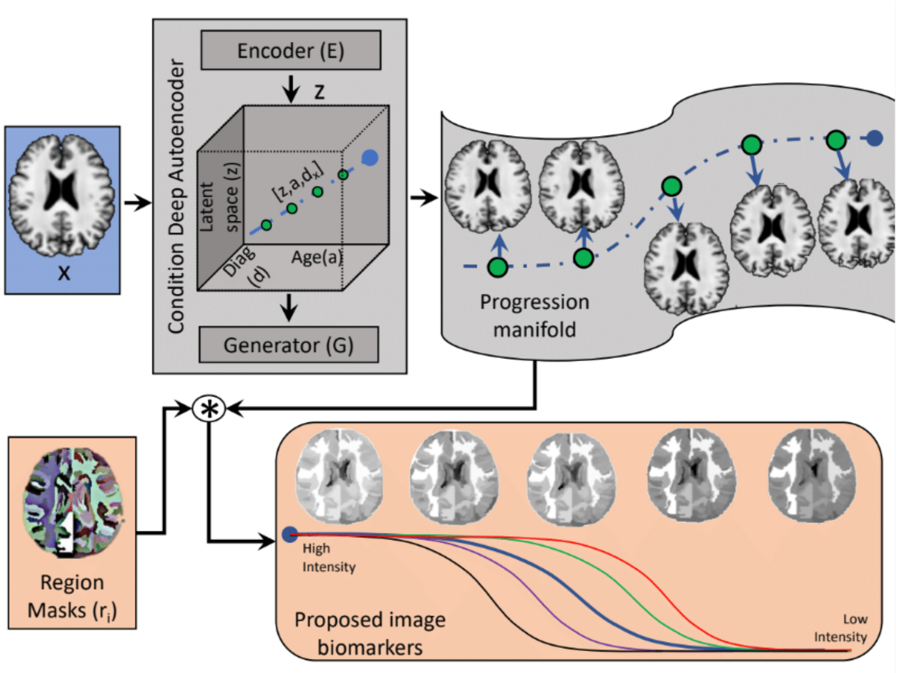
“Algorithm”

SuStaIn: Subtype and Stage Inference



Gain this extra information just by generalising event-based model – pretty neat

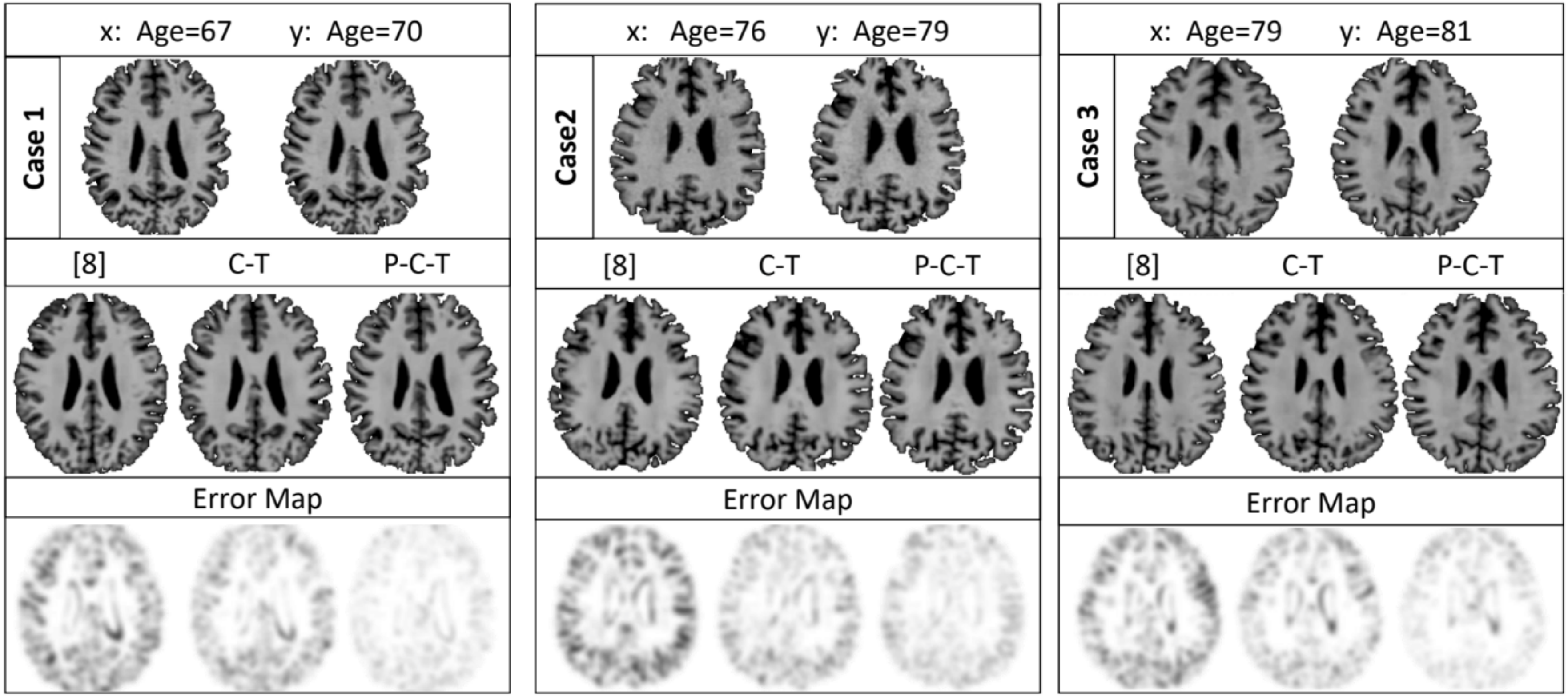
Simulating brain atrophy



Deep learning disease trajectories using generative adversarial networks

- also used in HEP e.g. CaloGAN, Paganini, Oliveira, Nachman. 2017.

Simulating brain atrophy

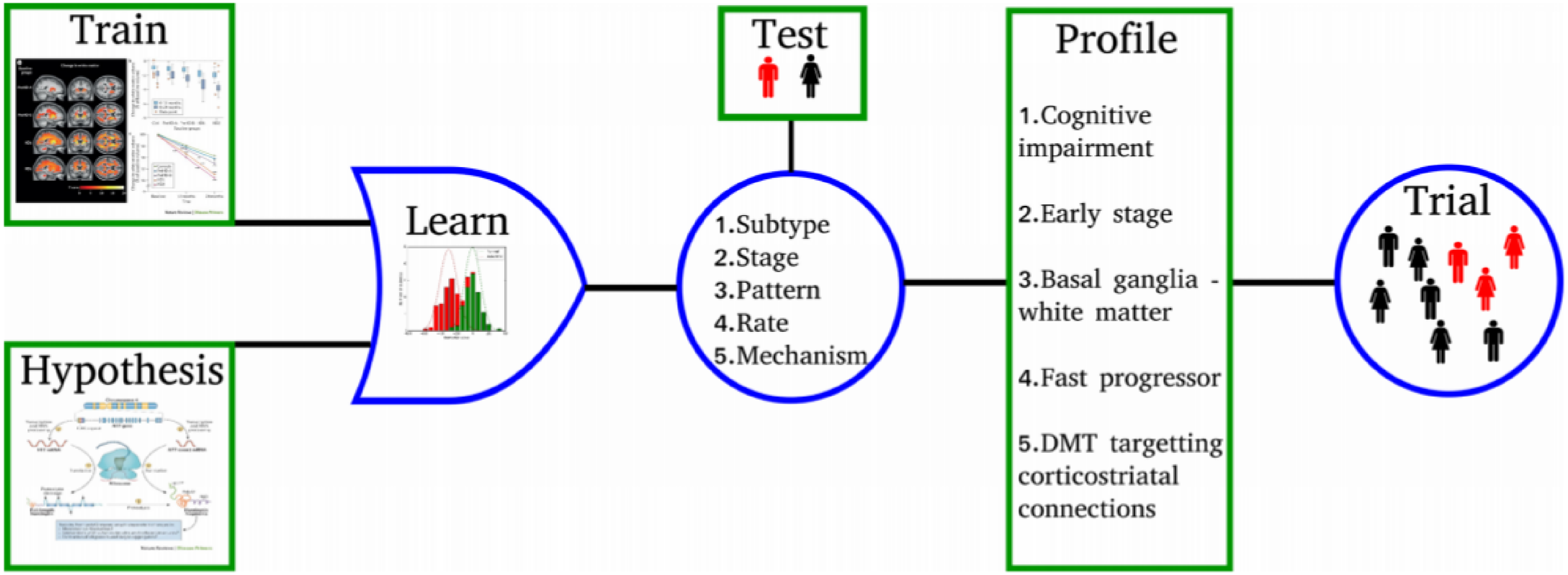


Deep learning disease trajectories using generative adversarial networks

- also used in HEP e.g. CaloGAN, Paganini, Oliveira, Nachman. 2017.

Big vision: computational models for clinical trials

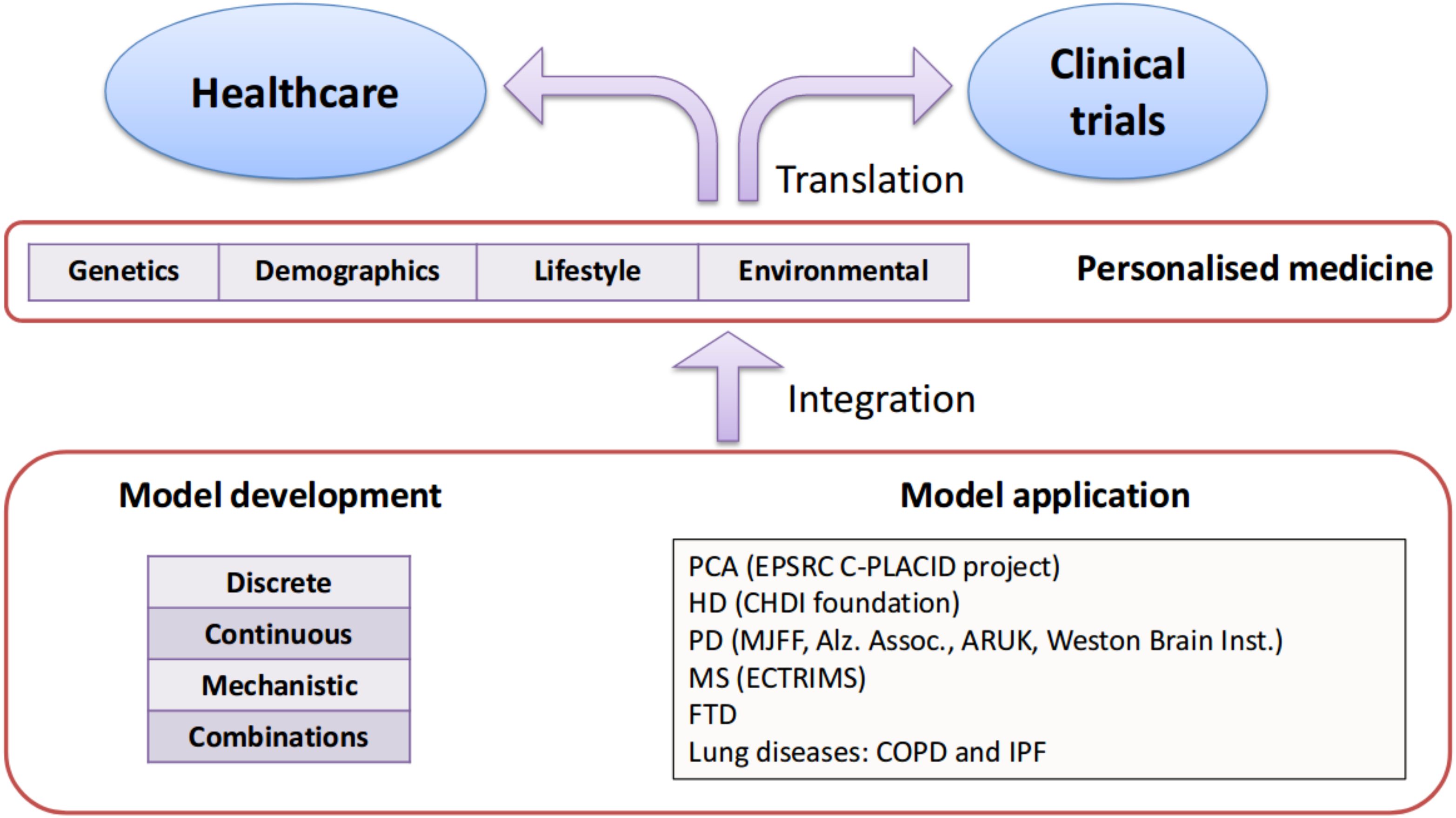
Patient data + machine learning = personalised profiles for clinical trial design



Model can be used for both prospective and retrospective analysis

→Save money and time
→Optimise trial design

Future



Healthcare

Clinical trials

Translation

Genetics | Demographics | Lifestyle | Environmental

Personalised medicine

Integration

Model development

- Discrete
- Continuous
- Mechanistic
- Combinations

Model application

PCA (EPSRC C-PLACID project)
HD (CHDI foundation)
PD (MJFF, Alz. Assoc., ARUK, Weston Brain Inst.)
MS (ECTRIMS)
FTD
Lung diseases: COPD and IPF

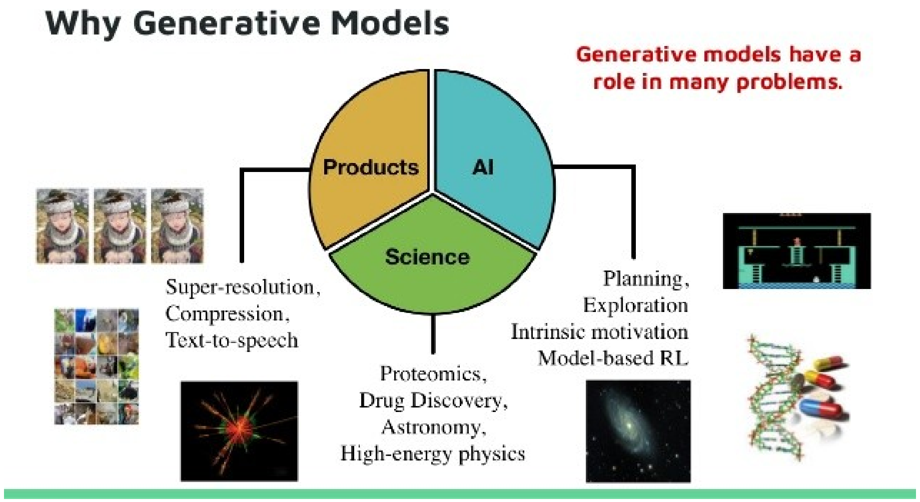
Discussion

- Presented computational methods to extract information from large and varied datasets

- Machine learning methods are suitable for medical problems – i.e. inferring patterns from complex systems

- Still much to do – can we understand the mechanisms themselves?

- What can HEP and CS learn from each other?



Why Generative Models

Generative models have a role in many problems.

Super-resolution,
Compression,
Text-to-speech

Proteomics,
Drug Discovery,
Astronomy,
High-energy physics

Planning,
Exploration
Intrinsic motivation
Model-based RL

https://www.slideshare.net/mlreview/tutorial-on-deep-generative-models