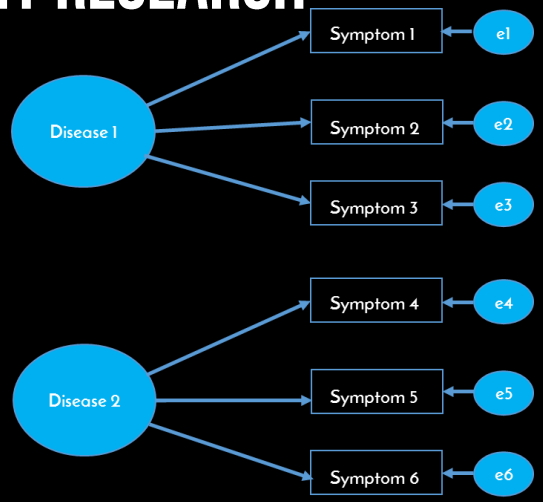
APPLICATIONS OF MACHINE LEARNING IN HEALTHCARE

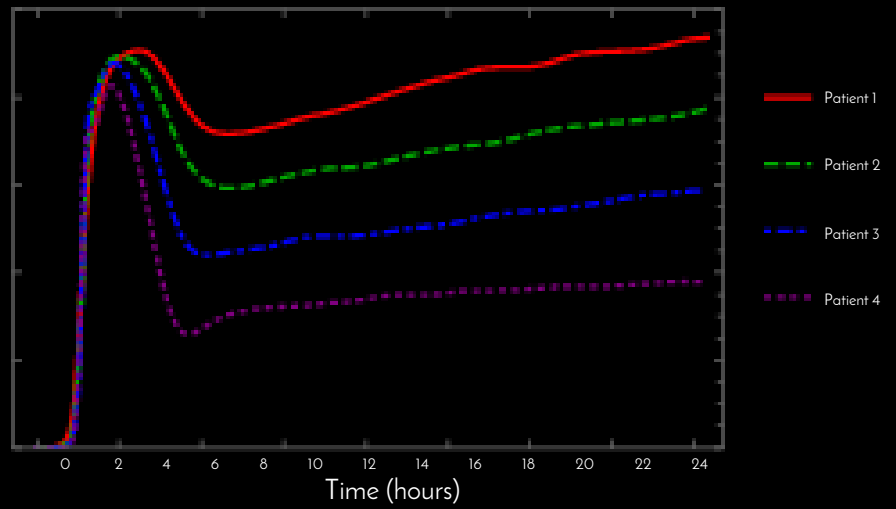
Danielle Belgrave
Imperial College London


62°W
61°W
0
25
50 miles
0
50 kilometres
Tobago
Scarborough
Caribbean Sea
11°N
11°N
TRINIDAD AND TOBAGO
VENEZUELA
Mt. Aripo 940m
Tunapuna
Arima
Port of Spain
Gualco
Trinidad
ATLANTIC OCEAN
Gulf of Paria
Rio Claro
San Fernando
Siparia
10°N
VENEZUELA
62°W
©Oxford Cartographers

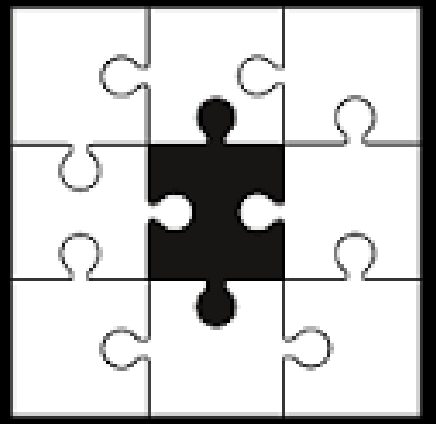
MY RESEARCH



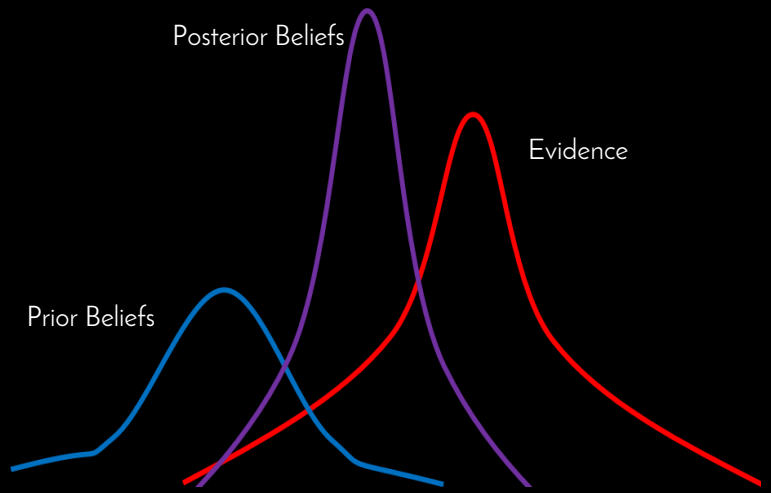
Latent Variable Modelling



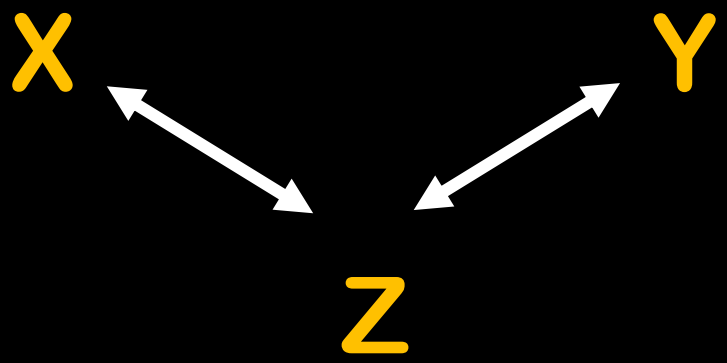
Longitudinal Data Analysis



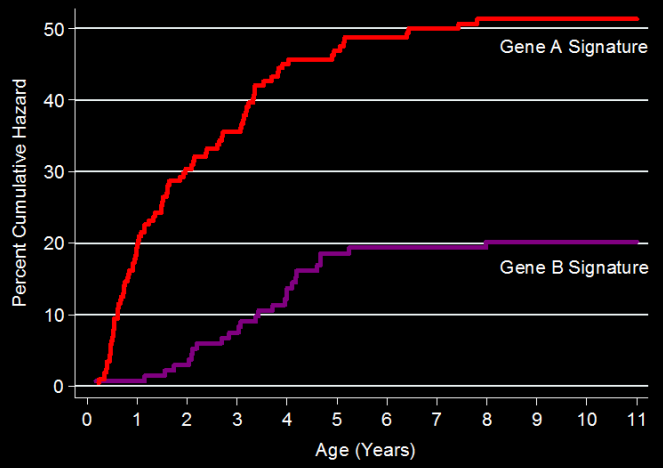
Missing Data



Bayesian Data Analysis



Causality



Survival Models

GENERAL STRUCTURE OF THIS TUTORIAL

Some **Ground Rules**: Laying the Basis

Motivation and Framework: **Endotype Discovery**

Focus: **Learning by Example**

Basic principles of **Causality**

Tips for **Team Science**

ELEMENTS OF THE PROJECT CYCLE

Understand the problem

Understand the data

Prepare the data

Evaluate Algorithms – Cross Validation

Finalise Models

WARNING: LITTLE FOCUS ON DEEP LEARNING

Deep Leaning gives excellent results on web-scale and image datasets

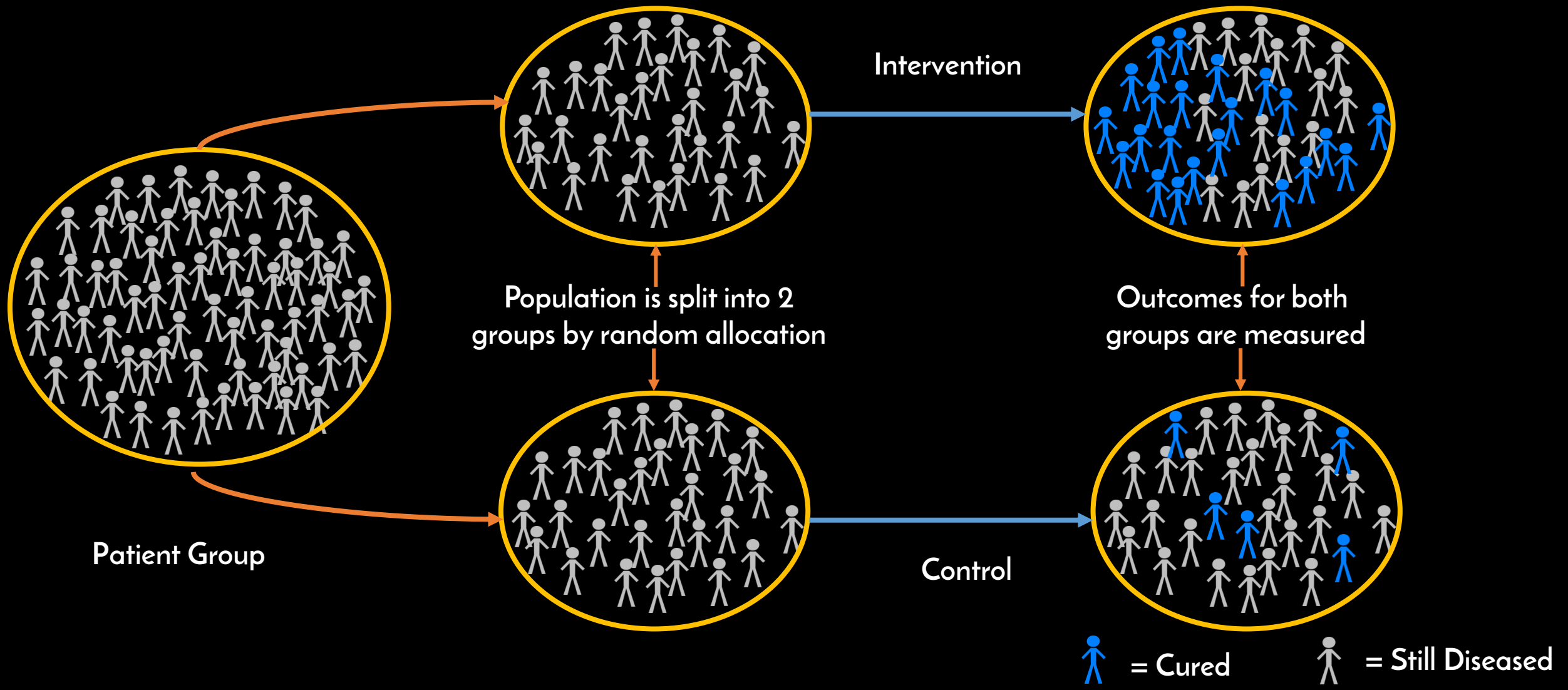
DL is very data hungry
Health data collection is (generally) expensive

Difficult to represent uncertainty

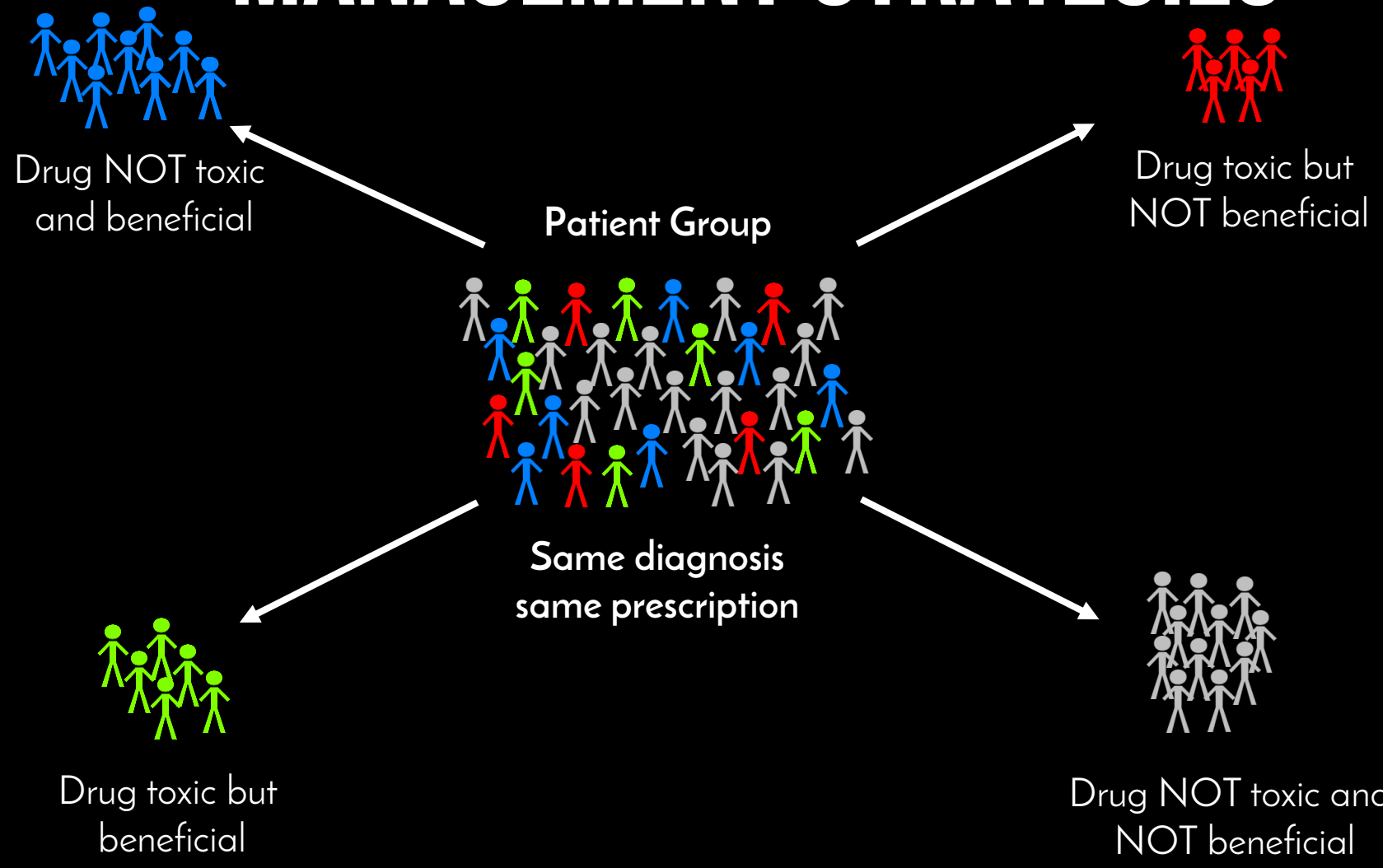
Interpretability

Model-Based approaches: **Focus on hypothesis generating**

RANDOMISED CONTROL TRIAL: TRADITIONAL APPROACH TO EVALUATING TREATMENT



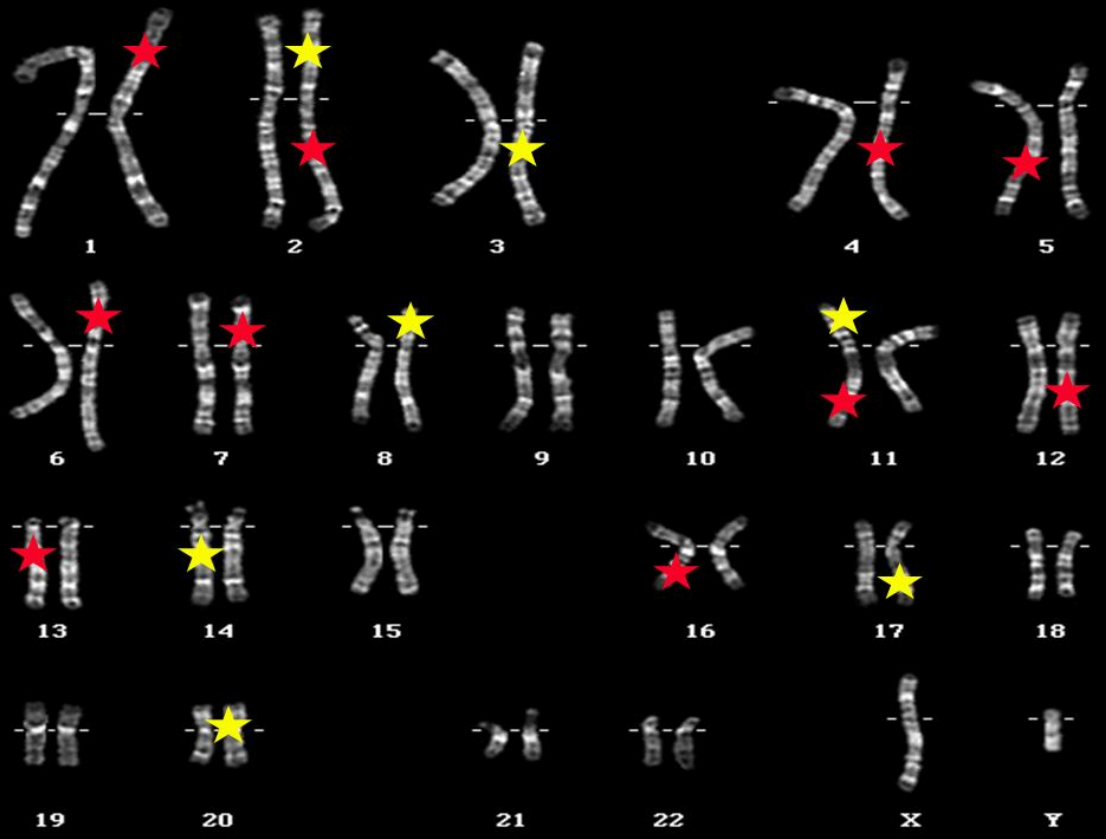
NEED FOR PERSONALIZED TREATMENT AND MANAGEMENT STRATEGIES



GENETICS: LOW YIELD

Legacy of non-replicated genetic epidemiology, typical of most common chronic disorders

- ★ Linkage in 1 study only
- ★ Linkage in >1 study



ENDOTYPE DISCOVERY: THE GRAND CHALLENGE

To identify **subgroups** of complex disease risk or treatment outcome explained by a **distinctive underlying mechanism** ("endotypes")

Foundation of **Stratified Medicine** - seeking better-targeted interventions

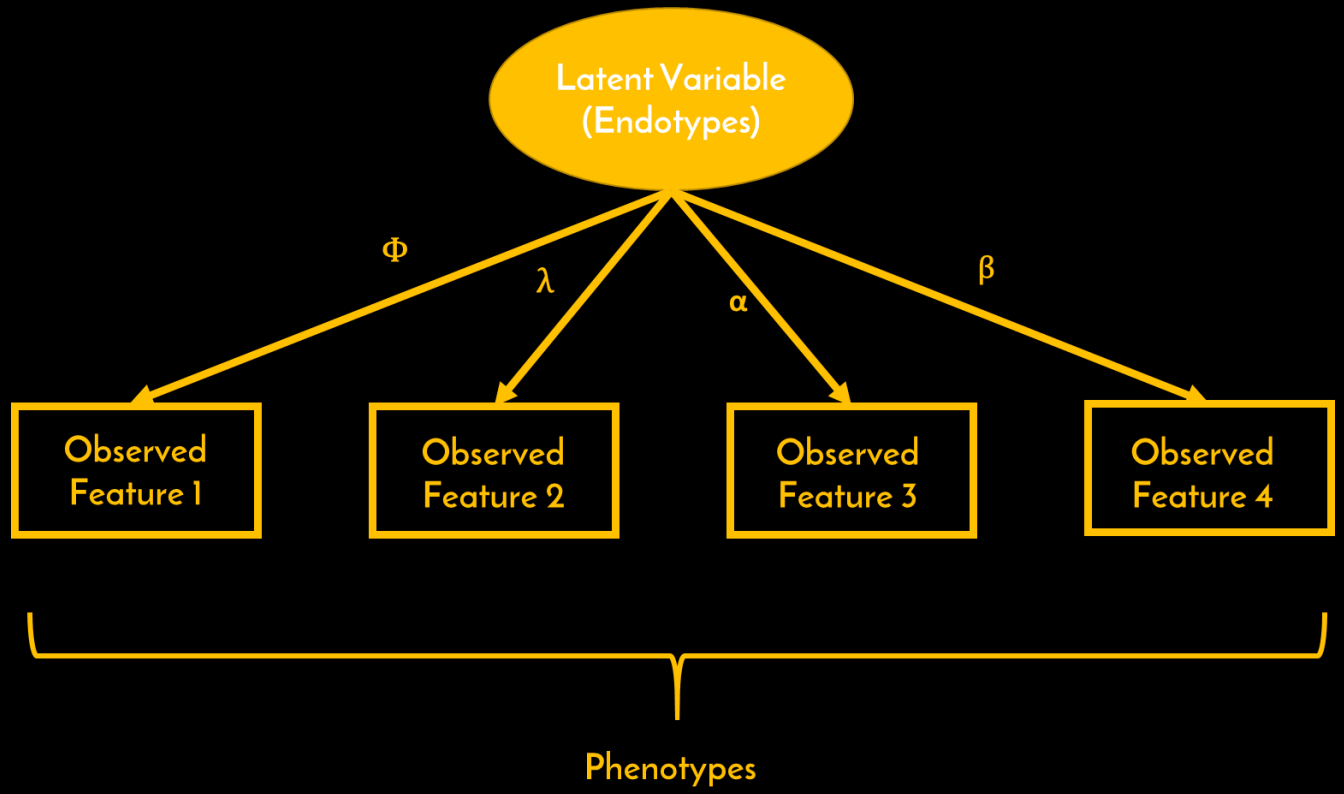
"We adore chaos because we love to produce order"

M.C. ESCHER

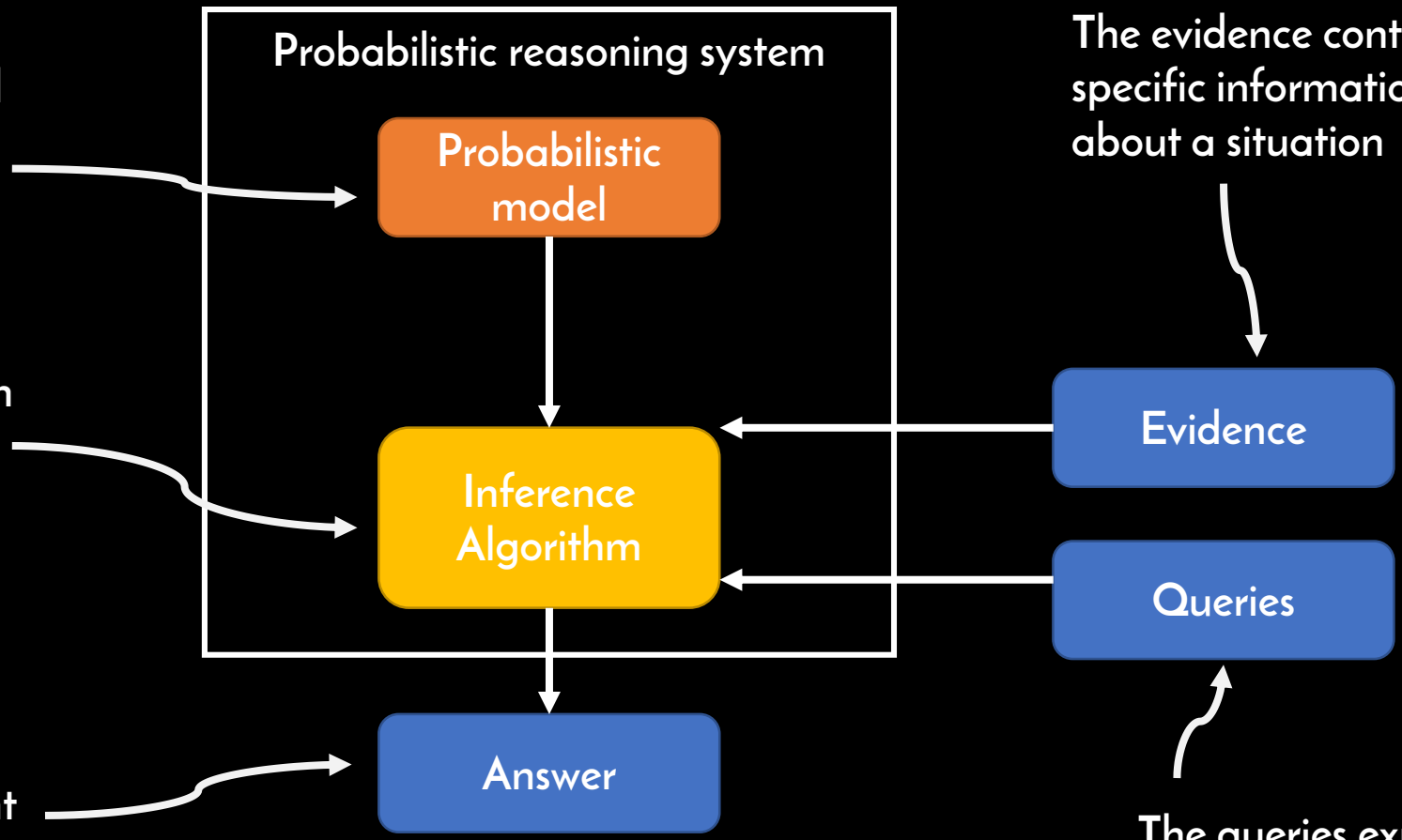
ORDER AND CHAOS, 1950

GENERALIZED FRAMEWORK FOR IDENTIFYING DISEASE ENDOTYPES

Parsimonious description of the data inferred from what is observed



PROBABILISTIC PROGRAMMING:
TOOL FOR IDENTIFYING LATENT STRUCTURE



The probabilistic model expresses general knowledge about a situation

The inference algorithm uses the model to answer queries given evidence

The answers to queries are framed as probabilities of different outcomes

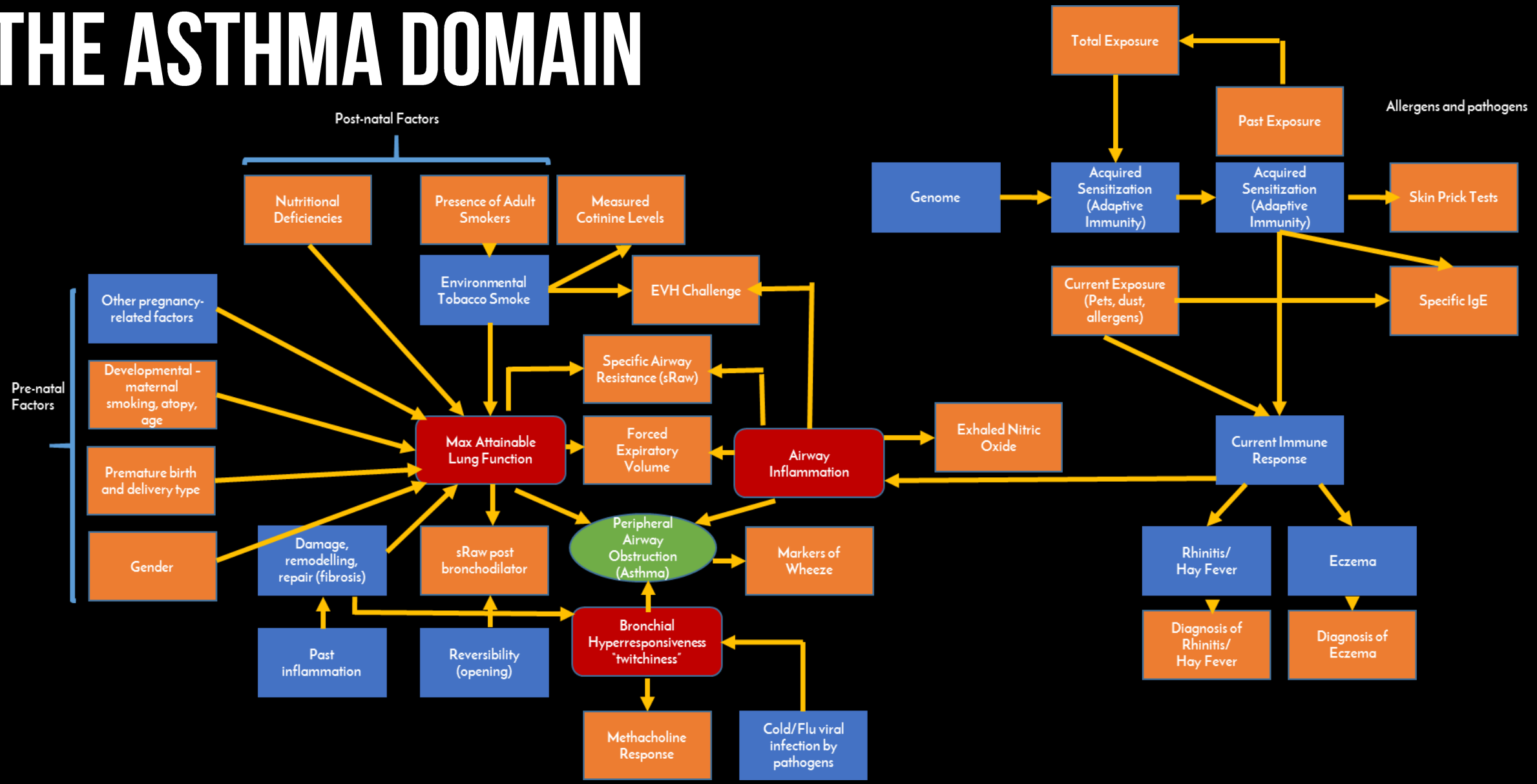
The evidence contains specific information about a situation

The queries express the things that will help you make a decision

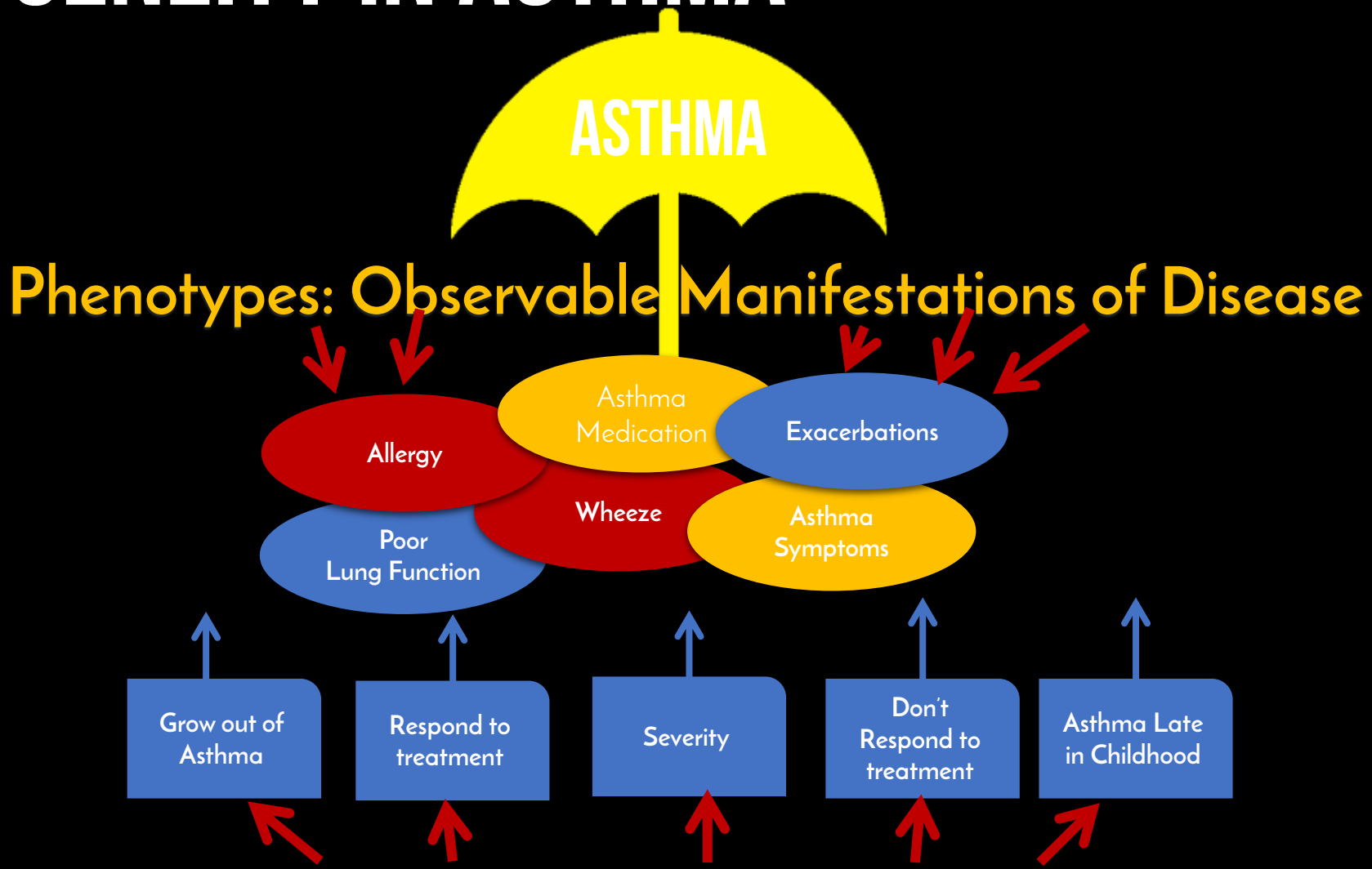
The basic components of a probabilistic reasoning system

Adapted from Pfeffer, Avi. "Practical probabilistic programming." *International Conference on Inductive Logic Programming*. Springer Berlin Heidelberg, 2010.

THE ASTHMA DOMAIN



HETEROGENEITY IN ASTHMA

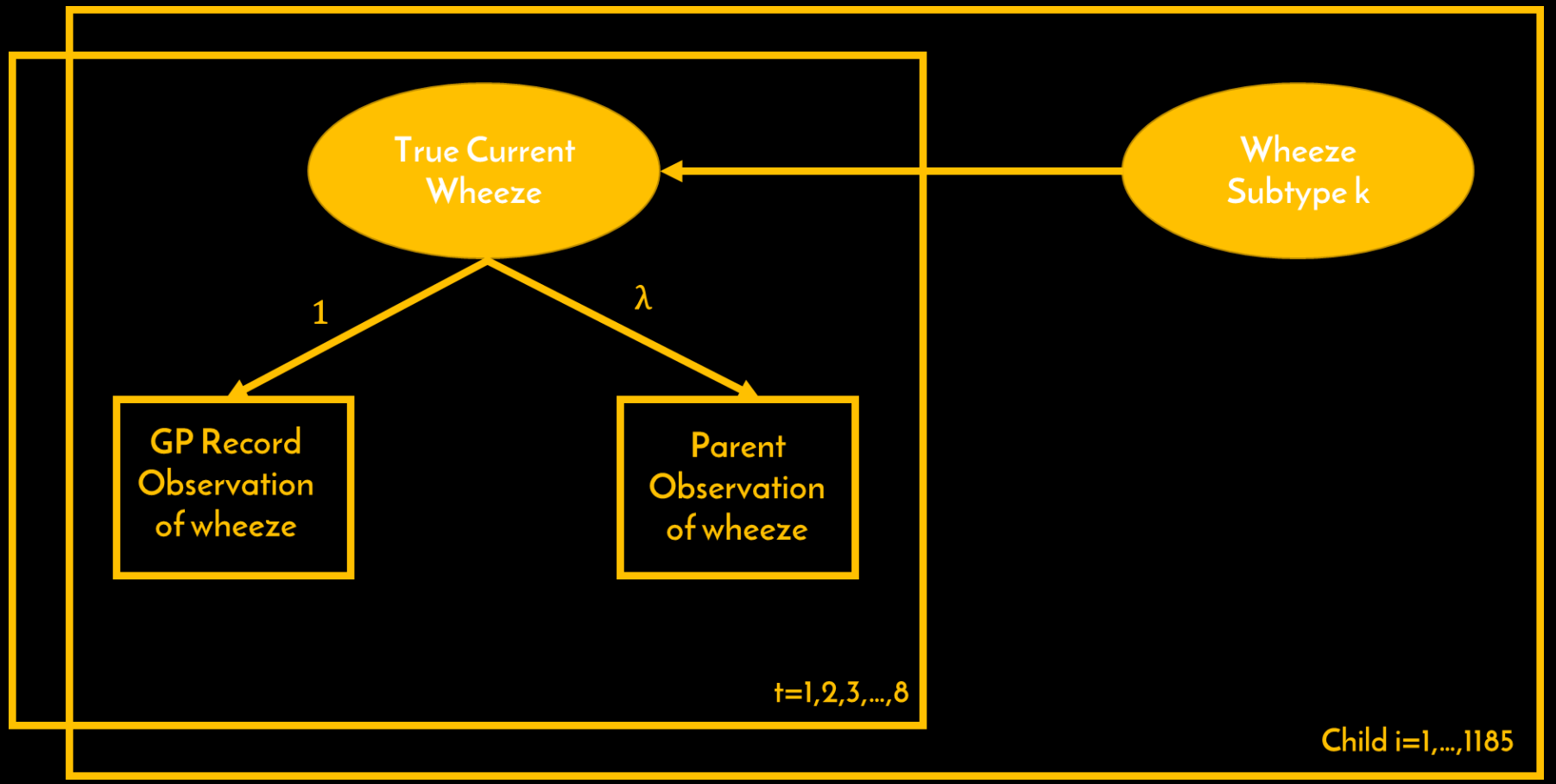


THE PROBLEM SPACE

To **define asthma subgroups (endotypes)** in a population-based birth cohort study **based on *both* parental reports and primary care consultation of wheeze** within the first 8 years of life

To **identify distinct genetic and physiological markers** which are associated with these phenotypes

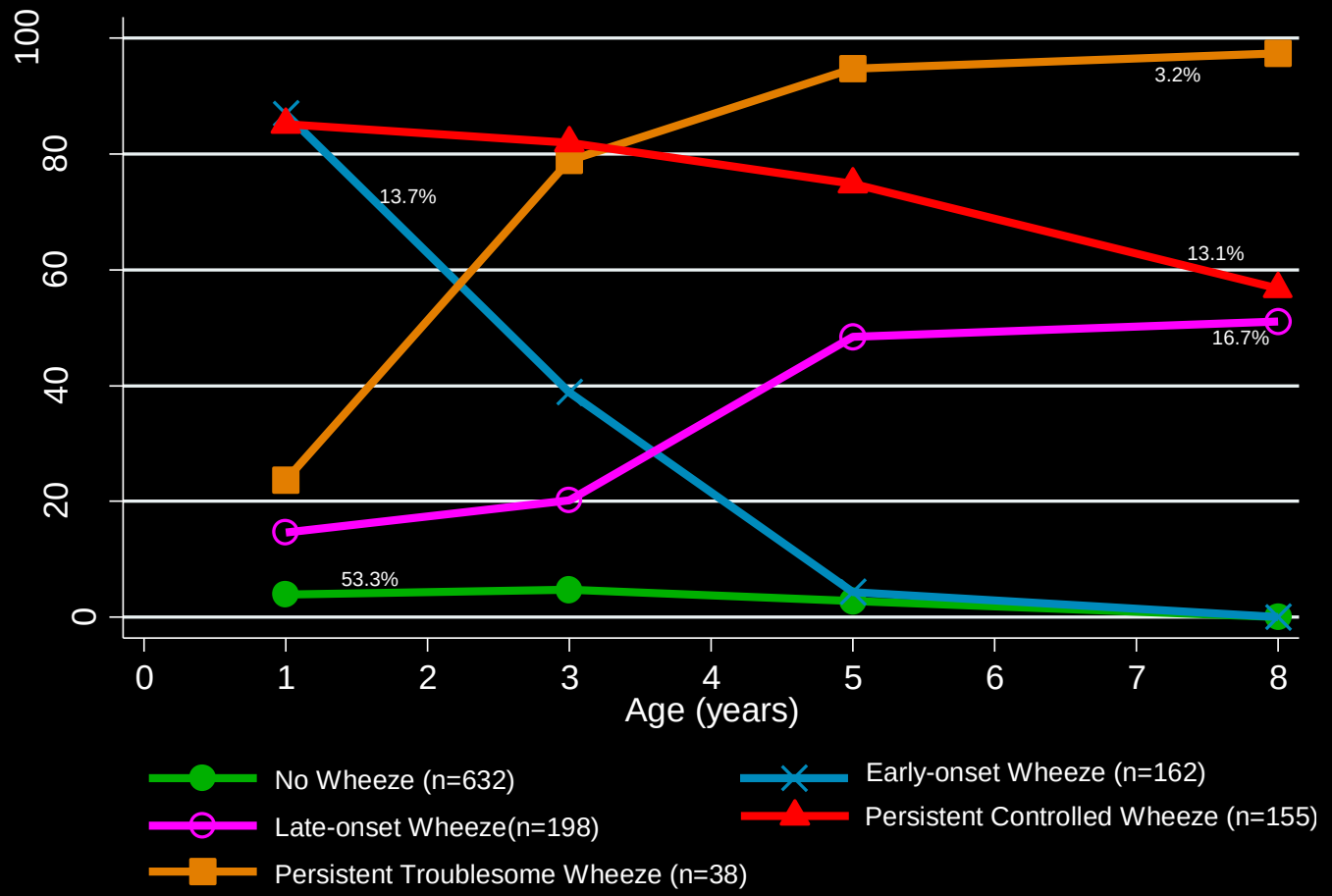
MODELLING STRATEGY FOR WHEEZE SUBTYPES



$Pr(y_{ij} = 1 | x_{ij}, c_i = k) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \xi_k + \lambda_k age + \varphi_k age^2$

x_{1i}= age; x_{2ij} rater at time j; x_{3ij} is gender $Pr(c_i = k)$ is multinomial over k classes and independent across children

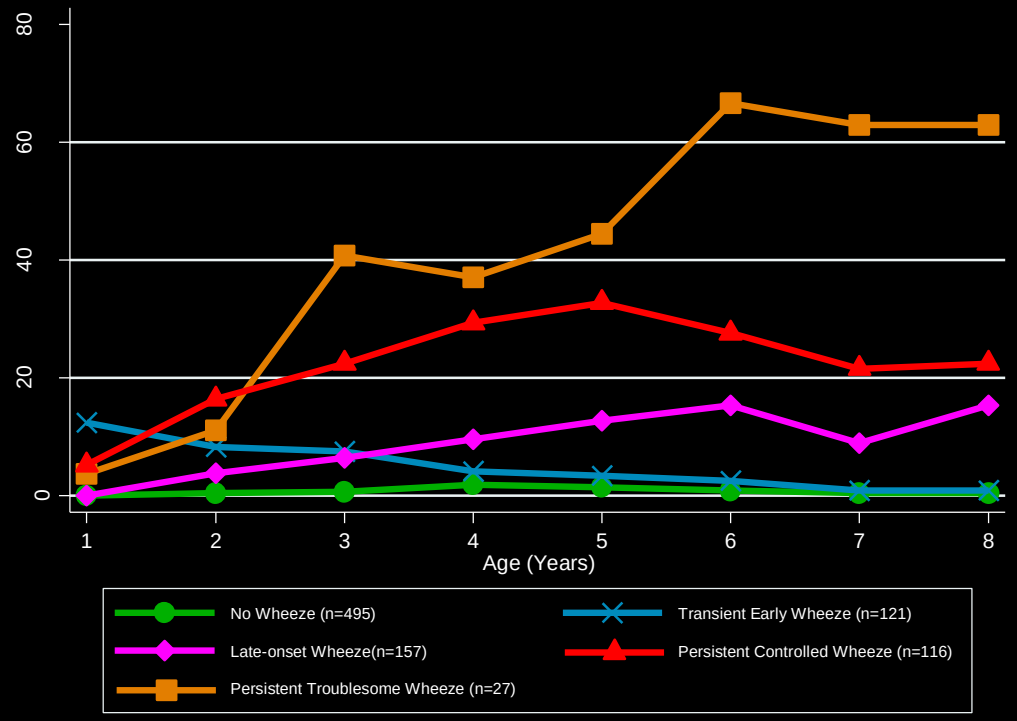
ASTHMA: A HETEROGENEOUS PHENOMENON

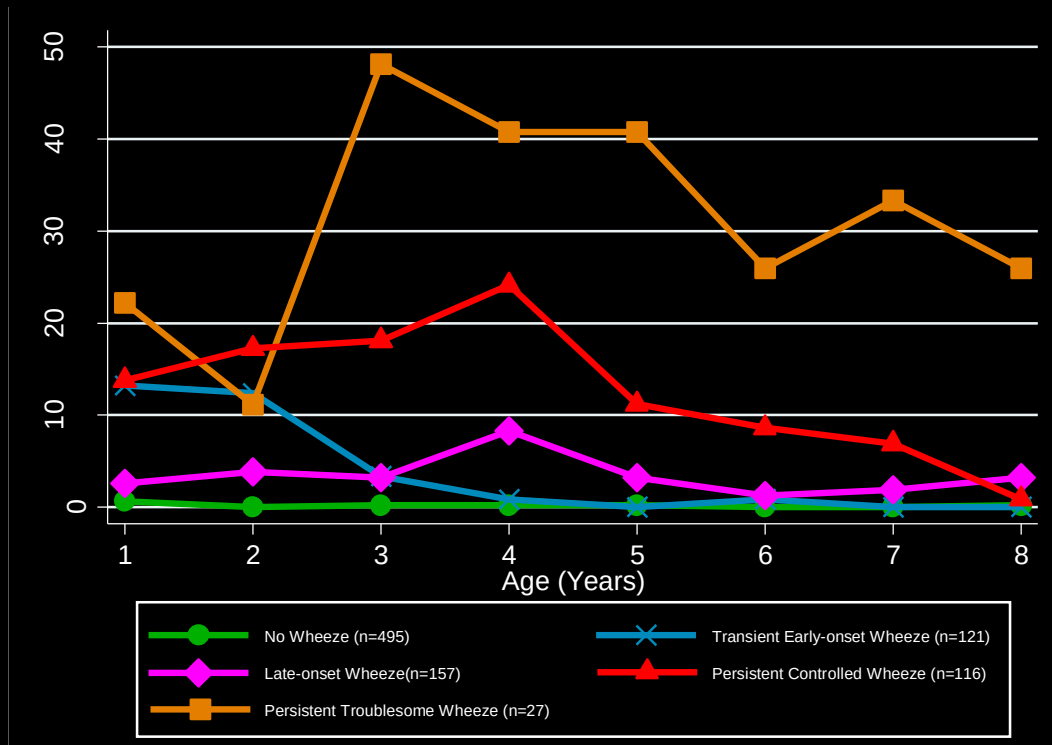


5 distinct latent classes with different genetic and environmental characteristics

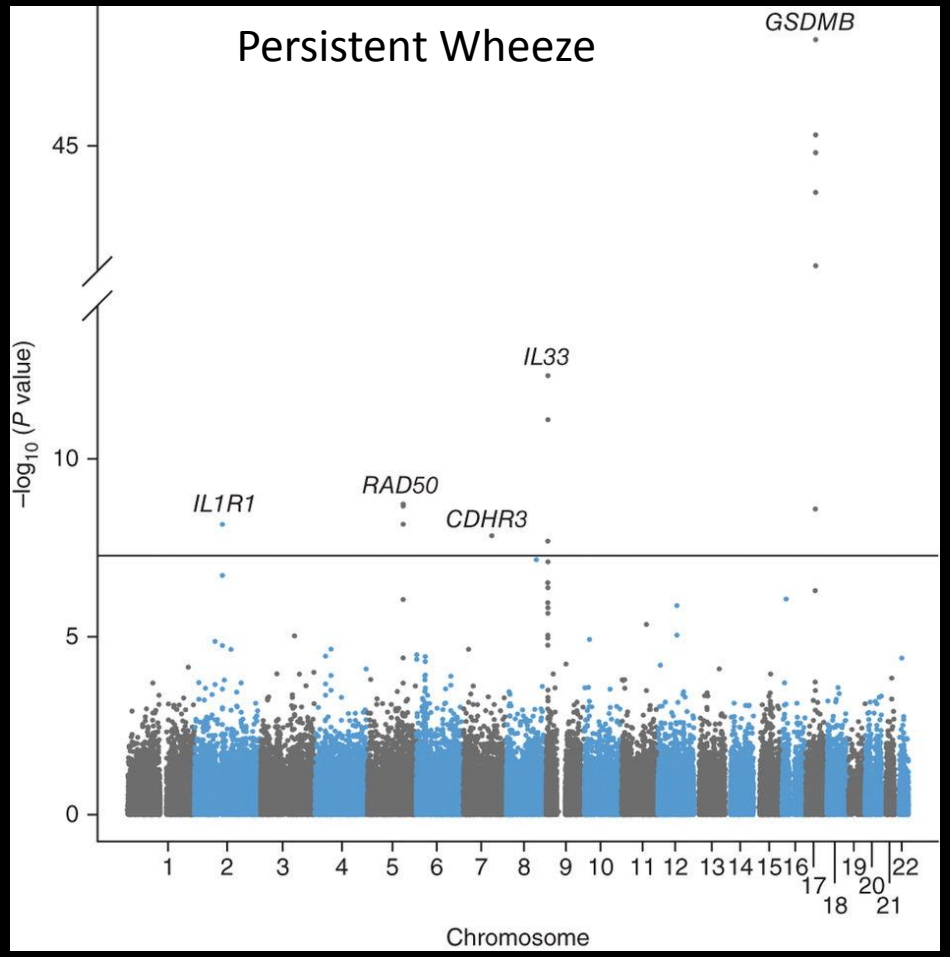
Belgrave, Danielle CM, et al. "Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing." Journal of Allergy and Clinical Immunology 132.3 (2013): 575-583.

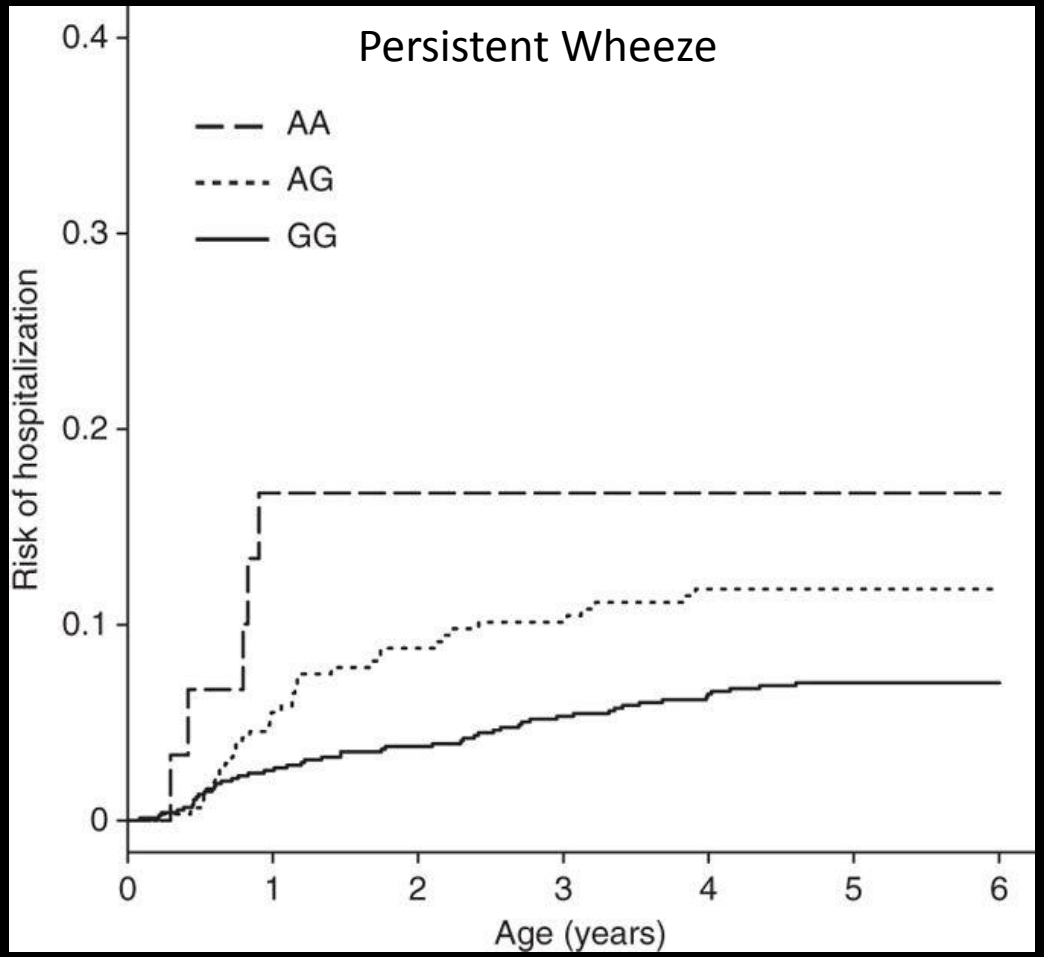
ASTHMA SUBTYPE-DEPENDENT RESPONSE TO TREATMENT





DISTINCT GENETIC PROFILE OF WHEEZE SUBTYPES





Bønnelykke, Klaus, et al. "A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations." *Nature genetics* 46.1 (2014): 51-55.

MOTIVATING ENDOTYPE DISCOVERY

Endotype discovery may have major implications for

Refining disease diagnosis

Identifying biomarkers that allow us to understand underlying **disease mechanisms**

More **personalised treatment** and management strategies of disease

RECEIVED WISDOM: CAUSALITY IN ALLERGY

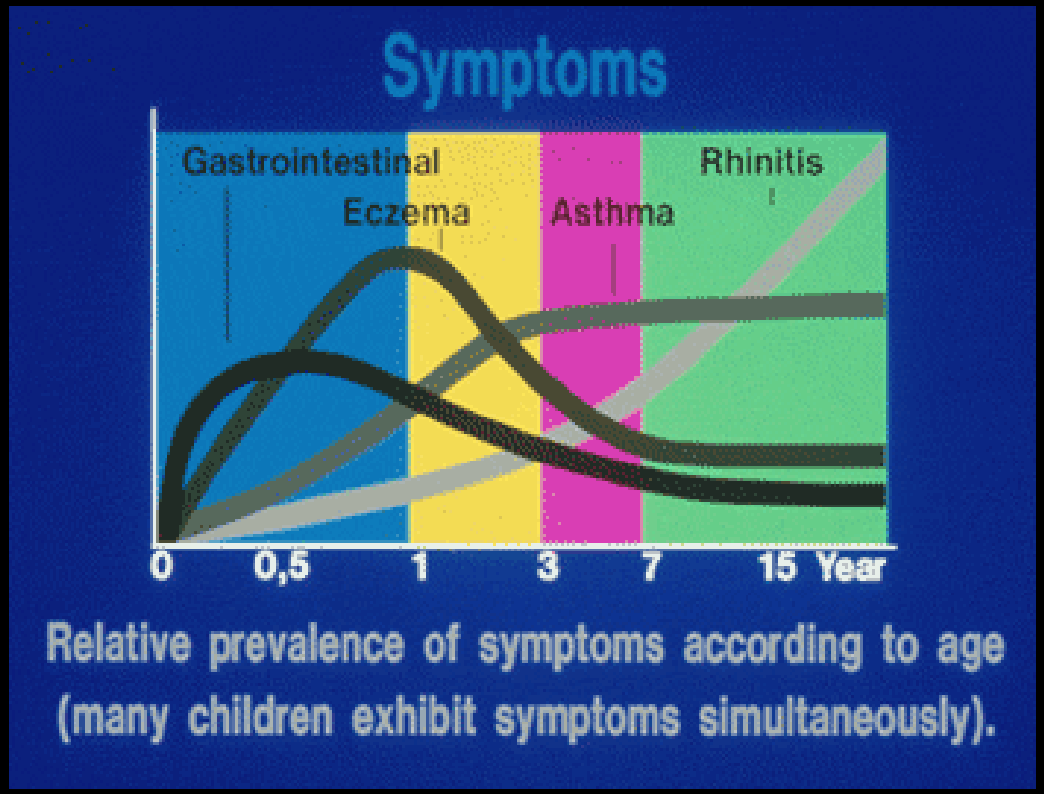
Progression of allergy:

Eczema -> Asthma -> Rhinitis

Symptoms Causally Linked

Prevention strategy:

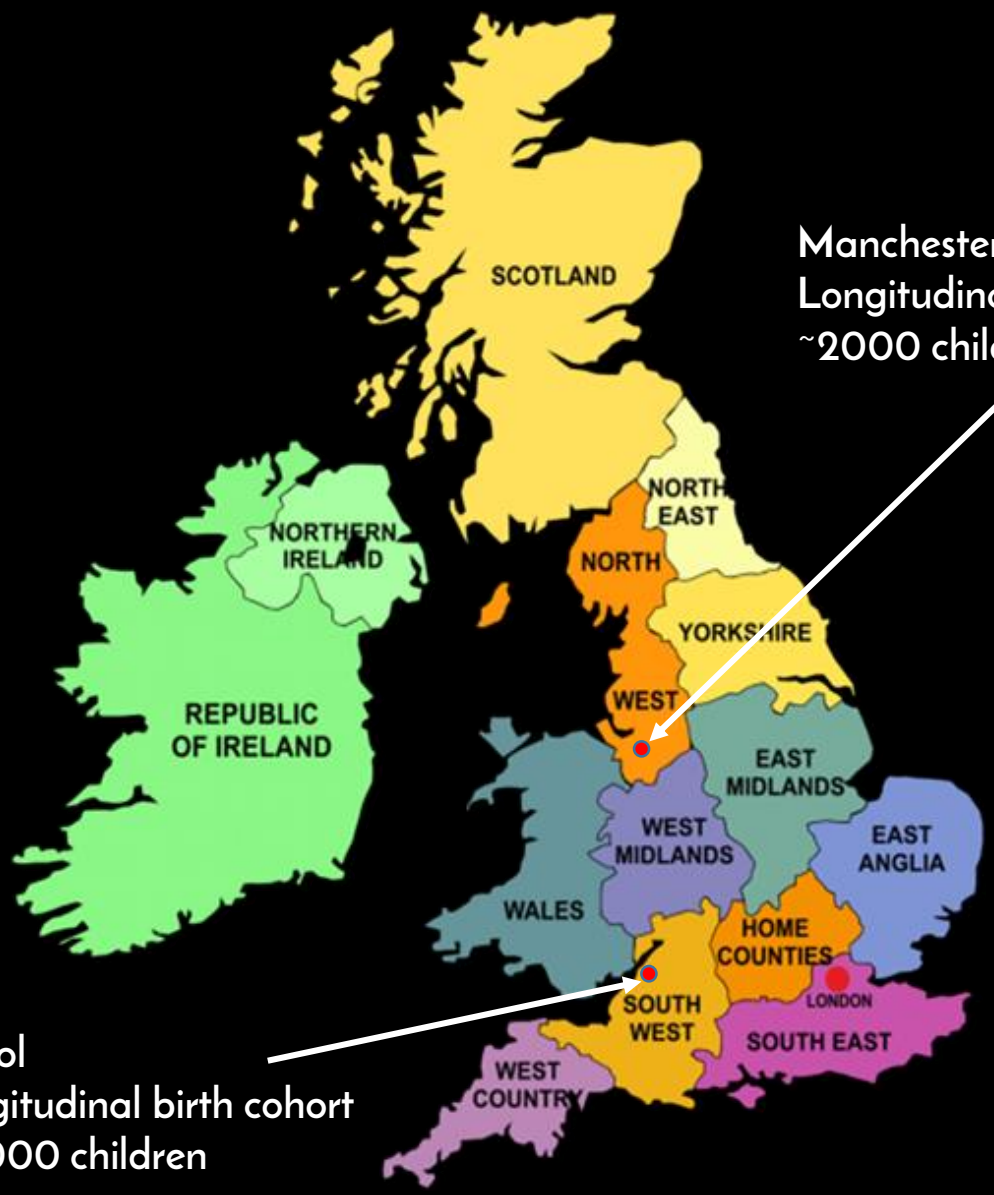
Target children with eczema to reduce progression to asthma and rhinitis



OBJECTIVE

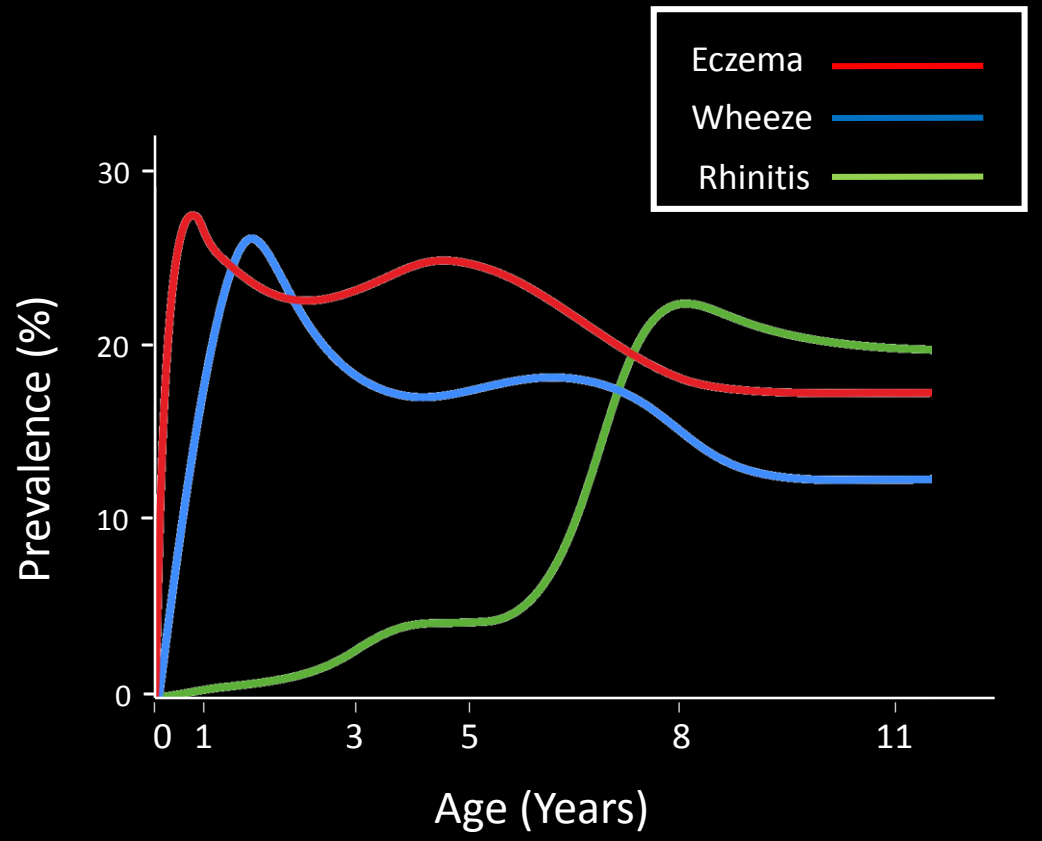
To **capture disease heterogeneity** and encapsulate different patterns of symptom progression during childhood using a probabilistic modelling approach.

THE DATA DOMAIN

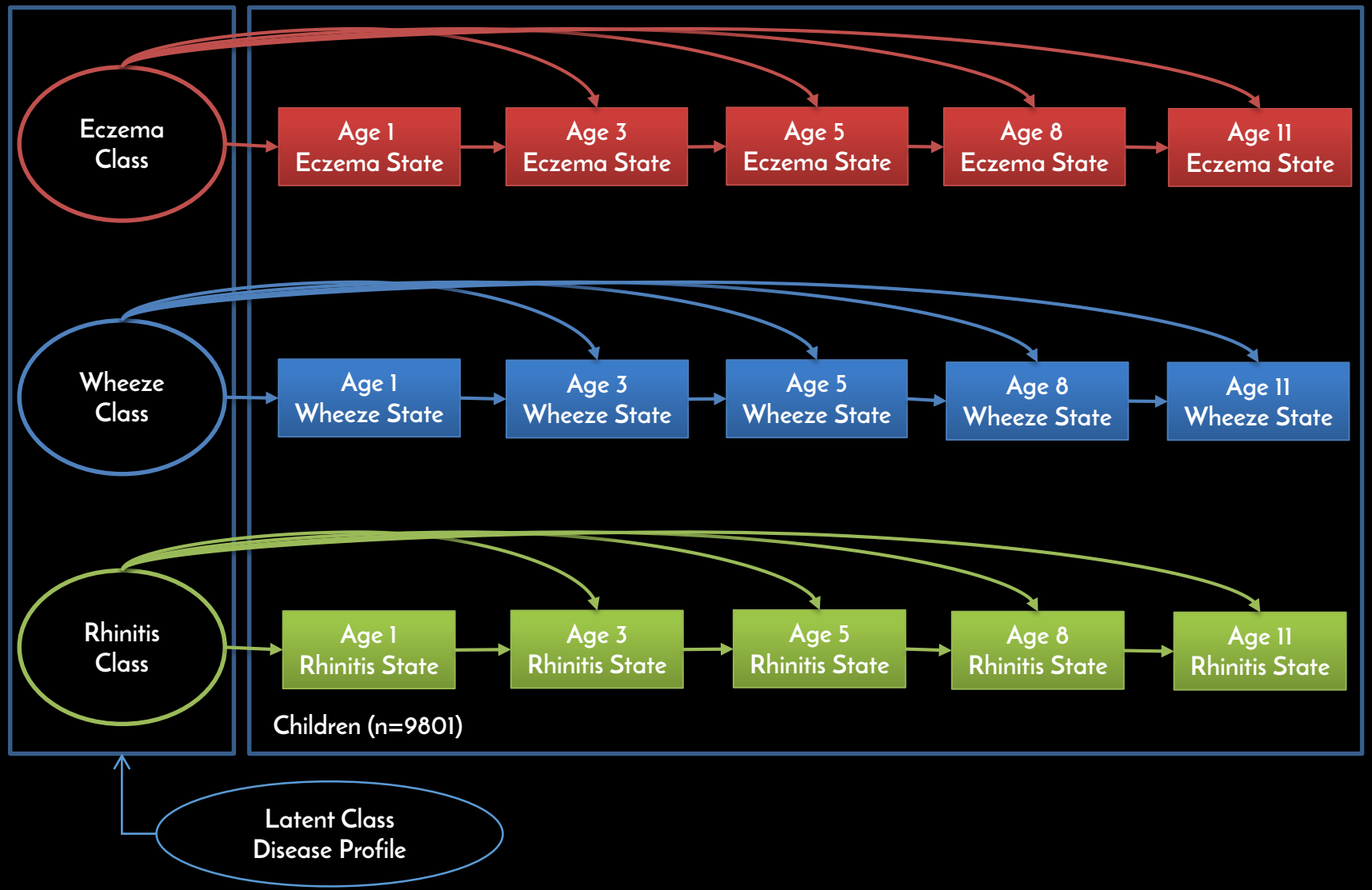


Manchester
Longitudinal birth cohort
~2000 children

Bristol
Longitudinal birth cohort
~10000 children



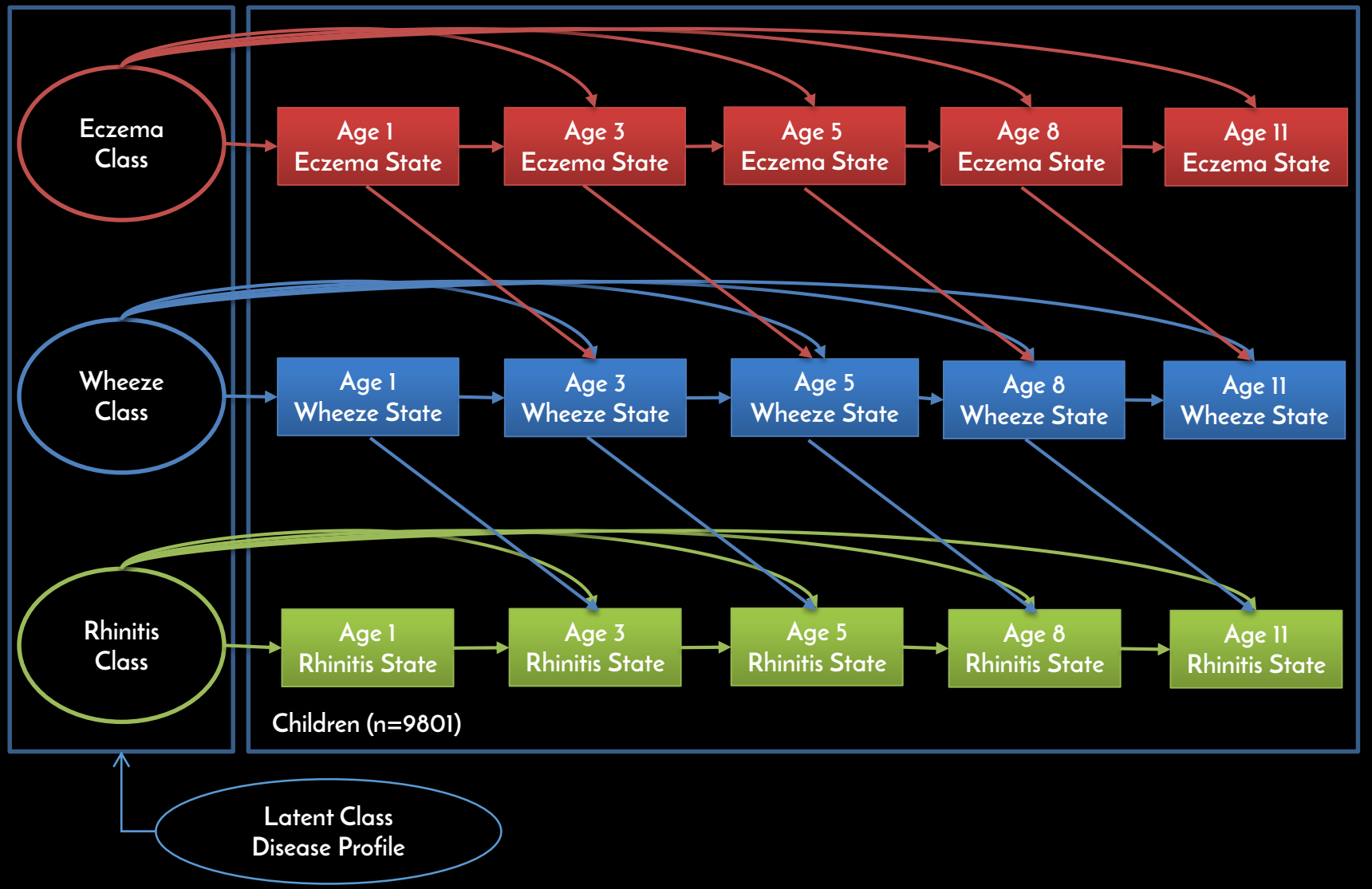
HIDDEN MARKOV MODEL 1: INDEPENDENT PROFILES



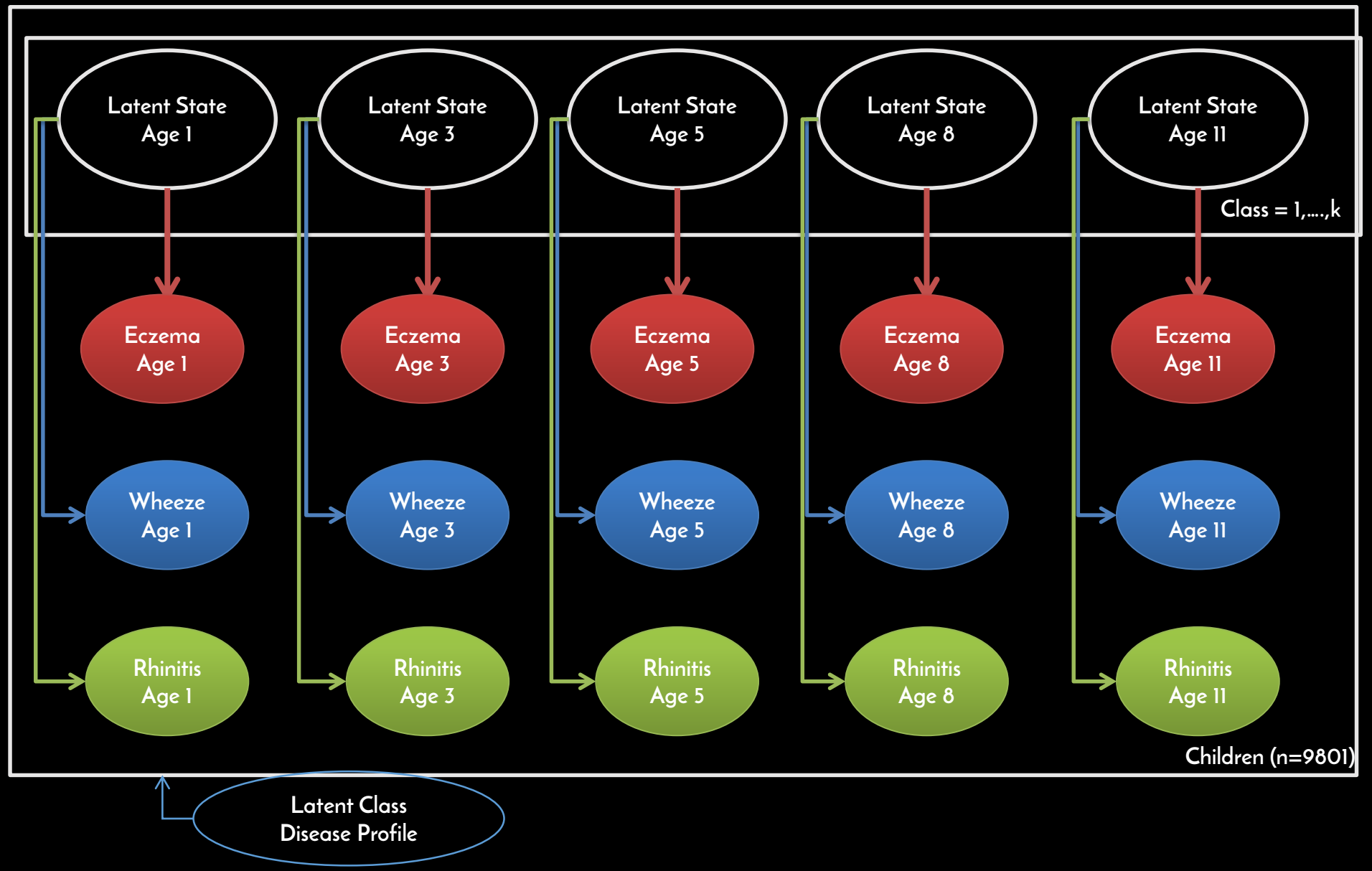
Manchester Asthma and Allergy Study
1184 subjects

Avon Longitudinal Study of Parents and Children
8665 subjects

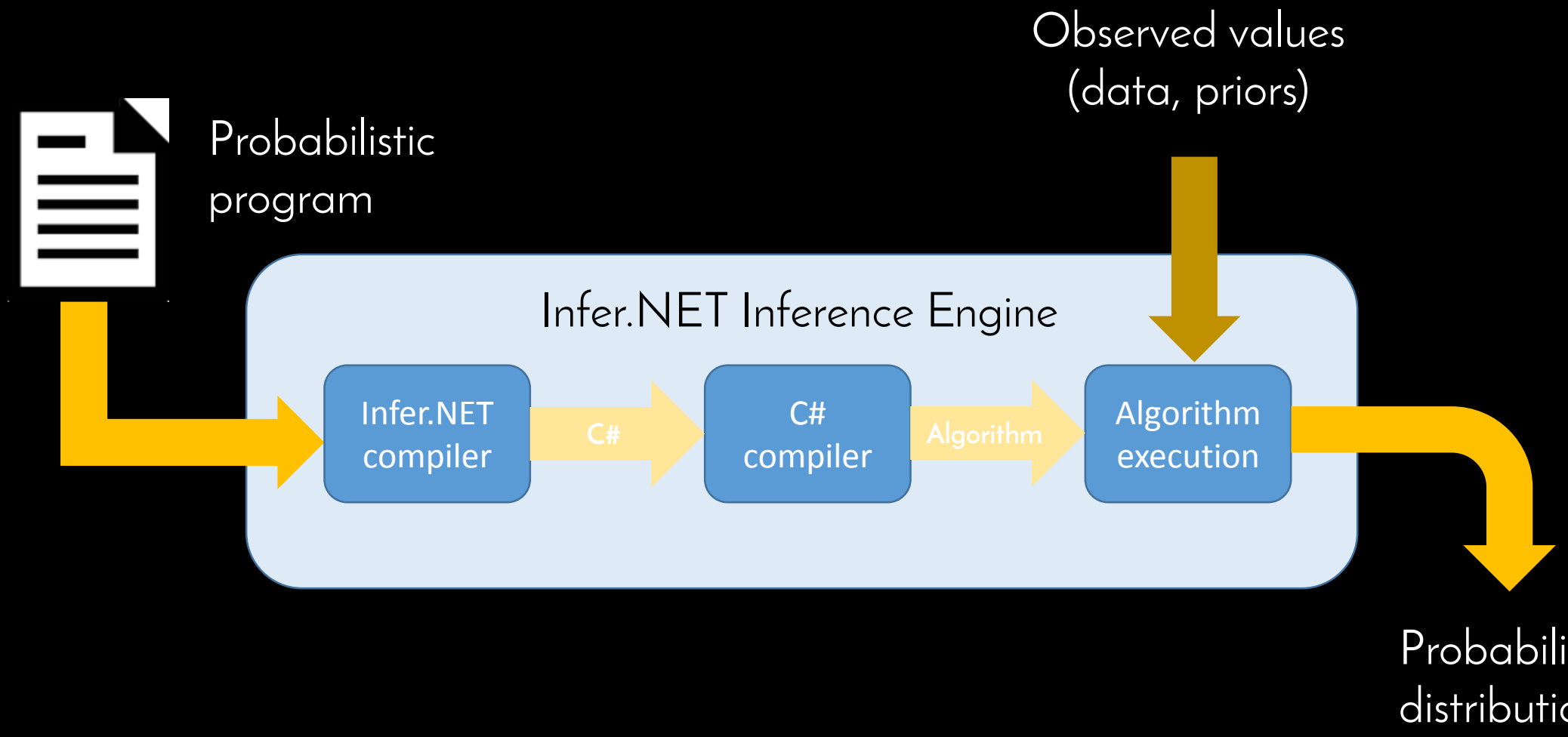
HIDDEN MARKOV MODEL 2: "ALLERGIC MARCH"



MODEL 3: LONGITUDINAL LATENT DISEASE PROFILE



INFER.NET INFERENCE ARCHITECTURE



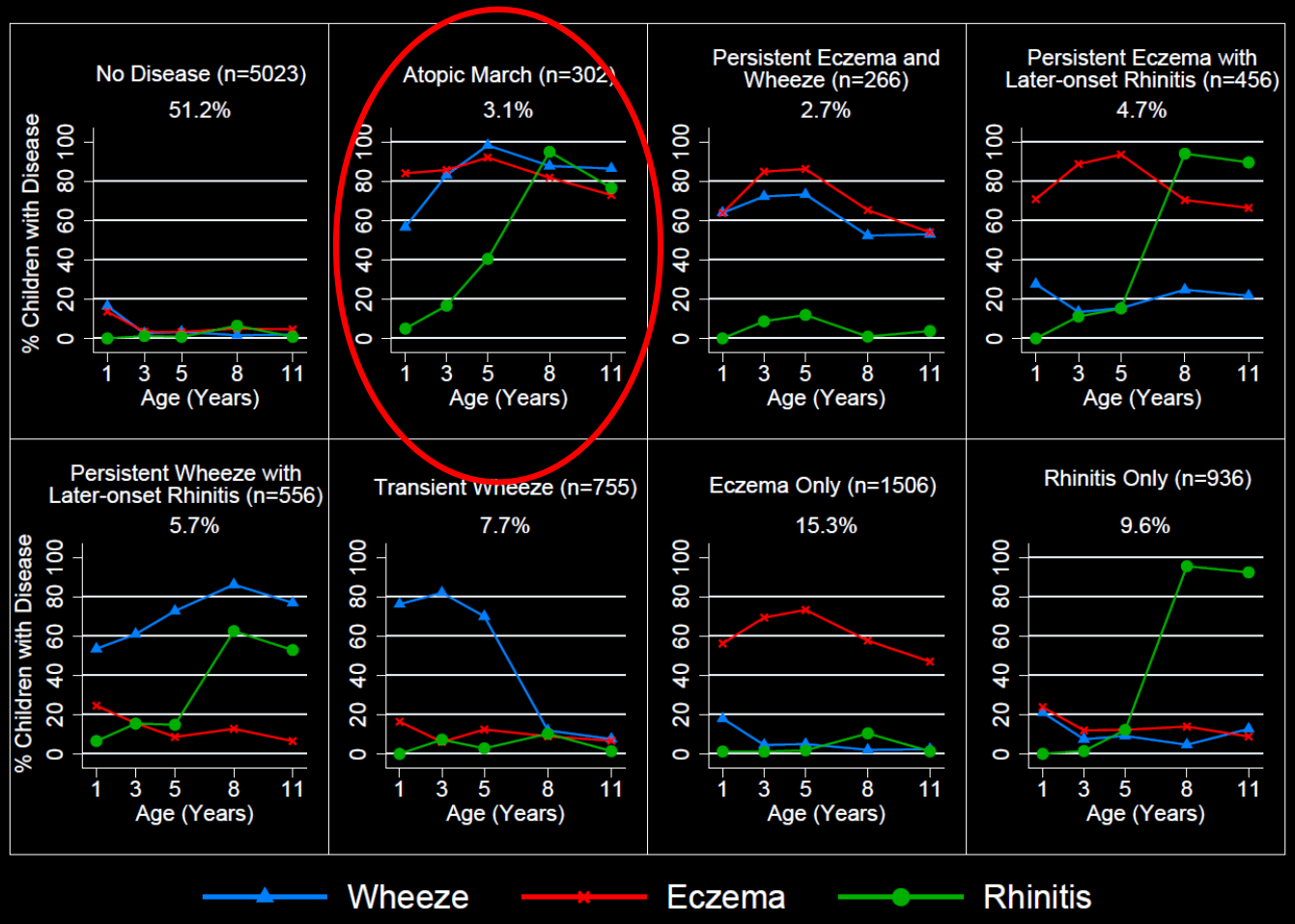
SENSITIVITY TO PRIORS

	Table of Model Evidence							
	Number of Inferred Classes							
Prior on the number of pseudo-counts	2	3	4	5	6	7	8	9
1/n	-50177	-49030	-48297	-47774	-47367	-47130	-46989	-47109*
2/n	-50200	-49104	-48310	-47797	-47357	-47143	-46994	-47334*
1	-49920	-48448	-47506	-46930	-46845	-46658	-46503	-46424*
2	-49920	-48448	-47506	-46930	-46845	-46733	-46596	-46431*

POSTERIOR PROBABILITY OF CLASS MEMBERSHIP

	Class							
1	2	3	4	5	6	7	8	
0.943	0.924	0.783	0.805	0.805	0.756	0.805	0.846	

DISAGGREGATING SYMPTOM HETEROGENEITY



From: Belgrave et al. Developmental Profiles of Eczema, Wheeze, and Rhinitis: Two Population-Based Birth Cohort Studies. PlosMedicine 2014

DISSECTING THE ATOPIC MARCH

The **Allergic March** reflects patterns at the population level, rather than the natural covariance of symptoms within individuals' life courses

Developmental profiles of **Eczema** → **Asthma** → **Rhinitis** are heterogeneous

Only a small proportion of children follow a trajectory profile similar to that of the atopic march

ANTIBIOTIC RESISTANCE: A GLOBAL PROBLEM

P. DUREUIL/PHOTOALTO/AGE FOTOSTOCK

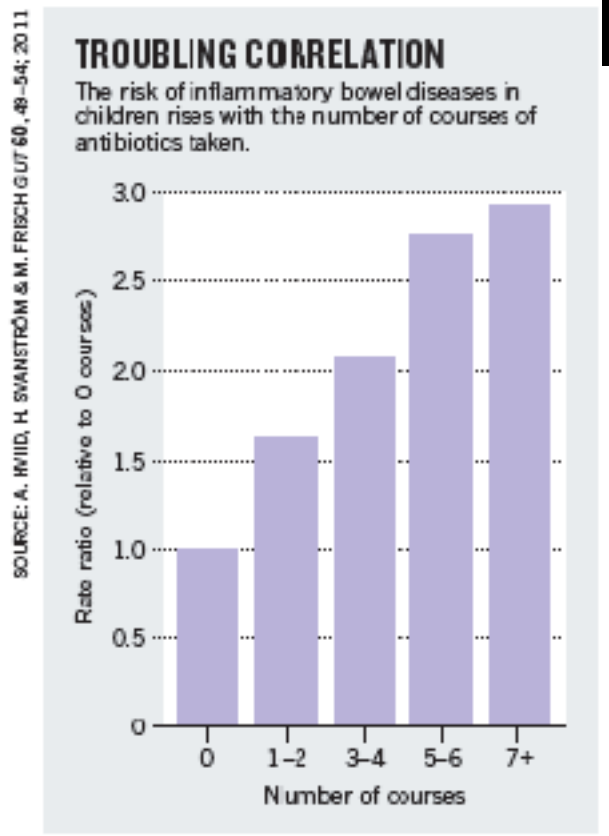
Dosed up: could excessive prescription of antibiotics be hampering children's ability to fight disease?

Stop the killing of beneficial bacteria

Concerns about antibiotics focus on bacterial resistance — but permanent changes to our protective flora could have more serious consequences, says **Martin Blaser**.

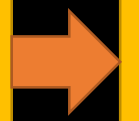
- Average child in developed countries takes 10-20 courses of antibiotics before age 18 yr

Blaser. 2011. Nature. 476:393



Factors affecting Early Respiratory Colonisation

- Antibiotics
- Delivery Method
- Bowel Colonisation
- Nasopharyngeal Colonisation
- Feeding Method



Neonatal Airway and Lung Bacterial Colonisation

Impact of Early Respiratory Colonisation

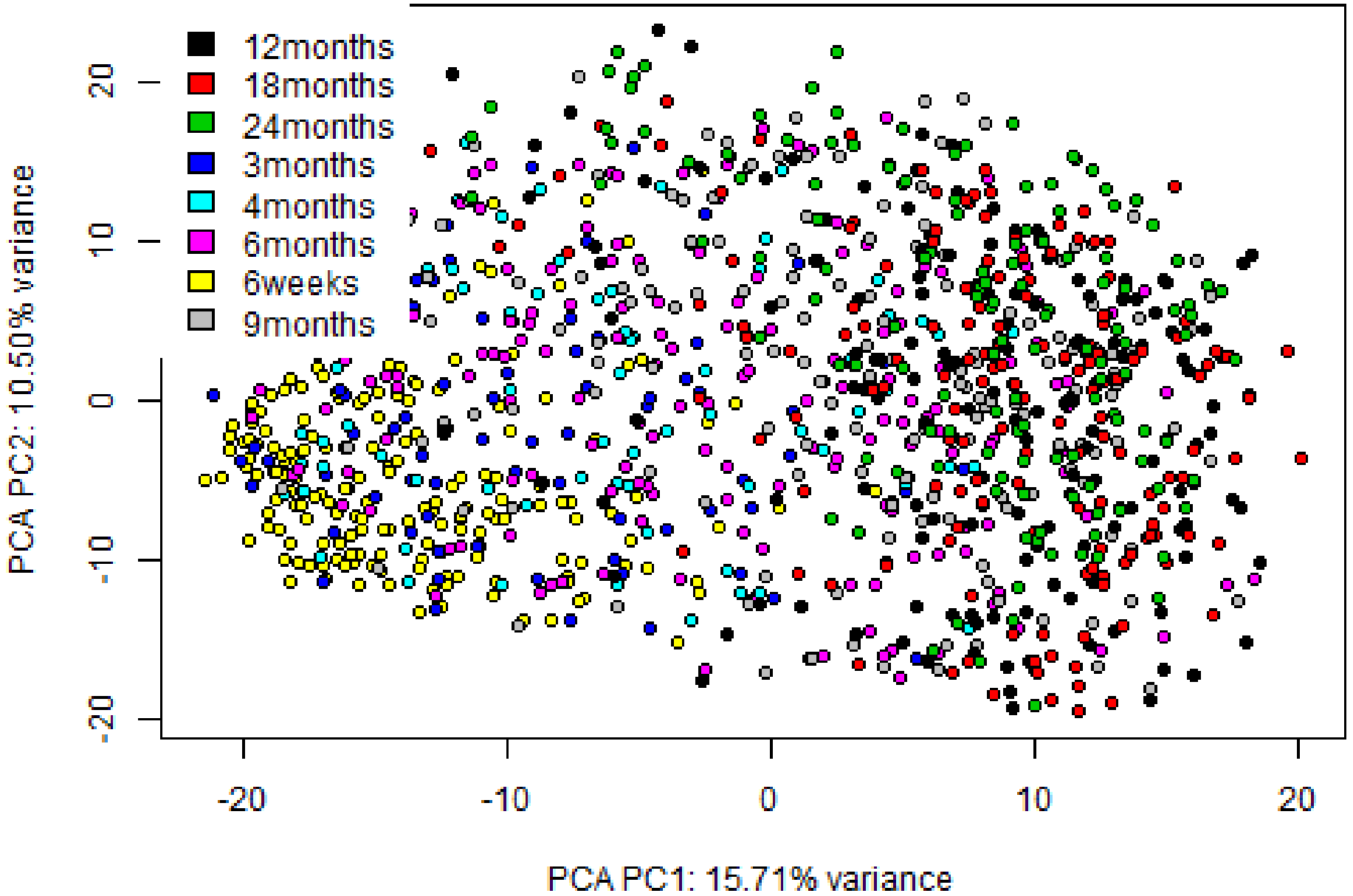
- Risk of Bronchiolitis
- Long Term Risk of Asthma
- Risk of Chronic Lung Disease in Preterm Infants

OPERATIONAL TAXONOMIC UNIT (OTU)

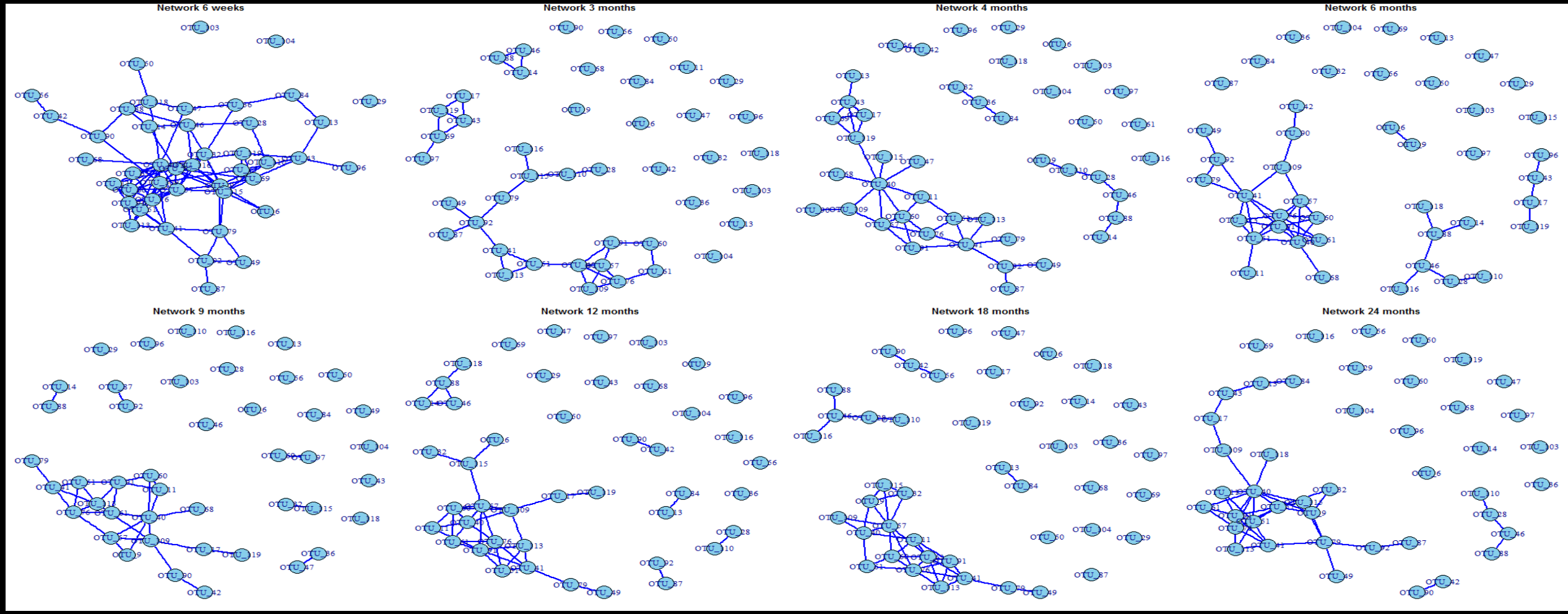
#OTU ID	D53~DRun8~24moSwab	D283~DRun15~24moSwabs	D173~DRun15~24moSwabs	D131~DRun15~24moSwabs	D225~DRun15~24moSwabs	D98~DRun15~24moSwabs
New.ReferenceOTU75	0	0	0	0	0	0
New.ReferenceOTU76	6	5	2	1	4	3
New.ReferenceOTU77	0	0	0	0	0	0
New.ReferenceOTU8	64	14	20	23	57	96
New.ReferenceOTU9	0	0	0	0	0	0
New.ReferenceOTU0	0	14	6	8	0	2
New.ReferenceOTU1	0	0	26	0	0	0
New.ReferenceOTU2	0	0	0	0	0	0
New.ReferenceOTU113	0	0	0	0	0	0
New.ReferenceOTU4	60	78	29	8	6	5

superkingdom	phylum	class	order	family	genus	species
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	uncultured bacterium
Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Oribacterium	uncultured bacterium
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces	uncultured bacterium
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas	uncultured bacterium
Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Leptotrichiaceae	uncultured	uncultured bacterium
Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus	uncultured bacterium
Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Veillonella	uncultured bacterium
Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Actinobacillus	uncultured bacterium
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas	uncultured bacterium
Bacteria	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	Neisseria	uncultured bacterium

OTU's are used to categorize bacteria based on sequence similarity.


12months
18months
24months
3months
4months
6months
6weeks
9months
PCA PC2: 10.50% variance
PCA PC1: 15.71% variance
20
10
0
-10
-20
-20
-10
0
10
20

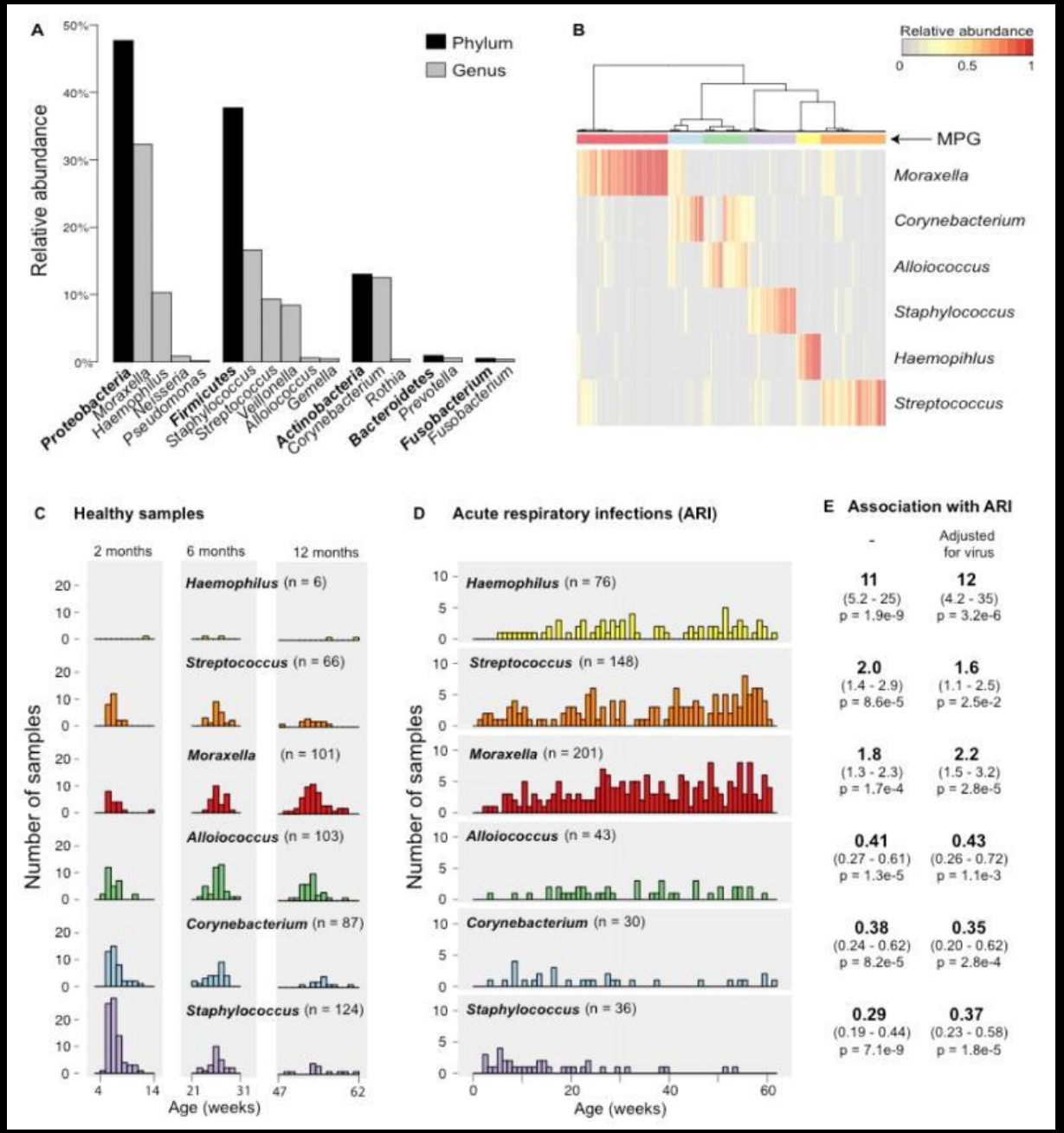
CORRELATION NETWORK ANALYSIS
EVOLUTION OF MICROBIOME PROFILE OVER TIME



MICROBIOME PROFILE AND RESPIRATORY DISEASE

Bacterial Composition of 1,021 Nasopharyngeal Aspirates Collected from 234 Infants during Periods of Respiratory Health and Disease

Clustering based on the 6 most common genera



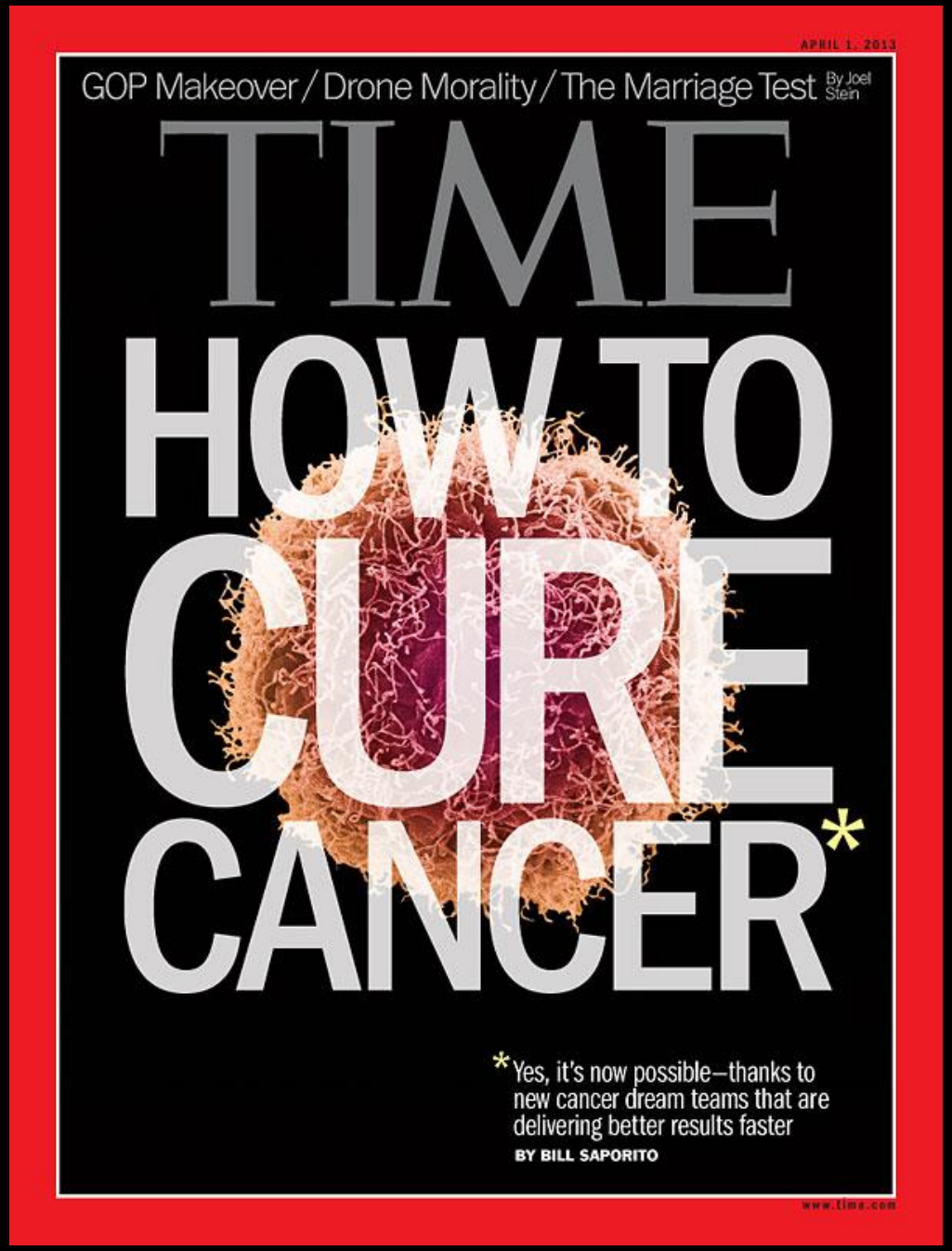
Teo, Shu Mei, et al. "The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development." Cell host & microbe 17.5 (2015): 704-715.

AN AGE-OLD PROBLEM...

12.7 million people discover they have cancer each year

7.6 million people die from cancer each year

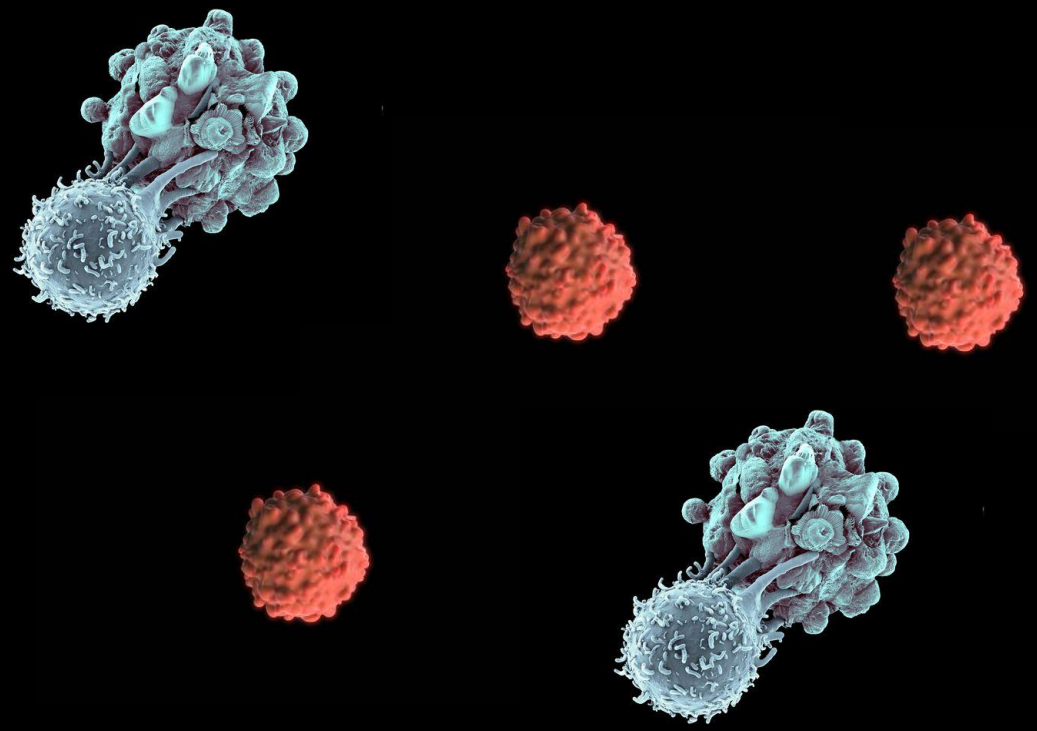
30 – 40% of these deaths can be prevented



THE PROBLEM WITH CANCER

Lack of tools for early detection and diagnosis

Cancer cells, even within the same tumor, are heterogeneous—that is, differences exist between the individual cells.



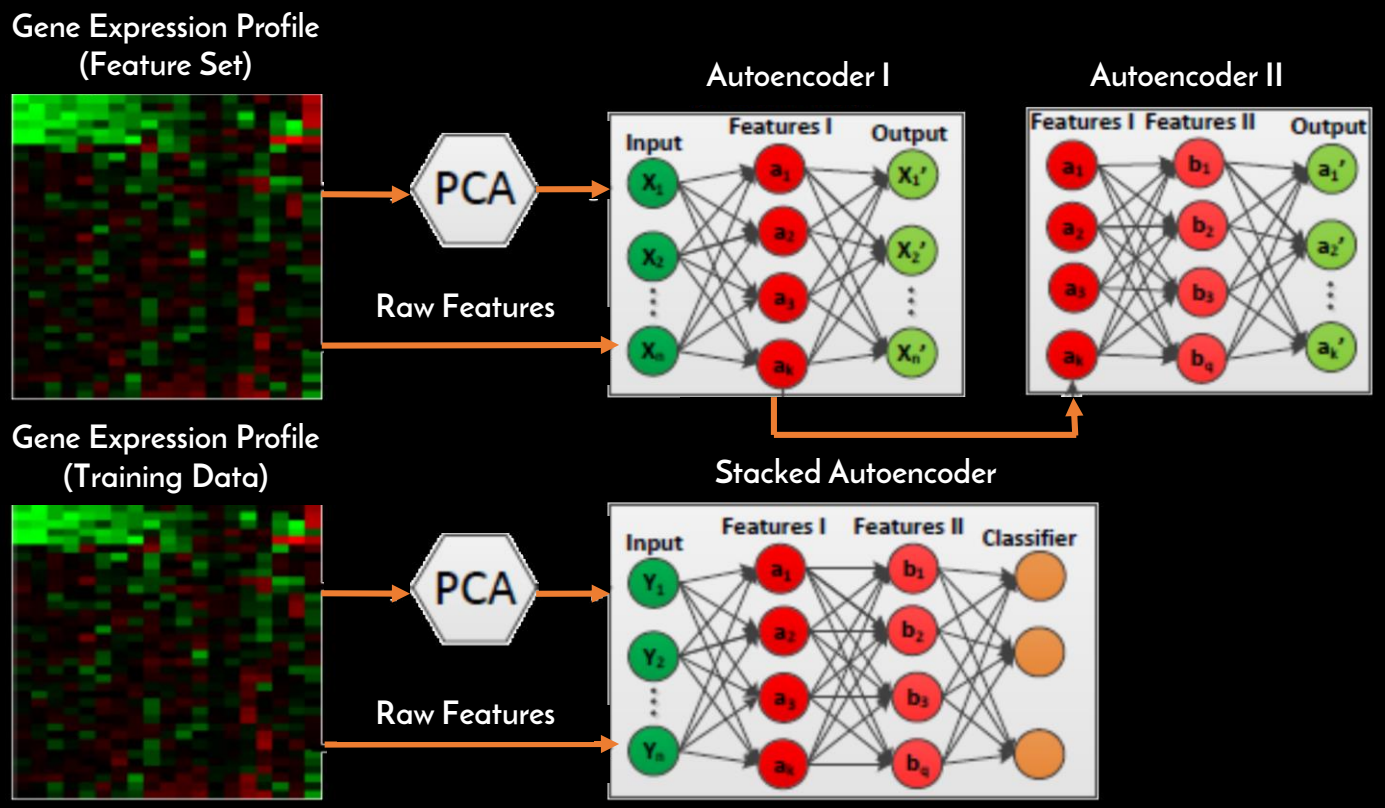
DEEP LEARNING TO ENHANCE CANCER DIAGNOSIS

Aim: To determine the difference between cancerous gene expression in tumour cells vs normal, non-cancerous tissues to obtain better insight into the disease pathology

To create a generalizable framework for new cancer types without the redesign of new features

Using Deep Learning to Enhance Cancer Diagnosis and Classification. **Rasool Fakoor**, Faisal Lahdak, Azade Nazi, Manfred Huber. ICML 2013 , WHEALTH workshop, Atlanta , GA, 2013.

CANCER DIAGNOSIS AND CLASSIFICATION



Using Deep Learning to Enhance Cancer Diagnosis and Classification. **Rasool Fakoor**, Faisal Lahdak, Azade Nazi, Manfred Huber. ICML 2013 , WHEALTH workshop, Atlanta , GA, 2013.

DELAYED INTENSIVE CARE UNIT (ICU) ADMISSION

Delayed ICU admission is correlated with mortality

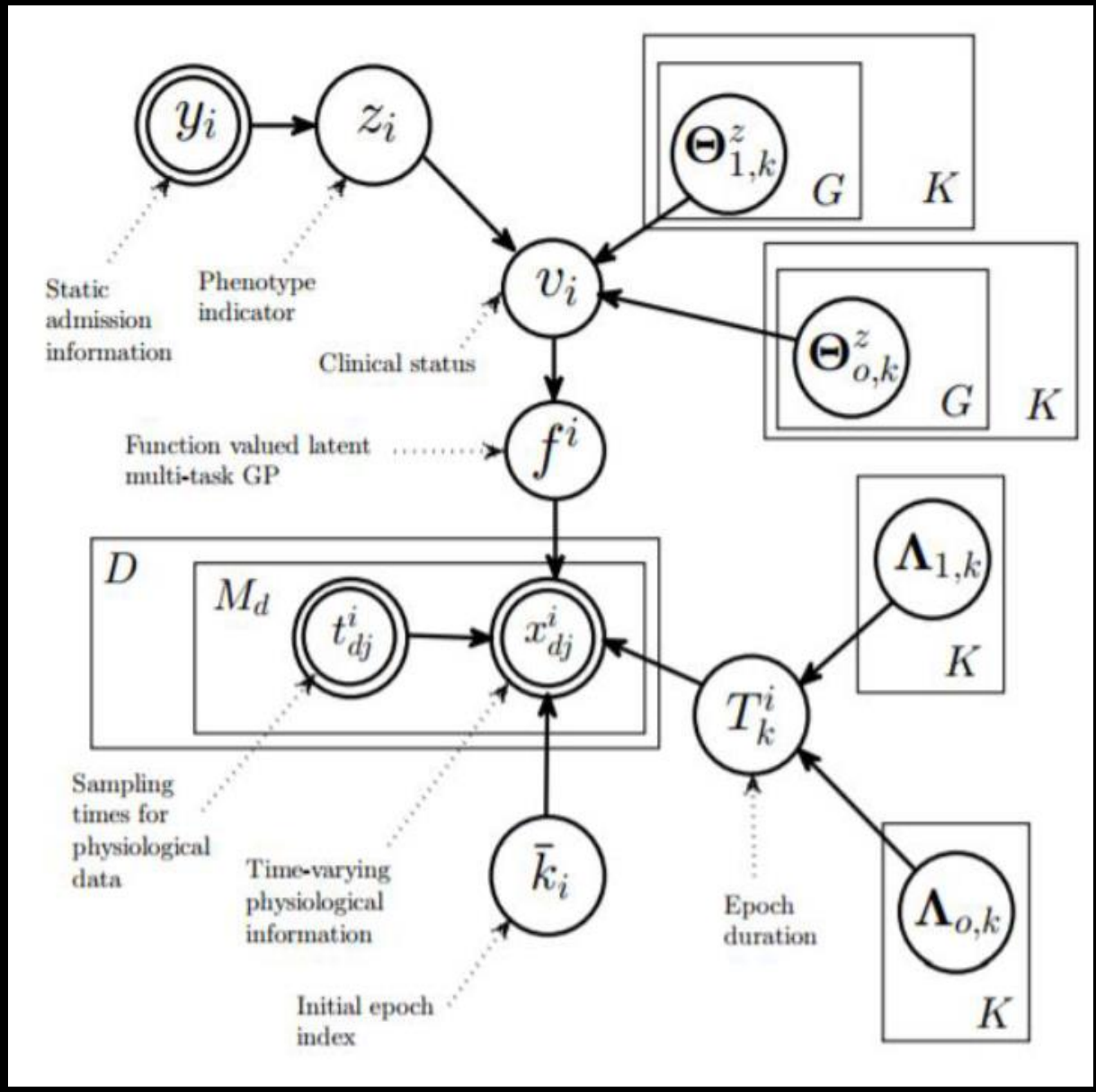
Ignoring correlations among vital signs, history and patient heterogeneity

Risk scoring methodology can confer huge clinical and social benefits on a massive number of critically ill inpatients who exhibit adverse outcomes including, but not limited to, cardiac arrests, respiratory arrests, and septic shocks.

A MULTI-TASK GAUSSIAN PROCESS MODEL FOR ICU ADMISSION

Results reflect the importance of adopting the concepts of personalized medicine in critical care settings; significant accuracy and timeliness gains can be achieved by accounting for the patients' heterogeneity.

Personalisation: Identify Endotypes via Latent Class Model



Alaa, Ahmed M., et al. "Personalized risk scoring for critical care prognosis using mixtures of Gaussian processes." *IEEE Transactions on Biomedical Engineering* (2017).

BRADFORD-HILL PRINCIPLES OF CAUSALITY

Plausibility

Consistency

Temporality

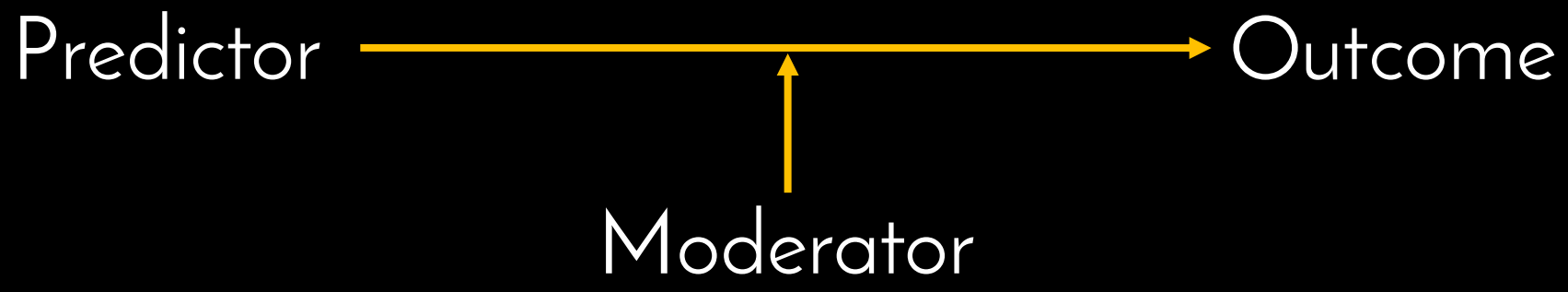
Strength

Specificity

Change in Risk Factor

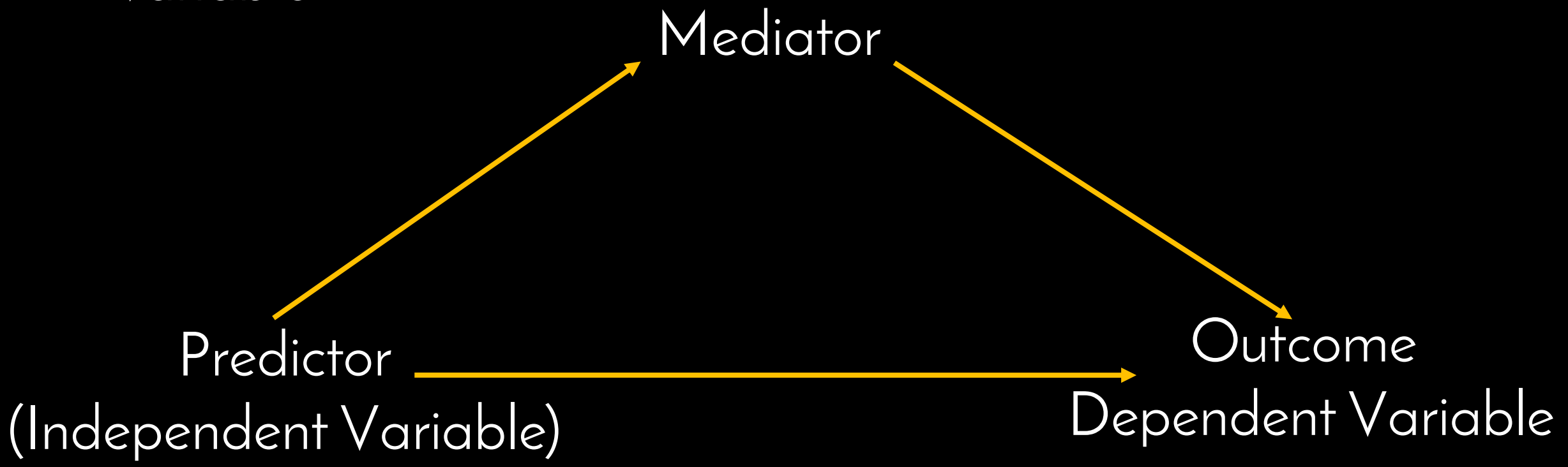
MODERATOR

A variable that **changes the impact** of one variable on another



MEDIATOR

A **mechanism** by which one variable affects another variable



TESTING MEDIATION

Step 1: Independent Variable → Dependent Variable

Step 2: Independent Variable → Mediator

Step 3: Mediator → Dependent Variable

Step 4: Effect of Independent Variable on Dependent Variable is significantly reduced by controlling for the mediator:

Sobel (1982) (http://www.unc.edu/~preacher/sobel/sobel.htm)

Goodman (1960) On the exact variance of products. *Journal of the American Statistical Association*, 55, 708-713.

INSTRUMENTAL VARIABLE (IV) ESTIMATION

Allows for consistent, unbiased estimation when the explanatory variables (covariates) are correlated with the error term in a regression model

Used to estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment

INSTRUMENTAL VARIABLE (IV) ESTIMATION

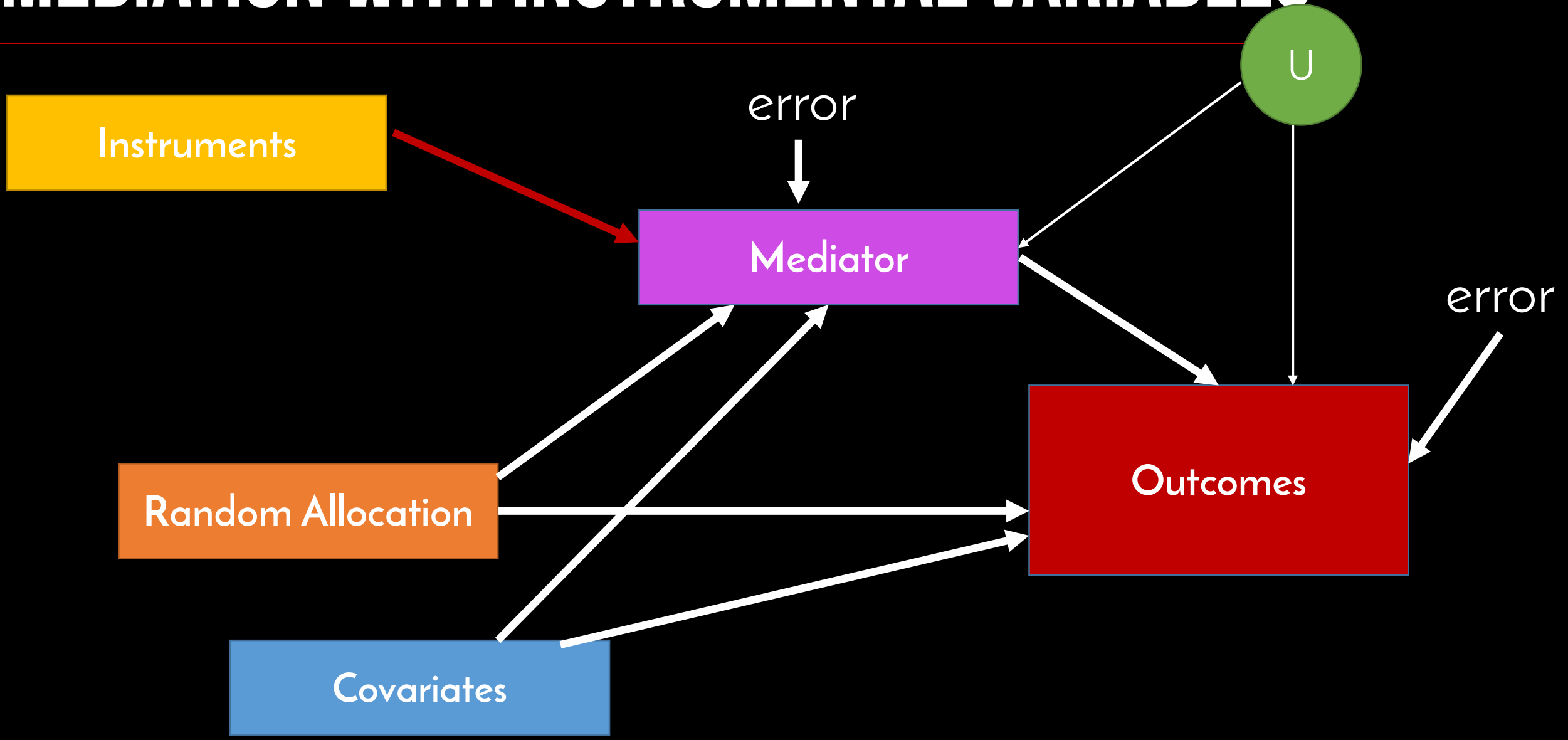
Scenarios:

- Change in the dependent variable change the value of at least one of the covariates (reverse causation)

- Omitted variables that affect both the dependent and independent variables

- Covariates are subject to measurement error

MEDIATION WITH INSTRUMENTAL VARIABLES



INSTRUMENTAL VARIABLE (IV) ESTIMATION

An instrumental variable is:

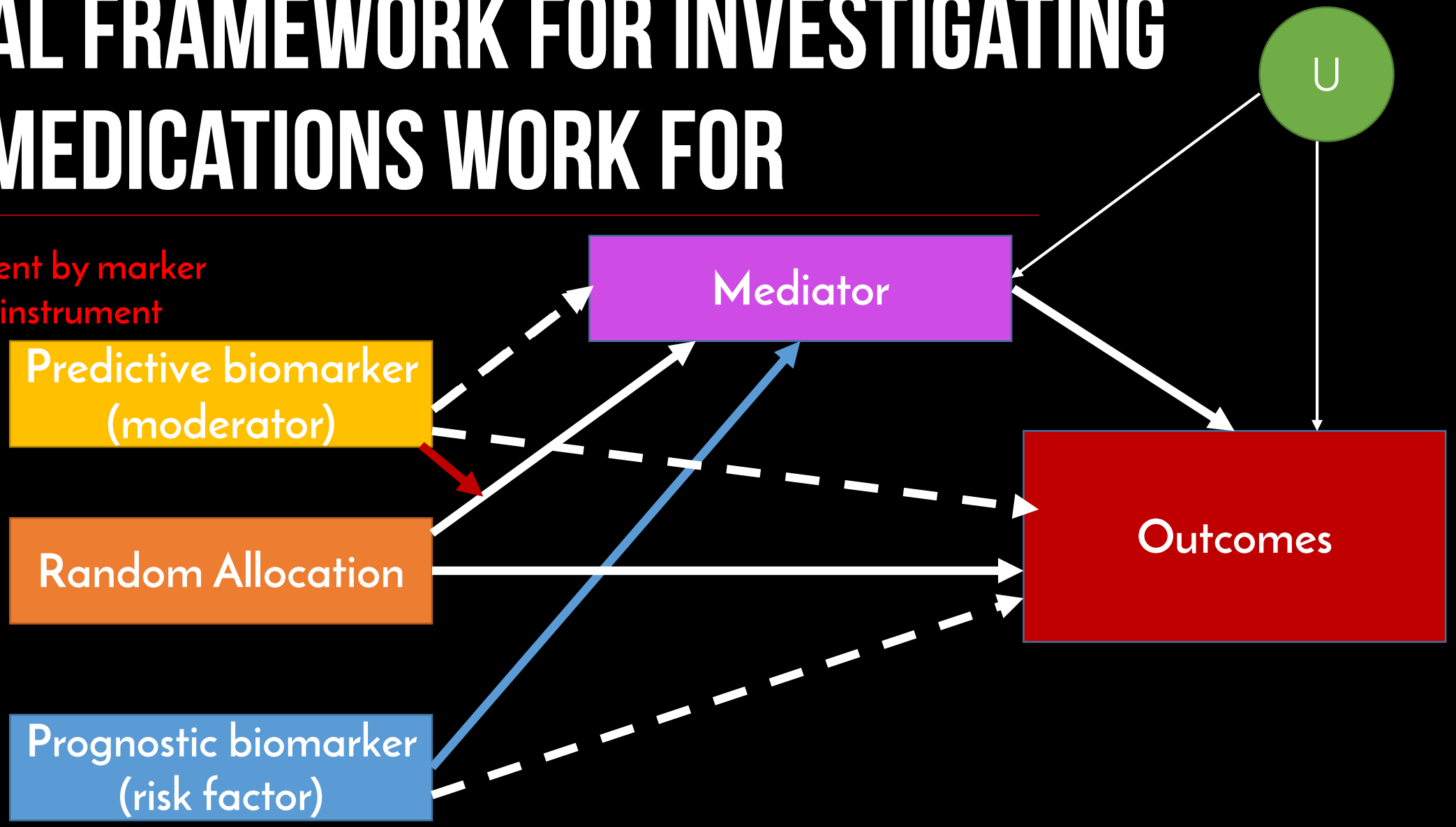
1. Strongly predictive of the mediating variable
2. Has no direct effect on the outcome except through the mediator
3. Does not share common causes with the outcome

Randomisation, where available, often satisfies this criteria when accounting for departures from randomised treatment.

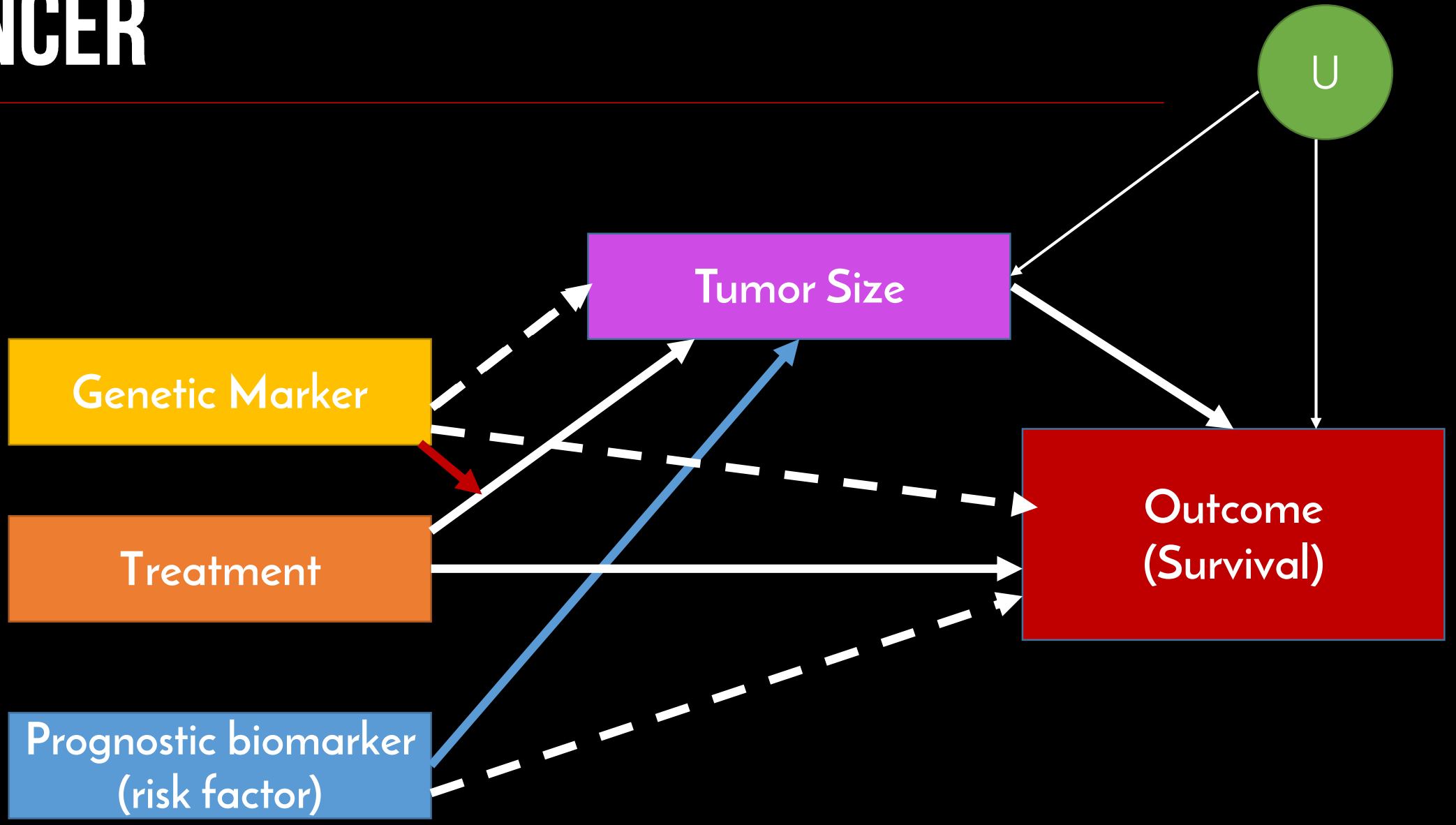
"Correlation and Causality" by David Kenny (1979)

EFFICACY AND MECHANISM EVALUATION: CAUSAL FRAMEWORK FOR INVESTIGATING WHO MEDICATIONS WORK FOR

Using the treatment by marker interaction as an instrument



EFFICACY AND MECHANISM EVALUATION: CANCER

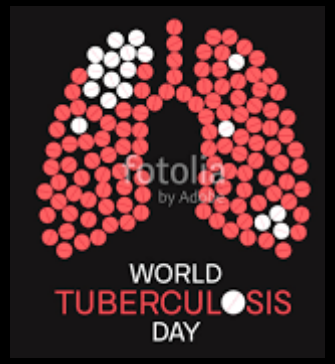


ML IN HEALTH: THERE IS STILL A LOT THAT NEEDS TO BE DONE...

"There is less attention paid to the more immediate problem of how we prevent these programs from amplifying the inequalities of our past and affecting the most vulnerable members of our society. "

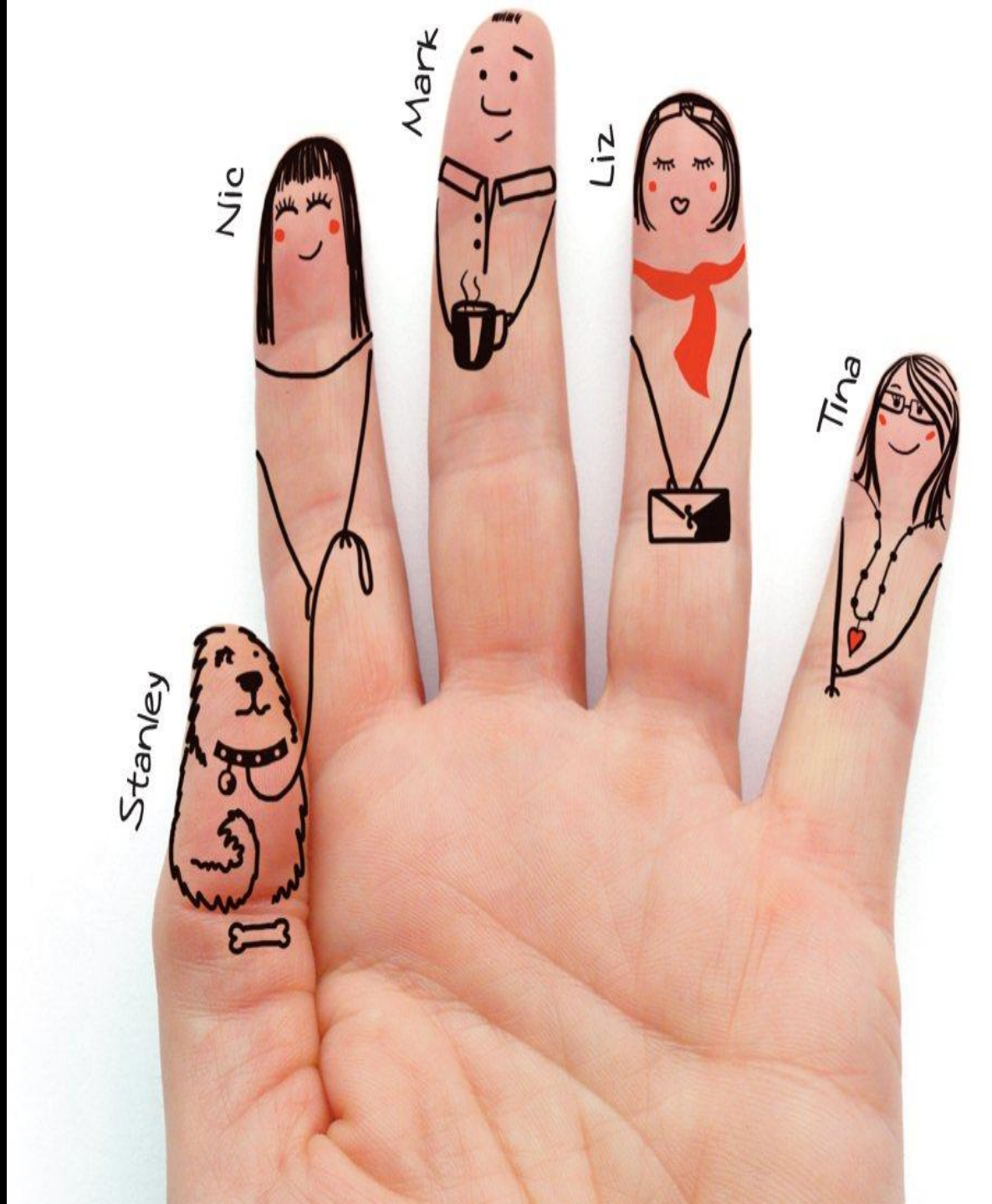
https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses

ML IN HEALTH: THERE IS STILL A LOT THAT NEEDS TO BE DONE...



The key to collaboration is effective communication

REFLECTIONS ON TEAM SCIENCE



Belgrave et al. Disaggregating asthma: Big investigation versus big data. *Journal of Allergy and Clinical Immunology* 139.2 (2017): 400-407.

Think deeply about the clinical context. Find solutions which are specific to the problem.

CONTEXT MATTERS

Good science is about merging different schools of thought for developing the bigger picture.

Data driven approach + Domain Knowledge = Holistic Approach to science

REFLECTIONS ON TEAM SCIENCE

Belgrave, Danielle, Angela Simpson, and Adnan Custovic. "Challenges in interpreting wheeze phenotypes: the clinical implications of statistical learning techniques." (2014): 121-123.

Principled epidemiology +
Biostatistics +
Machine Learning
= Heuristic Blend of Tools for understanding causality and clinical relevance

REFLECTIONS ON TEAM SCIENCE

Belgrave, Danielle, and Adnan Custovic. "The importance of being earnest in epidemiology." Acta Paediatrica 105.12 (2016): 1384-1386..

FROM INFORMATION TO KNOWLEDGE

1. **Team Science:** Discoveries about healthcare, **not hypothesised** a priori, have been made by experts explaining **structure** learned from **data** by algorithms tuned by those **experts**

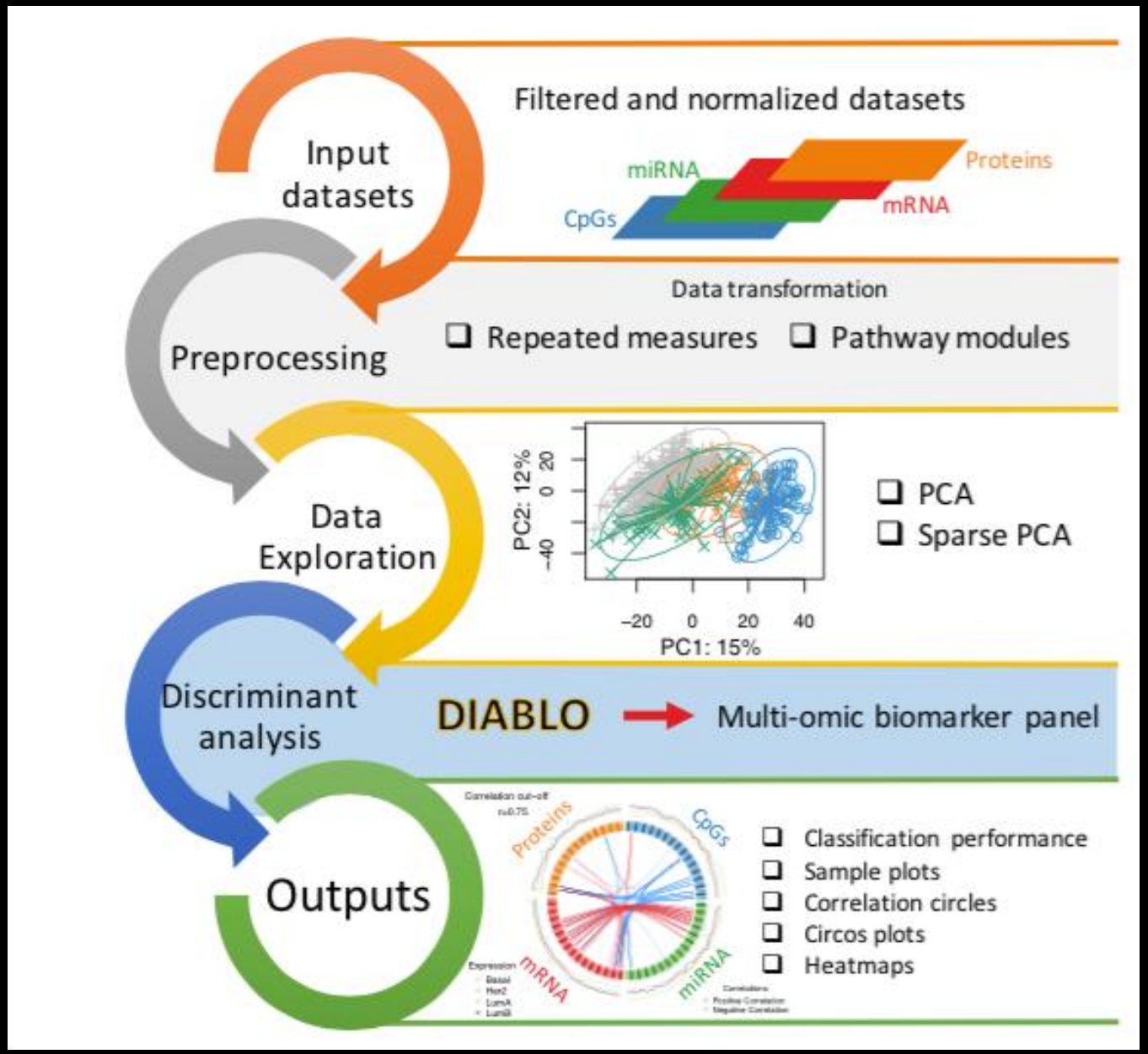
2. Heuristic blend of **biostatistics** and **machine-learning** reveals more than either method individually

3. An ML approach to extracting knowledge from information in healthcare requires persistent integration of
   Data
   Methods
   Expertise

Thank You

Deep Learning Indaba!

THE ROAD AHEAD...


Input datasets
Filtered and normalized datasets
miRNA
Proteins
CpGs
mRNA
Preprocessing
Data transformation
Repeated measures
Pathway modules
Data Exploration
PC2: 12%
PC1: 15%
PCA
Sparse PCA
Discriminant analysis
DIABLO
Multi-omic biomarker panel
Outputs
Proteins
CpGs
mRNA
miRNA
Classification performance
Sample plots
Correlation circles
Circos plots
Heatmaps

APPROXIMATING TRANSITION STATES AND CLASS MEMBERSHIP

Assumptions:

- Children in the same class have similar transitions of symptoms over time

```
public ClusterSimpleChain(int numYears)
    {
        …

        probState0 = Variable.Array<double>(k).Named("probState0");
        probState0Prior = Variable.Array<Beta>(k).Named("probState0Prior");
        probState0[k] = Variable<double>.Random(probState0Prior[k]);

        for (int y = 0; y < numYears; y++)
        {
…
#if clusterQ
            Q_T[y] = Variable.Array(Variable.Array<double>(s), k).Named("Q_T" + y);
            Q_F[y] = Variable.Array(Variable.Array<double>(s), k).Named("Q_F" + y);
            QTPriorArr[y] = Variable.Array(Variable.Array<Beta>(s),
k).Named("QTPriorArr" + y);
            QFPriorArr[y] = Variable.Array(Variable.Array<Beta>(s),
k).Named("QFPriorArr" + y);
            Q_T[y][k][s] = Variable<double>.Random(QTPrior[y][k][s]);
            Q_F[y][k][s] = Variable<double>.Random(QFPrior[y][k][s]);
#else
            Q_T[y] = Variable.Array<double>(s).Named("Q_T" + y);
            Q_F[y] = Variable.Array<double>(s).Named("Q_F" + y);
            QTPriorArr[y] = Variable.Array<Beta>(s).Named("QTPriorArr" + y);
            QFPriorArr[y] = Variable.Array<Beta>(s).Named("QFPriorArr" + y);
            Q_T[y][s] = Variable<double>.Random(QTPriorArr[y][s]);
            Q_F[y][s] = Variable<double>.Random(QFPriorArr[y][s]);
#endif
…
        }
```