# Statistical Challenges of Digital Twins

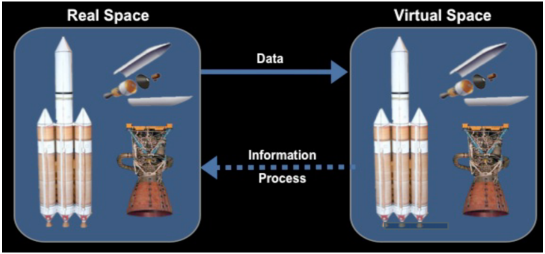
Richard Wilkinson

University of Nottingham

EPSRC
Engineering and Physical Sciences Research Council

# Digital twins

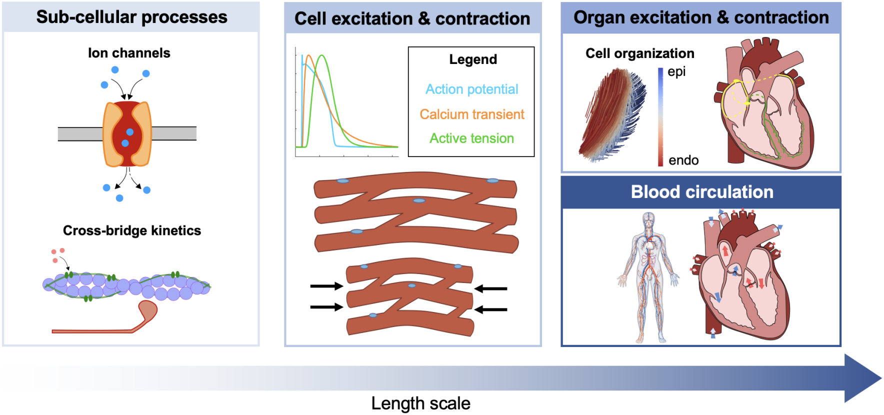
*~~A set of virtual information constructs that mimics the structure, context and behaviour of an individual or unique physical asset, that is dynamically updated with data from its physical twin throughout its life-cycle that informs decisions that realise value.~~*




A model of an individual, informed by data, that influences decisions.

# Motivating example: Cardiac physiology

Figures by Marina Strocchi, Steve Niederer, Richard Clayton



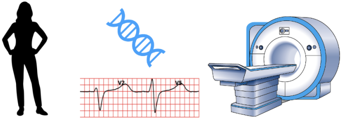
## Guidelines

Current treatment guidelines rely on statistics from **large and heterogeneous** patient groups

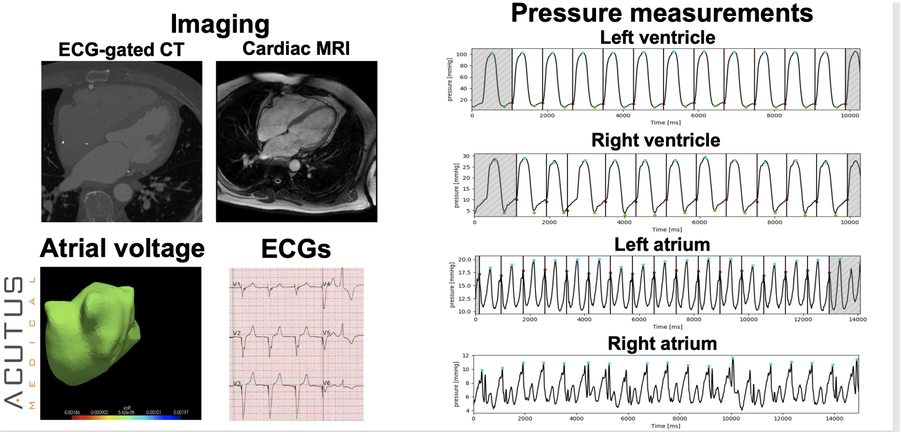
**vs**

## Precision medicine

Genetic information and data from **one individual patient** are analyzed to decide the best course of treatment

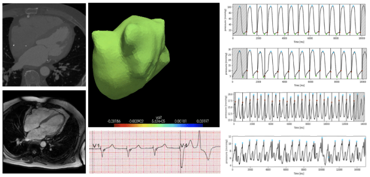


# Clinical data



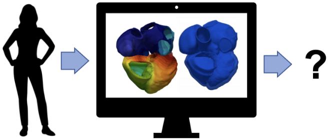
# Cardiac Digital Twins

**Clinical data from a specific patient**

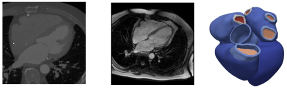


How do we **analyze** and combine **all this information?**

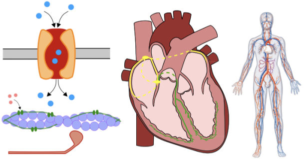
**Cardiac digital twin**




**Patient-specific anatomical model**



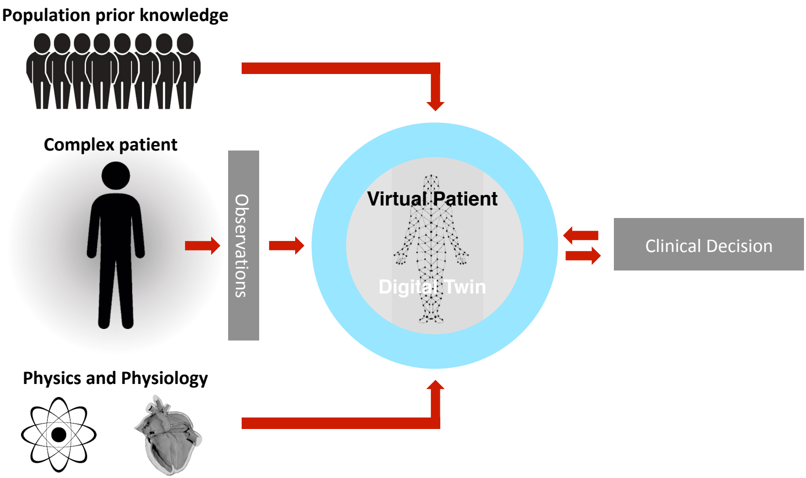
**Multi-scale heart model**



**Global sensitivity analysis**
Important parameters identification

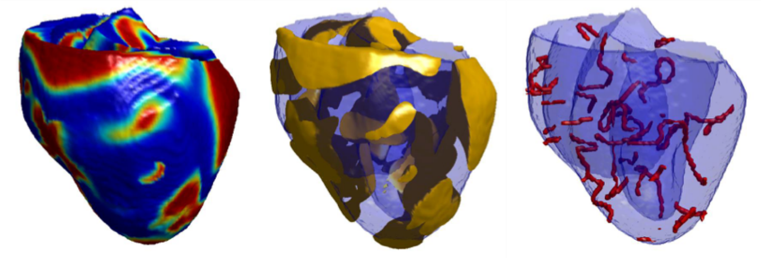
**Parameter fitting**
To replicate patient's clinical data

# Cardiac digital twin



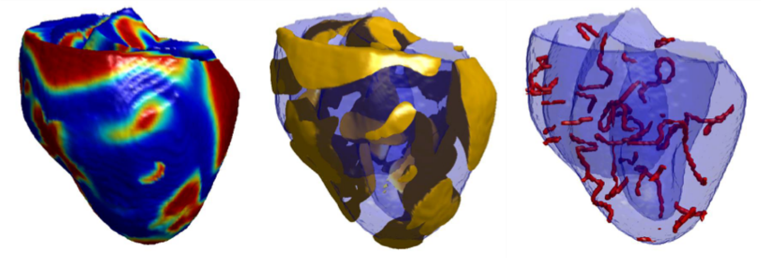

But how **confident** are we in our **prediction**

# Atrial fibrillation



Atrial fibrillation (AF) - rapid and uncoordinated electrical activation (arrhythmia) leading to poor mechanical function.

# Atrial fibrillation



Atrial fibrillation (AF) - rapid and uncoordinated electrical activation (arrhythmia) leading to poor mechanical function.

- Affects around 1,000,000 people in UK.
- Catheter ablation removes/isolates pathological tissue that sustain/initiate AF.
- 40% of patients subsequently experience atrial tachycardia (AT).

# Patient Specific Cardiac Models

Aim: predict whether an AF patient will develop AT following ablation, infer the reentry pathways, and then guide the surgical ablation to treat for both in a single procedure.

- Each intervention: 6% risk of major complication; cost ∼£8000.

# Patient Specific Cardiac Models

Aim: predict whether an AF patient will develop AT following ablation, infer the reentry pathways, and then guide the surgical ablation to treat for both in a single procedure.

- Each intervention: 6% risk of major complication; cost ∼£8000.

Cardiac models at forefront of personalised modelling

- Models are deterministic but clinical diagnosis is rarely definitive
  - uncertainty quantification/statistics challenge
- aim to consider costs and benefits across all potential outcomes weighted by their probability.

# Patient Specific Cardiac Models

Aim: predict whether an AF patient will develop AT following ablation, infer the reentry pathways, and then guide the surgical ablation to treat for both in a single procedure.

- Each intervention: 6% risk of major complication; cost ∼£8000.

Cardiac models at forefront of personalised modelling

- Models are deterministic but clinical diagnosis is rarely definitive
  - uncertainty quantification/statistics challenge
- aim to consider costs and benefits across all potential outcomes weighted by their probability.

# Statistical challenges

For a given patient, we want to select a model from our class of models $f(\theta, \omega)$ where

- $\omega$ are directly observable parameters specific to the patient such as geometry (ie for the computational mesh)
- $\theta$ are patient specific model parameters, eg diffusion parameters, which may be spatially varying ($\theta(x)$ for $x \in \omega$).

# Statistical challenges

For a given patient, we want to select a model from our class of models $f(\theta, \omega)$ where

- $\omega$ are directly observable parameters specific to the patient such as geometry (ie for the computational mesh)
- $\theta$ are patient specific model parameters, eg diffusion parameters, which may be spatially varying ($\theta(x)$ for $x \in \omega$).

Given data $D$ we want to solve the inverse problem

$$D = f(\theta, \omega) + e$$

to estimate

$$\pi(\theta, \omega \mid D) \propto \pi(\theta, \omega)\pi(D \mid \theta, \omega)$$

# Statistical challenges

In practice we need to be pragmatic

- Complex simulator and limited computational resource
- Large number of unknowns $\theta, \omega, f$
- Sparse noisy data
- Misspecification/discrepancy

# Statistical challenges

In practice we need to be pragmatic

- Complex simulator and limited computational resource
- Large number of unknowns $\theta, \omega, f$
- Sparse noisy data
- Misspecification/discrepancy

$$\mathbb{P}(\text{Event}|D) = \int \mathbb{P}(E|\theta,\omega,f)\pi(\theta,\omega,f|D)d\theta d\omega df$$

where

$$\pi(\theta,\omega,f|D) \propto \pi(D|\theta,\omega,f)\pi(\theta)\pi(\omega)\pi(f)$$

# Statistical challenges

In practice we need to be pragmatic

- Complex simulator and limited computational resource
- Large number of unknowns $\theta, \omega, f$
- Sparse noisy data
- Misspecification/discrepancy

$$\mathbb{P}(\text{Event}|D) = \int \mathbb{P}(E|\theta, \omega, f)\pi(\theta, \omega, f|D)d\theta d\omega df$$

where

$$\pi(\theta, \omega, f|D) \propto \pi(D|\theta, \omega, f)\pi(\theta)\pi(\omega)\pi(f)$$

We need to characterize variability at the

- population level $\pi(\theta), \pi(\omega)$ etc
- individual level $\pi(\theta, \omega, f, \ldots|D)$ – may need to be partially done in real time
- and the physics/simulator $\pi(D|\theta, \omega, f)$

# Surrogate models

If $f$ is slow/costly to evaluate standard methods such as MCMC are impracticable.

# Surrogate models

If $f$ is slow/costly to evaluate standard methods such as MCMC are impracticable. We can use surrogate models/emulators of $f$, e.g.

$$f(\cdot, \omega) \sim GP(m(\cdot), k(\cdot, \cdot))$$

which are trained on a small ensemble of simulator evaluations $C = \{\theta_i, f(\theta_i, \omega)\}_{i=1}^{n}$

- Currently run ∼1000 simulations for each new patient. Cost of £4-16k per patient.

We can then use the surrogate to estimate parameters etc

# Surrogate models

If $f$ is slow/costly to evaluate standard methods such as MCMC are impracticable. We can use surrogate models/emulators of $f$, e.g.

$$f(\cdot, \omega) \sim GP(m(\cdot), k(\cdot, \cdot))$$

which are trained on a small ensemble of simulator evaluations $C = \{\theta_i, f(\theta_i, \omega)\}_{i=1}^n$

- Currently run ∼1000 simulations for each new patient. Cost of £4-16k per patient.

We can then use the surrogate to estimate parameters etc
Note that this adds an additional uncertainty

$$\pi(f|C)$$

Other methods: NNs (e.g. PINNs), polynomial chaos, ROM, POD etc.

# Compact representation

If $\theta$ is high dimensional, we need to find a subset or transformation of the parameters $A\theta$ that we can estimate

- mesh used to simulate atrial electro-physiology has $\sim 30,000$ nodes, with 5 spatially varying parameters

# Compact representation

If $\theta$ is high dimensional, we need to find a subset or transformation of the parameters $A\theta$ that we can estimate

- mesh used to simulate atrial electro-physiology has $\sim 30,000$ nodes, with 5 spatially varying parameters

Typical methods

- Global sensitivity analysis: select a subset of the most important parameters (re contribution to variance).
- Basis expansions

$$\theta = \sum_{i=1}^{k} z_i \psi_i$$

  where $k \ll dim(\theta)$ and $\psi_i$ are basis vectors to be chosen
  - Imaging data, random projection, PCA/KL, active subspace methods...

# Non-identifiability

The huge number of parameters, sparse data, and limited computational power mean we can't hope to estimate everything.
How can we identify non-identifiabilities?

# Non-identifiability

The huge number of parameters, sparse data, and limited computational power mean we can't hope to estimate everything.
How can we identify non-identifiabilities?

- Difference between training and prediction tasks. We use data $D$

  $$D = h_1 f(\theta, \omega) + e$$

  to estimate $A\theta$.
  But suppose our prediction task is then

  $$h_2 f(\theta, \omega)$$

  How should we choose projection $A$?

# Fast and/or cheap inference

We want to calibrate in real time

- Catheter ablation: every additional 10mins of surgery increases stroke risk by x%

Even using a surrogate, MCMC is too expensive to use in-procedure, but we can pre-compute.

# Fast and/or cheap inference

We want to calibrate in real time

- Catheter ablation: every additional 10mins of surgery increases stroke risk by x%

Even using a surrogate, MCMC is too expensive to use in-procedure, but we can pre-compute.

Approximate inference methods

- Kalman inversion methods - estimate mean and variance of Gaussian approximation.
- Variational inference: instead of sampling, find variational approximation to the posterior

$$\arg\min_{\phi} KL(q_{\phi}(\theta)||p(\theta|D))$$

# Fast and/or cheap inference

- Amortized methods...

$$q(\theta|D) = N(m_\phi(D), s^2_\phi(D))$$

  where $m_\phi$ and $s^2_\phi$ are pre-trained neural networks.
- Neural posteriors. Eg use a normalizing flow to model

$$q(\theta \mid D)$$

  directly.

# Scalable DTs

At the moment, we create a new surrogate model for each new patient, e.g. estimating $\omega$ from imaging data

$$f(\cdot, \omega) \sim GP(m(\cdot), k(\cdot, \cdot)) \text{ trained with } C = \{\theta_i, f(\theta_i, \omega)\}_{i=1}^{n}$$

How can we reduce this cost?

# Scalable DTs

At the moment, we create a new surrogate model for each new patient, e.g. estimating $\omega$ from imaging data

$$f(\cdot, \omega) \sim GP(m(\cdot), k(\cdot, \cdot)) \text{ trained with } C = \{\theta_i, f(\theta_i, \omega)\}_{i=1}^n$$

How can we reduce this cost?

- Learn a statistical shape model $\omega = \sum z_i \phi_i$ e.g. via PCA and include $z$ in the inputs to the surrogate.
- Learn the discrepancy from a set of reference heart simulations to the new heart
$$f(\cdot, \omega') = f(\cdot, \omega^r) + \delta(\cdot)$$
- Learn diffeomorphism: hearts are topologically equivalent. If $\omega' = T\omega^r$, can we learn a $T'$ from $T$ such that $f(\cdot, \omega') = T' f(\cdot, \omega^r)$?

# Networked Digital Twins

Suppose we have DTs of 1000s of patients.

- How we we learn informative priors?
- How do we transfer knowledge through the network?
- How do we cheaply initialize new twins?

# Physics informed models

How can we incorporate relatively simple physics into data-models?

# Physics informed models

How can we incorporate relatively simple physics into data-models?
Suppose we want to infer forcing function $g$ in the system

$$\mathcal{L}u = g \text{ given observations } z_i = \langle h_i, u \rangle + e \qquad i = 1, \ldots, n$$

for example by solving constrained optimization problem

$$\min_g (z - Hu)^\top (z - Hu) \text{ subject to } \mathcal{L}u = g$$

# Physics informed models

How can we incorporate relatively simple physics into data-models?
Suppose we want to infer forcing function $g$ in the system

$$\mathcal{L}u = g \text{ given observations } z_i = \langle h_i, u \rangle + e \qquad i = 1, \ldots, n$$

for example by solving constrained optimization problem

$$\min_g (z - Hu)^\top (z - Hu) \text{ subject to } \mathcal{L}u = g$$

Introduce $n$ adjoint systems

$$\mathcal{L}^* v_i = h_i$$

then

$$\langle h_i, u \rangle = \langle \mathcal{L}^* v_i, u \rangle = \langle v_i, \mathcal{L}u \rangle = \langle v_i, g \rangle$$

If $g$ is a linear model (e.g. a RFF expansion of a GP) we can now do exact inference for $g$ at zero additional cost.

# Other topics

- Geometric uncertainty
  - Heart is never still, segmentation of MRI/CT image imperfect, images are obtained in unnatural situations.
  - Data are collected from an uncertain geometric location.
  - Need manifold valued models etc.
- Model discrepancy
  - How can we use the network of DTs to learn the model error?
- Multi-fidelity/multi-level methods
  - If we have models $f_1, f_2, \ldots,$ of varying costs and accuracies, how do we make the most accurate predictions we can within some given computational budget?

# Conclusions

Digital twins provide a fundable framework to work on many of the key mathematical/statistical challenges arising in UQ.

# Conclusions

Digital twins provide a fundable framework to work on many of the key mathematical/statistical challenges arising in UQ.

- At present, DTs aren't used to guide therapy.
  - We can currently build DTs for a single patient, but at great expense
  - Need to scale and speed up this process
- The huge number of uncertain parameters and cost of the simulations will mean we need to compromise:
  - find regularities in the problem to allow us to reduce dimension sufficiently in order to make inference possible
  - learn strong population structured prior distributions
  - develop fast method to approximately infer parameters.

# Conclusions

Digital twins provide a fundable framework to work on many of the key mathematical/statistical challenges arising in UQ.

- At present, DTs aren't used to guide therapy.
  - We can currently build DTs for a single patient, but at great expense
  - Need to scale and speed up this process
- The huge number of uncertain parameters and cost of the simulations will mean we need to compromise:
  - find regularities in the problem to allow us to reduce dimension sufficiently in order to make inference possible
  - learn strong population structured prior distributions
  - develop fast method to approximately infer parameters.

Thank you for listening!