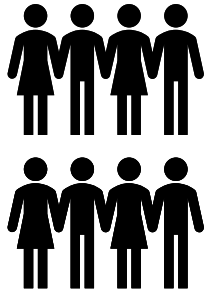




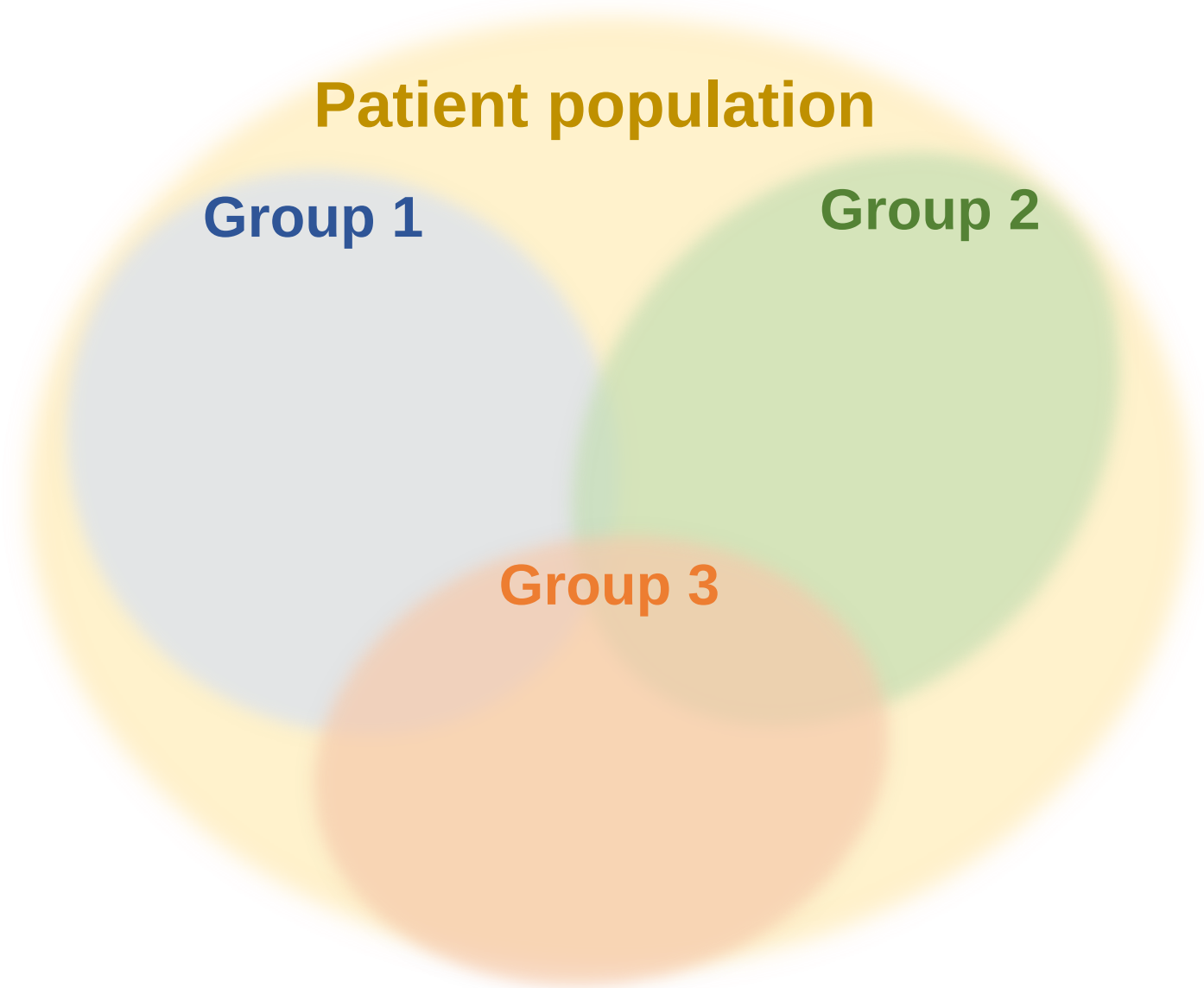
A Deep Variational Approach to Clustering Survival Data

Laura Manduchi[†], Ričards Marcinkevičs[†], Michela C. Massi, Thomas Weikert,
Alexander Sauter, Verena Gotta, Timothy Müller, Flavio Vasella, Marian Neidert,
Marc Pfister, Bram Stieltjes, Julia E. Vogt

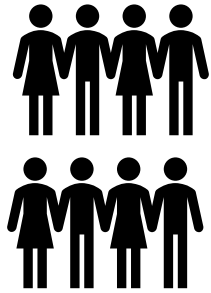
Survival Clustering: An Example



- Age
 - Treatment A
 - Treatment B
 - Lab results
 - ...
 - Time-to-event, t
- } x



Survival Clustering: An Example



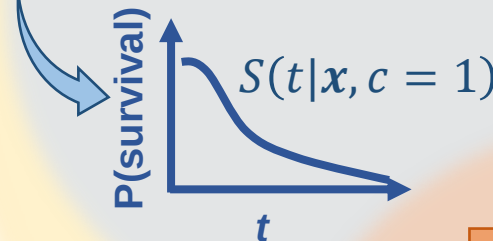
- Age
- Treatment A
- Treatment B
- Lab results
- ...
- Time-to-event, t

- Discover **groups** of patients characterised by **different associations** between the covariates x and survival time t
- Stratify patients by both risk and disease phenotype
- A more *personalised disease management*

Patient population

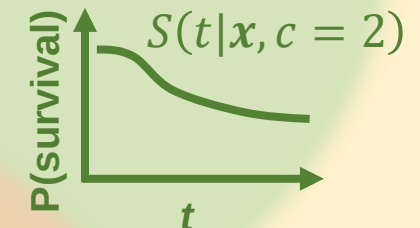
Group 1

Age
Treatment A
Treatment B
Lab results



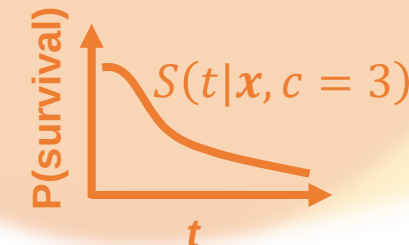
Group 2

Age
Treatment A
Treatment B
Lab results



Group 3

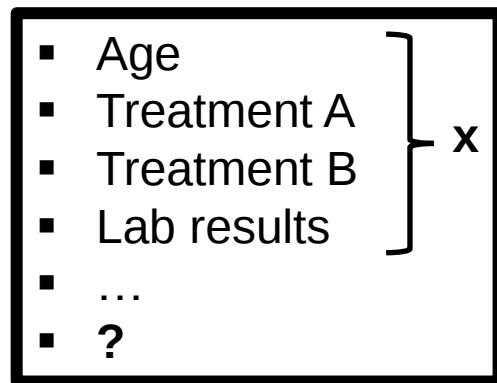
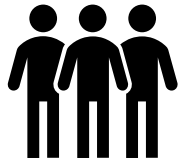
Age
Treatment A
Treatment B
Lab results



Survival Clustering: An Example

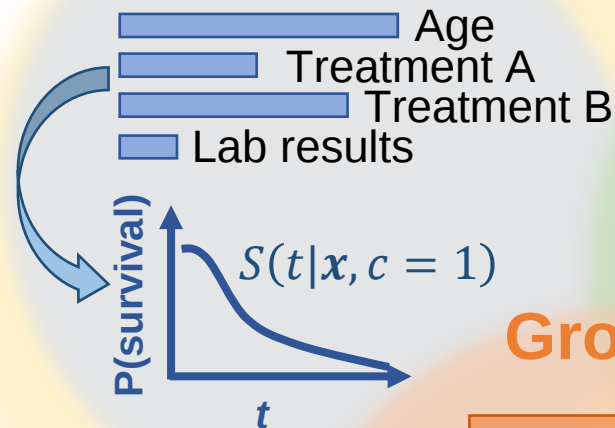
Some **observations**:

- $\neq$ unsupervised clustering
- survival time t as side information, not available at test time

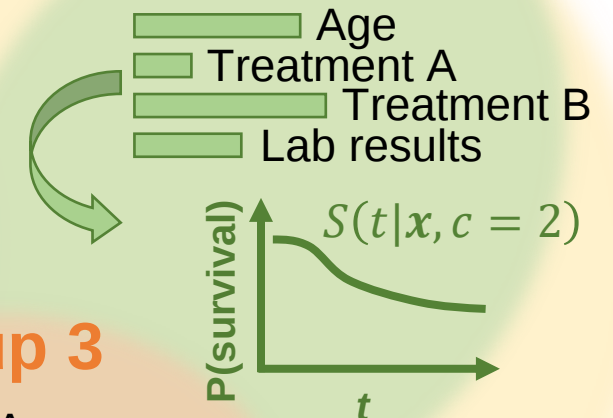


Patient population

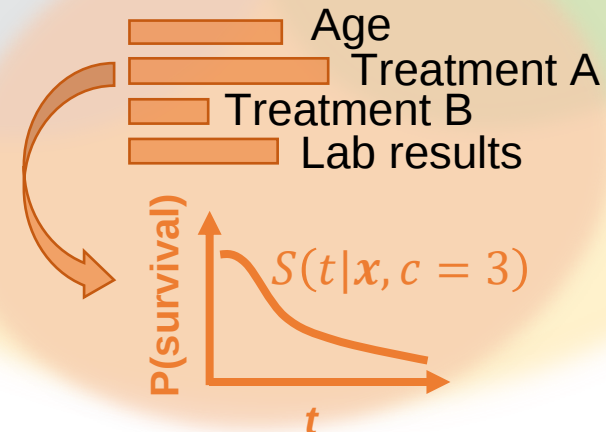
Group 1



Group 2



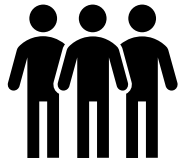
Group 3



Survival Clustering: An Example

Some **observations**:

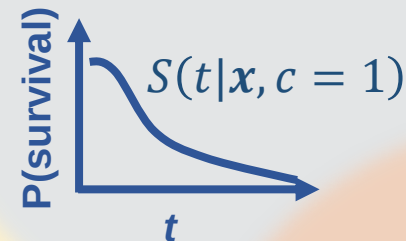
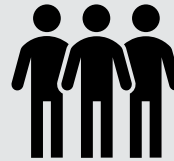
- $\neq$ unsupervised clustering
- survival time t as side information, not available at test time



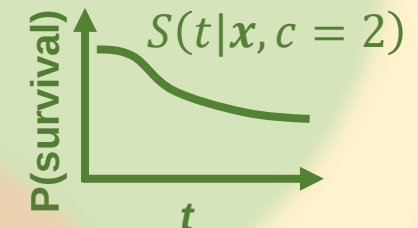
▪ Age	}	x
▪ Treatment A		
▪ Treatment B		
▪ Lab results		
▪ ...		
▪ ?		

Patient population

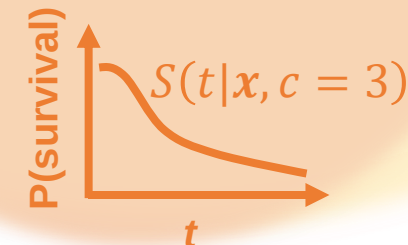
Group 1



Group 2



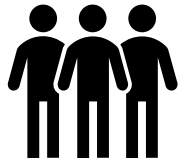
Group 3



Survival Clustering: An Example

Some **observations**:

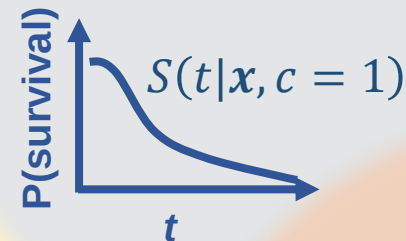
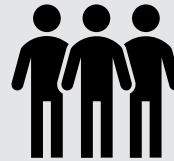
- $\neq$ unsupervised clustering
- survival time t as side information, not available at test time



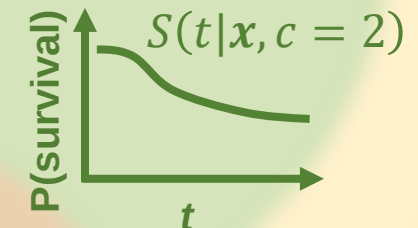
▪ Age ▪ Treatment A ▪ Treatment B ▪ Lab results ▪ ... ▪ Predicted time-to-event	x
--	-----

Patient population

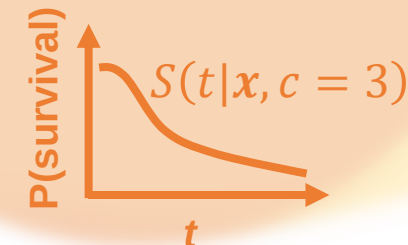
Group 1



Group 2



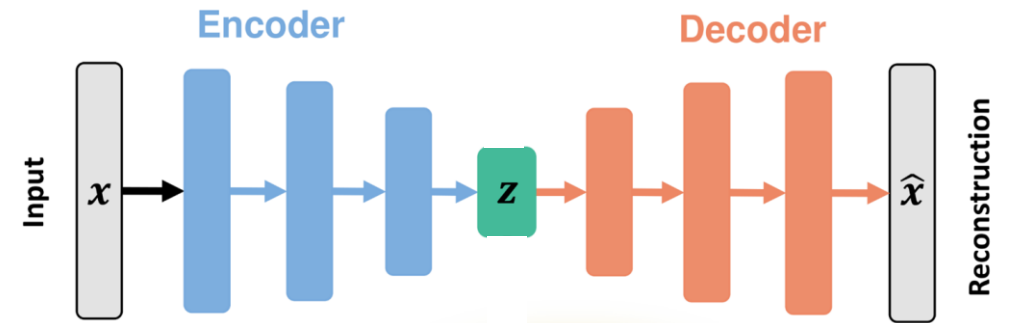
Group 3



Variational Deep Survival Clustering (VaDeSC)

Overview of our model:

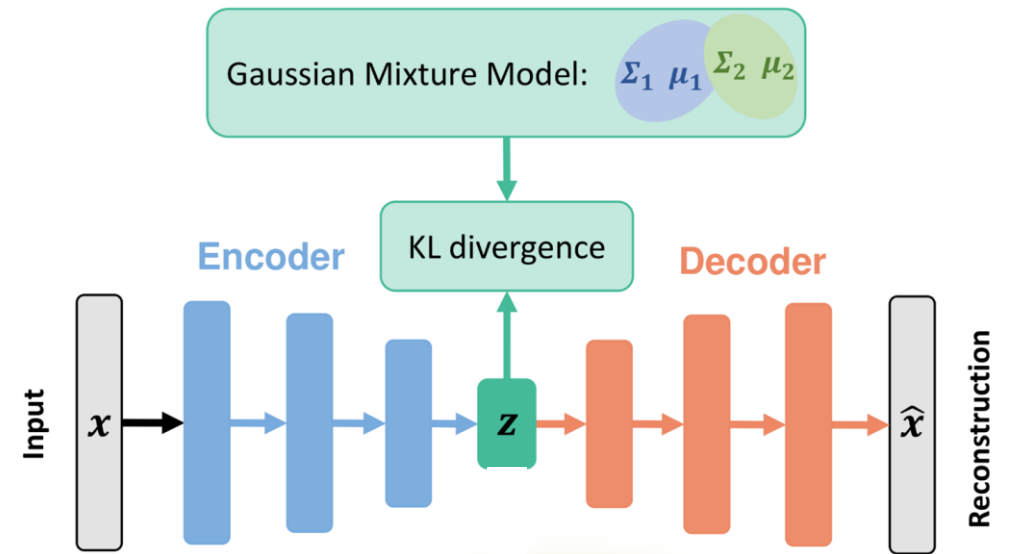
- based on a **variational autoencoder**



Variational Deep Survival Clustering (VaDeSC)

Overview of our model:

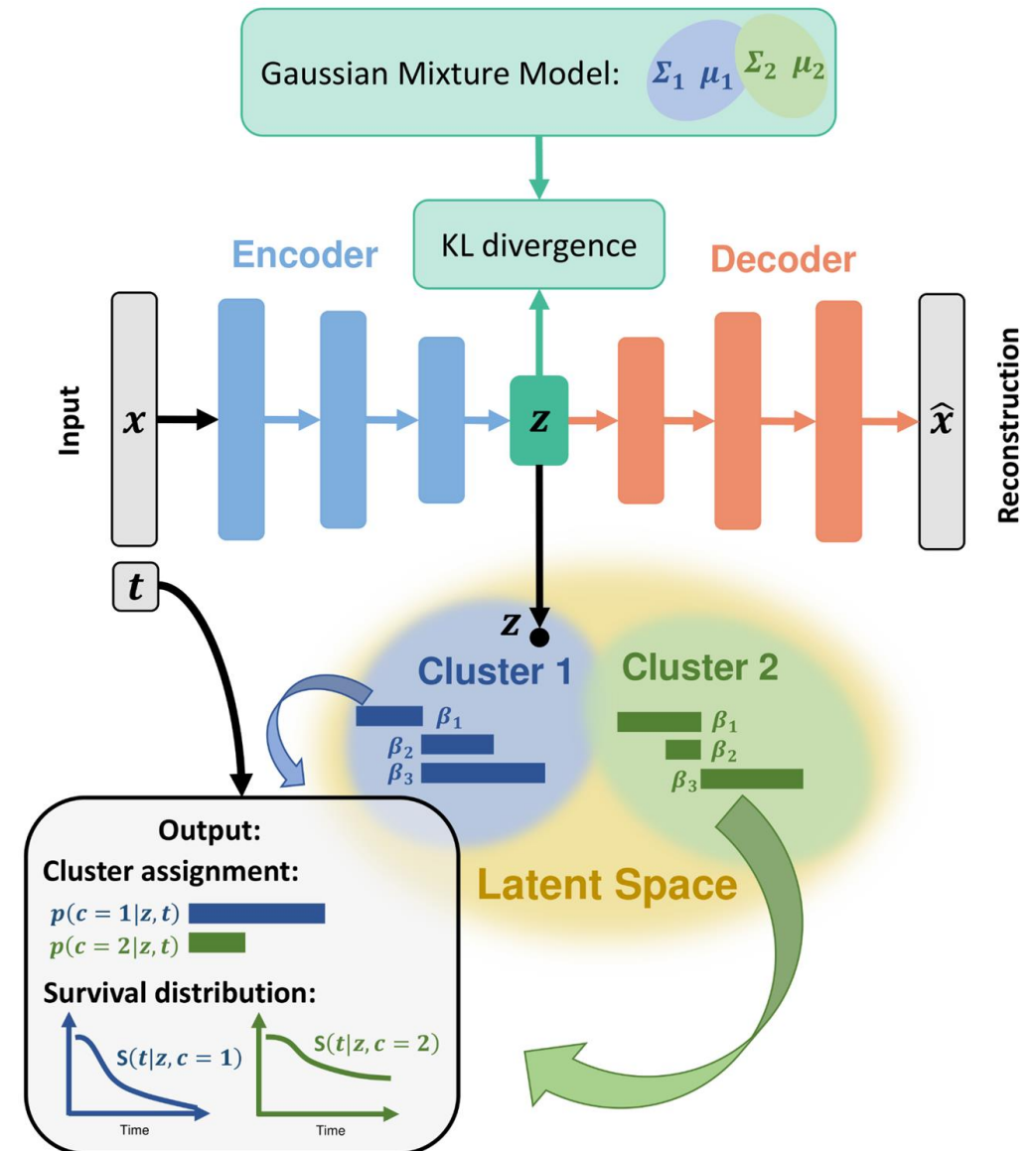
- based on a **variational autoencoder**
 - **Gaussian mixture** prior
 - clustering in the embedding space



Variational Deep Survival Clustering (VaDeSC)

Overview of our model:

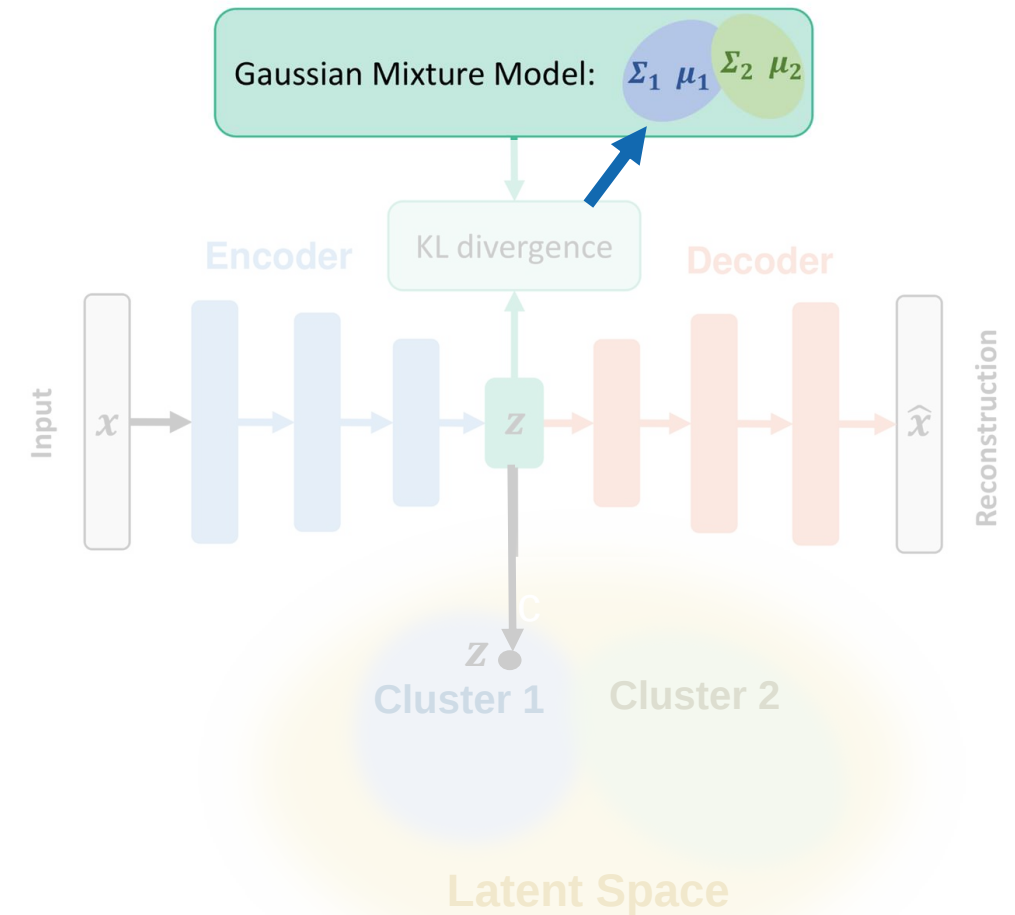
- based on a **variational autoencoder**
 - **Gaussian mixture** prior
 - clustering in the embedding space
- mixture of **survival distributions**
 - defined in the embedding space
 - **Weibull** with cluster-specific scales



VaDeSC: Generative Model

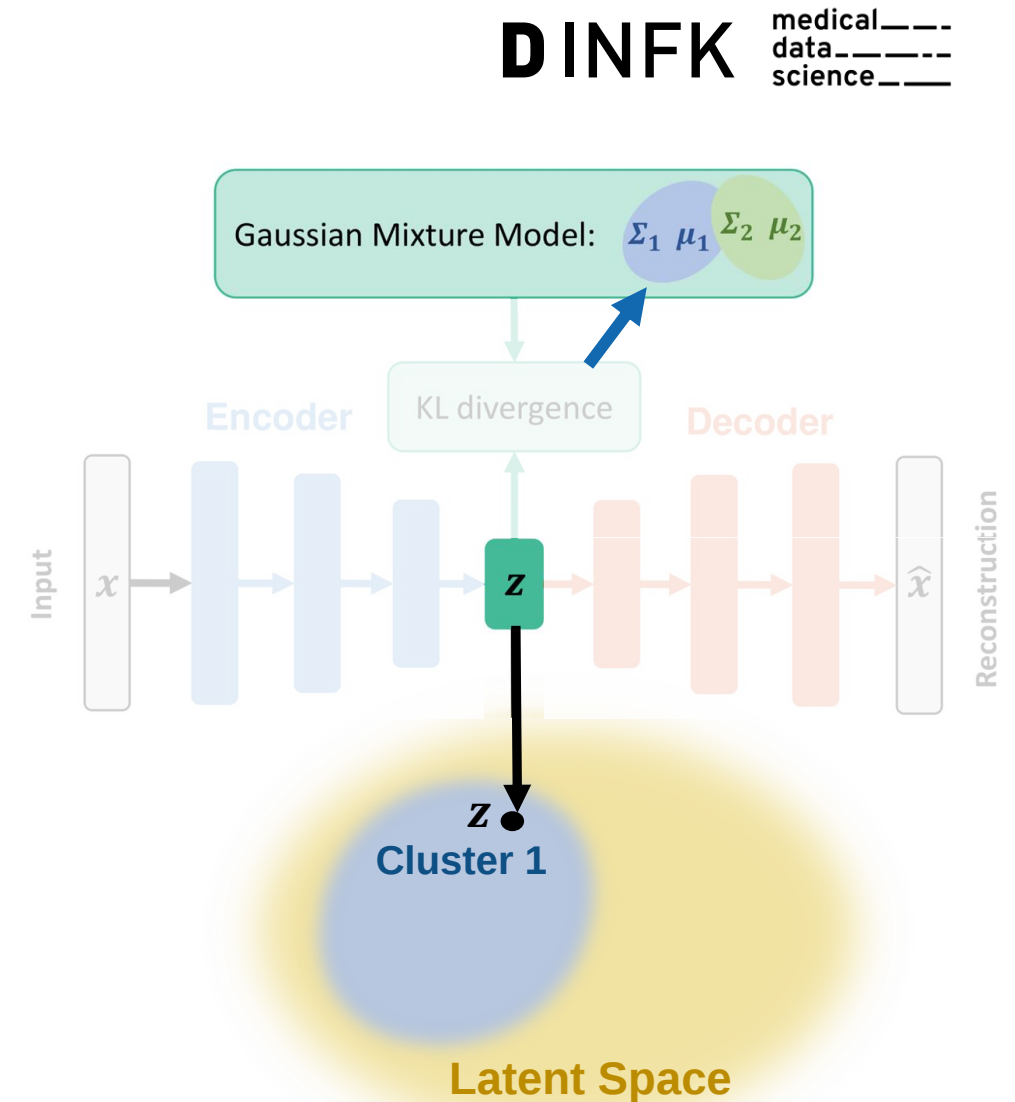
DINFK medical data science

1. Sample a **cluster**: $c \sim p(c; \boldsymbol{\pi}) = \pi_c$



VaDeSC: Generative Model

1. Sample a **cluster**: $c \sim p(c; \boldsymbol{\pi}) = \pi_c$
2. Sample the **latent embedding**
 $\mathbf{z} \sim p(\mathbf{z}|c; \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}, \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\})$
 $= \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$



VaDeSC: Generative Model

1. Sample a **cluster**: $c \sim p(c; \boldsymbol{\pi}) = \pi_c$

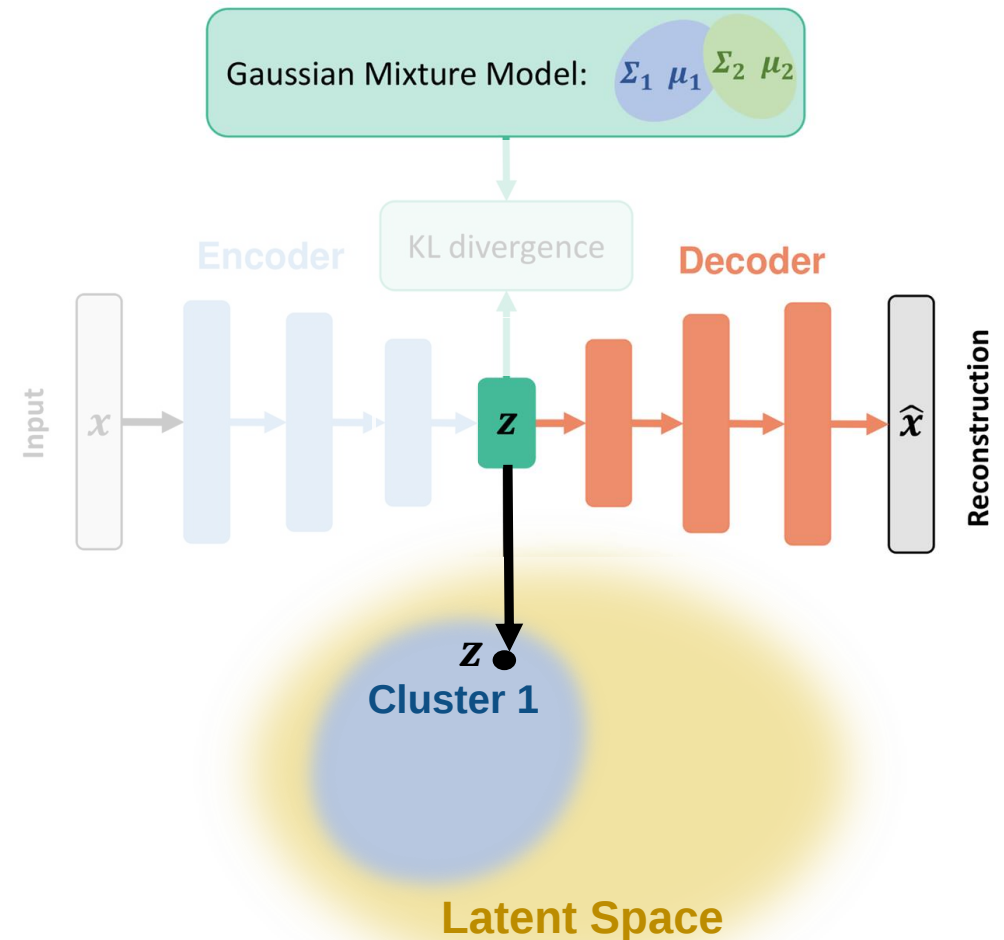
2. Sample the **latent embedding**

$$\mathbf{z} \sim p(\mathbf{z}|c; \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}, \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}) \\ = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

3. Sample the **explanatory variables**

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}), \text{ e.g. Bernoulli}(\mathbf{x}; \boldsymbol{\mu}_{\boldsymbol{\gamma}})$$

decoder output



VaDeSC: Generative Model

1. Sample a **cluster**: $c \sim p(c; \boldsymbol{\pi}) = \pi_c$

2. Sample the **latent embedding**

$$\mathbf{z} \sim p(\mathbf{z}|c; \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}, \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}) \\ = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

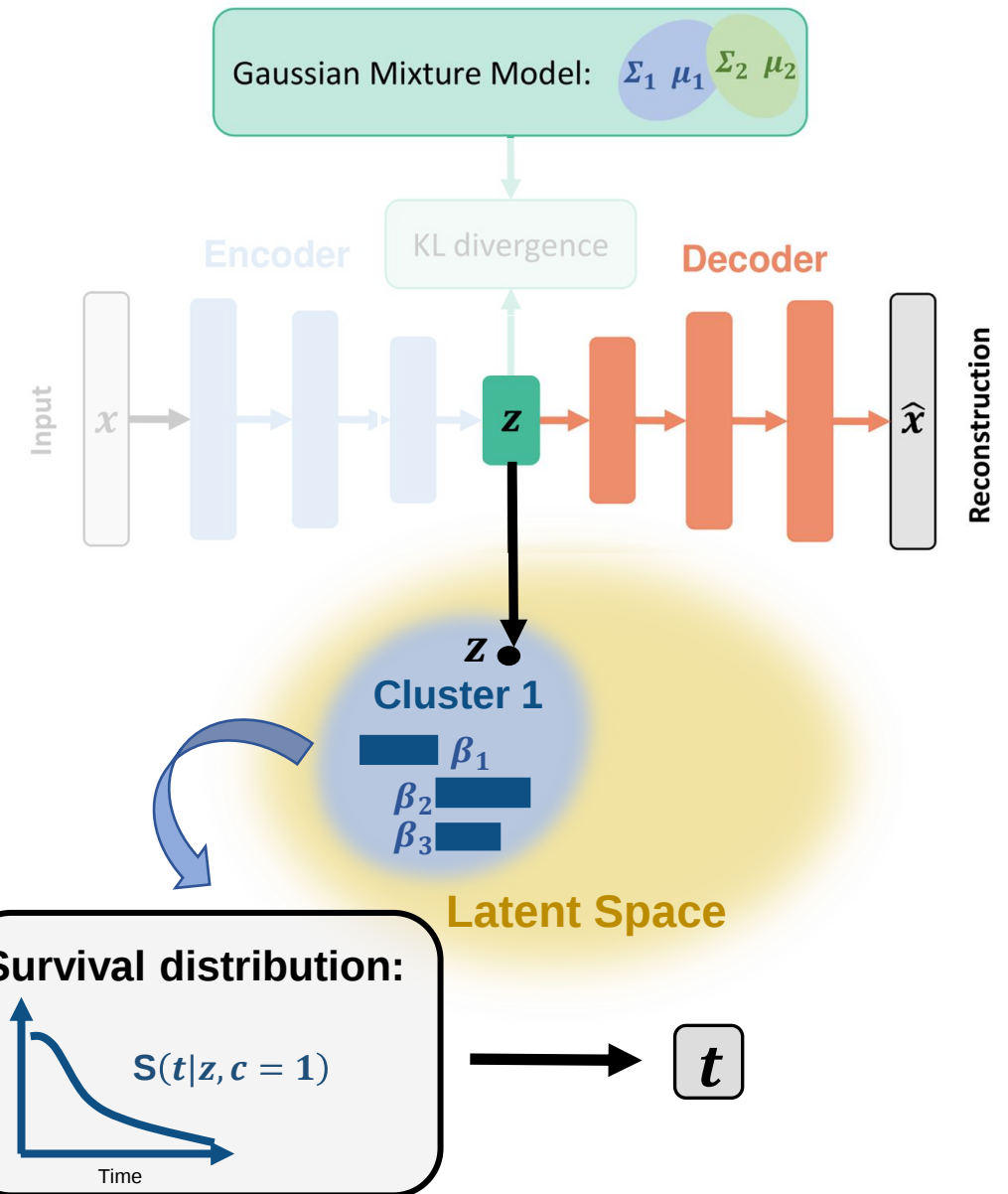
3. Sample the **explanatory variables**

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}), \text{ e.g. Bernoulli}(\mathbf{x}; \boldsymbol{\mu}_{\boldsymbol{\gamma}})$$

decoder output

4. Sample the **time-to-event**

$$t \sim p(t|\mathbf{z}, c; \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}, k)$$



1. Sample a **cluster**: $c \sim p(c; \boldsymbol{\pi}) = \pi_c$

2. Sample the **latent embedding**

$$\mathbf{z} \sim p(\mathbf{z}|c; \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}, \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}) \\ = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$



Parameters optimised by maximising a lower bound of the **likelihood** of $\{x_i, t_i\}_{i=1}^N$

3. Sample the **explanatory variables**

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{z}; \boldsymbol{\gamma}), \text{ e.g. Bernoulli}(\mathbf{x}; \boldsymbol{\mu}_{\boldsymbol{\gamma}})$$

decoder output

4. Sample the **time-to-event**

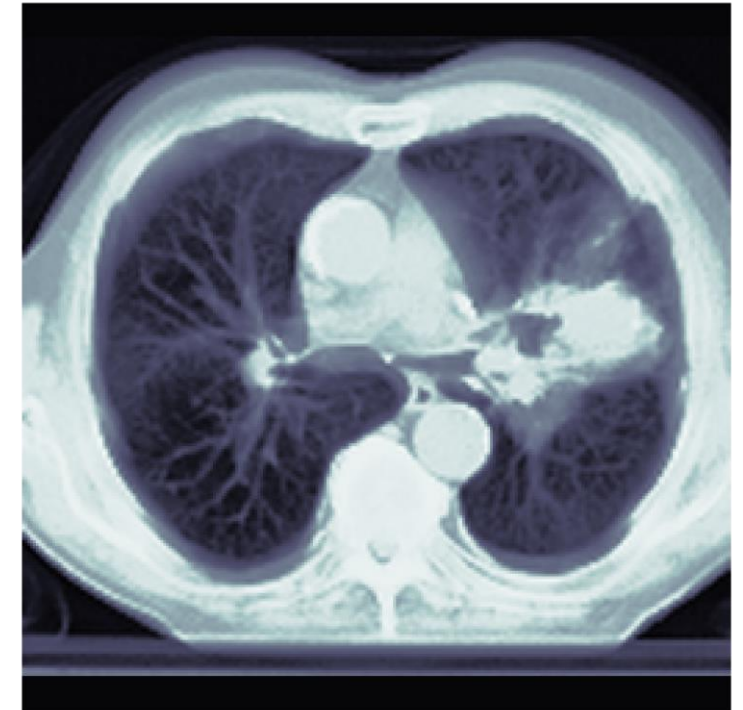
$$t \sim p(t|\mathbf{z}, c; \{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K\}, k)$$

Non-small Cell Lung Cancer

Non-small cell lung cancer (NSCLC): adeno-, squamous cell, large cell carcinomas, and other less differentiated types.

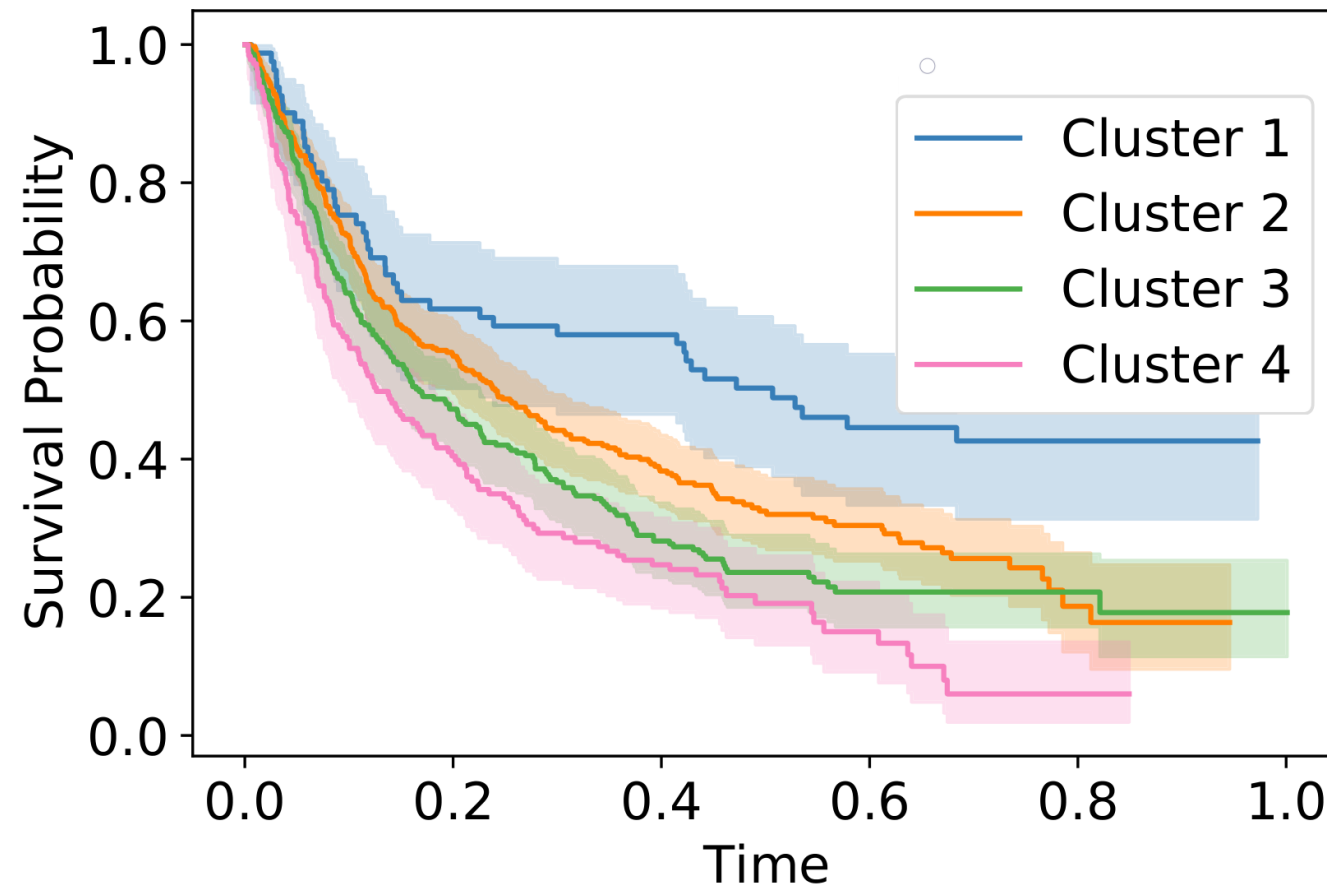
An **aggregated dataset** of **computed tomography (CT)** images:

- 392 patients at the University Hospital Basel, CH
- 511 patients at the Maastrro clinic, NL
- 211 patients at the Stanford University School of Medicine and Palo Alto Veterans Affairs Healthcare System, US

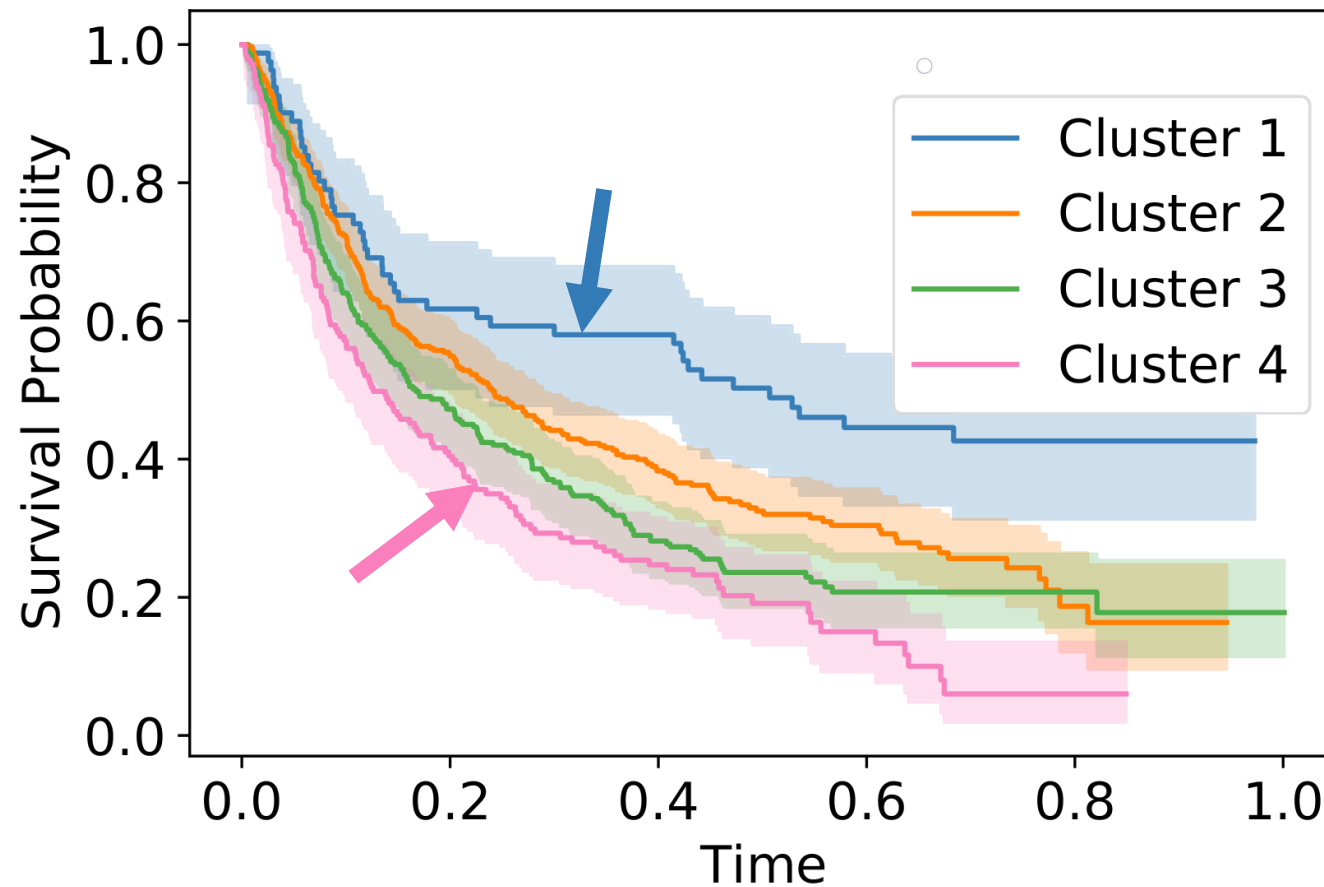


*a preprocessed CT slice
centred around the tumour*

NSCLC: Clustering

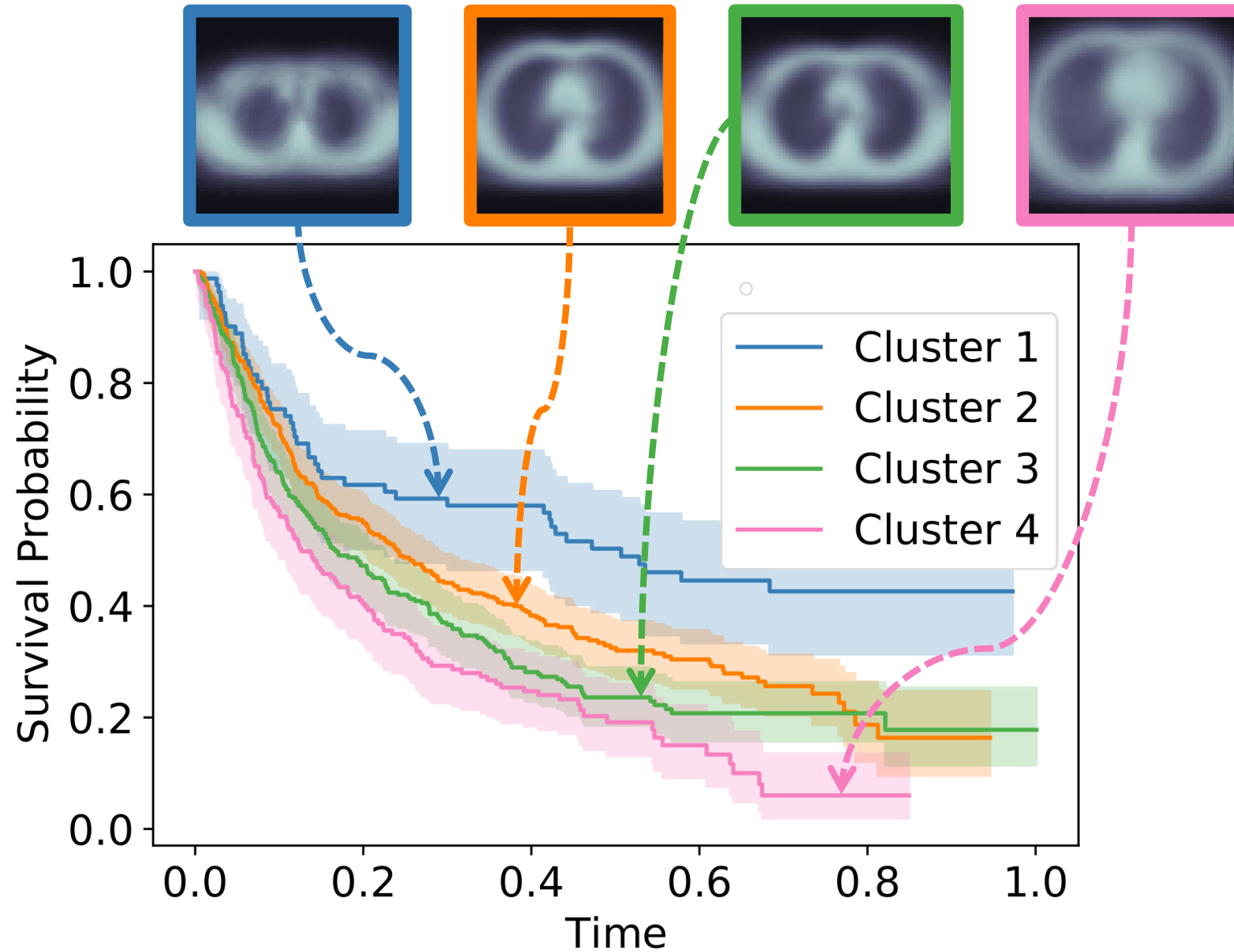


NSCLC: Clustering



NSCLC: Clustering

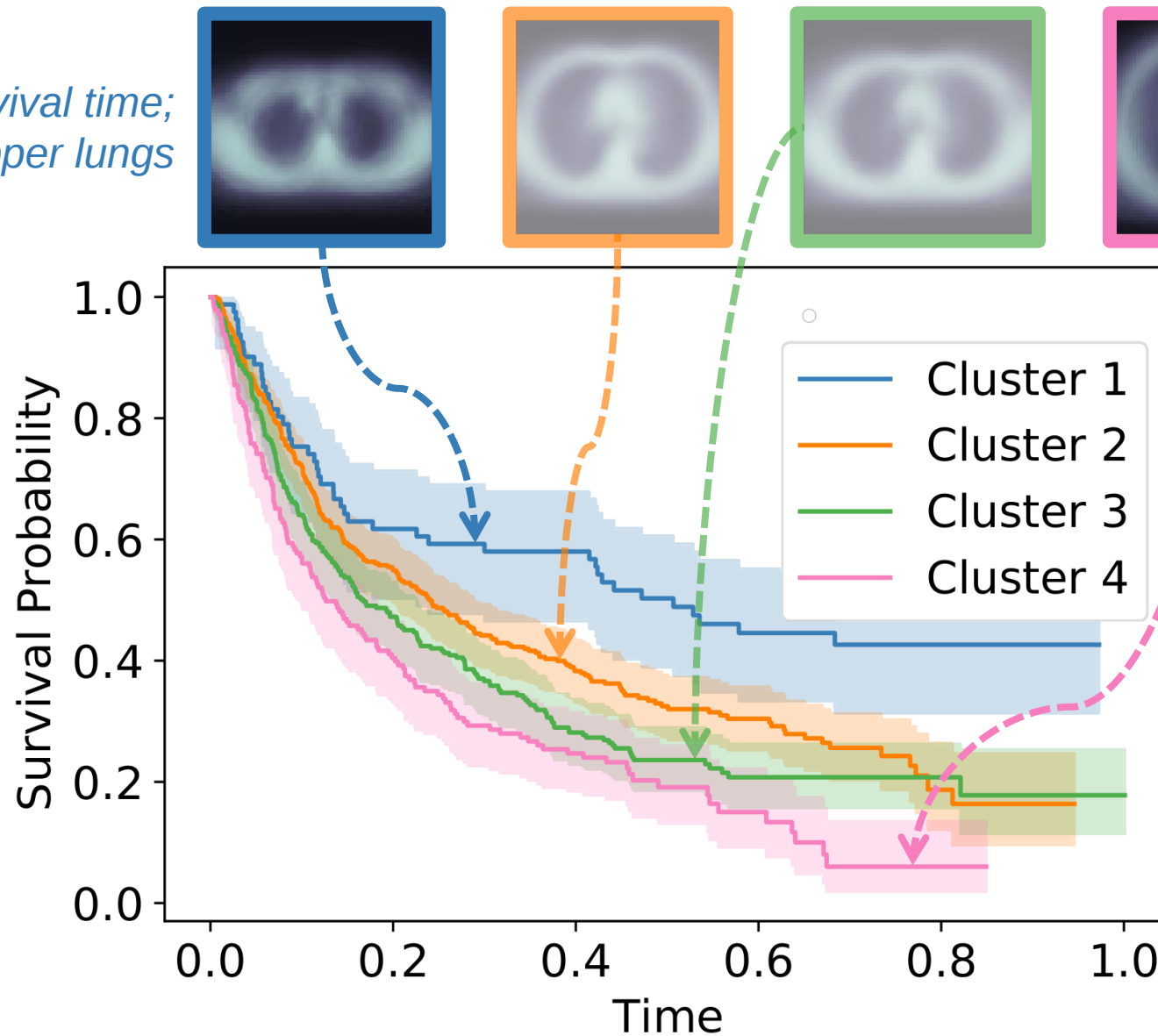
DINFK medical data science



NSCLC: Clustering

DINFK medical____
data_____
science____

*highest median survival time;
tumour in the upper lungs*

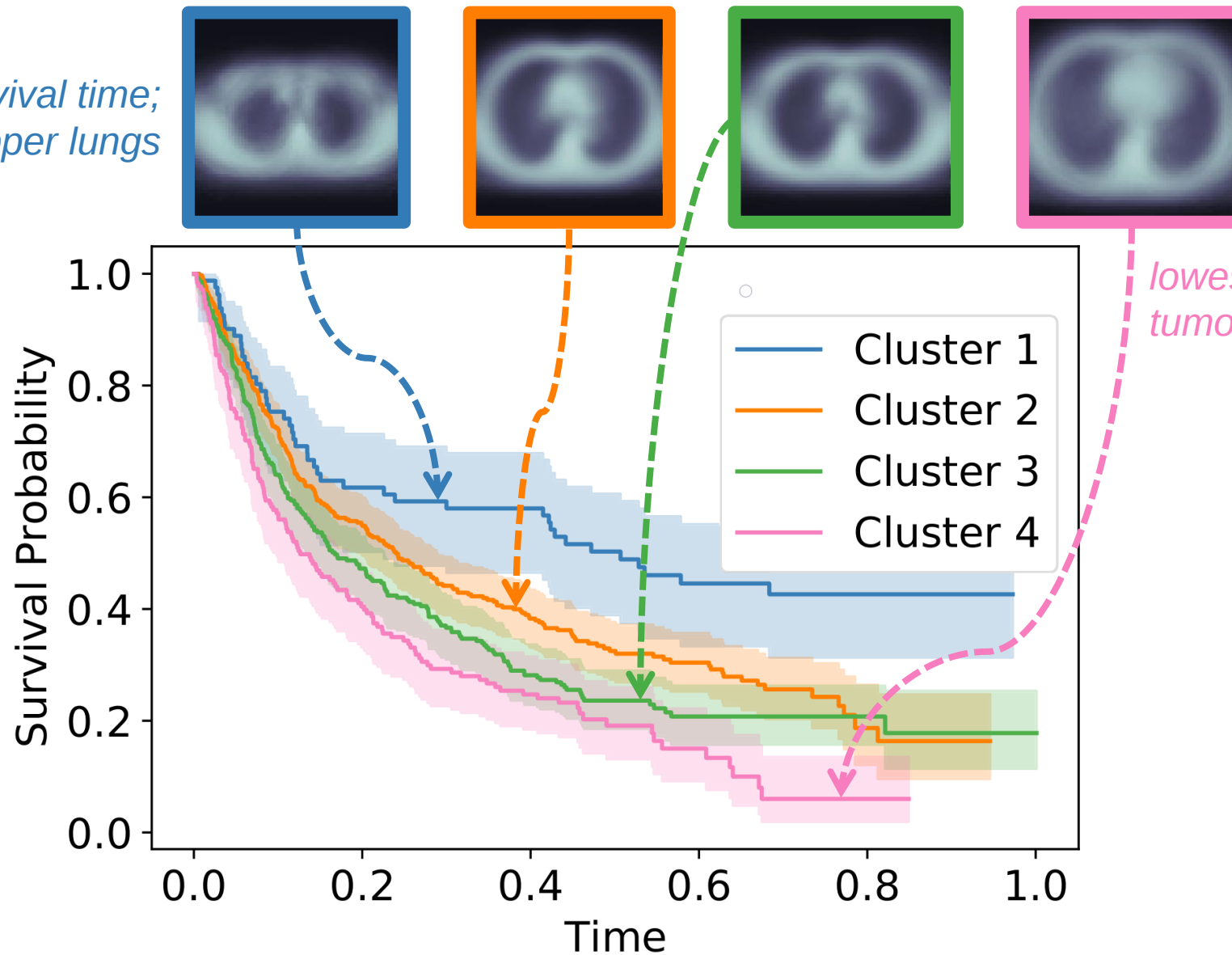


*lowest median survival time;
tumour in the lower lungs*

NSCLC: Clustering

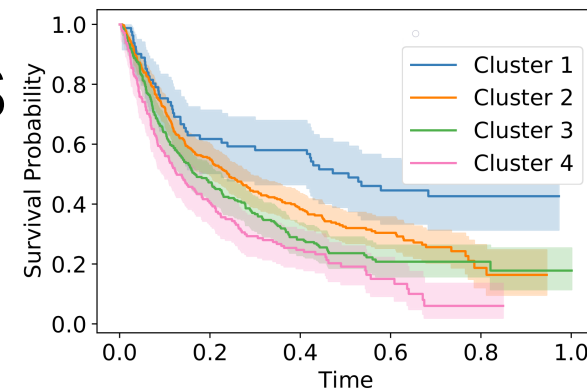
DINFK medical____
data_____
science____

*highest median survival time;
tumour in the upper lungs*



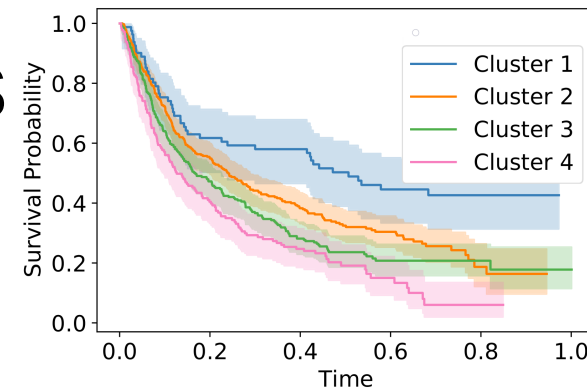
*lowest median survival time;
tumour in the lower lungs*

NSCLC: Demographic & Clinical Variables



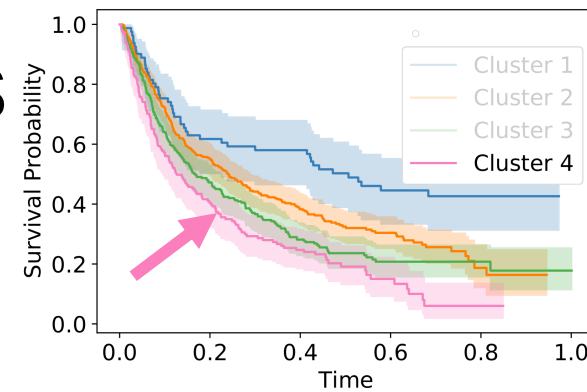
Variable					<i>p</i> -val.
	1	2	3	4	
T. Vol., cm³	43	36	40	63	$\leq 5e-2$
Age, yrs	62	69	67	70	$\leq 1e-3$
Female, %	36	19	38	23	$\leq 1e-3$
Smoker, %	67	94	87	100	0.12
M1, %	20	45	44	45	0.2
$\geq$ T3, %	10	29	35	31	0.7

NSCLC: Demographic & Clinical Variables



Variable					<i>p</i> -val.
	1	2	3	4	
T. Vol., cm³	43	36	40	63	$\leq 5e-2$
Age, yrs	62	69	67	70	$\leq 1e-3$
Female, %	36	19	38	23	$\leq 1e-3$
Smoker, %	67	94	87	100	0.12
M1, %	20	45	44	45	0.2
$\geq$ T3, %	10	29	35	31	0.7

NSCLC: Demographic & Clinical Variables



Variable	1	2	3	4	<i>p</i> -val.
T. Vol., cm³	43	36	40	63	$\leq 5e-2$
Age, yrs	62	69	67	70	$\leq 1e-3$
Female, %	36	19	38	23	$\leq 1e-3$
Smoker, %	67	94	87	100	0.12
M1, %	20	45	44	45	0.2
<u>$\geq$ T3, %</u>	10	29	35	31	0.7

Concluding Remarks

- A novel deep **probabilistic model** for clustering survival data
 - **Generative assumptions**
 - Improved **clustering performance**
 - Good for early **exploratory analysis** of complex datasets
- Application to **medical imaging** data
 - Prior research: **tabular & EHR data**
 - Meaningful findings



read on [OpenReview](#)



code on [GitHub](#)